Travel Time Forecasting

Ieva Bak

Lyngby 2007

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

Summary

The main objective of this thesis is the development of a forecast algorithm for short-term travel time forecasting. This algorithm is intended to become an inherent part of a new real-time traffic reporting system. This system is in the pipeline in the framework of the Danish Road Directorate. Practicability and operability are the central keywords that permeate every aspect of this thesis. Consequently great emphasis is placed on the data value chain from data collection and preparation of input data for the forecast algorithm to model deployment. The movement of data through a series of stages and the transformations that these data undergo in the process are illustrated. Data modeling is viewed as an intrinsic part of the complete data value chain with an end product in mind, combined with methods of heuristic nature. Insights into the nature of traffic data are provided by the use of clustering. The forecasting algorithm is subsequently based on the results of clustering of data. The developed algorithm is simple and its performance in terms of forecast accuracy is satisfactory. The result is considered to be superior to forecasting based on average travel times.

Preface

This thesis was prepared at Informatics and Mathematical Modeling, the Technical University of Denmark in partial fulfillment of the requirements for acquiring the master degree in engineering. This thesis was supervised by Bo Friis Nielsen at IMM and co-supervised by Jan Holm from the Road Directorate.

This thesis deals with the development of a forecast algorithm for real-time travel time forecasting and the establishment of a data value chain from data collection to model deployment in support of modeling. The main result is that clustering can be utilized in the context of travel time forecasting and that satisfactory results can be achieved by a relatively simple model.

A draft paper about real-time travel time forecasting based on the ideas presented in this thesis was submitted and accepted for presentation at the 14th World Congress on Intelligent Transport Systems in Beijing, China, October 2007.

Lyngby, July 2007

Ieva Bak

Contents

Chapter 1

Background information

The traffic flow on the motorway network spanning the Greater Copenhagen area is monitored by several hundred detectors and radars measuring vehicle speed and count. To date, traffic reports have been presented to the public on a website announcing traffic states (free flow, dense traffic, queuing) for each motorway segment [1]. A new traffic reporting system has been designed in order to improve the quality of the traffic reports. This system is intended to operate in real-time. After implementation the public will have access to the following information: three key tables showing long-term average travel times and speeds for the specific time of day, real-time travel times and speeds, 15-minute travel time and speed forecasts, respectively, for any pair of adjacent motorway exits covering the aforementioned motorway network. These values will be updated every minute. An excerpt from a key table showing real-time travel times and speeds for all adjacent motorway exits on Hillerødmotorvejen southbound is shown in Figure 1.1. Furthermore, the system will provide travel times between arbitrary motorway exits in the motorway network based on the real-time traffic situation. It can be used to forecast the travel time starting just now. The user selects a starting point A and a termination point B, after which the system calculates the expected travel time and speed on the selected route. A route is modeled as a sequence of segments between two adjacent motorway exits. The system calculates the travel time between motorway exits A and B as follows: for the first motorway segment on the route the real-time travel time is used. For the next segment on the route the real-time travel time,

the 15-minute forecast or the 30-minute forecast is used as segment travel time, depending on the running total travel time on the selected route. The travel times are accumulated along the motorway segments which make up the route. Depending on the accumulated travel time, the real-time, the 15-minute or the 30-minute forecast is used until the termination point B is reached. This process is sketched in Figure 1.2.

Historiske | Aktuelle | Forventet

Aktuel rejsetid - opdateret 16/5 16:12

Fra	Til	Distance Km.	Tid Min.	Hastighed Km/t
	Hillerødmotorvejen sydgående			
11 Allerød	10 Farum	1,6	1	109
10 Farum	8 Værløse	5,6	3	106
8 Værløse	7 Skovbrynet	3,0	2	99
7 Skovbrynet	X Gladsaxe	2,6	2	92
X Gladsaxe	1 Høje Gladsaxe	4,3	3	95

Figure 1.1: Key table - test version

Figure 1.2: User interface - test version

1.1 Scope and goal of the project

The central topic in this thesis is the development of a universal 15-minute travel time forecast algorithm that can be applied to each road segment between two motorway exits. If workable, the prepared forecast algorithm will be integrated

into the new traffic reporting system upon the completion of this thesis, after which this service will be released to the public. For this reason, great emphasis will be placed on the practical issues that arise when dealing with the development of a real-time large-scale application. Hence, there will be a trade-off between the complexity of the chosen methods of approach and the requirements, which working with real-time data in a large-scale application imposes on the possibilities of selecting the theoretically most desirable approach. This thesis will aim at ensuring compliance with the requirements outlined by the Road Directorate by the use of heuristic methods.

Chapter 2

Requirements specification

2.1 Functional and non-functional requirements

The development of a forecast model is subject to a number of functional and non-functional requirements, which need to be accounted for before, during and after the model building process. These requirements have informally been worked out by the Road Directorate in collaboration with the writer to ensure that the developed forecast algorithm complies with the proposed vision for the new traffic reporting system. Above all, the system requires that the forecasted travel times are reasonably trustworthy, and that they are reported to the public in real-time. Furthermore, the Road Directorate has expressed a desire that the preparation of input data, model building, model evaluation and model deployment is native to the database where the collected data will reside. The above requirements place a series of constraints on how the modeling process can be conducted, and what special purpose software tools should be used. The following paragraphs will outline these requirements and their particular details.

2.1.1 Service availability

The forecast function is required to have an excellent service availability. This means that the forecasted travel times must be considered reliable before they can be announced to the public. This has been defined as the maximum acceptable deviation that the forecasted travel times are allowed to have from the actual travel times in minutes. The maximum acceptable value is set to 5 minutes. This value has origins in the evaluation report prepared by COWI [2] of the 15-minute forecast model that was developed in connection with the extension of the M3 motorway [3]. The forecast functionality will be disabled if the difference between the actual and the forecasted travel times exceeds 5 minutes.

2.1.2 Computational considerations

The forecast algorithm is subject to the requirement that the computation time for the whole motorway network has to be containable to a few seconds. This is due to the fact that the forecasted travel times will be reported every minute.

2.1.3 Data preparation

This includes data cleaning and repair such as the handling of missing values, outliers and data transformation. In other terms, this step includes the creation of methods to process the collected data into a unified, consistent format before it is passed on as input to the forecasting algorithm.

2.1.4 Model building and evaluation

An inherent part of this process is choosing the software tool. The requirement states that this process needs to be native to the database where the collected data will reside. This means that, ideally, the process should be conducted in the database. This is due to the fact that the amount of modeling that needs to be done to create the forecast models for the whole motorway network is expected to be rather comprehensive given its size. The model building and evaluation process involves handling great amounts of data, which, in turn, would make the data extraction, transportation and loading process cumbersome and expensive if the input data had to be moved outside of the database.

2.1.5 Model deployment

This process involves the deployment of the prepared forecast models to the real-time application. It is a requirement that model deployment is native to the database to enable seamless model distribution without wasting resources on custom-designing a solution.

2.1.6 Recalibration

This process involves the maintenance and recalibration of the forecast models after deployment. This is necessary when the distribution of data has changed since the last time the models were built. It is a requirement that this process is automated, thereby minimizing manual work.

2.1.7 Interpretability

A capable person, but a non-expert, should be able to understand and conduct the previously mentioned steps once a routine for their execution has been defined and implemented.

2.2 Concluding remarks

The discussed requirements partially exclude the use of any prevalent stand alone software for model building. This is mainly due to the fact that this would require the movement of data outside the database, and subsequent implementation of the results in the application. The Road Directorate uses Oracle Database 10g [4] for storing the collected data. It was decided that all data handling pertaining to bringing the collected data into a format that can be used as input to the forecast algorithm would take place in a data warehouse that would be built inside the Oracle Database. For this reason, it makes sense to use Oracle Data Mining for subsequent model building, evaluation and deployment as this tool is an inherent part of the Oracle database. Chapter 5 will give an informal introduction on how Oracle Data Mining can be used for data modeling purposes.

Chapter 3

Conceptual project outline

The development of a forecasting algorithm, which is to be integrated in a commercial application, is limited not only to the selection of an appropriate forecast algorithm and an estimation of model parameters. A supporting framework needs to be created in which data handling and modeling is going to take place. Figure 3.1 outlines a conceptual data value chain from data collection, aggregation, transformation and preprocessing to production of 15-minute travel time forecasts. This figure will serve as a road map for the work that needs to be done in order to develop the requested forecast algorithm and make it ready for deployment. The box-shaped sections, which are marked with dotted lines, will not be considered in the project.

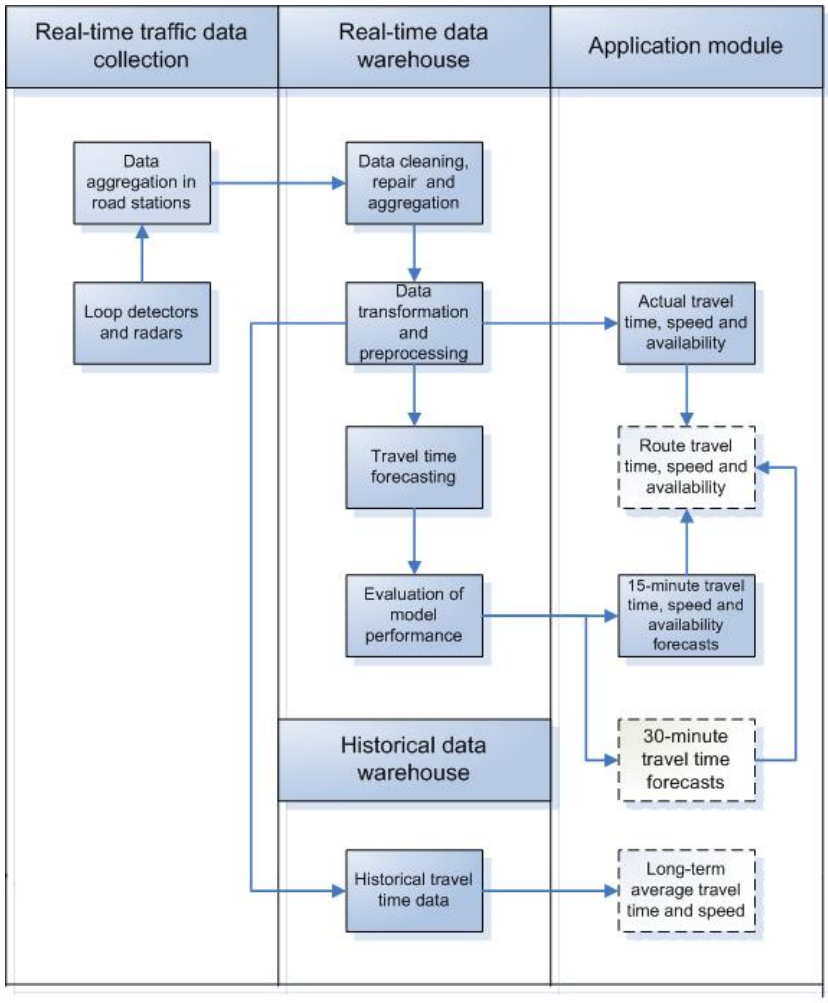

Figure 3.1: Conceptual project outline

Chapter 4

Review of the studied bibliography

4.1 Introduction

A number of articles about travel time forecasting were reviewed to gain insight into the studies that have been conducted in this field. Only articles that deal with models based on statistical analysis and machine learning have been reviewed. The outlined requirements in Chapter 2 will be used as a point of departure for the discussion of the theoretical aspects and the technical potentialities presented in the articles. This discussion includes the issues that are comparable with the ones that will need to be dealt with in the present project. These are listed below:

Introduction A brief introduction to the paper.

Data quality Method of approach towards dealing with the quality of input data (regarding, e.g., discarding and/or repairing corrupt and missing data, handling incident traffic patterns etc.)

Forecasting step The time interval upon which the forecasts are made.

Forecasting horizon The extent of time ahead to which the forecast is referring.

Type of input Speed or travel time and the interval upon which the input is based.

Forecast algorithm The selected methodology approach.

Validation The statistical and qualitative measures that are used for model validation.

Results

Scenarios for practical implementation

Following are a few articles that have been chosen for further examination to illustrate the variety and quality of the research that has been conducted in the area.

4.2 Reviewed articles

4.2.1 Road trafficking description and short term travel time forecasting, with a classification method [5]

Introduction The study is carried out on a road segment which is a branch of the Parisian highway network. This road segment is representative of the Parisian road traffic. It is 21.82 kilometers long and has 38 counting stations. The database used in the paper is composed of the daily evolution of the vehicles speed over 709 days. For each day and each counting station 180 measurements are available, corresponding to the average speed over a period of 6 minutes, ranging from 05:00:00 to 23:00:00.

Data quality Concrete action is taken to remove aberrant data and to complete missing data. These data cleaning techniques are applied to the 6-minute series containing the average speed.

Forecasting step Unspecified.

Forecasting horizon {18, 30, 48, 60, 78, 90, 108} minutes.

Type of input Average speed over a period of 6 minutes (6-minute series).

Forecast algorithm The forecasting methodology consists of two main parts: first, the estimation of standard speed profiles for each counting station; second, the matching of incoming observations for speed to the estimated profiles and hence, estimating speed at all counting stations to forecast travel time. Intercomparison of mixture models and the agglomerative clustering algorithm for the estimation of speed profiles.

Validation The difference between the actual and the estimated travel time, standardized by the actual travel time for each counting station using both forecasting methodologies. These are compared to the stationary pattern, which is the mean of all available traffic patterns for each counting station.

Results The best forecasts are achieved by the agglomerative forecast algorithm.

Scenarios for practical implementation None.

Comments The forecasting step is left unspecified. No differentiation is made between peak-hours and off-peak periods. Even if, a speed curve for one of the counting stations indicates that the travel times vary throughout the day. The discussion paragraph addresses the issue of outliers and rare events. Computational performance of the algorithms has also been taken into consideration.

4.2.2 Univariate Short-Term Prediction of Road Travel Times [6]

Introduction The study is conducted on data collected from Vehicle Information and Communication System in Japan over a 15 km long main road over a period of 14 months. For each day 262 measurements are available, corresponding to an average of travel times over a period of 5 minutes, ranging from

02:00:00 to 23:55:00. These travel times were derived from speed, flow and occupancy measurements collected from the detectors installed on this road.

Data quality Not considered.

Forecasting step 5 minutes.

Forecasting horizon {5, 10, 15, ..., 120} minutes.

Type of input Average travel time over a period of 5 minutes (5-minute series). The authors do not specify how the collected measurements are accumulated into the 5-minute travel time series.

Forecast algorithm The forecasting process consists of fitting a different model for each 5-minute interval. Experimental intercomparison of linear regression, neural networks, regression trees, k-nearest neighbors and locally weighted linear regression models.

Validation Root Mean Squared Error (RMSE). This error is averaged over all prediction points, i.e., over all predictive models.

Results Locally weighted linear regression models produce the best forecasts.

Scenarios for practical implementation None.

Comments Provides a remarkable insight into a variety of univariate methods for travel time forecasting for a single road segment.

4.2.3 Classification of Traffic Pattern [7]

Introduction The data used in this study was collected from a 12 km long expressway of the Tokyo Metropolitan Expressway Network over a period of 2

years. For each day 288 measurements are available, corresponding to an average of travel times on the expressway over a period of 5 minutes, ranging from 00:00:00 to 23:59:00. These travel times were derived from speed measurements collected from the detectors installed app. 300 meters apart.

Data quality Not considered.

Forecasting step Unspecified.

Forecasting horizon Unspecified.

Type of input The 5-minute travel time series are smoothed by use of a wavelet function. These series are subsequently divided into 2 periods, ranging from 07:00:00 to 13:00:00 and 15:00:00 to 20:00:00, respectively. The author does not specify how the collected speed measurements are accumulated into the 5-minute travel time series.

Forecast algorithm The forecasting methodology consists of two main parts: first, the segmentation of traffic patterns from the historical database into a fixed number of representative clusters (for each period); second, matching the incoming patterns with the representative clusters. Small large ratio clustering algorithm for market basket data was used for segmentation of traffic patterns. An exogenous database that for each traffic pattern contains information about day of the week, amount of rainfall, long weekend and vacations is used to facilitate the segmentation process.

Validation A coefficient that measures the correlation between each traffic pattern and the segmented patterns. Mean Absolute Error (MAE), mean absolute percentage %, percentage of forecasts within 5 % and 10 %. The traffic patterns are also matched individually with all traffic patterns from the non-segmented database (= the historical database).

Results Matching new traffic patterns to segmented historical traffic patterns performs better than matching new traffic patterns to all historical traffic patterns one by one.

Scenarios for practical implementation None.

