Limburgs
Universitair
Centrum

Modelling Activity-Diary Data: Complexity or Parsimony?

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Wetenschappen, Richting Wiskunde
aan het Limburgs Universitair Centrum te verdedigen door

ELKE MOONS

Promotor
Prof. dr. G. Wets
Copromotor
Prof. dr. M. Aerts

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This chapter provides an introduction to activity-based and to mode-choice models and an overview of this dissertation. Section 1.1 shortly reports the history of transportation modelling, while Section 1.2 gives a short discussion on 'complexity or parsimony' and on the problems that motivate this dissertation. Section 1.3 describes the organisation of the subsequent chapters.

1.1 History of Activity-Based Models

Modelling traffic patterns has always been a major area of concern in transportation research. Since 1950, due to the rapid increase in car ownership and car use in Western Europe and in the US; several models of transport mode, route choice and destination have been used by transportation planners. These models were necessary to predict travel demand on the long run and to support investment decisions in new road infrastructure that originated from this increased level of car use. In these days, travel was assumed to be the result of four subsequent decisions that were modelled separately. Those models are also referred to, within transportation literature, as Four-Step models.

1.1.1 The Trip-Based Approach: The Four-Step Model

The four-step model (Ruiter and Ben-Akiva, 1978) has been the primary tool for forecasting future demand of regional transportation services. It was introduced piecewise in the late 1950s, and was significantly modified since its first implementations.

Despite these modifications, the model still hangs on to the original standard framework.

The framework uses trips as an independent entity of analysis. This, however, leads to a number of serious limitations (Jovicic, 2001). A trip is defined as a one-way person movement by one of more modes of travel. Each trip has an origin and a destination. In the first step of the four-step model, i.e. *generation*, the model predicts the total number of trips generated and attracted to each zone in the study area. In the second step, i.e. *distribution*, the data, produced during the trip generation step, are used and the number of trips that will occur between one zone and another is predicted. These inter-zonal flows are represented in origin-destination (trip) tables. The third step of the modelling process is known as *mode choice*. This step assesses which transport mode is used for each trip. Once the number of trips and their origins and destinations are known, the last step (*assignment*), starts with allocating trips to particular routes in the transport network (McNally, 2000).

The major advantage is the simplification that is incorporated into these models, which made urban passenger travel demand forecasting relatively easy. The simplifying assumptions facilitated the quantitative analysis of travel demand, while this is in fact a result of complex travel behaviour.

However, many of these aggregate four-step models failed to make accurate predictions. The major drawback clearly is the focus on individual trips, where the spatial and temporal interrelationships between all trips and the characteristics of trips are ignored. Furthermore, the overall behaviour is represented as a range of constraints that define transport choice, while in fact it is an outcome of both real human decision making and a complex choice process. The last drawback clearly is the complete negation of travel as a demand derived from activity participation decisions.

1.1.2 The Tour-Based Approach

The original four-step models were replaced by theories about utility-maximising behaviour and individual choice behaviour. Multinomial logit models and more sophisticated techniques such as the nested logit and probit models formed the core of transportation modelling practice from the mid seventies onwards. Most of these techniques were implemented in so-called tour-based systems (Daly *et al.*, 1983). In the tour-based model, trips are explicitly connected in tours, i.e. chains that start and end at the same home or work base. This is done by introducing spatial constraints and by directions of movement. By means of this property, the lack of the spatial interrelationship, which was so apparent in the four-step trip based models, is dealt with. It is undisputable that much progress has been made in this research area compared to the aggregated four-step models. Nevertheless, these models did not escape criticism either. Especially in the eighties and early nineties, it was claimed by several researchers that very limited insight was offered into the relationship between travel and non-travel aspects. Indeed, travel has an isolated existence in these models and the question why people undertake trips is completely neglected. This is where activity based travel demand comes into play.

1.1.3 The Activity-Based Approach

The fundamental contributions of Hägerstrand (1970), Chapin (1974) and Fried *et al.* (1977) are the undisputed intellectual roots of activity analysis. Hägerstrand has put forward the time-geographic approach that characterises a list of constraints on activity participation. Chapin has identified patterns of behaviour across time and space. Fried *et al.* (1977) have dealt with the social structure and the question of why people participate in activities. These contributions came together in a study of Jones *et al.* (1983), where activities and travel behaviour were integrated. This was the first initial attempt to model complex travel behaviour. The major idea behind activity-based models is that travel demand is derived from the activities that individuals and households need or wish to perform. Travel is merely seen as just one of the attributes. Moreover, decisions with respect to travel are driven by a collection of activities that form an agenda for participation. Travel should therefore be modelled within the context of the entire agenda, or in other words, as a component of an activity scheduling decision. The concept of activity scheduling is an important one.

In short, travel patterns are the manifestation of the implementation of activity programs over time and space. In turn, activity patterns emerge as the interplay between the institutional context, the urban/physical environment, the transportation system and individuals' and households' needs to realise particular goals in life and to pursue activities (Ben-Akiva and Bowman, 1998). At least some of this complexity of travel decisions should be captured to make transportation models more reliable.

In order to summarise the above, we would like to cite the work of McNally (2000), who has nicely listed 5 themes which characterise the activity-based modelling framework:

- Travel is derived from the demand for activity participation.
- Sequences or patterns of behaviour, and not individual trips are the relevant unit of analysis.
- Household and other social structures influence travel and activity behaviour.

- Spatial, temporal, transportation and interpersonal interdependencies constrain activity/travel behaviour.

- Activity-based approaches reflect the scheduling of activities in time and space.

Activity-based approaches to transportation forecasting therefore aim at predicting which activities are conducted where, when, for how long, with whom, the transport mode involved and ideally also the implied route decisions.

Recently, models of time allocation to activities, activity duration and travel behaviour have been developed, sometimes in conjunction with travel behaviour models, such as activity-based transport models. Time use studies generally provide information on what individuals do over the course of a day, or, in some cases, over several consecutive days. Given the importance of activity participation in shaping travel behaviour, it is not surprising that in recent years there has been an increased interest in time use studies among travel behaviour researchers and travel demand modelers (Pas, 2002). The field of time use research has been reviewed by Pas and Harvey (1991) and they also examined the implications of this work for travel demand analysis and modelling. Conclusions indicated that mutual benefits would accrue from greater interaction between these related fields of research. This is why, during the past few years, the concept of time has moved from relative obscurity to centre stage in travel demand analysis and modelling. A large number of research efforts have been undertaken by travel behaviour researchers to develop models of how people use their time. In many ways, this research can be seen as taking the activity-based approach to travel analysis and modelling to a new level, since these models aim to predict what people do with their time, that is activity participation, which is the underlying rationale for travel and the basis for the activity-based approach (Golob, 1997; Fujii *et al.*, 1997; Lu and Pas, 1997; Bhat, 1996; Yamamoto and Kitamura, 1997).

Activity-based transportation models have certainly set the standard for the last decade of modelling travel demand. The models can be classified in a number of ways. Now we will focus on these comprehensive activity-based models of travel behaviour: a differentiation will be made between constraints-based, simultaneous (utility-maximisation) and sequential (computational process) models of activity scheduling behaviour. Constraints-based models have their roots in time geography, utility-maximisation models stem from microeconomic theory and psychology, while computational process models have been inspired by psychological decision process theories (Joh, 2004).

Constraints-based Models

These models typically examine whether particular activity patterns can be realised within a specific time-space environment. These models require as input activity programs, which describe a set of activities of a certain duration that can be performed at certain times. The space-time environment is defined in terms of locations, their attributes, available transport modes and travel times between locations for various transport modes. One of the attributes of interest is the opening hours of the facilities at that location. To examine the feasibility of a certain program, a combinatorial algorithm is typically used to generate all possible activity sequences. The feasibility of each sequence is then tested by checking whether (i) the interval between the end time of the previous activity and the start time of the next activity is sufficient to travel between location; (ii) the activity can start after the earliest possible start time and be finished before the latest possible end time and (iii) conditions about the sequencing of activities are not violated. The number of feasible activity schedules is often used as a measure of the flexibility that the time-space environment offers.

One of the first models in this tradition is Lenntorp's (1976) PESASP model. A similar model is CARLA, which basically is a combinatorial algorithm for generating feasible activity patterns (Jones *et al.*, 1983). BSP (Huigen, 1986) and MASTIC (Dijst and Vidakovic, 1997) are also similar models, and Kwan's Gisicas (1997) can be classified as a constraints-based model as well, although it makes also references to computational process models. Given an activity agenda, this GIS-based system begins scheduling by fitting activities on the agenda into the free time a person has, and orders them into a sequence. Activities with higher priority are ordered first, and the time constraints for performing certain activities are also taken into account.

A limitation of constraints-based models is that they lack the necessary mechanisms to predict adjustment behaviour of individuals. When faced with a changed time-space environment, individuals are likely to adjust/reschedule their activity programs. Constraints-based models, however, do not attempt to predict such behaviour.

Simultaneous Models (Utility-Maximisation Models)

Simultaneous models are based on observations of activity-travel patterns. The prediction is mainly done by means of utility-maximising econometric techniques (at first multinomial logit models were used, while later on nested logit models of increasing complexity were developed). These models assume that individuals evaluate a number of complete, one-day activity-travel patterns and choose the pattern that maximises their utility. In behavioural terms, these models do a rather moderate job, since they only adopt the assumption of utility-maximising behaviour.

One of the first known simultaneous models is Starchild, developed by Recker *et al.* (1986). It often has been referred to as the first operational activity-based model, but it was designed for research purposes and certainly not for general application. The primary weakness of Starchild is that it was designed to use data that, although essential to the theory of activity-based models, still is not available today. The only data set which was available for this model was a single activity diary data set that was collected to investigate the effects of carpooling (McNally, 2000). Another important, perhaps the most advanced, simultaneous model is the Daily Activity Schedule Model (Ben-Akiva *et al.*, 1996; Bowman, 1998). A prototype was developed for the Boston area (Bowman and Ben-Akiva, 1995), and later implemented for travel forecasting in Portland (Bowman *et al.*, 1998). The main difference here is that activity-travel patterns are treated in order of hierarchy. If a person completes a number of activities a day for instance, one of them is chosen to be the primary activity of the day. This means -in terms of the daily activity schedule model- that the utility of a primary activity (along with the travel that this activity brings along) is higher than the utility of secondary activities. Well-known methodologies such as nested logit models are used in this framework. Other examples of nested logit models -although considerably less complicated- include the work of Wen and Koppelman (1999) and the PETRA project (Fosgerau, 1998), which was funded by the Danish Transport Council and the Danish Energy Research Program.

The above models are all based on revealed preference data. In contrast, COBRA (Wang and Timmermans, 2000) has been based on conjoint choice experiments. Although their study demonstrated the potential of the newly proposed methodologies, it is doubtful whether conjoint experiments suffice to build a comprehensive activity-based model.

The Prism-Constrained Activity Travel Simulator (PCATS) has been developed by Kitamura and Fujii (1998). It is a system that simulates activity-travel behaviour, while considering prism constraints, availability of travel modes, and recognition of potential activity locations. Unlike the previous models that model the choice of activity pattern as a nested structure, PCATS assumes that individuals maximise the utility associated within the open periods, subject to the above constraints.

Bhat and Singh (2000) developed the Comprehensive Activity-Travel Generation for Workers (CATGW) model system. The activity-travel pattern is divided into several periods. Each of these patterns has its own characteristics and for every component a series of models is suggested to predict them. Although these have been largely published as isolated modelling efforts, when used in combination, it will result in a comprehensive modelling approach. Misra *et al.* (2003) extended this model to handle non-workers' activity schedules. Very recently, the Comprehensive Econometric Micro-simulator for Daily Activity-travel Patterns (CEMDAP) has been developed by Bhat *et al.* (2004). This system differs from its predecessors in that it is one of the first to comprehensively simulate the activity-travel patterns of workers as well as non-workers along a continuous time frame.

Sequential Models (Computational Process Models)

The models reviewed in the previous section simultaneously consider facets of travel patterns. The process, however, by which individuals arrive at their choices is not modelled at all. Sequential models (also called computational process models) represent an attempt of modelling this scheduling process. The utility-maximising framework, which is the most important assumption in the models described above, is now completely disregarded. After all, a lot of researchers have argued that people do not always necessarily arrive at 'optimal' choices, but rather use heuristics that may be context-dependent. In its most simple form, these modelling approaches use a set of simple IF-THEN rules, which take on the following form: IF (condition=X) THEN (perform action Y). Dependent on the context or the situation an individual faces, another outcome or another decision is taken. Such a set of rules makes up the model.

The first conceptual framework for understanding the process by which people organise their activities is Scheduler (Gärling *et al.*, 1989). Individuals and households are assumed to try and attain certain goals. Activities are defined as means, which the environment offers to attain these goals. The model Gisicas developed by Kwan (1997), can be best seen as an implementation of Scheduler. However, both Gisicas and Scheduler are not truly fully operational models (Timmermans, 2001).

Another model system that bears some resemblance with computational process modelling is AMOS, a dynamic micro-simulator of household activities and travel over time and space (Pendyala *et al.*, 1995, 1998). Although AMOS was designed with a rather specific policy application in mind and is not valid for general prediction, it nevertheless makes a good contribution in moving activity based transportation models toward operation status (McNally, 2000). Very recently, the model has been made operational for the State of Florida under the name FAMOS: Florida's Activity Mobility Simulator (Pendyala, 2004). The simulator is intended to serve as a comprehensive multi-modal activity-based micro-simulation model system that simulates activity and travel patterns at the level of the individual traveller.

Certain aspects of AMOS are very similar to Smash (Ettema *et al.*, 2000). This model concentrates on the process of activity scheduling. The scheduling process is assumed to be a sequential process consisting of a number of consecutive steps. This model was primarily developed as a process model and hence it does not offer much as a planning tool.

One of the most advanced operational process models in the transportation literature to date is Albatross (Arentze and Timmermans, 2000). Albatross has received significant attention and appreciation in literature (e.g. Axhausen, 2000; Arentze *et al.*, 2003; Janssens *et al.*, 2004a); it is the latest, most comprehensive and only fully operational computational process model at this moment. It can be considered as a multi-agent rule-based system that predicts activity patterns (see also Chapter 3). Several important extensions are realised in the second version of Albatross. The most important extension concerns the generation of schedule skeletons on a continuous time scale, which in the original version was taken as a given.

1.2 Complexity or Parsimony

Whether complex or parsimonious models should be preferred over the other is an old question. It has been investigated before in the data mining literature, in statistics, and perhaps even in all sciences (see e.g. Nock, 2002; Domingos, 1998; Zhang and Mühlenbein, 1995). Occam's razor (a plea for simplicity, see also Chapter 4) even dates back to the Middle Ages (Tornay, 1938).

The answer to this question all depends on the aim of the research. If your goal is to have a model with a high predictive performance and a large generalisability, perhaps complex models will serve this goal best. While, when coping with a large number of predictors, someone else might only be interested in a smaller subset of this predictor space that exhibits the strongest effects that influence a particular outcome. In order to get the 'big picture', that person is willing to sacrifice some of the small details. These small details can be disturbing, or, in real life, one simply does not have the time to check all different kinds of predictors in order to come up with a decision. Two small examples may illustrate this. Consider e.g. a doctor at an emergency room where a patient is just rushed in in the throes of a heart attack. Although this doctor's decision can save or cost a life, he/she does not have the luxury of extensive deliberation: just a few measurements will point to the action that will be undertaken (Gigerenzer *et al.*, 1999). Another example is situated at the stock market. Pedestrians and fund experts were asked in which domestic and international company they would invest their money. Most of the surveyed pedestrians had to make their choice solely on name recognition of the company. For the period considered in the study, the recognition knowledge of the pedestrians turned out to be even more profitable than the considered opinions of the experts! Thus, parsimonious models may provide a solution when one is only interested in the major effects that influence the outcome. These parsimonious models can be obtained by just applying simpler models or by using variable selection. This question of complexity and parsimony and which will serve better in what transportation context, is a central theme throughout this manuscript.

Recently, there is an increasing interest in the computational process model approach in order to model activity-diary data. Of these sequential models, the original Albatross system (Arentze and Timmermans, 2000) was the most complex and only fully operational model when this research was started. It aims to predict which activities will be conducted where, when, for how long, with whom and with which transport mode. These decisions determine the nine different choice facets of the model and a sequential execution of them provides activity patterns. Every choice facet can be regarded as a response variable with its own set of predictors that needs to be modelled (see also Chapter 3).

More specifically, this dissertation discusses two particular issues. With respect to the original Albatross model, the first issue concerns the following: 'Do simpler, and hence more parsimonious models perform better, or approximately as well as complex models in the context of activity-diary data?' This question will be regarded at choice facet level, thus for each of the nine dimensions separately, but also at a more aggregate level, at the level of the activity patterns. The data that were used to derive the original Albatross system are also used for the analyses provided in this part of the dissertation. In Moons *et al.* (2001, 2002a, 2005a, 2005b), two possible ways of simplification are examined and Chapter 4 summarises the results, while Chapter 5 is more concerned with a combination of simple classifiers.

The second issue is situated in one particular facet of activity-based models, i.e. the choice of transport mode. Several data sets (Dutch data, data from the San Francisco Bay area and from Southeast Florida) are used for this purpose. The question addressed within this framework focusses on the performance of nonlinear and semi-linear models, when compared to linear models (see Moons *et al.*, 2004b, 2004c). These different types of models (data mining techniques and statistical models) are compared to each other by means of three diagnostic measures in Chapter 7.
In Chapter 6, a new goodness-of-fit statistic is developed that measures the discrepancy between a parametric linear logistic regression model and the classification tree as its nonlinear, unrestricted nonparametric counterpart (see Moons *et al.*, 2002b,

2004a, 2005c). If the linear model is rejected by the test statistic, a close examination of the nonlinear model can help to improve the linear null model. These semi- and nonlinear models often lead to more parsimonious, but on the other hand also to more complex models (in terms of model definition, not in terms of the number of parameters).

To give some clear recommendations on the choice of a particular model is, of course, always difficult. Which model would be preferable over another? This question raises many questions, while, at first, the choice of a model depends, naturally, on the starting point of the modeler him-/herself. One can have as starting point to have interpretable models, while another prefers to have models with a very accurate prediction. This can lead already to very different choices. Therefore, these typical characteristics of the different models will also be discussed.

1.3 Organisation of the Subsequent Chapters

In Chapter 2, we present most of the data sets that are used throughout this manuscript. The data of the first Albatross system are described in Section 2.1. Because of the large amount of variables used, we will only consider the different dimensions of the system in Chapter 2, while the explanation and tabulation of all variables used is given in the Appendix. All the remaining data sets described in Chapter 2 are used to predict the choice of transport mode. The San Francisco Bay data are introduced in Section 2.2. Section 2.3 presents Dutch transport mode data for work related trips only. The data that are obtained from a household travel survey conducted in Southeast Florida are described in Section 2.4.

In Chapter 3, the original Albatross system will be reviewed (Arentze and Timmermans, 2000). The nine dimensions that describe which activity is conducted, where, when, with whom, for how long and which transport mode are discussed. Also the three different levels at which different models can be compared are introduced here.

In order to test whether more parsimonious models would result in models that are comparable in performance to the existing model structure, two possible ways to simplify the original Albatross structure are provided in Chapter 4. The first way deals with simple models, i.e. the application of Zero R, One R (Holte, 1993) and Naïve Bayes (Langley *et al.*, 1992, amongst others) within the context of the Albatross system and the comparison of the performance of these three simple heuristics with that of the originally used CHAID algorithm (Kass, 1980). A second manner in trying to simplify the structure is by applying a variable selection technique before using a recursive partitioning method (Kononenko, 1994; Quinlan, 1993). The results of this model are compared to the results of the technique without the feature selection.

Two other techniques that have often proven to ameliorate the results of a known classification technique are bagging and boosting (Breiman, 1996; Freund and Schapire, 1997). The results of both techniques are evaluated in the context of activity-diary data. Chapter 5 gives a short introduction to both techniques and reports the findings within the Albatross model.

In Chapters 3 - 5, the nine different outcomes of the nine different dimensions of the Albatross system are examined. In the subsequent chapters, we will focus on one particular aspect of the activity-diary data, i.e. on mode choice. The history of transportation modelling has shown that this particular part has always played a very important role.

The second part of this manuscript focuses on the second issue as discussed in Section 1.2. In Chapter 6, we propose a new lack-of-fit test statistic that contrasts the hypothesised model with a saturated model based on a sample space partitioning driven by the recursive partitioning algorithm as used in classification trees (Breiman *et al.*, 1984). These classification trees are nonparametric in nature and they can deal with large and complex data sets, a quality that other goodness-of-fit tests often lack. Simulation studies as well as studies in multidimensional settings will be presented to exemplify the proposed test procedures and, where possible, it will be compared to other, existing, lack-of-fit tests.

While investigating the question of complexity or parsimony in mode choice models in Chapter 7, we made use of semi- and nonlinear models. A very elegant extension to linear models has been provided by fractional polynomials as proposed by Royston and Altman (1994). The behaviour of fractional polynomials in this context is investigated on five data sets and the results are compared to that of a linear model and to those of nonlinear models. These nonlinear models are all nonparametric in nature. Support vector machines (Vapnik, 1996), that proven to be very good for prediction purposes and classification and regression trees (Breiman *et al.*, 1984) are the nonparametric models applied here.

We conclude with a summary on the major results in Chapter 8. The final conclusions with respect to transport modelling are divided into four different characteristics, relevant in transportation studies. All models are compared on grounds of predictive performance, interpretability, robustness and sensitiveness for policy measures.

Chapter 2

Motivating Examples

In this chapter, most data sets that are used throughout this manuscript, are introduced. The first series of data sets contains data from the original Albatross system, with its nine different choice facets (see Chapter 3). These data will be used to predict activity diaries in Chapters 4 and 5. Since we do not want this chapter to become too elaborate, we will discuss the different variables used in the nine dimensions in the Appendix. However, all information needed to comprehend the nature of the Albatross system is available here.

All the other data sets are used in Chapters 6 and 7. The second data set is collected in the San Francisco Bay area by the Metropolitan Transport Committee, here the selection of transport mode is studied more in particular. The same is true for the Dutch data set and for the fourth data set that was collected in Southeast Florida in 1999. Other data sets are introduced whenever the need arises.

2.1 The Albatross Data

We wish to thank the Urban Planning Group for the kind permission to use their data.

The activity diary data presented here, were used to derive the original Albatross system. The data were collected in February 1997 for a random sample of 1649 respondents in the municipalities of Hendrik-Ido-Ambacht and Zwijndrecht (South Rotterdam region) in the Netherlands. The activity diary asked respondents, for each successive activity, to provide information about the nature of the activity, the day, start and end time, the location where the activity took place, the transport mode (chain) and the travel time per mode, if relevant, accompanying individuals (alone, other member of household, other), and whether the activity was planned. Open time intervals were used to report the start and end times of activities. A pre-coded scheme was used for activity reporting. More details can be found in Arentze and Timmermans (2000).

The original induction algorithm, CHAID (see Chapter 4), requires that a limited number of discrete categories is defined for each variable. For the nominal variables, such as the socio-economic variables, the response categories are given. For continuous variables, such as for example travel times, one has to choose a method of discretising. Consistently, an equal frequency-interval has been used. This method divides a continuous variable into n parts, where each part contains approximately the same number of cases. Note that the induction algorithms may redefine the categories of continuous variables by merging contiguous categories.

The data were cleaned using a large set of rules incorporated in a dedicated computer program, called Sylvia (Arentze *et al.*, 1999). These cleaned activity-travel diaries are used throughout this manuscript. Each case is described in terms of a set of independent variables summarised in Tables 2.1 and A.1 to A.10.

2.1.1 General Characteristics

Table 2.1 summarises the general variables that were used for each choice facet of the model.

Table 2.1: General characteristics used in the various choice facets of Albatross

Name	Description	Categories
Day	Day of the week	1: Monday . . . 7: Sunday
Csec	Socio-economic class of the household	1: low . . . 4: high
Cage	Age of the oldest person in the household	$1 :< 25; 2 : 25 - 44; 3 : 45 - 64; 4 :> 64$
Ccomp	Household type	1: single, no work; 2: single, work
		3: double, one work; 4: double, two work;
		5: double, no work
Cchild	Presence of children in the household	1: none; 2: younger than 6
		3: 6-12; 4: older than 12
Gend	Gender of the person	1: male; 2: female
Ncar	Ratio between number of cars and number of adults	1: less than one; 2: one or more
Hwork1	Hours official work of the person per week	$0 : 0; 1 : 1 - 24; 2 : 25 - 32;$
		$3 : 33 - 38; 4 :> 38$
Hwork	Hours official work of the household per week	$0 : 0; 1 : 1 - 32; 2 : 33 - 38;$
		$3 : 39 - 60; 4 :> 60$

These include known household and person characteristics that can be relevant for the segmentation of the sample, including socio-economic variables, such as household type, age group, child index and socio-economic class; information about the (normal) activity program at a weekly basis with regard to time engaged in work at the household or person level; and car availability at the household level, indicated by a ratio between the number of cars and the number of adult members, so that for example a single-adult household with one car is equivalent to a double-adult household with two cars.

Furthermore, each dimension has its own list of more specific variables. They will be described in detail in Tables A.1 to A.10 in the Appendix.

2.1.2 Mode for Work

The 'Mode for Work' choice facet includes two different types of predictors (see Table A.1). A first series of variables describes the activity program at the level of the person's schedule skeleton and that of the partner, while a second series of covariates determines the work-chain for which the choice of transport mode needs to be made. These latter series include work and travel time information.

No extra constraints have to be taken into account in this dimension, the only constraint variable included is the number of cars. Obviously, if this number is zero, the car-driver mode is not a possible alternative. However, the car would also be infeasible in cases where there is only one car available that, in a previous step, has been assigned to a work-chain of the partner that is overlapping in time.

2.1.3 Activity Selection, Travel Party and Duration

The first set of variables in these three dimensions are program related (see Table A.2). They describe the time engaged in work activities, work-related travel and all different sorts of flexible activities. Note that at the present step, the travel time information for work-related trips is known, since the transport mode choice for the primary work activity is made in the previous step. The main idea for including these program-level variables is that they portray important conditions, such as the activity load of the current program, the possibility to combine activities, etc.

Furthermore, there are some variables at schedule level and some specific variables for each choice facet that determine some constraints. E.g. a shopping activity cannot be selected and planned in the schedule if the maximum time available within the opening hours of possible facilities for the activity is shorter than the duration of the activity itself.

2.1.4 Activity Start Time

For the specification of start time, the system distinguishes six episodes of the day, being before 10 AM, between 10 and 12 AM, between 12 and 2 PM, between 2 and 4 PM, between 4 and 6 PM and after 6 PM. For shopping, service and leisure activities, the opening hours of the facilities further restrict possible start times. Since the location is still unknown, the maximum opening hours of the municipality where the household lives, are used. For social activities, there are no timing constraints. In summary, the following information is available to describe the cases at program/schedule level:

- activity skeleton (selection, timing and location of fixed activities)
- mode for work activities
- selection, travel party and duration of flexible activities
- the start-time range and schedule position of processed activities.

The included household and person attributes are the same as in previous steps. Note that at this stage the schedule is complete in terms of the selection of activities to be done that day. At the schedule and activity level, there is a considerable overlap with independent variables used in the previous steps. Variables are added not only to cover the extra information given by the previous travel party and duration decisions, but also to describe specific conditions for start time decisions (see Table A.4).

2.1.5 Trip Chaining

For every free activity in the schedule, this component of the system determines whether it is possible to make a connection with an other out-of-home activity preceding the activity (After Stop), succeeding the activity (Before Stop) or both (In-Between Stop). The single stop option where the home location is both the origin of the trip to the activity location and the destination of the return trip is considered feasible in every case.

Each time an activity is added, the system defines three activity sets comprising the activities in the current schedule with which an After Stop, Before Stop and In-Between Stop can be realised, respectively. If any of the three sets includes more than one activity, the case is not considered for further analysis, because the trip chaining choice facet can only handle cases that involve a uni-dimensional choice. We emphasise that the chance of having multiple trip-connection possibilities of the same type within the given start-time interval is generally small so that only a small fraction of cases needs to be excluded for this reason. Adding an activity means that it is inserted in the observed schedule position. When next activities are considered, the added activity is considered in turn a candidate for establishing a trip connection in the same way as the activities in the schedule skeleton are. Thus, travel connections can be realised between flexible activities. It follows that the following information is available in this step:

- activity skeleton (selection, timing and location of fixed activities)
- mode for work activities
- selection, travel party, duration and start-time range of flexible activities
- the schedule position of activities that have been processed in this stage.

A connection is considered feasible only if there is a start-time and duration choice possible such that the activities can be connected in time by travelling between the locations. The set of variables that were used to describe the cases at the program-level, schedule-level and activity-level are summarised in Tables A.6 and A.7.

2.1.6 Activity Transport Mode

The final two decisions that need to be modelled concern transport mode and location choice. The mode used for the primary work activity is chosen in the first scheduling step. Furthermore, the locations of the fixed activities are considered given. Therefore, this subsection focuses on the transport mode for other-than-work and short-duration-work activities, while the location choice for flexible activities, will be regarded upon in the next subsections.

Transport mode decisions are made at the level of a tour. We consider car driver, car passenger, public transport and slow transport (walk or bike) as the categories of the dependent variable. A tour consists of a trip from home, a return trip and, in cases where the tour involves multiple activities at different locations, in addition one or more trips between out-of-home locations. The system assumes that individuals do not change mode across trips of the same tour reflecting the notion that possibilities to change mode are generally limited. The single-mode assumption is supported by the fact that only 4.4% of the tours in our data set involved multiple modes. In the future, one has to take into account that individuals may use multiple modes at the level of trips, but the present model does not account for a mode chain. Deterministic rules are used to determine the main mode and this is considered as the mode for the trip. Land-use, facility-opening hours and mode-specific travel time data of the study area are used to determine feasibility and relative speed of the mode alternatives.

The independent variables necessary to determine the transport mode describe the cases at household/individual, activity-program and tour level. As for the household/individual level, the same variables are used as in Table 2.1. The activity-program and tour-level variables are summarised in Table A.8.

2.1.7 Locations

After the mode for each tour has been specified, location decisions are sequentially made for each flexible out-of-home activity. When these decisions have to be made, all the other activity dimensions are known: schedule position, travel party, duration start time, trip chaining, transport mode and location of fixed activities. The location-choice set is dynamically defined as locations that are feasible given activity-timing constraints, activity-duration constraints, available facilities, opening hours of available facilities and speed of travel. Social activities are an exception as these activities are not dependent on (public) facilities. Therefore, all reachable locations are considered feasible in that case.

Just as in previous dimensions, each case is described at different levels including the household/individual, activity program/schedule, tour and activity level. The variables used to determine household/individuals are the same as in Table 2.1, while the other variables are described in Table A.10.

2.2 San Francisco Bay Area

The San Francisco Bay Area household travel survey (Purvis, 2003) has become established among the surveys in activity-based transport. It has already been conducted several years: in 1946/47, 1965, 1981, 1990, 1996 and recently in 2000. The demographic and travel behaviour data used here come from the detailed survey that was conducted by the region's metropolitan planning organisation (the Metropolitan Transportation Commission) across more than 5800 San Francisco Bay Area households for two day's trip making (including weekend days) in 1996. The travel survey was conducted in the nine counties of the San Francisco Bay Area. It is one of the few time-use and travel surveys that includes information on both commuters and non-commuters. Detailed information on both in-home and out-of-home activities and trips undertaken by an individual was recorded in the survey. While information on all trips and segments (in the case of chained trips) was collected, in-home activity information was requested only for those activities that were longer than 30 minutes in duration.

For our analysis, we opted to use a subset of the 1996 data. This original data set contains 34864 observations and 127 variables. After deleting all observations with missing data, there are 24752 records left. With 127 explanatory variables, it is almost impossible to investigate for each covariate the nature of the relationship (linear, quadratic, etc.) and considering all possible two-way interactions is also hardly feasible. Therefore, in the logistic regression models that we consider in Chapter 6, we limited ourselves to 26 variables that describe the different activity trips. As dependent variable in these analyses, we considered a binary variable indicating whether people were using the car as transport mode or not. Only trips by 'adults' (i.e. persons over 19 years of age) were considered. The independent variables are described in Table 2.2.

They comprise variables describing the reported person and household characteristics as well as typical activity/trip specific features.

Table 2.2: Description of Independent Variables for the San Francisco Bay Area Data

Name	Type	Description
	Person Characteristics	
Commuter Status (x_1)	Binary	0: non-commuter, 1: commuter
Gender (x_2)	Binary	1: male, 2: female
Age (x_3)	Continuous	in years
License (x_4)	Binary	0: no, 1: yes
Employed (x_5)	Binary	0: no, 1: yes
Student (x_6)	Binary	1: yes, 2: no
Day Activities (x_7)	Continuous	Total number of activities in a day
Day Trips (x_8)	Continuous	Total number of trips in a day
Race (x_9)	Binary	0: other, 1: white/Caucasian
	Household Characteristics	
Householdsize^{-1} (x_{10})	Continuous	Inverse of the household size
Workers (x_{11})	Continuous	Number of workers in household
Number of Vehicles (x_{12})	Continuous	Number of vehicles in household
Number of Bicycles (x_{13})	Continuous	Number of bicycles in household
Own-Rent (x_{14})	Binary	1: own house, 2: rent
Type of Home (x_{15})	Nominal	1: single family, detached unit; 2: duplex; 3: apartment; 4: condo or townhouse; 5: mobile home/trailer; 6: hotel/motel; 7: group quarters; 8: other
Carpool to Work (x_{16})	Binary	1: yes, 2: no
Years Residence (x_{17})	Continuous	Number of years at current residence
Auto Own (x_{18})	Continuous	Number of automobiles in household divided by number of members that are $>$ 5 years
Income per Member (x_{19})	Continuous	Household income (in 1996 dollars) divided by household size
	Activity/Trip Characteristics	
Start Hour (x_{20})	Continuous	Activity/Trip start time (hour)
Start Minute (x_{21})	Continuous	Activity/Trip start time (minutes)
End Hour (x_{22})	Continuous	Activity/Trip end time (hour)
End Minute (x_{23})	Continuous	Activity/Trip end time (minutes)
Duration (x_{24})	Continuous	Duration of activity/trip (minutes)
Origin Type (x_{25})	Nominal	1: home; 2: work; 3: school; 4: other
Destination Type (x_{26})	Nominal	1: home; 2: work; 3: school; 4: other

2.3 Dutch Data

As discussed in the first section and in the Appendix, we consider the activity diary data used to derive the original Albatross system. In this separate data set, in order to predict the transport mode used, only the work tours are considered. Some of the variables in Table A.1 are left out, others have turned back into their original continuous nature (instead of categorical) and some other variables are added. After cleaning the activity-travel diaries, 1025 cases were remaining and 39 variables. A binary variable that equals 1 if the mode choice is either car (being a car driver, not a passenger), slow transport or public transport, and zero otherwise will again be used as dependent variable (on three different data sets). Tables 2.3 and 2.4 summarise the 39 general and some specific explanatory variables that were used to model the data.

Note that AP represents the activity pattern of the concerned person on the day in which the concerned tour is embedded, while C is the concerned tour.

These general variables include known household and person characteristics that might be relevant for the segmentation of the sample, including socio-economic variables, such as household type, age group, child index and socio-economic class; information about the (normal) activity program at a weekly basis with regard to time engaged in work at the household or person level; and car availability at the household level, indicated by a ratio between the number of cars and the number of adult members, so that for example a single-adult household with one car is equivalent to a double-adult household with two cars. For a more detailed description we refer to Arentze and Timmermans (2000).

2.4 Southeast Florida

This data set is obtained from a household travel survey conducted in Southeast Florida in 1999. The survey was a part of the Southeast Florida Regional Travel Characteristics Study, which included an on-board transit survey, a visitor survey, a truck movement survey, and a work place survey in addition to the household survey. The household survey consisted of three steps, comprising: a CATI (computer-aided telephone interview) recruitment of survey participants, distribution by mail of survey instruments and travel diaries, and a CATI retrieval of the demographic data and travel diaries.

A total of 7500 households agreed to participate in the survey as a result of the CATI recruitment, and the survey instruments and travel diaries were mailed out for

Table 2.3: Description of Independent Variables for the Dutch Data: Part I

Name	Type	Description	Categories
		Person/Household Characteristics	
Day (x_1)	Nominal	Day of the week	1: monday - 7: sunday
Csec (x_2)	Binary	Socio-economic class of the household	1: low socio-econ. class, 0: other
Cage (x_3)	Ordinal	Age of oldest person in the household	$1 :< 25; 2 : 25 - 44;$ $3 : 45 - 64; 4 :> 64$
Ccomp (x_4)	Binary	Household composition and work status: At least one person works in household	0: no; 1: yes
Cchild (x_5)	Binary	Presence of children in the household	1: 3 children or more, 0: 2 or less
Gend (x_6)	Binary	Gender of the person	1: male, 2: female
Ncar (x_7)	Binary	Ratio between number of cars and number of driving licenses in the household	$1 :< 1, 2 :\geq 1$
Hwork1 (x_8)	Continuous	Hours official work of person per week	
Hwork (x_9)	Continuous	Hours official work of household per week	
		Activity/Tour Characteristics	
Nsec (x_{10})	Ordinal	Number of non-work, out-of-home activities in AP	$0 : 0; 1 : 1; 2 : 2;$ $3 : 3 - 4; 4 :> 4$
Avcar (x_{11})	Binary	Car available in terms of availability driving license and car in household	0: no; 1: yes
Two (x_{12})	Continuous	Total time of work in AP (in min.)	
Ttot (x_{13})	Continuous	Total time of primary and secondary work in AP (in min.)	
Yserv (x_{14})	Binary	There is at least one shopping or service activity in AP	0: no, 1: yes
YSoLei (x_{15})	Binary	Same for out-of-home social/ leisure activity	0: no, 1: yes
Ybget (x_{16})	Binary	Same for bring/get person or goods activity	0: no, 1: yes
CBT (x_{17})	Continuous	Earliest possible begin time of C	
CET (x_{18})	Continuous	Latest possible end time of C	
Cdur (x_{19})	Continuous	Difference between CET and CBT	

Table 2.4: Description of Independent Variables for the Dutch Data: Part II

Name	Type	Description	Categories
Cnout (x_{20})	Ordinal	Number of out-of-home activities in C	$1:1; 2:2$ $3:3-4; 4:>4$
Ctwo (x_{21})	Continuous	Total time of work in C (in min.)	
Cttot (x_{22})	Continuous	Total time of primary and secondary work in C (in min.)	
CyServ (x_{23})	Binary	There is at least one shopping or service activity in C	0: no, 1: yes
CySoLei (x_{24})	Binary	Same for out-of-home social/ leisure activity	0: no, 1: yes
CyBget (x_{25})	Binary	Same for bring/get person or goods activity	0: no, 1: yes

Aty1 (x_{26})	Nominal	Type of the first activity in C	1: work; 2: bget; 3: grocery;
			4: service; 5: non-grocery;
			6: leisure; 7: social; 8: other

Awith (x_{27})	Nominal	Person with whom first activity in C is conducted	0: none; 1: only others inside the household; 2: others outside the household
Pbrget (x_{28})	Binary	Partner has a bring/get activity during tour C	0: no, 1: yes
Pserv (x_{29})	Binary	There is a grocery, shopping or service activity in the partner's AP during tour C	0: no, 1: yes
		Transport Characteristics	
Ttbike (x_{30})	Continuous	Shortest travel time by bike for tour C (in min.)	+ 0.1 (to overcome problems with logarithms)
Rcabi (x_{31})	Continuous	Travel time ratio between car and bike (in %)	
Rpubi (x_{32})	Continuous	Travel time ratio between public transport and bike (in %)	
Rpuca (x_{33})	Continuous	Travel time ratio between public transport and car (in %)	
Textra2 (x_{34})	Continuous	Extra travel time to reach a location of order 2 (minutes bike time)	
Textra3 (x_{35})	Continuous	Same for order 3	
Textra4 (x_{36})	Continuous	Same for order 4	
Ptmax (x_{37})	Continuous	Maximum bike travel time across activities in the partner's AP during tour C	
YavSlo (x_{38})	Binary	Minimum sum of duration of activities in C plus minimum bike travel time $\leq$ maximum duration of C (=Cdur)	0: no, 1: yes
YavPu (x_{39})	Binary	Same for public transport travel time	0: no, 1: yes

each person in the household including visitors and infants. Parents were asked to fill out children's travel diaries. Of these, 5168 households and 11426 persons completed the survey. The survey day was set on one of the weekdays between Tuesday and Thursday for each household, and all travel-related activities were recorded for an entire 24-hour survey day. The detailed description of the survey and tabulations of simple statistics can be found in the survey report (The Corradino Group, 2000). The activity engagement and time allocation behaviour represented by these data are also analysed in Meka *et al.* (2002) and in Yamamoto *et al.* (2003).

In the analyses performed in Chapter 7, we focused on the activity data file. In total, 41955 activities were reported and after deleting the missing values, 14527 observations were remaining. The variables that describe household, person as well as activity characteristics can be found in Table 2.5.

Table 2.5: Description of Independent Variables for the Southeast Florida Data sets

Name	Type	Description	Categories
Hhldsize (v_1)	Continuous	Number of persons in the household	
Hhldemp (v_2)	Ordinal	Number of workers in the household	0–13
Numchild (v_3)	Ordinal	Number of children in the household	0–5
Numlic (v_4)	Ordinal	Number of licensed drivers in the household	0–13
Hvehicle (v_5)	Ordinal	Number of vehicles available in the household	0–9
Meminc (v_6)	Continuous	Annual income of the person	(×5000)
Memage (v_7)	Continuous	Age of the person	
Employed (v_8)	Nominal	Employed?	1: yes, 2: no
Acttype (v_9)	Nominal	Activity type	1: home, 2: work, 3: shop, 4: social recreation, 5: school, 6: other, 7: unknown
Actdur (v_{10})	Continuous	Duration of the activity (in minutes)	+ 0.1 (to overcome problems with logarithms)
Finadm (v_{11})	Nominal	Final activity?	1: if final activity, 0: otherwise
Midhmd (v_{12})	Nominal	Mid-day activity?	1: if mid-day home activity, 0: otherwise
Aggaty (v_{13})	Nominal	Aggregated activity purpose category	1: home, 2: subsistence (home and school), 3: maintenance (shopping), 4: leisure (social recreation), 5: other and unknown
Ampkdum (v_{14})	Nominal	AM peak?	1: if activity pursued in AM peak (7:15–9:15 AM), 0: otherwise
Midddum (v_{15})	Nominal	Mid-day peak?	1: if activity pursued in mid-day (9:16 AM–3:15 PM), 0: otherwise
Pmpkdum (v_{16})	Nominal	PM peak?	1: if activity pursued in PM peak (3:16–6:15 PM), 0: otherwise
Ofpkdum (v_{17})	Nominal	Off peak?	1: if activity pursued in off peak (6:16 PM–7:14 AM), 0: otherwise

Chapter 3

The Albatross System

In this chapter, the Albatross system will be introduced. At first, we will revise a short part of the history as provided in Chapter 1 and elaborate somewhat more on activity-based models, so as to put the Albatross model in the right perspective.
It should be noted that throughout this manuscript we work with the original data. In the meantime, however, an extended set of rules, based on a larger data set, has been derived (Arentze and Timmermans, 2002, 2004), though, when this research was started, the larger data set was not yet available.

3.1 Introduction

3.1.1 History

During the last decade, interest in spatial interaction patterns in transportation research and spatial sciences alike has shifted away from trips and tours to the analysis of complex daily activity-travel patterns (e.g. Bhat and Koppelman, 1999). This shift in interest was motivated by both methodological and policy considerations. It was realised that travel patterns are a manifestation of activity participation at different points in space (e.g. Axhausen and Gärling, 1992). A focus on daily activity patterns as opposed to single trips and multi-stop, multi-purpose tours was felt to lead to potentially better predictions of travel demand in time and space. The activity-based approach would allow one to better capture the interdependencies of activity participation and travel, within a particular spatial and institutional context. Furthermore, it would allow one to assess the impact of such new policy areas as teleworking and teleshopping that were virtually impossible to tackle with conventional models (see e.g. Timmermans *et al.*, 2002).

From a methodological perspective, the new focus on timing and duration of activities led to the application of statistical methods, including hazard (Bhat, 1996) and Tobit models that were rather new to the field. Moreover, and perhaps more challenging, were the attempts to develop more comprehensive models of activity-travel behaviour. These models typically attempt to predict various facets of travel behaviour. In addition to the traditional facets of destination, transport mode choice, and perhaps trip chaining (multi-purpose trips), activity-based models consider activity choice, timing, duration, travel party and route choice. Moreover, various types of constraints (spatial, temporal, institutional, spatial-temporal) were incorporated in the modelling efforts, and in some cases the decision-making unit was the household as opposed to the individual (e.g. Gliebe and Koppelman, 2000; Zhang *et al.*, 2002). This increased complexity (more choice facets, more choice alternatives, preferences and constraints, coordination of multiple persons, inter-temporal and spatial dependencies, etc) caused a major challenge for the modelling community.

