Machine Learning Meets Microeconomics: The Case of Decision Trees and Discrete Choice

Timothy Brathwaite[a,*], Akshay Vij[b], Joan L. Walker[c]

[a] *Department of Civil and Environmental Engineering, University of California at Berkeley*
116 McLaughlin Hall, University of California, Berkeley, CA, 94720-1720
[b] *Institute for Choice, University of South Australia*
Level 13, 140 Arthur Street, North Sydney, NSW 2060
[c] *Department of Civil and Environmental Engineering, University of California at Berkeley*
111 McLaughlin Hall, University of California, Berkeley, CA, 94720-1720

Abstract

In the 1960's, the logistic regression model from statistics and the binary probit model from psychology were linked with random utility theory, thereby connecting such methods with economic theory. Since then, the fields of statistics, computer science, and machine learning have created numerous methods for modeling discrete choices. However, these newer methods have not been derived from or linked with economic theories of human decision making. We believe this lack of economic interpretation is one reason discrete choice modelers have been slow to adopt these newer methods.

Our paper begins bridging this gap by providing a microeconomic framework for decision trees: a popular machine learning method. Specifically, we show how decision trees represent a non-compensatory decision protocol known as disjunctions-of-conjunctions and how this protocol generalizes many of the non-compensatory rules used in the discrete choice literature so far. Additionally, we show how existing decision tree variants address many economic concerns that choice modelers might have. Beyond theoretical interpretations, we contribute to the existing literature of two-stage, semi-compensatory modeling and to the existing decision tree literature. In particular, we formulate the first bayesian model tree, thereby allowing for uncertainty in the estimated non-compensatory rules as well as for context-dependent preference heterogeneity in one's second-stage choice model. Using an application of bicycle mode choice in the San Francisco Bay Area, we estimate our bayesian model tree, and we find that it is over 1,000 times more likely to be closer to the true data-generating process than a multinomial logit model (MNL). Qualitatively, our bayesian model tree automatically finds the effect of bicycle infrastructure investment to be moderated by travel distance, socio-demographics and topography, and our model identifies diminishing returns from bicycle lane investments. These qualitative differences lead the bayesian model trees to produce forecasts that directly align with the observed bicycle mode shares in regions with abundant bicycle infrastructure such as Davis, CA and the Netherlands. In comparison, the forecasts of the MNL model are overly optimistic.

Keywords: Decision Trees, Non-compensatory Decision Protocols, Discrete Choice, Two-stage Decision Making, Machine Learning, Semi-compensatory Models

1. Introduction

During the 1960s and 1970s, Daniel McFadden spearheaded the use of discrete choice techniques within economics, and in 2000, he was awarded a Nobel Prize for this work (University of California at Berkeley, 2000; Manski, 2001). By his own account (McFadden, 2001), McFadden's major contribution was *not* the creation of the conditional logit[1] model—a model that is still one of the most widely used discrete choice methods today. Indeed, the concept of a random utility maximization model was created earlier by Jacob

*Corresponding Author
Email addresses: timothyb0912@gmail.com (Timothy Brathwaite), Akshay.Vij@unisa.edu.au (Akshay Vij), joanwalker@berkeley.edu (Joan L. Walker)

[1] Note that the conditional logit model is also commonly referred to as the multinomial logit (MNL) model.

Marschak (1960), and statistical models that are nearly equivalent to McFadden's conditional logit model had already been introduced by David Cox (1966). According to McFadden,

> "The reason my formulation of the MNL model has received more attention than others that were developed independently during the same decade seems to be the direct connection that I provided to consumer theory [...]." (McFadden, 2001, p. 354).

Put simply, the great contribution of McFadden's work is that he connected an existing statistical model of discrete outcomes with economic theory (Manski, 2001).

In the more than fifty years since McFadden's pioneering efforts, the fields of machine learning and statistics have produced a vast array of methods that, like discrete choice models, predict the probability that a given discrete outcome will be realized out of a finite set of discrete alternatives. We now have decision trees, kernel machines, neural networks, and much more (Bishop, 2006; Friedman et al., 2008; Murphy, 2012). In general, these new techniques often display superior predictive ability compared to traditional discrete choice models (Fernández-Delgado et al., 2014; Wainer, 2016). However, despite this smorgasbord of accurate methods, discrete choice modelers have mostly restricted themselves to econometric techniques that are descended from McFadden's conditional logit model (Manski, 2001).