Comments The results are reported on a per day basis, which does not reflect the overall performance of the forecasting algorithm. Furthermore, they are hard to interpret due to the fact that essential information about the different parameters is left unspecified (forecasting step and forecasting horizon).

4.2.4 Travel Time Prediction with Support Vector Regression [8]

Introduction The data used in this study are travel times over 45, 78 and 350 kilometer long road segments in Taiwan collected over a period of 5 weeks. The 5 week period is chosen such that it does not contain any special events and holidays etc. as these factors, according to the authors, could bias the results. For each day 60 measurements are available, corresponding to an average of travel times on the expressway over a period of 3 minutes, ranging from 07:00:00 to 10:00:00. These travel times are derived from speed measurements collected from detectors installed at 1 km intervals along these road segments.

Data quality Days with missing and/or corrupted values are excluded from the study. Therefore, repair methods for improving the quality of data are not considered.

Forecasting step Unspecified.

Forecasting horizon Unspecified.

Type of input Average travel time over a period of 3 minutes (3-minute series). The authors do not specify how the collected speed measurements are accumulated into the 3-minute travel time series.

Forecast algorithm Support vector regression.

Validation Relative Mean Error (RME) and Root Mean Squared Error (RMSE). Comparison with current travel time forecast method and historical mean forecast method.

Results The support vector regression algorithm outperforms the other two methods.

Scenarios for practical implementation None.

Comments The results are unclear as information about the forecasting step and the forecasting horizon is not specified. The results of the study are subject to the condition that the input traffic patterns are "good" traffic patterns which is most unlikely in field conditions.

4.2.5 M3 forecast model [3]

The development of this model was conducted in the framework of the Road Directorate. The results of this study have not been published. This model has been included because research on Google did not return anything about real life implementation projects pertaining to travel time forecasting. For this reason, the writer will implicitly take this model as a baseline case when evaluating the results of the 15-minute forecast algorithm.

Introduction The data used in this study are travel times from 12 road segments spanning the M3 motorway. These segments are of varying length, ranging from 1.5 to 3.5 kilometers. The data have been collected over a period of 2 months. For each day 482 measurements are available for each road segment, corresponding to 1-minute travel time series, ranging from 06:00:00 to 10:00:00 and 14:00:00 to 18:00:00. These travel times are derived from measurements for speed and vehicle count collected from detectors installed at app. few hundred meter intervals along these road segments.

Data quality Observations with missing values are substituted with an average of observations from the other detectors which belong to the same road segment. Corrupt values are excluded from the study. Vacations and incidents are also removed.

Forecasting step 1 minute.

Forecasting horizon 15 minutes.

Type of input Difference between long-term (2 months) average travel times for the specific time of day of the forecast and travel times at 1-minute intervals. This value is calculated for each segment.

Forecast algorithm Linear autoregressive model. 120 models are made in total: two for each business day and road segment - one for the AM period and another for the PM period.

Validation Mean Squared Error (MSE). Percentage of error between the actual and forecasted travel times exceeding 2 and 5 minutes, respectively. These values are averaged over all road segments. Baseline predictors such as the 2-month historical average are included for comparison.

Results Model performance is evaluated on the M3 motorway stretch as a whole. Individual segments are not evaluated separately. The model with the lowest MSE value was chosen for implementation.

Scenarios for practical implementation The forecast algorithm was implemented in a real-time application. However, due to glitches in data relay from the data warehouse where the collected data reside, the service was never launched to the general public.

Comments The results are subject to the condition that all input traffic patterns are "good" traffic patterns. Individual road segment characteristics are not taken into account as one model structure is fit for all road segments.

4.3 Concluding remarks

The reflections of this section do not include the M3 forecast model. The comments are only related to the published articles.

A common feature in all of the reviewed articles is that none of them actually deal with a real life implementation case, suggesting that the proposed algorithms are not immediately intended for use in any kind of application. As a consequence hereof, the supporting framework in which data preprocessing and modeling takes place is not considered. The proposed forecast algorithms are evaluated under presumably ideal conditions in that aberrant traffic patterns are either excluded from the modeling process or ignored. How this affects the results, is not investigated in any great detail. Although the proposed forecast methods perform satisfactorily as reported in the studies, the results are not related to the type of application the forecasting algorithms potentially are going to get integrated into. Furthermore, the results of some of the studies are somewhat unclear in that they are blurred by the fact that crucial information about forecasting step and horizon is not revealed. Intercomparison between the studies is difficult as each study has a different basis in terms of topology of the examined road segments, the level of detail of input data, the forecasting step and horizon, and the utilized forecast method. The presented material is not backed up by even a hypothetical implementation scenario, which is another weak point. There is, however, no doubt that the reviewed articles have been informative in terms of introducing the writer to a wide range of methods and approaches that can be utilized when dealing with travel time forecasting. This will serve as an inspiration for further work.

Chapter 5

Practical issues

5.1 Introduction

Before beginning the data preparation and modeling, clarification about the supporting tools will be provided. It was mentioned in Chapter 2 that the Road Directorate has decided that the 15-minute forecast algorithm will be developed using the tools provided by Oracle Data Mining. This approach has been chosen to streamline processes pertaining to the development of the 15-minute forecast algorithm.

5.2 What is Data Mining with Oracle?

Oracle Data Mining (ODM) embeds data mining functionality in the Oracle Database [9]. This means that the data preparation, model building, evaluation and deployment remain in the database. ODM algorithms operate natively on the data residing in the database, thus eliminating the need for extraction and transfer into prevalent stand-alone tools for data modeling such as R, S-plus or MATLAB etc., resulting in a simpler and more streamlined data modeling process. Model deployment is also facilitated in that the results are already in the database. Also, the less data movement, the less time the entire process

takes. The direct coupling between the various stages of data modeling process is regarded as a step forward by the Road Directorate in terms of applicability of using data modeling in applications pertaining to traffic reporting.

5.3 Data Mining Functions in ODM

For a complete list of available data mining functions refer to Oracle Data Mining Concepts Guide [9].

Chapter 6

Real-time traffic data collection

A series of in-road loop detectors and roadside radars (698 to be exact) placed along motorways spanning the Greater Copenhagen Area supply the following information: vehicle presence, vehicle count and occupancy. In that detectors cannot directly measure speed, it is estimated from the time a vehicle spends between two detectors. These measurements are continuously relayed to centralized road stations set up on the sides of the roads. Although information is relayed many times per second, the measurements are accumulated and amplified at the road stations. Accumulated values for speed and vehicle count for each detector and radar are subsequently sent to the data warehouse at 1-minute intervals (hereinafter referred to as 1-minute accumulated measurements for speed and vehicle count). The accumulated value for speed is an estimate for the average of speed in the preceding minute.

Chapter 7

Real-time data warehouse

7.1 Introduction

A real-time data warehouse was built for the purpose of handling minute-to-minute data. All processing of data pertaining to the estimation of the 15-minute travel time forecasts will be conducted in this data warehouse. This chapter will outline the route of how the 1-minute accumulated measurements for speed and vehicle count can be transformed into prospective input to the forecasting algorithm.

7.2 Preliminary considerations

The reviewed bibliography in the area of travel time forecasting showed that different types of data were used as input to the forecasting algorithms (see Chapter 4 for examples). These data were accumulated and amplified to the desired level of detail from detector measurements of vehicle presence, vehicle count and occupancy. The choice of the type and the level of detail of input data depend to a great extent on the type and area of each application. Hence, the presented proposals cannot be extended to this project right away. The starting point in this project is the 1-minute accumulated measurements for

speed and vehicle count which are available for all detectors and radars. These measurements are updated at 1-minute intervals. The application requires the estimation of actual segment travel times and 15-minute forecasts between two adjacent motorway exits at 1-minute intervals. The segment travel times can be estimated by the use of aggregation, after which they can potentially be used as input to the forecasting algorithm. Hence, it was decided to aggregate the 1-minute accumulated measurements for speed and vehicle count to 1-minute travel times between two adjacent motorway exits. The aggregation steps are described in Section 7.3. The aggregation process itself is outlined in Section 7.4. Speed measurements for single vehicles could also have been used as input for the estimation of segment travel time. This is due to the fact that the average of speed may not be adequate to characterize the actual speed profiles experienced by vehicles traveling along a motorway segment. However, these measurements were not readily available.

7.3 Aggregation steps

The purpose of aggregation is to aggregate the 1-minute accumulated measurements for speed and vehicle count to 1-minute travel times between two adjacent motorway exits. This process consists of three steps, which are equivalent to the three levels of details in which the 1-minute accumulated measurements for speed and vehicle count will be found in during the aggregation process: the lowest level, the intermediate level and the highest level. The lowest level is equivalent to the 1-minute accumulated measurements for speed and vehicle count in each lane within a cross section. This is illustrated by the encircled detectors in Figure 7.1. The intermediate level is an aggregation of the measurements at the lowest level across the lanes within a cross section. A cross section is defined as a stretch of road that is covered by detectors in all lanes. The scope of a single cross section is illustrated in Figure 7.1. The highest level equals the aggregation of the measurements at the intermediate level across all cross sections between two adjacent motorway exits. The stretch between two motorway exits is defined as a motorway segment. This is illustrated in Figure 7.2. Data aggregation from the lowest level to the intermediate level produces cross section traffic data. Data aggregation from the intermediate level to the highest level produces segment traffic data. The applied formulas are described in Section 7.4.

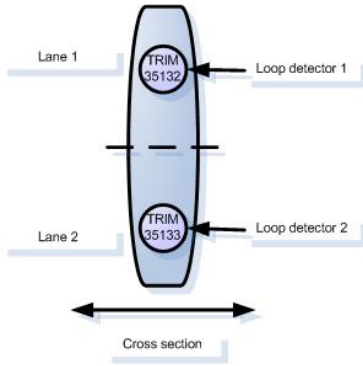

Figure 7.1: Detectors within a cross section

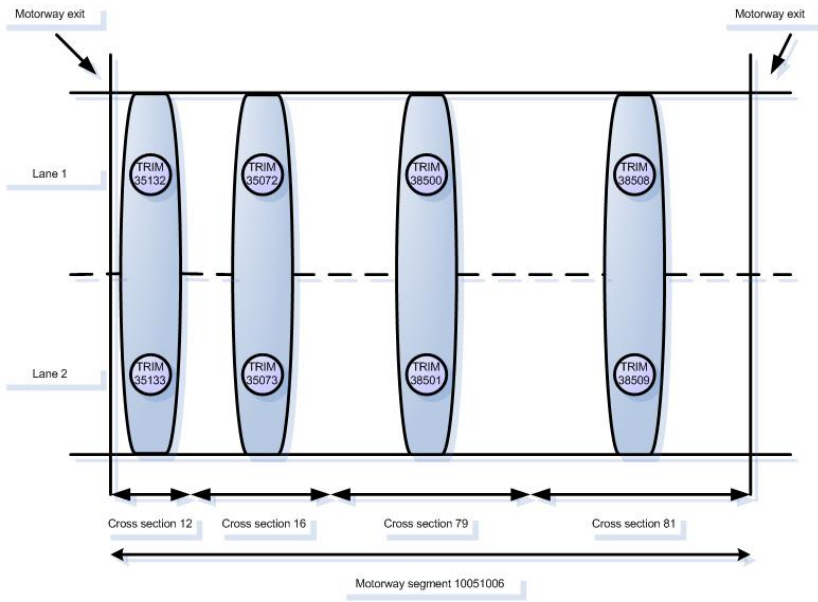

Figure 7.2: Aggregations levels

7.3.1 The lowest level

The lowest level is comprised of the following measurements: a detector identifier, timestamp, average speed and vehicle count. Table 7.1 shows these measurements for motorway segment 10051006 for one point in time. These mea-

surements should ideally exist for all active detectors. There are 698 active detectors and radars in the motorway network spanning the Greater Copenhagen Area. Values for average speed and vehicle count are missing for detectors with identifiers TRIM35132 and TRIM35133. This illustrates the fact that drop-outs occur quite frequently due to a number of reasons such as equipment failure, communication failure or extreme traffic bottlenecks.

Detector identifier	Timestamp	Average speed (km/h)	Vehicle count
TRIM35132	5-05-2007 11:55	*Missing*	*Missing*
TRIM35133	5-05-2007 11:55	*Missing*	*Missing*
TRIM35072	5-05-2007 11:55	115	2
TRIM35073	5-05-2007 11:55	85	10
TRIM38500	5-05-2007 11:55	118	1
TRIM38501	5-05-2007 11:55	104	7
TRIM38508	5-05-2007 11:55	92	14
TRIM38509	5-05-2007 11:55	88	8

Table 7.1: Measurements at the lowest level for motorway segment 10051006

7.3.2 The intermediate level

The intermediate level is comprised of the following measurements: a cross section identifier, timestamp, cross section speed and cross section length. Table 7.2 shows these measurements for motorway segment 10051006 for one point in time. This level is a collection of motorway cross sections, of which every cross section consists of a number of detectors. This number depends on the number of lanes on the given stretch of motorway. There are 298 cross sections in the motorway network spanning the Greater Copenhagen Area. Cross section with identifier 12, which consists of detectors with missing measurements, is included at the intermediate level for reasons explained in Section 7.4.

Cross section identifier	Timestamp	Cross section speed (km/h)	Cross section length (m)
12	5-05-2007 11:55	*Missing*	304
16	5-05-2007 11:55	90	933
79	5-05-2007 11:55	106	1422
81	5-05-2007 11:55	91	1643

Table 7.2: Measurements at the intermediate level for motorway segment 10051006

7.3.3 The highest level

The highest level is comprised of the following measurements: a segment identifier, timestamp, travel time, segment speed and segment length. Table 7.3 shows these measurements for motorway segment 10051006 for one point in time. This level is a collection of motorway segments, of which every segment consists of a number of cross sections. The motorway segments have been selected to be consistent with adjoining road entrances and exits. Hence, the number of cross sections in a motorway segment varies according to the stretch of motorway. There are 76 motorway segments in the motorway network spanning the Greater Copenhagen Area.

Segment identifier	Timestamp	Travel time (min)	Segment speed (km/h)	Segment length (m)
10051006	5-05-2007 11:55	2,72	95	4302

Table 7.3: Measurements at the highest level for motorway segment 10051006

7.4 Data cleaning, repair and aggregation

7.4.1 Introduction

Experience shows that real-time traffic data is always distorted by noise and usually include false or missing values, duplicate measurements, resulting from the malfunction of the data collection and relay mechanisms. For these reasons, the quality of the 1-minute accumulated measurements for speed and vehicle count needs to be checked, and corrective actions need to be taken before travel time can be calculated to ensure that the derived travel time is as reliable as possible. The data cleaning, data repair and data densification process is conducted on all three levels and is thus an inherent part of the aggregation process. The rules for data aggregation and cleaning have origins in the specification of requirements for intelligent traffic management in connection with the extension of the M3 motorway [10]. The rules for data repair were informally outlined during a series of meetings between the writer and the engineers at the Road Directorate, and are based on their experience. Detailed guidance about data densification has primarily been sought in data warehousing literature [11].

7.4.2 Cleaning rules

A series of cleaning rules were worked out in order to clean the 1-minute accumulated values for speed and vehicle count from unreasonable values, both individually and in combination. Table 7.4 shows an individual test rule for speed. This rule examines the value of speed for each individual detector at a time.

Cleaning rules Speed test	
Speed > 180 (km/h)	Discard

Table 7.4: Cleaning rules for each detector - Speed test

Table 7.5 shows combination test rules. These rules examine the combined values of speed and vehicle count for each individual detector at a time.

Cleaning rules Combination tests	
Speed = 1-180 (km/h), Vehicle count > 0	Accept
Speed = 0, Vehicle count = 0	Discard
Speed = 0, Vehicle count > 0	Discard
Speed = 0, Vehicle count < 0	Discard
Speed < 0, Vehicle count = 0	Discard
Speed < 0, Vehicle count > 0	Discard
Speed < 0, Vehicle count < 0	Discard
Speed > 0, Vehicle count = 0	Discard
Speed > 0, Vehicle count < 0	Discard

Table 7.5: Cleaning rules for each detector - Combination tests

The main deficiency of the individual test is that it assumes that the acceptable range of values for vehicle count for the same detector is independent of the value of the speed. Hence, unreasonable combinations of values for speed and vehicle count that are not listed in Table 7.5 will not be identified. If the 1-minute accumulated values for speed and vehicle count do not pass the combination test, the values for speed and vehicle count for the affected detector are fixed at null. A null value indicates a missing value [12].