The multitude of modelling attempts seems to converge now into two modelling approaches (disregarding constraints-based models for a while, since they have not achieved that much attention lately). First, the discrete choice utility-maximisation models, originally developed for trip and tour data, were extended to include more facets. Examples of such utility-maximisation models include the Daily Activity Schedule model (Bowman, 1998), the CATGW model system (Bhat, 1999), PCATS (Kitamura and Fujii, 1998), and Patricia (Borgers *et al.*, 2002), to name a few. Second, arguing that individuals do not necessarily maximise their utility, rule-based computational process models of activity scheduling behaviour have been developed. Unlike the econometric models, computational process models do not rely on algebraic equations but on a set of Boolean decision rules or neural networks to predict observed activity-travel patterns. Examples of such models include Scheduler (Gärling *et al.*, 1989), AMOS (Pendyala *et al.*, 1995), Albatross (Arentze and Timmermans, 2000, 2002), and Aurora (Joh *et al.*, 2001d), although the latter model uses a combination of algebraic equations and decision heuristics. A more detailed state-of-the-art review is given in Timmermans *et al.* (2002). These competing modelling approaches each have their specific advantages and pitfalls. Protagonists of computational process models argue that the utility-maximisation, econometric models do not reflect the true behavioural mechanisms underlying travel decisions and these models are based on too rigourous assumptions about travel behaviour (e.g. Gärling *et al.*, 1998). Likewise, advocates of econometric approaches argue that computational process models lack rigour, ease of interpretation, and the ability to statistically assess the significance of the decision rules. This leads to the paradox that although computational process models have been developed to better reflect the behavioural mechanism underlying activity-travel decisions, they are often viewed as black boxes, consisting of a large number of decision rules of which the specific influence on the final outcome of the model is impossible to identify. This discussion should be placed in the context of the purpose of the model. If the goal is to better understand behavioural mechanisms underlying travel behaviour, rules that better reflect actual decision-making seem paramount. On the other hand, if the goal is to predict travel patterns, the situation seems less clear as behaviourally more sound models do not necessarily also predict better. Unfortunately there is a lack of comparative studies in which the predictive performance of competing models, derived from the same data, is compared. Consequently, the discussion about future directions and pros and cons of the competing modelling approaches from a predictive point of view remains almost philosophical in nature.

3.1.2 The Albatross System

Recently, several studies indicate an increasing interest in the computational process model approach in order to model activity-diary data. These models derive choice rules (Boolean expressions) from activity-travel diary data. This process of rule induction is similar to the process of parameter estimation in algebraic, econometric models.

Albatross, the most complex fully-operational computational process model to date, was developed for the Dutch Ministry of Transportation (Arentze and Timmermans, 2000). It originally derives decision rules using a CHAID-based induction algorithm. This means that the set of condition variables, assumed to influence some facet of activity-travel behaviour, are successively split based on the chi-square measure, such as to find as homogeneous sets of conditions as possible until some stop criterion is met. This process can be represented in terms of a decision tree, which indicates which combination of condition states leads to a particular action (see e.g. Arentze *et al.*, 2000a, for more details).

The first task of an activity-based system is to generate individual activity programs. The process of activity program generation is of course dependent on several short term decisions such as the nature of a particular activity (mandatory or not), the urgency of completing a particular activity on a specific day and the desire to meet particular activity and time-related objectives. Other long-term decisions such as marital status, the number of children and the choice of residence are also likely

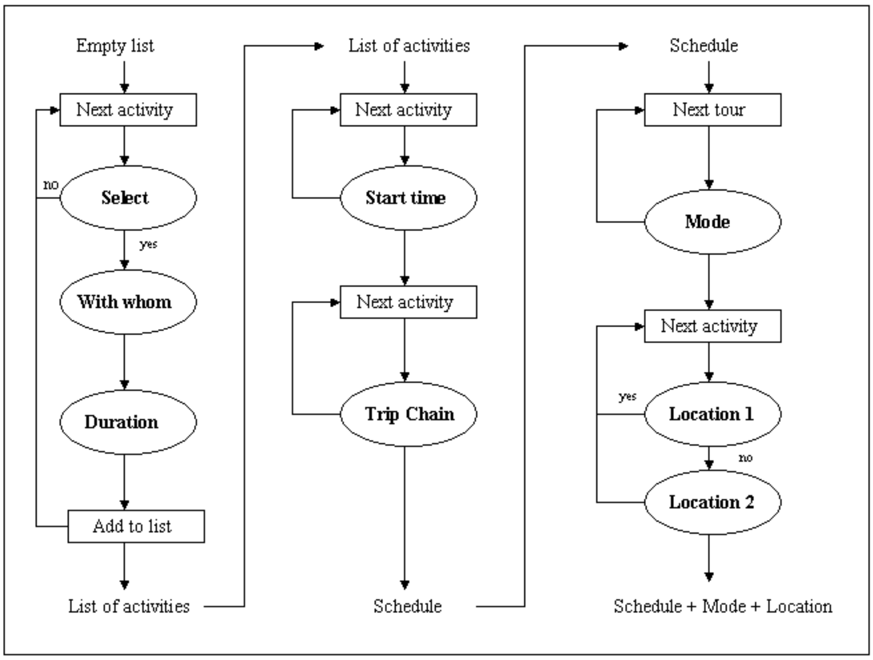

Figure 3.1: *Albatross' scheduling engine*

to have an important impact on the activity program generation. After the activity programs have been generated, the next step is to schedule these activities. This means that further dimensions have to be added to each activity.

The activity scheduling process happens sequentially at micro level. Figure 3.1 provides a schematic representation of the Albatross scheduling model. All relevant choice sets are considered through different agents of the Albatross model system: which activity is conducted, where, when, with whom, for how long and which transport mode is used. In order to schedule the activities, household interactions between individuals as well as constraints have to be taken into account in the system. These constraints can be of different types: (i) situational constraints (e.g. persons cannot be at different locations at the same time), (ii) institutional constraints (e.g. opening hours influence the activity which can be conducted), (iii) household constraints (e.g. bringing children to school limits the choice set of activities which can be conducted), (iv) spatial constraints (e.g. particular activities cannot be performed at particular locations) and (v) time constraints (each activity requires a minimum amount of time in order to be completed).

The activity scheduling agent of Albatross is based on an assumed sequential execution of decision trees to predict activity-travel patterns. The model first executes a set of decision rules to predict whether or not a particular activity will be inserted in the schedule. At the same time, the transport mode for the primary work activity is chosen. If the activity is added, the travel party and the duration of the activity are determined, based on another sets of rules, before a next activity is considered. The order in which activities are evaluated is pre-defined as: daily shopping, services, non-daily shopping, social and leisure activities. Time constraints are used in this step to determine the feasibility of the chosen activities. Subsequently, in order of priority, a general notion of time of the day (e.g. early morning, around noon, . . .) is determined for each activity. Based on this, for each activity, a preliminary position is determined in the schedule. Hereafter, trip links (i.e. trip chaining decisions) between activities are considered, which means that when tours are included in the schedule, they are identifiable as sequences of one or more out-of-home activities that start at home and end at home. These trip chaining decisions are not only important for timing activities but also for organising trips into tours. For each tour a transport mode is then determined. Note that if the activity is the primary work activity, then the transport mode was already chosen, if not, the choice of transport mode is made here in the scheduling process. Finally, the location of each activity is set. Possible interactions between mode and location choices are taken into account by using location information as conditions of mode selection rules. Institutional, spatial and time constraints are adopted in this step to determine which locations are feasible.

The general statistics for the decision tables for each of the choice facets can be found in Table 3.1. This table describes the statistics on the training set. A 75-25% split was made on the data set as a whole, where the first 75% are used to build the nine different models, whereas the remaining 25% was left to validate them.

The predictions for each model are based on a simulation procedure. This involves building an activity pattern for each person-day by successively making a decision on each choice dimension. A decision involves selecting a choice alternative based on the predicted probability distribution across alternatives on the choice facet concerned.

Instead of using the original CHAID-based induction algorithm, many other induction algorithms can be used to derive the set of decision rules for the nine different choice facets of the Albatross system. In the next chapters we will try to gain a better understanding in the influence of the use of respectively simple heuristics to derive the set of rules (Chapter 4), a smaller set of decision rules (trimmed decision tree) (see also Chapter 4) and the use of bagging and boosting (Chapter 5) on the predictive performance of sequential models of activity scheduling behaviour in general and the Albatross system in particular.

Table 3.1: General statistics for each choice facet

Dimensions	Number of cases	Minimum cases per column	Number of independent variables
Mode for work	858	15	32
Selection	14190	30	40
With-whom	2970	15	39
Duration	2970	15	41
Start time	2970	15	63
Trip chain	2651	15	53
Mode other	2602	15	35
Location 1	2112	15	28
Location 2	1027	15	28

However, in order to be able to compare these different models with the original CHAID-based induction algorithm and amongst each other, we need to have some evaluation criteria. These will be defined in the next section.

3.2 Model Comparison

Model performance tests can be conducted at three levels: the choice facet level, the activity pattern level and the trip matrix level. Recall that the Albatross system consists of nine different choice facets or dimensions and that each of them determines a different response variable. For every dimension, a separate model needs to be build. At the choice facet level, the attributes that remained in the final (decision tree) model for each approach will be discussed. The increase in predictive performance will also be calculated for each decision tree. At the activity pattern level, sequence alignment methods (Joh *et al.*, 2001a, 2001b, 2001c, 2002a) are used to assess the correspondence between the observed and predicted activity sequences. At the trip matrix level, correlation coefficients are calculated to measure the degree of correspondence between the observed and the predicted Origin-Destination matrices.

3.2.1 Choice Facet Level

In general, the performance of a particular model at choice facet level will be represented in a table. It will normally look as follows (see Table 3.2):

Table 3.2: Performance at choice facet level

Dimension	♯ alts	♯ attrs	♯ leafs	e	e_{ratio}
Mode for work	4				
Selection	2				
With-whom	3				
Duration	3				
Start time	6				
Trip chain	4				
Mode other	4				
Location 1	7				
Location 2	6				

The first column of these tables presents the 9 choice facets of Albatross. The second column lists the number of alternatives (levels of the dependent variable), while the third column gives the total number of attributes/independent variables that were considered to build the final decision tree. The fourth column depicts the total number of leafs of the decision tree. Column five reports the probability of a correct prediction, as defined in equation (3.1), and in the last column the predictive performance is compared to a null model (equations (3.2) & (3.3)). In the Albatross system, the null model assigns a new case to a category of the Y-variable with a probability, equal to the number of observed cases in the category divided by the total number of cases in the data set.

Following Arentze and Timmermans (2003), the probability of correctly predicting the choice for any given case j in the training sample space equals:

$$\begin{aligned}
e &= P(\text{correct prediction for random case } j)\\
&= \sum_k P(\text{correct prediction for case } j|j \text{ belongs to leaf node } k)\\
&\quad \times P(j \text{ belongs to leaf node } k)\\
&= \sum_k \sum_i P(\text{choice } i \text{ is observed and choice } i \text{ is predicted for case } j|\\
&\quad j \text{ belongs to leaf node } k) \times P(j \text{ belongs to leaf node } k)\\
&= \sum_k \sum_i \Big(\frac{f_{ik}}{f_k}\Big)^2 \times \frac{f_k}{n} \qquad \text{(due to probabilistic assignment rule)}\\
&= \frac{1}{n} \sum_k \frac{\sum_i (f_{ik})^2}{f_k} \qquad (3.1)
\end{aligned}$$

where,
n is the total number of cases,
f_k is the number of cases at leaf node k, and
f_{ik} is the number of cases at leaf node k with observed choice i.

Note that the sample space will be partitioned, either by the simple heuristics or by a recursive partitioning method. For the ease of notation, we denote every final split as a leaf node.

By comparing e with a simple prediction based on the frequency distribution of the choice possibilities in the sample, a measure of relative performance can be derived. The probability of correctly predicting the choice for a random case j, without applying any splitting criterion, can be found as:

$$e_0 = \frac{1}{n^2} \sum_i (f_i)^2 \qquad (3.2)$$

where, f_i is the overall frequency of choice i in the training sample.

The quotient

$$e_{ratio} = \frac{e - e_0}{1 - e_0} \qquad (3.3)$$

then indicates the increase in predictive performance as a ratio of the maximum increase that is possible given the information known in the sample. This is the measure that is provided in the last column of Table 3.2.

3.2.2 Activity Pattern Level

The assessment of goodness-of-fit at the activity pattern level requires the choice of an appropriate measure. Since observed and predicted sequences of activities will be compared, this measure needs to be able to capture the multi-facet aspect of activity patterns. In addition, and more critically, the measure should also be flexible in that it allows the inclusion of categorical and sequential information. Most of the facets of activity patterns, such as activity type and mode choice, are categorical in nature, but the facet of activity scheduling implies sequential information. The model should be successful in predicting the sequence of activity choices. In addition to the problem of being able to capture sequential information, the measure should also be sensitive to activity patterns of unequal length. The Sequence Alignment Method (SAM), introduced by Wilson (1998) in time use research, has the potential to capture all this. The SAM is one of various sequence comparison methods, originally introduced in disciplines such as molecular biology, chromatography, etc. The interesting feature of the SAM is that it employs 'biological' distance rather than geometric (Euclidean) distance as the basic concept of comparison. Biological distance can be defined as the amount of effort that is needed to equalise two strings of information and this is taken as an indicator of the (dis)similarity between strings (Kruskal, 1983). Unfortunately, the conventional SAM can only handle uni-dimensional strings. The uni-dimensional SAM can capture the intra-sequential relationships between elements of an attribute, but not the inter-relationships between elements of different attributes. Therefore, the Albatross system provides a multidimensional extension of the conventional sequence alignment method.

The system uses several sequence alignment methods to measure the goodness-of-fit. These methods measure the dissimilarity of the 2 sequences in terms of the effort required to make the two sequences identical by using insertion, deletion and substitution operators. In this way, sequences of unequal length can be compared. Insertion and deletion operations incur the same cost of one unit, while substitution of an element requires twice that cost. The lower the SAM measure, the more similar the sequences are.

The first set of four measures, that will be provided as goodness-of-fit measures, indicate the uni-dimensional SAM for the activity pattern attributes (being activity type, travel party, location and transport mode) separately. The UDSAM indicates a weighted sum of uni-dimensional SAM costs across the four dimensions, whereby activity type was given a weight of 2 units and the other attributes a weight of one unit, and the MDSAM indicates the multidimensional SAM using the same weights.

The MDSAM (Joh *et al.*, 2001a, 2002a) differs from the UDSAM in that it takes possible correlations between choice facets into account by allowing the alignment procedure to implement joint operations.

3.2.3 Trip Matrix Level

At the trip matrix level, the observed and predicted Origin-Destination (OD) matrices are compared. The basic unit for generating an OD-matrix is a trip. It contains the frequency of trips for each combination of origins (rows) and destinations (columns). Assuming that each tour starts from home and ends at home, within the 24-hour time frame, return trips (trips of which the destination is 'home') do not yield extra information and are, therefore, not included in the generated OD-matrices. Furthermore, trips with unknown origin or unknown destination, due to a missing value, are left out too. Obviously, in the predicted set missing values occur in exceptional cases only. This means that the number of valid cases will differ in the predicted and observed set, even if the model would predict the number of tours/trips accurately.

The Albatross system consists of 20 zones (i.e. origins and destinations) that are used as basis for each OD-matrix (see Table 3.3). Different matrices are generated varying a third dimension on which the interactions are broken down. This third dimension can include:

1. No third dimension
2. Transport mode (car driver, slow mode, car passenger, public transport, unknown transport mode)
3. Day of the week (weekday, Saturday, Sunday)
4. Primary activity (working activity outdoors, medical visit, bring-get activity, non-leisure activity outdoors, non-grocery shopping, grocery shopping, leisure activity outdoors, social visit outdoors, service activity, other activity outdoors, inhome activity).

Note that the number of cells and hence, the degree of disaggregation, differs between the matrices. For example, the OD-matrix by mode has 5 × 20 × 20 = 2000 cells, while the OD-matrix by activity type has 11 × 20 × 20 = 4400 cells.

The measure that will be used for determining the degree of correspondence between the observed and predicted matrices is the correlation coefficient. It will be calculated between observed and predicted matrix entries in general and for three trip matrices that are disaggregated, each time in a different way, based on some selected trip facets. The facets considered include transport mode, day-of-the-week

Table 3.3: Zones for OD-matrices in Albatross

Region	Zone	Postal code
Rotterdam-Noord	1	302,303,304,305
	2	306
	3	301
Rotterdam-Zuid	4	308
	5	307
	6	298
Hendrik-Ido-Ambacht	7	3341
	8	3342
	9	3343
Zwijndrecht	10	3331
	11	3332
	12	3333
	13	3334
	14	3335
	15	3336
Paependrecht, Sliedrecht, Dordrecht	16	335,336
	17	3311,3312
	18	3314,3316,3317
	19	3313,3315,3318, 3319,3328,3329
Elsewhere	20	other

and activity (purpose), as discussed above. What is now meant by a correlation coefficient between matrices? In all cases, the cells of the OD-matrices are rearranged into a single vector across categories and the correlation coefficient will be calculated by comparing the corresponding elements in the observed and the predicted vector. Thus, for the OD-matrices disaggregated on the day of the week, the cells of the matrices on weekday, Saturday and Sunday are rearranged into three separate vectors, and these three vectors are combined into one single vector. This occurs for the observed and the predicted matrices, and the correlation coefficient between this observed and predicted vector is the performance measure at trip matrix level. An advantage of the use of the correlation coefficient is that it is insensitive to the difference in scale between column frequencies (i.e. the difference in the total number of trips).

Chapter 4

Use of Simple Models in the Analysis of Activity-Diary Data

4.1 Introduction

In the past few years, activity-based forecasting of travel demand has become a major field of interest in transportation research. The aim of activity-based models is to predict which activities will be conducted where, when, for how long, with whom and with which transport mode. A comparison of a rule-based model and utility-maximising models on activity-travel patterns has been carried out (Arentze *et al.*, 2000b) and the rule-based system seems to be very flexible. The rule-based system also performs well in predicting transport choice behaviour if we used an induction technique (Wets *et al.*, 2000 and Doherty, 2001). Although these rule-based models perform very well, they also show some limitations. Most of them are based on quite complex rule sets. However, already in the Middle Ages, there was a call for trying to simplify things: William of Occam's razor states that 'Nunquam ponenda est pluralitas sin necesitate', meaning 'Entities should not be multiplied beyond necessity' (Tornay, 1938). It was born in the Middle Ages as a criticism of scholastic philosophy, whose theories grew ever more elaborate without any corresponding improvement in predictive power. In the intervening centuries it has come to be seen as one of the fundamental tenets of modern science and today it is often invoked by learning theorists as a justification for preferring simpler models over more complex ones. However, Domingos (1998) learned us that it is tricky to interpret Occam's razor in the right way. The interpretation "Simplicity is a goal in itself" is essentially correct, while "Simplicity leads to greater accuracy" is not.

On the one hand, research in the field of psychology shows that there is empirical evidence that simple models, based on fast and frugal heuristics that employ a minimum of time, knowledge and computation to make adaptive choices in real environments (Gigerenzer *et al.*, 1999), often predict human behaviour very well. This is simply because nowadays, people do not have the time to try to attain some optimal state in making choices, most choices have to be made very quickly (e.g. at the stock market, in a hospital). These heuristics have been tested extensively in psychological environments, and they have proved to work well, so now we will illustrate their use in a transport environment. Moons *et al.* (2001) examined the performance of simple classifiers for the transport mode dimension of the Albatross model system. We discovered that the predictive performance of these simple heuristics was only slightly less than that of a more complex induction algorithm. This chapter addresses the question what results these simple algorithms will give once they are built in the activity-diary scheme of Albatross.

On the other hand, a less drastic way of simplification is proposed as a second solution to the complex rule sets that are often derived from activity-diary data. While a larger number of rules may be valuable when one wishes to better understand the data, from a predictive perspective a large number of rules may imply that the decision tree induction algorithm has over-fitted the data. The obtained decision tree structure (set of decision rules) may then be very unstable and sensitive to highly correlated covariates.

Feature selection offers a solution to reducing the number of irrelevant attributes and as a consequence often the size of the decision tree will also be reduced. The key notion underlying feature selection is that the number of decision rules (size of the tree) is reduced by selecting and deleting irrelevant features (explanatory variables), based on some statistical measure. The impact of feature selection on the predictive performance of rule-based models is however not a priori clear. On the one hand, because the irrelevant conditions are deleted, feature selection may not have a substantial negative effect on predictive performance. However, a smaller decision tree may also result in a higher probability of misclassification, leading to worse predictive performance. It is against this background that the present chapter also reports the findings of a methodological study that was conducted to gain a better understanding of the influence of a smaller set of decision rules (trimmed decision tree) on the predictive performance of sequential models of activity scheduling behaviour in general and the Albatross model in particular. Moons *et al.* (2002a, 2002c, 2005b) investigated the influence of irrelevant attributes on the performance of the decision tree for the transport mode, the travel party, the activity duration and the location agent of the Albatross model system. We found that a trimmed decision tree, involving considerable less decision rules, did not result in a significant drop in predictive performance compared to the original larger set of rules that was derived from the activity-travel diaries. Similar techniques have been applied in completely different research domains: marketing (Buckinx *et al.*, 2004), artificial intelligence (Koller and Sahami, 1996; Kohavi *et al.*, 1994), bioinformatics (Zheng *et al.*, 2003), etc. In this chapter, the question 'To what extent can this result be generalised to the full set of decision trees, representing different choice facets, that make up the complete Albatross model system?' is inspected.

This chapter explores thus two examples of fast and frugal heuristics and the application of feature selection to decision rule induction and it compares their performance in Albatross. The predictive performance will also be evaluated at activity pattern level, where observed and generated sequences of activities are compared and at trip matrix level where the correlation coefficients that determine the strength of the associations between the observed and predicted origin-destination matrices are judged against each other.

In the next section, we will briefly introduce the different methods used in our analysis. At first, we will describe the simple heuristics, 1R and Naïve Bayes, while secondly the CHAID algorithm, used to determine the original set of rules of the Albatross system, is introduced. Finally, the C4.5 decision tree algorithm is introduced, together with the Relief-F feature selection method, since these two algorithms can be easily combined.

4.2 Methods

4.2.1 Simple Classifiers

There are in the literature some indications that very simple rules may achieve a surprisingly high accuracy on many data sets. For example, Rendell and Seshu (1990) occasionally remark that many real world data sets have 'few peaks (often just one)' and are therefore 'easy to learn'. Further evidence is provided by studies of pruning methods (e.g. Buntine and Niblett, 1992; Clark and Niblett, 1989; Mingers, 1989), where the accuracy is rarely seen to decrease as pruning becomes more severe. This is even so when the rules are pruned to the extreme, using only one or two variables. The most compelling initial indication that very simple rules often perform well, occurs in Weiss *et al.* (1990). In four of the five data sets studied, classification rules involving two or fewer attributes outperformed the more complex rules.

Therefore, in the next section, we will use two simple classifiers, 1R and Naïve Bayes, in order to set up the set of rules for each of the dimensions in the Albatross system and we will compare their performance to that of the standardly used CHAID algorithm.

One R

Holte developed a very simple classifier that provides a rule based on the value of a single attribute. This algorithm, which he called 'One R', may compete with state-of-the-art techniques used in the field (Holte, 1993). It turns out that simple rules frequently achieve surprisingly high accuracy. Perhaps this is because the structure underlying many real-world data sets is quite rudimentary and just one variable is sufficient to determine the outcome of an instance reasonably accurate.

Like other algorithms, One R takes as input a set of several attributes and a categorical output variable. Its goal is to infer a rule that predicts the class of the dependent variable given the values of the independent variables. The One R algorithm chooses the most informative single attribute and bases the rule solely on this attribute. Full details can be found in Holte's paper, but the basic idea is given below. The algorithm assumes that the attributes are discrete. If not, they must be discretised. Any method for turning a range of values into disjoint intervals must take care to avoid creating large numbers of rules with many small intervals. This is known as the problem of 'overfitting', because such rules are overly specific to the data set and do not generalise well. Holte achieves this by requiring all intervals (except the rightmost) to contain more than a predefined number of examples in the same class of the outcome variable. Empirical evidence (Holte *et al.*, 1989) led to a value of six for data sets with large number of instances and three for smaller data sets (with less than 50 instances).
The accuracy is measured by the number of correctly classified instances.

```
For each attribute $a$, form a rule as follows:
    For each value $v$ from the domain of $a$,
        Let $c$ be the most frequent class in the set of
            instances where $a$ has value $v$.
        Add the following clause to the rule for $a$:
            if $a$ has value $v$ then the class is $c$
    Calculate the classification accuracy of this rule.
Use the rule with the highest accuracy.
```

Holte (1993) performed a comprehensive study on the performance of the One R algorithm and results on sixteen data sets, that are frequently used by machine learning researchers, were reported. It turned out that despite its simplicity, the One R algorithm did astonishingly well in comparison with state-of-the-art induction algorithms. The rules that can be derived from the One R procedure are often a viable alternative to more complex structures. This strongly encourages a 'simplicity first' methodology in which the baseline performance is established using simple, rudimentary techniques before progressing to more sophisticated learning schemes, which inevitably generate output that is harder to interpret.

Naïve Bayes

The Naïve Bayes classifier (Good, 1965; Duda and Hart, 1973; Langley *et al.*, 1992), named after Bayes' rule and its naïve assumption of independence, is built on a conditional independence model of each attribute given the class. Again, this algorithm assumes that the attributes are discrete. If not, they must be discretised.It can be used to predict the class value of the outcome variable for a new instance. Formally, the probability of a class value C_i for an instance $X = [A_1, \ldots, A_n]$, consisting of n attribute values, is given by

$$P(C_i|X)$$

$$= \frac{P(X|C_i) \cdot P(C_i)}{P(X)} \qquad \text{by Bayes rule}$$

$$\propto P(A_1, \ldots, A_n|C_i) \cdot P(C_i) \qquad \text{proportional, since P(X) is the same for all values}$$

$$= \prod_{j=1}^{n} P(A_j|C_i) \cdot P(C_i) \qquad \text{by conditional independence assumption}$$

The above probability is computed for each class and the prediction is made for the class with the largest posterior probability. Instead of using just one attribute, as One

R does, this model thus uses all attributes and allows them to make contributions to the decision that are equally important and independent of one another, given the class. The model was shown to be surprisingly robust to obvious violations of this independence assumption, yielding accurate classification models even when there are clear conditional dependencies (Langley *et al.*, 1992; Domingos and Pazzani, 1996). It works particularly well when it is combined with variable selection procedures that serve to eliminate redundant and hence non-independent attributes.

Standard Albatross Classifier: CHAID

The popular technique of automatic interaction detection (AID) was described by Morgan and Sonquist (1963a, 1963b), amongst others. The technique used here, CHAID, is an offshoot of AID designed for a categorical dependent variable (Kass, 1980).

In AID, the data are successively bisected using a predictor, preserving the ordered nature of the categories where appropriate.

AID operates on an interval scaled dependent variable and maximises the between-group sum of squares at each bisection. In contrast, CHAID operates on a nominal scaled dependent variable and maximises the significance of a chi-squared statistic at each partition, which need not to be a bisection.

Standard AID is liable to misuse and it never really takes into account the sampling variability inherent in the data. CHAID tackles this problem by embedding the partitioning problem in a significance testing framework. This allows the formation and examination of multi-way splits which often lead to the conclusion that a predictor is indivisible according to the criterion.

The selection procedure of AID favours predictors with more categories since the maximisation criterion extends over more possibilities. A consequence of using significance testing in the decision-making process of CHAID is to nullify this bias.

CHAID proceeds in steps: it first detects the best partition for each predictor. Then the predictors are compared and the best one is chosen. The data are subdivided according to this chosen predictor. Each of these subgroups are then re-analysed independently, to produce further subdivisions for analysis.

Let the dependent variable Y have $d \leq 2$ response categories, and a particular predictor under analysis X have $c \leq 2$ categories. A subproblem in the analysis is to reduce the given $c \times d$ contingency table to the most significant $j \times d$ table by combining (in an allowable manner) categories of the predictor. Conceptually, one may first calculate statistics, $T_j^{(i)}$, the usual χ^2 statistics for the i-th method of forming a $j \times d$ table ($j = 2, 3, \ldots, c$; the range of i depending on the type of predictor). Then, if $T_j^* = \max_i T_j^{(i)}$ is the χ^2 statistic for the best $j \times d$ table, choose the most significant T_j^*. The full algorithm is as follows:

Step 1. For each predictor in turn: cross-tabulate the categories of the predictor with the categories of the dependent variable and do *steps 1a and 1b.*

> *Step 1a.* Find the pair of categories of the predictor (only considering allowable pairs as determined by type of the predictor) whose $2 \times d$ sub-table is least significantly different. If this significance does not reach a critical value, merge the two categories, consider this merger as a single compound category, and repeat this *step.*
>
> *Step 1b.* For each compound category consisting of three or more of the original categories, find the most significant binary split (constrained by the type of predictor) into which the merger may be resolved. If the significance is beyond a critical value, implement the split and return to *step 1a.*

Step 2. Calculate the significance of each optimally merged predictor, and isolate the most significant one. If this significance is greater than a criterion value, subdivide the data according to the (merged) category of the chosen predictor.

Step 3. For each partition of the data that has not yet been analysed, return to *step 1.* This step may be modified by excluding from further analysis partitions with a small number of observations (to ensure the validity of the significance test).

The traditional χ^2-based test of independence between two categorical variables is used to determine the significance of each partitioning (in steps 1a, 1b, 2) (e.g. using the 0.05 α-level). In step 2, the significance level of each optimally merged $j \times d$ contingency table is adjusted by a multiplier. These Bonferroni multipliers imply a more conservative test to account for the fact that several ways of splitting have been evaluated. The value of these Bonferroni multipliers depends on the number of ways in which c categories can be merged into j groups, and thus they are different for different type of predictors.

The three different types allowed by CHAID are:

1. *Monotonic predictors.* As in AID, a monotonic predictor is one whose categories lie on an ordinal scale. This implies that only *contiguous* categories may be grouped together. The Bonferroni multiplier is easily derived in this case:

$$B_{monotonic} = \binom{c-1}{j-1}.$$

2. *Free predictors.* Again as in conventional AID, a free predictor is one whose categories are purely nominal. This implies that any grouping of categories is permissible. By rephrasing it as an appropriate occupancy problem, Feller (1968) derived the multiplier to be:

$$B_{free} = \sum_{i=0}^{j-1} (-1)^i \frac{(j-i)^c}{i!(j-i)!}.$$

3. *Floating predictors.* In many practical cases, the categories of a predictor lie on an ordinal scale with the exception of a single category that either does not belong with the rest, or whose position on the ordinal scale is unknown. Kass (1980) calls this category a 'floating' category and the predictor a floating predictor. This situation typically arises when an investigation allows for an unknown or missing category.

   Except for the floating category, grouping is only allowed for contiguous categories as for the monotonic predictors. The floating category, however, may stand alone or be combined with any other category or group of categories. The Bonferroni multiplier comes from an extension of the monotonic case:

$$B_{floating} = \binom{c-2}{j-2} + j\binom{c-2}{j-1} = \frac{j-1+j(c-j)}{c-1} \times B_{monotonic}.$$

This partitioning can now be turned into a decision table (which is necessary as input for the Albatross system). Arentze and Timmermans (2000) propose a probabilistic rule for assigning responses to columns. If f_{jk} is the observed frequency of the k-th response in the j-th column of a particular contingency table and $n_j = \sum_k f_{jk}$ is the total number of cases assigned to the j-th column, then the probability of assigning any (new) case that falls in the j-th column to response k equals $\frac{f_{jk}}{n_j}$.

The χ^2-criterion finds a very accurate decision table and, at the same time, a parsimonious decision table for prediction. The criterion implies a partitioning of the sample space that maximises the difference in response distributions between columns. It yields a very accurate decision table in the sense that for none of the columns it is possible to define an additional one-way or multi-way split that would yield significantly different responses. The criterion results in a parsimonious decision table in that every implemented split is indeed significant. Alternative measures, such as entropy (see e.g. 4.2.2) are also appropriate, but the advantage of χ^2 is that its distribution is known so that significance testing is possible. On the other hand, there are also some negative points. First, there is no guarantee that the produced partitioning is optimal. Splits are considered for one independent variable at the time, and therefore, are not considered in interaction or reconsidered dependent on later decisions. Furthermore, despite the significance testing, the process may still capitalise on chance. At the level of each explanatory variable, the Bonferroni adjustments make sure that the search for possible ways to implement splits is taken into account. However, the test does not take into account that multiple predictors have been tested. In practice, however, the Bonferroni adjustments are very conservative so that it is unlikely that the differences in data are based on chance.

4.2.2 Decision Tree Induction and Feature Selection

Decision Tree Induction: C4.5

Decision tree induction can be best understood as being similar to parameter estimation methods in econometric models. The goal of tree induction is to find the set of Boolean rules that best represents the empirical data. The original Albatross system was derived using a Chi-square based approach (see subsection 4.2.1). In this subsection, however, the trees were re-induced using the C4.5 method (Quinlan, 1993) because this method can be easily combined with the Relief-F feature selection. Arentze *et al.* (2000a) found approximately equal performance in terms of goodness-of-fit of the two methods in a representative case study. The C4.5 algorithm works as follows. Given a set of I choice observations taken from activity-travel diary data, consider their values on n different explanatory variables or attributes $x_{i1}, x_{i2}, \ldots x_{in}$ and on the response variable $y_i \in \{1, 2, \ldots, p\}$ for $i = 1, \ldots I$. Starting from the root node, each node will be split subsequently into internal or terminal nodes. A leaf node is terminal when it has no offspring nodes. An internal node is split by considering all allowable splits for all variables and the best split is the one with the most homogeneous daughter nodes. More detailed information on recursive partitioning can be found in Chapter 6. The C4.5 algorithm recursively splits the sample space on X into increasingly homogeneous partitions in terms of the response variable Y, until the leaf nodes contain only cases from a single response class. Increase in homogeneity achieved by a candidate split is measured in terms of an information gain ratio. As stated in Quinlan (1993), the information theory on which the gain ratio criterion is based can be explained in the following statement:

Definition 4.2.1 *Information of a message*
The information conveyed by a message depends on its probability and can be measured in bits as minus the logarithm to base 2 of that probability.

For example, if there are four equally probable messages, the information conveyed by any of them is $-log_2(1/4) = 2$ bits.

Definition 4.2.2 *Information of a message that a random case belongs to a certain class*

$$-\log_2\Big(\frac{freq(C,T)}{|T|}\Big) bits$$

with T a training set of cases, C a class, $freq(C,T)$ the number of cases in T that belongs to class C and $|T|$ the number of cases in T.

Based on these definitions, the average amount of information needed to identify the class of a case in a training set (also called entropy) can be deduced as follows:

Definition 4.2.3 *Entropy of a training set*

$$info(T) = -\sum_{i=1}^{k} \frac{freq(C_i,T)}{|T|} \times \log_2\Big(\frac{freq(C_i,T)}{|T|}\Big) bits$$

with T a training set of cases, k the number of classes, C_i a class i, $freq(C_i,T)$ the number of cases in T that belongs to class C_i and $|T|$ the number of cases in T.

Entropy can also be measured after that T has been partitioned in n sets using the outcome of a test carried out on attribute X. This yields:

Definition 4.2.4 *Entropy after the training set has been partitioned on a test X*

$$info_X(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times info(T_i)$$

Using these two measurements, the *gain criterion* can be defined as follows:

Definition 4.2.5 *Gain criterion*

$$gain(X) = info(T) - info_X(T)$$

The gain criterion measures the information gained by partitioning the training set using the test X. In ID3, the ancestor of C4.5, the test selected is the one which maximises this information gain because one may expect the remaining subsets in the branches will be the most easy to partition. Note, however, that by no means this is certain because we have looked ahead only one level deep in the tree. The gain criterion has only proved to be a good heuristic. Although the gain criterion performed quite well in practice, the criterion has one serious deficiency, i.e. it tends to favour tests with many outcomes. Therefore, in C4.5, a somewhat adapted form of the gain criterion is used. This criterion is called the *gain ratio criterion.* In this criterion, the gain attributable to tests with many outcomes is adjusted using some kind of normalisation. In particular, the *split info*(X) measurement has to be defined.

Definition 4.2.6 *Split info of a test X*

$$\mathit{split\ info}(X) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \times \log_2\Big(\frac{|T_i|}{|T|}\Big)$$

This indicates the information generated by partitioning T into n subsets. Using this measure, the gain ratio is defined as follows:

Definition 4.2.7 *Gain ratio*

$$\mathit{gain\ ratio}(X) = \frac{\mathit{gain}(X)}{\mathit{split\ info}(X)}$$

This ratio represents how much of the gained information is useful for classification. In case of very small values of split info(X) (in case of trivial splits), the ratio will tend to infinity. Therefore, C4.5 will select the test which maximises the gain ratio, but subject to the constraint that the information gain must be at least as large as the average information gain over all possible tests. After building the tree, pruning strategies are adopted. This means that the decision tree is simplified by discarding one or more sub-branches and replacing them with leaves.

Feature Selection: Relief-F

Feature or variable selection strategies are often implied to explore the effect of irrelevant attributes on the performance of classifier systems. A feature selection method ranks all the attributes (features) in descending order of relevance. This relevance can be measured in several ways, leading to two large subclasses in feature selection methods: the filter and the wrapper approach. The fundamental difference between them is the evaluation criterion used to select or rank attributes. For wrappers, the selection or ranking results from the estimation of the performance on the associated induction algorithm, while the filter approach only makes use of the characteristics of the data itself. Both methods have been compared extensively (Hall, 1999a, 1999b; Koller and Sahami, 1996). In this analysis, the filter approach, more specifically the Relief-F feature selection method, is opted for since it can handle multiple classes of the dependent variable (the nine different choice facets that we are predicting range from two to seven classes) and above that it is easily combined with the C4.5 induction algorithm.

Feature selection strategies can be regarded as one way of coping with the correlation between the attributes. This is relevant because the structure of trees is sensitive to the problem of multi-collinearity, which implies that some variables would be redundant (given the presence of other variables). Redundant variables do not affect the impacts of the remaining variables in the tree model, but it would simply be better if they were not used for splitting. Therefore, a good feature selection method for this analysis would search for a subset of relevant features that are highly correlated with the class variable that the tree-induction algorithm is trying to predict, while mutually having the lowest possible correlations.

Relief (Kira and Rendall, 1992), the predecessor of Relief-F, is a distance-based feature weighting algorithm. It orders attributes according to their importance. To each attribute it assigns the initial value of zero that will be adapted with each run through the instances of the data set. The features with the highest values are considered to be the most relevant, while those with values close to zero or with negative values are judged irrelevant. Thus Relief imposes a ranking on features by assigning each a weight. The weight for a particular feature reflects its relevance in distinguishing the classes. In determining the weights, the concepts of *near-hit* and *near-miss* are central. A *near-hit* of instance i is defined as the instance that is closest to i (based on Euclidean distance between two instances in the n-dimensional variable space) and which is of the same class (concerning the output or dependent variable), while a *near-miss* of i is defined as the instance that is closest to i (based on

Euclidean distance) and which is of a different class (concerning the output variable). The algorithm attempts to approximate the following difference of probabilities for the weight of a feature X:

$$\begin{aligned} W_X &= P(\text{different value of } X|\text{nearest instance of different class}) \\ &- P(\text{different value of } X|\text{nearest instance of same class}). \end{aligned}$$

So, Relief works by random sampling an instance and locating its nearest neighbour from the same and opposite response class. The concept of a nearest neighbour is defined in terms of the Euclidean distance, so in an n-dimensional sample space, determined by the variables $X_1, \ldots, X_n$, the following distance measure will be used: $d(\mathbf{i}, \mathbf{j}) = \left(\sum_{k=1}^{n}(x_{ik} - x_{jk})^2\right)^{\frac{1}{2}}$, where $\mathbf{i} = (x_{i1}, \ldots, x_{in})$ and $\mathbf{j}$ are two n-dimensional vectors.

By removing the context sensitivity provided by the "nearest instance" condition, attributes are treated as mutually independent, and the previous equation becomes:

$$\begin{aligned} \text{Relief}_X &= P(\text{different value of } X|\text{different class}) \\ &- P(\text{different value of } X|\text{same class}). \end{aligned}$$

Relief-F (Kononenko, 1994) is an extension of Relief that can handle multiple classes and noise caused by missing values, outliers, etc. To increase the reliability of Relief's weight estimation, Relief-F finds the k nearest hits and misses for a given instance, where k is a parameter that can be specified by the user. For multiple class problems, Relief-F searches for nearest misses from each different class (with respect to the given instance) and averages their contribution. The average is weighted by the prior probability of each class.

4.3 Analysis and Results

The overall aim of this study is to investigate whether a simplification of the rule sets underlying the Albatross model leads to a significant loss in predictive power. This simplification can be obtained in two ways: either by the application of simple classifiers or by reducing the set of decision rules through the application of a feature selection method (Moons *et al.*, 2005a). The original model consists of nine choice facets. For each of these choice facets, a set of decision rules was extracted from activity-travel diaries. To predict activity-travel patterns, these decision trees are executed sequentially in the Albatross system according to some scheduling process model (see Chapter 3). We will investigate the effect of simpler rules for each choice facet.

4.3.1 Study Design

We have split the original data set into two subsets. A training set, containing the first 75% of the cases, on which the different models will be built and optimised for each choice facet. The remaining 25% of the cases make up the validation or test set that can be used to compute the accuracies (percentage of correctly classified instances), etc. These percentages are arbitrary but are common practice in validation studies (see e.g. Wets *et al.*, 2000).

Note that for reasons of comparison, the Zero R classifier has also been added in the analyses. This algorithm automatically classifies new cases to the majority class. For the second way of simplification, we will first build decision trees for each of the nine choice facets, using the C4.5 algorithm (Quinlan, 1993). This approach will be called the full approach. Next, we will first identify the relevant attributes for each of the nine choice facets separately, and then build the C4.5 trees incorporating only a subset of the most relevant attributes. This approach will be called the feature selection approach. In our first analysis (the 'full' approach), the C4.5 trees were induced based on one simple restriction: the final number of cases in a leaf node must meet a minimum. For eight out of the nine choice facets, this minimum was set to 15 (except for the very large data set of the 'select'-dimension, where this number was set to 30). In the second analysis (the feature selection approach), all the irrelevant attributes were first removed from the data by means of Relief-F feature selection (FS) method with the k parameter set equal to 10. Next, the C4.5 trees were built based on the same restriction as in the 'full' approach, though only the remaining relevant attributes were used. To determine the selection of variables, the following procedure was adopted. Several decision trees were built, each time removing one more irrelevant attribute, as they appeared lowest in the ranking that has been provided by the FS method. For each of these decision trees, the accuracy was calculated and compared to the accuracy of the decision tree of the full approach. The smallest decision tree, which resulted in a maximum decrease of 2% in accuracy compared to the decision tree including all features, was chosen as the final model for a single choice facet in the feature selection approach. This strategy was applied to all nine dimensions of the Albatross model.

To use a decision tree for prediction (note that the result of all different methods can be regarded as a decision tree), a rule needs to be specified that assigns a class Y_i to each case classified by the tree. Instead of just using the commonly used deterministic assignment rule of the decision tree, a probabilistic assignment rule was used since this might result in a better prediction of the aggregate distributions in the activity diary data. Each rule was assigned a probability distribution that was derived from the frequency distribution over the different alternatives in the training set for each leaf. These corresponding probabilities will be reflected in the predicted activity schedules. For each choice facet, this set of probabilistic rules gives us the decision tables that are used in the analysis. A simplified example of a decision tree and its corresponding probability distributions are presented in the Figure 4.1.

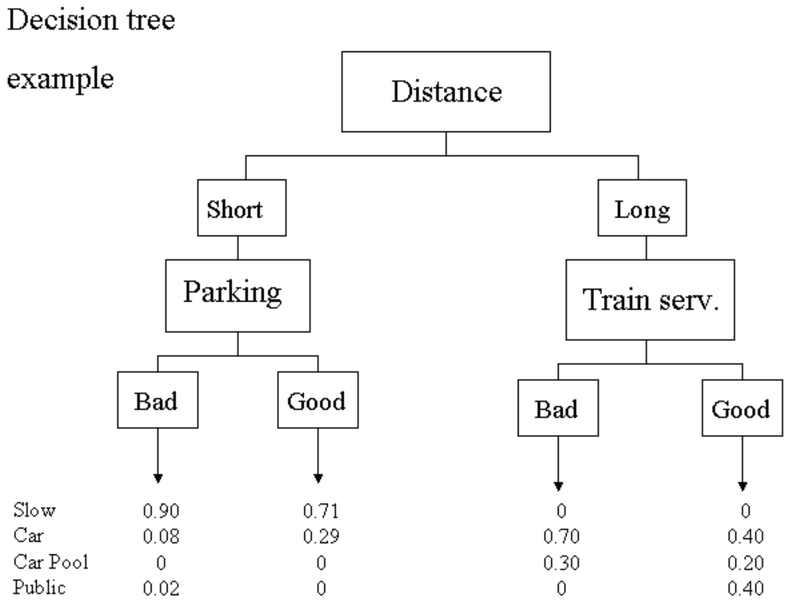

Figure 4.1: *Example of a decision tree*

4.3.2 General Results

At first, we will take a closer look at the average length of the observed and predicted sequences of activities. In the observed patterns, the average number of activities equals 5.160 for the training set and 5.155 for the test set. This average length offers room for 1-3 flexible activities complemented with 2-4 in-home activities. Considerable variation occurs, however, as indicated by the standard deviation of approximately 3 activities. Some descriptive statistics of the predicted patterns are shown in Table 4.1.

Table 4.1: Average number of predicted activities in the sequences (standard deviation between brackets)

	Training Set	Test Set
Zero R	5.217 (3.241)	5.199 (3.333)
One R	5.198 (3.182)	5.178 (3.128)
Naïve Bayes	5.071 (3.210)	5.088 (3.100)
CHAID	5.463 (2.970)	5.363 (2.783)
Full approach	5.286 (2.953)	5.286 (2.937)
Feature selection approach	5.014 (3.033)	4.907 (2.921)

On average, when comparing the simple classifiers, Zero R and One R overestimate the number of activities, however, this overestimation is somewhat less pronounced on the test set. Naïve Bayes on the other hand tends to underestimate the number of activities a little bit. All models seem to overestimate the variance a little, both on the training set and on the test set. We observe that in general the 'full' approach as well as the CHAID approach predict activity sequences that are somewhat too long, while those of the feature selection approach are rather a little bit too short. The variation is again overestimated by all three approaches.