We hypothesize that one reason machine learning models have not made greater inroads amongst discrete choice modelers is because these models have not been linked to economic theories of human decision-making. Moshe Ben-Akiva (1973), one of the earliest discrete choice researchers, once wrote that "a model can duplicate the data perfectly, but may serve no useful purpose for prediction[2] if it represents erroneous behavioral assumptions." Though written in the 1970's, we believe that this sentiment still pervades the field of discrete choice modeling and econometrics more broadly (Einav and Levin, 2014; Bajari et al., 2015a,b). As a result, econometricians do not make frequent use of alternative techniques from machine learning and statistics. Such methods may be useful for prediction under stationary conditions, but they are considered black-boxes that lack a theoretical basis for interpreting and understanding human behavior.

In contrast to newer techniques from statistics and machine learning, almost all discrete choice models in the literature are rooted in the theory of utility maximization (Train, 2009), and even competing discrete choice models are based on alternative behavioral theories such as regret minimization (Chorus, 2012). Overall, theory-based econometric techniques appear to have become dominant within econometrics because behavioral theories provide a way to understand and interpret one's model outputs beyond in-sample and out-of-sample predictive accuracy. Machine learning methods have yet to provide this additional framework and linkage with economic theory.

In this paper, we aim to bridge this method-versus-theory gap by continuing to merge existing quantitative techniques with economic principles. Our contributions to the literature are as follows. First, we take a popular machine learning method—decision trees—and we connect it to economic theory. To do so, we provide a microeconomic framework for the interpretation of decision trees. In particular, we show that decision trees correspond to a non-compensatory, microeconomic decision protocol known as "disjunctions-of-conjunctions" (Hauser et al., 2010). Using this perspective, we explain how many of the varieties of decision trees address and can be motivated by microeconomic considerations such as analyst uncertainty or heterogeneity in one's non-compensatory behaviors. Additionally, our economic viewpoint suggests new additions to the existing body of decision tree techniques—additions that should lead to not only richer econometric models, but to more accurate statistical models overall.

Second, by combining decision trees with traditional discrete choice models, we advance the state of the art in the modeling of semi-compensatory decision making. We discuss how decision trees allow us to more flexibly represent non-compensatory behaviors than previously possible. Moreover, we show that our two-stage, semi-compensatory model jointly models how non-compensatory decision protocols influence both choice set formation and preference heterogeneity[3].

[2]Note that the sort of prediction being referred to is prediction in the face of a policy change. This type of prediction is characteristic of causal inference whereby one predicts the effects of external manipulation of environmental conditions.

[3]We are aware that, in the discrete choice literature, the term preference heterogeneity has been used ambiguously. In some cases, preference heterogeneity refers to differences in the general preference for an alternative, irrespective of attributes of the alternative (Bhat, 1998). In other cases, preference heterogeneity is taken to also include the coefficients that are multiplied by an alternative's attributes when using a linear-in-parameters choice model specification (Kamakura et al., 1996). In still other

Finally, our third contribution is an empirical demonstration of the aforementioned techniques to the choice of travel mode in the San Francisco Bay Area. We show that the semi-compensatory models fit the data better than traditional models based solely on utility-maximization, and we show that the semi-compensatory models lead to a number of policy implications that are not readily uncovered by traditional discrete choice models. Through this application, we illustrate the quantitative and qualitative benefits that can come from combining economic theory with machine learning and modern statistical methods.

Structurally, the rest of our paper is organized as follows. In Section 2, we provide an econometrically accessible introduction to decision trees. Here, we focus on decision trees as a statistical tool. Next, Section 3 describes the microeconomic theories of non-compensatory decision making that are related to decision trees, and it shows how decision trees algorithmically represent these concepts. Here, we focus on the ways that decision trees are motivated by particular decision making principles. In Section 4, we review how the aforementioned microeconomic concepts have been operationalized in the discrete choice literature so far, and we make note of how decision trees address the theoretical and practical difficulties with these previous implementations. Section 5 then details the various types of decision trees, including combined decision-tree/discrete-choice models. Specifically, we orient our discussion around the ways these decision tree variants address economic considerations that might prevent choice modelers from using decision trees in their work. In Section 6, we formulate a new decision-tree/discrete-choice model, and we apply the model to the choice of travel mode in the San Francisco Bay Area. We describe the data used for this application, and we discuss the greater fit and unique insights provided by our semi-compensatory model in comparison to models based purely on utility-maximization. Finally, Section 7 concludes.

2. Decision trees explained

In this section, we provide a brief description of decision trees, targeting econometricians as our main audience. We will first provide an explanation of what a decision tree is, and we will use a highly simplified example to demonstrate how they can be used. After this, we will give a brief description of some of the (many) ways that decision trees are estimated from data. Here, again, we will focus on comparing and contrasting these estimation methods with techniques that econometricians are familiar with.