7.4.3 The lowest level

It is expected that the number of incoming 1-minute accumulated measurements for speed and vehicle count is consistent with the number of active detectors in the motorway network. However, this assumption turns out to be false quite frequently for reasons mentioned in Section 7.4.1 (see Table 1 for an example). In this case, the missing detector identifiers have to be inserted in the table where the other non-missing measurements reside. The missing detector identifier gets the same timestamp as measurements from the non-missing detectors. Values for average speed and vehicle count are fixed at null. The purpose with data densification at this level is to ensure that the following aggregation of the data from this level to the intermediate level is correct. This highlights an important issue with using aggregation for travel time estimation, which is that the aggregated travel time value is reliable and can be considered as prospective input to forecasting algorithms, only if the data at the lower levels is dense.

7.4.4 The intermediate level

The following algorithm has been employed to aggregate the 1-minute accumulated measurements for speed and vehicle count to the intermediate level: The number of vehicles in the motorway cross section is calculated from the following formula:

$$n = \sum_{j=1}^{M} n_j,$$

where n is the total number of vehicles in the affected cross sections, n_j is the number of vehicles in lane j and M is the total number of lanes. The number of lanes ranges from two to four lanes depending on the motorway segment. Cross section speed is calculated from the following formula:

$$v = \frac{\sum_{j=1}^{M} n_j v_j}{\sum_{j=1}^{M} n_j},$$

where v is the weighted average speed in the affected cross section, n_j is the number of vehicles in lane j, v_j is the speed in lane j and M is the total number of lanes. All corresponding pairs of measurements (v_j, n_j) need to have passed the combination test in order to calculate the aggregated values for average speed and vehicle count at the intermediate level. Another option could have been to substitute the negative and missing values for average speed and vehicle count

in the affected lanes with values from adjacent lanes. However, this option was dismissed given the differences in average speed between the fast and the slow lanes. Previously conducted research at the Road Directorate has shown that this substitution method has an impact on the reliability of the aggregated travel time values. It can be argued whether these results apply to rush hour traffic as the differences between the fast and the slow lanes are balanced out during this period. Furthermore, substitution and interpolation with values from adjoining detectors was also dismissed. A number of data repair methods have been implemented in order to densify the aggregated values. Missing values for speed at the intermediate level are substituted with non-missing values over a 5-minute time window preceding the timestamp at the same level. This technique has been chosen due to the fact that some data deliveries frequently lag behind the schedule in the range of a few minutes. The substitution fails if it turns out that all values for speed in the preceding 5 minutes are missing. In this case, the value for cross section speed is calculated as the average of speed values in the remaining cross sections which belong to the same motorway segment at the highest level:

$$v_{cross\ section} = (\frac{\sum_{j=1}^{M} v_j}{M}),$$

where $v_{cross\ section}$ is the average speed in the cross section, v_j is the speed in the remaining cross sections that belong to the same motorway segment and M is the number of cross sections in the affected motorway segment. The travel time at the intermediate level is calculated from the following formula:

$$t_{cross\ section} = \frac{s_{cross\ section}}{v_{cross\ section}},$$

where $t_{cross\ section}$ is cross section travel time, $s_{cross\ section}$ is the length of the cross section and $v_{cross\ section}$ is the cross section speed. The travel time $t_{cross\ section}$ in the cross section is fixed at null if the value for speed $v_{cross\ section}$ is missing.

7.4.5 The highest level

The travel time at this level is calculated by adding the individual cross section travel times:

$$t_{segment} = \sum_{j=1}^{M} t_j,$$

where $t_{segment}$ is the travel time in the motorway segment, t_j is the travel time in cross section belonging to the motorway segment and M is the number of cross sections. The speed is calculated by the following formula:

$$v_{segment} = \frac{s_{segment}}{t_{segment}},$$

where $v_{segment}$ is motorway segment speed, $s_{segment}$ is the length of the segment and $t_{segment}$ is the aggregated travel time in the segment. The travel time at this level will be hereinafter referred to as the aggregated travel time. Another measurement that is calculated at this level is the motorway segment availability rate. This measurement is used to quantify the available cross sections used in the segment traffic data production. A motorway segment availability rate is estimated based on the following formula:

$$Availability\ rate = \frac{\sum_{j=1}^{M''} s_j}{\sum_{j=1}^{M} s_i},$$

where M'' is the number of cross sections totally or partially included in the affected motorway segment with a valid cross section speed as determined in Section 7.4.4, s_j is the length of the jth cross section belonging to the affected motorway segment and having a valid speed, M is the number of cross sections constituting the motorway segment and s_i is the length of the ith cross section included in the motorway segment. The motorway segment availability rate is used as a quality measure to assess the reliability of the estimated segment travel times and speeds. Moreover, this measure can be used to make an estimate of the quality of the incoming data. This measurement will no be reported to the end users.

7.5 Concluding remarks

This chapter addressed a number of issues which need to be dealt with before embarking on the modeling process. A method of approach, by which the actual

travel times in a motorway segment can be estimated, was developed. These travel times will subsequently serve as prospective input to the forecasting algorithm. Furthermore, a set of rules for data cleaning and repair were devised to ensure that the quality of segment travel times is estimated to best effect.

Chapter 8

Historical data warehouse

A historical data warehouse was built for the purpose of storing the daily minute-to-minute cross section and segment traffic data, which can be recalled when examination of past data is required (e.g. when building a forecast model). This data warehouse contains data for each day from October 2006 through March 2007 (and growing with each passing day). For each motorway segment and each day 1440 travel times are stored in a table, corresponding to the aggregated travel time and speed over a period of 1-minute, ranging from 00:00:00 to 23:59:00 (the contents are identical to Table 7.3). For each motorway segment and each day 429120 travel times are stored in another table, corresponding to the cross-section travel time and vehicle count over a period of 1-minute, ranging from 00:00:00 to 23:59:00 (contents are identical to Table 7.2). The 1-minute accumulated measurements for speed and vehicle count corresponding to the lowest level are not retained (refer to the contents of Table 7.1). They are discarded at the end of each day.

Chapter 9

The examined motorway network

9.1 Results

The intensity of traffic varies widely throughout the day and for this reason, it can be hard to quantify. Consequently, it is not reasonable to expect that a single forecast model would model equally well travel times during the morning rush hour, afternoon rush hour, and off-peak periods. This means that for the purpose of forecasting, separate models should be fit to different periods of day. Different approaches can be taken towards dividing the day into a number of periods such as the morning and the afternoon peak hour [7] or dividing the day into a number of intervals of shorter duration [6]. Present study will look into the morning rush hour, which has been defined as the time interval between 06:30:00 and 09:45:00. Preliminary inspection has shown that not all motorway segments have morning rush hour traffic that spans the entire time interval. However, in order to standardize the data preparation, model building, evaluation and deployment processes for future use, the selected time interval had to capture the characteristics of the whole motorway network.

9.2 Description

The examined motorway network encompasses the motorway stretch called Hillerødmotorvejen. This motorway spans from motorway entrance Allerød in the north to motorway exit Høje Gladsaxe in the south. The examination will only encompass the morning rush hour in the southbound motorway stretch. The northbound motorway stretch will not be examined as it is not affected by the morning rush hour traffic and hence is of no immediate interest to the Road Directorate. The southbound stretch has been divided into five motorway segments, corresponding to the five motorway entrances and exits as shown in Table 9.1. Travel times at 110 km/h (free traffic flow) and 15 km/h (extreme traffic bottleneck) have been included to provide a basis for comparison for the reader. This stretch has been selected for examination for the following reasons: it consists of motorway segments of different lengths and varying intensities of traffic.

Segment identifier	Entrance	Exit	Travel time at 110 km/h (min)	Travel time at 15 km/h (min)	Length
10011002	Allerød	Farum	0,88	6,46	1616
10021003	Farum	Værløse	3,06	22,42	5606
10031004	Værløse	Skovbrynet	1,66	12,16	3040
10041005	Skovbrynet	Gladsaxe	1,44	10,56	2641
10051006	Gladsaxe	Høje Gladsaxe	2,35	17,21	4302

Table 9.1: Specifics on motorway segments on Hillerødmotorvejen southbound

Figure 9.1 illustrates travel speeds for each motorway segment. This figure has been included to illustrate the level of congestion on Hillerødmotorvejen during the morning rush hour. The segment travel speeds are better indicators of the level of congestion than travel times because the segments have different lengths. It can be seen that segments 10011002, 10021003, 10031004 and 10051006 are fairly equally affected by the rush hour traffic. These segments are severely congested during the rush hour period in that the average speed for much of the time is in the interval between 20 km/h - 40 km/h. Segment 10041005 is moderately congested. The average speed on this segment is approximately 70 km/h throughout the rush hour period. The level of congestion can be defined as the ratio between the number of vehicles through a motorway segment compared to the capacity of the motorway segment. The rest of the study will only detail the results for motorway segment 10051006.

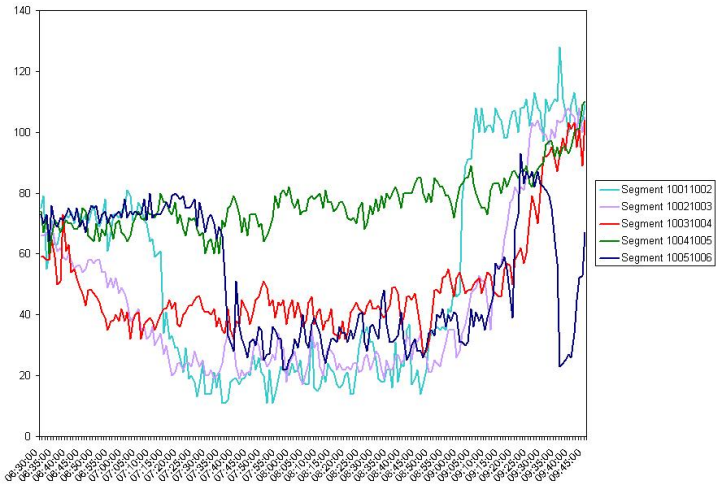

Figure 9.1: Evolution in average speed on Hillerødmotorvejen on Wednesday 28-02-2007

9.3 Results

9.3.1 Example: the aggregated travel times

Figure 9.2 shows the aggregated travel times for motorway segment 10051006 for a number of Mondays from October 2006 to March 2007. Days with missing values in the examined time interval have been discarded. The aggregated travel times fluctuate quite considerably over very short periods of time. This is mainly due to the fact that traffic is processed in a "stop-go" manner, meaning that localized queuing of short duration occurs very often. Furthermore, the aggregated travel times are not true travel times per se, but are rather the results of an aggregation - first, an accumulation of values of speed for all cars passing between two detectors over a period of 60 seconds; second, an aggregation of the accumulated values of speed across detectors and subsequently across cross sections. Both are a cause for considerable variations in speed which, in turn, is reflected in the aggregated travel times.

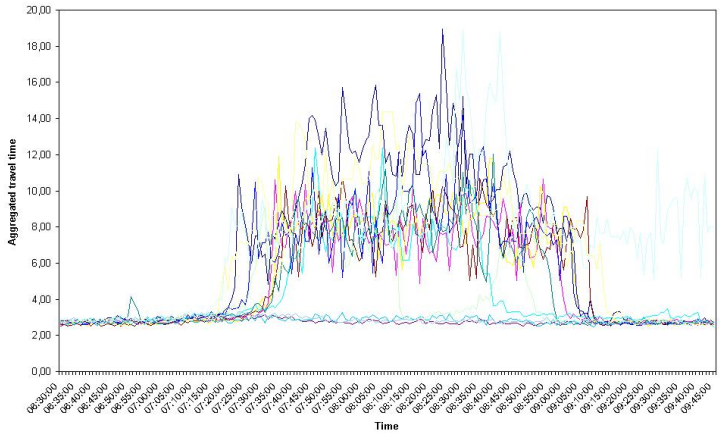

Figure 9.2: Aggregated travel times for different Mondays - motorway segment 10051006

9.3.2 Smoothing

The very stochastic nature of the aggregated travel time data could make the modeling process ineffective and in the worst case scenario, impact on the accuracy of the forecasts to an extent that would render them useless for the end user. To remedy this problem, it was decided to apply a simple moving average function, even if this would probably result in the loss of information. The purpose for this is to diminish the erratic minute-to-minute fluctuations in the aggregated travel times, which in overall terms have little significance, and allow the major trends in traffic flow to be made more visible. Erratic minute-to-minute fluctuations in the aggregated travel times were also deemed unacceptable by the Road Directorate in that these times would also be reported to the public through the website. The interpretation of minute-to-minute variations in the travel times would be confusing for the ordinary user. The fluctuations reflect the variations in cross section speed, not in segment travel time. Due to these reasons, a 10-minute moving average function was applied. For example, the calculation for travel time at time 06:45:00 consists of the average of travel times from 10 consecutive minutes preceding and including time point 06:45:00. This number is scientifically unsubstantiated per se; however, the inspection of the aggregated travel times, 1-minute (1 minute preceding and including the current travel time), 5-minute (5 minutes preceding and including the current travel time) and 10-minute moving average travel times indicated that the best results were obtained by choosing the 10-minute moving average

window function. The 1-minute and 5-minute moving average travel times were too responsive to the recent variations in the aggregated travel times, with the result that the minute-to-minute fluctuations in the travel times were still quite substantial. The 10-minute moving average function seemed to produce the desired results. A larger smoothing interval, however, would give a misleading picture of the current travel time times as it, from a theoretical viewpoint, would produce a more pronounced lag in the smoothed sequence. It can be discussed whether the proposed smoothing interval is already too large. This has, however, not been tested out in practice. Figure 9.3, 9.4 and 9.5 show the aggregated travel times after application of the 1-minute, 5-minute and 10-minute moving average function. It can be seen in Figure 9.5 that the course of the traffic flow is now illustrated more clearly. A number of traffic flow patterns have emerged, suggesting that the traffic flow might be grouped into a number of clusters. The disadvantage of using this smoothing technique is that all

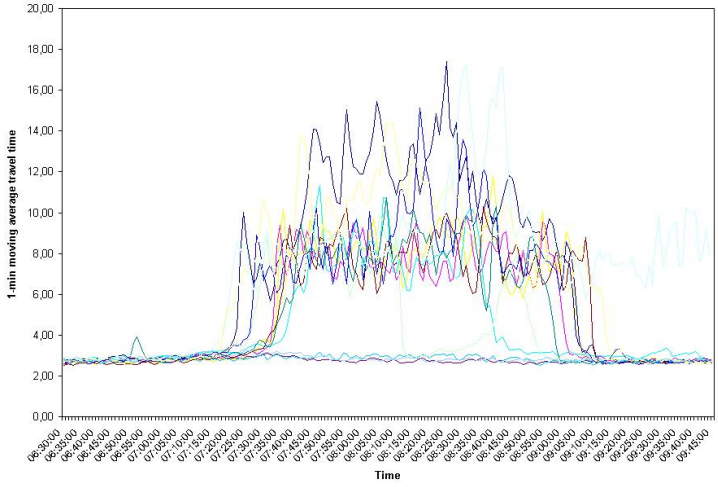

Figure 9.3: 1 - minute moving average travel time data for a number of Mondays - motorway segment 10051006

past observations are given the same weight, in which case, the resulting travel times might be downright misleading, especially as the size of the smoothing interval gets bigger. This applies also for congestion build-up and phase-out. Other smoothing techniques could have been used to diminish the fluctuations in the aggregated travel time values, such as the weighted moving average or the exponential smoothing functions. Both give more weight to recent observations and less weight to older observations [13]. However, these functions are only useful when there are trends. And there are no well-defined trends in the

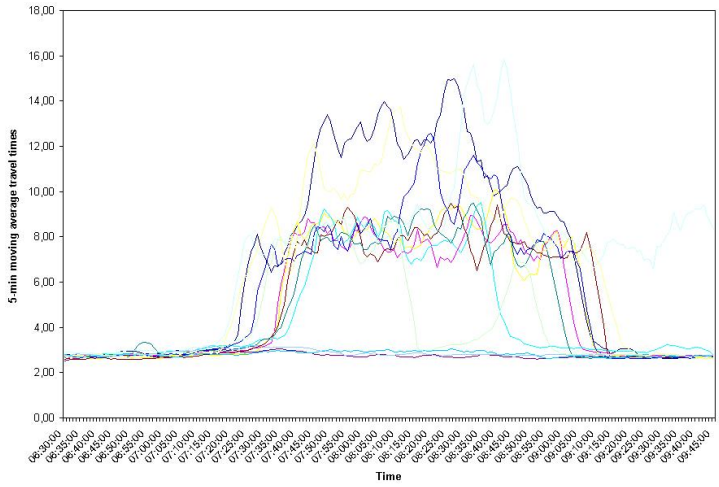

Figure 9.4: 5 - minute moving average travel time data for a number of Mondays - motorway segment 10051006

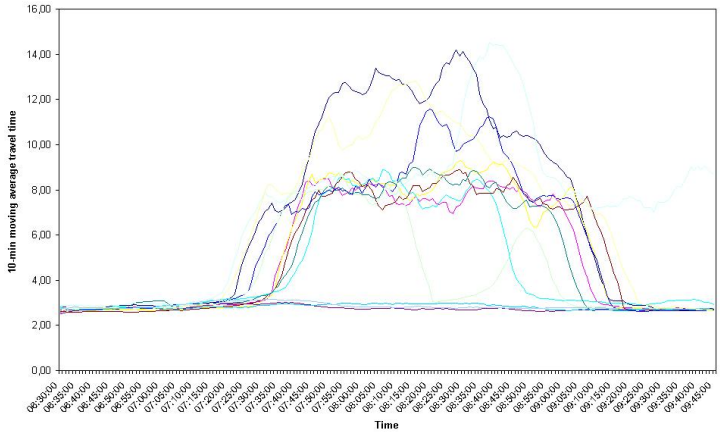

Figure 9.5: 10 - minute moving average travel time data for a number of Mondays - motorway segment 10051006

evolution in the aggregated travel times per se. Visual inspection of Figure 9.2 supports this claim. The aggregated travel times shift considerably throughout the rush hour period, even during congestion build-up and phase-out.