The results of these different methods will now be compared at three levels of aggregation (see Chapter 3): the choice facet level, the activity pattern level and the trip matrix level. At the choice facet level, we will discuss the attributes that remained in the final decision tree model of each of the two approaches. The probability of a correct prediction and a measure of relative performance are also calculated for each decision tree. At the activity pattern level, sequence alignment methods (Joh, *et al.*, 2001a, 2001b, 2001c, 2002a) were used to assess the correspondence between the observed and predicted activity sequences. At the trip matrix level, correlation coefficients are calculated to measure the degree of correspondence between the observed and the predicted Origin-Destination matrices.

4.3.3 Choice Facet Level

Tables 4.2 to 4.7 provide the results of the analyses conducted to assess model performance at the choice facet level. As stated in Chapter 3, the first column of these tables presents the nine choice facets of Albatross. The second column lists the number of alternatives (levels of the Y-variable), while the third column gives the total number of attributes that were considered to build the final decision tree. The fourth column depicts the total number of leafs of the decision tree, i.e. the number of probabilistic rules in the decision table. Column five reports the probability of a correct prediction and in the last column a measure of relative performance, where the probability of a correct prediction is compared to the probability of a correct prediction under a null model. This null model assigns a new case to a category of the Y-variable with a probability, equal to the number of observed cases in the category divided by the total number of cases in the data set.

Table 4.2: Performance at choice facet level (‘Zero R approach’)

Dimension	♯ alts	♯ attrs	♯ leafs	e	e_{ratio}
Mode for work	4	0	1	0.525	0.000
Selection	2	0	1	0.669	0.000
With-whom	3	0	1	0.355	0.000
Duration	3	0	1	0.334	0.000
Start time	6	0	1	0.172	0.000
Trip chain	4	0	1	0.533	0.000
Mode other	4	0	1	0.388	0.000
Location 1	7	0	1	0.375	0.000
Location 2	6	0	1	0.200	0.000

The results of the previous analyses show that, in general, the standardly used CHAID approach outperforms the other approaches on the dimensions separately. Overall, Naïve Bayes performs better than One R on predictive power. There is not much difference between the two partitioning algorithms, CHAID and C4.5 (or the ‘full’ approach), on some choice facets CHAID performs better, on other ones C4.5. The C4.5 generally needs less variables to build the trees. Also the complexity of both trees is comparable.

Table 4.3: Performance at choice facet level (‘One R approach’)

Dimension	♯ alts	♯ attrs	♯ leafs	e	e_{ratio}
Mode for work	4	1	6	0.595	0.147
Selection	2	1	5	0.677	0.025
With-whom	3	1	5	0.408	0.082
Duration	3	1	3	0.348	0.020
Start time	6	1	4	0.227	0.067
Trip chain	4	1	2	0.699	0.354
Mode other	4	1	4	0.413	0.040
Location 1	7	1	3	0.435	0.096
Location 2	6	1	3	0.234	0.043

Table 4.4: Performance at choice facet level (‘Naïve Bayes approach’)

Dimension	♯ alts	♯ attrs	♯ leafs	e	e_{ratio}
Mode for work	4	3	96	0.641	0.245
Selection	2	2	42	0.674	0.016
With-whom	3	3	140	0.458	0.160
Duration	3	3	60	0.370	0.053
Start time	6	3	64	0.318	0.176
Trip chain	4	2	4	0.765	0.497
Mode other	4	3	60	0.450	0.102
Location 1	7	2	15	0.475	0.161
Location 2	6	3	12	0.281	0.102

Table 4.5: Performance at choice facet level (‘CHAID approach’)

Dimension	♯ alts	♯ attrs	♯ leafs	e	e_{ratio}
Mode for work	4	32	23	0.648	0.259
Selection	2	40	106	0.724	0.166
With-whom	3	39	57	0.509	0.239
Duration	3	41	61	0.413	0.119
Start time	6	63	86	0.398	0.273
Trip chain	4	53	30	0.833	0.642
Mode other	4	35	65	0.528	0.229
Location 1	7	28	62	0.575	0.320
Location 2	6	28	34	0.354	0.193

Table 4.6: Performance at choice facet level ('full approach')

Dimension	♯ alts	♯ attrs	♯ leafs	e	e_{ratio}
Mode for work	4	3	8	0.598	0.155
Selection	2	15	35	0.686	0.052
With-whom	3	19	72	0.499	0.223
Duration	3	28	148	0.431	0.145
Start time	6	28	121	0.408	0.285
Trip chain	4	4	8	0.802	0.576
Mode other	4	15	63	0.524	0.222
Location 1	7	8	30	0.540	0.264
Location 2	6	15	47	0.372	0.214

Table 4.7: Performance at choice facet level ('feature selection approach')

Dimension	♯ alts	♯ attrs	♯ leafs	e	e_{ratio}
Mode for work	4	2	6	0.595	0.147
Selection	2	1	1	0.669	0.000
With-whom	3	4	51	0.467	0.173
Duration	3	4	38	0.368	0.051
Start time	6	8	1	0.172	0.000
Trip chain	4	10	13	0.811	0.596
Mode other	4	11	60	0.508	0.196
Location 1	7	6	15	0.513	0.222
Location 2	6	8	14	0.312	0.141

The results also indicate that feature selection generally generates considerably less complex decision trees than the full approach. One exception is the 'trip chaining' choice facet, which more leafs in the final tree with FS than in the tree without feature selection. A logical consequence of this result is that the measure of relative performance of the models with FS is smaller.

Another analysis at the choice facet level is concerned with comparing the most important attributes for the 'full' and the FS approach. In Tables 4.8 and 4.9, the maximally four most relevant attributes for predicting each choice facet are described for each approach. For the definition of the variables, we refer to the Appendix. The attributes on which the tree makes its first splits are considered to be more relevant than attributes on which splits are based further down in the partitioning process.

For the *Mode for work* facet, clearly only transport characteristics are important. For the One R and the feature selection approach even only the shortest travel time by bike seems to be relevant for the prediction of the transport mode for work. This might seem odd at first sight, but one has to bear in mind that the biking facilities in the Netherlands are very good, and thus if there is the possibility of going to work by bike, a lot of people will do so. In the other approaches, the variable selection seems to point especially at the distinction between bike as transport mode or car. There has to be added that the data set is quite skewed in favour of car as most frequent transport mode for work, this possibly gives an explanation why variables that are concerned with public transport do not appear among the four most important variables in predicting this choice facet. The one variable that is needed to make the splits in the feature selection approach (Tbike) does not occur in the list of the three variables needed to build the tree in the 'full' approach, however, this variable is fairly high correlated with one variable of the 'full' approach: ρ(Tbike, Rcabi) = -0.72.

In case of the *Selection* choice facet, the activity type and the day of the week appear to be important features as they are the only variables that appear on two different approaches. The FS approach uses the unconditional probabilities of the Zero R method. The reason for applying a single rule can be that the distribution of the choice variable in this data set is very skew. 79% of the entire data set can be explained by this default rule. Another reason might be that C4.5 only splits its decisions based on the modal class and not on the total frequency distribution over the classes of the response variable. If this latter would have been the case, there would be at least a split on the activity type, since the chance on a 'yes'-response varies highly over the different types of activities. The CHAID approach points more at person and household characteristics. In the 'full' approach, we observe that the probability of a correct prediction is not much higher than compared to the Zero R and the FS approach, although fifteen variables were used to build the tree. So probably there is information lacking to predict this choice facet accurately. Most methods take the activity type into consideration (especially grocery shopping appeared to have a high impact in the schedule making in the 'full' approach), which seems logical and apart from this, also the time component is a very crucial one. Surprisingly, these are not the variables that occur highest in the ranking of the feature selection method. However, we have to bear in mind that the C4.5 algorithm (in the 'full' approach) does not take any correlation into account (e.g. ρ(yAvail^3, Atype) = 0.22), so it can select correlated attributes to build the tree if they increase the homogeneity of the split.

Table 4.8: Description of the most important attributes for each approach for the first five dimensions

Choice facet	Attribute	One R	Naïve Bayes	CHAID	Full approach	FS approach
Mode for work	Rcabi				*	
	PTTMax				*	
	Ncar			*	*	
	Tbike	*	*	*		*
	Rpubi		*			
	Rpuca		*	*		
	Ccomp			*		
Selection	yAvail^3				*	
	Tmax(4)				*	
	yDshop				*	
	Atype	*		*	*	
	day		*	*		
	Nsec		*			
	Gend			*		
	Ccomp			*		
With-whom	yAvail^4				*	
	Atype	*		*	*	*
	yLeis				*	
	yCar(2)				*	
	Ccomp		*			*
	day		*	*		*
	Cchild		*	*		
	yCar(4)			*		
Duration	yAvail3^5				*	
	Awith^5	*	*	*	*	*
	yAvail2^5				*	
	Tleis				*	
	day			*		*
	Atype		*	*		*
	Csec		*			*
	Two			*		
Start time	Btwo(1)				*	*
	Tmax(1)			*		
	Tmax(2)	*	*	*	*	*
	Tmax(3)		*		*	*
	Iact			*	*	
	Tmax(4)		*			*
	Atype			*		

Table 4.9: Description of the most important attributes for each approach for the remaining four dimensions

Choice facet	Attribute	One R	Naïve Bayes	CHAID	Full approach	FS approach
Trip chain	yIBstop			*	*	*
	yAstop	*	*	*	*	*
	yBstop		*	*	*	*
	Optime				*	*
	Onwith			*		
Other mode	Awith1		*		*	*
	Rcabi				*	*
	Gend			*	*	*
	TTbike				*	*
	Hwork1	*				
	Csec		*			
	Ccomp		*	*		
	Avcar			*		
	Two			*		
Location 1	AvCmin			*	*	*
	AvCext10				*	*
	Mode	*	*	*	*	*
	AvCext20				*	
	Atype		*	*		*
	AvCext5			*		
Location 2	AvCext5		*		*	*
	Mode	*	*	*	*	*
	Atype			*		
	AvCext10			*		
	AvCext20				*	
	AvCmax		*		*	*
	Nout			*		*

As for the *With-whom* facet, the attributes that play a role in the choice whether the activity is performed alone or with others (a three-level variable), are more general (household) characteristics. It seems reasonable that the composition of the household, the presence of children, the day of the week and the activity type attributes play a prominent role in building the trees. The three variables in the feature selection approach all re-appear in the 'full' approach, which needs nineteen features to build its complete model. Again, the variables in the last two approaches do not match because of the possible high correlations between the most important variables in the 'full' approach and those of the feature selection approach, e.g. $\rho(\text{Ccomp, yAvail}^4) = 0.67$.

The next choice facet is *Duration.* Strangely enough the travel party and the activity type are the leading attributes in the most approaches, while also time concerning variables play a role in e.g. CHAID and the 'full' approach. Three variables make up the Naïve Bayes classifier, four attributes are necessary to build the feature selection model, while twenty-eight were needed in the C4.5 approach and even forty-one in the CHAID.

Slightly different results were obtained for *Start time.* In both approaches time concerning features were fundamental. The 'full' approach needed twenty-eight variables to build the model, CHAID even sixty-three, while for the feature selection approach only thirteen were necessary. In addition to the differences shown in the table, two relevant variables in the FS approach did not come up in the 'full' approach, i.e. the total time of work 1 including travel and the total time of work 1 and work 2, where work 1 stands for the primary work or school activity and work 2 denotes a voluntary work activity. Regarding the four most important variables, all approaches coincide largely with each other. The variables included in these trees can be regarded as being robust for the prediction of the start time dimension.

The next choice facet is *Trip chain.* Variables indicating whether there was time enough to include the activity in the corresponding place in the schedule are noticed to be valuable in building the tree and this accounts for all approaches. These attributes can be regarded as being robust for predicting the 'trip chain' dimension. Only two additional variables in each approach were needed to build the full models. For the feature selection approach, these variables described the number of mandatory out-of-home activities other than the work activity and whether there is travel party available in the schedule before activity X (this latter variable is also important in the CHAID approach), while in the 'full' approach the two extra variables denoted whether the first activity is a grocery activity and whether there is a bring/get activity at all in the total schedule.

For the simple classifiers, household characteristics appear to be the most relevant variables, while features measuring the travel times by bike and by car, as well as the travel party and the gender were used to predict the *mode for other than work trips* in the 'full' and the FS approach. Regarding the transport modes chosen, we come to the same conclusions as with the mode for work choice facet. The day of the week, the socio-economic status of the household and the total time of work 1 and work 2 in the schedule appear to be valuable features that are not incorporated in the 'full' model induction tree. These variables are rather highly correlated with variables in the 'full' approach e.g. $\rho(\text{Csec, Ncar}) = -0.51$.

Almost the same four crucial variables occurred in the decision tree approaches to predict the *Location 1* dimension and most of them were related to the feasibility of the location-selection heuristic given the schedule. Apart from these time related variables, also the activity type and the transport mode were prominent variables. All variables that appeared to be important in the feature selection approach, were also found in the 'full' approach, together with 4 other features. These variables are reasonably robust in predicting the location.

Finally, for the *Location 2* facet, also time related variables stayed the most important ones. The transport mode, the activity type and the number of out-of-home activities also play an important role in building the CHAID classifier. Six variables were necessary to model the feature selection approach, while 15 (among which the previous six) were needed in the full approach.

In summary, at the choice facet level, the methods do not differ dramatically in their predictive performance. The variables selected as being most important for some choice facets do not differ that much. However, a difference can be discerned in some other trees. These differences can then often be explained by high correlations between variables.

4.3.4 Activity Pattern Level

The performance of the two model approaches at the activity pattern level was assessed by comparing observed and predicted sequences of activities. Several sequence alignment methods (SAM) were used to measure the goodness of fit. SAM measures per dimension are added as indicators of the performance of the models on each dimension separately. In general, the lower these measures are, the less effort was needed to equalise the observed and the predicted sequences, and thus the better the prediction is. Note that SAM measures represent the costs of alignment of flexible elements of patterns only. In order to measure the dissimilarity between the observed and the predicted schedules, the ‘full’ approach, e.g., will require more deletion, while the FS approach will demand for more insertion. Since both have the same cost, this normally would result in not too much difference between the SAM measures, unless the predicted activities deviate heavily from the observed ones. Table 4.10 indicates the results on the training set.

Table 4.10: Model performance on training data: activity pattern level

Measure	Mean Distance (CHAID)	Mean Distance (Zero R)	Mean Distance (One R)	Mean Distance (Naïve Bayes)	Mean Distance (Full)	Mean Distance (FS)
SAM (activity type)	2.878	3.108	3.047	2.963	2.929	2.962
SAM (with)	3.209	3.500	3.363	3.246	3.205	3.189
SAM (location)	3.238	3.240	3.200	3.092	3.188	3.074
SAM (mode)	4.626	4.986	4.874	4.809	4.706	4.558
UDSAM	16.829	17.943	17.531	17.074	16.957	16.746
MDSAM	8.497	8.883	8.732	8.581	8.558	8.340

We can observe that overall the Zero R approach resulted in predicted sequences that were the furthest away from the observed sequences in any way, as could be expected. The One R approach has the second highest values on the SAM measures. The Naïve Bayes approach has a lower value than the CHAID and the ‘full’ approach for the uni-dimensional SAM for the location facet, though for all other measures the CHAID and the ‘full’ approach provide lower values. These two approaches are again comparable in performance. In general, the feature selection approach seems to predict the sequences better than the ‘full’ approach, except for the uni-dimensional SAM for the activity type facet.

Let us now consider one column in specific, e.g. the feature selection column. The average alignment cost according to the the MDSAM is approximately 8.3 units on the training set for the Albatross model. This is approximately 50% of the weighted sum of alignment costs across dimensions (UDSAM). This points at the fact that (since the MDSAM measure allows for joint operations to be performed) a reasonable amount of association between elements across the dimensions exists.

As the first three measures indicate, the alignment costs per dimension vary between 2.9 and 3.5 implying that, on average, the efforts of 1 substitutions and/or 2 insertion/deletion operations suffice to make two strings identical (on that dimension). The mode dimension is an exception in the sense that for this dimension the alignment costs are relatively high.

The results on the test data (see Table 4.11) indicate in general even lower SAM-measures when compared to the training data, especially for the One R, the CHAID and the 'full' approach. While the feature selection approach seems to outperform the other methods on the training data, the CHAID approach performs better on the test data, although the difference with the FS approach is small. Overall, the relative performance of the models seems comparable on the training and on the test set.

Table 4.11: Model performance on test data: activity pattern level

Measure	Mean Distance (CHAID)	Mean Distance (Zero R)	Mean Distance (One R)	Mean Distance (Naïve Bayes)	Mean Distance (Full)	Mean Distance (FS)
SAM (activity type)	2.777	3.130	3.027	3.022	2.903	2.929
SAM (with)	3.168	3.464	3.312	3.225	3.210	3.208
SAM (location)	3.127	3.251	3.184	3.107	3.166	3.033
SAM (mode)	4.626	5.018	4.592	4.781	4.497	4.600
UDSAM	16.475	17.993	17.142	17.156	16.678	16.699
MDSAM	8.333	8.951	8.474	8.671	8.374	8.373

This again confirms the our primary belief that people make their decisions on only a few simple heuristics instead of on a complex set.

4.3.5 Trip Matrix Level

At trip matrix level, we compare the number of trips made from a certain origin to a certain destination. Correlations were calculated between observed and predicted matrix entries in general and for trip matrices that are disaggregated each time in a different way based on some selected trip facets. The facets considered include transport mode, day-of-the-week and activity (purpose). The variation of the correlation coefficient within the columns can be largely explained by the variation in the number of cells between matrices. Recall that the general OD matrix has 400 cells, the OD matrix by day 1200, by mode 2000 and finally by activity 4400. As could be expected, the fit decreases with an increasing number of cells, i.e. the level of disaggregation of interactions.

In Table 4.12 the performance of the six different models on the training data set is given, while Table 4.13 illustrates the performance on the test data set. Both tables indicate that all correlation coefficients are similar. In the general case, the

Table 4.12: Model performance on training data: trip matrix level

Matrix	ρ(o, p) (CHAID)	ρ(o, p) (Zero R)	ρ(o, p) (One R)	ρ(o, p) (Naïve Bayes)	ρ(o, p) (Full)	ρ(o, p) (FS)
None	0.956	0.938	0.936	0.949	0.962	0.957
Mode	0.887	0.841	0.880	0.874	0.885	0.887
Day	0.959	0.943	0.939	0.953	0.959	0.956
Primary activity	0.892	0.806	0.834	0.867	0.899	0.883

‘full’ approach provides the highest correlation, although the difference with the FS approach does not exceed the 1% level. The highest correlation coefficient when the disaggregation is on transport mode is given by the CHAID and the FS approach. In the case of origin and destinations matrices with a difference made by day and by primary activity, the ‘full’ approach performs a little bit better than the FS approach. The test set is the most relevant data set for comparison of the models, therefore, we will focus on this latter one.

Table 4.13: Model performance on test data: trip matrix level

Matrix	ρ(o, p) (CHAID)	ρ(o, p) (Zero R)	ρ(o, p) (One R)	ρ(o, p) (Naïve Bayes)	ρ(o, p) (Full)	ρ(o, p) (FS)
None	0.937	0.925	0.928	0.917	0.942	0.947
Mode	0.836	0.787	0.862	0.842	0.856	0.849
Day	0.944	0.925	0.937	0.919	0.950	0.946
Primary activity	0.830	0.766	0.801	0.800	0.861	0.840

Table 4.14 shows that, for the disaggregation on day, the number of entries (trips made from a certain origin to a particular destination) made at weekdays tends to be overestimated by all methods, the number of trips predicted by the feature selection method agrees best with the observed number. The Naïve Bayes method overestimates the number of trips made on a Saturday, but underestimates them on Sunday. CHAID overestimates the trips undertaken at weekend days, while all other methods underestimate them. By taking a look at the disaggregation on primary activity, one can observe that the number of trips undertaken for a medical visit, a bring/get activity, a non-leisure or a leisure out-of-home activity is underestimated, while the number of service trips, social visits and other ‘out-of-home’ activities happens to be overestimated by all approaches. The number of work trips (out-of-home) is un-

Table 4.14: Number of trips at trip matrix level: test set in detail

Matrix	Observed	Predicted					
		Zero R	One R	Naïve Bayes	CHAID	C4.5 ('full')	feature selection
Day							
Weekday	2359	2572	2510	2414	2454	2564	2413
Saturday	356	276	277	380	416	331	262
Sunday	287	203	248	172	297	214	157
Primary activity							
Work out	970	901	937	944	960	942	971
Medical visit	44	32	34	30	36	36	38
Bring/get	538	485	497	508	526	524	502
Non-leisure out	106	83	89	91	94	89	85
Non-grocery	251	287	204	320	231	260	220
Grocery	319	281	316	321	398	399	264
Leisure out	466	329	440	264	447	371	269
Social visit out	241	353	378	304	352	331	280
Service	59	243	95	133	96	121	168
Other out	18	63	52	61	36	48	46
Transport mode							
Car	1609	1580	1609	1465	1771	1573	1466
Slow	814	1020	1013	1031	999	1038	920
Public	79	83	81	102	83	113	107
Car passenger	294	356	321	357	305	375	333

derestimated by all but the feature selection approach. The number of non-grocery shopping trips is overestimated by the Zero R, Naïve Bayes and the 'full' approach and underestimated by the others, and finally, the number of trips undertaken for grocery shopping is overestimated by the Naïve Bayes, CHAID and 'full' C4.5 approach, while underestimated by the other analyses. What the different transport modes are concerned, one notices that the number of trips undertaken as a car driver is correctly predicted by the One R approach, overestimated by the CHAID approach and underestimated by the remaining approaches, while the use of any other transport mode appears to be overestimated.

The relative performance can be indicated by a ratio of the fit scores of Table 4.13 between the three simple models and CHAID and we also compared the 'full' approach to the feature selection approach. As it appears, the ratios for the simple models range from 92.3 to 103% on the test set, those of the FS approach from 97.6 % to 100.5 % dependent on the OD-matrix. When test scores and training scores are compared, we can even see an increase in performance for the One R model. The Zero R model shows an increase in performance on the general OD matrix and on the OD matrix aggregated on primary activity. The Nïve Bayes model shows a little decline in relative performance on all but the transport mode aspect. Compared to the 'full' approach, the ratio of the fit scores of the FS approach in the general case is even higher on the test set, when compared to the training set.

The stability of the performance can be expressed as a ratio between test set and training set scores. For all models, the ratios range between 92.27-99.79% depending on the matrix. Therefore, we conclude that the extent to which over-fitting has occurred is approximately the same and at an acceptable level for all models. We would expect the performance of the test set to be worse, though this does not seem to be true.

4.4 Conclusion

In the last decade, computational process models that predict travel behaviour based on activity diary data have been suggested in the literature. These models usually perform very well, though, very often, they are based on a very complex set of rules. Moreover, research in the field of psychology has learned us that simple models often predict human behaviour very well. In fact, the call for simplicity is a question of all ages. Occam's razor, that has to be situated already in the Middle Ages, being an important example.

In addition, one has to be careful in interpreting these previous studies, they only support the proposition 'Simplicity is a goal in itself', not that simplicity would lead to greater accuracy or better models. It is in this light that this chapter should be regarded. We regarded two ways of simplifying the complex set of rules used to determine the Albatross system. On the one hand, we used simple classifiers to predict the nine dimensions, while on the other hand we performed two similar analyses: one with and one without irrelevant variables, while in the second analyses, at same time we cut back in the number of variables. The results of the tree-induction algorithms can namely be heavily influenced by the inclusion of irrelevant attributes. On the one hand, this may lead to over-fitting, while one the other hand, it is not evident whether the inclusion of irrelevant attributes would lead to a substantial loss in accuracy and/or predictive performance. The aim of the study reported in this chapter therefore was to further explore this issue in the context of the Albatross model system, currently the most comprehensive operational computational, rule-based process model of travel demand.

The results of the simple classifiers do indicate that the 'simpler' models do not perform better, but, on the other hand, it is also not the case that they are inferior to the complex CHAID approach. It is rather logical that the model that always takes the majority class (Zero R) does not perform that well, conversely, the models that make up their decisions based on one or a few variables are not in any case second to the complex analysis. This comes as a welcome bonus.

The results of the analyses conducted at the three different levels of performance, indicate that, also in the second way of simplification, the simpler models do not necessary perform worse. In fact, more or less the same results were obtained at the activity pattern level and at the trip matrix level. At the choice facet level, one can observe that a strong reduction in the size of the trees as well as in the number of predictors is possible without adversely affecting predictive performance too much. Thus, at least in this study, there is no evidence of substantial loss in predictive power in the sequential use of decision trees to predict activity-travel patterns.

The results indicate that using feature selection in a step prior to tree induction can improve the performance of the resulting model. It should be noted, however, that predictive performance and simplicity are not the only criteria. The most important criterion is that the model needs to be responsive to policy sensitive attributes and for that reason policy sensitive attributes, such as for example service level of the transport system, should have a high priority in the selection of attributes if the model is to be used for predicting the impact of policies. The feature selection method allows one to identify and next eliminate correlated factors that prevent the selection of the attributes of interest during the construction of the tree, so that the resulting model will be more robust to policy measures.

Similarly, the results of a trimmed decision tree should be assessed in terms of behavioural mechanisms. On the one hand, if one has strong theoretical reasons for including particular conditions, they should be kept in the decision tree.

These findings endorse the primary belief that people rely for their choices on some simple heuristics. In reality, one is limited in both knowledge and time and it is infeasible to go over all the different possibilities and then trying to make an optimal choice. Since, in the Albatross system, we are trying to predict nine different choices on travel behaviour made by human beings, this might give an idea on why these simple models do not necessarily perform worse than the complex models. In fact, this is not totally true. If simple models are able to predict the choices of a human being, this can mean two things: either the environment itself is perceived as simple, or the complex choice process can be described by simple models. Since activity-based transport modelers keep developing systems with an increasing complexity in order to try to understand the travel behaviour undertaken by humans, we acknowledge that the environment is not simple. However, whether it is perceived as simple by human beings, remains an open question.

Chapter 5

Bagging and Boosting within an Activity-Based Model

5.1 Introduction

In the previous chapters, we tried to obtain simple models for the nine choice facets that resulted in a fairly good prediction of the Albatross model. Two different ways of simplification were considered: on the one hand, simple classifiers were used, while on the other hand a more parsimonious version of a complex model has been determined by first applying a variable selection technique and then a tree induction algorithm. As could be observed, this latter method resulted in a reasonably good fit of the model. Therefore, in this chapter, we will try to improve the feature selection models from Chapter 4 by means of bagging and boosting techniques. Wickramaratna *et al.* (2001) and Maclin and Opitz (1997), however, suggest to use bagging and boosting especially with weak classifiers, so therefore, both techniques will also be applied to the One R models of the previous chapter.

Bagging and boosting are both very powerful learning ideas introduced in the last decade (see, e.g. Breiman, 1996, 1998). They are developed in order to improve the prediction. Both methods combine the output of several classifiers to produce an accurate prediction. Bagging does so by re-sampling the training set and averaging the result of the classifier on each of these bagged samples. In this way, bagging reduces the variance of the prediction and it improves its stability. Boosting, on the other hand, combines the predictions acquired on repeatedly modified versions of the data through a weighted majority vote to produce the final prediction. More details can be found in the next section.

5.2 Methods

5.2.1 Bagging

Before treating bagging more into detail, we will make a side jump to bootstrapping, since this is a very important aspect of bagging.

The *bootstrap* is a general tool for assessing statistical accuracy (Efron, 1979). The basic idea is to randomly draw data sets with replacement from the training data, each sample the same size as the original training set. This is done B times, producing B bootstrap data sets. Then we refit the model to each of the bootstrap data sets and examine the behaviour of the fits over the B replications.

Suppose that the tree induction algorithm leads to the prediction $\hat{f}(x)$ at input x. Bootstrap aggregation or *bagging* (Breiman, 1996) averages this prediction over a collection of bootstrap samples, thereby reducing its variance and improving the stability of the prediction. For each bootstrap sample, we fit our model, giving prediction $\hat{f}^{\star b}(x)$. The bagging estimate is then defined by

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{\star b}(x). \tag{5.1}$$

Each bootstrap tree will typically involve other features than the original, and might have a different number of terminal nodes. The bagged estimate is now the average prediction at x from these B trees. Actually, a tree produces a classifier $\hat{G}(x)$ for a d-class response. Hence it is useful to consider an underlying indicator-variable $\hat{f}(x)$, with value one and $d-1$ zeroes, such that $\hat{G}(x) = \arg\max_i \hat{f}(x)$. Then the bagged estimate $\hat{f}_{bag}(x)$ (5.1) is a d-vector $(p_1, p_2, \ldots, p_d)$, with p_i equal to the proportion of trees predicting class i at x. Treating these as estimates of the class probabilities, our predicted class is the one with the most 'votes' from the B trees, $\hat{G}_{bag}(x) = \arg\max_i \hat{f}_{bag}(x)$.

5.2.2 Boosting

In this section, we will describe the most popular *boosting* algorithm due to Freund and Schapire (1997), called 'AdaBoost'. Following Hastie *et al.* (2001), consider a d-class output variable y, a vector of predictor variables x and a classifier $G(x)$ that produces a prediction taking values $\{1, \ldots, d\}$. The error rate on the training sample is

$$\overline{err} = \frac{1}{N}\sum_{i=1}^{N} I(y_i \neq G(x_i)).$$

The purpose of boosting is to sequentially apply the classification algorithm to repeatedly modified versions of the data, thereby producing a sequence of classifiers $G_m(x), m = 1, \ldots, M$. The predictions of all these classifiers are then combined through a weighted majority vote to produce the final prediction:

$$G(x) = \arg\max_{y \in Y} \sum_{m=1}^{M} \alpha_m I(y \neq G_m(x)), \tag{5.2}$$

where Y denotes the set of possible outcomes for the response variable. Here $\alpha_1, \ldots, \alpha_M$ are computed by the boosting algorithm, they weigh the contribution of each respective $G_m(x)$. Their effect is to give higher influence to more accurate classifiers in the sequence. The data modifications at each boosting step consist of applying weights $w_1, \ldots, w_N$ to each of the training observations $(x_i, y_i), i = 1, 2, \ldots, N$. Initially, all the weights are set equal to $w_i = \frac{1}{N}$, so that the first step simply trains the classifier on the data in the usual manner. For each successive iteration $m = 2, 3, \ldots, M$, the observation weights are individually modified and the classification algorithm is re-applied to the weighted observations. At step m, those observations that were misclassified by the classifier $G_{m-1}(x)$ induced at the previous step have their weights increased by a factor $\exp(\alpha_m)$, whereas the weights are decreased for those that were classified correctly. So, as the iterations proceed, observations that are difficult to classify correctly receive ever-increasing influence. Each successive classifier is hereby forced to concentrate on those training observations that are missed by previous ones in the sequence. The algorithm is shown in detail below:

```
1. Initialize the observation weights $w_i = \frac{1}{N}, i = 1, \ldots, N$
2. For m = 1, ..., M:
    (a) Fit a classifier $G_m(x)$ to the training data using weights $w_i$.
    (b) Compute $err_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i}$
    (c) Compute $\alpha_m = \log\left[\frac{1-err_m}{err_m}\right]$
    (d) Set $w_i \leftarrow w_i \cdot \exp[\alpha_m \cdot I(y_i \neq G_m(x_i))], i = 1, \ldots, N$.
3. Output $G(x) = \arg\max \sum_{m=1}^{M} \alpha_m I(y \neq G_m(x))$
```

Since its introduction, much has been written to explain the success of AdaBoost in producing accurate classifiers. Most of this work was centred on the use of classification trees as base learner $G(x)$, where improvements are often dramatic. In fact, Breiman (1998) referred to AdaBoost with trees as the ‘best off-the-shelf classifier in the world’.

Both techniques have been used extensively in the literature, especially in statistics, learning theory and in artificial intelligence (see, e.g. Nock and Lefaucheur, 2002; Friedman *et al.*, 2000; Oza and Russell, 2001).

5.3 Analysis and Results

In this chapter we want to investigate whether an improvement of a simple model by means of bagging and boosting will lead to a better performance of the Albatross model. The nine dimensions are represented by the decision tables derived from the combination of trees with AdaBoost and trees combined with bagging. Under ‘trees’, we understand both the One R and the feature selection models. We will investigate the effect of bagging and boosting for each level of comparison. In order to compare performances, the results of the One R and the feature selection approach in Chapter 4 are added for each analysis.

Note that for the bagging algorithm, we use 50 bootstrap samples, so B equals fifty. The bootstrap samples are determined as follows: at first, we considered all training cases and then selected only these variables that were used in the feature selection approach. This reduced data set was then used for bootstrapping. Probably, different results would be obtained if the selection of variables was carried out after bootstrapping was applied to the whole training data set. However, the number of rules in the decision tables could then have become very large, consider e.g. the

'mode 2' dimension where eleven variables can be chosen. Nevertheless, this can be an interesting topic for further research. In this way, we could also determine if the selected variables are robust in predicting the dimensions.
For the boosting algorithm, the number of iterations (M) is set equal to 10 in the analyses.

At first we will take a look at some descriptive statistics, while in the following subsections the performance of the model will be investigated subsequently at choice facet level, activity pattern level and at trip matrix level.

Table 5.1: Average number of activities in the predicted sequences (standard deviation between brackets)

	Training Set	Test Set
	Predicted	Predicted
One R-Bagging	4.769 (2.956)	4.782 (3.029)
One R-Boosting	4.715 (2.971)	4.712 (2.885)
One R	5.198 (3.182)	5.178 (3.128)
FS-Bagging	5.064 (3.205)	5.092 (3.147)
FS-Boosting	5.037 (3.230)	4.975 (3.163)
Feature selection	5.014 (3.033)	4.907 (2.921)

Some descriptive statistics of the predicted patterns are shown in Table 5.1. As in Chapter 4, the average number of activities in the observed patterns equals 5.160 for the training set and 5.115 for the test set, with a standard deviation of 2.807 and 2.709, respectively. This average length offers room for 1-3 flexible activities complemented with 2-4 in-home activities. Considerable variation occurs, however, as indicated by the standard deviation of approximately 3 activities. On average, bagging and boosting seem to underestimate the number of activities, even though the application of One R itself causes an overestimation. Bagging and boosting on the FS approach clearly improves the predicted number of activities both on training and on test set. Both techniques overestimate the variance a little, but here bagging and boosting applied to the One R models seem to give a better prediction. Wickramaratna *et al.* (2001) and Maclin and Opitz (1997) state that bagging and boosting should primarily be used with weak classifiers. The combination of weak classifiers tends to ameliorate the performance, while the combination of rather powerful classifiers (such as tree induction algorithms) may end up in providing worse results. To investigate whether we can subscribe to their viewpoint, the performance of bagging and boosting on the three levels will be investigated.

5.3.1 Choice Facet Level

Tables 5.2 and 5.3 yield the results at choice facet level. As can be observed, the number of attributes used and the number of leafs in the feature selection bagging and boosting analyses are the same. This is because these bagging and boosting models do not result in one particular tree, but they only provide a prediction for each training case. In order to obtain rules in the decision table format, we considered all training cases and regarded only at the variables that were selected after feature selection. Then for each combination of categories amongst these selected variables, we averaged the prediction outcome, which lead us to a probabilistic statement for each rule. E.g., consider the ‘Mode for work’ choice facet. After the feature selection procedure, two relevant variables remain: Tbike and Rpubi. Tbike and Rpubi both have four categories, this leads to $4 \times 4 = 16$ rules. For each of these combination of categories i and j ($i, j \in \{1, \ldots 4\}$), we select the number of training cases ($= f_{ij}$). The dependent variable for this choice facet can take four different categories. Thus, for each rule we can determine the number of cases that have a certain outcome predicted ($= f_{ijk}$) with $k \in \{1, \ldots, 4\}$. The ratio of these frequencies leads to probabilities for each possible outcome on each rule. In this way, we can still derive decision tables, so that the Albatross system can be used.

As can be observed, there is not much difference in the results of bagging and boosting. The probability of correctly predicting the outcome variable is higher using the boosting learning method compared to bagging in the ‘Mode for work’, the ‘Activity selection’, ‘Trip Chaining’ and the ‘Location 1’ dimension, in the other dimensions bagging outperforms boosting in the One R models. For the feature selection models, boosting shows a slightly better performance than bagging on three dimensions: the ‘Mode for work’, ‘Activity selection’ and ‘Location 2’, but both methods clearly perform better than the original analysis. The increase in performance as a ratio of the maximum increase that is possible given a null model (see Chapter 4) is as a consequence also larger (or equal) in these dimensions for the boosting method. Note

Table 5.2: Performance at choice facet level (bagging & boosting on One R models)

Dimension	♯ alts	♯ attrs	♯ leafs	e	e_{ratio}
Bagging					
Mode for work	4	2	13	0.605	0.169
Selection	2	1	1	0.678	0.026
With-whom	3	2	35	0.430	0.117
Duration	3	4	255	0.418	0.126
Start time	6	3	63	0.318	0.176
Trip chain	4	1	2	0.699	0.354
Mode other	4	7	1175	0.724	0.549
Location 1	7	1	3	0.435	0.096
Location 2	6	3	16	0.296	0.120
Boosting					
Mode for work	4	3	54	0.640	0.243
Selection	2	1	1	0.734	0.196
With-whom	3	1	5	0.408	0.082
Duration	3	1	3	0.348	0.020
Start time	6	1	4	0.227	0.067
Trip chain	4	4	14	0.807	0.586
Mode other	4	1	4	0.413	0.040
Location 1	7	3	43	0.501	0.202
Location 2	6	1	3	0.234	0.043
One R					
Mode for work	4	1	6	0.595	0.147
Selection	2	1	5	0.677	0.025
With-whom	3	1	5	0.408	0.082
Duration	3	1	3	0.348	0.020
Start time	6	1	4	0.227	0.067
Trip chain	4	1	2	0.699	0.354
Mode other	4	1	4	0.413	0.040
Location 1	7	1	3	0.435	0.096
Location 2	6	1	3	0.234	0.043

that the null model for the bagging may deviate somewhat compared to all the other analyses, since the magnitude of the training set for bagging equals fifty times the original training set (sampling with replacement). For comparison purposes, we have added Tables 4.3 and 4.7 from Chapter 4, these denote the results without bagging and boosting.

Table 5.3: Performance at choice facet level (bagging & boosting on FS models)

Dimension	♯ alts	♯ attrs	♯ leafs	$e_{bagg.}$	$e_{ratio,bagg.}$	$e_{boost.}$	$e_{ratio,boost.}$
Mode for work	4	2	16	0.611	0.188	0.614	0.188
Selection	2	1	1	0.672	0.011	0.673	0.011
With-whom	3	4	415	0.566	0.327	0.564	0.324
Duration	3	4	359	0.452	0.177	0.451	0.175
Start time	6	8	1375	0.704	0.642	0.703	0.641
Trip chain	4	10	512	0.896	0.777	0.894	0.773
Mode other	4	11	2026	0.931	0.887	0.930	0.886
Location 1	7	6	83	0.564	0.301	0.561	0.299
Location 2	6	8	532	0.708	0.635	0.710	0.638

Table 5.4: Performance at choice facet level (‘feature selection approach’)

Dimension	♯ alts	♯ attrs	♯ leafs	e	e_{ratio}
Mode for work	4	2	6	0.595	0.147
Selection	2	1	1	0.669	0.000
With-whom	3	4	51	0.467	0.173
Duration	3	4	38	0.368	0.051
Start time	6	8	1	0.172	0.000
Trip chain	4	10	13	0.811	0.596
Mode other	4	11	60	0.508	0.196
Location 1	7	6	15	0.513	0.222
Location 2	6	8	14	0.312	0.141

We can conclude that there is not much difference between bagging and boosting on the choice facet level, though both approaches clearly perform better than the One R and the feature selection procedures on each of these dimensions separately.

5.3.2 Activity Pattern Level

In this subsection, we investigate whether the bagging and boosting models are successful in predicting the sequences of activity choices.

Table 5.5: Model performance on training data: activity pattern level

Measure	Mean Distance (Bagging)	Mean Distance (Boosting)	Mean Distance (approach without)
	One R		
SAM (activity type)	2.783	2.773	3.047
SAM (with)	3.126	3.114	3.363
SAM (location)	2.882	2.789	3.200
SAM (mode)	4.452	4.352	4.874
UDSAM	16.027	15.802	17.531
MDSAM	7.993	7.898	8.732
	Feature selection		
SAM (activity type)	2.975	2.980	2.962
SAM (with)	3.217	3.258	3.189
SAM (location)	3.113	3.115	3.074
SAM (mode)	4.454	4.464	4.558
UDSAM	16.735	16.798	16.746
MDSAM	8.293	8.292	8.340

In Table 5.5 one can observe that the SAM measures on the One R boosting (and bagging) models are lower than those of the FS bagging and boosting models, despite the fact that the predicted number of activities was better in these latter cases. Apart from the uni-dimensional SAM measure disaggregated on activity type, for which the CHAID approach showed the best results, all other SAM measures on the training data were lowest for the regular feature selection approach in Chapter 4. If one combines the results of Chapter 4 with the results of the analyses performed here, the best results overall are clearly shown by the One R boosting model. For the feature selection models, bagging apparently led to predicted sequences that were closer to the the observed sequences when compared to boosting, while for the One R models the opposite was true.

Consider now e.g. the One R boosting column more in detail: the average alignment cost according to the the MDSAM is approximately 7.9 units on the training set for the Albatross model. Their ratio indicates that this figure equals approximately

Table 5.6: Model performance on test data: activity pattern level

Measure	Mean Distance (Bagging)	Mean Distance (Boosting)	Mean Distance (approach without)
	One R		
SAM (activity type)	2.821	2.805	3.027
SAM (with)	3.204	3.088	3.312
SAM (location)	2.889	2.844	3.184
SAM (mode)	4.347	4.189	4.592
UDSAM	16.081	15.732	17.142
MDSAM	7.944	7.782	8.474
	Feature selection		
SAM (activity type)	2.988	3.022	2.929
SAM (with)	3.270	3.342	3.208
SAM (location)	3.158	3.096	3.033
SAM (mode)	4.588	4.540	4.600
UDSAM	16.990	17.022	16.699
MDSAM	8.471	8.356	8.373

49.98% of the weighted sum of alignment costs across dimensions (UDSAM). This means that a considerable degree of association between elements across dimensions exists.

As the first four measures of bagging and boosting show, the alignment costs per dimension vary between 2.8 and 4.4 implying that, on average, the efforts of 1-2 substitutions and/or 1-3 insertion/deletion operations are sufficient to make the observed and the predicted strings identical.

Table 5.6 learns that the SAM-measures on the test set are a little bit higher when compared to the training data. Though, overall, the relative performance of both models seems comparable on the training and on the test set. Again the One R boosting model performs best, and if the uni-dimensional SAM disaggregated on activity type is disregarded for a moment, this best performance is across both Chapter 4 and 5. This particular uni-dimensional SAM measure is lowest for the CHAID approach. This clearly depicts that although the performance on each dimension separately does not show the highest accuracy value on the One R boosting analysis, the sequential execution of each of these dimensions can be best in predicting the sequences of activities. Thus, a best performance on dimensions separately does not need to lead to a best performance when they are executed together and visa versa. Now, we will investigate whether the same is true at trip matrix level.

5.3.3 Trip Matrix Level

At trip matrix level, the observed and predicted origin-destination (OD) matrices are compared via a correlation coefficient. This coefficient measures the relation between the observed and predicted number of trips. Recall that the variation of the correlation coefficient within the columns is largely explained by the variation in the number of cells between matrices. As one might expect, the fit decreases with an increasing number of cells, i.e. the level of disaggregation of interactions. Table 5.7 illustrates the performance of the four different models on the training data set. With the exception of the One R bagging correlation coefficients disaggregated on day and on primary activity, bagging and boosting do not seem to outperform the original One R or feature selection approaches. For the training data, the regular feature selection approach gives the best results among these six analyses. One can observe that for the feature selection approach the deviation from the original values becomes larger with an increasing level of disaggregation. Though, in general, one can state that the results do not differ very much from the original values.

Table 5.7: Model performance on training data: trip matrix level

Matrix	ρ(o, p) (Bagging)	ρ(o, p) (Boosting)	ρ(o, p) (approach without)
		One R	
None	0.935	0.935	0.936
Mode	0.876	0.872	0.880
Day	0.950	0.937	0.939
Primary activity	0.837	0.834	0.834
		Feature selection	
None	0.952	0.955	0.957
Mode	0.885	0.879	0.887
Day	0.952	0.955	0.956
Primary activity	0.863	0.846	0.883

In Table 5.8 the performance on the test data set is provided. This test set is the most relevant data set for comparison of the models.