2.1. What are decision trees and how do we use them?

In simple terms, decision trees are a set of "if-then" statements that are used to predict a given quantity[4] (Loh, 2011). Etymologically, decision trees get their name because they are often represented graphically as a tree: an acyclic set of nodes connected by directed edges, with each node connected to at most one preceding node, beginning with a single "root" node that has no edges pointing into it, and terminating with a set of "output" nodes (Meila and Jordan, 2000; Rokach and Maimon, 2005). Each path from the root node to an output node represents one of the "if-then" statements that make up the tree. These if-then statements must partition the space of explanatory variables into a set of mutually exclusive regions (corresponding to the output nodes) that span the entire space of explanatory variables (Lemon et al., 2003). Then, when making predictions about a decision maker, the "if" condition is used to determine the region/output-node the decision maker is in, and the corresponding "then" statement is used to provide the desired prediction. In a discrete choice context, such predicted quantities might be (1) the probability that a particular alternative is considered or (2) the probability with which an alternative is chosen.

To continue our explanation of what decision trees are and how they can be used, we will now provide a concrete, but highly stylized example of choice set generation, conditional on a given decision tree. In the discussion that follows, we realize that modelers may have many valid reservations about the realism of our example. It suffices to say that concerns about the deterministic nature the choice sets generated by our tree

cases, preference heterogeneity is taken to also include choice set heterogeneity (Vij and Walker, 2014). In this paper, we use preference heterogeneity to include all of the coefficients in one's linear-in-parameters choice model specification. If one is using a non-linear or non-parametric choice model specification, we are also using preference heterogeneity to include differences in the systematic utility functions for different individuals.

[4]We realize that our definition of decision trees is broad. Our definition includes models such as regression trees, classification trees, decision lists, and decision tables (Rivest, 1987; Loh, 2011). For this paper's purposes, these models are similar enough to merit a joint description.

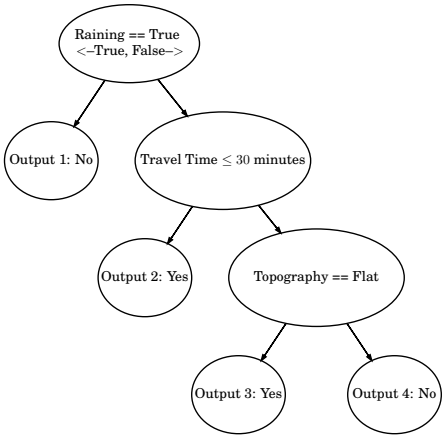

Figure 1: Example decision tree for bicycle consideration

(shown in Figure 1), concerns about the explicit discontinuities in the tree, and concerns about how such a tree could be estimated can all be addressed. Our example only features these qualities for simplicity of discussion. We note that in some contexts, deterministic choice sets are not uncommon: for example, when individuals are making residential location choices, some housing options may be deterministically excluded because the rents violate the individuals' income constraints (Kaplan et al., 2012; Zolfaghari et al., 2013; Bhat, 2015). Moreover, decision trees that probabilistically predict an individual's choice set can be estimated. These considerations will be discussed in Section 5. Concerns about the explicit discontinuities in our tree can be relaxed by considering individual heterogeneity in the split points of a tree or in the very structure of the tree being used. Like the issue of estimating trees that probabilistically predict an individual's choice set, concerns about individual heterogeneity are discussed in Section 5. Lastly, the estimation of decision trees will be discussed in Subsection 2.2.

Now, disclaimers aside, imagine that we are modeling the choice set formation behavior of travellers who are choosing the mode by which they will travel. Further, assume that our population of individuals has only two commuting alternatives: bicycle and public transit, and assume that public transit is always considered. Finally, Figure 1 shows the decision tree that represents the assumed choice set formation process in our hypothetical population. Here, *Raining* is either *True* or *False*, *Travel Time* is measured in minutes, *Topography* is either *Flat* or *Hilly*, and the dependent variable (bicycle consideration) is either *Yes* or *No*. From the tree in Figure 1, a number of useful observations can be made. First, there are four output nodes, two of which result in bicycle being considered and two that result in bicycle not being considered. Secondly, we see that bicycle consideration is a function of weather (raining or not), travel time, and topography. Now, to use the tree to make predictions for a given individual, one must traverse the tree from top to bottom, ending at one of the tree's output nodes. The rules for traversing the given[5] decision tree are that if the condition in a decision node (i.e. a non-output node) is True, then one goes to the left and if the condition is False, one goes to the right.