9.3.3 Application of the aggregated travel time data

As a starting point, the aggregated travel times can be used to form a general view of the traffic patterns that govern the examined motorway segment. The 15-minute forecast model developed in connection with the extension of the M3 motorway assumed that the traffic followed a weekly pattern, which meant that the traffic pattern on any Monday morning resembled those of the other Mondays; the traffic pattern on any Tuesday morning resembled those of the other Tuesdays etc [3]. It is of interest to find out whether this assumption also applies to this motorway segment. Examination of the 10-min moving average travel times from October 2006 though March 2007 immediately disproves this assumption, as shown in Figure 9.6.

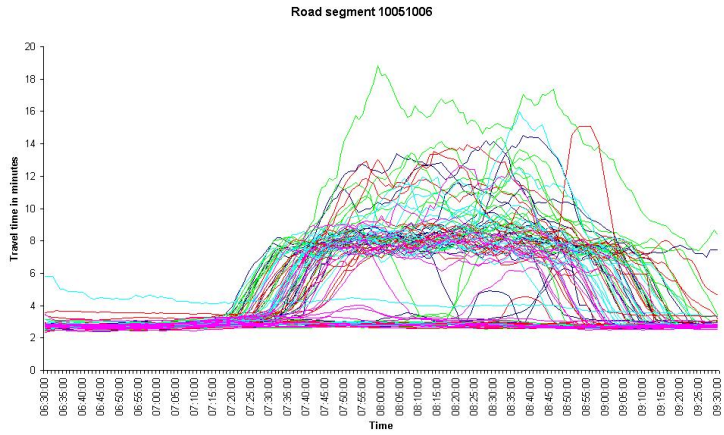

Figure 9.6: 10-min moving average travel times for motorway segment 100510006. Mondays: blue lines, Tuesdays: green lines, Wednesdays: red lines, Thursdays: light blue lines, Fridays: pink lines

It can be seen that congestion build-up intervals range from 07:15:00 and 07:30:00 and congestion phase-out intervals range from 09:00:00 to 09:30:00. The 10-minute moving average travel times range on average from 8 minutes up to 14 minutes. Some days reach peak travel times that are as high as 18 minutes. The travel time on the majority of days is, however, around 8 minutes. It can be concluded that the travel times do not follow a distinct weekly pattern, but are rather governed by the location of congestion build-up, the length of the rush hour, the travel times and the location of congestion phase-out.

9.4 Concluding remarks

Inspection of the aggregated travel times suggested that a smoothing technique was called for in order to eliminate the substantial minute-to minute fluctuations in the aggregated travel times. A 10-minute moving average function was applied to good effect as major trends in traffic flow patterns were uncovered. The selected approach leaves, however, plenty of room for improvement in that the results are heuristic. A sensitivity analysis could, for instance, be conducted in support (or rejection) of the chosen time lag. The application of other smoothing techniques could also be used in order to smoothen the minute-to-minute variations in the aggregated travel times. An entirely different approach could be employed that states that the travel times at the lower levels are used instead of the aggregated travel times. The examination of a series of traffic patterns indicated that the 10-minute moving average travel times might be grouped into a number of clusters based on the shape of the traffic patterns, rather than their respective belonging to the business days and (or) vacations.

Chapter 10

Clustering

10.1 Introduction

Clustering will be used as a means to detect whether the traffic flow on any given day has similarities with the traffic flow on other days. The inspection of Figure 9.6 suggests that traffic patterns could perhaps be grouped into a number of clusters based on the shape of the curve for the traffic flow in the considered time interval. The search for these trend curves will be entirely based on the traffic patterns in the historical data warehouse. There is hope that the algorithm will merge similar traffic patterns into clusters, such that patterns that belong to the same cluster are more similar than patterns that belong to different clusters. The influence of exogenous variables on clustering will also be examined. Each pattern can be characterized by four exogenous variables: business day (Monday through Friday), season (autumn, winter, spring, summer), vacation (fall recess, winter holidays, public holidays, summer vacation etc.), incident (yes/no). It is of interest to find out whether the clustering algorithm can to detect patterns that strongly deviate from the majority of traffic patterns. These traffic patterns are denoted incidents. An incident can be a road accident, bad weather, road works and the like, briefly described, patterns that are not predictable in advance.

10.2 Clustering in Oracle Data Mining

Clustering will be performed using Oracle Data Mining (ODM) clustering algorithms. ODM provides two algorithms for this purpose:

Enhanced k-Means algorithm, which is an enhanced version of the traditional k-Means algorithm [14]

Orthogonal partitioning clusters (the O-Cluster algorithm), which is an Oracle proprietary algorithm [15]

Both algorithms support identifying naturally occurring groupings within the input data set. The enhanced K-means algorithm creates hierarchical clusters and groups the input traffic patterns into the user specified number of clusters. The O-cluster algorithm selects the most representative clusters without the user prespecifying the number of clusters. Both algorithms provide detailed information about the generated clusters, which includes placement of all traffic patterns in the clustering model hierarchical tree, cluster rules, which capture the main characteristics of the data assigned to each cluster and cluster centroid values. The generated clusters can subsequently be used to score new input patterns on their cluster membership. They are also used to generate a Bayesian probability model that is used during scoring for assigning data points to clusters. Clustering can also be used for incident detection by building a clustering model, applying it and then finding items that do not fit in any cluster. Both algorithms were initially tested out on a sample data set. The O-cluster algorithm did not return any usable results due to the fact that all patterns were clustered in the same cluster which, according to the inspection of the 10-minute moving average travel times in Figure 9.6, can be dismissed. In addition, the lack of adequate documentation on the O-cluster algorithm made it almost an impossible task to tune the numerous parameter settings which presumably are required for this algorithm to run properly. Neither Google search nor the browsing of the Oracle Technology Network discussion forums [16] shed more light on how to apply this algorithm in practice. For this reason, the enhanced k-Means algorithm was selected for further examination. The start-up phase was difficult due to the lack of adequate documentation in terms of explaining the theoretical applicability of the numerous parameter settings and interpreting the output. Once these obstacles were overcome, the algorithm was used to gain insight into the behaviour of traffic patterns.

10.2.1 Enhanced k-Means algorithm

The enhanced k-Means algorithm is a distance-based clustering algorithm that partitions the data into a predetermined number of clusters, provided that there are enough distinct patterns in the data set. The algorithm relies on a distance metric to measure the similarity between data points which is either Euclidean, Cosine, or Fast Cosine distance (refer to [17] for an informal explanation of the applicability of Cosine and Fast Cosine distance metrics). The former and the latter distance metrics have not been tested out. Data points are assigned to the nearest cluster according to the distance metric used. The algorithm builds models in a hierarchical top-down manner and is reminiscent of the algorithms for divisive clustering [18]. The algorithm begins by placing all traffic patterns in a single cluster. This single cluster is denoted as the first level of the hierarchy. Each level of the hierarchy represents a particular grouping of data into disjoint clusters of traffic patterns. It then chooses the traffic pattern whose average dissimilarity from all the other patterns in the data set is largest. This pattern becomes the first member of the second cluster. At each successive step that pattern in the first cluster whose average distance from those in the second cluster, minus that for the remaining patterns in the first cluster is largest, is transferred to the second cluster. This continues until the corresponding difference in averages becomes negative. When the transfer of patterns from the first cluster to the second cluster stops, there are no longer any patterns in the first cluster, that are, on average closer to those in the second cluster. The result is thus a split of the original cluster into two child clusters, one containing the patterns transferred to the second cluster, and the other those remaining in the first cluster. These two clusters represent the second level of hierarchy. After the children of the parent node have converged, the traditional k-Means algorithm is run on all leaves until convergence, that is, either when the change in error between two consecutive iterations is less than the minimum error tolerance or the number of iterations exceeds the maximum number of iterations (both are parameter settings that need to be specified before running the k-Means algorithm, refer to Section 10.2.2). Each successive level is produced by applying this splitting procedure to one of the clusters at the previous levels. There are two different strategies on how to choose which child cluster to split in order to increase the size of the tree, until the desired number of clusters has been reached. This strategy only applies from the second level of the hierarchy and onward. The child clusters can be split based on either largest variance or largest size. Variance is computed as the sum of squared distances from all patterns belonging to the child cluster to the cluster centroid. Size is computed as the number of patterns belonging to each child cluster. Recursive splitting continues until all child clusters either become single patterns or the specified number of clusters has been reached. The centroids of the inner clusters in the hierarchy are updated as the construction of the tree evolves. The whole tree is returned.

Moreover, the algorithm returns, for each cluster, the place in the clustering model hierarchical tree, a cluster prototype (consisting of the mean and variance) and histograms (one for each feature). The clusters discovered by the algorithm are then used to create rules that capture the main characteristics of the data assigned to each cluster. The rules represent the bounding boxes that envelop the data in the clusters discovered by the clustering algorithm. The antecedent of each rule describes the clustering bounding box. The consequent encodes the cluster ID for the cluster described by the rule. Furthermore, the algorithm provides probabilistic scoring and assignment of data to clusters. For a given record, the clustering models return the probability of the record belonging to a given cluster P (cluster | record). This probability is similar to what is obtained from using a classification model. From this perspective, clustering is treated as a classification problem where first the class labels (clusters) are identified and then the records are classified into clusters from a predefined set of clusters [19]. Patterns can be affiliated to several clusters with the same probability as long as the total probability does not exceed 1. This might occur when splitting a natural grouping into a number of constituents. In addition, this will occur whenever an attempt is made to cluster an incident pattern. Missing values are not automatically handled. If missing values are present in the data set and are not treated before the clustering algorithm is run, the algorithm will handle the input data incorrectly with the result that the data will be grouped erroneously. This algorithm is not susceptible to initialization issues, that is, the results of the algorithm are reproducible for identical parameter settings. This is useful when dealing with a real life implementation in that the results need not be stored but can be obtained any time. However, the algorithm also suffers from a number of deficiencies such as the vagueness of termination criteria in terms of choosing the right number of clusters in the model. It is then up to the user to decide which model (if any) actually represents a natural clustering in the sense that patterns within each of its groups are sufficiently more similar to each other than patterns assigned to different groups at that level.

10.2.2 The applied settings

The algorithm has the following settings (for a full list of available settings refer to [21]):

Number of clusters. This setting specifies the number of clusters in the model. The value must be between 2 and the number of distinct cases in the data set.

Distance function. This setting specifies how the algorithm calculates the distance. The distance function can be Euclidean, Cosine or Fast Cosine.

Clustering will be performed with the Euclidean distance function.

Split criterion. This setting specifies how the algorithm splits the node. The split criterion can either be variance or size. Clustering will be performed with size as the split criterion. Reasons for this will be explained in Section 10.3.2.

Minimum error tolerance. This setting specifies the minimum percentage change in error between iterations to consider that the clusters have converged. The minimum error tolerance must be a non-negative number that is less than 1. Preliminary runs of the algorithm showed that this setting did not impact the outcome of the algorithm at all. For this reason clustering will be performed with the minimum error tolerance set to .001 (default value).

Maximum Iterations. This setting specifies the maximum number of iterations for the k-Means algorithm. The value must be between 2 and 30. Clustering will be performed with the number of iterations set to 30.

A pattern is affiliated to a cluster only if the probability of affiliation is larger than 50 %. Patterns, for which the probability of affiliation is less than 50 %, are not affiliated to any clusters in order to avoid affiliating that pattern to several clusters with the same probability. In most cases, this occurs when the number of clusters in the model exceeds the amount of natural groupings in the data set. If so, there are several "most-likely" affiliations with probabilities less than 50 %. Aside from this, incident patterns would normally be affiliated with a probability less than 50 % in that they do not belong to any clusters per se. This approach has been chosen in order to keep tabs on the affiliations as the number of clusters in the model is increased, which otherwise would not be feasible. The probabilities are provisionally assigned only for the purpose of evaluation of the validity of the emerged clusters and in order to determine the optimal number of clusters.

10.2.3 Concluding remarks

The main detriment attributable to using ODM is the lack of adequate documentation on how the clustering algorithms work in practice. Oracle only provides limited informative material, which for the most part does not go beyond the very basic introduction to the algorithms, making the initial starting cumbersome. The lack of elaborate documentation prohibited the use of the O-cluster algorithm due to the fact that the algorithm produced clusters that were deemed incorrect. For this reason, it was chosen to further investigate the enhanced k-Means algorithm.

10.3 Results

10.3.1 Input data

The input data for the clustering algorithm is the 10-minute moving average travel times. The data set consists of 86 traffic patterns from October 2006 through March 2007. Each pattern Y_j consists of a sequence of pairs (y_i, t_i) in the interval from 06:30:00 through 09:45:00, corresponding to 196 features, where y_i denotes the 10-minute moving average travel time value and t_i the time that has elapsed since midnight. Weekends have been discarded due to the fact that the traffic flows at the speed limit and are hence of no immediate interest for modeling purposes. Patterns with missing data in the chosen time interval have also been discarded. This is due to the fact that the enhanced k-Means algorithm lacks the ability to handle missing data. Approximately 40 traffic patterns have been deselected. The number of consecutive missing data points for each deselected traffic pattern was not investigated. The number of "complete" patterns was deemed sufficient, and hence no measures were taken in order to densify patterns with missing data points. No prior assumptions are made about the nature of each traffic pattern. The patterns are approximately evenly distributed between business days and months.

10.3.2 Split criterion

Originally, clustering analysis was performed on an ad hoc basis by trying out both split criterions and by randomly varying the number of clusters. It turned out that the best results were achieved by using size as the split criterion in that traffic patterns with similar characteristics were grouped together early, that is, when the number of clusters was small in comparison to the size of the data set. Exceptional traffic patterns were separated out, that is, were not affiliated to any clusters, when the number of clusters in the model was low relative to the number of well-separated groups. Setting the split criterion to variance, resulted in the creation of clusters that were comprised of single traffic patterns when the number of clusters in the model was low in comparison to the size of the data set. This meant that the number of clusters in the model had to be increased substantially in order to capture the underlying characteristics of the data. More often than not, patterns that at first sight resembled each other were separated out as single clusters, whereas the other criterion was capable of grouping the exact same traffic patterns into homogeneous groupings. As a consequence hereof, it was chosen to proceed with size as the split criterion in that this criterion was able to form well-separated groups relatively early, that

is, when the number of clusters relative to the size of the data set was small. Exceptional patterns were segregated as opposed to the other split criterion meanwhile maintaining the natural groupings in the data.

