Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Transportation Research Part B 41 (2007) 772–783
TRANSPORTATION RESEARCH PART B
www.elsevier.com/locate/trb

Parametric action decision trees: Incorporating continuous attribute variables into rule-based models of discrete choice

Theo Arentze, Harry Timmermans *

Eindhoven University of Technology, Department of Building and Architecture, Urban Planning Group, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Received 10 January 2006; received in revised form 20 October 2006

Abstract

Rule-based models, such as decision trees, are ideally suited to represent discontinuous effects of independent variables on discrete choice behavior in transport or spatial systems. At the same time, however, the models require that continuous attributes, such as for example travel time and travel costs, are discretisized, which may decrease the sensitivity of predictions for policy measures that involve these attributes. To overcome this problem and combine the specific strengths of the rule-based and parametric modeling approaches, this paper introduces a hybrid approach. The so-called parametric action decision tree (PADT) replaces the conventional action-assignment rule of the decision tree by a logit model or any other parametric discrete choice model. The PADT includes alternative-specific constants to take the impact of leaf-node membership into account in addition to terms for the continuous attributes. As an illustration, we show how the approach can be used to incorporate travel-costs sensitivity in Albatross – a rule-based model of activity-travel choice. The results indicate that the enhanced, hybrid model can reproduce realistic ranges of price elasticities of travel demand.

Keywords: Discrete choice models; Rule-based models; Decision-tree induction; Activity-based models

1. Introduction

The problem of dealing with non-compensatory decision rules within a random-utility-maximizing framework of discrete choice has received continuous attention in travel-behavior research since Tversky's seminal work on behavioral decision making (Tversky, 1972). Swait (2001) proposed a non-compensatory choice model incorporating attribute cut-offs into the decision model and tested it as an extension of the traditional compensatory utility maximization framework. Cantillo and Dios Ortúzar (2005) proposed a hybrid semi-compensatory two-stage model incorporating thresholds for the acceptance of attributes in the process of

* Corresponding author. Tel.: +31 40 247 3315; fax: +31 40 243 8488.
E-mail address: eirass@bwk.tue.nl (H. Timmermans).

0191-2615/$ - see front matter
doi:10.1016/j.trb.2007.01.001

discrete choice, similar to earlier work by Timmermans et al. (1986). The rule-based formalism, such as the production system, provides an alternative framework, which is particularly suitable to represent discontinuous effects of attributes on choice behavior (Gärling et al., 1994). Rule-based models are founded on theories of bounded rationality (Newell and Simon, 1972) and can be extracted from choice observations through rule-induction methods that emerged from recent work in artificial intelligence, statistics and other disciplines. Several studies demonstrated the use of rule-induction methods to model spatial and travel choice behavior (Thill and Wheeler, 2000; Gahegan, 2000; Arentze et al., 2000; Wets et al., 2000; Yamamoto et al., 2002; Xie et al., 2003; Moons et al., 2005).

Compared to parametric methods, the strength of the rule-based approach is that it allows one to represent various types of interactions between variables, such as conditional relevance and conditional classification (Van der Smagt and Lucardie, 1991). Conditional relevance occurs if an attribute variable is relevant conditional upon some other variable. For example, in an elimination-by-aspects heuristic, the relevance of a certain attribute (i.e., aspect) is dependent on the performance of alternatives on a higher-level attribute. Only if the choice alternatives perform equally well on the first attribute, a second attribute is taken into consideration. The same holds for the next attribute in the hierarchy and so on. Conditional classification, on the other hand, is related to the phenomenon that an individual may be indifferent for variation on an attribute within certain sections of the domain of the attribute. In such cases, the number and position of cut-off points on the range of the variable determines how the individual classifies an outcome of the attribute. The classification may be dependent on other variables. An example is the perception of travel time of a trip in terms of labels such as short and long. The perception may be dependent on the transport mode used for the trip or situational settings such as the weather or time pressure. Conditional relevance/classification gives rise to discontinuous effects of attribute variables on choice probabilities. A change in situation or attributes of choice alternatives may have less or more than proportional impacts depending on whether or not they lead to a qualitative shift in perception.

This strength of rule-based approaches may at the same time be a weakness. The discrete representation of variables is not only a power but also a restriction of the formalism for variables measured on a continuous (interval or ratio) scale.[1] Some rule-induction methods are able to optimize the choice of cut-off-points simultaneously with constructing the tree.[2] However, even if classifications are flexible, the assumption of indifference between outcomes within classes is still imposed on the model. The property is a power if such limited sensitivity accurately describes the choice behavior modelled. However, if classifications underlying choice behavior in a population studied differ between individuals or situations not captured by condition states in the tree, the response behavior of the population may be better approximated by allowing for some sensitivity across the continuous range. In transport and spatial systems, continuous service-level variables, such as travel time and travel costs, are often important explanatory variables for choice facets of trips such as frequency, destination, transport mode and even departure time. If (condition-dependent) ranges of indifference on such service-level variables are not uniformly distributed in the population, the response to changes may vary in an approximately continuous fashion at an aggregate level and the rule-based model would fail to reproduce this. Furthermore, elasticities (i.e., marginal substitution rates) related to continuous variables, which are often relevant for policy making and theory development, are not revealed when their values are discretisized.[3]

To address this problem and combine the specific strengths of rule-based and parametric methods, we propose a more flexible formalism defined as an extension of the decision tree. In principle, several formalisms including decision trees, production systems, association rules, causal networks and decision tables can be applied (Witten and Frank, 2005; Keuleers et al., 2001; Janssens et al., 2004, 2005). The decision tree has

[1] Some rule-induction methods, such as CART (Breiman et al., 1984) and M5 (Quinlan, 1992), do allow *dependent* variable to be continuous and to be treated as continuous. Our argument here, however, is related to *independent* variables.