So, what can one use the tree in Figure 1 for? First, the tree and its predictions can be directly used to inform policies. For instance, a municipality trying to increase bicycle usage must first ensure that bicycle is considered as a mode of travel. Based on this example's tree, the municipality might subsidize the relocation costs for individuals that wish to move to a location that is 30 minutes away or closer to their workplace. Such subsidies would help push bicycle into the choice sets of individuals, thereby increasing the expected number of bicycle commuters. Secondly, the tree in Figure 1 might be used as part of a larger model building effort. For instance, one might use the tree in Figure 1 to inform a two-stage model of travel mode choice. At

[5]We note in passing that the traversal rules may change from tree to tree, based on author preference, but they should always be explicitly stated.

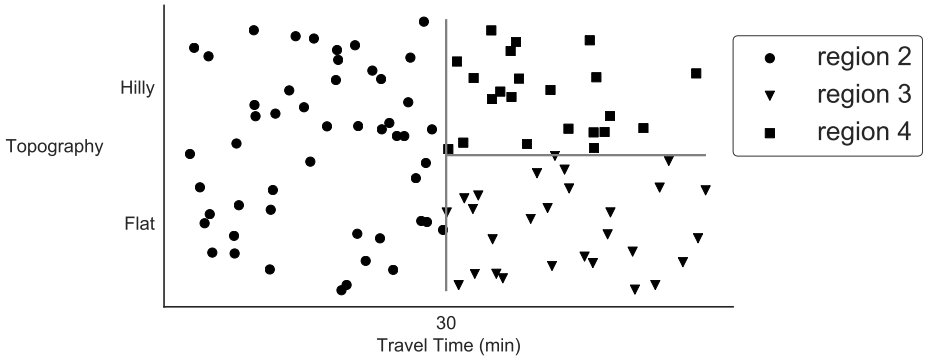

Figure 2: Regions 2-4 of example decision tree for bicycle consideration

the first stage, an individual's choice set is modeled. By assuming that individuals must travel to work and that public transit is always considered, our example is left with two possible choice sets: {Public Transit} and {Public Transit, Bicycle}. The choice sets in this example are based on whether bicycle is considered or not, and the probabilities of these choice sets (i.e. the first stage in Manski's two-stage models) can be written as follows:

$$\begin{aligned} P\left(C = \{\text{Public Transit, Bicycle}\} \mid x, \text{tree}\right) &= P\left(\text{Bicycle considered} \mid x, \text{tree}\right) \\ &= \sum_r P\left[\text{Bicycle considered} \mid T\left(x\right) = r\right] P\left[T\left(x\right) = r\right] \\ \text{where } C &= \text{An individual's choice set.} \\ r &\in \{1, 2, 3, 4\} \\ r &= \text{A specific region demarcated by the decision tree.} \\ T\left(x\right) &= \text{The region an individual belongs in based on } x \text{ and the tree.} \\ x &= \text{Explanatory variables for an individual.} \end{aligned} \tag{1}$$

For most decision trees, $T\left(x\right)$ is a deterministic function[6] such that, given explanatory variables x, an observation is deterministically assigned to a given region/output-node r. For our example, regions 2-4 are graphically depicted in Figure 2. Because our $T\left(x\right)$ is deterministic, $P\left[T\left(x\right) = r\right]$ is either 1 or 0, and the same is true of the probability of bicycle consideration, conditional on being in a given region. In all cases, we can expand $P\left[T\left(x\right) = r\right]$ to more explicitly show how each explanatory variable contributes to the likelihood of an observation being in a given region.

Specifically, we note that each "if" statement in the decision tree can be written as the union of elementary conditions, typically[7] with one such elementary condition per explanatory variable. For instance, let x_1 denote whether it is raining, let x_2 denote the bicycle travel time between an individual's home and work, and let x_3 denote the topography between an individual's home and work. Additionally, let S_{rk} denote the set that variable x_k must be in for an individual to belong to region r. Using these variables, we can write the region corresponding to the first output node as $S_{11} = \{\text{True}\}$, $S_{12} = [0, \infty)$, and $S_{13} = \{\text{Flat, Hilly}\}$. These sets reflect the fact that output node 1 is the region of the variable space where *Raining* is True and where any values of *Travel Time* or *Topography* are valid. With this notation, we can express the probability of bicycle consideration as follows:

[6]The primary exception to this is a "probabilistic" decision tree, also known as a "soft" or "fuzzy" decision tree, where $T\left(x\right)$ is a probabilistic function. These decision tree variants will be discussed in Section 5. The other exception is where the case of measurement error where the value x is unknown and modeled with a probability distribution of its own.