10.3.3 Estimating the number of clusters

Visual inspection of traffic patterns belonging to motorway segment 10051006 gives a vague estimate of the prospective number of clusters (see Figure 9.6). In that the number of clusters in the model is not automatically determined by the chosen clustering algorithm, but has to be set before the algorithm can be run, the determination of the right number of clusters called for another technique than visual inspection. This number will be determined using the following cluster validity measurement techniques: the within-point scatter (within sum of squares function) [22] and the elbow criterion [23]. Separate solutions will be obtained for each number of clusters K. An estimate for the optimal number of clusters is then obtained by identifying a kink in the plot of the within sum of squares values as a function of K. However, this approach is somewhat heuristic and the kink cannot always be unambiguously identified [20]. Following are the results for the within sum of squares function for $K \in \{2, 3, ..., 15\}$. The value for K_{MAX} was chosen in view of initial trial runs of the enhanced k-Means algorithm. This was due to the fact that as K_{MAX} was increased the algorithm was unable to assign the majority of the traffic patterns to any of the clusters with a probability that was greater than 50 %. This can be accounted for by the fact that, as the number of clusters in the model is increased, naturally occurring groupings are split into their subgroups, and if patterns from two or more sub-groups resemble each other (in that they, in fact, belong together), the algorithm assigns a pattern to several clusters with the same probability. Hence, these patterns would be neglected in the calculation of the within sum of squares function. Thus, increasing K_{MAX} would not necessarily decrease its value. These observations entail that the data might, in fact, be divided into a manageable number of well-separated groups. Figure 10.1 shows the within sum of squares function for motorway segment 10051006. The first strong "break" in the value of the within sum of squares function occurs at 3 clusters, followed by another strong "break" at 4 clusters and a less pronounced "break" at 5 clusters, indicating that the optimal number of clusters is to be found in this range. After 5 clusters the drop in the within sum of squares value levels off, suggesting that the data set most likely is comprised of at least five natural groupings. For this reason, it is a reasonable assumption that a feasible lower bound for the optimal number of clusters can be estimated at 5 clusters. This within sum of squares function is useful when determining the optimal number of clusters for a particular motorway segment. This function is, however, not easily comparable across motorway segments in that the magnitude of the values

of the within sum of squares function will depend on the travel time for each motorway segment. Another method, denoted the elbow criterion, contains basically the same information about the potential clusters in the data. It uses a different scaling in that the function values are expressed as percentages of variance explained by the clusters, which are easier to compare across motorway segments.

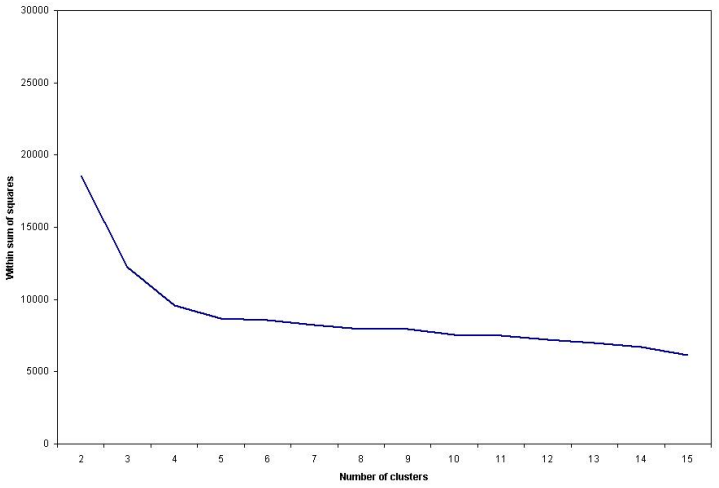

Figure 10.1: Within sum of squares function - motorway segment 10051006

The elbow criterion is the percentage of variance explained by the clusters against the number of clusters, which is the ratio of within-point scatter to total-point scatter. The number of clusters should be chosen so that adding another cluster doesn't add sufficient information [23]. The first clusters will add much information, but at some point the marginal gain in adding a new cluster will drop, giving an "elbow" in the graph. This is approach is also heuristic in that this elbow can not always be unambiguously identified.

Same conclusion as for the within sum of squares function can be drawn from the inspection of the elbow criterion function in Figure 10.2. 85 % percent of the variance is explained when the number of clusters is 5, 15 % up from when the number of clusters is 2, the marginal gain, however, from adding extra clusters remains in the range of 4 % when going from 5 to 15 clusters. Given the results from both cluster validity measurements techniques, it was decided to start up with using 5 clusters to illustrate the grouping of the traffic patterns. Models with 6, 7 and 8 clusters are included for illustration purposes and also to give assurance that the clustering algorithm performs as intended.

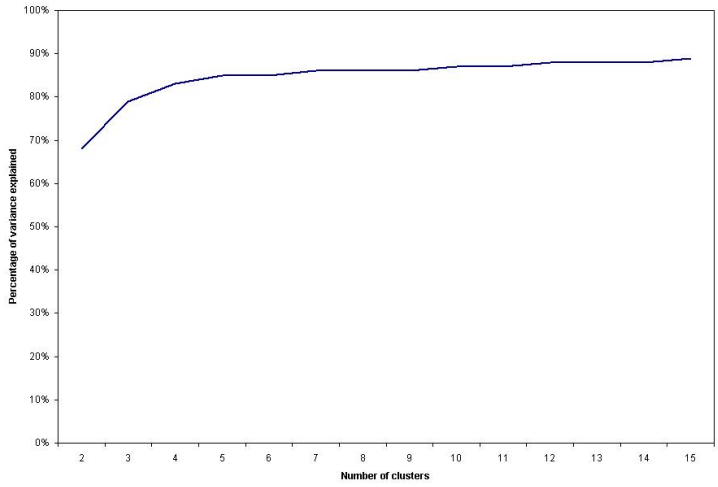

Figure 10.2: Percent of variance explained - motorway segment 10051006

10.3.4 Example: five clusters

The enhanced k-Means clustering algorithm has been applied to the 10-minute moving average travel time data for motorway segment 10051006 as described in Section 10.3.1. The data are a 86×196 table of 10-minute moving average travel times, each representing a measurement for a date-stamp (row) and point in time (column). The enhanced k-Means clustering algorithm is applied with $K = 5$ in light of the outcome of the within sum of squares function and the elbow criterion for each clustering with K running from 2 to 15. Figure 10.3 shows the clusters that have emerged from running the enhanced k-Means algorithm. It can be seen that the shape of the clusters is mostly governed by the intensity of the traffic flow during peak hours and the length of the peak hour period. The rush hour traffic begins approximately at the same time, namely, in the time interval between 07:15:00 and 07:30:00. Also, the slope of congestion build-up is approximately the same. It can be seen that congestion build-up times do not differ significantly between clusters 1, 2, 3 and 4. Cluster 5 does not exhibit a rush hour traffic pattern. The average travel time during the rush hour ranges from 7 minutes to 9 minutes for clusters 2, 3 and 4, with the exception of cluster 1 where the average peak travel times are approximately 12 minutes. There is more variation in congestion phase-out times than in congestion build-up times. The rush hour traffic begins to halt around 09:05:00 for cluster 4, around 09:10:00 for cluster 3, and around 09:15:00 for clusters 1 and 2. This could be explained by the fact that people tend to leave their houses around the

same time in the morning, but the traffic flow since then can be affected by a number of events which might have an impact on the phase-out process.

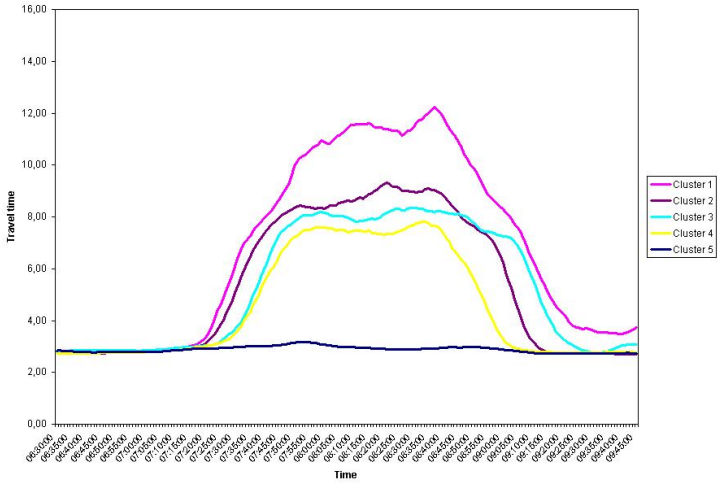

Figure 10.3: 5 clusters - motorway segment 10051006

Table 10.1 shows the distribution of traffic patterns between the five clusters. It can be seen that all business days are distributed more or less evenly between the five clusters, and that all clusters contain approximately the same number of days. Hence, the previously made assumption that the traffic flow follows a pattern that is governed by Mondays through Thursdays as well as Fridays and holiday traffic can be dismissed (see Section 9.3.3 for a contributory cause to this assumption). The vacation's column lists the 5 weekdays from fall recess in October 2006 and 3 weekdays from the winter holidays in February 2007. The remaining two weekdays were deselected due to missing travel time values. All vacations belong to the same cluster, namely, cluster 5. The intensity of traffic on days belonging to this cluster approximately equals traffic at free flow. Moreover, it can be seen that Mondays through Thursdays have been distributed evenly between the five clusters with the exception of the lack of presence of Tuesdays in cluster 4. Table 10.2 shows the compound distribution of the examined business days across clusters. The trend is towards that Mondays through Thursdays predominantly belong to clusters with high travel times, namely, clusters 1, 2 and 3. Over 50 % of Mondays through Thursdays belong to these clusters, whereas only 13 % of Fridays. All Fridays, except for two, have been grouped in cluster 4 and 5. The sizeable presence of Fridays in cluster 5 makes sense as it is a common belief at the Road Directorate that the traffic flow on Fridays proceeds differently than the traffic flow on the other business days,

and for the most part resembles vacation traffic. This assumption is, however, only partially true as a sizeable number of Fridays was also grouped into cluster 4 where the average travel time during peak hours reaches approximately 8 minutes, which is four times higher than the travel time at free flow (see Table 9.1).

	Monday	Tuesday	Wednesday	Thursday	Friday	Vacation	Total
Cluster 1	2	4	2	2	1		11
Cluster 2	3	4	1	6	1		15
Cluster 3	4	5	4	2			15
Cluster 4	2		4	4	7		17
Cluster 5	2	2	3	4	6	8	25
No cluster	1	1	1				3

Table 10.1: Distribution of business days between clusters - motorway segment 10051006

	Monday	Tuesday	Wednesday	Thursday	Friday
Cluster 1	15%	27%	14%	11%	7%
Cluster 1, 2	38%	53%	21%	44%	13%
Cluster 1, 2, 3	69%	85%	50%	56%	13%
Cluster 1, 2, 3, 4	85%	85%	79%	78%	60%
Cluster 1, 2, 3, 4, 5	100%	100%	100%	100%	100%

Table 10.2: Compound distribution of business days across clusters - motorway segment 10051006

In the following traffic patterns affiliated to each of the five clusters along with the ones which were not affiliated to any clusters will be shown in order to visually assess the quality of the resulting groupings in terms of how well these traffic patterns are separated into the five clusters. It can be seen from Figure 10.4, 10.5, 10.6, 10.7 and 10.8 that the enhanced k-Means algorithm is successful at grouping together traffic patterns of the same shape. Inspection of all five clusters suggests that the clusters are well-separated, and it can therefore be assumed that the examined data set is comprised of at least five naturally occurring groupings. There are, however, a few exceptions. One traffic pattern in the third and fourth cluster, and three patterns in the fifth cluster are inappropriately placed in these clusters in that they profoundly deviate from the other patterns in this group. These traffic patterns are marked with green, red and purple. The reason why these patterns are affiliated to the respective clusters is the probabilistic nature of pattern affiliation in that probabilistically these patterns are quite close to the clusters they have been assigned to. Three traffic patterns have been affiliated to all five clusters with a 20 % probability (see Figure 10.9). The inspection of these traffic patterns immediately suggests that this is not due to the fact that a well-separated grouping has been split

up into it's constituents, but rather indicates that an incident might have taken place on the affected days. It is possible that another conclusion will be reached as the amount of traffic patterns in the historical data warehouse increases, and the clustering algorithm is rerun.

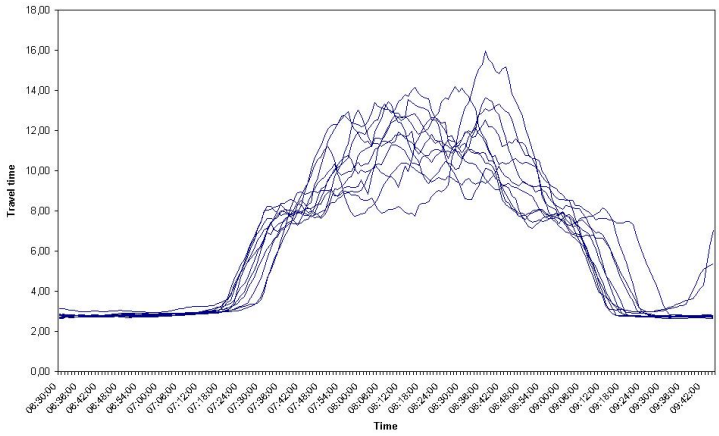

Figure 10.4: Cluster 1

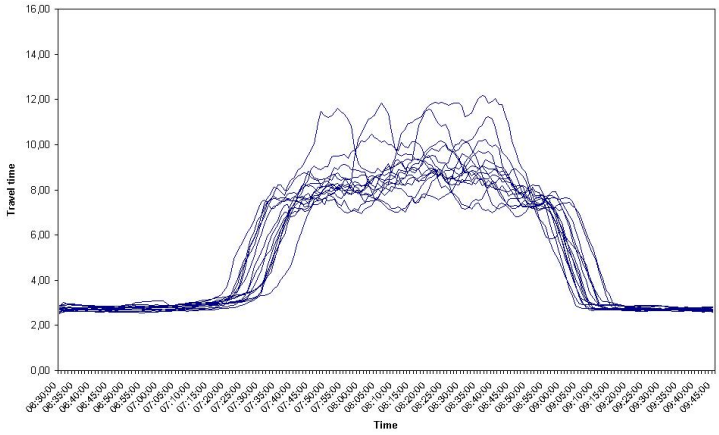

Figure 10.5: Cluster 2

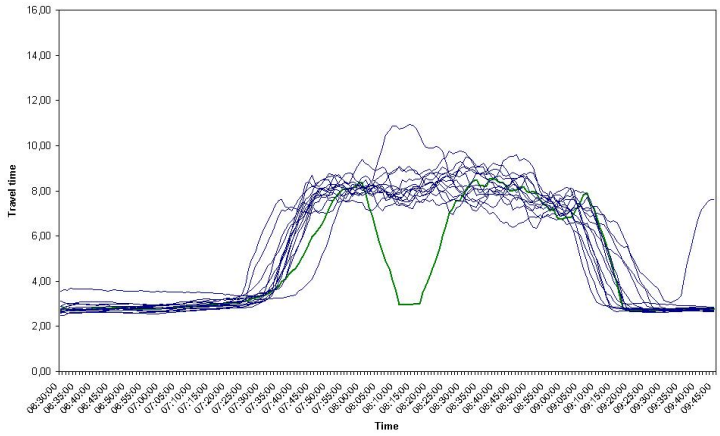

Figure 10.6: Cluster 3

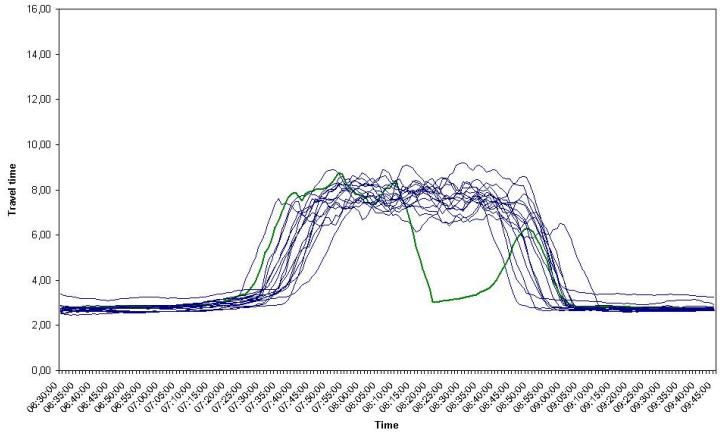

Figure 10.7: Cluster 4

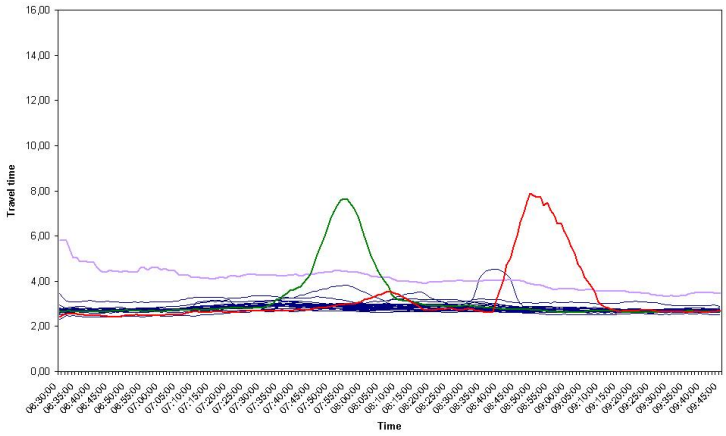

Figure 10.8: Cluster 5

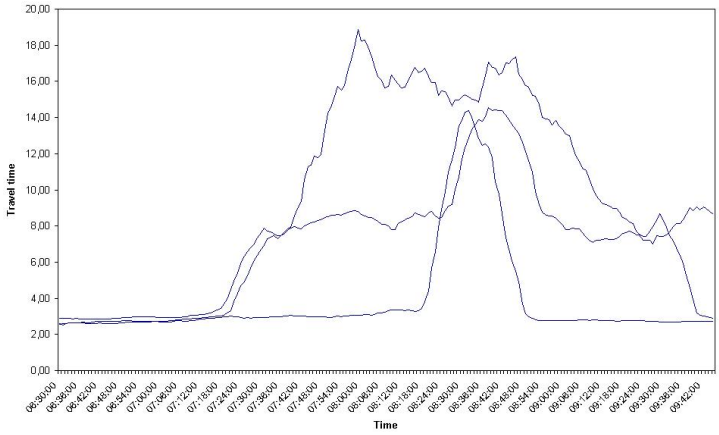

Figure 10.9: Traffic patterns without cluster affiliation

10.3.5 The season factor

Table 10.3 shows the distribution of traffic patterns across months. All months are present in each cluster except for December. This is most likely due to the fact that only seven days qualified as input to the clustering algorithm.