[2] An example of a rule-induction method that optimizes the discretization of attribute variables simultaneously with the structure of the tree is C4.5 (Quinlan, 1993).

[3] This is not to say that no elasticity information can be derived from rule-based models. Simulation methods have been proposed to extract elasticity information from rule-based model in a pre-processing step (Arentze and Timmermans, 2003a). However, for continuous variables elasticity information obtained in that way is limited if these variables are discretisized.

perhaps obtained most attention, for good reasons. It is a straightforward representation formalism that guarantees the consistency and completeness of the decision rules that are represented by the tree. In the following we will refer to the proposed formalism using the term Parametric Action Decision Tree (or, in short, PADT). A PADT avoids the need to discretisize continuous attribute variables by incorporating them in the action-assignment rule rather than the condition section of the tree. The action-assignment rule is complementary to a decision tree and determines a choice (of an action) for each new case classified by the tree. A commonly used action-assignment rule is the so-called plurality rule. Given the leaf node of a case, this rule assigns the action that is most frequently observed at the leaf node in the data set used to induce the tree (called, training data). The probabilistic variant of this rule assigns choice probabilities to action alternatives rather than a single action. The advantage of the probabilistic approach is that unexplained variance in behavior is reproduced in predictions. The PADT adopts a probabilistic assignment rule too. However, rather than a static probability distribution, it uses a parametric model for each leaf node. In the model, the influence of the condition state at the leaf node is represented by alternative-specific constants and the influence of continuous attributes by coefficients. In this way, predictions are maximally sensitive to both categorical and continuous attribute variables.

The purpose of this paper is to develop and provide an empirical illustration of the proposed PADT method. The illustration demonstrates how the method can be used to make activity-travel choice models sensitive to transportation pricing policies. The structure of the paper is straight-forward. First, we describe the proposed method and next we consider an application to illustrate the approach. We conclude the paper by discussing the major conclusions.

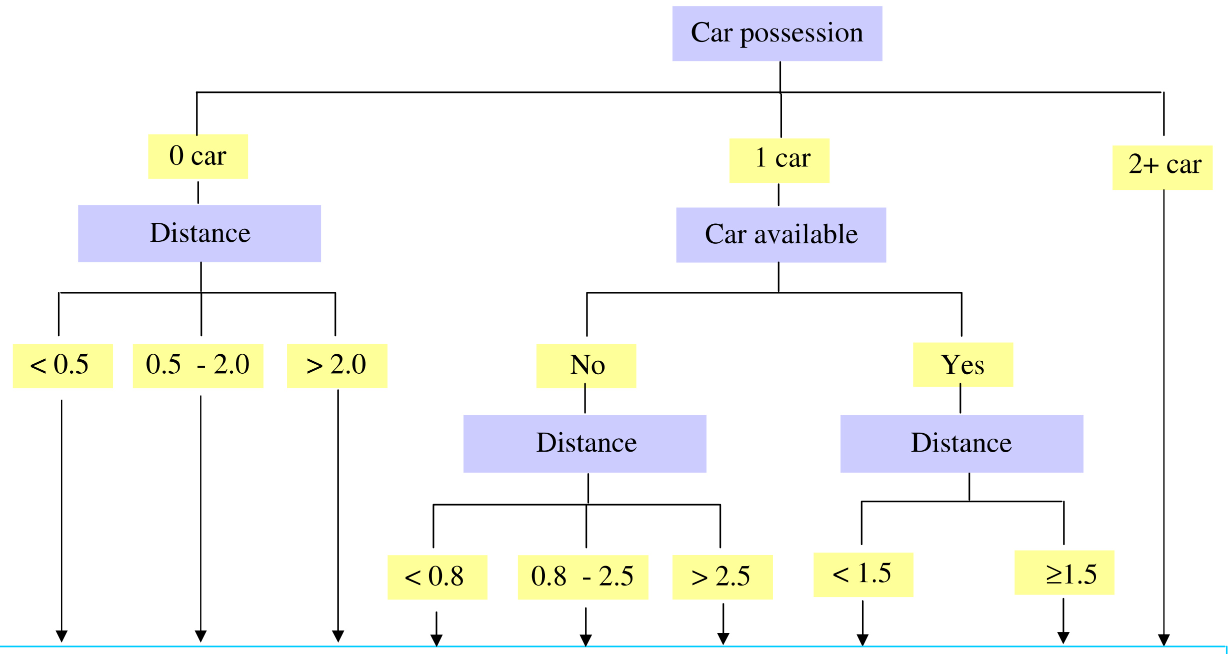
2. Method

In this section, we describe the proposed method focusing on the formalism and issues of empirical estimation. First, we discuss the conventional decision tree and the problem addressed in this study in more detail.

2.1. The conventional decision tree and problem

A decision tree is a directed acyclic graph satisfying the following properties (Safavian and Landgrebe, 1991). There is exactly one node, called the root, which no edges enter. The root contains all the class labels. Every node except the root has exactly one entering edge and there is a unique path from the root to each node. There are two types of nodes: decision nodes and leaves. Decision nodes specify a test which should be carried out on the value of an attribute (i.e., independent variable). Each possible outcome of the test results in a branch of the decision tree. Leaves are the terminal nodes of the tree. They specify to which class an unlabeled instance belongs. The problem of deriving a decision tree from a set of observations is defined as one of finding the most parsimonious tree structure that maximizes some measure of goodness-of-fit. As the problem is NP hard, heuristic methods have been proposed. Generally, they construct a tree by recursively splitting the sample of observations on attribute variables into partitions that are as homogeneous as possible in terms of the target variable. The various methods proposed differ in terms of the split criterion used and whether a stopping criterion or a two-staged process of growing and pruning a tree is implemented (Witten and Frank, 2005; Quinlan, 1993; Safavian and Landgrebe, 1991).