[7]Exceptions to this statement come from decision trees that are not "axis-aligned," such as oblique decision trees that use inequalities with linear combinations of variables for their "if" conditions (Murthy et al., 1994; Ittner and Schlosser, 1996).

$$
\begin{aligned}
P\left(\text{Bicycle considered} \mid x, \text{tree}\right) &= \sum_r P\left[\text{Bicycle considered} \mid T\left(x\right) = r\right] P\left[T\left(x\right) = r\right] \\
&= \sum_r P\left[\text{Bicycle considered} \mid T\left(x\right) = r\right] P\left[\bigcap_k x_k \in S_{rk}\right] \\
&= \sum_r \left\{ P\left[\text{Bicycle considered} \mid T\left(x\right) = r\right] \prod_k P\left[x_k \in S_{rk}\right] \right\}
\end{aligned} \tag{2}
$$

The equation above shows how, conditional on a given decision tree, one can form the sorts of probability statements that are common in the first stage of two-stage choice models with non-compensatory rules for choice set formation (Gilbride and Allenby, 2004; Cantillo et al., 2006). Moreover, if one's decision tree was being used to directly predict the probability of a given alternative, one's likelihood function would be formed analogously. Besides being transparent about how the structure of the tree translates to one's likelihood equations, Equation 2 highlights the link to the non-compensatory decision protocol known as disjunctions-of-conjunctions (Hauser et al., 2010). Though we will delay a detailed discussion of this protocol to Section 3, we point out here that logical disjunctions are algebraically represented as summations and logical conjunctions are algebraically represented as products (Gilbride and Allenby, 2004). Equation 2 shows that when modeling bicycle consideration with a decision tree, our probabilities of interest are explicitly given as a summations of products (i.e as disjunctions-of-conjunctions). Importantly, such a decision protocol generalizes the typical conjunctive or disjunctive rules that are used in choice models that represent non-compensatory processes. See Section 3 for further discussion and explanation of this point.

2.2. How do we estimate decision trees?

In the previous subsection, we explained what decision trees are and (conditional on a specific decision tree) what one can do with them. In this subsection, we turn to the question of how such decision trees are estimated from data and how such estimation techniques differ from those commonly employed in the discrete choice literature.

To begin, discrete choice modelers are most likely to be familiar with estimation techniques such as maximum likelihood, method of moments, and bayesian Markov Chain Monte Carlo (MCMC) methods (Train, 2009). Of these techniques, only bayesian MCMC methods have been applied to the estimation of decision trees (Chipman et al., 1998; Denison et al., 1998; Letham et al., 2015; Pratola, 2016). We believe that the main reason for this discrepancy in estimation methods is that decision trees are not continuous functions. Instead, they are explicitly discontinuous functions of the explanatory variables (e.g. at a particular node, should we split on *Travel Time* or *Travel Cost*?). Maximum likelihood, if it is to be performed at all can no longer rely on gradients and hessians, so enumeration and comparison of all decision trees is necessary. However, enumeration of all possible decision trees is NP-hard (Ruggieri, 2017). Since it is computationally prohibitive to enumerate all possible decision trees and assess their log-likelihoods, maximum likelihood estimation of decision trees is typically viewed as infeasible. Similarly, since the method of moments and its generalizations require continuous moment functions (Hansen, 1982), these estimation techniques cannot be used to estimate decision trees.

As highlighted in the last paragraph, estimation of decision trees is severely hindered by the discontinuous nature of the trees and the fact that explicit enumeration of all possible trees is computationally prohibitive. Due to these challenges, most estimation techniques (both bayesian and frequentist) use approximations and heuristics. By far, the most common frequentist heuristic is to use a greedy algorithm to estimate the tree (Rokach and Maimon, 2005). Here, one recursively performs a search over all variables and values of those variables to pick the variable and value combination that best meets some "splitting criteria." After finding the best variable and value pair, the dataset is split according to the chosen pair. The process is then repeated for each subset of the data: those meeting the chosen condition and those not meeting the condition. The greedy estimation of the decision tree will terminate once some stopping criteria is met (e.g. no output node should contain less than 5 observations). After the initial estimation of the decision tree, some estimation methods "prune" the initial tree by removing nodes according to a "pruning criterion" (Mingers, 1989; Esposito et al., 1997). Differing methods and criteria for splitting, stopping, and pruning all lead to different types of decision trees (Loh, 2014; Rokach and Maimon, 2014). Moreover, besides the

greedy approach just described, there exist a number of other frequentist tree estimation techniques such as using genetic algorithms (Barros et al., 2012) or branch-and-bound algorithms (Angelino et al., 2017). Though we cannot perform an exhaustive review of the various decision tree estimation techniques, good surveys of this material can be found in Murthy (1998); Rokach and Maimon (2005); Barros et al. (2012), and Lomax and Vadera (2013).