	October	November	December	January	February	March
Cluster 1	3	4	1	1	1	1
Cluster 2	1	3		5	2	4
Cluster 3	1	1		4	5	4
Cluster 4	3	2	1	3	3	5
Cluster 5	5	6	5	1	5	3
No cluster	1	1				1
Total	14	17	7	14	16	18

Table 10.3: Distribution of months between clusters - motorway segment 10051006

Cluster 1 for the most part consists of traffic patterns from October and November, whereas patterns from January, February and March mostly constitute clusters 2 and 3. Clusters 4 and 5 consist of patterns from all of the examined months. Cluster 5 has a notable presence of days from all months, except for January and March. October is due to fall recess, December to the fact that a lot of people tend to take time off before Christmas and February to winter holidays. These observations suggest that the level of congestion might depend on time of the year. Experience shows that there is, indeed, a season effect in terms of the amount of traffic. It is, however, too early to jump to conclusions in that the amount of available data in the historical data warehouse for the time being is deemed insufficient to make a qualified assessment of this effect.

10.3.6 Example: six, seven and eight clusters

The formed clusters in Section 10.3.4 had some shortcomings due to the fact that several traffic patterns were seemingly misplaced. To check the validity of the model, it was decided to apply the enhanced k-Means algorithm to the same data set with $K = 6, 7, 8$. The formed clusters with $K = 6$ are shown in Figure 10.10. Cluster 5 consists of a single traffic pattern, which is one of the patterns that was not affiliated to any clusters with $K = 5$ (see Figure 10.9). The remaining clusters are identical to clusters 1, 2, 3, 4 and 5 with $K = 5$. The percentage of variance explained remains the same in that the added cluster is a single day, which has no influence on the remaining clusters (see Figure 10.2).

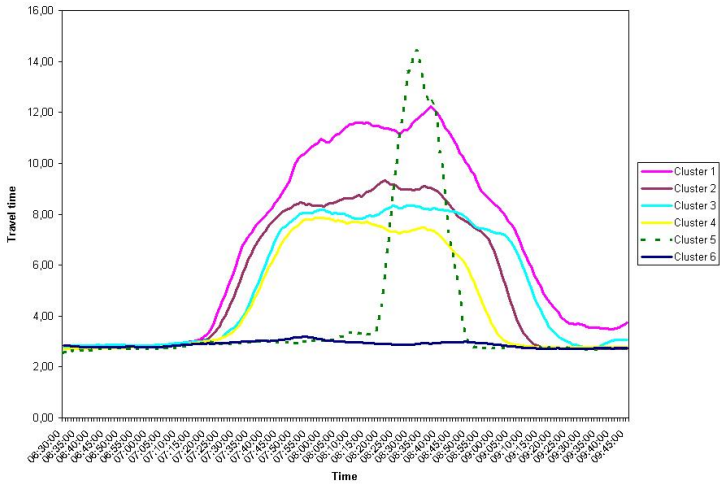

Figure 10.10: 6 clusters - motorway segment 10051006

The formed clusters with $K = 7$ are shown in Figure 10.11. Cluster 5 is identical to cluster 5 in Figure 10.10. Cluster 6 corresponds to the green traffic pattern in Figure 10.8. The remaining clusters are identical to clusters 1, 2, 3, 4 and 5 with $K = 5$, except for the change in the number of traffic patterns in cluster 6. The percentage of variance explained increases by 1 % in that the added cluster is a single day, which with $K = 5$ was assigned inappropriately.

The formed clusters with $K = 8$ are shown in Figure 10.12. Cluster 5 is identical to cluster 5 in Figure 10.10. Cluster 6 is identical to cluster 6 in Figure 10.11. Cluster 7 consists of the red and lavender traffic patterns in Figure 10.8. The remaining clusters are identical to clusters 1, 2, 3, 4 and 5 with $K = 5$ except for the change in the number of days in cluster 7. There is no gain in the percent

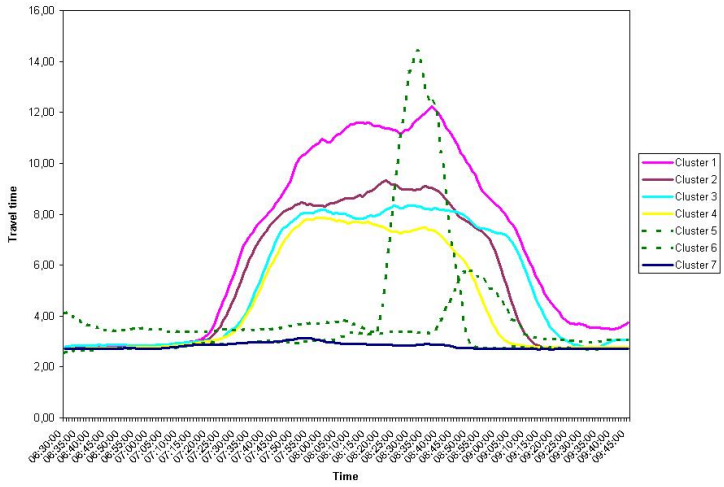

Figure 10.11: 7 clusters - motorway segment 10051006

of variance explained (see Figure 10.2).

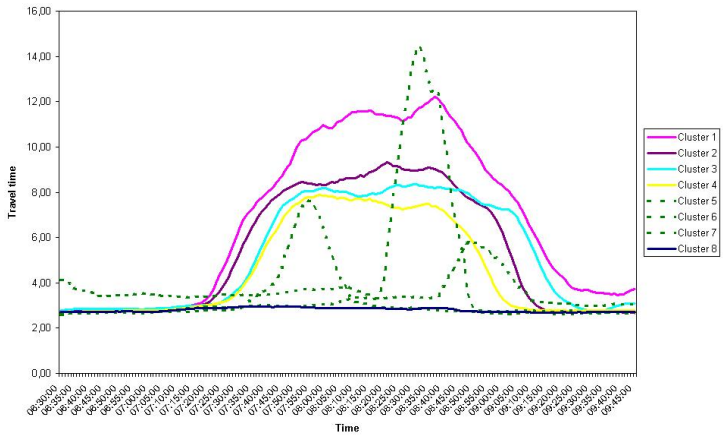

Figure 10.12: 8 clusters - motorway segment 10051006

The results show that increasing the number of clusters only marginally influences the formed clusters with $K = 5$. Traffic patterns that were classified as incidents and patterns that stood out from the crowd have been segregated.

10.4 Concluding remarks

The enhanced k-Means algorithm was applied in order to determine whether the available traffic patterns could be grouped into a fixed number of representative patterns. The optimal number of representative patterns was determined by use of the within sum of squares function and elbow criterion. The estimation of the number of clusters was conducted on a single training set, no independent tests sets were utilized for the validation of the estimate. This is due to the fact that this cross-validation technique cannot be utilized in the context of clustering in that the within sum of squares function value would also decrease with increasing the number of clusters in the model [24].

Detailed cluster analysis was conducted for the estimated optimal number of clusters. It showed that during fall recess, winter holidays and on the majority of Fridays travel times are fairly constant, and the traffic resembles free flow conditions. The other clusters had approximately the same shape but different plateau levels. For the presented distribution of traffic patterns quite plausible explanations exist. The assumption that business days resemble each other was dismissed as there were no well-separated groups of Mondays, Tuesdays, Wednesdays or Thursdays. The previously made assumption that traffic flow on Fridays differs from traffic flow on the other business days was partially disproved as Fridays were also mixed with other business days. Moreover, fall recess and winter holidays were visible and grouped together in the same cluster along with other traffic patterns that exhibited travel times at the speed limit. Incident traffic patterns either formed their own cluster or were not affiliated to any clusters. Furthermore, a vague seasonal effect was observed, which will require further investigation once the amount of data in the historical data warehouse permits it.

The advantage of using the enhanced k-Means algorithm is that no assumptions need to be made about the data a priori. All traffic patterns can enter the analysis on an equal footing. There is no need to store exogenous attributes about each pattern. It was shown that all these effects are elucidated automatically.

Chapter 11

Forecasting

11.1 Introduction

Chapter 10 showed that the 10-minute moving average travel time data could be divided into a fixed number of representative traffic patterns. This suggests that a new traffic pattern could perhaps be compared to the pool of representative traffic patterns to identify which pattern it resembles. For now, information about these patterns will be used to determine whether it can be used in the context of travel time forecasting.

11.2 Data preparation

The input data set consists of 86 traffic patterns in the interval between 06:30:00 and 09:45:00, corresponding to 196 features. It is identical to the data set used to perform clustering in Section 9.3. No data preprocessing such as the removal of previously detected incidents has taken place. The input data set is divided into a training set (50 % of the input data) and a test set (50 % of the input data). It is a prior assumption that the amount of available data in the historical data warehouse is sufficient to make a fairly informed guess about the optimal

number of clusters for travel time forecasting. The training set will be used to estimate representative patterns. The test set will be used to estimate the optimal number of clusters. Subsequently, the test set will become the training set and the training set will become the test set, and the estimation process will be repeated. The patterns are distributed randomly between the training and test sets. Business days are uniformly distributed between the training set and the test set. This is, however, not expected to influence the performance of the clustering algorithm given that business days were distributed between all clusters. The discovered season effect, albeit vague, has not been accounted for. The enhanced k-Means algorithm will be run twice - once for each training set. Model performance will also be evaluated twice - once for each test set. Data will be clustered into K clusters ranging from 2 to 15 clusters (see Section 10.3.3 for an explanation). Hence 14 models will be trained and subsequently tested. This approach has been chosen in order to test the applicability of clustering for forecasting purposes. If deemed applicable, the k-Means algorithm will only be run on the training set in the future, after which the estimation of the optimal number of clusters will be conducted on the test set. Adjustments will also most likely be made to the distribution ratio of input data between training and test sets.

11.3 Assumption

The motorway segments on Hillerødmotorvejen will be considered independent. This means that the state of neighboring segments will not be taken into account when determining future travel times for motorway segment 10051006.

11.4 The forecast algorithm

The models are evaluated based on the clusters found in the training set. Clusters, which contain a single traffic pattern, are excluded from model testing and the forecasts are performed on the remaining clusters in the model. This is due to the fact that a single day most likely represents an exceptional event, and hence is of no immediate interest for further analysis in that the occurrence of each exceptional event is unique. The occurrence of an incident would in most cases trigger the disablement of the forecast functions as it is unlikely that the formed clusters would be able to capture it. The cluster centroids are stored in a $196 \times K$ table of real numbers, each representing a centroid measurement for a point in time (row) in the interval between 06:30:00 and 09:45:00 and a cluster ID running from 1 to K (column). These tables are generated by running the

enhanced k-Means algorithm on the training set - once for each number of clusters. 14 such tables are generated for each training set for evaluation purposes. This number will be reduced to one, once the optimal number of clusters has been selected and the model is ready for deployment. Cluster centroids and the 10-minute moving average travel times are used as input to the clustering algorithm. The forecasting step is put to 1 minute, which means that the 10-minute moving average travel time will be recalculated each minute. The forecasting horizon is put to 15 minutes, meaning that the produced forecast will be in force 15 minutes later. The forecasting step and horizon have been selected as per requirements for the new traffic reporting system. The proposed algorithm covers the following steps:

Calculate the distance between the 10-minute moving average travel time values to each of the cluster centroids at time t. The squared error has been chosen as the distance metric and is calculated from the following formula: $(x_{observed_t} - x_{cluster_t})^2$, where $x_{observed_t}$ is the 10-minute moving average travel time value and $x_{cluster_t}$ is the cluster center value at time t.

Select the cluster ID K, which has the smallest squared error.

Use this cluster as a forecast cluster for the expected travel time, such that the centroid value of this cluster offset 15 minutes from time t becomes the forecasted travel time at time $t + 15$.

The size of the squared error is evaluated for different window functions. This means that the calculation of the squared error now includes the squared error of the current time as well as a summation of the squared errors for a number of time points preceding the current time. The accumulated squared error is calculated from the following formula: $\sum_{i=1}^{T}(x_{observed_i} - x_{cluster_i})^2$, where T is the number of time points preceding the current time, $x_{observed_i}$ is the observed travel time and $x_{cluster_i}$ is the cluster centroid value. The value for T is set to 10 minutes. Moreover, all times which have elapsed since 06:30:00 including the current time are included in the calculation - in this case the time window is termed unbounded. However, it is a prior assumption that the accumulation of the preceding 10-minute moving average travel time values, regardless of the accumulation window, might not have an impact on the 15-minute forecasts as they already embody an average of the aggregated travel times for 10 consecutive minutes preceding and including the current value.

11.5 Evaluation criteria

The performance of the model will be evaluated by comparing the mean squared error and the percentage of errors between the 10-minute moving average travel time and the 15-minute forecast that exceed two and five minutes, respectively. Mean squared error (MSE) measures the expected value of the square of the error, which is the amount by which the observed travel time on average differs from the predicted travel time [25] and is calculated from the following formula: $MSE = \frac{\sum_{i=1}^{T}(x_{observed_i} - x_{pred_i})^2}{N_{dataset}}$, where $x_{observed_i}$ is the 10-minute moving average travel time value at time t, x_{pred_i} is the forecasted travel time value and $N_{dataset}$ is the number of observations in the dataset. The values of the mean squared error might give a misleading impression as they, in fact, represent the average over all time points in the examined time interval. The largest values of the mean squared error are expected to occur under congestion build-up and phase-out, and during localized queuing of short duration due to the fact that the traffic flow exhibits major fluctuations during these time periods and hence are harder to forecast accurately. The percentage of error between the 10-minute moving average travel time values and the forecasted travel time value that exceeds 2 minutes (small errors) and 5 minutes (large errors) will also be calculated. These measures quantify the amount of discrepancies as a function of the total number of input features. The determination of small errors enables the practitioner to evaluate the quality of forecasts on a microscopic level. The determination of large errors is important in that these errors will be recognized by the drivers. Moreover, the forecast function will be disabled if the difference between the 10-minute moving average travel time value and the forecasted travel time value exceeds 5 minutes.

11.6 Results

It can be seen from Figure 11.1 and 11.2 that, based on the values of MSE for the various time windows, the best travel time forecasts are achieved by using a time window of zero. This means that the performance of the model only depends on the current observation and the immediate history of observations need not be taken into account for future travel time estimation. This result was anticipated as the 10-minute moving average travel times already embody the immediate history of observations. For these reasons all further work with regard to travel time forecasting for this motorway segment will be done using a time window of zero. The reddish brown curve shows the model with five clusters. This model has been highlighted because it was shown in Section 10.3 that, based on the results of cluster validation techniques, five clusters could

be used as a lower bound for the optimal number of clusters for this motorway segment.