Table 5.8: Model performance on test data: trip matrix level

Matrix	ρ(o, p) (Bagging)	ρ(o, p) (Boosting)	ρ(o, p) (approach without)
	One R		
None	0.925	0.930	0.928
Mode	0.877	0.883	0.862
Day	0.927	0.929	0.937
Primary activity	0.807	0.803	0.801
	Feature selection		
None	0.951	0.950	0.947
Mode	0.863	0.861	0.849
Day	0.948	0.950	0.946
Primary activity	0.818	0.798	0.840

So, if we take a more detailed look at the test set, Table 5.9 learns us that the number of trips conducted at weekdays is underestimated by bagging and boosting in the One R approach, while bagging and boosting in the feature selection approach overestimates this number. The trips undertaken at weekend days are underestimated by all analyses. Considering the disaggregation on primary activity, one can observe that the number of trips undertaken for a medical visit, a bring/get activity, a non-leisure or a leisure out-of-home activity is underestimated, while the number of service trips and social visits or other out-of-home activities is overestimated. The number of work trips (out-of-home) happens to be overestimated by the feature selection bagging approach and underestimated by the other three. The number of trips undertaken for non-grocery shopping is underestimated by the One R analyses and overestimated by the FS analyses, and finally, the number of grocery shopping trips is overestimated by the One R boosting approach and underestimated by the other analyses. If one finally takes a closer look at the different transport modes, one can observe that the number of trips undertaken as a car driver or car passenger are overestimated by the FS approaches and underestimated by the One R analyses. The number of trips undertaken by bike or on foot is overall overestimated, while the number of trips by public transport is seriously underestimated.

Table 5.9: Number of trips at trip matrix level: test set in detail

Matrix	Observed	Predicted			
		One R Bagging	One R Boosting	FS Bagging	FS Boosting
Day					
Weekday	2359	2356	2318	2507	2498
Saturday	356	208	251	278	252
Sunday	287	171	124	190	137
Primary activity					
Work out	970	924	890	994	955
Medical visit	44	32	31	31	37
Bring/get	538	501	492	493	497
Non-leisure out	106	88	84	86	87
Non-grocery	251	243	149	279	259
Grocery	319	257	327	243	213
Leisure out	466	284	278	314	318
Social visit out	241	255	261	302	306
Service	59	111	125	186	169
Other out	18	58	65	51	52
Transport mode					
Car	1609	1446	1522	1643	1688
Slow	814	1182	1141	989	848
Public	79	9	4	20	18
Car passenger	294	100	2	309	316

In general, bagging and boosting on the analyses conducted in the previous chapter apparently do improve the results at trip matrix level, especially for the feature selection approach. If these results are compared to those of Chapter 4, one can observe that the highest general correlation coefficient of the regular feature selection approach is improved by applying the techniques of bagging and boosting. The correlation coefficient when disaggregated for transport mode was highest for the regular One R approach, it has been improved here by both bagging and boosting and by bagging the feature selection models. The highest correlation coefficient when the OD matrices are disaggregated on day was provided by the 'full' C4.5 model, and the same coefficient was obtained by boosting the feature selection technique. Only the correlation coefficient on primary activity could not be improved. Again this shows that although the performance on each dimension separately does not show the highest accuracy value of all analyses, the sequential execution of each of these dimensions can be best in predicting the number of trips made from a particular origin to a certain destination. Hence, as a consequence, in this case, one can conclude that a combination of simple or relatively simple models can provide better decision rules and hence perform better than the more complex models.

The relative performance is measured by the ratio of the fit scores between the bagging and boosting models on the one hand and the One R and the feature selection approach (see Chapter 4) on the other hand. It appears that the ratios range from 98.93 to 101.74% on the test set for One R bagging, from 99.15% to 102.44% for One R boosting, from 97.38% to 101.65% for FS bagging and from 95 to 101.41% for FS boosting, depending on the OD-matrix. When we compare the test scores and training scores, one can even observe an increase in performance on the transport mode dimension for all methods and for the day dimension for the feature selection and for the primary activity dimension for the One R analyses as well.

The stability of the performance is here indicated by the ratio between the test and training set scores. For all four models, the ratios range between 94.33-101.26% depending on the matrix. Therefore, one may conclude that, again, over-fitting occurs approximately in a similar way and it is at an acceptable level for both models.

5.4 Conclusion

Bagging and boosting have been introduced in the last ten years as being very powerful learning ideas that improve the accuracy of the prediction. Bagging does so by lowering the variance of the prediction, while boosting provides an averaged prediction by means of voting the results of different classifiers.

In Chapter 4, two different ways were proposed to simplify complex models. On the one hand, simple heuristics were used, while on the other hand, a complex model was simplified by means of variable selection that was used before the model was built. The aim of this chapter was to investigate whether an improvement of a simple model by means of bagging and boosting will lead to a better performance of the Albatross model. For that reason, we chose two simple models: one simple heuristic (the One R approach) and hence a more weak classifier, and the model that was built after feature selection has been applied (the feature selection approach), a somewhat stronger classifier. Bagging and boosting techniques have been applied to both simple models and the results of both methods when applied to the Albatross system are rather promising. The outcomes are compared at three levels: at choice facet level, at activity pattern level and at trip matrix level. At choice facet level, both methods clearly performed better than the original analysis (the One R and the feature selection approach in Chapter 4). In six out of the nine dimensions, the models derived from applying bagging and boosting after feature selection provided even the highest accuracies overall. At activity pattern level, the best results are obtained from boosting the One R approach. Considering the few number of rules that were necessary to make up these nine models for the dimensions of the Albatross system, one can truly state here that a combination of simple models can perform better than a complex model. And the same applies at trip matrix level, especially bagging and boosting applied to the feature selection models seem to provide nice results. One may conclude that even though the bagging and boosting models do not necessarily show the best performance on all dimensions separately, the sequential execution of these models can lead to a best performance at activity pattern or at trip matrix level.

Wickramaratna *et al.* (2001) and Maclin and Opitz (1997) suggested to use bagging and boosting only with weak classifiers. Combining weak classifiers tends to ameliorate the performance, however a combination of rather powerful classifiers (such as tree induction algorithms) might end up in providing worse results is what they stated. Boosting also tends to be more susceptible to noise and it quickly over-fits the data, therefore, applying it to simpler models might result in a better combined classifier. On the other hand, it is only logical that a combination of weak classifiers outperforms the weak classifier itself, whereas a combination of not-so-weak classifiers can introduce noise in the prediction. The analyses performed in this chapter showed that combining the boosting algorithms with the One R algorithm did result in a better performance at activity pattern level, while combining bagging and boosting with the feature selection technique did provide nice results at trip matrix level. Thus we can not completely endorse to their viewpoint.

Of course, one might then ask the philosophical question if there does not exist a 'best' classifier, one whose performance cannot be beaten, not even by bagging and boosting . . . . Though, even the performance of this 'best' classifier can drop down in the sequential execution of the Albatross system. This remains an open question.

Chapter 6

Lack-of-fit on Mode Choice Models

6.1 Introduction

Over the last decade, two major changes have taken place in transportation analysis. On the one hand (as described in Chapter 3), activity-based models became more and more important when modelling travel demand. The research interest shifted away from trips and tours to the analysis of complex activity travel patterns. This change grew from the primary belief that travel is a derived demand and more emphasis is put now on the activities behind it that induce this movement in space and time. Travel is no longer seen as an isolated facet of the decision making process, it is now regarded as the result of activity patterns in space and time. The connection between people/household characteristics and activity/trip features is thus of great concern to transport modelers and for this purpose, logit models are very frequently used in transportation analysis in order to model the probability of choosing a particular mode choice above the others. On the other hand, the transportation analyst encounters far more large data sets to model when compared to e.g. 15 years ago. This is partly a consequence of the fact that more variables are considered to be relevant now. Not only transportation characteristics and demographical variables will play an important role, also activity features will come into play. This leads to a growth of the data sets in magnitude, but thanks to the improvement of the survey techniques (e.g. Virgil, Janssens *et al.*, 2004b; Chase, Doherty and Miller, 2000), data sets grew also more and more elaborate in depth. The assessment of the model fit on such complex data sets requires special attention.

Focusing on the special case of a binary response Y (e.g. preference of a certain mode choice above others in mode choice models), a well-established model is the logistic regression model in which a logit link is used, see Yamamoto *et al.*, 2000; Kockelman, 1997; among others. It has also been applied manifold in the research area of activity-based (and hence mode choice) modelling, e.g. Ben-Akiva and Lerman (1985) and Bhat (1998). The reasons are manifold: the possibility of determining prognoses for the event of interest, the ease of interpretation of the parameters in terms of odds ratios, the availability of standard software tools, etc.

There exist several methods to assess the adequacy of parametric models, see e.g. Hart (1997). For generalised linear models, a well-known method is based on the deviance, which is the likelihood ratio test contrasting the hypothesised model with a saturated model, or the Pearson test statistic (Pearson, 1900). Both are asymptotically chi-square distributed. This is a valid test procedure in case all variables are categorical, but in case of continuous explanatory variables, the number of distinct covariate patterns (which serve as cells in a contingency table) grows with the sample size and the test is no longer valid. Categorising all continuous variables might solve this problem, but at the cost of power and, more importantly, it is not clear how classes should be constructed. Note that in the Albatross data sets (Chapter 2 and Appendix) all variables are categorised as well, since this was a condition to use the CHAID induction algorithm. However, perhaps a better model fit can be found if all variables could retain their original nature?

The onset to the use of lack-of-fit tests for logistic regression models with continuous predictors was given by Hosmer and Lemeshow (1980). Many other methods and approaches were examined, some of them in very general likelihood or moment estimation methods, see e.g. Azzalini *et al.* (1989); others using order selection criteria (Eubank and Hart, 1992); or nonparametric concepts like smoothing (Kuchibhatla and Hart, 1996; le Cessie and van Houwelingen, 1991, 1993, 1995); or orthogonal series approximation (Aerts *et al.*, 1999, 2000), among others.

Most of the above mentioned methods have been shown to be rather easily implementable and to have good power characteristics, especially for low dimensional sample spaces. In case there are many explanatory variables however, most of these methods are faced with problems related to the so-called curse of dimensionality and often also related to practical difficulties of implementation. As a consequence, they lose many of their desirable properties when there are three or more explanatory variables, as discussed e.g. in Aerts *et al.* (2000).

Hosmer and Lemeshow's approach (HL) is based on a Pearson-like statistic by forming 10 equally sized groups (deciles of risk). The choice of 10 is somewhat arbitrary but simulations have shown that to be a reasonable rule of thumb. The null model has a large impact on this test statistic, since the groups are based on the fitted probabilities under this null model. In this way they solve the aforementioned problem of categorising and in the meantime they implicitly deal with the dimensionality problem. It is well-known however that this is at the cost of power.

A recent article by Pulkstenis and Robinson (2002) on the goodness-of-fit for the logistic regression setting proposes a methodology similar to that of Hosmer and Lemeshow. The grouping is also based on the deciles of risk, but it is made within the cross-classification of all categorical covariates in the model. A possible disadvantage of this method, similar to the Hosmer and Lemeshow's approach, is that it still uses the null model to define the groups. The test statistic in these methods uses the fitted probabilities under the null model in order to determine whether or not there is a lack-of-fit in the proposed model. Since we are investigating lack-of-fit, one might expect a more powerful procedure using another grouping method, based on a nonrestricted model.

Here, we propose a new method resembling the HL test statistic in that it also uses a Pearson-like statistic but now contrasting the hypothesised model with a saturated model based on a sample space partitioning driven by the recursive partitioning algorithm as used in classification trees (see e.g. Breiman *et al.*, 1984 and Zhang and Singer, 1999). Classification trees are nonparametric in nature and they can deal with large and complex data sets.

Studies in multidimensional settings as well as simulation studies will be presented to exemplify the test procedures presented next.

6.2 A Tree-Based Lack-of-Fit Test

Consider the general setting: binary response data on n subjects and a logistic regression setting with q potential covariates. Let $\pi = P(Y = 1)$, then the null hypothesis H_0 states

$$H_0 : \text{logit}(\pi) = \log\Big(\frac{\pi}{1-\pi}\Big) = \beta\mathbf{X} \tag{6.1}$$

as the correct model for our data, where $\mathbf{X}$ is a $n \times (p+1)$ matrix consisting of the n measurements on the p $(\leq q)$ covariates $x_1, \ldots, x_p$ in the p last columns (the first column of $\mathbf{X}$ consists of ones). The alternative hypothesis H_1 is not a specific alternative model. We are interested in so called omnibus tests for H_0 with power against a wide range of alternative models.

6.2.1 Pearson Statistic

First, consider the case all q explanatory variables are categorical. An $I \times 2$ contingency table can then be constructed for the observed counts with I indicating the total number of possible combinations of the q categorical variables along the rows. Let n_i be the number of subjects in covariate pattern i, y_i the number of observed events and $\hat{\pi}_i$ the fitted probability under the null model in the i-th of all I possible combinations. The Pearson goodness-of-fit test statistic (Pearson, 1900) is given by:

$$X^2 = \sum_{i=1}^{I} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}. \tag{6.2}$$

The test has an approximate chi-square null distribution with $I - p - 1$ degrees of freedom (*df*) and the null hypothesis of no lack-of-fit is rejected when the test statistic exceeds the corresponding upper α critical value.

6.2.2 Hosmer-Lemeshow Test

Consider again the null hypothesis (6.1) but now one or more explanatory variables are measured on a continuous scale. Since Pearson's chi-square test is not applicable anymore when continuous covariates are present, Hosmer and Lemeshow (1980) proposed a new statistic where the grouping strategy is based on the values of estimated probabilities. One first orders all responses according to their fitted probabilities under the null model and then classifies them into g groups. Hosmer *et al.* (1988) advocate the use of $g = 10$ 'deciles of risk' groups. More precisely, the first group contains those subjects (10 % of the sample size) with the smallest estimated probabilities, etc. Their test statistic has the same form as the Pearson test statistic, although the grouping is different. The formula is identical to (6.2) but now $I = g$, the number of groups and $\hat{\pi}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\hat{\pi}_j^0$ is the average of the probabilities $\hat{\pi}_j^0$ for all covariate values $(x_{j1}, \ldots, x_{jp})$ in group i, fitted under the null model. Also note that the HL statistic is restricted to the case $q = p$. Indeed the categorisation is based on the null model, implying no other variables are used at any place in the construction of the HL test.

As mentioned before, since the test statistic is based on the fitted probabilities under the null model, we expect in general a rather moderate power performance for this test. Because of its particular way of grouping, the cells in the contingency table are nicely balanced but random. Therefore the HL statistic has a complicated nontrivial null distribution. Hosmer and Lemeshow (1980) showed that, based on simulations, the asymptotic distribution is approximately chi-square with $g-2$ degrees of freedom where g is the number of groups.

6.2.3 A Tree-Based Test

Similar in nature to the HL test statistic, the proposed tree-based (TB) test statistic is obtained by calculating the Pearson chi-square statistic based on a $2 \times g$ table. We propose to construct groups based on the recursive partitioning algorithm underlying classification trees (see Moons *et al.*, 2002b, 2004a, 2005c). The choice of g, the number of groups, will affect the power characteristics of the method. Given a particular grouping approach, the TB test statistic T is defined in the same way as in (6.2). Whereas the partition for the HL test is based on the null model, our approach is based on recursive partitioning as applied in classification trees, which can be considered as a flexible nonparametric alternative model. This will have a beneficial effect on the power characteristics of the test statistic.

In general, we expect this approach to be more powerful than the HL test due to the fact that the structure of the individual covariate patterns is based on a classification tree. The test statistic actually measures the discrepancy between the parametric null model and the classification tree as its unrestricted nonparametric counterpart. It also covers the case that $q > p$. Indeed, the classification tree can be based on all or part of the q potential explanatory variables.

Recursive Partitioning

There are several methods to perform recursive partitioning on a data set and basically the algorithms can be classified into 2 groups: those that yield binary trees and those that result in multiway splits. CART (Breiman *et al.*, 1984) and Quest (Loh and Shih, 1997) are members of the first group, while the second group includes e.g. C4.5 (Quinlan, 1993), Cruise (Kim and Loh, 2001) and CHAID (Kass, 1980) (see also Chapter 4). Classification and regression trees (CART, see e.g. Breiman *et al.*, 1984), that we use here, is thus just one way of partitioning. It is available in many software packages and it is often used as standard reference. In summary, a tree consists of different layers of nodes. It starts from the *root node* in the first layer, the first parent node. In a binary tree, a parent node is split into 2 *daughter nodes* on the next layer. Each of these 2 daughter nodes become in turn parent nodes. This recursive partitioning algorithm continues until a node is *terminal* and has no offspring (determined by a stopping criterion). Nodes in deeper layers are getting more and more homogeneous, less ‘impure’, with respect to the response. An internal node is split by considering all allowable splits for all variables and the best split is that one with the most homogeneous daughter nodes. The 'goodness' of a split can be defined as the reduction in impurity

$$\Delta i(\tau) = i(\tau) - P(\tau_L)i(\tau_L) - P(\tau_R)i(\tau_R)$$

with $i(\tau)$ denoting the impurity of the node τ and $P(\tau_L)$ (and $P(\tau_R)$) the probability that a subject falls into the left (resp. right) daughter node τ_L (resp. τ_R) of node τ. A popular example of such an impurity measure is the entropy measure $i(\tau) = -p_\tau \log(p_\tau) - (1-p_\tau)\log(1-p_\tau)$, with $p_\tau = P(Y=1|\tau)$. In the pruning process, the initial tree is then pruned recursively, leading to a sequence of pruned and nested subtrees. From this sequence of trees, we choose the subtree with g terminal nodes. These final nodes define the groups for our test statistic. For more details on classification trees and the recursive partitioning and pruning process, we refer to Breiman *et al.* (1984) and Zhang and Singer (1999).

Modified TB Tests

Whereas the Hosmer and Lemeshow procedure leads by definition to balanced groups, the tree-based method typically results in unbalanced groups. So, certain final nodes might contain only a few observations and this might de-stabilise the distributional properties of the tree-based test statistic. A simple remedy is to adapt the stopping rule used in the recursive partitioning algorithm, namely not to split a node any further if it contains less than a certain percentage of the full sample size. Additional to that, we studied two variations of statistic (6.2). A first modification is a weighted Pearson statistic

$$T_W = \sum_{i=1}^{g} w_i \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$$

with weight $w_i = gn_i/N$, giving less weight to small groups. If we take $w_i = 1$ the test statistic reduces to the original version T. Also, for the HL test $N/n_i = g$ such that $w_i = 1$ (but with a different grouping).

As a second modification, the TB test can be based on the Cressie and Read (1984) family of power divergence statistics, i.e.

$$T_{CR} = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{g} \Big\{ y_i \Big(\Big(\frac{y_i}{n_i\hat{\pi}_i} \Big)^\lambda - 1 \Big) + (n_i - y_i) \Big(\Big(\frac{n_i - y_i}{n_i(1-\hat{\pi}_i)} \Big)^\lambda - 1 \Big) \Big\}$$

with $-\infty < \lambda < \infty$. For $\lambda = 1$ it equals the Pearson based formulation (6.2). Cressie and Read (1984) recommend the statistic with $\lambda = \frac{2}{3}$, which they found less susceptible to effects of sparseness. (Although, under certain circumstances, this figure has been countered by García-Pérez and Núñez-Antón (2004).) We will compare the three versions of the TB test in the simulations and the data examples.

6.2.4 Distributional Properties

As for the HL test, the distribution of the TB tests cannot be obtained from a straightforward application of the usual theory for chi-square goodness-of-fit tests. Indeed the categorisation, the sample space partitioning, is random. Moore and Spruill (1975) considered the distribution of chi-square test statistics in this situation and their main result is that the distribution, under appropriate regularity conditions, is that of a chi-square distribution with the usual reduction in degrees of freedom due to estimated parameters plus a weighted sum of independent chi-square random variables each with one degree of freedom where the weights are eigenvalues of a particular matrix.

As in the case of the HL method, it is very hard to turn this theoretical result into a practical form. Even if one would be able to fully specify this distribution in all detail, it would be difficult to estimate all unknown parameters and to implement it in practice. Alternatively one could try, similar to what Hosmer and Lemeshow did, to empirically investigate the null distribution in different settings and to derive a simple, practical and reasonable approximation. From the simulations, part of which are presented in the next section, we learned that a chi-square distribution with $2 \times g - p$ degrees of freedom is a reasonable choice.

One does not need to rely on the asymptotic distribution of the TB statistics to conduct the test. A null distribution can always be simulated by a parametric bootstrap method (see e.g. Section 4.2.3 in Davison and Hinkley, 1997). This approximate Monte Carlo approach is illustrated in the data examples and in the second simulation study.

6.3 Simulation Study 1: Two Covariates

In this section, we examine the null distribution of the TB tests. We compare the power characteristics of the T, T_W and T_{CR} tests with the HL test and with an oracle test (a likelihood ratio test, testing the null model against the true alternative, only known by an 'oracle').

6.3.1 Setting 1

This first setting shows the power of the tree-based test on one particular misspecification: one forgot to take a variable into account. This setting is very important, since this occurs quite often in reality, especially for higher order terms of a variable (quadratic is shown here) or for an interaction variable. First, we examine the null distribution.

The Null Model

Consider $Y_i \sim \text{Ber}(\pi(x_i, z_i)), i = 1, \ldots, 100$, where x_i and z_i are fixed values, uniformly distributed on $(-6, 6)$. The null hypothesis is given by $\text{logit}(\pi(x_i, z_i)) = \beta_0 + \beta_1 x_i + \beta_2 z_i$, $\beta_0 = 0.0, \beta_1 = 0.8$ and $\beta_2 = 0.3$. We consider four test statistics: the HL test, with partitioning up to 10 groups as advised by Hosmer and Lemeshow (1980); the TB test T, the weighted version T_W, and the Cressie-Read version T_{CR}. All three versions of the TB statistic are based on final trees pruned up to 7 final nodes. Simulated critical points (1000 runs) of the null distributions are presented in Table 6.1, together with exact critical points of the χ^2 approximations.

Table 6.1: Simulated 1, 5 and 10% critical points of the HL, T, T_W and T_{CR} test statistics for the null hypothesis of the setting in the first simulation study

Null model			
Test	0.10%	0.05%	0.01%
HL	12.46	15.97	28.85
$\chi^2(8)$	13.36	15.51	20.09
T	20.86	25.13	32.45
T_W	16.40	18.93	23.68
T_{CR}	19.76	22.97	27.60
$\chi^2(12)$	18.55	21.03	26.22

The simulations illustrate that the null distributions are reasonably well approximated by the χ^2 distribution with 10-2=8 degrees of freedom for the HL test and by the χ^2 distribution with $(7 \times 2) - 2 = 12$ degrees of freedom for the TB test statistics. The χ^2_{2g-p} approximation seems to work fairly well. There are some deviations from the approximate distributions for the tree-based statistics (especially the T version) as well as for the HL test statistic. We recommend to use a bootstrap simulation of the null distribution, next to this approximate chi-square distribution.

Table 6.2: Simulated rejection percentages of the HL, T, T_W and T_{CR} tests for two alternative models

	Interaction Model					
	$\beta_3 = 0.10$		$\beta_3 = 0.20$		$\beta_3 = 0.30$	
	Test(0.10)	Test(0.05)	Test(0.10)	Test(0.05)	Test(0.10)	Test(0.05)
HL	27.14	14.29	68.47	51.20	83.17	72.95
T	36.33	24.90	93.17	84.54	100	100
T_W	45.71	32.04	97.19	94.58	100	100
T_{CR}	40.20	26.73	96.18	91.16	100	100
Oracle Test	77.02	73.08	99.59	99.59	100	100
	Quadratic Model					
	$\beta_3 = 0.10$		$\beta_3 = 0.20$		$\beta_3 = 0.30$	
	Test(0.10)	Test(0.05)	Test(0.10)	Test(0.05)	Test(0.10)	Test(0.05)
HL	13.83	5.01	17.60	6.80	24.45	9.62
T	29.66	15.43	89.2	78.4	99.80	99.20
T_W	47.70	32.06	95.40	91.40	100	99.80
T_{CR}	36.47	23.85	94.40	86.60	99.80	99.80
Oracle Test	86.97	84.52	99.59	99.59	100	99.80

Power Characteristics

We study the power characteristics for two alternative models: an interaction model $\text{logit}(\pi(x_i, z_i)) = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i$ with $\beta_0, \beta_1, \beta_2$ as in the previously stated null hypothesis and a quadratic model $\text{logit}(\pi(x_i, z_i)) = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i^2$ with $\beta_0, \beta_1, \beta_2$ as stated before. For both models we take $\beta_3 = 0.1, 0.2$ and 0.3 and all simulated rejection rates are based on 500 runs.

The results in Table 6.2 clearly indicate that, as expected, all rejection rates increase with the value of β_3. All TB tests show much higher power results than the HL test. The performance of the HL test is particularly poor for the quadratic model. From the TB tests, the weighted version has the highest power, in both models. The T_{CR} test is also improving the T test, but less pronounced. All results are (not surprisingly) far below the rejection rates of the oracle test, especially for alternatives close to the null. This is the price to pay for an omnibus type of test. For alternatives further away from the null value, the TB tests reach higher powers, closer to that of the oracle test.

6.3.2 Setting 2

In this second setting, we will focus on a change in functional form of one of the variables in the null model.

The Null Model

Consider again $Y_i \sim \text{Ber}(\pi(x_i, z_i)), i = 1, \ldots, 100$, where x_i and z_i are fixed values, x_i uniformly distributed on $(-6, 6)$ and z_i uniformly distributed on $(1, 20)$. The null model yields $\text{logit}(\pi(x_i, z_i)) = \beta_0 + \beta_1 x_i + \beta_2 z_i$, $\beta_0 = 0.0, \beta_1 = 0.8$ and $\beta_2 = 0.3$. We consider again the four test statistics: the HL test, with partitioning up to 10 groups; and the three versions of the the TB test pruned up to 5 ànd 7 final nodes. In this way, we can investigate if an increase in the number of final nodes causes an improvement in the approximation of the null distribution and/or a change in the power characteristics. Table 6.3 shows the simulated critical points (1000 runs) of the null distributions together with the exact critical points of the χ^2 approximations.

Table 6.3: Simulated 1, 5 and 10% critical points of the HL, T, T_W and T_{CR} test statistics for the null hypothesis of the second setting in the first simulation study

Null model			
Test	0.10%	0.05%	0.01%
HL	10.66	14.79	34.53
$\chi^2(8)$	13.36	15.51	20.09
	5 final nodes		
T	11.95	15.16	22.23
T_W	10.05	11.87	16.44
T_{CR}	12.55	15.32	20.11
$\chi^2(8)$	13.36	15.51	20.09
	7 final nodes		
T	17.56	20.48	35.49
T_W	15.19	17.36	24.44
T_{CR}	17.64	20.04	26.99
$\chi^2(12)$	18.55	21.03	26.22

The simulated null distributions are reasonably well approximated by the respective χ^2 distributions. One cannot say for certain that the results with seven final nodes are better approximated than those with five final nodes. The null distribution

Table 6.4: Simulated rejection percentages of the HL, T, T_W and T_{CR} tests for the alternative functional form model

	Test(0.10)	Test(0.05)
HL	18.29	5.49
	5 final nodes	
T	31.40	20.80
T_W	17.40	9.00
T_{CR}	30.20	20.20
	7 final nodes	
T	38.01	27.64
T_W	27.03	14.63
T_{CR}	38.41	27.44
Oracle Test	88.56	85.81

of the Cressie-Read version of the tree-based statistic appears to be closest to the χ^2 distribution. Still, there are some deviations, especially at the 1% critical point of the T and the HL statistic. Therefore, we still recommend to use a bootstrap simulation of the null distribution, next to this approximate chi-square distribution.

Power Characteristics

The alternative model for this functional form setting is given by $\text{logit}(\pi(x_i, z_i)) = \beta_0 + \beta_1 x_i + \beta_2 \log(z_i)$, with β_0, β_1 and β_2 as specified in the accompanying null model. The assumption in this setting is thus that an incorrect functional form is used for one of the explanatory variables.

The results in Table 6.4 indicate that the power of the test increases with an increasing number of final nodes. The proposed tree-based tests clearly have rejection rates that are much higher when compared to the Hosmer and Lemeshow test. The T_{CR} and the T test show the best results. All rejection rates are far below the results of the oracle test, but then again, the tree-based statistic is an omnibus type of test, not tailored for a particular alternative model.

6.3.3 Further Discussion

The simulation results shown in this section are a selection of a wide range of settings and models we studied. The TB test shown here is based on $g = 5$ and $g = 7$ final nodes whereas the HL used $g = 10$ deciles of risk groups. We also considered the

TB test with the same number of 10 groups or final nodes. The results were quite comparable with the ones shown in Table 6.1-6.4. With the choices $g = 5$ and $g = 7$ we show that the TB test is able to have, even based on a smaller $g \times 2$ table, a substantially higher power than the HL with the recommended $g = 10$ deciles of risk. In the examples below, as in the second setting above, we varied the number of final nodes. In general, it has an effect on the power of the TB test. For a larger sample size and for a more complex null model, we recommend to use a larger number of final nodes. Finding the optimal number of final nodes from a theoretical point of view is of course related to the distributional properties of the TB statistic and it is expected to be a difficult theoretical problem. This is not the focus here. This issue is of more interest to the theoretician than the practitioner. Our experience says that the latter will rarely find it necessary to take g larger than 20 or so, regardless of sample size.

Simulations not shown here include settings in which the true unknown model contains more variables than the null model ($q > p$). Simulations do very convincingly confirm that the HL test has completely no power to detect such kind of deviations from the null model, whereas the TB test does show good power, at least if the tree is constructed incorporating the missing covariates.

A small technical complication is that the sequence of pruned subtrees might not contain a subtree with exactly g final nodes (because a larger subbranch is pruned away). The pruning algorithm could be modified but our experience is that it rarely happens. Simulations showed that the algorithm leading to a final tree with number of final nodes at least equal to but as close as possible to g sometimes results in a tree with $g + 1$ final nodes but it had a negligible effect on the final tree and the value of the TB test statistics.

6.4 Data Examples

In this section, some bio-statistical data sets will be introduced, because goodness-of-fit tests have a large background in this type of literature. In this way, we are able to compare our proposed tree-based tests to some existing test statistics.

In order to have a data set that is comparable in complexity to the data sets that are often used in transportation research, we have also added an example on the Belgian Health Interview Survey data. And finally, a last example illustrates the TB test in the context of transportation research. In this example, we also show the use of the tree-based test statistic as a model selection tool.

6.4.1 Example 1: The GVHD Data

This data set comes from a clinical trial with 166 patients receiving a bone marrow transplant from an HLA-identical family donor. It is a subset of data from a study described in Gratama *et al.* (1992) and it was used by le Cessie and van Houwelingen (1995) to illustrate a lack-of-fit test based on a score test in a random effects model.

The patients had one of the three following diagnoses: (a) severe aplastic anemia, (b) acute non-lymphoblastic leukemia, or (c) acute lymphoblastic leukemia and who had a complete serology of four herpes-type viruses for donor and recipient. The outcome of interest is the occurrence of grades II-IV acute graft versus host disease (GVHD). Possible risk factors for GVHD considered here are the age of the donor, the age of the recipient , diagnosis of the disease of the recipient (three categories as described above), sex of the donor and sex of the recipient. A variable derived from these latter two variables is whether the sex of the donor matches the sex of the recipient (sex match). The age of the donor and recipient are treated as continuous variables, the other variables as categorical.

As an illustration we examine the goodness of fit of the same sequence of models as chosen by le Cessie and van Houwelingen (1995). Table 6.5 shows the p-values obtained by le Cessie and van Houwelingen CH_i ($i = 1, 2$ according to two different distance measures), together with those obtained by Hosmer-Lemeshow (HL, using deciles of risk), the TB test T, the weighted version T_W and the Cressie-Read version T_{CR} (all with pruning up to 7 terminal nodes). All p values in the last four columns were obtained by a parametric bootstrap (1000 runs).

Table 6.5: Test results GVHD data: p-values for four null models with diagnosis D, age donor A and sex match S. Left part (2 left columns) from le Cessie and van Houwelingen (1995); right part (4 right columns) from a parametric bootstrap simulation with 1000 runs

Null model	CH_1	CH_2	HL	T	T_W	T_{CR}
D	.058	.028	.694	.032	.039	.025
D, A	.140	.050	.182	.104	.081	.096
D, A, S	.390	.046	.721	.266	.228	.235
D, A, S, A^2	-	.260	.229	.365	.334	.340

The table shows that the TB tests yield p values comparable to those obtained by the test of le Cessie and van Houwelingen (1995). The Hosmer-Lemeshow test clearly suffers from lack of power to reject poor null models. The modified tree-based tests T_W and T_{CR} confirm their better power characteristics.

6.4.2 Example 2: The POPS Data

The POPS dataset originates from the Project on Preterm and Small-for-Gestational-Age Infants in the Netherlands (POPS), a Dutch follow-up study on preterm infants by Verloove and Verwey (1988), and was extensively analysed by le Cessie and van Houwelingen (1991, 1993). Data were collected on 1338 infants, born in 1983 in The Netherlands with a gestational age of less than 32 completed weeks and/or a birth weight of less than 1500 g. After deleting the observations with missing data, a data set of 1310 infants remained. We consider the situation after 2 years. The response variable indicates whether or not the infant has died within 2 years or has survived but with a major handicap. The explanatory variables are gestational age and weight of the babies at birth. As an illustration, we consider three null models with at least one quadratic effect. Table 6.6 shows the results for the TB test and compares them with the results from several other tests from literature. We restricted attention to the TB test using the Cressie-Read statistic.

Table 6.6: Test results POPS data: p-values for three null models with gestational age A and birth weight W

Null model	CH	BR	ACH_1	ACH_2	B_S	B_N	T_{CR}	HL
A, A^2, W	.02	.01	-	-	.000	.006	.012	.125
A, W, W^2	-	-	-	-	.000	.000	.044	.002
A, A^2, W, W^2	.45	.06	.07	.02	.126	.138	.090	.207

The last four columns show p-values for two Bayesian motivated tests, a singleton test B_S and a nested test B_N (Aerts *et al.*, 2004) using a sequence of alternative models including up to fifth order main and interaction effects (using orthogonal polynomials), the TB test T_{CR} based on the Cressie-Read statistic with pruning up to 15 terminal nodes, and the HL test based on deciles of risk. All p values were simulated using the parametric bootstrap (1000 runs).

The first four columns show some results from other test statistics proposed in literature: a kernel based goodness of fit method (CH) proposed by le Cessie and van Houwelingen (1991, 1993), the Brown statistic (BR, Brown (1982), and an order selection score test ($ACH1$) and the value of a score based AIC criterion ($ACH2$) as reported by Aerts *et al.* (2000).

The p-values in Table 6.6 show that there is clear evidence against any model without both quadratic terms (model 1 and 2). Only the HL test does not reject model 1. As also discussed in Aerts *et al.* (2000), there is some evidence against model 3 with both quadratic terms, but the different test results disagree.

6.4.3 Example 3: The HIS Data

The Belgian Health Interview Survey (HIS) was conducted in 1997. The main objective of this survey was to determine the population's health, life style and use of health services. A total sample of 10,000 interviews (0.1% of the Belgian population) was planned, equally spread over the year 1997. A detailed description of the sampling scheme used in the HIS was published in Quataert *et al.* (1998). Using classification trees in comparison to logistic regression, Hens *et al.* (2002) examined the profiles of persons who are at risk to obtain certain diseases or who do not respond to prevention programs as e.g. cervix cancer screening via smears. The Belgian communities are responsible for cervix cancer screening as a part of the preventative health care. According to the Belgian National Cancer Registry (Haelterman, 1999), cervix cancer is the fifth most common cancer among women in Belgium in the period of 1993-1995. Therefore it is not surprising that for health policy goals cervix cancer is an important point of attention. In an early stage cervix cancer can already be detected by means of a simple smear. Because early detection decreases the mortality substantially, women between 25 to 64 years old should have a smear every 3 to 5 years, according to the European guidelines (Coleman *et al.* (1993), Advisory Committee on Cancer Prevention (2000), Arbyn *et al.* (2001). The question investigated in Hens *et al.* (2002) was in what respect the group of women, aged 25-64, not having a smear is different from the group of women that did have a smear taken in the past three years. Special interest goes out to whether an invitation letter increases the probability of undergoing screening. For more details on this application, we refer to Hens *et al.* (2002).

Here, as an illustration, we examine the lack of fit of some logistic regression models, similar to the ones used in Hens *et al.* (2002). From the HIS data file, only women aged between 25 and 64 were selected. Women without uterus are excluded from the analysis, leading to 2893 subjects. After deleting all observations with missing data, there are 1945 observations left. The binary response variable is 'screening status' (a smear taken in the last three years, yes or no). The general topics of the explanatory variables are shown in Table 6.7.[1]

Table 6.7: General topics of the explanatory variables

Lifestyle	Physical Activity
	Nutritional Habits
	Alcohol Consumption
	Smoking
Health Problems	Subjective Complaints
	Chronical Conditions
	Mental Health
	Functional Limitations
Prevention and Health Promotion	Vaccination
	Cardiovascular Prevention
	Aids Prevention
Use of Health Care	Contacts with GP
	Contacts with specialists
	Contacts with dentist
	Paramedics
	Alternative Methods
	Hospital Admissions
	Use of Medication
Health and Society	Social Health
	Access to Health Care

With 85 explanatory variables (essentially all of them are categorical) it is almost impossible to investigate for each covariate the nature of the relationship (linear, quadratic, etc.). Moreover, there are 7140 possible two-way interactions, considering all of them is hardly feasible. The logistic models we consider here, contain the following 25 explanatory variables: Age Category (ordinal, 9 categories), Income (ordinal, 5 categories), Household Type (nominal, 5 categories), Consumed Bread (binary), Snack Eating (ordinal, 3 categories), Province (nominal, 11 categories), Hospital Admission (binary), Blood Pressure Control (binary), Profession (nominal, 9 categories), Lack of Physical Activity (binary), BMI Category (ordinal, 6 categories), Cholesterol Control (binary), Frequency Heavy Drinking (ordinal, 6 categories), Milk Consumption (binary), Daily Drinker (binary), Medication (binary), Appreciation

[1] The full questionnaire can be consulted at
http://www.iph.fgov.be/epidemio/epien/crospen/hisen/table.htm.

Social Relationships (binary), Invitation Letter (binary), Knowledge HIV protection (binary), Frequency Dentist Visits (ordinal, 5 categories), Preventive Tooth Control (binary), HIV Screening (binary), Breakfast Eating (ordinal, 4 categories), Marital Status (nominal, 4 categories) and Educational Level (ordinal, 5 categories). Most of the explanatory variables in the models are variables indicating an awareness of the patient towards his own health status, e.g. cholesterol control, heavy drinking moments, blood pressure control, etc. Other predictor variables are of a demographic nature, e.g. age, income and province. Appreciation of social relationships seems to have a small influence in the models. For health policy purposes the effect of a screening invitation is of great importance. This specific explanatory variable indicates whether a person received an invitation letter advising her to have a cervix cancer screening. For more details, we again refer to Hens *et al.* (2002).

Our first model (referred to as model 1) contains the above explanatory variables with a single linear effect for all ordinal variables. As a second model (model 2), we consider model 1 extended with a quadratic effect for Age Category, Frequency Heavy Drinker, Income, Breakfast Eating and Educational Level. A third model (model 3) incorporates the interaction effects Age × Breakfast Eating, Age × Educational Level, Breakfast eating × Educational level, Educational level × Province and Educational Level × Household Type. As a last model, model 4, we reconsider the model with only main effects (as in model 1), but now with all explanatory variables as nominal (saturated main effects model).

This is clearly a setting with a lot of variables and observations, a setting where many other lack-of-fit tests would encounter serious problems, also in the practical implementation of the method. The HL test and the TB test can be used in the same fashion as in the other two examples. For all analyses we took 20 final nodes, for both test statistics. Simulated p-values, based on 500 bootstrap replications, are shown in Table 6.8.

Table 6.8: Test results HIS data: p-values for four null models

	Hosmer Lemeshow		Tree-based test	
Null model	statistic	p-value	statistic	p-value
Model 1	30.63	0.03	68.51	0.00
Model 2	10.45	0.94	65.36	0.00
Model 3	14.19	0.69	60.53	0.02
Model 4	12.09	0.83	59.94	0.10

The results of the weighted and the Cressie-Read version of the tree-based test are not included, since the results are very similar to the ordinary version. Model 1 is rejected by both tests. The Hosmer-Lemeshow test however fails in rejecting model 2 and 3, which are clearly rejected by the tree-based test. There is not much evidence against model 4, by either lack-of-fit test. This example illustrates the use of the tree-based tests in complex logistic regression models and confirms the higher power as compared to the Hosmer-Lemeshow test.

6.4.4 Example 4: Dutch Car Driver data

The paper by le Cessie and van Houwelingen (1993) brought to our attention that these tests can function as goodness-of-fit statistics on one hand, in that they show the power of a certain test to detect a specific lack-of-fit, while on the other hand, the tests can serve as model selection criteria to choose a certain model out of a series as well. Thus, in this section, we examine the use of the weighted tree-based test statistic (T_W) as a model selection tool on the Dutch Car Driver data (see section 2.3 in Chapter 2). A series of logistic regression models is fit, and we will investigate the power of the T_W statistic to reject a model that is not sufficiently suited to fit the data. We compare the power of T_W to that of the HL statistic as well as to some well-established model selection criteria, such as Akaike's Information Criterion (AIC) (Akaike, 1973) and the Bayesian Information Criterion (BIC) (Schwartz, 1978). As we mentioned above, not the approximate null distribution, but the null distribution based on the parametric bootstrap (500 runs) was used to obtain the p-values. We choose 10 groups for both statistics in this second example, since this data set is much smaller and more final nodes in the classification tree are not necessary to make the sample segmentation. An overview of the models that were used and the corresponding p-values of the HL and the T_W statistics as well as the AIC and the BIC values can be found in Table 6.9.

We started with the model that was chosen by applying the forward stepwise regression technique. The model that came up as resulting model from this technique was chosen as initial model. This initial model contains 25 variables, 12 categorical and 13 continuous ones. A general remark is that transport characteristics are most important to determine the choice of the vehicle that is used to make the trip. In second place we have person and household characteristics, while activity and tour characteristics appear to be less relevant to the decision of mode choice. The first initial model has just a single linear effect for all the variables. The results on the weighted tree-based statistic clearly indicate that this first model is not adequate

Table 6.9: Test results Mod012 data: p-values and goodness-of-fit measures on different null models

Variables used in model	HL		T_W		AIC	BIC
	Statistic	P-value	Statistic	P-value		
Model 1: $x_2, x_4, x_6, x_7, x_8, x_9$ $x_{11}, x_{13}, x_{14}, x_{17}, x_{19}, x_{20}, x_{21}, x_{22}$ $x_{25}, x_{27}, x_{28}, x_{30}, x_{31}, x_{32}, x_{33}, x_{36}$ x_{37}, x_{38}, x_{39}	6.09	0.648	64.71	0	926.98	1050.29
Model 2: Model 1 + interaction $x_{31} \times x_7$	4.47	0.800	64.77	0.002	928.91	1057.16
Model 3: Model 2 + interactions $x_7 \times x_{30}$ and $x_7 \times x_{38}$	14.33	0.058	64.89	0.006	925.42	1063.53
Model 4: Model 3 + interactions $x_{30} \times x_{14}$ and $x_{30} \times x_{33}$	5.31	0.670	45.02	0.028	910.23	1058.90
Model 5: Model 4 + interactions $x_{38} \times x_{36}$ and $x_{14} \times x_{38}$	5.63	0.616	45.14	0.034	912.38	1070.22
Model 6: Model 5 + interactions $x_{14} \times x_9$ and $(x_{30} \times x_{33})^2$	7.35	0.434	39.63	0.044	910.41	1078.11
Model 7: Model 6 + interactions $x_{33} \times x_{28}$ and $x_{38} \times x_{13}$	8.46	0.358	38.62	0.048	902.98	1080.55
Model 8: Model 7 + interactions $x_{38} \times x_{20}$ and $x_{20} \times x_{17}$	8.62	0.328	35.64	0.066	896.94	1084.38
Model 9: Model 8 + interaction $x_{20} \times x_{32}$	9.55	0.274	35.14	0.074	898.56	1090.92

enough to fit this large amount of data. Hosmer and Lemeshow's statistic cannot reject this model. Now, this initial model will be chosen as a starting-point model to add quadratic effects as well as interactions. These terms were suggested by the tree built on the same variables that were used to make up the initial model. A different split on the same variable in a left and a right part of the tree might suggest an interaction term for instance, while a second split on the same variable might indicate a quadratic effect (see Figure 6.1).

As we expected, in general, the value of the tree-based test statistic decreased as more variables come into play, however, there is evidence that until the seventh model, none of the models performs good enough. The Hosmer and Lemeshow statistic, on the contrary, fails in rejecting each of these models. Strangely enough, its value decreases first and increases again afterwards, when more variables are added to

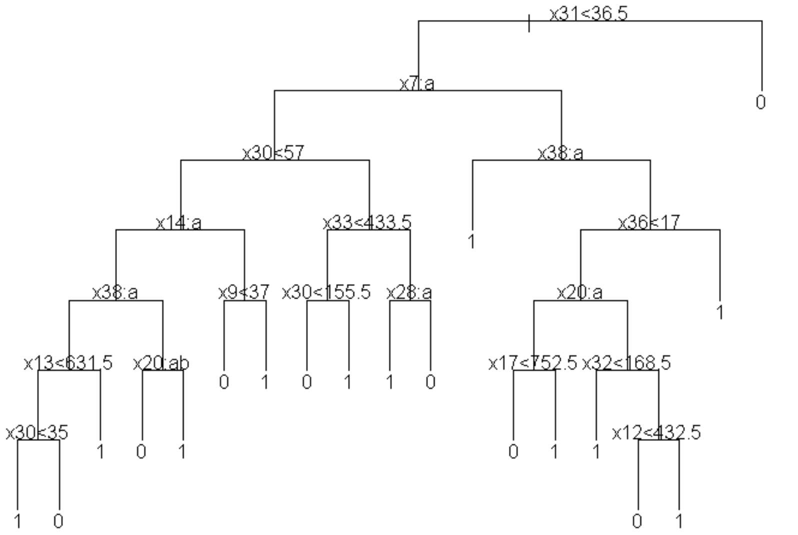

Figure 6.1: *Classification tree on the mod012 Car driver data*

the models and this behaviour cannot be explained. But this is clearly a setting with a large number of variables, a setting where many of the goodness-of-fit tests would encounter serious problems, also in the practical implementation of the method. Therefore, it is our belief that the Hosmer and Lemeshow was not developed for such applications, while the tree-based statistic was especially designed for this type of large data-mining situations that we will encounter more and more in the future. The Akaike Information Criterion reaches its minimal value at model 8, while the Bayesian Information Criterion attains its minimal value at the first model. This fourth example illustrates that the use of the tree-based test as an instrument for model selection behaves very similar to Akaike's AIC criterion. Indeed, model 8, the model selected by AIC, is the first model (in the sequence of nested models) that is not rejected by the tree-based test (at the 5% level). Compared to AIC, the use of the tree-based test for model selection is of course computationally much more extensive. On the other hand, when selecting a final model from a family of candidate models, the tree-based test might indicate, by its p-value, that the model selected by AIC is still inappropriate and needs to be further adapted and extended and it can give a suggestion how.