For the bayesian estimation of decision trees, a prior is placed over the space of possible decision trees, the likelihood is formed using equations similar to Equation 2, and then an MCMC algorithm is used to sample from the posterior distribution of possible decision trees (Chipman et al., 1998; Denison et al., 1998; Letham et al., 2015; Pratola, 2016). At first glance, this seems exactly the same as what is always done in a bayesian estimation. However, since the set of all possible decision trees is huge and discrete, the MCMC algorithms do not typically "explore" the entire posterior distribution of trees (Chipman et al., 1998). The approximation is that the MCMC methods typically only explore part of the posterior since these algorithms are limited by however much time an analyst has to let the algorithm run. If the MCMC algorithm is run for long enough, the hope is that "high accuracy" sections of the posterior are explored, such that one samples from the trees that are most predictive of the choices in one's dataset. Note, unlike the frequentist estimation methods where trees are defined based on how they are estimated, differing priors or differing MCMC methods lead to differences in how the space of decision trees is explored, but it is uncommon to speak of "different" bayesian decision trees. Such differentiation is likely unnecessary because, given an impractically long time, all bayesian MCMC techniques will explore the entire posterior of trees.

Finally, we pause to make a few passing remarks about the properties of the various estimators for decision trees. In standard discrete choice modeling, much importance is placed on having consistent and efficient estimators. The greedy estimation techniques described above for decision trees have long been proven to be consistent, non-parametric estimators of underlying data-generating processes (Gordon and Olshen, 1980, 1984; Toth and Eltinge, 2011). Bayesian techniques have also demonstrated their consistency in simulation (Letham et al., 2015), though formal proofs are still missing. In terms of efficiency, however, it is not clear that this notion is meaningful for decision tree models. In particular, the notion of an "efficient" estimator being one that achieves the Cramer-Rao lower bound is no longer meaningful since the parameter space (the number of splits in the tree, variables and values being split on, and the tree structure) is discrete and increases with the size of one's data set (i.e. it is not fixed). If one views efficiency as being inversely related to the variance of one's estimator, then it is known that estimation techniques that generate a large number of candidate trees and then select the best one tend to be less variable than the greedy methods described above (Tibshirani and Knight, 1999). Nevertheless, whether or not other variations on the notion of efficiency can be shown to apply to decision trees is beyond the scope of this paper and will not be investigated.

3. Decision trees: The link with microeconomics

In Section 2, we described what decision trees are, how a given decision tree can be used, and how decision trees might be estimated. Additionally, in both Sections 1 and 2, we noted that decision trees correspond to a non-compensatory decision protocol known as disjunctions-of-conjunctions (Hauser et al., 2010). In this section we will review this microeconomic interpretation of decision trees in detail. Initially, we will briefly describe standard discrete choice models and their use of compensatory decision protocols. Then we will motivate the need for non-compensatory decision protocols, and in Subsection 3.1, we will proceed to describe a number of such behavioral strategies. We will begin with simple non-compensatory protocols and proceed to describe further generalizations of such strategies until we arrive at disjunctions-of-conjunctions: a focal point of our paper. Finally, in Subsection 3.2, we will mathematically show how decision trees represent disjunctions-of-conjunctions.

To start, we note that compensatory decision protocols are decision making strategies where, for a given alternative, "high levels of satisfaction with one attribute compensate for low levels of satisfaction with [other]" attributes (Foerster, 1979). As readers are likely aware, almost all discrete choice models used in practice and research are based on compensatory decision processes, with utility-maximization being the most

common example[8] (Swait, 2001b; Truong et al., 2015). However, counter to prevailing practices, behavioral economists and psychologists have presented much evidence that individuals frequently depart from standard notions of utility maximization and rationality (Foerster, 1979; Bronner, 1982; Tversky and Kahneman, 1986; Conlisk, 1996). Spurred by these observations, a steady but small stream of research has both called for and proposed new models of human decision making that explicitly incorporates the possibility of non-utility maximizing choice behavior (Simon, 1955; Tversky, 1972; Gigerenzer and Goldstein, 1996; Leong and Hensher, 2012). Such alternative methods of decision making are typically referred to as non-compensatory decision rules or non-compensatory decision protocols. They are called non-compensatory because they do not always allow positive attributes of a given alternative to compensate for negative attributes of that same alternative. Additionally, since non-compensatory decision rules do not typically require the evaluation of all attributes of all alternatives, they better capture the limited cognitive resources of decision makers (Simon, 1955; Young, 1984; Swait, 2001b) and are therefore thought to be more behaviorally realistic.