In the context of the decision problems we consider here, the classes represent choice options on some facet of an activity pattern. An arbitrary example of a decision tree is shown in Fig. 1. The tree determines the choice of transport mode for a given trip with known travel distance. The attribute variables of the tree include: number of cars in the household the trip-maker belongs to (Car-possession), availability of car during the time-window of the trip (Car available) and trip length (Distance, for example, in kilometers). The choice alternatives are the transport modes for trips that possibly can be observed in the domain considered. In the example, these include a slow mode, car and public transport (PT). The decision tree defines a choice-probability distribution across the transport modes for each leaf node. As splitting the training data is the basic operation of a decision-tree induction method, a leaf node defines a partition of the data. The choice-probability distribution at each leaf node is obtained by calculating the relative frequency distribution of choices in the partition corresponding to the leaf node. Using the model to predict the choice in a new case involves the



P(slow)	1.00	0.60	0.10	0.85	0.45	0.00	0.50	0.01	0.00
P(car)	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.85	0.98
P(PT)	0.00	0.40	0.90	0.15	0.55	1.00	0.00	0.14	0.02

Fig. 1. Arbitrary example of a conventional decision tree: choice of transport mode for a trip.

steps of assigning the case to a leaf node based on its attributes and next assigning an action using the probability distribution at the leaf node. In a deterministic action assignment, the action having the highest probability is selected. On the other hand, a probabilistic assignment returns the entire probability distribution as a prediction. Possibly, Monte Carlo simulation is used successively to make a decision for the case. In this way, unexplained variance is reproduced in predictions.

In this example, Distance is the only continuous variable. The variable occurs at several decision nodes of the tree and is classified in different ways depending on its predecessor nodes. Even though the influence of other variables is captured, there may remain variation in perceptions of outcomes in the studied population not captured by attributes at predecessor nodes, with as a consequence that predictions are biased. But even if the population is homogeneous or the model relates to a single agent (individual, household or company), discretization of continuous variables may still lead to unrealistic behavior of the model. As Yuan and Shaw (1995) argue, a small change in a continuous variable, if that leads to a reclassification of the situation, may give a quite drastic change in response. For example, consider for the model of Fig. 1 a small change in trip length from just below to just above 2.5 km in case of a 1-car household and car is not available for the trip. Then, the model would predict a rather abrupt shift in the probability distribution from slow mode to public-transport mode. Yuan and Shaw argue that such sudden changes may not always be realistic. As a solution, they propose fuzzification of the underlying decision tree. Using that method, a membership function is defined for each class used in the decision table such that cases are allocated to a condition state (i.e., a leaf node) with a certain degree instead of in an all-or-nothing fashion. Predicted probabilities are, then, calculated as a weighted mean of action probabilities across condition states using the membership value as a weight. They illustrate the approach with an application.[4]

As the Yuan and Shaw study shows, fuzzification of condition states potentially is an effective method to reduce the crisp character of rule-based systems in general. Fuzzification is also widely applied in models using a set of if-then rules rather than a decision tree. A disadvantage of fuzzification, however, is that membership functions are hard to estimate empirically. The genetic algorithm has been used as a heuristic to find the parameter values that best fit a set of observations. The rules themselves, however, are a priori defined instead

[4] Fuzzification can be seen as a way of smoothing probability distributions at leaf nodes of a decision tree. Another form of smoothing is sometimes also used as part of the method to induce a decision tree (see Quinlan, 1992), namely as an alternative for pruning a tree. The aim of smoothing in that case is to obtain more reliable estimates of distributions at those leaf nodes that are based on only few observations in the training data. Often this involves adjusting the probability of the leaf node by averaging the probability estimates along the path leading from the root node to this leaf node. In sum, smoothing as an alternative of pruning and fuzzification of condition states serve different purposes and are based on different techniques.

of empirically derived (by induction). Even if the estimation problem would be solved, the problem of discretization still remains. Although they are fuzzy, cut-off-points still need to be pre-defined ad-hoc and need to be limited in number. This may be justifiable in case of a homogeneous target population or a single agent. In other cases, however, the method would not solve the problem identified earlier. The PADT approach that we propose circumvents the problem by adopting continuous attributes as independent variables in a dynamic action section of rules, as explained below.

2.2. Formalism

In this section, we derive the PADT from first principles. Assume that subjects choose the alternative that maximizes a utility. We can write the utility of choice alternative j in a case i as:

$$U_{ij} = f_j(X_{ij}) + \varepsilon_{ij} \tag{1}$$

where U_{ij} is the utility of choice alternative j in case i, X_{ij} is a set of attributes and ε_{ij} is an error term. Index j is adopted in the subscript of f to represent the most general case where the utility function is alternative specific. The PADT approach relies on a distinct treatment of discrete and continuous variables. Assuming that utility is an additive function of the two sets of attributes, we can rewrite Eq. (1) as:

$$U_{ij} = f_j(X_{ij}^{\mathrm{D}}) + f_j(X_{ij}^{\mathrm{C}}) + \varepsilon_{ij} \tag{2}$$

where $X_{ij}^{\mathrm{D}} \subseteq X_{ij}$ is the subset of discrete attributes and $X_{ij}^{\mathrm{C}} \subseteq X_{ij}$ is the subset of continuous attributes. In the stage of inducing the decision tree, the continuous attributes can be left out of consideration since we may rewrite Eq. (2) as:

$$U_{ij} = f_j(X_{ij}^{\mathrm{D}}) + \varepsilon'_{ij} \tag{3}$$

where

$$\varepsilon'_{ij} = f_j(X_{ij}^{\mathrm{C}}) + \varepsilon_{ij} \tag{4}$$

In words: in generating the optimal tree, we consider variation of utilities related to continuous attributes as a component of error variance. We emphasize that decision-tree induction methods do not rely on any assumption regarding the distribution of the error terms. The purpose of induction is to create groups that are as homogeneous (i.e., pure) as possible regarding (action) choice probabilities. In the result, therefore, we can replace the set of discrete attributes by a single attribute representing leaf-node membership. The utility effect of leaf-node membership can be represented by an alternative-specific constant, so that we can write:

$$U_{ij} = \alpha_{kj} + \varepsilon'_{ij} \tag{5}$$

where k is the index of the leaf node in which case i falls and α_{kj} is the utility constant of alternative j at leaf node k. The constant represents the utility effect of the combination of attributes included in the path from the root of the tree to the leaf node. As implied by Eq. (4), a component of the error term relates to continuous attributes. In the action-assignment we can use the information of continuous attributes and reduce the error in predictions. For example, if we assume an additive functional form, we can further specify the model given by Eq. (5) as:

$$U_{ij} = \alpha_{kj} + \beta'_j X_{ij}^{\mathrm{C}} + \varepsilon_{ij} \tag{6}$$

where β_j is an attribute vector of coefficients and, as before, X_{ij}^{C} is an attribute vector of scores. As indicated by index j in the subscript of the beta coefficient, this equation assumes the most general case where the coefficient for each continuous attribute is alternative-specific. Finally, if we assume that error terms are IID Gumbel distributed, we can calculate the choice probabilities of a case i falling in leaf node k as the well-known MNL model:

$$P_{kj} = \frac{\exp(\alpha_{kj} + \beta'_j X_{ij}^{\mathrm{C}})}{\sum_{j'} \exp(\alpha_{kj'} + \beta'_{j'} X_{ij'}^{\mathrm{C}})} \tag{7}$$

Using this equation, we can calculate the choice probability distribution for an unseen case, given that we know the leaf node in which the case falls. Thus, deriving a prediction from a PADT consists of two steps: (1) determining the leaf node based on discrete attributes and (2) calculating the probabilities based on the leaf node found and the continuous attributes using Eq. (7).

Eq. (7) replaces the static action-assignment rule used in a conventional decision tree (DT). In that regard, it is worth noting that a DT can be represented as a special case of a PADT, namely a PADT in which the action-assignment model is defined as:

$$P_{ik} = \frac{\exp(\alpha_{kj})}{\sum_{j'} \exp(\alpha_{kj'})} \tag{8}$$

In this simple model, the maximum-likelihood estimates of the constants can be solved analytically as:

$$\alpha_{kj} = \ln(n_{kj}/n_{k0}) \quad j \neq 0 \tag{9}$$

$$\alpha_{k0} = 0 \tag{10}$$

where $j = 0$ is the base alternative and n_{kj} is the frequency of observing action j at leaf node k in the training set. Substituting (9) and (10) in (8) and rewriting results in

$$P_{ik} = \frac{n_{kj}}{\sum_{j'} n_{kj'}} \tag{11}$$

which is identical to the standard probabilistic action-assignment rule in conventional decision trees.

As an example, Fig. 2 schematically shows the PADT-alternative for the DT of Fig. 1. In agreement with the PADT method, the (continuous) distance attribute is no longer included in the induction stage and, consequently, does not recur as a variable at decision nodes of the tree. In this arbitrary example, this omission resulted in a reduction of the number of leaf nodes from 9 to 4. Furthermore, the static probability distributions at leaf nodes are replaced by a set of alternative-specific constants denoted as $\alpha_{11} - \alpha_{34}$. Since there is only one continuous attribute in this example, the logit model used for action-assignment includes only one beta parameter. Using the PADT to predict the transport mode for a given (new) trip involves first determining the leaf-node membership based on the Car-possession and Car-available attributes and next applying the logit model to predict the choice probabilities of modes taking into account the leaf-node membership as well as the distance traveled on the trip.

2.3. Estimation

As it follows from the foregoing, deriving a PADT from choice observations involves the steps of (i) deriving a decision tree using some decision-tree induction method and (ii) estimating the parameters of the