These four data examples illustrate the use of the different tree-based tests in (complex) logistic regression models and they confirm the higher power when compared to the Hosmer and Lemeshow test.

6.5 Simulation Study 2: High Dimensional Data Examples

6.5.1 Example 1: San Francisco Bay Data

This first example discusses the results of a small simulation study in a high dimensional covariate space based on the San Francisco Bay data (Purvis, 2003, see Section 2.2 in Chapter 2). It covers the use of the weighted tree-based test statistic (T_W) as a goodness-of-fit statistic. We want to investigate the power to detect an interaction term. In order to get more insight on this power behaviour, we considered the 24752 × 27 design matrix **X** (first column consists of ones) and we generated new response values according to the null model and the model extended with one interaction term Age × Auto Own

$$\text{logit}P(Y=1) = \beta_0 + \sum_{j=1}^{26} \beta_j x_j + \beta_{27} x_3 \times x_{18}$$

with $\beta_{27} \in \{0.05, 0.5, 50, 500, 1000, 1500\} \times 10^{-4}$. This interaction variable can determine whether there is a difference in car use amongst young drivers that possess a car compared to older drivers. Within each simulation run, the p-values were simulated based on 1000 bootstrap null samples for the HL test and the T_W test. Next, percentage rejections (determined on 500 runs) were calculated at 0.10, 0.05 and 0.01 significance levels. The HL statistic was based on 10 groups (the deciles of risk), while the tree-based test T_W is based on a final tree pruned up to 15 final nodes.

Table 6.10 shows that, as expected, the rejection rates of the tree-based statistic increase with the value of β_{27}. If more emphasis is put on the forgotten term, it will be detected better as it should. The tree-based test clearly outperforms the Hosmer and Lemeshow test here. The performance of the latter one is extremely poor, they show very low, almost no power in detecting a forgotten interaction term.

Table 6.10: Rejection rates for different values of β_{27}

$10^{-4}\times$	$\beta_{27} = 0.05$			$\beta_{27} = 0.5$		
	test(0.10)	test(0.05)	test(0.01)	test(0.10)	test(0.05)	test(0.01)
HL	10.0	5.6	1.4	9.2	6.4	1.0
T_W	10.6	5.8	0.6	13.2	7.0	0.8
$10^{-4}\times$	$\beta_{27} = 50$			$\beta_{27} = 500$		
	test(0.10)	test(0.05)	test(0.01)	test(0.10)	test(0.05)	test(0.01)
HL	12.8	4.8	0.8	14.0	7.4	2.4
T_W	18.2	9.0	1.0	59.0	44.4	19.8
$10^{-4}\times$	$\beta_{27} = 1000$			$\beta_{27} = 1500$		
	test(0.10)	test(0.05)	test(0.01)	test(0.10)	test(0.05)	test(0.01)
HL	10.2	4.4	0.4	4.8	2.4	1.8
T_W	91.0	87.0	68.0	94.0	88.2	65.0

6.5.2 Example 2: The HIS Data

Inspired by the HIS example, this section discusses the results of a small simulation study in a high dimensional covariate space (based on the HIS data). We consider the setting of model 1 as discussed in the previous section, with 25 explanatory variables and model 1 serves as null model. The analysis in the previous section showed that model 3 with some interaction terms was not rejected. But how large is the power to detect an interaction term in this specific situation? To get some more insight on this power behaviour, we took the 1945 × 26 design matrix $\mathbf{X}$ (first column existing of ones) and generated new response values according to the fitted model 1 (the null model) and model 1 extended with one interaction term Age Category × Income

$$\text{logit}\{P(Y_i = 1)\} = \beta_0 + \sum_{\ell=1}^{25} \beta_\ell x_{\ell i} + \beta_{26}(x_{1i} \times x_{2i})$$

with $\beta_{26} \in \{0.0, 0.1, 0.2\}$. Figure 6.2 shows the results based on 150 simulation runs. Within each simulation run, p-values were simulated based on 100 bootstrap null samples for the HL test and for the three versions of the TB test. Next, percentage rejections were calculated at significance levels 0.10, 0.05 and 0.01. Typically a larger number of bootstrap samples is needed to accurately estimate p values and associated power, but we believe our results are indicative in their comparison between the different tests.

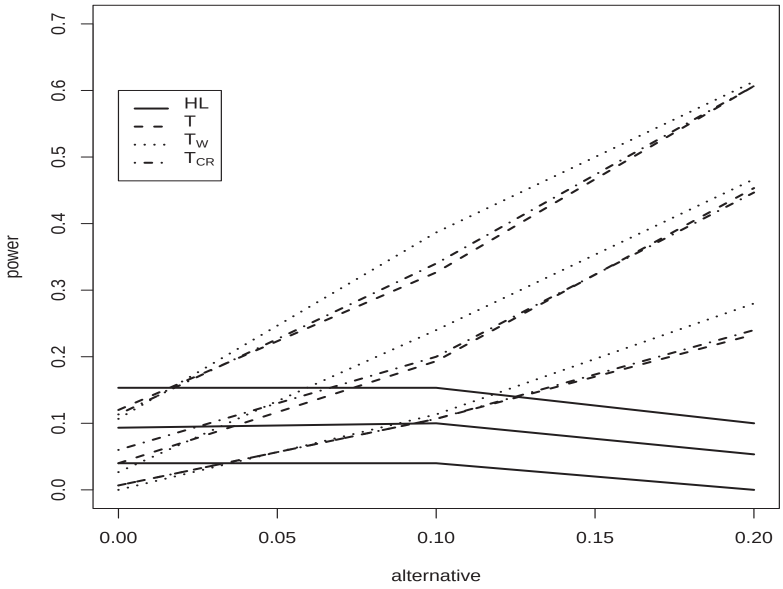

Figure 6.2: *Simulated power curves for the HIS simulation study. Each test HL, T, T_W, T_{CR} has three curves: an upper curve for level 0.10, a middle curve for level 0.05 and a lower curve for level 0.01. Each curve connects the rejection rates for $\beta_{26} \in \{0.0, 0.1, 0.2\}$, for a particular test at a specific level*

Each test HL, T, T_W, T_{CR} has three curves: an upper curve corresponding with level 0.10, a middle curve with level 0.05 and a lower curve for level 0.01. Each curve connects the rejection rates for $\beta_{26} \in \{0.0, 0.1, 0.2\}$, for a particular test at a specific level. This figure shows a very poor power behaviour for the Hosmer-Lemeshow test, in this simulation setting. The tree-based tests show reasonable power curves, increasing with the value of the alternative. The three versions of the TB test are very comparable, with some slight advantage for the weighted version.

6.6 Conclusion

A tree-based test statistic can be used to assess the goodness-of-fit of a logistic regression model containing continuous covariates or a mixture of continuous and categorical covariates. It is a nice example showing how a nonparametric method can be used to confirm or improve a parametric model. If the tree-based test rejects the parametric null model, a closer inspection of the classification tree might reveal a particular deviation of the null model. A different split on the same variable in a left and a right part of the tree might suggest an interaction term for instance.

Simulations indicate that the proposed tree-based test statistic has, compared to the Hosmer-Lemeshow test, very promising power characteristics in detecting incorrectly modelled variables, omitted interaction effects as well as higher order effects, even in a high dimensional covariate space. Further theoretical research is needed to investigate the asymptotical null distribution in more detail. In small sample studies the proposed rule of thumb appears to work reasonably well, though on higher dimensional data sets, the asymptotic distribution can be quite wrong. For all practical purposes, however, one can rely on the bootstrap approach.

The Hosmer-Lemeshow test has been generalised to repeated binary observations using generalised estimating equations by Horton *et al.* (1999) and to clustered binary data by Geys *et al.* (2002). Also tree-based models have been developed for these settings, see e.g. Zhang and Singer (1999). Future research will take a closer look at extensions of the tree-based test to multivariate, longitudinal or clustered responses. Also the application of other classifiers using multi-way instead of binary splits is a possible avenue for future research.

Chapter 7

Use of Nonlinear Models in Determining Mode Choice

7.1 Introduction

Activity-based models now held a prominent place when modelling travel demand, but the transportation analyst is faced with a new problem: he/she encounters a much wider variety of types of models to choose from. However, the use of statistics in this field is still dominated by linear models, a legacy of pre-computer times. And apart from this, many users of multiple regression (based on data sets that include at least one continuous covariate) include only linear terms in the covariate(s). If curvature in the relationship between the outcome variable is suspected, the model may be extended to include a quadratic term, but cubic or higher order polynomials are rarely used. Nonparametric and spline smoothers are powerful and flexible tools which indeed pose few limitations on the functional form, however many users do not require such sophistication, but they do need models that are reasonably flexible, easy to understand and parsimonious. Consequently, over the last decennia, nonlinear as well as semi-linear statistical models became popular in other fields such as medicine. They have been applied widely in pharmacokinetics and -dynamics, epidemiology, survival analysis, clinical trials, . . . (Royston and Altman, 1994; Royston *et al.*, 1999; Faes *et al.*, 2003). Though, also the nonlinear machine learning techniques have their application in various areas, such as pattern recognition, medicine, bioinformatics, cryptography, . . . (Ben-Yacoub, 1999; Burges, 1998; Veropoulos *et al.*, 1999; Brown *et al.*, 1999; Chen *et al.*, 2001) The idea of this chapter is to investigate what the semi-linear and nonlinear models could add to transportation analysis. We know, e.g. that the income is inversely related to the probability of using slow transport. But whether this inverse relationship is cubic, quadratic, linear, or determined by some square root is not sure? This is only a small problem that can be solved by fractional polynomials. For some variables, one does not even have a clue what their relationship to the response variable might be. Therefore, these semi- and nonlinear models can provide a solution. Note that by nonlinear models we mean also nonparametric models here, hence nonlinear means nonlinear in the covariates. 'Do these models perform better or worse than the widely used logit model or do they yield better results?' - that is the key question throughout this chapter.

In the previous chapter and in this, we will focus on one particular aspect of modelling activity diary data, i.e. the choice of transport mode. This chapter will concentrate on different kinds of nonlinear models. Some of these models have their background in the statistical literature (fractional polynomials), while others (support vector machines, CART) stem from the data mining/machine learning community.

Linear models (logistic regression) are well established in the field of statistical modelling. There are two different ways to extend these models, parametrically, by means of statistical models, and non-parametrically, by means of machine learning algorithms. A first parametric extension to linear models that provides more flexibility can be found in fractional polynomials. The advantage of this extension is that it offers a better prediction, because of its flexibility and it is still interpretable, just as linear models. Fractional polynomials are therefore often called a semi-linear modelling approach.

Apart from this parametrical extension, there are also other models from statistical learning theory that have proven to be very useful in other research fields. Some examples are: classification and regression trees (CART, Breiman *et al.*, 1984), support vector machines (SVM's, Vapnik, 1996), neural networks (Haykin, 1994), multivariate adaptive regression splines (MARS, Friedman, 1991), ... CART are only partly interpretable and rather unstable, two solutions for this are applied in Chapter 5. Boosting should ameliorate the accuracy, while bagging stabilises the trees. SVM's are very good for prediction purposes, though it is rather a black box approach and the results are not interpretable. Neural networks, MARS, and Bayesian networks are other possible extensions, but in this chapter, we will confine the number of extensions to fractional polynomials on the one hand and to support vector machines and CART on the other hand, since these techniques have not been explored yet in the field of mode choice modelling.

We will try to find a way to compare the performance of both types of models by means of a small adjustment on a well-known goodness-of-fit measure.

The next section discusses the semi-linear modelling approach: fractional polynomials and the nonlinear machine learning algorithm: support vector machines.

7.2 Models

7.2.1 Fractional Polynomials

Linear models have been used extensively, almost routinely, by applied statisticians and researchers (Royston and Altman, 1994). However, often is the relationship between a dependent variable and one or more continuous covariates curved. Usually, one attempts to represent curvature in regression models by means of polynomials of the covariates, typically quadratics. Cubic or higher order polynomials are used/useful only rarely. In general, low order polynomials offer a limited family of shapes, while high order polynomials may fit poorly at the extreme values of the covariates. Various attempts have been made to devise more acceptable models. Box and Tidwell (1962) developed an appropriate linearisation of each variable in a multiple regression model, though for models with more than one covariate, there are considerable difficulties in estimating the powers reliably. A cubic spline can be seen as the link between conventional polynomials and the modern methods of nonparametric smoothing. Splines are developed as for interpolation purposes (Whittaker, 1923). Later, the smoothing spline was developed as a method for fitting curves to data (Reinsch, 1967; Silverman, 1985). Nonparametric smoothers are an attempt to 'let the data show us the appropriate form' (Hastie and Tibshirani, 1990) rather than imposing a limited range of forms on the data. Nonparametric and spline smoothers are flexible and powerful tools that impose few limitations on the functional form, though the fitting process may be computationally intensive.
In this manuscript, we search for models that are reasonably flexible and more importantly, easy to understand and parsimonious.

Fractional polynomials (Royston and Altman, 1994; Royston *et al.*, 1999; Faes *et al.*, 2003; ...), an extended family of curves whose power terms are restricted to a small predefined set of values, may provide a solution for this. They provide much more flexible shaped curves than conventional polynomials, but in cases where the extension is not necessary, this family essentially reduces to conventional polynomials. A particular feature of the fractional polynomials is that they provide a wide class of functional forms, with only a small number of terms (Royston and Altman, 1994; Sauerbrei and Royston, 1999).

Let x be a continuous covariate, a *fractional polynomial of degree m* is then defined to be

$$\phi_m(x, \zeta, \mathbf{p}) = \zeta_0 + \sum_{i=1}^{m} \zeta_i H_i(x),$$

where m is a positive integer, $\mathbf{p} = (p_1, \ldots, p_m)$ a real-valued vector of powers with $p_1 \leq \ldots \leq p_m$ and $\zeta = (\zeta_1, \ldots, \zeta_m)$ coefficients. We set $H_0(x) = 1$, $p_0 = 0$ and then, for $i = 1, \ldots, m$

$$H_i(x) = \begin{cases} x^{(p_i)} & \text{if } p_i \neq p_{i-1} \\ H_{i-1}(x)\ln(x) & \text{if } p_i = p_{i-1} \end{cases}$$

The round bracket notation indicates the Box-Tidwell transformation:

$$x^{(p_i)} = \begin{cases} x^{p_i} & \text{if } p_i \neq 0 \\ \ln(x) & \text{else} \end{cases}$$

This full definition includes possible 'repeated powers' which involve powers of ln(x). E.g. a fractional polynomial of the third degree ($m = 3$) with powers (1,1,2) is of the form $\zeta_0 + \zeta_1 x + \zeta_2 x\ln(x) + \zeta_3 x^2$. Experience suggests (Royston and Altman, 1994) that $p_i \in \{-2, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, 2, \max(3, m)\}$ is sufficiently rich to cover many practical cases adequately.

In this chapter, we will consider fractional polynomials of the first and second degree applied to a multiple logistic regression setting with r continuous and s categorical covariates. For $m = 1$, this means that for every continuous covariate a fractional polynomial of degree one will be fitted. The model equation then yields:

$$\text{logit}(\pi_x) = \log\Big(\frac{\pi_x}{1-\pi_x}\Big) = \beta_0 + \sum_{k=1}^{r} \beta_k x_i^{(p_i)} + \sum_{l=1}^{s} \gamma_l z_l$$

hereby are the x-variables continuous covariates and the z-variables are categorical.

For $m = 2$, all continuous covariates will have the form of a fractional polynomial of degree two. This means that first

$$\text{logit}(\pi_x) = \beta_0 + \sum_{j=1}^{2} \beta_{1j} H_j(x_1) + \sum_{k=2}^{r} \beta_k x_k + \sum_{l=1}^{s} \gamma_l z_l$$

will be fitted. I.e. the relationship between the dependent variable and all x-variables will be taken as a straight line, except for y and x_1. At step 2, we fix the fractional polynomial function of the first variable (not the coefficients !), and repeat the same for the second variable, etc. After x_r is reached, the first iteration is completed, the model now only consists of β_0 and fractional polynomials of the second degree and it no longer contains the term $\sum_k \beta_k x_k$. In the first step of the second iteration, the constant term and the fractional polynomial models for variables $x_2, \ldots x_r$ and the categorical covariates are fixed, while a new fractional polynomial function for variable x_1 is defined, etc. This iteration continues again until variable x_r is reached. Subsequent iterations are analogous. Convergence is attained when the fractional polynomial functions for each of the continuous variables do not change from one iteration to the next. This iterative procedure used to fit a model with multiple covariates is closely related to back-fitting (see, e.g. Breiman and Friedman, 1985) and it is described in general in Section 3.5 of Royston and Altman (1994). The iterative procedure used in the following analyses to fit a model with multiple covariates follows the suggested strategy as described in Roystan and Altman (1994, p. 436–437), which has been implemented in the R package mfp. It is based on their experience that the conditional relationships between dependent variable and predictors in multiple covariate models are not often sufficiently complex to require fractional polynomials with $m > 2$. The algorithm is a type of stepwise regression, much like that described by Hastie and Tibshirani (1990, p. 260–261). When considering variable x_i at each iteration, we only consider $m_i \leq 2$. We choose $m_i = 2$ if the fit of the $\tilde{\mathbf{p}}_\mathbf{i}$-model for $m_i = 2$ is significantly better than the $\tilde{p}_i$-model for $m_i = 1$. This is called the test for simplification: test the fractional polynomial of degree two against the best fractional polynomial of degree one at alpha level ($\alpha = 0.1$) on two degrees of freedom. If this test is significant, choose $m_i = 2$, otherwise choose $m_i = 1$. Similarly for $m_i = 1$, we only choose $p_i = \tilde{p}_i$ in preference to $p_i = 1$ if $\tilde{p}_i$ is a significantly better fit according to the criterion set in the non-linearity test. This tests the fractional polynomial of degree one in x against a straight line on one degree of freedom. Finally, in the inclusion test, if $p_i = 1$ is obtained, we omit x_i (at this iteration) if the resulting increase in deviance is not statistically significant ($df = 1$). Likewise, at each step we omit z_i if the increase in deviance is not significant. Any omitted variables are retested in the next iteration. Convergence is achieved when the set of fractional polynomial functions (and omitted variables) does not change.

All significance tests are carried out using an approximate P-value calculation based on a difference in deviances (-2 × loglikelihood) having a chi-squared distribution. Therefore, each of the tests in this procedure maintains a significance level only approximately equal to select. For a given significance level, it provides some protection against over-fitting, in that it protects against choosing over complex mfp models.

7.2.2 Support Vector Machines

Support vector machines (SVM's) are learning systems that use a hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from optimisation theory that implements a learning bias derived from statistical learning theory (Hastie *et al.*, 2001). This learning strategy, that has been introduced by Vapnik and co-workers (1996), is a principled and very powerful method that in the few years since its introduction has already outperformed most other systems in a wide variety of applications.

In order to clearly understand the procedure of support vector machines, one first has to discuss the technique for constructing an *optimal separating hyperplane* between two classes that are perfectly separable by a linear boundary. Then extensions to the nonseparable case, where the classes overlap are considered and these techniques are then generalised to what is known as *support vector machines.* The SVM produces nonlinear boundaries by constructing a linear boundary in a large, transformed version of the feature space. These nonlinear boundaries then define the maximum margin hyperplane.
Now, consider N pairs or the training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, with $x_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$ (for logit models, usually $y_i \in \{0, 1\}$).
Define a hyperplane by

$$\{x : f(x) = x^T\beta + \beta_0 = 0\},$$

where β is a unit vector: $\|\beta\| = 1$. The *optimal separating hyperplane* separates the two classes and maximises the distance to the closest point from either class (Vapnik, 1996). Not only does this provide a unique solution to the separating hyperplane problem, but by maximising the margin between the two classes on the training data, this leads to better classification performance on the test data. Dropping the norm constraint on β, this hyperplane can be found by solving the following optimisation problem:

$$\begin{array}{l} \min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 \\ \text{subject to } \; y_i(x_i^T\beta + \beta_0) \geq 1, \qquad i = 1, \ldots, N. \end{array} \tag{7.1}$$

This is a convex optimisation problem, which can be solved using Lagrange multipliers α_i. The Lagrange function then equals

$$L = \min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{N} \alpha_i[y_i(x_i^T\beta + \beta_0) - 1].$$

Setting the derivatives to zero, we obtain:

$$\beta = \sum_{i=1}^{N} \alpha_i y_i x_i \tag{7.2}$$

$$0 = \sum_{i=1}^{N} \alpha_i y_i, \tag{7.3}$$

and substituting these in the previous equation, one obtains the so-called Wolfe dual

$$L = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j, \tag{7.4}$$

subject to $\alpha_i \leq 0$. In addition, the solution must satisfy the Karush-Kuhn-Tucker conditions, which include (7.2), (7.3), (7.4) and

$$\alpha_i [y_i (x_i^T \beta + \beta_0) - 1] = 0 \ \forall i. \tag{7.5}$$

From this equation, one can see that

- if $\alpha_i > 0$, then $y_i(x_i^T \beta + \beta_0) = 1$, or in other words, x_i is on the boundary of the margin;
- if $y_i(x_i^T \beta + \beta_0) > 1$, x_i is not on the boundary of the margin, and hence $\alpha_i = 0$.

The band that is $\frac{1}{\|\beta\|}$ units away from the hyperplane on either side is what is called the margin. Thus the margin indicates the distance of the support points to the hyperplane.

One can observe from (7.2) that the solution vector β is defined in terms of a linear combination of the *support points (vectors)* x_i, these are the points that are defined to be on the boundary of the margin via $\alpha_i = 0$. Likewise, β_0 is obtained by solving (7.5) for any of the support points.

This *optimal separating hyperplane* produces a function $\hat{f}(x) = x^T \hat{\beta} + \hat{\beta}_0$, for classifying new observations (from the test sample):

$$\hat{G}(x) = \text{sign} \hat{f}(x).$$

Although none of the training cases fall in the margin (by construction), this will not be the case for the test observations. Intuition learns that a large margin on the training data will lead to good separation of the test data.

When the data are not separable, there will be no feasible solution to this problem, and an alternative formulation is necessary. This alternative formulation will be provided by the *support vector machine* that allow for an overlap, but minimise a measure of the extent of this overlap.

One way to deal with this overlap is still to maximise the margin (i.e. to minimise $\|\beta\|$), but to allow for some points to be on the wrong side of the margin. Define the slack variables ξ_i. The value ξ_i in the constraint $y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i$ is the proportional amount by which the prediction $f(x_i) = (x_i^T\beta + \beta_0)$ is on the wrong side of its margin. Hence by bounding the sum $\sum \xi_i$, we bound the total proportional amount by which the predictions fall on the wrong side of their margin. A misclassification occurs when $\xi_i > 1$, thus bounding $\sum \xi_i$ at a value, say W, bounds the total number of training misclassifications at W.
Dropping again the norm constraint on β, the equivalent form of (7.1) becomes

$$\min \|\beta\| \quad \text{subject to} \quad \begin{cases} y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, \ \forall i, \\ \xi_i \geq 0, \sum \xi_i \leq \text{constant} \end{cases} \tag{7.6}$$

This is the usual way the support vector classifier is defined for the nonseparable case (Hastie *et al.*, 2001; Cristianini and Shawe-Taylor, 2000). By the nature of the criterion (7.6), one can observe that points well inside their class margin will not play a big role in shaping the boundary. We can re-express (7.6) again by means of the Lagrange multipliers α_i, μ_i and γ, and the Lagrange function becomes:

$$L = \frac{1}{2}\|\beta\|^2 + \gamma \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i[y_i(x_i^T\beta + \beta_0) - (1 - \xi_i)] - \sum_{i=1}^N \mu_i \xi_i.$$

Again, setting the respective derivatives equal to zero, one obtains

$$\beta = \sum_{i=1}^N \alpha_i y_i x_i, \tag{7.7}$$

$$0 = \sum_{i=1}^N \alpha_i y_i, \tag{7.8}$$

$$\alpha_i = \gamma - \mu_i, \ \forall i, \tag{7.9}$$

as well as the positivity constraints, $\alpha_i, \mu_i, \xi_i \geq 0, \ \forall i$. Substituting (7.7) - (7.9) in the Lagrange function, we obtain again the Wolfe dual function

$$L = \sum_{i=1}^N \alpha_i - \frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j x_i^T x_j.$$

The solution must, in addition to (7.7) - (7.9), also satisfy the Karush-Kuhn-Tucker conditions, that also include the following constraints:

$$\begin{aligned} \alpha_i[y_i(x_i^T\beta+\beta_0)-(1-\xi_i)] &= 0, \qquad (7.10)\\ \mu_i\xi_i &= 0, \qquad (7.11)\\ y_i(x_i^T\beta+\beta_0)-(1-\xi_i) \geq 0, & \qquad (7.12) \end{aligned}$$

for $i = 1, \ldots, N$. Together these equations (7.7) - (7.12) define a unique solution to the optimisation problem (7.6).

From (7.7), one can observe that the solution for β looks like

$$\hat{\beta} = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i,$$

with nonzero coefficients $\hat{\alpha}_i$ only for those observations i for which the constraints in (7.12) are exactly met (due to (7.10)). These observations are called the *support vectors*, since $\hat{\beta}$ is represented in terms of them alone. Among these support vectors, some of them will lie on the edge of the margin ($\xi_i = 0$), and hence from (7.11) and (7.9) will be characterised by $0 < \alpha_i < \gamma$; the remainder ($\xi_i > 0$) have $\alpha_i = \gamma$. From (7.10), one can see that any of these margin points ($0 < \alpha_i, \xi_i = 0$) can be used to solve for β_0, and one typically uses an average of all these solutions for numerical stability.

Given the solutions $\hat{\beta}_0$ and $\hat{\beta}$, the decision function (*the support vector classifier*) can be written again as

$$\begin{aligned} \hat{G}(x) &= \text{sign } [\hat{f}(x)] \\ &= \text{sign } [x^T\hat{\beta}+\hat{\beta}_0]. \end{aligned}$$

The tuning parameter of this procedure is γ. Larger values of γ focus attention more on (correctly classified) points near the decision boundary, while smaller values involve data points further away. Either way, misclassified points are given weight, no matter how far away.

The *support vector classifier* that has been described so far, finds linear boundaries in the input feature space. As with other linear methods, one can make the procedure more flexible by enlarging the feature/covariate space using basis expansions such as polynomials or splines. Generally linear boundaries in the enlarged space achieves better training-class separation, and translate to nonlinear boundaries in the original space. Once the basis functions $h_m(x), m = 1, \ldots, M$ are selected, the procedure is the same as before. One fits the SV classifier using input features $h(x_i) = (h_1(x_i), \ldots, h_M(x_i)), i = 1, \ldots, N$ and produces the (nonlinear function $\hat{f}(x) = h(x)^T\hat{\beta} + \hat{\beta}_0$. The classifier is $\hat{G}(x) = \text{sign } [\hat{f}(x)]$ as before.

The *support vector machine* is an extension of this idea, where the dimension of the enlarged space is allowed to get very large, infinite in some cases. It might seem that the computations become prohibitive. We can represent the latter optimisation problem and its solution in a special way that only involves the input features via inner products. The Wolfe dual function then has the form

$$L = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle h(x_i), h(x_j) \rangle. \tag{7.13}$$

The solution can now be written as

$$\begin{aligned} f(x) &= h(x)^T \beta + \beta_0 && (7.14) \\ &= \sum_{i=1}^{N} \alpha_i y_i \langle h(x), h(x_i) \rangle + \beta_0. && (7.15) \end{aligned}$$

As before, given α_i, β_0 can be determined by solving $f(x_i) = 0$ in (7.15) for all x_i for which $0 < \alpha_i < \gamma$.

As can be observed, both (7.13) and (7.15) involve $h(x)$ only through inner products. Thus, the transformation $h(x)$ does not need to be specified, only the knowledge of the kernel function

$$K(x, x') = \langle h(x), h(x') \rangle$$

that computes the inner product in the transformed space is required. Note that K should be symmetric positive (semi-) definite function. Three popular choices for K in the SVM literature, that are also implemented in the R package $e1071$, are

d-th degree polynomial: $K(x, x') = (\kappa + \gamma \langle x, x' \rangle)^d$,
Radial basis: $K(x, x') = \exp(-\frac{\|x - x'\|^2}{\gamma})$,
Neural network: $K(x, x') = \tanh(\kappa_1 \langle x, x' \rangle + \kappa_2)$.

The role of the parameter γ is clearer in an enlarged feature space, since perfect separation is typically achievable there. A large value of γ will discourage any positive ξ_i, and lead to an over-fit wiggly boundary in the original feature space. A small value of γ will encourage a small value of $\|\beta\|$, which in turn causes $f(x)$ and hence the boundary to be smoother. (Often a value of 1 for γ seems to work fine in practice.)

7.2.3 Classification And Regression Trees (CART)

This subsection will be very brief, since CART has already been introduced in Chapter 6 in the subsection on recursive partitioning. In general, it describes a way of partitioning the parameter space, just as C4.5 is manner of recursive partitioning (see also Chapter 4). The general settings of C4.5 and CART are thus very much alike. The main difference in partitioning, as discussed in Chapter 6, is that C4.5 allows for multi-way splits and CART (Breiman *et al.*, 1984) only allows for binary splits to be conducted.
Classification trees are used for categorical dependent variables, while regression trees are applied to continuous y-variables. In this chapter, only classification trees are considered. The Gini-index (for multiclass responses) $i(\tau) = \sum_{j \neq i} P(Y = j|\tau)P(Y = i|\tau),\ (i, j \in \{1, \ldots, J\})$ or the entropy measure (for binary responses, see Chapter 6) can be used as a splitting criterion. Several methods exist for controlling the tree: the minimum number of observations that must exist in a node in order for a split to be attempted (this is set to 20 in the performed analyses), the minimum number of observations in any terminal leaf node (usually about a third of the minimum split size), the maximum depth of the tree, etc. For more detail, we refer to Breiman *et al.* (1984).

7.3 Data and Model Comparison

7.3.1 The Data

The data sets used in this chapter are derived from the Albatross data set (Dutch data, see Chapter 2, Section 2.3): transport mode for work and from the Southeast Florida data set as described in Chapter 2, Section 2.4. For both data sets, only a limited number of covariates has been considered. This was done, on the one hand, because of practical issues (computational complexity: the magnitude of the data set combined with a large number of explanatory variables can cause problems for obtaining the results in a reasonable amount of time (see e.g. in the analyses on the Southeast Florida data set in the next section); too many variables with missing values, etc.). But, on the other hand, Chapter 4 learned us also that some variables can be irrelevant and that they can disturb the performance of the models. Thus, respectively the 20 and 17 'best' (according the the Relief-F feature selection algorithm) variables have been chosen to perform the analyses.

The first data set (Dutch data) contains 1025 observations. It has been split into two different data sets: one for the prediction of using ‘slow’ transport in order to go to work and one for the prediction of using public transport. Both being a car passenger and all different kinds of public transport are considered as public transport in this case. For the Southeast Florida data set, that contains 14527 cases, the same split into slow and public transport has been made, and here a car passenger, taking the bus, metro-rail, metro-mover, tri-rail, jitney, school bus and taxi are considered as public transport. The distribution of the different transport modes over the data sets can be found in Table 7.1.

Table 7.1: Distribution of different transport modes over the data sets

	Number of cases	Slow	Public	Car driver	Other
Dutch data	1025	18.93%	12.29%	68.78%	0.00%
Southeast Florida	14527	2.71%	9.67%	87.13%	0.49%

Table 7.2 shows for each of the three Dutch data sets which twenty selected variables are used to determine the three possible transport modes. For a description of the variables, we refer to Chapter 2.

Table 7.2: Selection of the different variables for the three Dutch the data sets

Variable	Public	Slow	Car Driver	Variable	Public	Slow	Car Driver
x_1	*	*	*	x_{22}	*		
x_3	*	*	*	x_{23}	*	*	*
x_5	*	*		x_{24}			*
x_6	*	*	*	x_{25}	*	*	*
x_7	*	*	*	x_{26}	*	*	*
x_{10}	*	*	*	x_{27}	*	*	*
x_{14}	*	*	*	x_{30}	*	*	*
x_{15}	*	*	*	x_{31}		*	*
x_{16}	*	*	*	x_{32}		*	*
x_{19}	*			x_{33}		*	*
x_{20}	*	*	*	x_{38}	*	*	*
x_{21}	*			x_{39}	*	*	*

In the analyses of the Southeast Florida data sets, variable v_{17} is completely determined by $v_{14} - v_{16}$ and v_{13} is a simplification of v_9, thus either one of the variables will be used in the analyses, while v_{17} will, of course, be left out to ensure the validity of the model.

All data sets are split in a training and a test set. The training set contains a random sample comprising 70% of the total data set, while the remaining 30% makes up the test set, as commonly used in practice. The training set will be used to build the model, the test set is used for validation.

7.3.2 How to Compare Non-Parametric Machine Learning Algorithms and Statistical Parametric Models?

In order to make an 'honest' comparison between the results of the machine learning algorithm (that only provide accuracies as a result of the classification) and the semi-linear models, we have to come up with some kind measure that can be used for both machine learning and for statistical models.

Two classical diagnostics that are often used in logistic regression are sensitivity and specificity. The *sensitivity* is defined as probability that a positive case ($y = 1$) is predicted, given that it is observed, hence

$$\text{Sensitivity} = P(y_{predicted} = 1 | y_{observed} = 1).$$

In the same way, the *specificity* is the conditional probability on a negative ($y = 0$) predicted case, given that the observed case is also negative. The *prevalence* is determined by the number of positive observed cases respective to the total number of cases, such that the accuracy can be written as:

$$\text{Accuracy} = \text{Prevalence} \times \text{Sensitivity} + (1 - \text{Prevalence}) \times \text{Specificity}.$$

Thus, if the parametric models can be turned into some sort of classification, we can use the above equation as well. So define

$$\begin{aligned} y_{i,predicted} &= 1 \text{ if } \pi_i \geq \text{cut-off} \\ &= 0 \text{ if } \pi_i < \text{cut-off.} \end{aligned}$$

As stated in Neter *et al.* (1996) using 0.5 as cut-off does not always work for the best. Using 0.5 is only appropriate when it is equally likely in the population of interest that outcomes zero and one will occur. This certainly is not the case here. Therefore, since we do not know whether the data set is a random sample from the population in Southeast Florida, Neter *et al.* (1996) suggest to take the following proportion as a cut-off value

$$\text{cut-off} = \frac{\sum_{i=1}^{N_{\text{train}}} y_i}{N_{\text{train.}}},$$

i.e. the proportion of ‘successes’ at the training set. This can be seen as a Bayesian estimate (the use of prior information) for the number of ‘successes’.

For the purpose of information, the Akaike Information Criterion (AIC) (Akaike, 1973) and the Bayesian Information Criterion (BIC) (Schwartz, 1978) are also added for parametrical models. Let p be the number of variables used in the model, N be the number of cases in the data set and L be the likelihood function, then

$$\begin{aligned} AIC &= -2\log L + 2 \times p \\ BIC &= -2\log L + \log(N) \times p \end{aligned}$$

For the application to mode choice models, the sensitivity can be regarded as the most important diagnostic of the three. The fact is that the main purpose of the models in the next section is the prediction of the positive cases. That the negative cases are predicted well comes in handy, though the aim is on the prediction of the positive cases. Suppose that the data set is very skewed and that you have a rather low prevalence (as is the case in the following data sets). If most negative cases are predicted well, but none of the positive cases is correctly predicted, then you have a very high specificity and accuracy, but this was not the intention of the model.

7.4 Results

In this chapter, our focus is on the application of semi- and nonlinear models in the context of mode choice models. The use of fractional polynomials, as a parametrical extension to linear models will be investigated on the one hand, and on the other, the use of two non-parametrical techniques, support vector machines and classification and regression trees, will be explored as well. The performance of these models will be compared to that of the widely used logit model (see Moons *et al.*, 2004b, 2004c).

For the parametric models, ordinal categorical variables have been considered as nominal (saturated main effects model) as well as continuous. Each time, the best results are presented in the next subsections.

For the support vector machines, the kernels used in training and predicting are those kernels described in Section 7.2: the linear optimal separating hyperplane, the polynomial kernel of degree 3 (to provide some extra nonlinearity compared to the linear hyperplane), the radial basis kernel and finally, the neural net kernel.

7.4.1 Dutch Data: Public Transport

Parametric Models

The multiple linear logistic regression model, expressing the saturated main effects, has an AIC value of 536.59 and a BIC value of 719.65. The sensitivity on the training set equals 0.739, the specificity 0.703 and the accuracy 0.708 (cut-off value of 0.128). On the test set, the sensitivity even increases up to 0.824, the specificity yields 0.648 and the accuracy 0.668. The final multiple fractional polynomial (mfp) model (also considering ordinal variables as nominal) is defined by:

$$\begin{aligned}\text{logit}(\pi(x)) &= \beta_0+\beta_1(x_6=2)+\beta_2(x_7=2)+\beta_3(x_{10}=3)+\beta_4(x_{16}=1)+\beta_5(x_{27}=1)\\ &\quad+\beta_6(x_{27}=2)+\beta_7x_{30}^3+\beta_8\ln(x_{30})x_{30}^3+\beta_9(x_{39}=1)\end{aligned}$$

The continuous variables x_{19}, x_{21} and x_{22} do not appear in this final model, while covariate x_{30} has repeating powers (3,3) for its fractional polynomial.
The parameter estimates and their standard errors can be found in Table 7.3. Taking the ordinal variables as a single linear effect does not improve the model.

Table 7.3: Parameter estimates of the semi-linear model on Dutch data - Public

Parameter	Estimate	Standard error
β_0	-2.906	0.290
β_1	0.828	0.265
β_2	-0.997	0.263
β_3	-1.517	1.087
β_4	-1.501	0.466
β_5	1.814	0.486
β_6	0.775	0.251
β_7	1.066×10^{-6}	0.220
β_8	-1.373×10^{-6}	0.336
β_9	0.629	0.276

One can observe that, when all remaining variables are kept constant, the probability of choosing public transport to go to work is higher if the traveller is a women, if the activity is conducted with other people and if the sum of the duration of the activities plus the minimum public transport travel time is less or equal then the maximum duration. On the other hand, the probability of choosing public transport to go to work decreases when 3 or 4 non-work out-home activities are conducted during the activity pattern, when there are 1 or more cars available per adult or if there is a bring/get activity in the activity pattern.

Nonparametric Models

The linear optimal separating hyperplane needs 380 support vectors to make up its classifier. It automatically classifies each case to the outcome value zero, this leads to an accuracy of 0.872 on the training set and 0.889 on the test set. Note that the sensitivity in this case equals 0.000 and the specificity 1.000.

The polynomial kernel of degree 3 requires 199 support vectors to make up the classification, and it almost makes a perfect separation on the training set (accuracy of 0.999), while on the test set, it leads to an accuracy of 0.801. The sensitivity of the training is equal to 1.000 and the specificity yields 0.998. One can observe that a third degree of the polynomial kernel probably leads to over-fitting on the training set, since the sensitivity on the test set has decreased to 0.294 and the specificity to 0.864.

When the radial kernel is chosen for the support vector machine, we end up with 320 support vectors and a specificity of 0.033, an accuracy of 0.876 on the training data. The results on the test set indicate a zero sensitivity and an accuracy of 0.889. It means here that 3 positive cases of the training set are better classified, when compared to the linear kernel.

Finally, the neural net kernel takes only 185 support vectors in order to make the classification, but the accuracy on training (0.845) and test set (0.857) are less when compared to the other support vector machines.

In a second example of nonparametric models, we focus on the classification tree on this data set. The CART algorithm results in a final tree of depth eight with nine final nodes. The detailed tree can be found in Figure 7.1. Variable x_{30} appears to be the only common variable in the best fractional polynomial model and in the resulting tree here. The accuracy on the training set is 0.900 and on the test set 0.863. The sensitivity on the training set (0.293) indicates that at least some positive cases are correctly predicted, the value on the test set equals 0.118.

All results are summarised in Table 7.4. Note that here the simple linear logistic regression model provides the best result. The Polynomial SVM comes in second best, despite the obvious over-fitting on the training data. The non-parametrical models do not perform well. This is because they are trained on accuracy, and not on sensitivity.

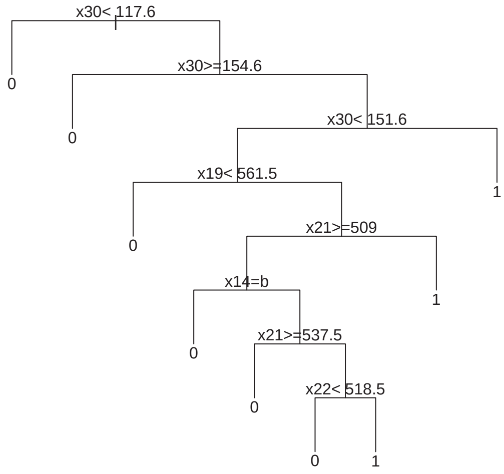

Figure 7.1: *Final tree on Dutch public transport data*

Table 7.4: Performance values for the models on Dutch data - Public Transport

	accuracy		sensitivity		specificity	
	training set	test set	training set	test set	training set	test set
Linear	0.708	0.668	0.739	0.824	0.703	0.648
Mfp	0.699	0.847	0.652	0.029	0.706	0.949
SVM - Linear	0.872	0.889	0.000	0.000	1.000	1.000
SVM - Polynomial	0.999	0.801	1.000	0.294	0.998	0.864
SVM - Radial basis	0.876	0.889	0.033	0.000	1.000	1.000
SVM - Neural net	0.845	0.857	0.011	0.000	0.968	0.963
CART	0.900	0.863	0.293	0.118	0.989	0.956

This bad performance is also due to the skewness of the data set, only 12.29 % of the total sample takes public transport as the travel mode to go to work.

7.4.2 Dutch Data: Slow Transport

Parametric Models

There was no difference in the final multiple fractional polynomial model with the ordinal categorical variables taken as nominal compared to continuous attributes. Therefore, the results of the linear regression model presented here, are those with the ordinal variables considered as continuous-type covariates. The value of Akaike's Information Criterion of the linear model is equal to 455.93 and the value of the Bayesian Information Criterion yields 606.95. The accuracy on the training set holds 0.843 (the cut-off value is equal to 0.199), while on the test set it is 0.847. Though, more important is the sensitivity, on the training set, the value equals 0.853 and on the test set 0.882. Considering the final multiple fractional polynomial model, the accuracy and the sensitivity on the test set do not change, while on the training set, both decrease a little bit, the accuracy to 0.834 and the sensitivity to 0.839. The AIC value, on the other hand, decreases up to 428.02, indicating a better model fit, and the BIC value of 473.78 only confirms this.

The final multiple fractional polynomial model is determined by:

$$\begin{aligned} \text{logit}(\pi(x)) &= \beta_0 + \beta_1 x_7 + \beta_2(x_{14} = 1) + \beta_3(x_{15} = 1) + \beta_4(x_{16} = 1) + \beta_5(x_{26} = 2) \\ &\quad + \beta_6(x_{27} = 1) + \beta_7 x_{30} + \beta_8(x_{38} = 1) + \beta_9(x_{39} = 1) \end{aligned}$$

Only the continuous covariate x_{30} appears in this model, and then even as a single linear effect, no fractional polynomials are necessary in this case. The ratio's between the travel time of car, slow transport and public transport do not seem to have an important impact in predicting the usage of slow transport modes.

The difference with the linear model lies in the fact that in the mfp model non-significant variables have been deleted due to the iterative procedure as described in Section 7.2.

The parameter estimates and corresponding standard errors can be found in Table 7.5. All ordinal variables are taken as single linear effects.

When interpreting this model, it turns out the the probability of choosing a slow transport mode (walk/bike) to go to work decreases if there is one or more cars per adult in the household, if there is at least one shopping or service activity in the activity pattern. The same is true for social/leisure out-home activities, while on the other hand a bring/get activity in the activity pattern increases the probability of slow transport, except when this activity is the first in the concerned tour. The probability decreases also when the activity is pursued with other members of the household. If

Table 7.5: Parameter estimates of the semi-linear model on Dutch data - Slow

Parameter	Estimate	Standard error
β_0	1.730	0.475
β_1	-1.400	0.285
β_2	-1.117	0.378
β_3	-0.760	0.288
β_4	1.177	0.477
β_5	-1.664	1.154
β_6	-1.453	0.778
β_7	-0.031	0.450
β_8	2.152	0.327
β_9	-1.138	0.325

the shortest travel time by bike of the tour increases, logically, the probability of using slow transport for the tour will decrease. If the sum of the duration of the activities in the tour plus the minimum bike travel time is less or equal then the maximum duration of the tour, then the probability increases, while the opposite occurs when the instead of the minimum bike travel time, the minimum public transport travel time is considered.