3.1. Non-compensatory decision rules

Thus far, some of the non-compensatory decision processes that have been detailed in the discrete choice literature include: dominance (Cascetta and Papola, 2009), lexicography (Kohli and Jedidi, 2007), elimination-by-aspects (Tversky, 1972), satisficing (Stüttgen et al., 2012), conjunctive rules, disjunctive rules, subset-conjunctive rules, and disjunctions-of-conjunctions. Of these, conjunctive and disjunctive rules are quite prevalent in the literature, and all of the last four non-compensatory rules are related to decision trees. We therefore describe the last four non-compensatory decision protocols below, and in Section 4, we review how these four protocols have been previously incorporated into discrete choice models.

Conjunctive Rules (Coombs, 1951; Dawes, 1964)

Using a conjunctive decision rule, an individual only considers alternatives that meet all of a given number of requirements. For instance, an individual making a residential location choice may only consider housing options that meet his or her requirements on the maximum amount of rent **and** the distance from the individual's workplace location. The "and" statement is what distinguishes this decision rule as conjunctive. As noted in Subsection 2.1, conjunctive statements are algebraically represented using products.

Disjunctive Rules (Coombs, 1951; Dawes, 1964)

Using a disjunctive decision rule, individuals only consider alternatives that meet at least one of a given set of requirements. For instance, continuing with the residential choice example, an individual may only consider housing options that are within a given distance from their workplace location **or** that are within a given distance from major public parks. The "or" statement is what distinguishes this decision rule as disjunctive. As noted in Subsection 2.1, disjunctive statements are algebraically represented using sums.

Subset-Conjunctive Rules (Jedidi and Kohli, 2005)

Subset-conjunctive rules are a generalization of both conjunctive rules and disjunctive rules. Using a subset-conjunctive decision rule, an individual only considers alternatives that meet a certain number of requirements. Using another residential location choice example, consider an individual who would like to live within one mile of a major public park, who would like to live within two miles of his or her workplace, who would like to pay less than $1,000 per month in rent (but is flexible), and who would like to live within one mile of a subway station. Under a subset-conjunctive rule, this individual would consider any housing units that meet some number of these four requirements. For instance, this individual might consider any housing units that meet at least three of these four requirements. Note that if this individual only considered housing units that met all four requirements, then this would be equivalent to a conjunctive decision rule with four requirements. Likewise, if this individual only required housing units to meet one of the four requirements, then this would be equivalent to a

[8] We are aware of the increasing number of discrete choice models that are being estimated under the assumption of regret-minimizing behavior. However, such models are still compensatory in nature, and therefore retain many of the properties we describe in the context of utility-maximization.

disjunctive decision rule. Algebraically, subset-conjunctive rules are therefore sums of products, with the restriction that each product term have a given number elements (one for each requirement that should be met).

Disjunctions-of-Conjunctions (Hauser et al., 2010)

Disjunctions-of-conjunctions generalize the conjunctive, disjunctive, and subset-conjunctive decision rules. Under a disjunctions-of-conjunctions decision protocol, an individual will consider any alternative that meets at least one of a given set of conjunctive conditions. Each condition may differ in the number of requirements that compose the conjunction. Algebraically, then, disjunctions-of-conjunctions are expressed as sums of products with no constraints on the number of elements in each product.

Consider once more the residential choice example. If, for instance, our decision maker was more concerned about rent than the other requirements, he or she might consider any housing unit that required less than $1,000 per month in rent and that met one of the remaining three requirements. Additionally, he or she might consider any housing unit that was simultaneously within one mile of a major park, within one mile of a subway station, and within two miles of his or her workplace. In this case, only one of the following four conjunctive conditions needs to be met in order for a housing unit to be considered:

- rent less than $1,000 per month and housing unit within one mile of a major public park
- rent less than $1,000 per month and housing unit within two miles of the individual's workplace
- rent less than $1,000 per month and housing unit within one mile of a subway station
- housing unit within one mile of a major public park and within one mile of a subway station and within two miles of the individual's workplace.

As can be seen from the example above, if the individual had only one condition for consideration, we would have a conjunctive rule. If the individual had only one requirement in each of the four conditions above, then we would have a disjunctive rule. Similarly, if we expanded the first three conditions above so that they each included a third requirement, we would once again have the subset-conjunctive rule whereby any housing unit with three of the four requirements would be considered.