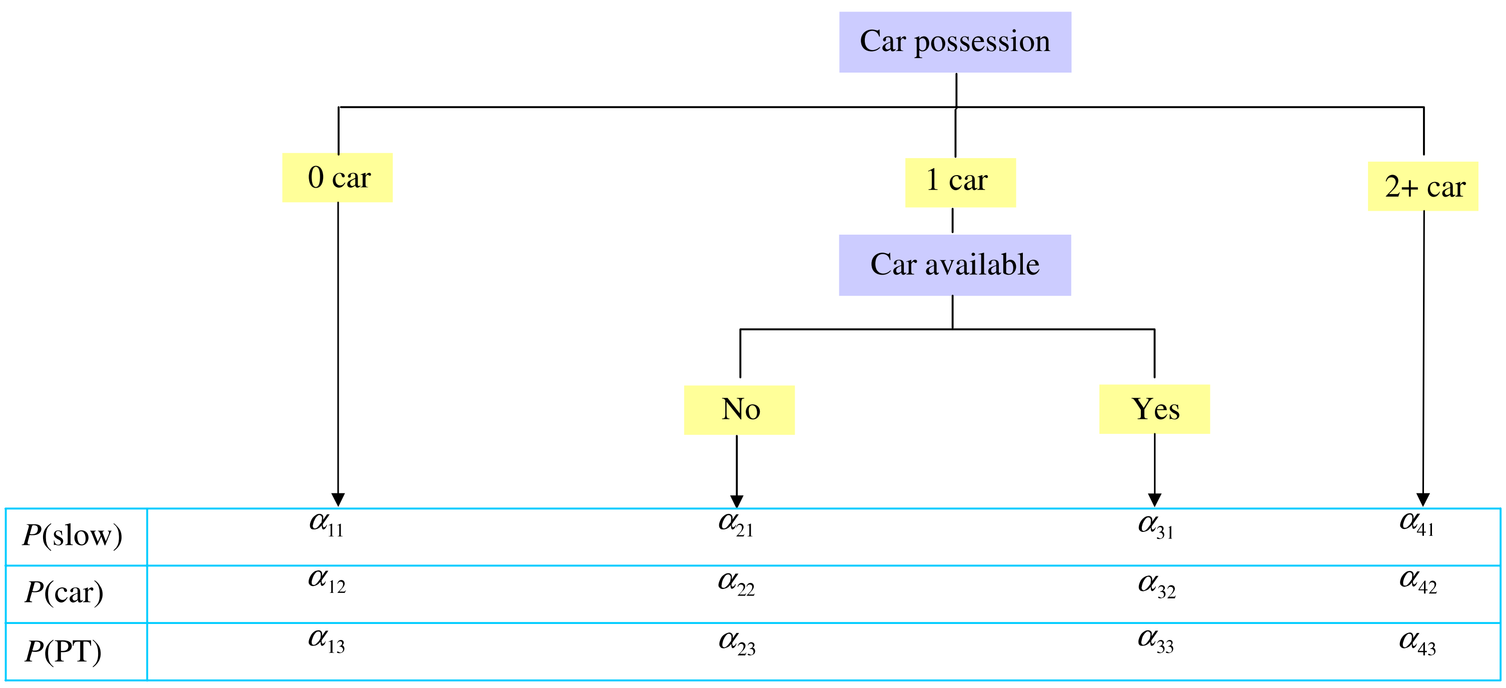


Fig. 2. Arbitrary example of a parametric action decision tree: choice of transport mode for a trip.

action-assignment rule. The decision tree defines a new categorical variable that has as many levels as leaf nodes, K. This variable together with the continuous variables constitute the independent variables used in the estimation in the second step. We emphasize that this latter step (just as the first step) does not require new methodologies. The same log-likelihood methods and statistical tests apply as those that are generally used in log-likelihood estimation. Assuming such a methodology as given, we focus here on some issues that deserve more attention.

The set of constants (alphas) and coefficients (betas) in the action-assignment rule, given by Eq. (7), are parameters that need to be estimated. Since utilities are measured on an interval scale, the constant of one of the choice alternatives per leaf node can be set to zero. This means that $K(J-1)$ constants and $J|C|$ coefficients needs to be estimated in the most general specification (K is number of leaf nodes of the tree, J the number of choice alternatives and C is the set of continuous attribute variables). The same sample of cases used in the induction stage of the PADT can also be used to estimate the parameters of the action-assignment rule using a standard maximum-likelihood method. Because of the relatively large set of constants, the degrees of freedom and hence the number of cases needed for estimation is relatively large. However, the expected minimum number of cases required is not larger compared to a DT (with probabilistic assignment). The constants replace the static choice probabilities at leaf nodes. This means that compared to a conventional DT, only the beta parameters introduce additional degrees of freedom. At the same time, excluding continuous attributes from the induction stage will lead to a reduction of the tree in terms of the number of leaf nodes. As a consequence, the additional parameters coincide with a decrease in the number of alpha parameters. As a likely result, the PADT is even more efficient in terms of number of parameters than an equivalent DT. Besides the total number of cases, also the number of cases at each leaf node deserves attention. To make sure that reliable estimates of constants are obtained at each leaf node, it is important to set a minimum to the number of cases per leaf node as a threshold for splitting in the induction stage. This minimum number does not need to be larger in case of a PADT compared to a DT for the same reason. Therefore, the PADT method has no implications for settings of a stopping criterion or pruning criterion, depending on the induction method used. Finally, it is worth noting that in many applications, there is no reason to assume that beta parameters need to be estimated alternative specific. In those cases, a reduction of degrees of freedom can be realized by estimating generic parameters reducing the number of coefficients from $J|C|$ to $|C|$.

2.4. Conditional dependence

As implied by Eq. (2), the PADT model relies on the assumption that utility effects of continuous attributes, on the one hand, and discrete attributes, on the other, are additive. Under this assumption, the optimal decision tree based on discrete attributes is independent of continuous attributes and, hence, continuous attributes can be ignored in the induction stage. In reality, the assumption may not hold. It is conceivable that the dependence of choice probabilities on some discrete attribute is conditional upon a continuous attribute. For instance, in the earlier example this would occur if the influence of Car-possession on transport-mode choice is dependent on the distance class in which the trip falls. In that case, a split on Car-possession may not be identified if the distance variable is not used for splitting in the induction procedure. We emphasize, however, that a situation of *pure* conditional dependence of the choice variable on an attribute variable is rather unlikely in real data sets. In conventional modeling terms, this would correspond to a situation where a variable has no significant main effect and only through an interaction with another variable has an impact on choices. Moreover, it is possible to control for this by means of a variant of the PADT. In this variant, continuous attributes are included in the induction stage as well as in the action-assignment stage. In this way, splits on attributes can be identified after splits on continuous variables have been implemented that otherwise would remain hidden. A potential problem of this variant is that it creates a correlation between leaf-node membership, on the one hand, and continuous attributes that have been used for splitting, on the other. If the correlation is substantial, multi-collinearity may lead to unreliable estimates of the beta parameters. The presence of (this form) of multi-collinearity can be easily tested just as in any model-estimation procedure. If it is problematic, the beta coefficients could be estimated leaf-node specific, provided that the partition of the sample at each leaf node is large enough. In this way, the beta parameters capture the systematic variation in choice probability remaining within subranges of the continuous variables that hold at the leaf node considered.