Nonparametric Models

The linear optimal separating hyperplane is made up based on 209 support vectors. This time it performs better then only classifying each case to the majority class, leading to a sensitivity of 0.608 and an accuracy of 0.876 on the training data set and on the test set the values yield 0.549 and 0.863 respectively.

The polynomial kernel of degree three needs two support vectors less, but again one can observe some over-classification on the training set. The sensitivity on the training set is again equal to 1.000, while on the test set merely the value of 0.588 can be achieved. We have here an accuracy of 0.999 compared to 0.824 on the test. It does not even attain the level of the simple linear classifier there.

Applying the radial kernel on the support vector machine leads to an accuracy of 0.918 on training and of 0.880 on the test set. 69.2 % of the positive cases are correctly predicted on the training data, while on the test data about half of them (0.510) are correctly predicted. It needs 297 support vectors to make up the margin.

The neural net kernel with parameters $\kappa_1 = 0$ and $\kappa_2 = 0.05$ takes 213 support

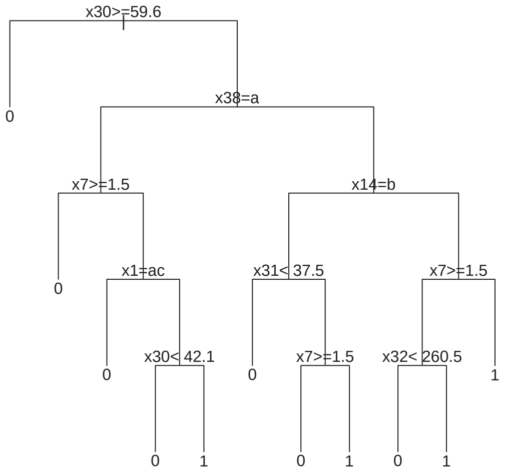

Figure 7.2: *Final tree on Dutch slow transport data*

vectors to end up with an accuracy of 0.805 on the training set, a rather moderate performance, when compared to the others, while on the test set, it provides an accuracy of 0.863, exactly the same as the linear optimal hyperplane. The former, however, does a better job in predicting the positive cases correctly. If we take a look at the sensitivity on the test data, the value equals 0.569 compared to 0.549 in the linear optimal hyperplane.

The second nonparametric model, the classification tree algorithm (again, with ordinal categorical variables regarded as if they were continuous), results in a tree with eleven final nodes and of depth five. This final tree is shown in Figure 7.2.

The sensitivity on the test data is the same as that of the linear optimal hyperplane, on the training data, CART performs better. The accuracy on the training set equals 0.897, while on the test set it yields 0.860. Variables x_{14}, x_{30} and x_{38} are common in the nonparametric CART model and in the parametric mfp model.

To get a clear overview of all results on this data set, a summary is provided in Table 7.6.

Table 7.6: Performance values for the models on Dutch data - Slow Transport

	accuracy		sensitivity		specificity	
	training set	test set	training set	test set	training set	test set
Linear	0.843	0.847	0.853	0.882	0.840	0.840
Mfp	0.834	0.847	0.839	0.882	0.833	0.840
SVM - Linear	0.876	0.863	0.608	0.549	0.943	0.926
SVM - Polynomial	0.999	0.824	1.000	0.588	0.998	0.871
SVM - Radial basis	0.918	0.880	0.692	0.510	0.974	0.953
SVM - Neural net	0.805	0.863	0.476	0.569	0.887	0.922
CART	0.897	0.860	0.678	0.549	0.951	0.922

7.4.3 Dutch Data: Car Driver

We added the analysis on the Car Driver data set in this chapter, since Chapter 6 indicates which final interaction model is not rejected by the tree-based goodness-of-fit test. The performance of this interaction model on training and test set can be found in the final table of this subsection.

Parametric Models

Again, the best results are obtained, when the ordinal categorical variables are considered as continuous variables, we will thus do so in the mfp analysis as well as in the CART analysis for the nonparametric models. The model comparison criteria for the linear model comprise: an AIC of 742.75, a BIC of 902.92 and an accuracy of 0.731 and a sensitivity of 0.747 on training and respectively 0.733 and 0.730 on the test set (the cut-off value here is 0.673). Note that the cut-off value does not exactly agree with the distribution percentages of the transport modes over the data set as provided in Table 7.1. These values do not agree completely, because the cut-off value is based on the values of the training set alone, which is a random sample of the total data set.

The multiple fractional polynomial model looks like:

$$\begin{aligned}\text{logit}(\pi(x)) &= \beta_0 + \beta_1(x_6 = 2) + \beta_2 x_7 + \beta_3(x_{14} = 1) + \beta_4(x_{24} = 1) + \beta_5(x_{26} = 2) \\ &\quad + \beta_6(x_{26} = 8) + \beta_7 x_{30}^{-1} + \beta_8 x_{30}^{3} + \beta_9 x_{31} + \beta_{10} x_{33}^{-1} + \beta_{11} x_{33}^{-\frac{1}{2}} \\ &\quad + \beta_{12}(x_{38} = 1) + \beta_{13}(x_{39} = 1)\end{aligned}$$

A fractional polynomial with powers (-1,3) was obtained for variable x_{30}, x_{31} just has a single linear effect, x_{32} apparently has no effect, while again a fractional polynomial this time with powers $(-1, -1/2)$ was necessary for variable x_{33}.

The parameter estimates and their respective standard errors can be found in Table 7.7.

Table 7.7: Parameter estimates of the semi-linear model on Dutch data - Car Driver

Parameter	Estimate	Standard error
β_0	-13.410	2.944
β_1	-0.594	0.232
β_2	1.549	0.222
β_3	0.797	0.231
β_4	0.819	0.461
β_5	1.306	0.743
β_6	-1.471	0.590
β_7	4.321	7.155×10^{-3}
β_8	1.529×10^{-7}	6.212×10^{-2}
β_9	-7.356×10^{-2}	0.161
β_{10}	-1.018×10^{4}	1.872
β_{11}	7.676×10^{2}	4.727
β_{12}	-1.338	0.305
β_{13}	0.701	0.342

According to these estimates, when all other variables are kept constant, the probability of using the car as transport mode to go to work is lower amongst females, when compared to men. The probability increases with the number of cars per adult, if there is at least one shopping or service activity in the activity pattern, at least one social or leisure out-home activity in the concerned tour, if the first activity of the tour is a bring/get activity and if the travel time by bike increases. The probability decreases when the first activity of the tour is not really determined, if the ratio between car and bike travel time increases, if the ratio between public transport and car travel time decreases and if the minimum sum of the duration of the activities in the tour plus the minimum bike travel time is less or equal to the maximum duration of the concerned tour, whereas it increases if the minimum sum of the duration of the activities in the tour plus the minimum public transport travel time is less or equal to the maximum duration of the tour.

This mfp model has an AIC of 665.58 and a BIC of 729.65 which are considerably less than these values for the linear model. The mfp model also shows a better performance, when one takes a look at the sensitivity: 0.795 on the training and 0.811 on the test set. The accuracy on the training set yields 0.769 and 0.788 on the test set.

The best interaction model according to the goodness-of-fit test in Chapter 6 incorporates forty-one terms. It has an AIC of 666.23 and a BIC of 858.44. Both values lay in between the criteria value for the linear and the mfp models. The accuracies slightly outperform the mfp model (0.790 on training and 0.792 on test set), conversely, the number of covariates used is more than three times the number used in the mfp model. The sensitivity of the training set is exactly the same as that of the mfp model, though on the test set, a value of 0.788 indicates a somewhat lower performance here, but it is still better than the linear regression model.

Nonparametric Models

336 support vectors are necessary to make up the linear optimal separating hyperplane. It performs rather well in this context, leading to a sensitivity of 0.936 and an accuracy of 0.815 on the training set and respective measure of 0.928 and 0.831 on the test set.

The support vector machine that tries to find a classification of the cases by means of a polynomial kernel of the third degree is based on 335 support vectors. The resulting accuracy, specificity and the sensitivity on the training set equal all three 1.000, indicating something similar as in the previous two data sets, probably an over-classification on the training set. As a consequence, the accuracy on the test is lower, even when compared to all other means of classifying, i.e. 0.723. The sensitivity of the test set yields 0.793.

The radial kernel support vector machine generally needs the highest number of support vectors. 450 support vectors are necessary here and with the γ-parameter set to 0.019, the SVM leads to an accuracy of 0.859 on the training and 0.831 on the test set. A sensitivity of 0.973 on training and of 0.946 on test set depict that this SVM is the best model so far.

The neural net kernel SVM produces the following results: a training-accuracy of 0.730 and a test-accuracy of 0.792 are the outcome based on 343 support vectors. 86.5% of the positive cases are correctly predicted on the training data, while even 90.1% is correctly predicted on the test set.

The resulting classification tree has nine final nodes and a depth of six. The recursive partitioning of the data set can be found in Figure 7.3. Variable x_1 is the only variable that appears in the classification tree and not in the final multiple fractional polynomial model.

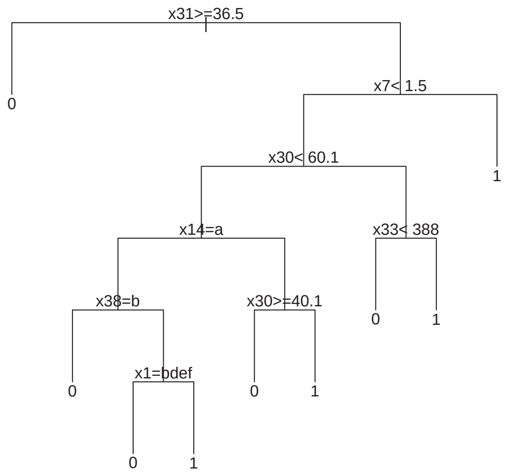

Figure 7.3: *Final tree on Dutch car driver data*

The performance of this classification tree can be situated in between the parametric models and the SVM. The training-sensitivity equals 0.915 and on the test data the sensitivity equals 0.864. The accuracy on the training set of the classification based on this final tree equals 0.813 and on the test set it yields 0.769.

All results of the different classifications are summarised in Table 7.8.

Table 7.8: Performance values for the models on Dutch data - Car Driver

	accuracy		sensitivity		specificity	
	training set	test set	training set	test set	training set	test set
Linear	0.731	0.733	0.747	0.730	0.698	0.741
Mfp	0.769	0.788	0.795	0.811	0.715	0.729
Interaction	0.790	0.792	0.795	0.788	0.779	0.800
SVM - Linear	0.815	0.831	0.936	0.928	0.566	0.576
SVM - Polynomial	1.000	0.723	1.000	0.793	1.000	0.541
SVM - Radial basis	0.859	0.831	0.973	0.946	0.626	0.529
SVM - Neural net	0.730	0.792	0.865	0.901	0.451	0.506
CART	0.813	0.769	0.915	0.864	0.604	0.518

7.4.4 Southeast Florida: Public Transport

For the analyses on this data set, all ordinal variables are used as a single linear effect in models described below.

Parametric Models

Using v_{13} instead of v_9 to make the model more parsimonious, appears to be a good strategy, since the AIC and BIC of these models are lower. The linear model has an AIC of 5764.89 and a BIC of 5902.20. The cut-off value equals 0.096 leading to a accuracy of 0.689 on the training and 0.691 on the test set. The sensitivity on the training data equals 0.662 and it decreases a little on the test set to 0.649.

The final multiple fractional polynomial model (which took more than 2.5h CPU time to be processed!) yields an AIC of 5291.68 and a BIC value of 5428.99. The accuracy on the training set is 0.747 and on the test set it is 0.903. The sensitivity on the training data is higher than the linear regression model (0.694), but on the test set none of the positive cases has been predicted correctly. The final mfp model is specified by:

$$\begin{aligned}
\text{logit}(\pi(x)) \quad = \quad & \beta_0 + \beta_1 v_1^{-2} + \beta_2 v_1 + \beta_3 v_2 + \beta_4 v_3 + \beta_5 v_4 + \beta_6 v_5 + \beta_7 v_6^{\frac{1}{2}} \\
& + \beta_8 v_6^2 + \beta_9 v_7^{-\frac{1}{2}} + \beta_{10} v_7^{-\frac{1}{2}} \times \ln(v_7) + \beta_{11} v_{10}^{-2} + \beta_{12} \ln(v_{10}) \\
& + \beta_{13}(v_{13} = 3) + \beta_{14}(v_{13} = 4) + \beta_{15}(v_{13} = 5) \\
& + \beta_{16}(v_{14} = 1) + \beta_{17}(v_{15} = 1) + \beta_{18}(v_{16} = 1)
\end{aligned}$$

v_1, v_6, v_7 and v_{10} are the continuous covariates. The fractional polynomial of degree 2 of v_1 has powers (-2,1), that of $v_6(1/2, 2)$, of $v_7(-1/2, -1/2)$ and of v_{10} (-2,0).
The parameter estimates and standard errors of this semi-linear model are provided in Table 7.9.

Table 7.9: Parameter estimates of the semi-linear model on Southeast Florida Public Transport

Parameter	Estimate	Standard error
β_0	15.610	1.387
β_1	-1.098	2.046×10^{-3}
β_2	0.176	0.623
β_3	-0.107	5.144×10^{-2}
β_4	-0.386	6.425×10^{-2}
β_5	0.153	5.602×10^{-2}
β_6	-0.726	5.725×10^{-2}
β_7	-0.753	0.250
β_8	3.885×10^{-3}	7.301×10^{-2}
β_9	-30.950	1.137
β_{10}	-45.720	1.246
β_{11}	1.448×10^{-2}	2.994×10^{-9}
β_{12}	0.261	3.308×10^{-2}
β_{13}	0.798	0.181
β_{14}	0.543	0.179
β_{15}	0.854	0.114
β_{16}	-0.693	0.119
β_{17}	-0.678	9.891×10^{-2}
β_{18}	-0.373	0.104

The probability of using public transport in Southeast Florida increases with a growing household size, with the number of driving licenses, with an increasing duration of the activity and if the activity type is maintenance, leisure or other, with respect to a home activity. On the other hand, the probability of using public transport decreases with an increasing number of employed people in the household, with an increasing number of children, with an increasing number of cars, with increasing income and age and if the activity is pursued in the AM or PM peak or at mid-day.

Nonparametric Models

The CPU time needed for obtaining these nonparametric models is very high. For the polynomial kernel of degree two and three, we were not able to obtain the results in three weeks time! We therefore skipped these analyses.

The linear optimal separating hyperplane can be formed using 2469 support vectors, and here it appears to be nothing more than the majority class or zero R classifier. The accuracy on the training set equals thus 0.904 and on the test set 0.903. The sensitivity on training and test set equal thus 0.000.

The support vector machine based on the radial kernel takes 7162 support vector machines to accomplish a performance of 0.943 on the training and 0.905 on the test set. 40.8 % of the positive cases on the training data is correctly predicted, but only 5.2 % on the test set.

Finally, the neural net kernel support vector machine only needs 1962 support vectors to build the classifier, but its accuracies do not reach further than those of the linear hyperplane.

To round off this series of nonparametric models, the performance of the classification tree on this data set will be investigated. As it turns out, the tree is only of depth one, with 2 final nodes. Figure 7.4 shows that one split has been made on the age of the person. This final tree resulted in an accuracy of 0.912 on the training set and of 0.911 on the test set. The sensitivity of the tree on the validation set was 0.097.

All results are summarised in Table 7.10.

Table 7.10: Performance values for the models on Southeast Florida - Public Transport

	accuracy		sensitivity		specificity	
	training set	test set	training set	test set	training set	test set
Linear	0.689	0.691	0.662	0.649	0.691	0.695
Mfp	0.747	0.903	0.694	0.000	0.753	1.000
SVM - Linear	0.904	0.903	0.000	0.000	1.000	1.000
SVM - Radial basis	0.943	0.905	0.408	0.052	1.000	0.997
SVM - Neural net	0.904	0.903	0.000	0.000	1.000	1.000
CART	0.912	0.911	0.095	0.097	0.999	0.998

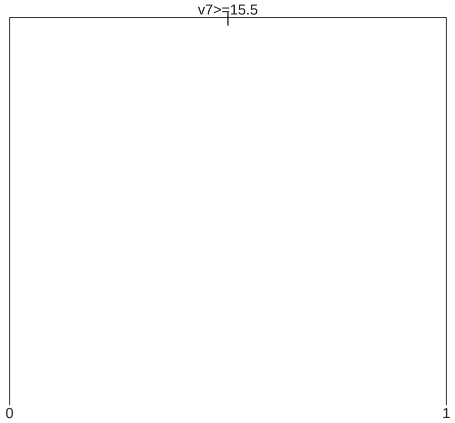

Figure 7.4: *Final tree on Southeast Florida public transport data*

7.4.5 Southeast Florida: Slow Transport

Parametric Models

The 'best' parametric models on this data set are obtained again by considering single main effects for the ordinal categorical variables and by using v_{13} instead of v_9.

The linear logistic regression model takes a value of 2249.38 on Akaike's Information Criterion and a value of 2386.70 on Schwartz Criterion (BIC). The cut-off value for determining the accuracies is 0.026, while the accuracy on the training set itself yields 0.690 and on the test set 0.693. The sensitivity on the training data yields 0.634 and on the test data 0.627.

The multiple fractional polynomial setting ends up with an AIC of 2236.89 and a BIC of 2330.84. The training set's sensitivity is a little bit lower (0.604), but on the test set, it has improved (0.659). The accuracy on the training set is equal to 0.694 and that of the test set 0.617. The mfp model itself is set by:

$$\begin{aligned}\text{logit}(\pi(x)) &= \beta_0 + \beta_1 v_1 + \beta_2 v_2 + \beta_3 v_3 + \beta_4 v_4 + \beta_5 v_5 + \beta_6 v_7^{-2} \\ &\quad + \beta_7 v_7^{-2} \times \ln(v_7) + \beta_8(v_8 = 2) + \beta_9 \ln(v_{10}) + \beta_{10}(v_{12} = 1) \\ &\quad + \beta_{11}(v_{13} = 4) + \beta_{12}(v_{15} = 1)\end{aligned}$$

Fractional polynomials of degree one appear to be necessary for the variables v_1 and v_{10}, the former with power one, the latter with power zero. Covariate v_7 takes a

fractional polynomial of degree 2 with powers (-2,-2).
The parameter estimates and their standard errors can be found in Table 7.11.

Table 7.11: Parameter estimates of the semi-linear model on Southeast Florida Slow Transport

Parameter	Estimate	Standard error
β_0	-3.089	0.236
β_1	0.221	0.829
β_2	0.237	0.092
β_3	-0.320	0.105
β_4	-0.536	0.118
β_5	-0.864	0.115
β_6	143.554	0.438
β_7	394.588	1.314
β_8	0.523	0.160
β_9	-0.070	0.029
β_{10}	0.394	0.186
β_{11}	1.618	0.202
β_{12}	0.401	0.134

Interpreting these model parameters learns us that the probability of using slow transport in Southeast Florida increases with increasing household size, with an increasing number of employed people in the household, if the person is not-employed compared to employed, if the activity is pursued at mid-day and if the type of the activity is leisure. The probability of using slow transport decreases on the other hand with an increasing number of children, with an increasing number of driving licenses and an increasing number of cars in the household, with an increasing age and with an increasing duration of the activity.

Nonparametric Models

The support vector machine that tries to find a linear optimal separating plane in the transformed covariate space is made out of 1601 support vectors. Again as a classifier it appears to be the Zero-R classifier in that it classifies each case to the majority class. This leads to an accuracy of 0.974 on the training data and 0.971 on the test data.

Again, for the same reasons as above, we did not obtain results on the polynomial

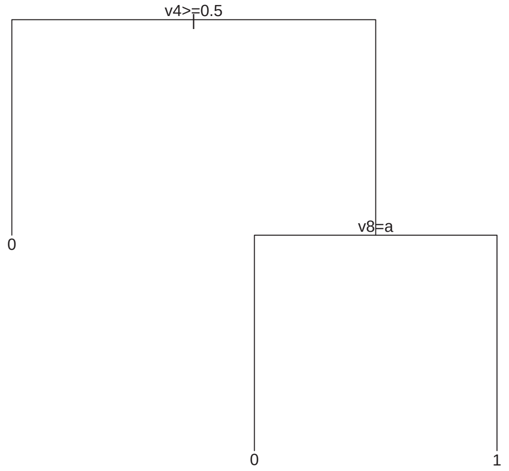

Figure 7.5: *Final tree on Southeast Florida slow transport data*

kernel SVM.

The radial kernel SVM takes 5850 support vectors to make up the SVM classifier. 16.4 % of the positive instances were correctly predicted on the training data, though only 0.8 % was correctly predicted on the test set. This classifier results in a training-accuracy of 0.978 and in a test accuracy of 0.971.

The neural net kernel SVM with parameters κ_1 equal to zero and κ_2 equal to 0.056 needs only 536 support vectors to determine an accuracy of 97.36 % on the training and 97.11 % on the test set. Again, this classifier does not outperform the simplest classifier possible, but on the other hand, since these classifiers are trained to have a very high accuracy without taking sensitivity into account, it is hardly not feasible to do better.

The classification tree (see Figure 7.5) is a little bit more complicated than this simplest classifier (3 final nodes) and one can observe that the performance on the test set is slightly worse. The training accuracy yields 0.975 and the test accuracy equals 0.970. The tree classifies unemployed people without driving license as the only persons that use slow transport. The sensitivity on the training set is equal to 0.097 and that of the test set equals 0.056.

All results are summed up in Table 7.12.

Table 7.12: Performance values for the models on Southeast Florida - Slow Transport

	accuracy		sensitivity		specificity	
	training set	test set	training set	test set	training set	test set
Linear	0.690	0.693	0.634	0.627	0.691	0.695
Mfp	0.694	0.617	0.604	0.659	0.696	0.616
SVM - Linear	0.974	0.971	0.000	0.000	1.000	1.000
SVM - Radial basis	0.978	0.971	0.164	0.008	1.000	1.000
SVM - Neural net	0.974	0.971	0.000	0.000	1.000	1.000
CART	0.975	0.970	0.097	0.056	0.998	0.998

7.5 Conclusion

In this chapter, we tried to discover whether semi- and nonlinear models could add something to transportation analysis in general and to mode choice analysis in the Netherlands as well as in Florida in particular. Linear, semi-linear and nonlinear models were fitted and compared to each other by means of three diagnostics (sensitivity, accuracy and specificity, in decreasing order of importance).

General conclusions that can be drawn report that on very skewed data sets, the performance of linear regression and the multiple fractional polynomial model are usually better than the results of the support vector machines and CART. The main idea of models applied to a setting with a binary response variable is to predict the positive cases well. Since the SVM models are especially derived to achieve an accuracy (instead if sensitivity) as high as possible, this may conflict with the purpose of the modeler. On better balanced data sets (as the Dutch Car Driver data), the performance of the SVM and the CART models are comparable and usual somewhat better than the results of the (semi-)linear models. These latter models have the advantage of being better interpretable, while the SVM's are simply a black box approach. These support vector machines, and more in specific the polynomial kernel SVM, sometimes tend to over-fit the data on the training set. CART is also interpretable, but not in terms of the parameters as in linear models. Also here one has to apply pruning strategies in order to avoid over-fitting.

Further research will take a closer look at extensions of semi-linear and nonlinear statistical models. Apart from support vector machines and classification and regression trees, there are also other nonlinear models, well-known in machine learning: Neural (Zurada, 1992; Haykin, 1994) and Bayesian (Pearl, 1988) networks, to name a few. A possible avenue for future research is to investigate whether these models are transferable to other countries. One could question whether the variables that appeared important in order to model e.g. the probability of using public transport, are the same in all countries, and if so, perhaps policy measures can be undertaken in order to advise certain groups of people (e.g. commuters with flexible working hours) to use the public transport system.

Chapter 8

Final Conclusions

8.1 Introduction

The purpose of this thesis was to find an answer to two particular questions. At first whether simpler, and hence more parsimonious models would perform better, worse or approximately as well as complex models in the context of activity-diary data. And secondly: how well is the performance of nonlinear and semi-linear models, as compared to linear models at the selection of a transport mode? These semi- and nonlinear models often lead to more parsimonious, but on the other hand also to more complex models (in terms of model definition, not in terms of the number of parameters). Another objective that fits within this second main topic of linear and nonlinear models, is concerned with testing parametric linear models on their goodness-of-fit. A test was developed to investigate lack-of-fit of a linear model based on a nonparametric classification tree. This test clearly shows the value of a nonlinear model, and how it can serve to improve a linear model. Of course, it is difficult to give clear recommendations on the choice of a particular model. Which model would be preferable? It raises many questions and there are several possible grounds for preferring one model above another. In transportation studies, predictive performance, interpretability, robustness and sensitiveness for policy measures are generally considered to be relevant criteria for model comparison. These different characteristics will be discussed in the next two sections.

In this final chapter, considerations are made about the consequences for transportation modelling. Which models considered here can be of use for the transportation modelling community, what are the pros, cons and restrictions for each of the models? Are they robust, interpretable, can their parameters be influenced, etc? In the first section, some final remarks are given on the activity-based models used in Chapters 4 and 5. The next section concerns a summary of the results on mode-choice models as presented in Chapters 6 and 7.

8.2 Activity-Based Models

This manuscript has explored the relevance and performance of four different simple models, two recursive partitioning methods (CHAID and C4.5) as well as the application of bagging and boosting in building activity-based models of transportation demand. The four simple models are the Zero R model, the One R model (which can be regarded as a very simple tree structure), the Naïve Bayes model, and the Feature Selection (FS) approach as described in Chapter 4. Furthermore, in Chapter 5, bagging and boosting have been applied to the One R models and to the feature selection models. This allows us to make a comparison between bagging and boosting on a rather weak classifier (One R) and a stronger one, i.e. a tree induction algorithm (C4.5 after feature selection).

8.2.1 Predictive Performance

Let us consider again the probability on a correct prediction (or the predictive accuracy) on each dimension for each approach. Figures 8.1 and 8.2 show the performance for each of the methods on every dimension.

Figure 8.1 shows the accuracies of the three simple classifiers (Zero R, One R and Naïve Bayes), together with those of CHAID, C4.5 (the 'full' approach) and of the second type of simple models, i.e. the C4.5 trees after feature selection. The Zero R method clearly sets the lowest standard above which all other methods should perform. The simple models (going from One R up to FS) clearly outperform Zero R and the recursive partitioning methods (CHAID and C4.5) outperform these simple methods by just a few percentages. Both bagging and boosting models achieve better results than their original analysis, even better than the recursive partitioning methods. Only on the 'Mode for work' and the 'Location 1' dimensions, the results of the CHAID analysis could not be improved. Though, the number of rules necessary to improve on the CHAID analysis by means of bagging and boosting the FS models also outnumbers the number of rules of the CHAID analysis. In general, one could say that the performance of the models is approximately the same. The simple models perform a little worse than the complex, but they compensate for this fact by

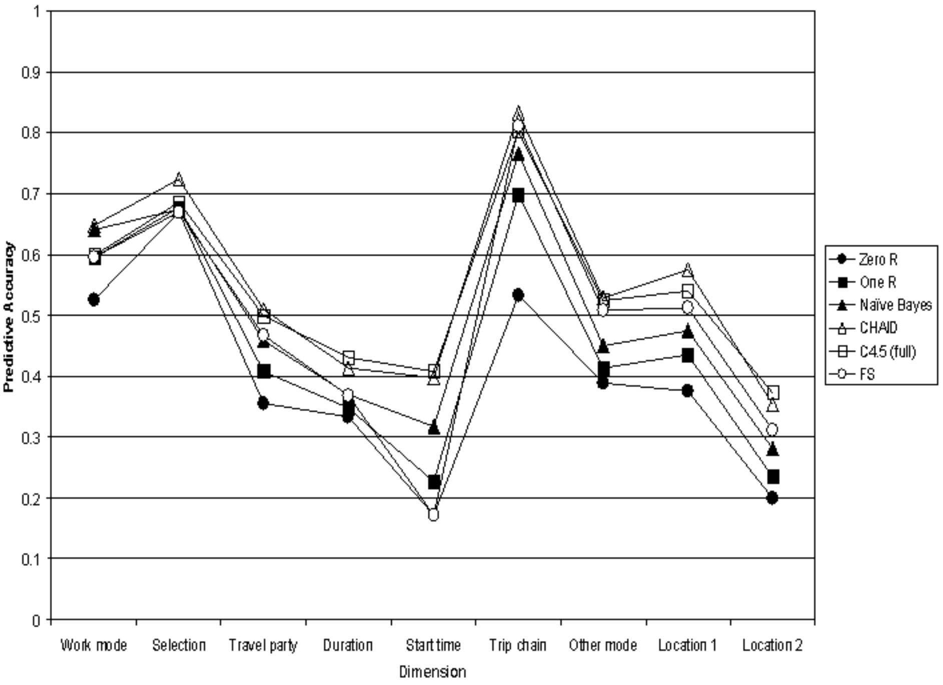

Figure 8.1: *Performance of 'simple' activity-based models*

requiring fewer rules to build them. This holds solely for predicting each dimension separately.

If we consider the aggregate behaviour, (i.e. the sequence alignment measures (SAM) and the correlation coefficients of the OD matrices that capture the performance over the nine different dimensions as discussed in Chapters 4 and 5), apparently the results differ somewhat. These SAM measures determine the dissimilarity between the observed and predicted sequences of activities and should be as low as possible. On the test set, apart from the SAM measure disaggregated on location, which is lowest for the feature selection approach, all other SAM measures are best for either the CHAID or the 'full' C4.5 approach. Though, very often, the feature selection approach or the One R approach come as a close second best. Therefore, it is not surprising that these measures can be improved by bagging and boosting on the One R method. At trip matrix level, the One R method shows the highest correlation coefficient between the observed and expected origin-destination matrices, when disaggregated on transport mode. The full approach has the highest coefficient overall and when we disaggregate on primary activity, while the feature selection approach shows the best results when

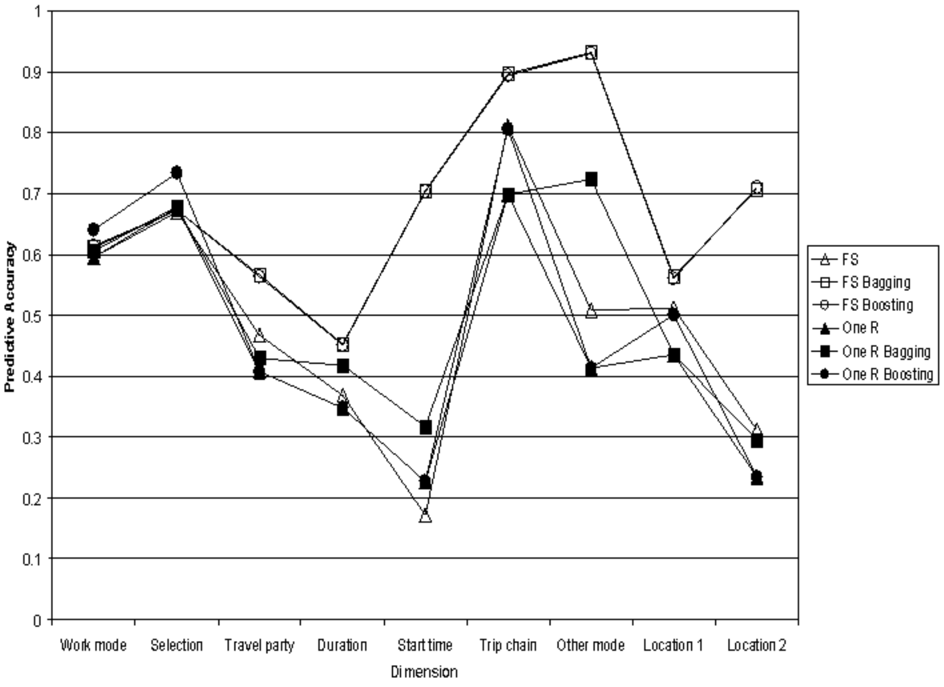

Figure 8.2: *Performance of bagging and boosting on activity-based models*

one disaggregates on day. Most of these results can again be equaled or improved by bagging and boosting. Conclusion: (combinations of) simple models do not necessarily perform worse than the more complex models and they are able to capture the most important information in trying to predict the activity-travel behaviour.

There are also some other grounds on which some modelling approach might be preferred above another. Interpretability, robustness and sensitiveness for policy measures are considered to be relevant criteria in transportation models. All these characteristics will be discussed in the next three subsections.

8.2.2 Interpretability

With respect to this characteristic, there is a notable uniformity amongst the different approaches. In fact, all simple models can be regarded as being tree structures. In the Zero R approach, there is only the root node, in the One R approach, there is just one branch, while in the Naïve Bayes approach trees up to depth three are possible. The CHAID and C4.5 approach are recursive partitioning methods, hence they provide tree structures. In general, the CHAID method provides the most complex trees, followed by the C4.5, the Naïve Bayes and the feature selection approach, in this order.

The different states, represented by the branches of the tree, are often readily recognisable in terms of context-specific decision rules or behavioural patterns. The legibility of the tree structures may assist policy makers in identifying those groups of individuals or contexts for which certain transport-demand measures are applicable. For example, a tree may reveal circumstances in which a particular measure would or would not lead to a change in mode choice, travel party, ...

Bagging and boosting models are harder to interpret. They do not lead to one particular tree, either they provide fifty possibly different tree structures on which the results are averaged, or they provide weighted results. After taking a closer look at the fifty trees that bagging provides per dimension, one may conclude that they usually differ, sometimes this difference starts already at the root node. This does not allow for immediate recognition of the important variables for prediction of the outcome variables.

8.2.3 Robustness and Stability

The advantage of all these tree structure models (when compared to e.g. parametric models) is that they are 'free of assumptions', thus their structure is completely determined by the data. Above that, the tree structure models are also less sensitive to possible sources of bias, like outliers and multicollinearity (Wets *et al.*, 2000). Outliers are simply cases that may increase the heterogeneity in a branch under given conditions, but they rarely effect the variables on which the split is made or the modal response outcome in that particular branch. Tree induction models do not solve the problem of multicollinearity, but this type of imperfections in the data has probably less disruptive impacts on the outcome when e.g. compared to parametric models. Since tree structure models determine the selection of the explanatory variables together with their impact on the outcome variable, multicollinearity means that some variables are simply redundant given the presence of others. These redundant variables will not be used for the formation of the branches in the tree structure.

Removing the irrelevant variables to increase robustness is also the purpose when a variable selection technique was applied before building the C4.5 trees. The results learn us that a strong reduction in the complexity of the tree do not necessarily lead to a decrease in performance, sometimes even quite the reverse happens. Experience learns that quite different structures may fit almost equally well on a given data set. In order to increase the robustness, bagging was applied to two of the simple approaches, with successful results.

8.2.4 Sensitiveness for Policy Measures

In general, tree structure models predict discontinuous behavioural changes if policy measures lead to a shift from one condition state to another (e.g. if the price of petrol raises above 1.25 €, public transport will be chosen as transport mode). In other words, people's behaviour, as predicted by the model, is to some extent insensitive to (small) changes (e.g. if the petrol prices increases from 1.05 to 1.15 €, the model will not predict a change in the transport mode used). However, whether a rather low sensitiveness also leads to a low prediction of behavioural change is an empirical question. It might very well be the fact that individuals are indifferent to small variations in conditions (e.g. travel distances), and that they make changes in their behaviour (e.g. mode choice) only if the changes make a qualitative difference (e.g. is my destination within walking distance or not). Investigating this question about changes in travel behaviour requires data about the behaviour before and after a change is implemented or data about the response of individuals to (hypothetical) situations. These data are not at hand, but future research could focus on this questions by comparing the ability of the different methods to predict behavioural change.

8.3 Mode-Choice Models

The second main question posed in this manuscript was whether semi- and nonlinear models perform better than the standardly used linear models in the context of mode choice models? The semi-linear approach of multiple fractional polynomials (mfp) is compared to the nonlinear approaches of support vector machines (SVM's) and classification and regression trees (CART) on the same four criteria as the activity-based models, i.e. predictive performance, interpretability, robustness and sensitiveness to policy measures.

Another study aim that fits within this context is concerned with testing parametrical linear models on their goodness-of-fit. We developed a test, based on a nonparametric classification tree, that examined linear models on lack-of-fit. This section clearly puts its value in perspective.

8.3.1 Predictive Performance

The mode choice data sets that have been used are very skewed. It means that the probability on observing a case is very low, the majority class of the response variable is zero. Therefore, it would not have been 'honest' to make a comparison of the different models on e.g. predictive *accuracy* alone. Therefore, we have added the *sensitivity* and the *specificity* measure as well. The sensitivity measures (in the case of public transport) the conditional probability of predicting public transport given that public transport was observed, while the specificity equals the conditional probability of predicting the 'other' transport mode, given that is was observed. These two additional measures give a more complete picture of the relative performance on the five different mode choice data sets. Consider e.g. the data set of Southeast Florida on public transport: only 9.67% of the cases do use public transport. Thus if a particular model predicts default the 'other' transport mode, it has an accuracy of over 0.900, but none of the cases that you actually want to predict is predicted correctly. Figures 8.3 to 8.7 show the results on the different data sets. The models on the horizontal axis are ordered according to an increasing value of the accuracy. The predictive performance measures (all on a scale from zero to one) will again be compared on the test set. Some general characteristics are discussed below.

For all skew data sets, the specificity is rather high, since this 'other' transport mode is rather easy to predict. For all public transport data sets, the sensitivity of the multiple fractional polynomial model and the support vector machines is especially low. The sensitivity measure of CART is somewhat higher, while the linear model appears to be best in predicting predict public transport for these data sets. For both slow transport data sets, the same results can be observed as in the public transport case, apart from the multiple fractional polynomial models. In the slow transport data sets, the mfp model clearly predicts the 'slow transport' choice much better. In the Dutch Car Driver data set, which is more balanced than the others, the results are somewhat better, as expected. The nonlinear models outperform the linear and semi-linear on both accuracy and sensitivity, while the semi-linear model also performs better than the two linear models.

To conclude, one could advise to consider nonlinear models for prediction, certainly if the data set is balanced or if the number of 'Y=1' cases is large enough. Otherwise, i.e. if the probability of observing a case is rather low, a linear or semi-linear model will probably serve best.

This illustrates once more the value of the tree-based test statistic. Overlooking the results on the different data sets, the parametric models (linear and semi-linear) ap-

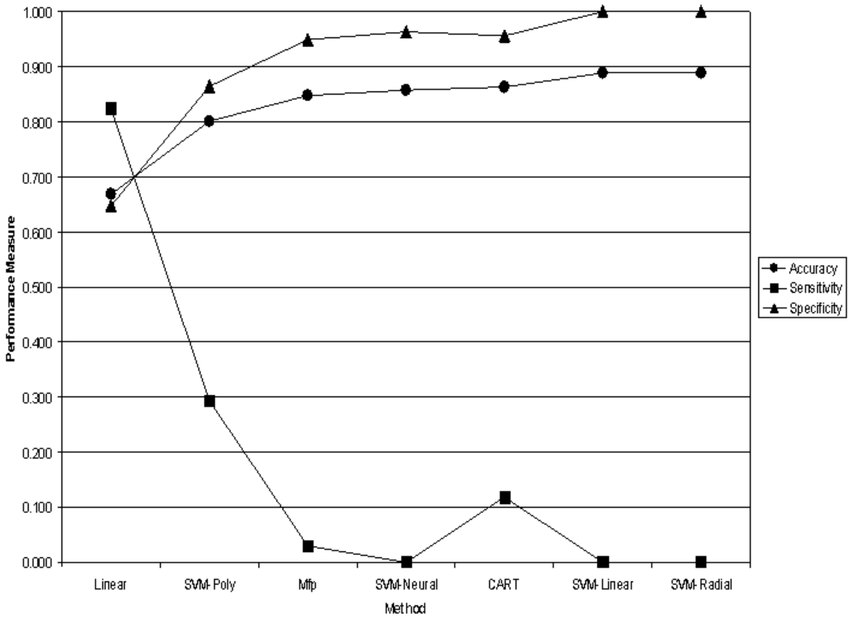

Figure 8.3: *Performance of Different Methods on Mode Choice Models: Dutch Data - Public Transport*

pear to be the only models with a reasonable predictive performance in terms of sensitivity. And sensitivity seems a very important measure in this context. Apparently, only a small number of people use slow or public transport, thus it is important to come up with an interpretable model that is able to predict the transport usage of this minority. The tree-based lack-of-fit test is able to evaluate the fit of parametric models (regardless of the type of explanatory variables) and if the null model is rejected, a close inspection of the classification tree can reveal a particular deviation from the null model. It is a nice example that shows how a nonlinear, nonparametric technique can be used to confirm or improve a parametric (linear or semi-linear) model.

The next three subsections will discuss the interpretability, the robustness and the sensitiveness for policy measures. With respect to these properties sometimes the parametric models will have advantages over the nonparametric models, while with respect to others it might be the other way around.

8.3.2 Interpretability

For interpretability, we argue that the nonparametric CART model and the linear and semi-linear models are comparable. The impact of a certain explanatory variable (ceteris paribus) on the response can easily be determined from the corresponding parameter value in the parametric models. On the other hand, in the CART model, the branches of the tree represent different condition states and the sequence of splits that make up these branches indicate which variables have an impact on the response and what this impact is. When it comes to readability, the structure that is provided by a tree has an advantage over the additive model that is provided by the (semi-) linear models. Due to this legibility, policy makers can easier identify groups of individuals or contexts wherein similarities or changes in the choice of transport mode can be identified.
The SVM models will be hard to compare with the other models, since they can be regarded as a black box approach. One can change the values of the parameters that make up the different SVM models, but with respect to interpretability, robustness and sensitiveness to policy measures, there is very little to discuss.

8.3.3 Robustness

Again, as discussed in the section on activity-based models, the CART model has two possible advantages. At first, the parametric models assume a predefined additive functional form, the CART model is 'free of assumptions' in the sense the data determine the tree structure.

Furthermore, the stability of the parameter estimates of a parametric model tends to be sensitive to the presence of outliers and multicollinearity in the data used for estimation. Careful preprocessing is required in order to try to eliminate these possible sources of bias. Tree structure models, on the other hand, are less sensitive to such sources of bias (Wets *et al.*, 2000). As discussed above, outliers only increase the heterogeneity in the nodes, whereas multicollinearity means that some variables are irrelevant given the presence of others and consequently they are not used in the formation of the tree.

8.3.4 Sensitiveness for Policy Measures

With respect to sensitiveness to policy measures, parametric models seem to have an advantage as they are designed to predict the size of the independent impacts of the explanatory variables on the choice of the transport mode (response variable).

Even the impact of a very small change in one of the variables can be determined. The design of (semi-)linear models makes them ideal for policy makers to test the impact of different scenario's (e.g. what is the effect on the transport mode if we increase the number of part-time workers?). In contrast, the CART model can only predict a change in the response variable if the change in behaviour caused by a new policy measure leads to a shift from one branch to another. Thus, the tree model is less sensitive to policy measures, though this does not mean that the prediction of a change in the mode choice will be worse in the tree models. It depends on whether the individual regards the change as having a qualitative impact for which a change in transport mode is needed. The extent to which this property represents an advantage of tree models over parametric models depends on the degree to which the transport mode to be predicted can be governed by the branches of the tree that are extracted from the data. Investigating this question is an interesting topic for future research.