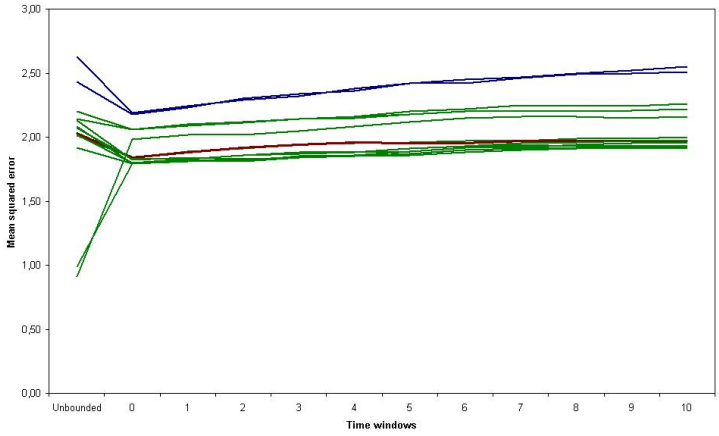

Figure 11.1: Mean squared error - test set 1 motorway segment 10051006; the blue curves represent models with 2 and 3 clusters; the green curves represent models with 4 clusters and above; the reddish brown curve represents a model with 5 clusters

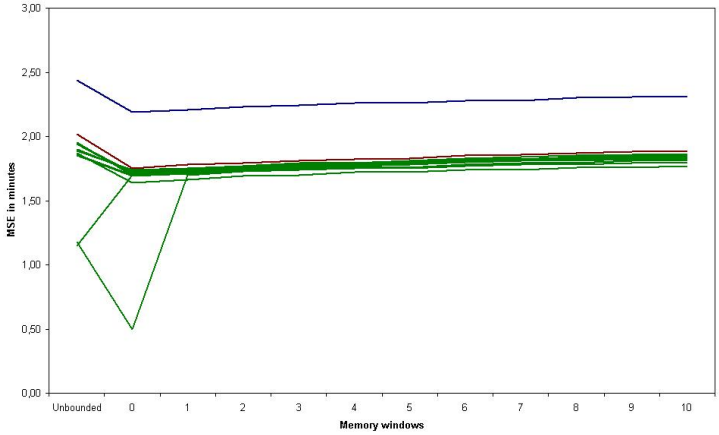

Figure 11.2: Mean squared error - test set 2 motorway segment 10051006; the blue curve represents a model 2 clusters; the green curves represent models with 3 clusters and above; the reddish brown curve represents a model with 5 clusters

Figure 11.3 shows the values of MSE for both test sets. It can be seen that the values of MSE are larger for test set 1 than test set 2. This discrepancy can most likely be ascribed to the fact that the formed clusters in training set 1 do not represent the data in test set 1 very well, and that the formed clusters in training set 2 are more in sync with the data in test set 2.The season factor which was observed when performing clustering on all 86 traffic patterns might have influenced the outcome of these tests. The best forecast performance is achieved by employing models where the number of clusters is 7 for test set 1 and 4 for test set 2 (except for 13 clusters), after which the values of MSE level off. This can be attributable to the fact that the algorithm has a tendency to segregate single daya and put them in an independent cluster when the number of clusters is increased. This does not impact travel time forecasts as clusters, which consist of a single day are excluded from the modeling process (see Section 10.3.6 for discussion).

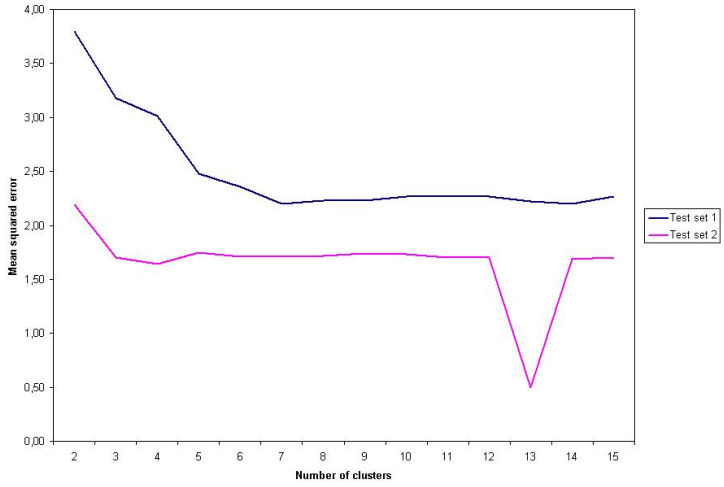

Figure 11.3: Mean squared error - motorway segment 10051006

Figure 11.4 and 11.5 shows the percentage of errors exceeding two and five minutes, respectively. The percentage of errors exceeding two minutes is smallest when the number of clusters is 4 and 3, respectively. The percentage of errors exceeding five minutes is smallest when the number of clusters is 10 and 15, respectively.

Figure 11.6, 11.7, 11.8 and 11.9 illustrate the results of applying the forecasting algorithm to a number of days from both test data sets.

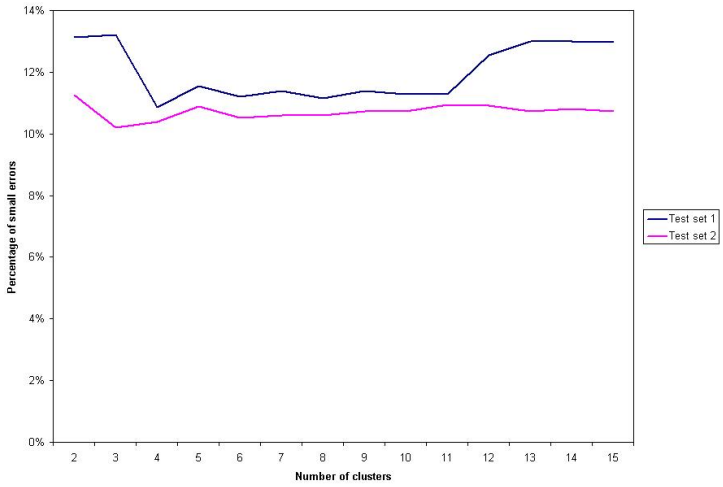

Figure 11.4: Percentage of small errors - motorway segment 10051006

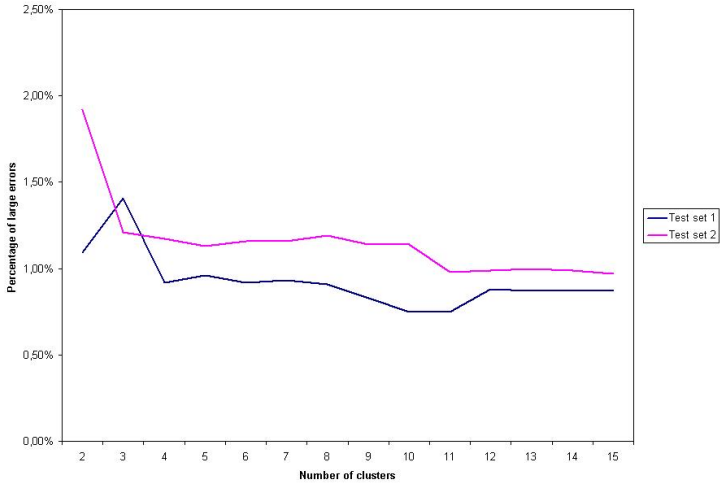

Figure 11.5: Percentage of large errors - motorway segment 10051006

The results are illustrated with 4 (minimum percentage errors $>$ 2 minutes) and 7 (minimum MSE) clusters for test set 1, and with 3 (minimum percentage errors $>$ 2 minutes) and 4 (minimum MSE) clusters for test set 2. Curves for

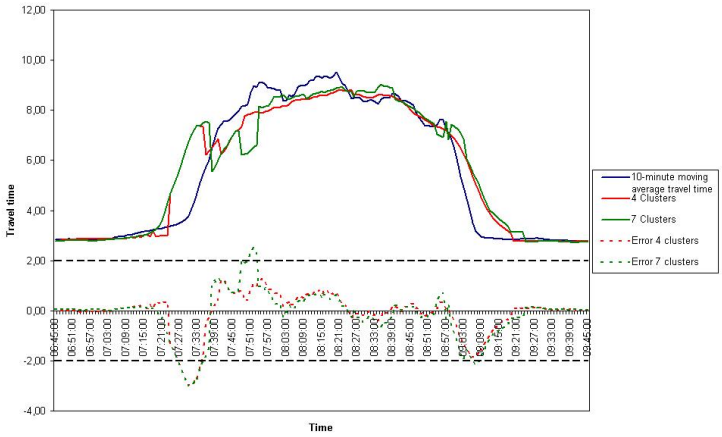

Figure 11.6: Test set 1 - 10-minute moving average travel times vs. forecasts using models with 4 and 7 clusters - Friday 26-01-2007

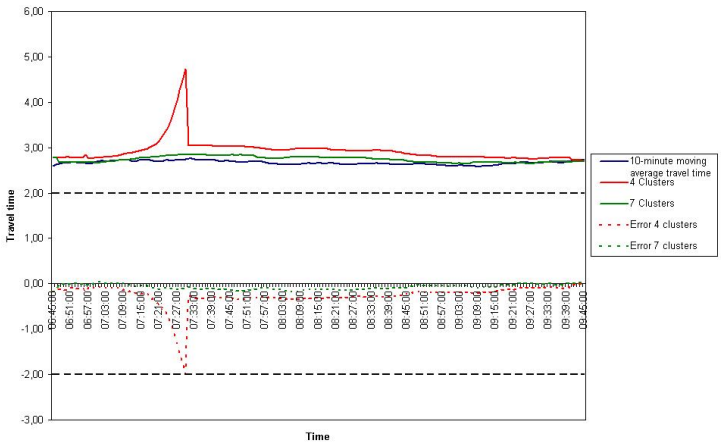

Figure 11.7: Test set 1 - 10-minute moving average travel times vs. forecasts using models with 4 and 7 clusters - Friday 16-02-2007 (winter vacation)

the difference between the 10-minute moving average travel times (hereinafter actual travel times) and the forecasted travel times have also been included. Results with 10 and 15 clusters (minimum percentage errors $>$ 5 minutes) for test set 1 and 2, respectively, have not been included. This is due to the fact that the number of large errors is small and fairly constant for all models, except for

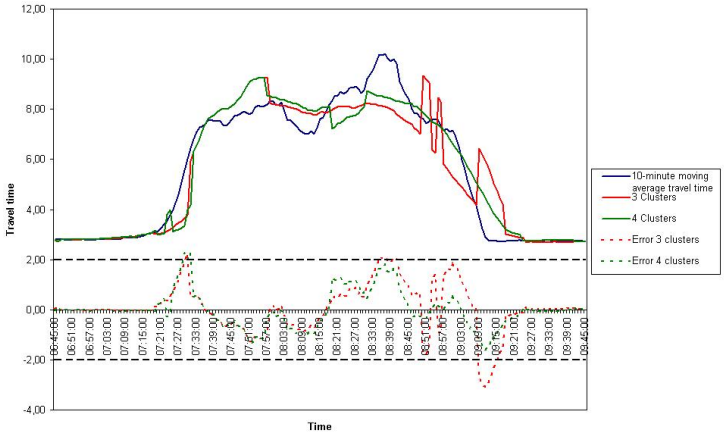

Figure 11.8: Test set 2 - 10-minute moving average travel times vs. forecasts using models with 3 and 4 clusters - Thursday 16-11-2006

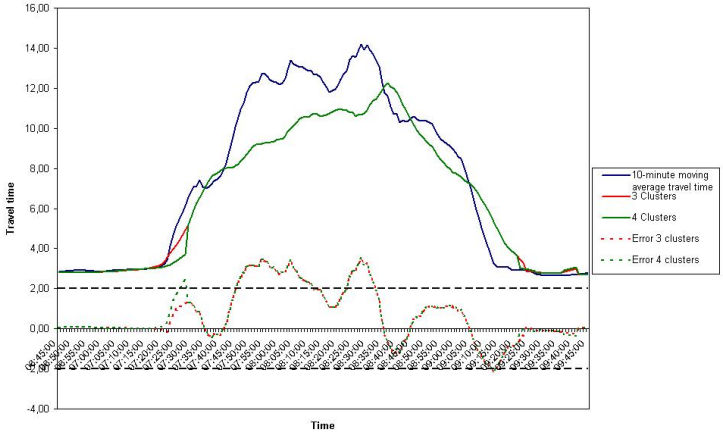

Figure 11.9: Test set 2 - 10-minute moving average travel times vs. forecasts using models with 3 and 4 clusters - Monday 06-11-2006

the first two models. The majority of deviations are in the range of 0-2 minutes and the difference between using 4 or 7 clusters in the model for test set 1, and 3 or 4 clusters in the model for test set 2 is negligible. In most cases the discrepancies between the actual and forecasted travel time values occur under congestion build-up and phase-out. It can be seen that the forecasting

algorithm has a tendency to slightly underestimate the forecasted travel times. The forecasts also have a tendency either to lag behind or be ahead, especially during congestion build-up and phase-out. There are, however, no systematic differences between the actual and the forecasted travel time values, and it takes usually less time than the forecast horizon to contain severe divergences between the actual and the forecasted values. Friday 16-02-2007 (see Figure 11.7) has been included to illustrate the effect of using clustering as means for travel time forecasting on a traffic pattern which is a business day and a winter holiday simultaneously. The latter factor was not taken into account when forecasting travel times for this traffic pattern. The model with 4 clusters does not initially pick up the intensity of the traffic flow as the forecasted travel times begin to rise around 07:20:00. This mishap is, however, remedied approximately 10 minutes later, after which the forecasted travel times are in sync with the actual travel times. A forecast model solely based on modeling this day as any other business day (or as any other Friday) would perhaps be unable to detect this feature. Figure 11.7 and 11.8 have rapid upward and subsequently downward shifts in the forecasted travel times. This is for two reasons. First, if the traffic flow on a given day is between two clusters the algorithm will shift between these clusters and, as a consequence thereof, oscillations will occur. Second, clusters overlap under free flow conditions, which give rises to discrepancies between the actual and the forecasted travel time values right after the onset of congestion build-up and in the final stages of congestion phase-out. The algorithm does, however, usually take remedial action immediately, after which the discrepancies are reduced or eliminated. The results of the forecasting analysis are not directly comparable to the results obtained in the studied bibliography for a number of reasons. First, the forecasting steps and the forecasting horizons are different. Second, statistical and quantitative measures that are used for model validation are different. Third, the starting point for each study in terms of dividing the day into morning/afternoon rush hour, the week into business days and the removal/retain of incident patterns is different. The results can, however, be put into perspective by comparing them with the forecast model developed in connection with the extension of the M3 motorway [3]. The starting point for the development of the forecast algorithm is the same - the days were also divided into a morning and an afternoon rush hour period of approximately the same length, and the forecasting step and horizon were the same. There are, however, also a series of dissimilarities. First, a different model was fit for each business day in comparison to fitting one model for all business days regardless of any exogenous factors. Second, all incident traffic patterns were removed before model parameters were estimated. And the results were also profoundly different. The M3 forecast algorithm did not have the ability to automatically elucidate exogenous effects. The presented forecast algorithm looks only at the actual travel times without taking account of any exogenous factors and elucidates all effects automatically.

11.7 Implementation issues

Section 11.6 showed that the proposed forecast algorithm can be utilized for the purpose of travel time forecasting. To extend the application of the forecast algorithm to all segments in the motorway network, a method of approach needs to be worked out in order to automate the evaluation of the performance of the prospective forecast models in terms of choosing the model which will be implemented in the application. In the following guidelines for selection of input data for training and test, and the optimal number of clusters will be presented. These guidelines have been worked out based on the results in Section 11.6. The amount of data currently residing in the historical data warehouse has also been taken into account given that implementation will be conducted upon completion of the thesis.

11.7.1 Selection of input data

All available traffic patterns in the historical data warehouse will be utilized as input for training and test. The distribution ratio will be changed to 80 % for training and 20 % for test. 80 % of the data in the historical data warehouse should enable the enhanced k-Means algorithm to detect all representative patterns. The remaining 20 % is deemed sufficient to make an informed guess about the optimal number of clusters based on the representative patterns. Patterns for training and testing will still be chosen at random due to the fact that only vague season effects were observed. The number of prospective forecast models will initially be kept at 14. This number might increase as the amount of data in the historical data warehouse grows and a larger spectrum of traffic patterns becomes available.

11.7.2 Selection of the best forecast model

It was decided that the optimal number of clusters will be chosen when the percentage of errors exceeding 2 minutes is at a minimum. This is due to the fact that the majority of errors were found in the interval between 2 and 5 minutes. It is, however, possible that this strategy will not work equally well for some of the remaining motorway segments. This is yet to be tested out. In worst case scenario the practitioner might be forced to think along entirely new lines. It is important to remember that these guidelines have been worked out based on the situation as it was in March 2007. As time goes by and the amount of available data in the historical data warehouse grows, an entirely different

approach might be called for. These guidelines should be revised whenever a recalibration of the model takes place.

11.8 Other methods for travel time forecasting

The Road Directorate does not presently make use of any tools, aside from the employed verification techniques in Section 11.5 that would enable the writer to verify the quality of the forecasts made by the proposed forecast algorithm. This is due to the fact that the travel times are presently not forecasted per se. However, a number of trivial forecasting methods for future travel times have been suggested that can be utilized for comparison with the proposed forecast algorithm. Furthermore, the quality of the forecasts produced by these methods is also of interest in terms of determining whether they can be used for reporting purposes. This is of major relevance in case data relay from the road stations is interrupted. If so, the current assessment is that no forecasts would be announced on the website. However, if the quality of the forecasts produced by one of these methods (if any) is deemed acceptable, an estimate for the 15-minute forecast value could be provided to the end users as a substitute for the forecasts produced by the proposed algorithm. Following are the proposed forecast methods:

Speed limit forecast method. This method suggests making use of the speed limit and using it as a forecast. This method assumes that the future travel times can always be modeled as travel times in free flow conditions thus completely ignoring the fact that travel times during morning rush hour substantially differ from travel times at free flow. For this reason they are of little or no value for future travel time forecasting during the morning rush hour due to heavy congestion. This method has been included as it is widely used in various route finders and navigation systems for travel time estimation.