3. Empirical illustration

The illustration considered in this section focuses on using the PADT method to deal with travel costs in a rule-based and activity-based model of travel demand. Sensitivity for travel costs is an important characteristic of a model if it is to be used to assess impacts of price scenarios such as, for example, an increase in fuel price. In realistic scenarios of this kind, price changes are typically small (e.g., often not larger than 10%) and expected responses in terms of travel demands even smaller. For example, the typical price elasticity for car use is of the order of magnitude of 0.20 and, in case of a 10% increase in fuel price, this would mean an expected decrease of car kilometers of 2%. Needless to say that such small changes in conditions and responses impose high demands on the sensitivity of a model. In that sense, incorporating such costs sensitivity in a rule-based model is a rather strict test case for the PADT method.

3.1. The case

The rule-based model we consider here is the Albatross model – an activity-based model developed for the Dutch Ministry of Transportation and Public Works (Arentze and Timmermans, 2000, 2004; Arentze et al., 2003). Albatross implements a sequential decision making process to generate a schedule from scratch for a given individual (in the context of a household) and a given day of the week. Activity-diary data are used to induce a decision tree for each step in the scheduling process using a CHAID-based method. The data set used includes 9985 person-days collected in the period of 1997–2000 in different regions of the Netherlands. The data are combined with nation-wide data of the transport and land-use system (Arentze et al., 2003).

A price scenario may have an impact on various choice facets of activity-travel patterns. For a given activity, people may adapt the frequency or the location of the activity or change the transport mode and/or route of the trips for the activity. This means that costs sensitivity should be incorporated in several decision trees. As a case, we focus here on two decision trees both dealing with the work activity. The first tree determines for each individual and day whether or not a work activity is included in the schedule (for the day considered) and the second tree determines for each work activity in the schedule the transport mode used for the trip to the work location. In sum, we isolate the first two decision steps in the process model of Albatross and we show how a PADT can be developed as an alternative for the original DT for each of these two decisions. The PADTs are derived using the same activity and study-area data as used for the original DTs. Assuming that frequency and mode are the most relevant short-term adaptations to price changes people may consider, predicted price elasticities can be compared to known estimates of price elasticities (of work-related trips).

3.2. Method

Regarding the measurement of the travel-costs attribute we make the following assumptions. The (monetary) travel costs of a trip are defined as the *variable* costs. The variable costs of slow mode are zero. The variable costs of a trip by car-driver are calculated as the product of kilometers traveled and a nation-wide average fuel price per kilometer. With respect to Car-passenger, we roughly set the travel costs to half the costs estimated for the car-driver mode (assuming that the costs of the trip can be shared among multiple persons). Finally, public-transport costs of a trip are estimated as the least costs for the trip across the different mode alternatives within this category, i.e. bus, tram and metro taken together and train. In Albatross and most other activity-scheduling models, the transport mode decision is made after the location of the activity has been determined. This means that the destination of, in this case, a work-trip is known at the moment the transport mode decision is to be made. Therefore, for the mode decision, the travel costs related to choice alternatives can be readily calculated as explained. For the decision to include a work activity, on the other hand, determining the travel-costs attribute of each choice alternative (i.e., yes or no) is less straightforward. Clearly, the travel costs of not including a work activity are zero. The difficulty arises in estimating the costs associated with the positive choice. Since the location and transport mode are unknown at the moment of the decision, we must rely on *expected* costs. The analyses below use the mean travel costs across all work activities included in the activity diaries as an estimate of travel-costs of the yes-choice.

Furthermore, the linkage between the two decisions deserves attention. A change in the frequency of work activities in response to a price scenario, changes the base-rates of mode-choice alternatives in a next step. For example, if fuel costs increase, the model will respond by reducing the frequency of work activities in the population (e.g., people will work at home more often). Then, even if there would not be any effect of the costs increase on mode choice, we still would expect a decrease of car-trips because car-trips rather than trips by other modes will have been canceled. This illustrates the general phenomenon that work-selection adaptation has a consequence for base probabilities of mode choice alternatives. We take this into account by adjusting the predicted choice probabilities of modes as follows:

$$P_i'(m) = P_i(m) \times [1 + \mathrm{d}P(m)] \times \frac{1}{1 + \mathrm{d}P} \quad \forall m \tag{12}$$

where $P_i(m)$ is the predicted probability of mode m in case i before adjustment, $\mathrm{d}P(m)$ is the proportional change in number of m-trips and $\mathrm{d}P$ is the proportional change in number of trips overall. The last term on the right-hand side of the equation makes sure that the adjusted probabilities sum up to one across modes. Note that $\mathrm{d}P(m) = 0$ if there is no price change for m, $\mathrm{d}P(m) < 0$ if the price for m has increased and $\mathrm{d}P(m) > 0$ if the price for m has decreased. The value of $\mathrm{d}P$ is given by:

$$\mathrm{d}P = \sum_m \mathrm{d}P(m)P(m) \tag{13}$$

where $P(m)$ is the base choice probability of mode m. In many cases, a price scenario will relate to a single mode only. Then, $\mathrm{d}P(m) = 0$ for all m except the mode that is subject to the price change. To estimate the value of $\mathrm{d}P(m)$ for the single mode that is subject to the price change, the activity prediction should be run both under the price scenario and the null scenario. The difference in number of m trips between the runs equals the difference in number of trips predicted. In the general case where the price scenario affects multiple modes simultaneously the same principle applies. The price change for the different modes should be simulated separately so that the change found in trip rate indeed can be attributed to a change in number of trips of the mode subject to the price change. In the single and multiple price change case, the estimates can be found as:

$$\mathrm{d}P(m) = \begin{cases} \frac{\mathrm{d}P_m}{P(m)} & \text{if the scenario involves a price change for } m \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

where $\mathrm{d}P_m$ is the proportional change in trip rate under the scenario where only the change in price for m is implemented.

We emphasize that the issues of calculating costs and adjusting choice probabilities are not specific for the PADT method. Rather they are independent of the technique used to model the individual decisions and occur in any travel-behavior model that assumes a sequential decision process.