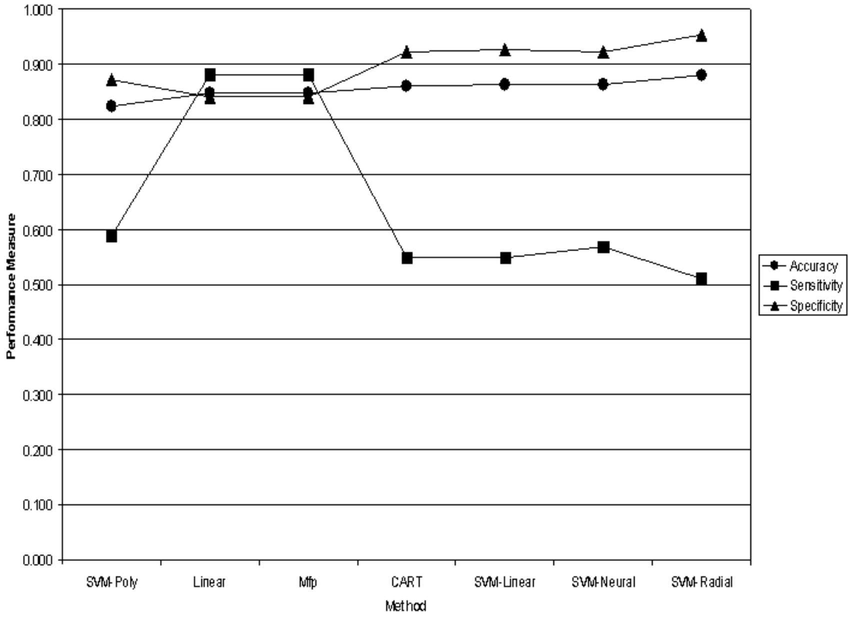

Figure 8.4: *Performance of Different Methods on Mode Choice Models: Dutch Data - Slow Transport*

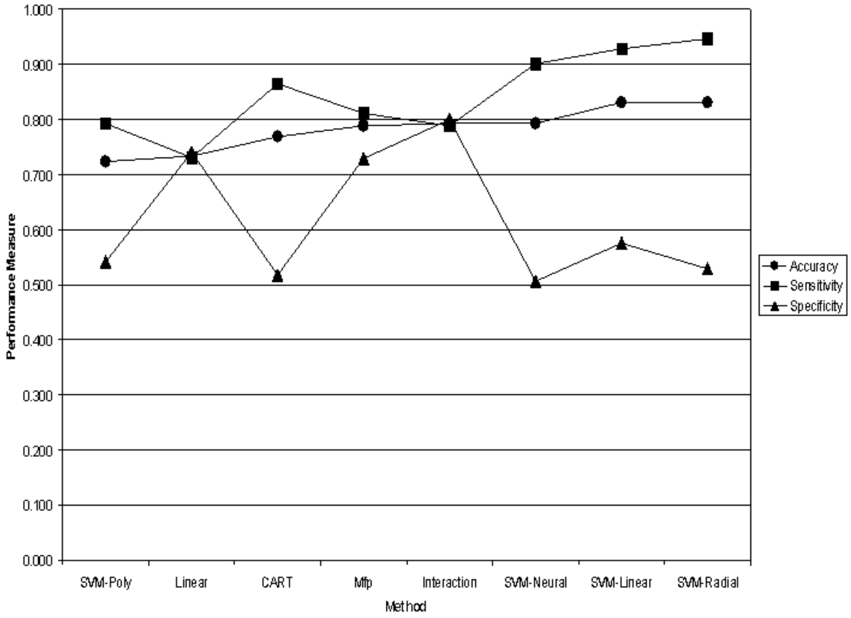

Figure 8.5: *Performance of Different Methods on Mode Choice Models: Dutch Data - Car Driver*

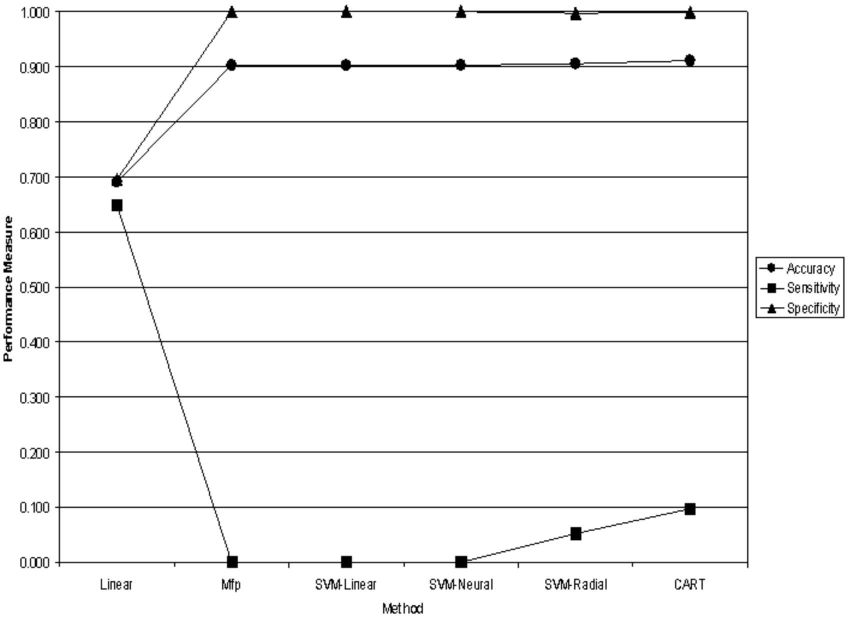

Figure 8.6: *Performance of Different Methods on Mode Choice Models: Southeast Florida - Public Transport*

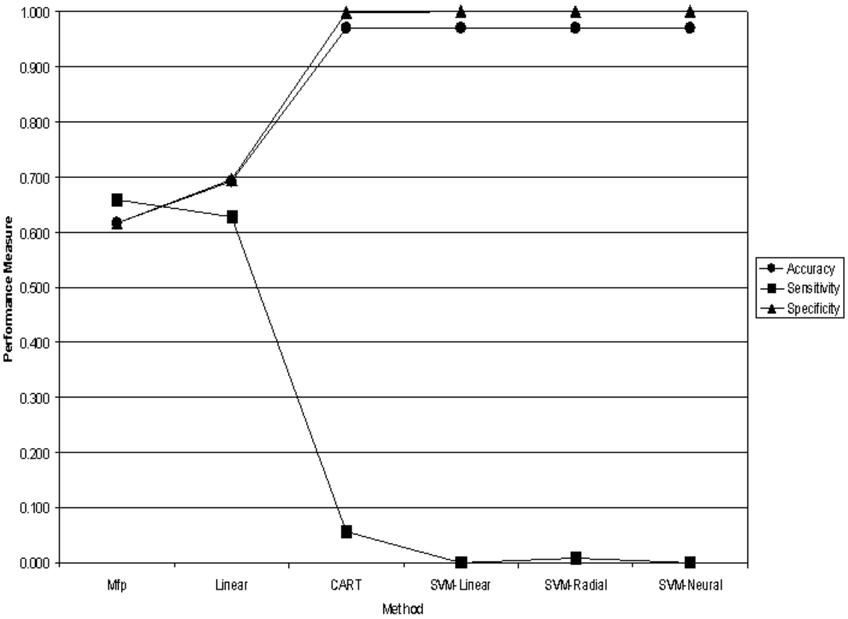

Figure 8.7: *Performance of Different Methods on Mode Choice Models: Southeast Florida - Slow Transport*

Appendix A

In order not to overload Chapter 2, we will discuss the variables that are important in the Albatross system (Arentze and Timmermans, 2000). Note again that all relevant choices are considered through different dimensions in the Albatross system: which activity is conducted, where, when, with whom, for how long and which transport mode is used. The variables that determine these nine choice facets will be discussed separately in the following sections.

A.1 Mode for Work

Table A.1 shows the list of independent variables for the 'mode for work' choice facet.

The first variable 'group' is included to allow the system to distinguish between cases where there is no partner, the partner's schedule for that day is unknown or that schedule is known. The next series of variables describe the activity program at the level of the schedule skeleton (S)[1]. These include the total time engaged in Work1, in Work1 and Work2 together the number of mandatory, out-of-home activities other than work and the presence of a bring/get activity. Work1 includes work/school activities and Work2 voluntary work activities. For the partner, the variables have zero values if there is no partner, or if the partner's schedule is unknown.

The succeeding variables describe the work-chain (W)[2] for which a mode choice is to be made. These include work time and travel time information. Bike travel time is taken as an indicator of travel distances. Furthermore, travel time ratios between modes are included as indicators of the relative speed of each mode on the (shortest) route between locations.

Then a series of descriptors at the level of the work-chain are included. First, the

[1] Schedule skeleton

[2] Chain of work episodes for which transport mode is selected

Table A.1: Independent variables used in the ‘mode for work’ choice facet of Albatross

Name	Description	Categories
group	Partner status	1: no partner 2: schedule partner unknown 3: partner schedule known
Two	Total time of Work1 in minutes in S	$0:0; 1:\leq 240; 2:241-360;$ $3:361-480; 4:>480$
Ttot	Total time of Work1 and Work2 in S	$1:\leq 240; 2:241-360;$ $3:361-480; 4:>480$
Nsec	Number of mandatory, out-of-home activities other than work in S	$0:0; 1:1; 2:2; 3:3-4;$ $4:4-5; 5:>5$
yBget	There is a bring/get activity in S	0: yes; 1: no
Pwo	Total time of Work1 in minutes in S of partner	$0:0; 1:\leq 240; 2:241-360;$ $3:361-480; 4:>480$
PTtot	Total time of Work1 and Work2 in S of partner	$1:\leq 240; 2:241-360;$ $3:361-480; 4:>480$
PNsec	Number of fixed out-of-home activities other than work in S of partner	$0:0; 1:1; 2:2; 3:3-4;$ $4:4-5; 5:>5$
PyBget	There is a bring/get activity in S of partner	0: yes; 1: no
Tbike	Objective travel time by bike to location of W in minutes	$1:\leq 10; 2:11-20;$ $3:21-30; 4:31-50;$ $5:51-100; 6:>100$
Rcabi	Ratio car/bike travel time in %	$1:\leq 25, 2:26-50;$ $3:51-75; 4:>75$
Rpubi	Ratio public transport/bike travel time in %	$1:\leq 100; 2:101-150;$ $3:151-200; 4:>200$
Rpuca	Ratio public transport/car travel time in %	$1:\leq 300; 2:301-500;$ $3:501-700; 4:>700$
Peak1	Start time of W falls in 7:30-9:00 AM	0: yes; 1: no
Peakn	End time of W falls in 17:00-18:00	0: yes; 1: no
Two2	Total time of W in minutes	$1:\leq 300; 2:301-500;$ $3:501-700; 4:>700$
Nloc	Number of different locations in W	1: one; 2: more than one
Avo	Activity in S with end time within 1-hour interval before first work episode in W	1: none 2: bring/get; 3: other
Ana	Activity in S with start time within 1-hour after last work episode in W	1: none 2: bring/get; 3: other
Pywork	Partner has work activity during work time	0: no; 1: yes
Pybget2	There is a bring/get activity in S of partner during W	0: no; 1: yes
PNfix	Number of out-of-home activities in S of partner during W	0: none; 1: one 2: more than one
PTTmax	Maximum bike travel time across activities in S of partner during W (minutes)	0: none; 1: 1-15; $2:16-30; 3:>30$

start time of the first work episode and end time of the last work episode of the chain determine whether travel time takes place during the morning and/or evening rush hours. Second, the number of different work locations involved serve as a measure of the amount of travel involved apart from the first and last commute. Third, activities included in the skeleton that are closely related in time to the start of the first work episode or the end of the last work episode are recorded as possible condition for trip-chaining during the first and the last commute. Finally, the last set of variables tends to cover travel demands of the partner during the work-chain. These include the number of out-of-home activities in the schedule skeleton, maximum travel time across locations and the presence of a bring/get activity.

A.2 Activity Selection, Travel Party and Duration

Table A.2 shows the list of independent variables for the 'activity selection', 'travel party' and 'duration' choice facets. The footnotes [3], [4] and [5] distinguish between the three different choice facets, while S stands for the evolving schedule.

The program-level variables are partly dynamically and partly statically defined. The fixed activities, which belong to the skeleton of the schedule, are given and remain constant during the process. Therefore, the variables Two and Ttot are defined statically as the total time scheduled for Work1 and Work1 and Work2 together, respectively. Twincl is added to take observed travel time as well as activity time related to Work1 activities into account. In the present step, the travel time information is considered known, given our assumption that transport mode choice for primary work activity is made in the previous step.

The other program-related variables (except yBget) are dynamically defined. A first set of variables defines for each flexible activity the total time scheduled (T-variants) or, simply, the presence of the activity (y-variants) in the current schedule. The variable values are initially zero and updated each time an activity is added. The Nsec variable is a summary variable representing the number of flexible or fixed out-of-home activities other than work in the current schedule. The rationale for including the program-level variables in general is that they describe conditions, such as activity load of the current program, possibilities to combine activities, and so on.

Some of the variables at this level need further explanation. The Iact variable defines for the current activity type the number of instances of that type present in

[3] available for selection decisions only

[4] available for travel party decisions only

[5] available for duration decisions only

Table A.2: Independent variables used in the ‘activity selection’, ‘travel party’ and ‘duration’ choice facets of Albatross

Name	Description	Categories
Iact	Number of instances of the current activity type in S	0: 0; 1: 1 $2 :> 1$
Two	Total time of Work1 in S (in minutes)	$0 : 0; 1 :\leq 240; 2 : 241 - 360;$ $3 : 361 - 480; 4 :> 480$
Twincl	Total time of Work1 incl. travel in S	$0 : 0; 1 :\leq 260; 2 : 261 - 380;$ $3 : 381 - 500; 4 :> 500$
Ttot	Total time of Work1 and Work2 in S	$0 : 0; 1 :\leq 240; 2 : 241 - 360;$ $3 : 361 - 480; 4 :> 480$
Nsec	Number of out-of-home activities other than work in S	$0 : 0; 1 : 1; 2 : 2;$ $3 : 3; 4 : 4; 5 :> 4$
yBget	There is a bring/get activity in S	0: no; 1: yes
yDshop	There is a daily shopping activity in S	0: no; 1: yes
yServ	There is a service activity in S	0: no; 1: yes
yNDshop	There is a non-daily shopping activity in S	0: no; 1: yes
ySoc	There is an out-of-home social activity in S	0: no; 1: yes
yLeis	There is an out-of-home leisure activity in S	0: no; 1: yes
Tsoc	Total time of social activities (in-home and out-of-home) in S	$0 : 0; 1 :\leq 30; 2 : 31 - 60;$ $3 : 61 - 120; 4 :> 120$
Tleis	Total time of out-of-home leisure activities in S	$0 : 0; 1 :\leq 30; 2 : 31 - 60;$ $3 : 61 - 120; 4 :> 120$
Td-shop	Total time of daily shopping activities in S	$0 : 0; 1 :\leq 20; 2 : 21 - 40;$ $3 : 41 - 60; 4 :> 60$
Tserv	Total time of service activities in S	$0 : 0; 1 :\leq 20$ $2 : 21 - 40; 3 : 41 - 60; 4 :> 60$
Tnd-shop	Total time of non-daily shopping activities in S	$0 : 0; 1 :\leq 30$ $2 : 31 - 60; 3 : 61 - 120; 4 :> 120$
A1dur	Total relative time of current activity in S	0: none; 1: short 2: average; 3: long
Tmax(t)	Maximum available time in t-th time interval in S^{fix} (in minutes)	0: 0; 1: 1-30 $2 : 31 - 60; 3 :> 60$
yCar(t)	Availability of car in t-th time interval in S^{fix}	0: no; 1: yes; 2: schedule partner is unknown
Atype	Activity type	1: daily shopping; 2: service; 3: non-daily shopping; 4: social; 5: leisure
yAvail[3]	Selection of activity is feasible given S and minimum duration for the activity type	0: no; 1: yes
yAvail[4]	‘Others in the household’ option is available given the household composition	0: no; 1: yes
yAvail2[5]	The ‘average’ duration class is feasible given S and the minimum duration for that class	0: no; 1: yes
yAvail3[5]	The ‘long’ duration class is feasible given S and the minimum duration for that class	0: no; 1: yes
Awith[5]	Travel party	0: none; 1: only others inside household; 2: others outside household involved

the current schedule. This is a powerful variable if the probability of adding a next activity decreases with the number of instances already scheduled. A1dur represents an alternative way of encoding activity time. The variable defines a short, average and long time relative to the activity type under concern, such that for example the long category of one type may still be shorter than the average of another type. The definition of the duration categories corresponds with the alternatives considered for the duration choice (see Table A.3).

The next set of variables describes cases at the schedule level. First, the Tmax(t) variables represent the maximum time available across available time slots in the schedule skeleton. The index t defines a particular time period among six distinguished time periods: before 10 AM; 10-12 AM; 12-2 PM; 2-4 PM; 4-6 PM and after 6 PM. The time for each time slot and each time period is determined by the overlap between time ranges given by opening hours of available facilities for the activity type, the time between fixed activities and the time period t. Second, the yCar(t) variables represent the availability of the car in each time period t, as a function of the number of available cars in the household, the number of adult members of the household and the mode used by the partner for work. For example, the car is considered not available if the car is in use for work by the partner and there is less than one car per adult available in the household. As in the previous table, the equal-frequency method was used to discretise continuous variables.

Finally, besides activity type, the activity-level variables are specific for each of the three considered choice facets. The variables at this level represent feasibility conditions for choice alternatives. Selecting an activity is considered infeasible if the maximum available time across the time slots that are available within opening hours of available facilities for the activity is shorter than the minimum duration for the activity type. For travel party decisions, the options 'alone' and others outside the household' are considered to be always available. The 'other(s) inside the household' option, however, is considered available only in multi-person households. With respect to the duration choice, the exact definitions of alternatives are shown for each flexible activity type in Table A.3.

The shortest duration class is available by definition (given the positive selection decision). The average and long duration alternatives, however, are available only if the minimum duration defined for the concerned class fits in the schedule (evaluated in the same way as in the case of selection). Apart from the feasibility conditions, the travel party dimension is an additional variable for the duration choice facet. This dimension is considered known at the time that the duration decision is made.

Table A.3: Classification of activity duration

	short		average		long	
type	range	mean	range	mean	range	mean
daily shopping	[10 - 20]	15	[21 - 45]	35	[46 - 90]	50
service	[5 - 10]	5	[11 - 20]	15	[21 - 40]	30
non-daily shopping	[10 - 30]	20	[31 - 80]	60	[81 - 160]	90
social	[10 - 75]	60	[76 - 150]	120	[151 - 300]	180
leisure	[10 - 60]	40	[61 - 120]	90	[121 - 240]	150

A.3 Activity Start Time

Table A.4 shows the list of independent variables for the 'start time' choice facet. The footnote [6] points to earlier definitions of duration alternatives in the table for duration decisions, whereas A denotes the concerned activity and S^{all} the complete observed schedule.

The first set of variables, labelled Tmax(t), represents for each distinguished time interval t the available time in the current schedule given start and end time times of the fixed activities, the opening hours of available facilities for the concerned activity and estimated travel times for the free as well as for the fixed activities in the current schedule. Tmax represents the maximum time across feasible positions in the current schedule. Because the location, mode and trip chains are not yet known in this stage, the travel time estimates are based on activity-type specific ratios between activity type derived from the entire data set. These ratios are represented in Table A.5.

On the other hand, the used facility opening hours are specific for origins and day of the week. The time periods t correspond to the alternatives for the start-time choice (i.e. before 10 AM, 10-12 AM, 12-2 PM, 2-4 PM, 4-6 PM and after 6 PM).

The Tmax variables are updated after each start time decision. Initially, only the schedule skeleton is given and the fixed start and end times determine the available time in each position.

As said before, once a start-time decision is made for a flexible activity, its schedule position is taken as given (i.e. taken as observed). Tmax accounts for the assumed flexibility in timing and duration choices of flexible activities by calculating the maximum time available per position. The maximum represents the available time under the most favourable duration and start-time choice within given duration and start-time constraints. Still, the sequential procedure implies that high-priority activities

[6] see earlier definitions of duration alternatives in Section A.2

Table A.4: Independent variables used in the ‘start time’ choice facet of Albatross

Name	Description	Categories
Nsec	Number of mandatory out-of-home activities other than work in S^{all}	$0:0; 1:1; 2:2;$ $3:3-4; 4:>4$
Two	Total time of Work1 in S^{all} (in minutes)	$0:0; 1:\leq 240$ $2:241-360; 3:361-480$ $4:>480$
Twincl	Total time of Work1 incl. travel in S^{all}	$0:0; 1:\leq 260;$ $2:261-380; 3:381-500; 4:>500$
Ttot	Total time of Work1 and Work2 in S^{all}	$0:0; 1:\leq 60;$ $2:61-120; 3:121-240; 4:>240$
yBget	There is a bring/get activity in S^{all}	0: no; 1: yes
yDshop	There is a daily shopping activity in S^{all}	0: no; 1: yes
yServ	There is a service activity in S^{all}	0: no; 1: yes
yNDshop	There is a non-daily shopping activity in S^{all}	0: no; 1: yes
ySoc	There is an out-of-home social activity in S^{all}	0: no; 1: yes
yLeis	There is an out-of-home leisure activity in S^{all}	0: no; 1: yes
Tsoc	Total time of social activities (in-home and out-of-home) in S^{all}	$0:0; 1:\leq 30; 2:31-60;$ $3:61-120; 4:121-240; 5:>240$
Tleis	Total time of out-of-home leisure activities in S^{all}	$0:0; 1:\leq 30; 2:31-60;$ $3:61-120; 4:121-240; 5:>240$
Td-shop	Total time of daily shopping activities in S^{all}	$0:0; 1:\leq 20; 2:21-40;$ $3:41-60; 4:>60$
Tserv	Total time of service activities in S^{all}	$0:0; 1:\leq 20; 2:21-40;$ $3:41-60; 4:>60$
Tnd-shop	Total time of non-daily shopping activities in S^{all}	$0:0; 1:\leq 30; 2:31-60;$ $3:61-120; 4:>120$
Tmax(t)	Maximum available time in t-th time interval in S^{all} (possible duration for A)	0: < minimum; 1: minimum-average 2: average-maximum; 3: > maximum[6]
Btwo(t)	There is a Work1 activity with start time in $t = 1, \ldots, 3$ in S	0: no; 1: yes
Etx(t)	There is an out-of-home activity with end time in $t = 1, \ldots, 6$ in S	0: no; 1: yes
DBT(t)	Saved bike travel time if A is linked with out-of-home activity with start time in $t = 1, \ldots, 3$	0: 0 or no such activity; $1:\leq 10$ $2:11-30; 3:>30$
DET(t)	Saved bike travel time if A is linked with out-of-home activity with end time in $t = 1, \ldots, 6$	0: 0 or no such activity; $1:\leq 10$ $2:11-30; 3:>30$
yCar(t)	Availability of car in t-th time interval in S	0: no; 1: yes; 2: schedule partner is unknown
Atype	Activity type of A	1: daily shopping; 2: service; 3: non-daily shopping; 4: social; 5: leisure
Awith	Travel party of A	0: none; 1: only others inside household; 2: others outside household involved
Iact	Number of the current activity type of A	$1:1; 2:>1$
Adur	Duration of A	1: short; 2: average; 3: long[6]
Ad1	Shortest bike travel time across possible locations for A (minutes)	$0:0; 1:\leq 10$ $2:11-30; 3:>30$

Table A.5: Travel time/duration ratios used to estimate travel times based on activity duration

activity	ratio
daily shopping	0.33
service	0.65
non-daily shopping	0.28
social	0.14
leisure	0.14
unknown	0.30

(e.g. shopping) may reduce the start-time options for activities lower in the assumed hierarchy (e.g. leisure activities).

The levels for Tmax are defined dependent on the duration class of the activity under concern. The zero level means that there is no feasible schedule position for the t-th start-time range even if the minimum duration of the activity is taken. The levels 1 and 2 denote respectively, that there is a feasible position for implementing an average and long duration type of activity. Hence, the Tmax variable has two functions. First, it defines the feasibility condition for each start-time option and second, it indicates the extent to which each time period allows flexible choice of activity duration.

The next set of schedule-level variables allows the system to anticipate on possibilities to establish connections with other out-of-home activities. Various indicators are included. First, Btwo(t) indicates whether the current schedule includes a work activity with a start time falling in the t-th time period. Second, the ETx(t) denotes the same for the end time of any out-of-home activity. For existing flexible activities possible end times given duration and start-time constraints are taken. Note that for other than work activities only the end times are taken into account. This is done to reduce redundancy in the set of variables. Other-than-work activities tend to be short so that start and end times often fall within the same time period and only one value can serve as an indicator for both.

Second, the DBT(t) and DET(t) variables more specifically indicate the travelled distance that could be saved by establishing a travel connection. Hereby, DBT refers to the work activity with start time falling into time period t, if any, DET, relates to the out-of-home activity of any type with the end time falling in the t-th interval, if any. As in previous models, bike travel time is taken as indicator of distance. Let O denote the existing out-of-home activity, A the activity for which the start time choice is made and H the home location, then the saved time is determined by comparing the sum of travel time across H-O-H and H-A-H tours with the travel time of H-O-A-H or H-A-O-H trip. In all trip types, the location that minimises travel time across location alternatives for A is taken as the location for A.

The final set of schedule-level variables is given by yCar(t). As explained before, this variable represents the availability of the car in the *t*-th time period, given the number of cars present per adult member of the household and the mode used for the work activity in the partner's schedule (if any). Finally, the remaining variables all relate to dimensions of the activity for which the start-time decision is currently made. These are restricted to the dimensions considered known at this stage, i.e. the activity type, travel party, duration and shortest home-based distance.

A.4 Trip Chaining

The set of variables that were used to describe the cases at the program-level, schedule-level and activity-level are summarised in Tables A.6 and A.7. The footnote [7] points to earlier definitions of duration alternatives in the table for duration decisions, S^{all} denotes the complete observed schedule, whereas S is the current schedule. Finally, A is the concerned activity and O is an existing previous or next activity.

The program-level variables are largely the same as in the previous step. Only the variables that are specific for the trip-chaining step are considered here.

First, the yAstop, yBstop and yIBstop denote the feasibility of the trip-chaining options. The rules for determining the feasibility take the spatial, temporal and institutional constraints into account. The next set of variables, then, describes the concerned flexible activity regarding the dimensions that are considered known in this step. First activity step is defined in two alternative ways by a single nominal variable and a binary variable for each activity type respectively. Binary encoding is added to allow the system to distinguish between certain types also if significant splits on the nominal variable cannot be found due to the Bonferroni adjustments. The Awith, Adur and Astart variables describe the travel-party, duration and start-time dimensions in terms of the choice alternatives of the choice facets in previous steps. Finally Ad1 measures the shortest distance from the home location across the possible locations for the activity. Note that in the case of social activities every zone in the area is by definition zero. For the other activities, the shortest distance depends on locations of available facilities.

[7] see earlier definitions of duration alternatives in Section A.2

Table A.6: Independent variables used in the 'trip chaining' choice facet of Albatross: Part I

Name	Description	Categories
Nsec	Number of mandatory out-of-home activities other than work in S^{all}	$0 : 0; 1 : 1; 2 : 2; 3 : 3 - 4; 4 :> 4$
Two	Total time of Work1 in S^{all} (in minutes)	$0 : 0; 1 :\leq 240; 2 : 241 - 360;$ $3 : 361 - 480; 4 :> 480$
Twincl	Total time of Work1 incl. travel in S^{all}	$0 : 0; 1 :\leq 260; 2 : 261 - 380;$ $3 : 381 - 500; 4 :> 500$
Ttot	Total time of Work1 and Work2 in S^{all}	$0 : 0; 1 :\leq 60; 2 : 61 - 120;$ $3 : 121 - 240; 4 :> 240$
yBget	There is a bring/get activity in S^{all}	0: no; 1: yes
yDshop	There is a daily shopping activity in S^{all}	0: no; 1: yes
yServ	There is a service activity in S^{all}	0: no; 1: yes
yNDshop	There is a non-daily shopping activity in S^{all}	0: no; 1: yes
ySoc	There is an out-of-home social activity in S^{all}	0: no; 1: yes
yLeis	There is an out-of-home leisure activity in S^{all}	0: no; 1: yes
Tsoc	Total time of social activities (in-home and out-of-home) in S^{all}	$0 : 0; 1 :\leq 30; 2 : 31 - 60;$ $3 : 61 - 120; 4 : 121 - 240; 5 :> 240$
Tleis	Total time of out-of-home leisure activities in S^{all}	$0 : 0; 1 :\leq 30; 2 : 31 - 60;$ $3 : 61 - 120; 4 : 121 - 240; 5 :> 240$
Td-shop	Total time of daily shopping activities in S^{all}	$0 : 0; 1 :\leq 20; 2 : 21 - 40;$ $3 : 41 - 60; 4 :> 60$
Tserv	Total time of service activities in S^{all}	$0 : 0; 1 :\leq 20; 2 : 21 - 40;$ $3 : 41 - 60; 4 :> 60$
Tnd-shop	Total time of non-daily shopping activities in S^{all}	$0 : 0; 1 :\leq 30; 2 : 31 - 60;$ $3 : 61 - 120; 4 :> 120$
yCar	There is a car available in the selected time-of-day, given work activity of partner	0: no; 1: yes; 2: schedule partner is unknown
yBstop	Feasibility of a Before Stop, given space-time constraints	0: no; 1: yes
yAstop	Feasibility of an After Stop, given space-time constraints	0: no; 1: yes
yIBstop	Feasibility of a Between Stop, given space-time constraints	0: no; 1: yes
Atype	Activity type of A	1: daily shopping; 2: service; 3: non-daily shopping; 4: social; 5: leisure

Table A.7: Independent variables used in the ‘trip chaining’ choice facet of Albatross: Part II

Name	Description	Categories
yAd-shop	A is a grocery activity	0: no; 1: yes
yAserv	A is a service activity	0: no; 1: yes
yAnd-shop	A is a non-grocery activity	0: no; 1: yes
yAsoc	A is a social activity	0: no; 1: yes
yAleis	A is a leisure activity	0: no; 1: yes
Awith	Travel party of A	0: none; 1: only others inside household; 2: others outside household involved
Adur	Duration of A	1: short; 2: average; 3: long[7]
Astart	Start time of A	1: $<$ 10 AM; 2: 10-0 AM; 3: 0-2 PM; 4: 2-4 PM; 5: 4-6 PM; 6: $>$ 6 PM
Ad1	Shortest bike travel time across possible locations for A (minutes)	$0:0; 1:\leq 10;$ $2:11-30; 3:>30$
Ontime Optime	Available time for A before (On) or after O (Op), given the timing of fixed activities	0: $<$ minimum; 1: minimum - average; 2: average - maximum; 3: $>$ maximum[7]
On-/Optype	Activity type of O	1: bring/get; 2: work1; 3: other
Onwith Opwith	Travel party of O	0: none; 1: only others inside the household; 2: others outside the household involved
On-/Opdu	Duration of O	1: $\leq$ 10; 2: 11-40; 3: 41-120; 3: $>$ 120
Ondu1 Opd1	Bike travel time to (nearest) location of O from home	0: 0; 1: $\leq$ 10; 2: 11-20; 3:$>$ 20
Ond3 Opd3	Shortest bike travel time between location of O and possible locations for A	0: 0; 1: $\leq$15; 2: 16-30; 3:$>$ 30
On-/Opd13	Saved bike travel time of A is linked with O	0: 0; 1: $\leq$ 10; 2: 11-30; 3:$>$ 30

The final set of variables describe the (uniquely) identified feasible activities, if any, for making a before or an after connection respectively. Note that if A can be positioned before as well as after an certain activity, the variables refer to the same activity. First, the Otime variables represent the available time for completing A in that position. In fact, the maximum available time is calculated if there is flexibility in determining the start time and duration for existing activities in the current schedule. The Otype, Owith and Odu describe the activity type, the travel party and duration of the activity, again, in terms of the same categories that are used throughout the model. Finally, the next variables describe the spatial context in terms of the (shortest) distances to O from home, the distance between A and O and the saved travel distance when the connection would be made (H-A-O-H or H-O-A-H) compared to the single stop option (H-A-H).

A.5 Activity Transport Mode

The transport mode of the cases can be determined by independent variables at household/individual, activity-program and tour level. The same variables as in Table 2.1 can be used for the household/individual level, while the activity-program and tour-level variables are summarised in Table A.8. The footnote [8] points to earlier definitions of duration alternatives in the table for duration decisions, S is the current schedule and C is the concerned tour.

Activity-program variables concern the total time engaged in work and the number of second, mandatory activities, such as service, shopping and the bring/get activities. Together, these variables indicate the 'workload' of the program. Possibly, a high workload may lead to a preference for fast modes so as to increase the remaining time for leisure and social activities.

The tour-level variables cover various aspects. First, the time-of-day when the tour is undertaken is potentially relevant as it may determine the degree of congestion on the road network during travelling. However, at this stage the start time of the tour is not exactly known. The exact departure time will be dependent on the mode used for the tour. For example, a fast mode allows one to delay the departure time, while keeping the time engaged in the activities itself constant. Moreover, the start time and duration of flexible activities are flexible. To account for the freedom of choice on all these dimensions, we included a variable that determines the earliest possible start time of the tour.

[8] see earlier definitions of duration alternatives in Section A.2

Table A.8: Independent variables used in the ‘transport mode for other than work activities’ choice facet of Albatross

Name	Description	Categories
Nsec	Number of mandatory out-of-home activities other than work in S	$0:0; 1:1; 2:2; 3:3-4; 4:>4$
Two	Total time of Work1 in S (in minutes)	$0:0; 1:\leq 240; 2:241-360;$ $3:361-480; 4:>480$
Ttot	Total time of Work1 and Work2 in S	$1:\leq 120; 2:121-240; 3:241-360;$ $4:361-480; 5:>480$
CBT	Earliest possible begin time of C	1: $<$ 10 AM; 2: 10-0 AM; 3: 0-2 PM; 4: 2-4 PM; 5: 4-6 PM; 6: $>$ 6 PM
Aty1	Type of the first activity in C	1: work; 2: bring/get; 3: grocery; 4: service; 5: non-grocery; 6: leisure; 7: social; 8: other
Aty2	Type of the second activity in C	0: home; 1: work; 2: bring/get; 3: (non-)grocery or service; 4: leisure or social; 5: other
Adur1	Duration of the first activity in C	1: short; 2: average; 3: long[8]
Awith1	Travel party of the first activity in C	0: none; 1: only others inside the household; 2: others outside the household involved
Cbrget	Bring or get activity is part of C	0: no; 1: yes
Cgroc	Grocery activity is part of C	0: no; 1: yes
Cserv	Service activity is part of C	0: no; 1: yes
Cshop	Non-daily shopping activity is part of C	0: no; 1: yes
Csoco	Social activity is part of C	0: no; 1: yes
Cleiso	Leisure activity is part of C	0: no; 1: yes
Cnlout	Non-leisure activity is part of C	0: no; 1: yes
TTbike	Shortest travel time by bike for tour C (in minutes)	$0: 0; 1:\leq 5; 2:6-15; 3:16-25$ $4:26-35; 5:36-60; 6:>60$
Rcabi	Travel time ratio between car and bike (%)	$1:\leq 25, 2:26-33$ $3:34-85; 4:>85$
Rpubi	Travel time ratio between public transport and bike (%)	$1:\leq 100; 2:101-200$ $3:201-260; 4:>260$
Rpuca	Travel time ratio between public transport and car (%)	$1:\leq 100; 2:101-700$ $3:701-900; 4:>900$
Textra2	Extra bike travel time to reach location of order 2 (minutes)	0: 0; 1: $\leq$ 10; 2: 11-15; 3: 16-30; 4: $>$ 30
Textra3	Same for order 3	0: 0; 1: 1-15; 2: 16-20; 3: 21-35; 4: $>$ 35
Textra4	Same for order 4	0: 0; 1: 1-20; 2: 21-30; 3: 31-40; 4: $>$ 40
Pbrget	Partner has a bring/get activity during tour C	0: no; 1: yes
Pserv	Partner has a shopping or service during tour C	0: no; 1: yes
PTmax	Partner’s maximum bike travel time across activities during tour C (minutes)	0: 0; 1: 1-10; 2: 11-20; $3:21-40; 4:>40$
Avcar	Car is available given the work activity of the partner	0: no; 1: yes; 2: there is no partner or schedule partner is unknown

A second potentially important aspect of the tour is the tour's purpose. The dimensions of the first activity in the tour, such as the activity type, travel party and duration, are included as descriptors of the tour's purpose. In the majority of tours, which involve only one out-of-home activity, these variables suffice. To cover the general case where a tour involves multiple activities, we additionally use a series of variables indicating the absence or presence of an activity of each distinguished activity category in the tour.

Third, the required travel distance on the tour is a potential moderator of mode choice. We use the shortest-route bike time as an indicator of distance. Because locations of flexible activities are still unknown in this stage, the shortest-travel time across possible locations for the activity is taken as an index here[9] Additionally, the tour-specific relative speed of each mode is measured in terms of travel-time ratios between car/bnike, public transport/bike and public transport/car.

Mode choice may further depend on the required travel distance to reach locations of higher order. Fast modes probably reduce the disutility of travel and therefore may be preferred in cases where the individual wishes to visit a higher-location at a relative long marginal distance. The fourth set of tour-level variables, therefore, defines the extra bike-travel time required to reach locations for each higher-order location. The bike times calculated relate to the first activity only and assume a home-based trip[10] In case of fixed activities (no location choice) and social activities (no higher-order locations), the marginal distances are set to zero. Note that these variables describe location choice options and, consequently, allow the system to anticipate on location choices in choosing a mode.

Furthermore, the activity schedule of the partner, if any, may compete with the use of car in households where there is only one car available. The fifth set of tour-level variables describe the presence of a bring/get activity, the presence of a shopping or service activity and the maximum bike-travel time across the partner's tour that necessarily overlap in time with the tour concerned. Overlapping tours are identified by comparing latest possible start times and earliest possible end times. If there is no partner, the partner's schedule is unknown or there are no overlapping tours, the variables are set to zero.

The final independent variable defines the availability of the car-driver mode. Car

[9] If the tour involves more than one flexible activity, the shortest travel time is calculated based on home-based distances. This was done to avoid the computational complexity of optimising the choice of multiple locations simultaneously.

[10] This covers the majority of cases in our data sets as about 70 % of the tours involve only a single activity.

driver is considered not available if there is no car available in the household or the person has no driving license or is otherwise incompetent to drive. Furthermore, the option is not available if there is only one car and this car is in use by the partner for work (the first schedule decision). Hence, three groups are distinguished: there is no car available (Avcar = 0), there is a car available (Avcar = 1) and there is no partner of the schedule of the partner is unknown (Avcar = 2). The other mode alternatives - car passenger, public transport and slow mode - are considered always available. At least in these data set, (almost) every person has a bike and a public transport link exists between every origin-destination in the study area (although the required travel time may vary strongly). Nevertheless, the time available for the tour may rule out public transport or slow modes and dictate the use of the car. In the present system, however, this is not taken into account, because uncertainties about available times, possible locations and travel times at this stage of the process make it hard to evaluate this constraint.

A.6 Locations

The different categories of the choice facet location1 are:

- Hmin: the nearest location from home;
- Cmin: the nearest location in the context of the tour;
- Cext5: the highest-order location within 5 minutes;
- Cext10: the highest-order location within 10 minutes;
- Cext20: the highest-order location within 20 minutes;
- Cmax: the highest-order location;
- Other: none of the foregoing.

The required travel time and the order of the location are used as selection criteria. Travel times are based on the observed mode. They are derived from travel-time matrices considering either the home location (Hmin) or the entire tour (Cmin-Cmax) as the basis for the trip(s). The Cext5 - Cext20 categories assume a maximum acceptable travel time and select the highest order location within reach. We assume that individuals define acceptable travel times relative to the travel time required for the nearest location (tour based). Therefore, the maximum levels - 5, 10 and 20 minutes - are defined as extra travel time over the minimally required travel time (tour-based).

Table A.9: Definition of location orders in Albatross

Order	daily shopping (floor space) in m^2	non-daily shopping (floor space)	service (♯ of outlets)	leisure (♯ of outlets)
1	1-1000	1-4000	1-40	1-5
2	1001-2000	4001-10000	41-80	6-10
3	> 2000	10001-50000	81-800	11-100
4	-	50001-70000	-	> 100
5	-	> 70000	-	-

Location order, on the other hand, is defined dependent on activity type and the size of facilities available at the location. The exact definitions are given in Table A.9.

An observed location belongs to the 'other' category, if it matches none of the others, because the location is inferior in terms of travel time and order or it does not meet the maximum travel-time constraints of Cext5, Cext10 or Cext20.

Just as in previous choice facets, each case is described at different levels including the household/individual, activity program/schedule, tour and activity level. The variables used to describe household/individuals are the same as in Table 2.1. The other variables are described in Table A.10. The footnote [11] points to earlier definitions of duration alternatives in the table for duration decisions, S is the current schedule, A the current activity and C is the concerned tour.

Activity-program variables describe the activity load of the schedule in terms of the amount of time engaged in and number of out-of-home, non-leisure activities, such as work/school, shopping, personal business and others. Tour-level variables determine the number of activities conducted on the tour, the nature of the previous and the next activity and whether the trip to the activity starts and/or ends at home. Activity-level variables describe the profile of the (flexible) activity in terms of activity type, duration, travel party and start time (time of day). The classifications used for these dimensions correspond to the actions of choice facets used in earlier stages.

The final set of variables, AvCmin - AvCmax, has a special status. They determine, e.g., that Cmin is not available in cases where the nearest home location (Hmin) is identical with the nearest tour-based location (Cmin).

For the location2 choice facet, we consider only the 'other' locations from the previous choice facet. This facet now selects a travel-time band comprising the location where the activity is to be performed. travel times are evaluated exclusively in the

[11] see earlier definitions of duration alternatives in Section A.2

Table A.10: Independent variables used in the 'location' choice facets of Albatross

Name	Description	Categories
Twincl	Total time of Work1 inclusive travel in S (in minutes)	$0:0; 1:\leq 260; 2:261-380;$ $3:381-500; 4:>500$
Ttot	Total time of Work1 and Work2 in S	$1:\leq 240; 2:241-360$ $3:361-480; 4:>480$
Nsec	Number of mandatory out-of-home activities other than work in S	$0:0; 1:1; 2:2; 3:3-4;$ $4:4-5; 5:>5$
Atype	Activity type	1: daily shopping; 2: service; 3: non-daily shopping; 4: social; 5: leisure
Mode	Transport mode	1: car (driver or passenger); 2: slow; 3: public
Adur	Activity duration	1: short; 2: average; 3: long^{11}
Awith	Travel party of A	0: none; 1: only others inside the household; 2: others outside the household involved
Tiday	Start time of A	1: $<$ 10 AM; 2: 10-0 AM; 3: 0-2 PM; 4: 2-4 PM; 5: 4-6 PM; 6: $>$ 6 PM
Tmax	Maximum available time in the schedule position of the activity (inclusive travel times)	0: 0; 1: 1-30; 2: 31-60; 3: $>$ 60
Nout	Number of out-of-home activities in C	1: 1; 2: 2; 3: $>$ 2
fromH	Trip to A starts from home	0: no; 1: yes
toH	Trip from A ends at home	0: no; 1: yes
Aprev	Type of previous activity	0: home; 1: work; 2: other mandatory; 3: social or leisure
Anext	Type of next activity	0: home; 1: work; 2: other mandatory; 3: social or leisure
AvCmin	Cmin location is available in choice set	0: no; 1: yes
AvCext5	Cext5 location is available in choice set	0: no; 1: yes
AvCext10	Cext10 location is available in choice set	0: no; 1: yes
AvCext20	Cext20 location is available in choice set	0: no; 1: yes
AvCmax	Cmax location is available in choice set	0: no; 1: yes

context of the concerned tour (as opposed to home-based). As before, the time limits are defined in terms of the extra travel time relative to the minimum travel time across locations of the choice facet. In that way, the following bands are defined: 0-5 (Cext5), 6-10 (Cext10), 11-20 (Cext20), 21-30 (Cext30) and more than 30 minutes (Cext>30). Furthermore, locations outside the study area are considered as an additional choice category (Othout). We use the same set of variables to describe the cases as in the location1 choice facet (see table A.10). Only the availability variables are redefined to indicate the presence of locations in the (reduced) choice set lying within each of the successive time bands.

Samenvatting

Het modelleren van verplaatsingen is altijd van groot belang geweest in transport onderzoek. In de jaren 50 was er, wegens de snelle toename in autogebruik, nood aan modellen die de vraag naar transport konden voorspellen. In die dagen werden verplaatsingen beschouwd als het resultaat van vier achtereenvolgende stappen, die apart gemodelleerd werden: het genereren van trips, het verdelen van trips over de verschillende zones, de keuze van vervoermiddel en de toekenning van de route. Deze modellen staan ook bekend als het trip-gebaseerde of vierstapsmodel (Ruiter en Ben-Akiva, 1978). Een nadeel van deze modellen is dat de interacties tussen trips in tijd en ruimte genegeerd worden. Daarom werden er vanaf het midden van de jaren zeventig toer-gebaseerde systemen ontwikkeld (Daly *et al.*, 1983). Deze modellen combineren verschillende trips in een toer die vertrekt en aankomt thuis of op het werk. Doch, opnieuw kwam er kritiek op deze modellen, omdat de verplaatsing nog steeds als een geïsoleerd gegeven beschouwd werd en de reden waarom men trips onderneemt, werd nog steeds verwaarloosd.

Dit heeft ertoe geleid dat er in de loop van het laatste decennium een verandering opgetreden is binnen het verkeersonderzoek. De interesse in ruimtelijke interactiepatronen is verschoven van trip- en toer-gebaseerde modellen naar de analyse van complexe dagelijkse activiteitenpatronen en de hiermee gepaard gaande verplaatsingen (Bhat en Koppelman, 1999). De basisgedachte achter activiteitengebaseerde modellen is dat het verplaatsingsgedrag van personen bepaald wordt door de activiteiten die individuen of huishoudens wensen te ondernemen. De verplaatsing op zich is slechts één van de componenten binnen het rooster van activiteiten dat in tijd en ruimte gemodelleerd moet worden. Deze activiteitengebaseerde modellen pogen ook de duur van de activiteiten te voorspellen, wanneer en op welke locatie ze worden uitgevoerd, welk vervoermiddel wordt gebruikt om tot op deze locaties te komen, enz. Het huishouden en andere sociale structuren hebben natuurlijk een invloed op het verplaatsings- en activiteitengedrag van personen, net zoals de onderlijke afhankelijkheid tussen personen, vervoermiddelen, etc.

Deze activiteitengebaseerde aanpak van verplaatsingsgedrag heeft geleid tot verschillende modellen waarbij voornamelijk 2 grote groepen tot stand zijn gekomen, deze van de simultane en deze van de sequentiële modellen. De eerst groep van de nuts-maximaliserende modellen vindt zijn oorsprong in de micro-economie en psychologie, de tweede groep van computationele procesmodellen werd geïnspireerd door psychologische beslissingsprocestheorieën. De simultane modellen gaan ervan uit dat een individu per dag een complete set patronen van activiteiten met bijbehorende verplaatsingen evalueert en er vervolgens het patroon uitkiest dat zijn of haar nutsfunctie maximaliseert. Een grote groep onderzoekers was echter gekant tegen dit soort modellen, ze betwisten het feit dat personen altijd tot een 'optimale' keuze zouden komen en gaan ervan uit dat individuen heuristieken gebruiken die verschillend kunnen zijn al naar gelang de context waarin men zich bevindt. In zijn meest simpele vorm zijn deze modellen opgebouwd uit een verzameling van ALS-DAN regels: ALS aan voorwaarde X voldaan is, DAN wordt actie Y ondernomen.