Historical travel time forecast method. Historical travel times are long-term average travel times for the specific time of day of the forecast. The historical travel time forecast method suggests making use of the average travel times as measured on all working days at every minute between 06:45:00 and 09:45:00 and using them as a forecast. The historical travel times have been computed based on the same dataset as used for clustering. This forecast would be exact, if future travel times were equal to their historical averages, and furthermore, the current and past travel times had no impact on future travel times. This method completely ignores the current traffic situation.

Current travel time forecast method. This method suggests making use of the current 10-minute moving average travel time value and using it as a forecast. This prediction would be exact if future travel times were equal to the current ones. Current travel time forecast method is the only method which accounts for the current traffic situation; however this method only accounts for the situation right now and assumes that the situation in the future resembles the situation right now. This assumption is, however, valid in free flow conditions but as congestion starts building up, the instantaneous travel time starts lagging. The method suffers from the lack of a built-in means that would account for the traffic flow in the future.

The following measures will be calculated to assess the performance of the previously mentioned forecast methods: the mean squared error (MSE) and the percentage of error between the 10-minute moving average travel times and the forecasted travel times exceeding two and five minutes, respectively. A description of these methods can be found in Section Section 11.5. Table 11.1 shows that the best performance is achieved by using the proposed forecast algorithm. The current travel time method is only slightly inferior to the proposed forecast algorithm. Historical travel time forecast method comes in third. This reconfirms the belief that the historical travel times are a weak estimate of the current traffic situation as the travel times differ considerably between working days and weeks. Speed limit forecast method has the worst performance.

Motorway segment	Speed limit	Historical travel times	Current travel time	Test set 1	Test set 2
10051006	13	4,31	2,56	2,20	1,64

Table 11.1: MSE for current travel time prediction methods

Table 11.2 shows the percentage of encountered errors between the 10-minute moving average travel times and the forecasted travel times which exceed two minutes. The results bear resemblance to the reached conclusions based on inspection of Table 11.1. The ranking of the methods is the same as before. The percentage of errors exceeding two minutes doubles when using the current travel time method for forecasting in comparison to the proposed forecast algorithm, whereas the difference in performance between the proposed method and the current travel time method was in the range of a few percent. The gain from using the proposed forecast method compared to the historical travel times and speed limit methods is threefold. Table 11.3 shows the percentage of errors between the 10-minute moving average travel times and the forecasted travel times which exceed two minutes. The proposed forecast algorithm outperforms the trivial forecast methods.

Motorway segment	Speed limit	Historical travel times	Current travel time	Test set 1	Test set 2
10051006	28,1	27,87	18,1	10,87	10,21

Table 11.2: Percentage of errors exceeding 2 minutes

Motorway segment	Speed limit	Historical travel times	Current travel time	Test set 1	Test set 2
10051006	2,17	1,51	1,36	0,92	1,13

Table 11.3: Percentage of errors exceeding 5 minutes

It can be discussed whether the trivial methods can be used as substitutes in case forecasts cannot be estimated. Speed limit and historical travel times method are not acceptable as substitutes due to their poor performance in the context of travel time forecasting. The performance of the current travel time method is comparable to the proposed forecast algorithm on average as there is not much difference between the values of MSE. However, due to the fact that the increase in the percentage of errors exceeding 2 minutes is twofold, this method is not recommended for the purpose of substitution.

11.9 Model recalibration

Traffic flow on some motorway segments might change with time due to changing seasons, road works or changes in travel patterns etc. For these reasons, from time to time, a recalibration of model parameters will be under consideration. The recalibration process includes the recreation of representative traffic patterns and the estimation of the optimal number of clusters, after which the table that stores the cluster centroid values is updated. A method of approach is called for that would automatically trigger the recalibration of model parameters. An initial suggestion would be to track the evolution in the percentage of errors exceeding 2 minutes. This method was used for the selection of the optimal number of clusters. An appropriate threshold value needs to be determined that can be used to monitor the performance of the model. This threshold value could be put to the minimum value of percentage of error exceeding 2 minutes (as per the current number of clusters in the model). In that this value is subject to uncertainty a few per cent need to be added for slack. When the percentage of errors exceeds this threshold value a warning can be issued to the practitioner to initiate model recalibration. A procedure could also be set up that automatically recalibrates the model parameters. An automatic recalibration would be associated with zero costs, once a routine for conducting it is implemented. This

is due to the fact that all routines that handle the data from data collection to generation of the aggregated travel time values at 1-minute intervals are already an inherent part of the data warehouse.

11.10 Concluding remarks

Clustering was utilized in the context of travel time forecasting. An algorithm was proposed that outperformed other trivial methods for forecasting. It was shown that one forecast model can be utilized to make forecasts for all business days, meaning that there is no need to prepare separate models for each business day or for handling vacations and incident patterns. This implies that it is not necessary to preprocess the input data (aside from deselecting traffic patterns with missing values) before creating the representative traffic patterns. The clustering algorithm will take care of grouping traffic patterns together based on the intensity of traffic. This means that there is no need to keep tabs on an event log for the purpose of input data preprocessing, which can be cumbersome. The inspection of single traffic patterns before forming clusters can be time consuming as the amount of data in the historical data warehouse grows. The simplicity of the forecast algorithm in terms of the required input means that its implementation is straightforward. This is due to the fact that there is only a need to store the values of the cluster centroids which can be stored in a table per motorway segment. Furthermore, it is only necessary to store the 10 latest aggregated travel times immediately preceding the current travel time value. The calculation of the squared error is computationally cheap which means that the forecasts will be made in real-time as per requirements. The percentage of error exceeding five minutes is small, which means that the forecast function would have an excellent service availability as the algorithm would be switched off only a fraction of the time. The recalibration of this algorithm is also straightforward and can be conducted any time or when the percentage of 2 minute errors exceeds the specified threshold value. There are, however, also a number of deficiencies. The accuracy of the forecasts might be comprised as a result of clustering and due to the fact that the input values are the 10-minute moving average travel times. Moreover, the forecasts tend to lag behind or ahead, especially under congestion build-up and phase-out. Furthermore, if the traffic pattern lies in the interval between two clusters, oscillations in the forecasted travel times will occur. In addition, it has been assumed that motorway segment 10051006 is stochastically independent from segment 10041005, which means that the forecasting algorithm only takes into account the situation on one segment at a time, thus completely ignoring the traffic on the other segment.

Chapter 12

Future work

There are a number of issues pertaining to further development and perhaps the improvement of forecasts which, for the time being, will remain unsolved. Before modeling began, it was assumed that the aggregated segment travel times were stochastically independent. However, it is obvious to hypothesize that their covariance should also be identified. This is due to two reasons: first, theoretically it can be expected that their use would result in increased forecast accuracy, because the state of the neighboring segments would be taken into consideration when determining future travel times; second, it should be noted that the forecasting of segment travel times is not only an end product in itself, but also an input to the route travel time forecasting application. The accumulation of forecasted segment travel times across segments might result in misleading forecasts in that travel time forecasts for an individual segment are already subject to some degree of uncertainty per se.

Increasing the forecast horizon to 30-minutes is another topic of interest that requires further study. It is expected that the performance of the proposed forecast algorithm will deteriorate as the forecasting horizon is increasing. Furthermore, the appropriate level of aggregation of the accumulated 1-minute measurements for speed and vehicle count should be investigated. It was assumed a priori that the aggregation level should be the 1-minute segment travel times. The choice of the forecasting step should also be subject to further research. It can be hypothesized that the quality of the produced forecasts will improve by increasing

the level of aggregation or by increasing the interval in which the forecasts are made. The studied bibliography showed that all of the studies utilized higher levels of aggregation. This would also be relevant when developing the 30-minute forecast model in order to minimize the degree of uncertainty. The smoothing interval of the 1-minute aggregated travel time values was chosen based on visual inspection of the smoothed out travel time curves. The optimal size of the smoothing interval should be determined, after which forecast performance can be assessed. Experiments using exponential smoothing functions should also be conducted. There are also a number of issues that will remain unresolved in the short term. The season effect should be investigated when the amount of data in the historical data warehouse permits it. Furthermore, the creation of more sophisticated methods for data cleaning and repair should be looked into. It is hypothesized that applicable solutions cannot be proposed until the amount of data in the historical data warehouse is larger.

Chapter 13

Conclusion

The main objective of this thesis was to develop a universal algorithm for forecasting travel times 15 minutes ahead in time which was going to be embedded in the new real-time traffic reporting system. Although the main focus was on developing a forecast algorithm, the process was not exclusively confined to selecting an appropriate algorithm and estimating model parameters. The Road Directorate had outlined a number of requirements primarily pertaining to computational performance, data handling and model deployment, which had to be honored in this work. The preparation of input data was an issue of special importance. Development of operational scenarios for model deployment and recalibration were also requested. Moreover characteristics in modeling such as consideration of the type of input data, the type of desired output and the quality of data, which are factors that strongly affect the ability of the forecasting algorithm in providing accurate and efficient forecasts, needed to be taken into consideration. These requirements were prepared to ensure that the recommended solution was operational in a practical framework. The initial scope of activities was comprehensive if all issues involved were going to be closely examined. Consequently emphasis was put on developing a product as a result in which practicability and operability rather than theoretical research was made a priority. In order to accommodate the requirements a conceptual project outline was developed, which charted the course for the tasks that needed to be accounted for before, during and after model development. First, this included the development of a supporting framework in which all data handling was going

to be conducted. It was decided that all processes pertaining to data handling would be confined to the Oracle Database. A real-time data and a historical data warehouse were built up for transforming the collected data into a format that could be used as input for forecasting. Oracle Data Mining was utilized for data understanding purposes and for model building and evaluation. This tool was chosen in order to streamline the modeling process because it is embedded in the Oracle Database where the data reside. Clustering was chosen for the purpose of data understanding. The start-up phase was somewhat challenging due to the fact that the documentation about the clustering algorithms was scanty. The scope of the implemented algorithms was unclear. Oracle Technology Network was utilized in order to gain more insight on how to set and tune the parameters, which was required before the algorithms were run. Apart from that the algorithms were tested out on a trial and error basis. Clustering gave insight into how the input data could be structured (or handled) for the purpose of travel time forecasting. It was demonstrated that preprocessing of input data was rendered needless, as all exogenous effects were elucidated automatically. Four possible exogenous variables were investigated: the effect of working days, of seasons, of vacations and incidents. The applicability of clustering in the area of travel time forecasting was evidenced. A simple and flexible algorithm forecasting algorithm was proposed. The only parameter that needs to be determined for each motorway segment is the number of clusters in the model. The cluster centroids are stored in a table, which basically constitutes the model. Suggestions for model recalibration were proposed. This can be conducted at any time. The simplicity and the flexibility of the forecast algorithm means that a forecast model can be worked out for all motorway segments, even if there is no immediate justification for that, which would be the case if the variation in the aggregated 10-minute moving average travel times during the morning (and afternoon) rush-hour is insignificant. This will without doubt facilitate the preparatory work pertaining to model building and streamline later model deployment. The results were satisfactory. The amount of large errors was deemed insignificant. The amount of small errors was acceptable. Although the forecasts under congestion build-up and phase-out involved a certain amount of uncertainties, there is no doubt that the obtained results are better than the previously gained knowledge in the Road Directorate about travel time forecasting, as a result of which at this stage the proposed algorithm is going to be implemented "as-is". The utilization of clustering in travel time forecasting has shown that satisfactory results can be achieved by a relatively simple model. The amount of data which was available in March 2007 was sufficient in order to obtain workable results. However, as more data becomes available the forecast performance of the proposed forecast algorithm can be improved. The strength of this project is that satisfactory results were obtained even though the main emphasis was put on practicability rather than model complexity. The flow of data from data collection to model deployment was considered. A sound knowledge of Oracle Data Mining as a prospective tool for data modeling was achieved

despite start-up difficulties. One additional feat of note is that it was impossible to find a single thread in the Oracle Technology Network data mining discussion forum that even remotely approached a success story in terms of applying the clustering algorithm in a commercial application. This seems to indicate that perhaps the application of data mining in Oracle data warehouse environments is still in its early stages. The Road Directorate has given approval to implement the proposed forecasting algorithm into the new traffic reporting system. The outlined strategies for model selection and model recalibration will also be put into practice.

Bibliography

[1] www.trafikken.dk/wimpdoc.asp?page=document&objno=77436

[2] Wendelboe, J. T. 2006, 'Rejsetidsprognoser for Motorring 3 - Evaluering', Internal memorandum, The Road Directorate (contact person Ieva Bak)

[3] Dehlendorff, C. 2006, 'Prognosemodel for M3', Internal memorandum, The Road Directorate (contact person Ieva Bak)

[4] *Oracle Database*, Available at http://www.oracle.com/database/index.html

[5] Loubes, J., Maza, E. & Lavielle, M. 2003, 'Road trafficking description and short term travel time forecasting, with a classification method', *The Canadian Journal of Statistics*, Vol. 31, No. ?, Pages ???-???

[6] Nikovski, D., Nishiuma, N., Goto Y. & Kumazawa, H. 2005, 'Univariate Short-Term Prediction of Road Travel Times', *IEEE Intelligent Transportation Systems Conference*, Vienna, Austria

[7] Chung, E. Year ????, 'Classification Of Traffic Pattern', Center for Collaborative Research, University of Tokyo

[8] Wu, C., Wei, C., Su, D., Chang, M. & Ho, J. 2003, 'Travel Time Prediction with Support Vector Regression', *IEEE*(Unknown)

[9] *Oracle Data Mining Concepts*, 2005, Oracle Technology Network, 10g Release 2, Available at download-uk.oracle.com/docs/pdf/B14339_01.pdf

[10] Steria, 'IT System for M3 - Interface Control Document M3/PROGNOSES', 2005, Internal memorandum, The Road Directorate (contact person Ieva Bak)

[11] Hobbs, L., Hillson, S., Lawande, S. & Smith, P. 2005, *Oracle Database 10g Data Warehousing*, Elsevier Digital Press, Burlington, MA, USA, Pages 293-297

[12] *Oracle Data Mining Concepts*, 2005, Oracle Technology Network, 10g Release 2, Available at download-uk.oracle.com/docs/pdf/B14339_01.pdf, Chapter 2-2, Page 17

[13] *Exponential Smoothing*, Available at en.wikipedia.org/wiki/Exponential_smoothing

[14] *Oracle Data Mining Concepts*, 2005, Oracle Technology Network, 10g Release 2, Available at download-uk.oracle.com/docs/pdf/B14339_01.pdf, Chapter 4-2, Page 36

[15] *Oracle Data Mining Concepts*, 2005, Oracle Technology Network, 10g Release 2, Available at download-uk.oracle.com/docs/pdf/B14339_01.pdf, Chapter 4-2, Page 37

[16] *Oracle Data Mining Forum*, Available at forums.oracle.com/forums/forum.jspa?forumID=55&start=0

[17] *Oracle Data Mining Forum*, Available at forums.oracle.com/forums/thread.jspa?messageID=1583611

[18] Hastie, T., Tibshirani, R. & Friedman, J. 2003, *The Elements of Statistical Learning*, Springer-Verlag, Canada, Pages 478-480

[19] *Finding the Most Typical Record in a Group*, Marcos. M. Campos, Available at http://oracledmt.blogspot.com/2006/07/finding-most-typical-record-in-group.html

[20] Hastie, T., Tibshirani, R. & Friedman, J. 2003, *The Elements of Statistical Learning*, Springer-Verlag, Canada, Page 472

[21] *Oracle Data Mining Application Developers Guide*, 2005, Oracle Technology Network, 10g Release 2, Available at download-uk.oracle.com/docs/pdf/B14340_01.pdf, Chapter 3-8, Page 34

[22] Hastie, T., Tibshirani, R. & Friedman, J. 2003, *The Elements of Statistical Learning*, Springer-Verlag, Canada, Pages 461-462

[23] *Data Clustering*, Available at en.wikipedia.org/wiki/Data_clustering

[24] Hastie, T., Tibshirani, R. & Friedman, J. 2003, *The Elements of Statistical Learning*, Springer-Verlag, Canada, Pages 471

[25] *Mean Squared Error*, Available at en.wikipedia.org/wiki/Mean_squared_error