3.3. Results

The two decision trees were induced using observations from the entire set of 9985 diaries. Table 1 shows some statistics of the results of tree induction. A total of 8455 schedules provided the observations for the work-selection PADT. Not all diaries were useable because of missing values. The diaries included a total

Table 1
Some statistics of the tree induction results

	Work-selection PADT	Work-transport-mode PADT
Number of cases	8455	3275
Number of condition variables	21	39
Number of leaf nodes	45	52
Expected hit ratio	0.78	0.59
Expected hit ratio compared to NULL model	0.51	0.38

of 3275 work trips that provided observations for induction of the mode-choice PADT. The number of discrete attributes equaled 21 for the work-selection PADT and 39 for the mode-choice PADT. The attributes describe characteristics of the household, person, study area (accessibility of locations), and facets of the activity pattern as far as known in the stage of the scheduling process in which the decision tree operates. Since work-selection is the first scheduling decision, nothing is known yet about the activity pattern at the moment the decision is made in contrast to the mode-decision. This explains why more attribute variables are used for the latter PADT. The number of leaf nodes gives an indication of the size of each decision tree induced. The hit ratios give an indication of the goodness-of-fit of each model on the training data when a static probabilistic assignment rule is used as in a conventional DT. The first ratio expresses the expected percentage of correct predictions. The second ratio indicates the increase of the hit ratio compared to a null model, as a percentage of the maximum increase possible. The null model is defined as a root-node only tree meaning that aggregate choice probabilities are used for action-assignment. Thus, the latter hit ratio indicates the improvement in prediction achieved by splitting on discrete attributes (see Arentze and Timmermans, 2003b for a detailed discussion of the goodness-of-fit measures).

The beta parameter for each PADT represents the extent to which the choice under concern is sensitive (i.e., elastic) to costs. The sensitivity may be different for the two PADTs. For example, people may be more reluctant to change the frequency of work trips, in which case the mode decision would prove to be more sensitive. For illustration purpose, we did not estimate the beta parameters, but rather show for a range of beta values the price-elasticities predicted by the model in a scenario of a 10% increase of variable-costs for car. For the beta of the work-selection PADT two possible values (-2×10^{-4}, -4.5×10^{-4}) and for the beta of the mode-choice PADT four possible values (in the range from -1.4×10^{-4} to -4.5×10^{-4}) were considered. Note that price elasticity expresses the decrease in total kilometers traveled as a combined effect of a decrease in work trips and a switch from car to alternative transport modes. The elasticities were calculated based on the same sample of individuals that provided the data for inducing the tree. In summary, the procedure involves the steps: (1) assign values to the beta parameters; (2) generate maximum-likelihood estimates of the alpha parameters at leaf nodes for each PADT under the given beta value; (3) use the two PADTs to predict the work-selection and mode-choice decisions under the null scenario and the price scenario; (4) calculate the price elasticities by comparing the total distance traveled between the two scenarios.

Table 2 shows the predicted price elasticity and cross elasticities for each combination of beta values for work-selection and mode-choice. The results indicate that predicted cross-elasticities (i.e., Slow, Public, Car-passenger) are *in*sensitive for the value of the beta parameter for work-selection. This makes sense, since we expect that only car-based work trips are canceled in response to the scenario. The cross-elasticities *are* sensitive to variation of the beta parameter for mode-choice: they increase with increasing negative value of the parameter. This makes sense too, as the beta parameter for mode-choice determines the extent to which mode switches from car to an alternative mode will take place. Car-driver elasticity is sensitive for both parameters. Estimates of fuel price elasticities vary widely across studies. Graham and Glaister (2002) found on the basis of an extensive literature review that short-term price elasticities tend to be between -0.2 and -0.3. If we roughly take -0.25 as the best estimate, then the combination of values of -4.5×10^{-4} and -1.8×10^{-4} for the activity and mode costs parameter would correspond to best estimates of the betas.

Table 2
Predicted car price elasticities of travel demand for different settings of beta parameters for a 10% increase in car-travel costs

Beta work activity	Beta transport mode	Slow	Car-driver	Public	Car-passenger
-2×10^{-4}	-1.4×10^{-4}	0.05	−0.14	0.13	0.18
	-1.6×10^{-4}	0.06	−0.15	0.15	0.20
	-1.8×10^{-4}	0.07	−0.17	0.17	0.22
	-2.0×10^{-4}	0.08	−0.18	0.18	0.25
-4.5×10^{-4}	-1.4×10^{-4}	0.05	−0.23	0.13	0.18
	-1.6×10^{-4}	0.06	−0.24	0.15	0.20
	-1.8×10^{-4}	0.07	−0.25	0.17	0.22
	-2.0×10^{-4}	0.08	−0.26	0.18	0.25

4. Conclusion and discussion

The parametric action decision tree (PADT) introduced in this study intends to combine the specific strengths of rule-based models and parametric models of discrete choice. The method involves replacing the conventional assignment rule by a parametric model (e.g., a logit model) to better deal with continuous attribute variables. In the new model, the impact of discrete attributes is represented by alternative-specific constants for each leaf node and the impact of continuous attributes by coefficients. In the prediction stage, the tree is used to classify cases and the logit model to determine choice probabilities. In this way, a PADT overcomes the weakness of rule-based models of being less sensitive to continuous variation of level-of-service attributes of a transport system or land-use system, while it remains able to capture all context-dependencies and discrete variation. As an illustration, we showed how PADTs can be specified and derived to incorporate travel-costs sensitivity in an activity-based and rule-based model of travel demand called Albatross. Taking the work activity as a test case, the study indicated that the ranges of price elasticities found in empirical studies can be reproduced in a rule-based model using this approach.