Eén van de meest geavanceerde én het enige operationele sequentiële model tot op heden is het Albatross systeem (Arentze en Timmermans, 2000). Dit systeem, dat we zullen gebruiken doorheen dit proefschrift, wordt uitgebreid toegelicht in Hoofstuk 3. De oorspronkelijke data uit Hendrik-Ido-Ambacht en Zwijndrecht, twee gemeentes uit de regio ten zuiden van Rotterdam, worden ook hier gebruikt. Deze staan beschreven in Hoofdstuk 2 en in de Appendix. Het systeem heeft als doel te voorspellen welke activiteiten waar uitgevoerd worden, wanneer, voor hoe lang, met wie en welk transportmiddel voor de verplaatsing gebruikt wordt. Deze beslissingen bepalen de negen verschillende keuzefacetten (ook wel dimensies genoemd) van het systeem en een sequentiële uitvoering van de negen modellen die op hun beurt telkens één van deze facetten trachten te voorspellen, levert de voorspelde activiteitenpatronen. De performantie van het systeem wordt getest op drie niveaus: op het niveau van de keuzefacetten, van de activiteitenpatronen en van de tripmatrices. Op dit eerste niveau wordt er per dimensie nagegaan hoe goed elk model de respons kan voorspellen, dit wordt gemeten aan de hand van de accuraatheid. Verder wordt er ook wat dieper ingegaan op welke verklarende variabelen het belangrijkst zijn om tot deze voorspelling te komen. Op het niveau van de activiteitenpatronen vergelijkt men de geobserveerde en de voorspelde sequenties van activiteiten door middel van Sequence Alignment Methodes (SAM) (Joh, *et al.*, 2001a, 2001b, 2001c, 2002a). Deze methodes meten het verschil tussen 2 sequenties in termen van de inspanning die nodig is om de twee sequenties gelijk te maken met behulp van invoeging, verwijdering en substitutie. Hoe lager de SAM maat, hoe gelijker de sequenties zijn. Tot slot test men de performantie van het Albatross systeem ook op tripmatrix niveau. De geobserveerde en voorspelde oorsprong-bestemming matrices worden met elkaar vergeleken door middel van een correlatiecoëfficiënt. Deze matrices bevatten het aantal trips ondernomen van een bepaalde oorsprong (rij) naar een bepaalde bestemming (kolom). Een bepaald type van modellen presteert dus goed binnen het systeem als de accuraatheid per dimensie behoorlijk is, als de SAM maten laag zijn en de correlatiecoëfficiënten op tripmatrix niveau dicht bij 1 liggen.

De vraag of men de voorkeur moet verlenen aan complexe of aan meer eenvoudige modellen, is waarschijnlijk even oud als de wetenschap zelf. Occam's scheermes ('Nunquam ponenda est pluralitas sin necesitate' - Niets moet onnodig verveelvuldigd worden), een smeekbede voor eenvoud, dateert zelfs uit de Middeleeuwen (Tornay, 1938), doch men moet erover waken dat deze stelling juist verklaard wordt. 'Eenvoud is een doel op zich' is in essentie de juiste interpretatie, terwijl 'Eenvoud leidt tot een grotere accuraatheid' dit niet is. Het antwoord op die vraag naar eenvoud hangt natuurlijk af van het doel dat men met een bepaald model wil bereiken. Als men een model beoogt met een hoge voorspellings- en veralgemeningsgraad, dan zullen complexere modellen dit doel waarschijnlijk het best dienen. Indien men echter het grote geheel wil zien en bereid is om hiervoor wat kleinere details op te offeren, dan kan een eenvoudiger model soelaas brengen. Indien men geconfronteerd wordt met een grote verzameling van verklarende variabelen die een bepaalde uitkomst moeten voorspellen, dan kunnen deze kleinere details storend werken. In de psychologie (Gigerenzer *et al.*, 1999) beroept men zich ook vaak op het feit dat men in het dagelijkse leven eenvoudigweg de tijd niet heeft om de verschillende mogelijkheden in overweging te nemen, en daarom kiest men voor eenvoudigere modellen. Denk bv. aan een dokter in het operatiekwartier, waar net een patiënt wordt binnengevoerd die een hartaanval krijgt. De beslissing van deze arts kan een leven redden of er een kosten, en hij/zij heeft niet de tijd om uitgebreid overleg te plegen. Slechts een paar metingen moeten uitwijzen welke acties ondernomen moeten worden. Dus, eenvoudigere modellen kunnen een oplossing bieden waneer men vooral geïnteresseerd is in de hoofdeffecten die de uitkomst beïnvloeden. Men kan deze eenvoudigere modellen aan de ene kant bekomen door eenvoudige heuristieken toe te passen, of door technieken toe te passen die een selectie maken uit een grote verzameling van verklarende variabelen.

Deze vraag naar complexiteit of eenvoud, en welke van de twee beter dienst doet in de context van transportmodellen vormt een rode draad doorheen dit proefschrift. Met betrekking tot het activiteitengebaseerde Albatross systeem gaan we in Hoofdstuk 4 en 5 na of eenvoudige modellen beter, even goed of slechter presteren dan het standaard geïmplementeerde CHAID inductie algoritme (Kass, 1980) dat we classificeren onder de complexere modellen. Dit zal onderzocht worden op niveau van de keuzefacetten, dus voor elk van de dimensies apart, maar ook op een meer geaggregeerd niveau, op basis van de voorspelde activiteitenpatronen. Gesteund door onderzoek binnen het domein van de psychologie dat ons leert dat menselijk gedrag vaak goed voorspeld wordt door eenvoudige modellen, gaan we deze stelling in Hoofdstuk 4 toetsen in een transportomgeving. Twee verschillende manieren om de complexe verzameling van regels per dimensie te vereenvoudigen worden bekeken. De eerste manier steunt op eenvoudige heuristieken (Zero R, One R (Holte, 1993) en Naïve Bayes (Langley *et al.*, 1992)) die gebruikt worden om de negen Albatrossdimensies te voorspellen, terwijl in de tweede manier twee gelijkaardige analyses uitgevoerd worden. In de eerste analyse wordt een C4.5 boom (Quinlan, 1993) gebouwd met behulp van alle variabelen, terwijl in de tweede analyse enkel een bepaalde set van variabelen gebruikt wordt, die geselecteerd werden aan de hand van een variabele selectie techniek, Relief-F (Kononenko, 1994). Het is immers zo dat irrelevante variabelen een sterk effect kunnen hebben op de structuur van de boom. De resultaten van de verschillende analyses leren ons dat een sterke reductie in het aantal verklarende variabelen dat gebruikt wordt (én bijgevolg in de complexiteit van de bomen) niet noodzakelijk leidt tot een verlies in de voorspellende kracht van het systeem.

Een uitbreiding op deze eenvoudige modellen wordt gebracht in Hoofdstuk 5. Bagging en boosting (Breiman, 1996; Freund en Schapire, 1997) zijn twee technieken die in de laatste tien jaar geïntroduceerd werden als ideeën om de accuraatheid van een voorspelling te verbeteren. Bagging doet dit door de variantie van de voorspelling te verminderen, terwijl boosting eigenlijk een gewogen gemiddelde voorspelling oplevert. Deze bagging en boosting technieken worden toegepast op de One R modellen uit Hoofdstuk 4 en op de modellen na variabele selectie. De resultaten worden opnieuw vergeleken op de drie niveaus: per dimensie apart, op niveau van de activiteitenpatronen en op tripniveau. We mogen concluderen dat de resultaten zeker die van de modellen zonder bagging of boosting overtreffen, en dat ze zelfs op geaggregeerd niveau het beste resultaat over Hoofdstuk 4 en 5 heen behalen.
Deze resultaten bevestigen de stelling dat (een combinatie van) eenvoudige modellen niet noodzakelijker slechter presteert dan de complexere, ook binnen een activiteitengebaseerd transportmodel.

Een tweede onderzoekspunt situeert zich binnen één specifiek facet van activiteitengebaseerde modellen, nl. de keuze van vervoermiddel. De verschillende datasets die hiervoor gebruikt worden, zijn weerom allemaal beschreven in Hoofdstuk 2. Door evolutie op verschillende domeinen wordt de onderzoeker nu geconfronteerd met twee nieuwe problemen: aan de ene kant zijn de datasets waarmee gewerkt moet worden veel groter geworden, zowel wat het aantal variabelen betreft als het aantal gegevens en aan de andere kant is er in het laatste decennium een overvloed aan nieuwe technieken ontwikkeld. Hoofdstuk 6 focust op het eerste probleem: hoe kan men de fit van een lineair model nagaan in dergelijke hoogdimensionele datasets, terwijl Hoofdstuk 7 een aantal van deze nieuwe technieken toepast binnen vervoersmodellen.

Een typische setting die vaak gebruikt wordt om na te gaan in welke omstandigheden een bepaald vervoermiddel geprefereerd wordt boven de andere is de logistische regressie setting. Om na te gaan in hoeverre het model de data goed fit, wordt vaak gebruik gemaakt van de Pearson teststatistiek, doch deze kan enkel gebruikt worden als alle verklarende variabelen categorisch van aard zijn. Verschillende methoden (zie bv. Hosmer en Lemeshow, 1980; Hart, 1997) werden ontwikkeld om ook continue variabelen mee in rekening te kunnen brengen, zonder dat men deze eerst moet categoriseren, maar de meeste technieken verliezen aan kracht van zodra het aantal verklarende variabelen toeneemt. We stellen in Hoofdstuk 6 een nieuwe teststatistiek voor, die het nulmodel contrasteert met een niet-parametrisch model dat gebaseerd is op het algoritme dat gebruikt wordt voor classificatiebomen (Breiman *et al.*, 1984). Deze teststatistiek laat toe dat zowel categorische als continue variabelen gemodelleerd worden. Verschillende simulatiesettings en de resultaten op werkelijke datasets tonen aan dat deze teststatistiek een veel belovende kracht heeft om fout gemodelleerde variabelen of vergeten variabelen te ontdekken, zelfs in hoge dimensies (i.e. veel gegevens en veel variabelen). Bovendien, als het lineaire nulmodel niet aanvaard wordt, dan geeft de classificatieboom, als niet-parametrische tegenhanger van het parametrisch nulmodel, aan in welke mate het nulmodel verbeterd kan worden. Om een dieper inzicht te verkrijgen in de asymptotische nulverdeling van deze boom-gebaseerde teststatistiek, is echter verder theoretisch onderzoek nodig.

In Hoofdstuk 7 wordt gefocust op de performantie van niet-lineaire en semi-lineaire modellen en de resultaten worden vergeleken met deze van de lineaire modellen. Deze semi- en niet-lineaire modellen leiden vaak tot een eenvoudiger model wat het aantal parameters betreft, maar van de andere kant ook tot een moeilijker wat betreft de definitie van het model. Fractionele polynomen (Royston en Altman, 1994) (semi-lineair) toegepast binnen een logistische regressie setting, classificatie en regressie bomen (Breiman *et al.*, 1984) en support vector machines (Vapnik, 1996) (niet-lineair) worden met elkaar vergeleken op basis van sensitiviteit, accuraatheid en specificiteit in dalende volgorde van belangrijkheid. Veronderstel bv. dat er slechts weinig cases zijn waarin openbaar vervoer gebruikt wordt, dan zijn het toch juist die cases die men juist wil voorspellen. Uit Hoofdstuk 7 volgt dat (semi-)lineaire modellen in het algemeen goed presteren op deze scheve datasets en dat het bijgevolg de moeite loont om

hier wat dieper op in te gaan. Dit bevestigt tevens het nut van de boom-gebaseerde teststatistiek die in Hoofdstuk 6 ontwikkeld werd.
Support vector machines (SVM) werden ontwikkeld om een zo hoog mogelijke accuraatheid te behalen en presteren bijgevolg duidelijk beter bij meer gebalanceerde datasets. Bij dergelijke SVM-modellen schuilt dan wel het gevaar dat ze gaan overfitten op de traningset, bij classificatie en regressiebomen kan dit wat opgevangen worden door snoeitechnieken toe te passen. Het nadeel van de SVM-methode is dat het echt een 'zwarte doos' techniek is: het resultaat wordt verkregen, maar de manier waarop wordt niet getoond en het is niet interpreteerbaar. Bomen zijn wel interpreteerbaar, maar niet in dezelfde mate als de voorgestelde (semi-)lineaire modellen waar de impact van elke variabele op de respons afzonderlijk kan bepaald worden.

Hoofdstuk 8 vat alle resultaten nog eens samen in een conclusiehoofdstuk. Er wordt dieper ingegaan op de consequenties voor het modelleren van verplaatsingen en op de voor- en nadelen van de verschillende modellen. Verder worden de gebruikte technieken onderling vergeleken op basis van vier criteria die van belang zijn bij transportonderzoek: kracht om het gevraagde te voorspellen, interpreteerbaarheid, robuustheid en gevoeligheid voor beleidsmaatregelen.

References

Advisory Committee on Cancer Prevention. (2000) Recommendations on cancer screening in the European Union. *European Journal of Cancer*, **36**, 1473–1478.

Aerts, M., Claeskens, G. and Hart, J. (1999) Testing the fit of a parametric function. *Journal of the American Statistical Association*, **94**, 869–879.

Aerts, M., Claeskens, G. and Hart, J. (2000) Testing lack of fit in multiple regression. *Biometrika*, **87**, 405–424.

Aerts, M., Claeskens, G., Hart, J., Moons, E. and Wets, G. (2003) Two lack of fit tests for multiple logistic regression. *Proceedings of the 18th International Workshop on Statistical Modelling*, Verbeke, G., Molenberghs, G., Aerts, M., and Fieuws, S. (Eds.), Leuven: Katholieke Universiteit Leuven, 15–20.

Aerts, M., Claeskens, G. and Hart, J. (2004) Bayesian-motivated tests of function fit and their asymptotic frequentist properties. *Annals of Statistics*, **32(6)**, 2580–2615.

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. *Proceedings of the 2nd International Symposium on Information Theory*, Petrov, B. and Cski, F. (eds.), Akadémiai Kiadó, Budapest, 267–281.

Arbyn, M., Van Oyen, H., Lynge, E. and Mickshe, M. (2001) European consensus on cancer screening should be applied urgently by health ministers. *British Medical Journal*, 323–396.

Arentze, T.A., Hofman, F., Kalfs, N. and Timmermans, H.J.P. (1999) (sylvia) System for logical verification and inference of activity diaries. *Transportation Research Record*, **1660**, 156–163.

Arentze, T., Hofman, F., van Mourik, H., Timmermans, H. and Wets, G. (2000) Using decision tree induction systems for modeling space-time behavior. *Geographical Analysis*, **32**, 330–350.

Arentze, T.A., Borgers, A., Hofman, F., Fujii, S., Joh, C., Kikuchi, A., Kitamura, R., Timmermans, H.J.P. and van der Waerden, P. (2000) Rule-based versus utility-maximizing models of activity-travel patterns. *Proceedings of the 9th international association for travel behaviour research conference*, Gold Coast, Queensland, Australia.

Arentze, T.A. and Timmermans, H.J.P. (2000) *Albatross: A Learning-Based Transportation Oriented Simulation System.* Eindhoven University of Technology, EIRASS.

Arentze, T.A. and Timmermans, H.J.P. (2002) *Albatross 2.0.* Eindhoven University of Technology, EIRASS.

Arentze, T.A. and Timmermans, H.J.P (2003) Measuring the goodness-of-fit of decision-tree models of discrete and continuous activity-travel choice: methods and empirical illustration. *Journal of Geographical Systems*, **4**, 1–22.

Arentze, T., Hofman, F. and Timmermans, H. (2003) Reintroduction of Albatross' decision rules using pooled activity-travel diary data and extended set of land use and cost-related condition states. *Proceedings of the 82nd Annual Meeting of the Transportation Research Board*, Washington, D.C., USA.

Arentze, T.A. and Timmermans, H.J.P. (2004) A learning-based transportation oriented simulation system. *Transportation Research B*, **38(7)**, 613–633.

Axhausen, K. and Gärling, T. (1992) Activity-based approaches to travel analysis: conceptual frameworks, models and research problems. *Transport Reviews*, **12**, 324–331.

Axhausen, K.W. (2000) Activity-based modelling: Research directions and possibilities. Internal paper n. 48, IVT ETH Zurich.

Azzalini, A., Bowman, A.W. and Haerdle, W. (1989) On the use of nonparametric regression for model checking. *Biometrika*, **76**, 1–11.

Ben-Akiva, M. and Lerman, S. (1985) *Discrete Choice Analysis.* M.I.T. Press, Cambridge, MA.

Ben-Akiva, M.E., Bowman, J. and Gopinath, D. (1996) Travel demand model system for the information area. *Transportation*, **25**, 241–266.

Ben-Akiva, M.E. and Bowman, J.L. (1998) Integration of an activity-based model system and a residential location model. *Urban Studies*, **35(7)**, 1231–1253.

Ben-Yacoub, S. (1999) Multi-modal data fusion for person authentication using SVM. *Proceedings of the 2nd International Conference on Audio- and Video-based Biometric Person Authentication*, Chellapa, R. (ed.), 25–30.

Bhat, C.R. (1996) A generalized multiple durations proportional hazard model with an application to activity behavior during the work-to-home commute. *Transportation Research*, **30B**, 465–480.

Bhat, C.R. (1998) Accommodating variations in responsiveness to level-of-service variables in travel mode choice modeling. *Transportation Research A*, **32(7)**, 495–507.

Bhat, C.R. (1999) A comprehensive and operational analysis framework for generating the daily activity travel profiles of workers. *Paper presented at the 78th Annual Meeting of the Transportation Research Board*, Washington, D.C., USA.

Bhat, C.R. and Koppelman, F. (1999) A retrospective and prospective survey of time use research. *Transportation*, **26**, 119–129.

Bhat, C.R. and Singh, S.K. (2000) A comprehensive daily activity-travel generation model system for workers. *Transportation Research A*, **34(1)**, 1–22.

Bhat, C.R., Guo, J., Srinivasan, S. and Sivakumar, A. (2004) Comprehensive econometric microsimulator for daily activity-travel patterns. *Proceedings of the 83rd Annual Meeting of the Transportation Research Board*, Washington, D.C., USA.

Borgers, A.W.J., Timmermans, H.J.P. and van der Waerden, P.J.H.J. (2002) Patricia: predicting activity-travel interdependencies using a suite of choice-based, inter-linked analyses. *Transportation Research Record*, **1807**, 145–153.

Bowman, J.L. and Ben-Akiva, M.E. (1995) Activity-based model system of urban passenger travel demand. *Paper presented at the 74th Annual Meeting of the Transportation Research Board*, Washington, D.C., USA.

Bowman, J.L. (1998) *The Day Activity Schedule Approach to Travel Demand Analysis.* Ph.D. dissertation, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge.

Bowman, J.L., Bradley, M., Shiftan, Y., Lawton, T.K. and Ben-Akiva, M.E. (1998) Demonstration of an activity-based model system for Portland. *Paper presented at the 8th World Conference on Transport Research*, Antwerp, Belgium.

Box, G.E.P. and Tidwell, P.W. (1962) Transformations of the independent variables. *Technometrics*, **4**, 531–550.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) *Classification and Regression Trees.* Wadsworth Statistics/Probability Series.

Breiman, L. and Friedman, J.H. (1985) Estimating optimal transformations for multiple logistic regression and correlation (with discussion). *Journal of the American Statistical Association*, **80**, 580–619.

Breiman, L. (1996) Bagging predictors. *Machine Learning*, **26**, 123–140.

Breiman, L. (1998) Arcing classifiers (with discussion). *Annals of Statistics*, **26**, 801–849.

Brown, C.C. (1982) On a goodness of fit test for the logistic model based on score statistics. *Communications in Statistics-Theory and Methods*, **11**, 1087–1105.

Brown, M.P.S., Grundy, W.G., Lin, D., Cristianini, N., Sugnet, C., Ares, M. and Haussler, D. (1999) Support vector machine classification of microarray gene expression data. University of California, Santa Cruz, Technical report UCSC-CRL-99-09.

Buckinx W., Moons, E., Van den Poel, D. and Wets, G. (2004) Customer-adapted coupon targeting using feature selection. *Expert Systems with Applications*, **26(4)**, 509–518.

Buntine, W. and Niblett, T. (1992) A further comparison of splitting rules for decision-tree induction. *Machine Learning*, **8**, 75–86.

Burges, C.J.C. (1998) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, **2(2)**, 121–167.

Chapin, F.S. (1974) *Human Activity Pattern in the City.* Wiley, New York.

Chen, S., Samingan, A.K. and Hanzo, L. (2001) Support vector machines multiuser receiver for DS-CDMA signals in multipath channels. *IEEE Transactions on Neural Networks*, **12(3)**, 604–611.

Clark, P. and Niblett, T. (1989) The CN2 induction algorithm. *Machine Learning*, **3**, 261–283.

Coleman, D., Day, N., Douglas, G., Farmery, E., Lynge, L., Philip, J. and Segnan, N. (1993) European guidelines for quality assurance in cervical cancer screening. *European Journal of Cancer*, **29A, suppl. 4**, 1–38.

Cressie, N. and Read, T. (1984) Multinomial goodness of fit tests. *Journal of the Royal Statistics Society, Series B*, **46**, 440–464.

Cristianini, N. and Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge University Press.

Daly, A.J., van Zwam, H.H. and van der Valk, J. (1983) Application of disaggregate models for a regional transport study in The Netherlands. *Paper presented at the 3rd World Conference on Transport Research*, Hamburg, Germany.

Davison, A.C. and Hinkley, D.V. (1997) *Bootstrap Methods and their Application.* Cambridge University Press.

Dijst, M. and Vidakovic, V. (1997) Individual action space in the city. *Activity-Based Approaches to Activity Analysis*, Ettema, D.F. and Timmermans, H.J.P. (eds.), Pergamon Press, Oxford, 73-88.

Doherty, S.T. and Miller, E.J. (2000) A computerized household activity scheduling survey. *Transportation*, **27**, 75–97.

Doherty, S. (2001) Classifying activities by time horizon using machine learning algorithms. *Paper presented at the 80th Annual meeting of the Transportation Research Board*, Washington, D.C., USA

Domingos, P., Pazzini, M. (1996) Beyond independence: conditions for the optimality of the simple Bayesian classifier. *Machine learning: Proceedings of the thirteenth international conference*, Saitta, L. (ed.), Morgan Kaufmann, 105–112.

Domingos, P. (1998) Occam's two razors: The sharp and the blunt. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, 37–43.

Duda, R.O. and Hart, P.E. (1973) *Pattern Classification and Scene Analysis.* New York, NY, Wiley.

Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Annals of Statistics*, **7**, 1–26.

Ettema, D.F., Borgers, A.W.J. and Timmermans, H.J.P. (2000) SMASH (A Simulation Model of Activity Scheduling Heuristics): An empirical test. *Geographical and Environmental Modelling*, **4**, 175–187.

Eubank, R.L. and Hart, J.D. (1992) Testing goodness-of-fit in regression via order selection criteria. *Annals of Statistics*, **20**, 1412–1425.

Faes, C., Geys, H., Aerts, M. and Molenbergs, G. (2003) Use of fractional polynomials for dose-response modeling and quantitative risk assessment in developmental toxicity studies. *Statistical Modelling*, **3**, 109–125.

Feller, W. (1968) *An Introduction to Probability Theory and Its Applications.* Wiley, New York.

Fosgerau, M. (1998) PETRA: an activity based approach to travel demand analysis. *Paper presented at the 8th World Conference on Transport Research*, Antwerp, Belgium.

Fried, M., Havens, J. and Thall, M. (1977) Travel behavior – A synthesized theory. Final Report, NCHRP, Transportation Research Board, Washington, D.C., USA.

Friedman, J.H. (1991) Multivariate adaptive regression splines. *Annals of Statistics*, **19(1)**, 1–67.

Friedman, J., Hastie, T. and Tibshirani, R. (2000) Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *Annals of Statistics*, **28(2)**, 337–407.

Freund, Y. and Schapire, R.E. (1997) A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences*, **55(1)**, 119–139.

Fujii, S., Kitamura, R. and Monma, T. (1997) A study of commuter's activity patterns for the estimation of induced trips. Manuscript, Kyoto University.

García-Pérez, M.A. and Núñez-Antón, V. (2004) Small-sample comparisons for goodness-of-fit statistics in one-way multinomials with composite hypotheses. *Journal of Applied Statistics*, **31(2)**, 161–181.

Gärling, T., Brännäs, K., Garvill, J., Golledge, R.G., Gopal, S., Holm, E. and Lindberg, E. (1989) Household activity scheduling. *Transport Policy Management and Technology Towards 2001: Selected Proceedings of the 5th World Conference on Transport Research*, **4**, Western Periodicals, Ventura, CA, 235–248.

Gärling, T., Laitila, T. and Westin, K. (1998) Theoretical foundations of travel choice modeling: introduction. *Theoretical Foundations of Travel Choice Modeling*, Gärling, T. Laitila, T. and Westin, K. (eds.), Elsevier, Oxford, 1–32.

Geys, H., Molenberghs, G. and Williams, P. (2002) Analysis of toxicology data with covariates specific to each observation. *Journal of Agricultural, Biological and Environmental Statistics*, **7**, 176–190.

Gigerenzer, G., Todd, P.M. and the ABC Research Group. (1999) *Simple Heuristics That Make Us Smart.* Oxford University Press, New York.

Gliebe, J.P. and Koppelman, F.S. (2000) A model of joint activity participation. *Paper presented at the 9th International Association for Travel Behavior Conference*, Gold Coast, Queensland, Australia.

Golob, T.F. (1997) A simultaneous model of activity participation and trip chain generation by households. *Prepared for presentation at the 8th International Conference on Travel Behavior Research*, Austin, Texas, USA.

Good, I.J. (1965) *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* M.I.T. Press.

Gratama, J.W., van der Nat, H., Weiland, H.T., Stijnen, T., Fibbe, W.E., Vossen, J.M.J.J., Willemze, R. and Verdonck, L.F. (1992) Intensification of GVHD prophylaxis interferes with the effects of pretransplant herpes virus serology on the occurence of grades II-IV acute graft-versus-host-disease. *Annals of Hematology*, **64**, A137–A139.

Haelterman, M. (1999) Kanker in België. National Cancer Registry.

Hägerstrand, T. (1970) What about people in regional science? *Papers of the Regional Science Association*, **24**, 7–21.

Hall, M.A. (1999a) *Correlation-based Feature Selection for Machine Learning.* Ph.D. dissertation, Department of Computer Science, University of Waikato, Hamilton.

Hall, M.A. (1999b) Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper. *Proceedings of the Florida Artificial Intelligence Symposium (FLAIRS)*, Orlando, Florida, USA.

Hart, J.D. (1997) *Nonparametric Smoothing and Lack-of-fit Tests.* New York: Springer Verlag.

Hastie, T.J. and Tibshirani, R.J. (1990) *Generalized Additive Models.* London: Chapman and Hall.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning; Data Mining, Inference, and Prediction.* Springer Series in Statistics.

Haykin, S. (1994) *Neural Networks: A Comprehensive Foundation.* Macmillan College Publishing Company, Englewood Cliffs.

Hens, N., Bruckers, L., Arbyn, M., Aerts, M. and Molenberghs, G. (2002) Classification tree analysis of cervix cancer screening in the Belgian health interview survey 1997. *Archives of Public Health*, **60**, 275–294.

Holte, R.C., Acker, L. and Porter, B.W. (1989) Concept learning and the problem of small disjuncts. *Proceedings of the eleventh international joint conference on artificial intelligence*, Morgan Kaufmann, 813–818.

Holte, R.C. (1993) Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, **11**, 63–90.

Horton, N.J., Bebchuk, J.D., Jones, C.L., Lipsitz, S.R., Catalano, P.J., Zahner, G.E.P. and Fitzmaurice, G.M. (1999) Goodness-of-fit for gee: an example with mental health service utilization. *Statistics in Medicine*, **18**, 213–222.

Hosmer, D.W. and Lemeshow, S. (1980) A goodness-of-fit test for the multiple regression model. *Communications in Statistics*, **A10**, 1043–1069.

Hosmer, D.W., Lemeshow, S. and Klar, J. (1988) Goodness-of-fit testing for the logistic regression model when the estimated probabilities are small. *Biometrical Journal*, **30**, 911–924.

Hosmer, D.W. and Lemeshow, S. (1989) *Applied Logistic Regression.* Wiley, New York.

Hosmer, D.W., Hosmer, T., Lemeshow, S. and le Cessie, S. (1996) A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, **16**, 965–980.

Huigen, P.P.P. (1986) Binnen of buiten bereik?: Een sociaal-geografisch onderzoek in Zuid-West Friesland. Nederlandse Geografische Studies 7, University of Utrecht, Utrecht, The Netherlands.

Hunt, E.B., Marin, J. and Stone, P.J. (1966) *Experiments in Induction.* Academic Press, New York.

Janssens, D., Wets, G., Brijs, T. and Vanhoof, K. (2004a) Improving the performance of a multi-agent rule-based model for activity pattern decisions using Bayesian networks. *Proceedings of the 83rd Annual Meeting of the Transportation Research Board*, Washington, D.C., USA.

Janssens, D., Wets, G., De Beuckeleer, E. and Vanhoof, K. (2004b) Collecting activity-travel diary data by means of a new computer-assisted data collection tool. *Proceedings of the 11th European Concurrent Engineering Conference*, Hasselt, Belgium, 85–89.

Joh, C-H, Arentze, T.A. and Timmermans, H.J.P. (2001a) Multidimensional sequence alignment methods for activity-travel pattern analysis: a comparison of dynamic programming and genetic algorithms. *Geographical Analysis*, **33**, 247–270.

Joh, C-H, Arentze, T.A. and Timmermans, H.J.P (2001b) A position-sensitive sequence alignment method illustrated for space-time activity diary data. *Environment and Planning A*, **33**, 313–338.

Joh, C-H, Arentze, T.A. and Timmermans, H.J.P. (2001c) Pattern recognition in complex activity-travel patterns: a comparison of Euclidean distance, signal processing theoretical and multidimensional sequence alignment methods. *Transportation Research Record*, **1752**, 16–22.

Joh, C-H, Arentze, T.A. and Timmermans, H.J.P. (2001d) Towards a theory and model of activity-travel rescheduling behavior. *Proceedings of the 9th World Conference on Transportation Research*, Seoul.

Joh, C-H, Arentze, T.A., Hofman, F. and Timmermans, H.J.P. (2002a) Activity-travel pattern similarity: a multidimensional alignment method. *Transportation Research B*, **36**, 385–403.

Joh, C-H, Arentze, T.A. and Timmermans, H.J.P. (2002b) Modeling individuals' activity-travel rescheduling heuristics: theory and numerical experiments. *Transportation Research Record*, **1807**, 16–25.

Joh, C-H. (2004) *Measuring and Predicting Adaptation in Multidimensional Activity-Travel Patterns.* Ph. D. Dissertation, Urban Planning Group, Technische Universiteit Eindhoven.

Jones, P.M., Dix, M.C., Clarke, M.I. and Heggie, I.G. (1983) *Understanding Travel Behavior.* Gower, Aldershot.

Jovicic, G. (2001) *Activity Based Travel Demand Modelling.* Danmarks Transport-Forskning.

Kass, G.V. (1980) An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, **29**, 119–127.

Kim, H. and Loh, W.-Y. (2001) Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, **96**, 589–604.

Kira, K. and Rendall, L.A. (1992) A practical approach to feature selection. *Proceedings of the 9th International Conference on Machine Learning*, Aberdeen, Scotland, UK, Sleeman, D.H. and Edwards, P. (eds.), Morgan Kaufmann Publishers, San Mateo, 249–256.

Kitamura, R. and Fujii, S. (1998) Two computational process models of activity-travel choice. *Theoretical Foundations of Travel Choice Modeling*, Gärling, T., Laitila, T. and Westin, K. (eds.), Elsevier, Oxford, 251–279.

Kockelman, K.M. (1997) Travel behavior as function of accessibility, land use mixing, and land use balance: Evidence from San Francisco bay area. *Transportation Research Record*, **1607**, 116–125.

Kohavi, R., Becker, B. and Sommerfield, D. (1997) Improving simple bayes. *Poster papers of the 9th European conference on machine learning*, 78–87.

Koller, D. and Sahami, M. (1996) Toward optimal feature selection. *Proceedings of the 13th International Conference on Machine Learning*, Bari, Italy, Saitta, L. (ed.), 284–292.

Kononenko, I. (1994) Estimating attributes: analysis and extensions of relief. *Proceedings of the 7th European Conference on Machine Learning*, Catania, Italy, Bergadano, F. and De Raedt, L. (eds.), Springer Verlag, 171–182.

Kruskal, J.B. (1983) An overview of sequence comparison. *Time Warps, String Edits, and Macromolecules*, Sankoff, D. and Kruskal, J.B. (eds.), Addison-Wesley, London, 265–310.

Kuchibhatla, M., and Hart, J.D. (1996) Smoothing-based lack-of-fit tests: variations on a theme. *Journal of Nonparametical Statistics*, **7**, 1–22.

Kwan, M.-P. (1997) GISICAS: An activity-based travel decision support system using a GIS-interfaced computational process model. *Activity Based Approaches to Activity Analysis*, Ettema, D.F. and Timmermans, H.J.P. (eds.), Pergamon Press, Oxford, 263–282.

Langley, P., Iba, W. and Thompson, K. (1992) An analysis of Bayesian classifiers. *Proceedings of the 10th national conference on artificial intelligence*, AAAI Press and MIT Press, 223–228.

Le Cessie, S. and van Houwelingen, J.A. (1991) A goodness-of-fit test for binary data based on smoothing residuals. *Biometrics*, **47**, 1267–1282.

Le Cessie, S. and van Houwelingen, J.A. (1993) Building logistic models by means of a non parametric goodness of fit test: a case study. *Statistica Neerlandica*, **47**, 97–109.

Le Cessie, S. and van Houwelingen, H.C. (1995) Testing the fit of a regression model via score tests in random effects models. *Biometrics*, **51**, 600–614.

Lenntorp, B. (1976) Paths in space-time environment: A time geographic study possibilities of individuals. *Lund Studies in Geography, Series B. Human Geography*, **44**, Department of Geography, The Royal University of Lund.

Lindsey, J.K. (2001) *Nonlinear Models in Medical Statistics.* Oxford University Press.

Loh, W.-Y. and Shih, Y.-S. (1997) Split selection methods for classification trees. *Statistica Sinica*, **7**, 815–840.

Lu, X. and Pas, E.I. (1997) An examination of activity time allocation on two consecutive days. *Prepared for presentation at the 8th International Conference on Travel Behavior Research*, Austin, Texas, USA.

Maclin, R. and Opitz, D. (1997) An empirical evaluation of bagging and boosting. *Proceedings of AAAI/IAAI*, 546–551.

McNally, M.G. (2000) The activity-based approach. Center for Activity Systems Analysis. Paper UCI-ITS-AS-WP-00-4.

Meka, S., Pendyala, R.M. and Wasantha Kumara, M. (2002) Structural equations analysis of within-household activity and time allocation between two adults. *Proceedings of the 81st Annual Meeting of the Transportation Research Board*, Washington, D.C., USA.

Mingers, J. (1989) An empirical comparison of pruning methods for decision tree induction. *Machine Learning*, **4(2)**, 227–243.

Misra, R., Bhat, C.R. and Srinivasan, S. (2003) A continuous time representation and modelling framework for the analysis of non-worker activity-patterns: Tour and episode attributes. *Paper presented at the 82nd Annual Meeting of the Transportation Research Board*, Washington, D.C., USA.

Moons, E., Wets, G., Vanhoof, K., Aerts, M. and Timmermans, H. (2001) How well perform simple rules on activity diary data. *Proceedings of the 7th International Computers in Urban Planning and Urban Management Conference*, Honolulu, USA.

Moons, E., Wets, G., Aerts, M. and Vanhoof, K. (2002a) The role of Occam's razor in activity based modeling. *Computational Intelligent Systems for Applied Research - Proceedings of the 5th International FLINS Conference*, Ruan, D., D'hondt, P. and Kerre, E.E. (Eds.), Gent, Belgium, 153–162.

Moons, E., Wets, G. and Aerts, M. (2002b) Goodness-of-Fit test based on decision trees. *Proceedings of the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences*, Valafar, F. (Ed.), Las Vegas, USA, 594–600.

Moons, E., Wets, G., Vanhoof, K., Aerts, M., Arentze, T. and Timmermans, H. (2002c) The impact of irrelevant attributes on the performance of classifier systems in generating activity schedules. *Proceedings of the 81st Annual Meeting of the Transportation Research Board*, Washington D.C., USA.

Moons, E., Aerts, M. and Wets, G. (2004a) A tree based lack-of-fit test for multiple logistic regression. *Statistics in Medicine*, **23**, 1425–1438.

Moons, E., Wets, G. and Aerts, M. (2004b) Nonlinear models in transportation. *Proceedings of Conference on Progress in Activity-Based Analysis*, Maastricht, The Netherlands,.

Moons, E., Aerts, M. and Wets, G. (2004c) The application of fractional polynomials and support vector machines in transportation analysis. *Paper accepted for presentation at the Joint Statistical Meetings*, Toronto, Canada.

Moons, E.A.L.M.G., Wets, G.P.M., Aerts, M., Arentze, T.A. and Timmermans, H.J.P. (2005a) The impact of simplification in a sequential rule-based model of activity scheduling behavior. Accepted for publication in: *Environment and Planning A*.

Moons, E., Wets, G., Aerts, M., Arentze, T. and Timmermans, H. (2005b) The impact of irrelevant attributes on the performance of classifier systems applied to certain aspects of the activity-based context. Conditionally accepted in: *Transportation*.

Moons, E., Aerts, M. and Wets, G. (2005c) Improving mode choice models: A tree based lack-of-fit test for multiple logistic regression. Submitted to *Transportation and Statistics*.

Moore, D.S. and Spruill, M.C. (1975) Unified large-sample theory of general chi-squared statistics for tests of fit. *Annals of Statistics*, **3**, 599–616.

Morgan, J.A. and Sonquist, J.N. (1963a) Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, **58**, 415–434.

Morgan, J.A. and Sonquist, J.N. (1963b) Some results from a non-symmetrical branching process that looks for interaction effects. *Proceedings of the Soc. Stats. Sec., ASA*, 40–53.

Neter, J., Kutner, M.H., Nachtsheim, C.J. and Wasserman W. (1996) *Applied Linear Statistical Models.* Irwin.

Nock, R. (2002) Complexity in the case against accuracy estimation. *Theoretical Computer Science (A)*, **301**, 143–165.

Nock, R. and Lefaucheur, P. (2002) A robust boosting algorithm. *Proceedings of the 13th European Conference on Machine Learning*, Elomaa, T., Mannila, H. and Toivonen, H. (eds.), Springer, 319–330.

Oza, N.C. and Russell, S. (2001) Online bagging and boosting. *Artificial Intelligence and Statistics*, Richardson, T. and Jaakkola, T. (eds.), 105–112.

Pas, E.I. and Harvey, A.S. (1991) Time use research: Implications for travel demand analysis and modelling. *Understanding Travel Behaviour in an Era of Change*, Stopher, P.R. and Lee-Gosselin, M. (eds.), Pergamon Press.

Pas, E.I. (2002) Time use and travel demand modeling: Recent developments and current challenges. *In Perpetual Motion: Travel Behavior Research Opportunities and Application Chalenges*, Mahmassani, H.S. (ed.), Pergamon Press, 307–331.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, California.

Pearson, K. (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophy Magazine Series (5)*, **50**, 157–172.

Pendyala, R.M., Kitamura, R. and Reddy, D.V.G.P. (1995) A rule-based activity-travel scheduling algorithm integrating neural networks of behavioral adaptation. *Paper presented at the EIRASS Conference on Activity-Based Approaches*, Eindhoven, The Netherlands.

Pendyala, R.M., Kitamura, R. and Reddy, D.V.G.P. (1998) Application of an activity-based travel demand model incorporating a rule-based algorithm. *Environment and Planning B*, **25**, 753–772.

Pendyala, R.M. (2004) FAMOS: Application in Florida. *Paper presented at the 83rd Annual Meeting of the Transportation Research Board*, Washington, D.C., USA.

Pulkstenis, E. and Robinson, T.J. (2002) Two goodness-of-fit tests for logistic regression models with continuous covariates. *Statistics in Medicine*, **21**, 79–93.

Purvis, C. (2003) Household travel survey reports. http://www.mtc.ca.gov/maps_and_data/datamart/survey/ (accessed February 18, 2005).

Quataert,, P., Van Oyen, H., Tafforeau, J. et al. (1998) Health interview survey 1997. Protocol for selection of the households and the respondents. SPH/Episerie No. 12, SPH, Brussels.

Quinlan, J.R. (1993) *C4.5 Programs for Machine Learning.* Morgan Kaufmann Publishers, San Mateo.

Recker, W.W., McNally, M.G. and Root, G.S. (1986) A model of complex travel behavior: Part 2: an operational model. *Transportation Research A*, **20**, 319–330.

Reinsch, C.H. (1967) Smoothing by spline functions. *Numerische Mathematik*, **10**, 177–183.

Rendell, L. and Seshu, R. (1990) Learning hard concepts through constructive induction. *Computational Intelligence*, **6**, 247–270.

Royston, P. and Altman, D.G. (1994) Regression using fractional polynomials of continuous covariates: parsimonious parametric modeling (with discussion). *Applied Statistics*, **43**, 429–467.

Royston, P., Ambler, G. and Sauerbrei, W. (1999) The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology*, **28**, 964–974.

Ruiter, E.R. and Ben-Akiva, M. (1978) Disaggregate travel demand models for the San Francisco bay area. *Transportation Research Record*, **673**, 121–128.

Sauerbrei, W. and Royston, P. (1999) Building multivariable prognostic and diagnostic models: transformation of the predictors using fractional polynomials. *Journal of the Royal Statistical Society, Series A*, **162**, 71–94.

Schwartz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Silverman, B.W. (1985) Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society, Series B*, **47**, 1-52.

The Corradino Group (2000) Southeast Florida regional travel characteristics study: household travel characteristics survey plan and findings. Technical report prepared for Florida department of transportation, Miami-Dade MPO, Broward County MPO, Palm Beach County MPO.

Timmermans, H.J.P., Arentze, T.A. and Joh C-H. (2000) Modeling learning and evolutionary adaptation processes in activity settings: theory and numerical simulations. *Transportation Research Record*, **1718**, 27–33.

Timmermans, H.J.P. (2001) Models of activity scheduling behaviour. *Stadt Regional Land*, **71**, 63–78.

Timmermans, H.J.P., Arentze, T.A. and Joh, C-H. (2002) Analyzing space-time behavior: new approaches to old problems. *Progress in Human Geography*, **26**, 175–190.

Tornay S. (1938) *Ockham: Studies and Selections.* La Salle, IL: Open Court.

Vapnik, V. (1996) *The Nature of Statistical Learning Theory.* Springer-Verlag, New York.

Verloove, S. and Verwey, R.Y. (1988) Project on preterm and small-for-gestational age infants in the Netherlands, 1983. University Microfilms International, no. 8807276. Ann Arbor, MI.

Veropoulos, K., Cristianini, N. and Campbell, C. (1999) Controlling the sensitivity of support vector machines. *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden.

Wang, D. and Timmermans, H.J.P. (2000) A conjoint-based model of activity engagement, timing scheduling and stop pattern formation. *Paper presented at the 79th Annual Meeting of the Transportation Research Board*, Washington, D.C., USA.

Weiss, S.M., Galen, R.S. and Tadepalli, P.V. (1990) Maximizing the predictive value of production rules. *Artificial Intelligence*, **45**, 47–71.

Wen, C-H. and Koppelman, F.S. (1999) An integrated system of stop generation and tour formation for the analysis of activity and travel patterns. *Transportation Research Record*, **1676**, 136–144.

Wets, G., Vanhoof, K., Arentze, T. and Timmermans, H. (2000) Identifying decision structures underlying activity patterns: an exploration of data mining algorithms. *Transportation Research Record*, **1718**, 1–9.

Whittaker, E.T. (1923) On a new method of graduation. *Proceedings of the Edinborough Mathematical Society*, **41**, 63–75.

Wickramaratna, J., Holden, S. and Buxton, B. (2001) Performance degradation in boosting. *Lecture Notes in Computer Science*, Kittler, J. and Roli, F. (eds.), 11–21.

Wilson, C. (1998) Analysis of travel behaviour using sequence alignment methods. *Paper presented at the 77th Annual Meeting of the Transportation Research Board*, Washington, D.C., USA.

Yamamoto, T. and Kitamura, R. (1997) An analysis of time allocation to in-home and out-of-home activities across working days and non-working days. Manuscript, Kyoto University.

Yamamoto, T., Fujii, S., Kitamura, R. and Yoshida, H. (2000) An analysis of time allocation, departure time and route choice behavior under congestion pricing. *Paper presented at 79th Annual Meeting of the Transportation Research Board*, Washington, D.C., USA.

Yamamoto, T., Kitamura, R. and Pendyala, R.M. (2003) Comparative analysis of time-space prism vertices for out-of-home activity engagement on working and non-working days. *Proceedings of the 82nd Annual Meeting of the Transportation Research Board*, Washington, D.C., USA.

Zhang, B-T and Mühlenbein, H. (1995) Balancing accuracy and parsimony in genetic programming. *Evolutionary Computation*, **3(1)**, 17–38.

Zhang, H.P. and Singer, B. (1999) *Recursive Partitioning in the Health Sciences.* Springer-Verlag, New York.

Zhang, J., Timmermans, H.J.P. and Borgers, A.W.J. (2002) A utility-maximizing model of time use incorporating group decisions mechanisms. *Proceedings of the 82nd Annual Meeting of the Transportation Research Board*, Washington, D.C., USA.

Zheng, C.L., de Sa, V.R., Gribskov, M. and Murlidharan Nair, T. (2003) On selecting features from splice junctions: an analysis using information theoretic and machine learning approaches. *Genome Informatics*, **14**, 73–83.