In this paper we considered the PADT method as an extension of the conventional decision tree. However, positioning it alternatively as an extension of the logit model would be equally valid. For the specification of choice models in transport or spatial systems, generally, a large number of categorical variables is available as candidate independent variables. The number of degrees of freedom is often restrictive for the number of categorical variables and number of categories per variable that one can include in the model. Automated methods of interaction detection, such as for example CHAID, are well suited and have specifically been developed to identify the variables and classifications that are most significant for explaining choice observations. When used in a pre-processing stage, such methods allow one to make well-informed choices regarding the selection and categorization of discrete attribute variables. The PADT method proposed in this study can be seen as a more rigorous instance of this approach. In a PADT, the full set of discrete variables is replaced by a single composite categorical variable, i.e. leaf-node membership, constructed by decision tree induction. To the extent the heuristic behind decision-tree induction is effective, the new composite variable represents the most efficient parameterization of the discrete variables, given significance thresholds for splitting set by the modeler. The higher the threshold set, the more degrees of freedom available for continuous attribute variables and vice versa.

Acknowledgement

Arnout Schoemakers and Frank Hofman (Dutch Ministry of Transportation, Public Works and Water Management) are acknowledged for their support of this research project.

References

Arentze, T.A., Hofman, F., Timmermans, H.J.P., 2003. Re-induction of Albatross' decision rules using pooled activity-travel diary data and an extended set of land use and costs-related condition states. Transportation Research Record 1831, 230–239.

Arentze, T.A., Hofman, F., van Mourik, H., Timmermans, H.J.P., Wets, G., 2000. Using decision tree induction systems for modeling space–time behavior. Geographical Analysis 32, 330–350.

Arentze, T.A., Timmermans, H.J.P., 2000. Albatross: A Learning Based Transportation Oriented Simulation System. European Institute of Retailing and Services Studies, Eindhoven.

Arentze, T.A., Timmermans, H.J.P., 2003a. Measuring impacts of condition variables in rule-based models of space–time choice behavior: method and empirical illustration. Geographical Analysis 35, 24–45.

Arentze, T.A., Timmermans, H.J.P., 2003b. Measuring the goodness-of-fit of decision-tree models of discrete and continuous activity-travel choice: methods and empirical illustration. Journal of Geographical Systems 5, 185–206.

Arentze, T.A., Timmermans, H.J.P., 2004. A learning based transportation oriented simulation system. Transportation Research Part B 38, 613–633.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont (CA).

Cantillo, V., Dios Ortúzar, J. de, 2005. A semi-compensatory discrete choice model with explicit attribute thresholds of perception. Transportation Research Part B 39, 641–657.

Gahegan, M., 2000. On the application of inductive machine learning tools to geographical analysis. Geographical Analysis 32, 113–139.

Gärling, T., Kwan, M.-P., Golledge, R.G., 1994. Computational-process modelling of household activity scheduling. Transportation Research Part B 28, 355–364.

Graham, D.J., Glaister, S., 2002. The demand for automobile fuel: a survey of elasticities. Journal of Transport Economics and Policy 36, 1–25.

Janssens, D., Wets, G., Buys, T., Vanhoof, K., Timmermans, H.J.P., Arentze, T.A., 2004. Improving the performance of a multi-agent rule-based model for activity pattern decisions using Bayesian Networks. Transportation Research Record 1894, 75–83.

Janssens, D., Wets, G., Buys, T., Vanhoof, K., Timmermans, H.J.P., Arentze, T.A., 2005. Integrating Bayesian networks and decision trees in a sequential rule-based transportation model. European Journal of Operations Research 175, 16–34.

Keuleers, B., Wets, G., Arentze, T.A., Timmermans, H.J.P., 2001. Association rules in identification of spatial temporal patterns in multiday activity diary data. Transportation Research Record 1752, 32–37.

Moons, E., Wets, G., Aerts, M., Arentze, T.A., Timmermans, H.J.P., 2005. The impact of simplification in a sequential rule-based model of activity-scheduling behavior. Environment and Planning A 37, 551–568.

Newell, A., Simon, H.A., 1972. Human Problem Solving. Prentice-Hall, Englewood Cliffs, NJ.

Quinlan, J.R., 1992. Learning with continuous classes. In: Adams, Sterling (Eds.), Proceedings AI '92. World Scientific, Singapore, pp. 343–348.

Quinlan, J.R., 1993. C4.5 Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo (CA).

Safavian, S.R., Landgrebe, D., 1991. A survey of decision tree classifier methodology. IEEE Transactions on Systems Man and Cybernetics 21, 660–674.

Swait, J.D., 2001. A non-compensatory choice model incorporating attribute cut-offs. Transportation Research Part B 35, 903–928.

Timmermans, H.J.P., Borgers, A.W.J., Veldhuisen, K.J., 1986. A hybrid compensatory–noncompensatory model of residential preference structures. Netherlands Journal of Housing and Environmental Research 1, 227–233.

Thill, J.C., Wheeler, A., 2000. Tree induction of spatial choice behavior. Transportation Research Record 1719, 250–258.

Tversky, A., 1972. Elimination by aspects: a theory of choice. Psychological Review 79, 281–299.

Van der Smagt, T., Lucardie, L., 1991. Decision making under not-well-defined conditions: from data processing to logical modelling. Tijdschrift voor Economische en Sociale Geografie 82, 295–304.

Wets, G., Vanhoof, K., Arentze, T.A., Timmermans, H.J.P., 2000. Identifying decision structures underlying activity patterns: an exploration of data mining algorithms. Transportation Research Record 1718, 1–9.

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufman, Amsterdam.

Xie, C., Lu, J.Y., Parkany, E., 2003. Work travel mode choice modeling with data mining: decision trees and neural networks. Transportation Research Record 1854, 50–61.

Yamamoto, T., Kitamura, R., Fujii, J., 2002. Driver's route choice behavior: analysis by data mining algorithms. Transportation Research Record 1807, 59–66.

Yuan, Y., Shaw, M.J., 1995. Induction of fuzzy decision trees. Fuzzy Sets and Systems 69, 125–139.