Before moving on to Subsection 3.2, we pause to briefly summarize why we believe the link between disjunctions-of-conjunctions and decision trees is important. First, as noted above, conjunctive rules and disjunctive rules are seen as important information processing strategies, and they have been applied in many choice modeling efforts (Foerster, 1979; Swait, 2001b; Gilbride and Allenby, 2004; Elrod et al., 2004; Martínez et al., 2009; Hauser et al., 2010; Hess et al., 2012; Kaplan et al., 2012; Zolfaghari et al., 2013; Truong et al., 2015). Being a generalization of these two rules, disjunctions-of-conjunctions may also be an important decision making strategy, but it has seldom been tested in choice modeling contexts. We think a major reason for this lack of choice modeling application is because there have not been easy or straightforward ways to estimate such rules. Linking disjunctions-of-conjunctions to decision trees gives researchers a way to estimate disjunctions-of-conjunctions by drawing upon well established methods of estimating decision trees. Additionally, once disjunctions-of-conjunctions can be estimated by themselves, it is then possible to estimate such strategies in combination with the compensatory procedures used in standard discrete choice models. We pursue this strategy later, in Section 5 and Section 6.

3.2. Linking decision trees with disjunctions-of-conjunctions

As described in the previous subsection, disjunctions-of-conjunctions are highly flexible non-compensatory decision protocols. Here, we highlight how decision trees mathematically represent the relationships implied by disjunctions-of-conjunctions.

First, we define the necessary notation. Let b represent a primitive boolean statement, i.e. a specific requirement. Such a statement is an equality or inequality that is not composed of any other equalities or inequalities. For instance, $x == 2$ and $x \leq 5$ are primitive boolean statements but $((x_1 == 2) * (x_2 \leq 5))$ is not a primitive boolean statement because it is composed of two boolean statements. Additionally, if b is True, then we say that $b = 1$, and if b is False, then we say that $b = 0$.

With this notation, conjunctive rules can be expressed as:

$$\begin{aligned} &\text{if } \left(\prod_{i=1}^{B} b_i\right) == 1 \Rightarrow y \\ &\quad \text{where } B = \text{the total number of requirements in the rule.} \\ &\quad\quad \text{“}\Rightarrow\text{”} = \text{“then” or “implies”.} \\ &\quad\quad\quad y = \text{some outcome.} \end{aligned} \tag{3}$$

In words, this is read as “if all requirements, b_i, are met, then y”. This follows because each b_i must be True (i.e. must be met) in order for that b_i to equal 1, and we need all b_i to equal 1 in order for $\prod_{i=1} b_i$ to evaluate to 1.

Similarly, a disjunctive rule can be expressed as:

$$\text{if } \left(\sum_{i=1}^{B} b_i\right) \geq 1 \Rightarrow y \tag{4}$$

In words, this is read as “if at least one (i.e. if any) of the requirements b_i are met, then y”. This follows because any requirement b_i that is not met will cause that b_i to evaluate to 0. If at least one requirement is met, then the corresponding b_i's will evaluate to 1, and then $\sum_{i=1} b_i$ will be greater than or equal to 1.

With these building blocks, we turn immediately to the case of disjunctions-of-conjunctions[9]. In words, the use of disjunctions-of-conjunctions requires statements such as “if at least one of some set of conjunctive conditions is met, then y.” To mathematically express such a statement, we will introduce additional symbols. The first symbol, p, will represent conjunctive conditions, i.e. products of primitive boolean statements. As noted in Subsection 3.1, in disjunctions-of-conjunctions, the various conjunctive conditions need not have the same number of requirements. To account for this, we will index the various conjunctive conditions by i, and we will use $\mid p_i \mid$ to denote the number of requirements that make up p_i. Finally, we will use the symbol, b_j^i, to indicate the j'th primitive boolean statement (i.e. the j'th requirement) in conjunctive statement p_i. With this additional notation, our disjunctions-of-conjunctions statement can now be expressed as:

$$\begin{aligned} &\text{if } \left(\sum_{i=1}^{D} p_i\right) \geq 1 \Rightarrow y \\ &\text{if } \left(\sum_{i=1}^{D} \prod_{j=1}^{|p_i|} b_j^i\right) \geq 1 \Rightarrow y \\ &\quad \text{where } D = \text{the total number of conjunctive conditions.} \end{aligned} \tag{5}$$

From the first line, we mathematically see the disjunction (i.e. the summation) of conjunctive conditions. The second line shows the conjunction (i.e. the product) of requirements. Now, for subset-conjunctive rules, we merely impost the constraint that $\mid p_i \mid$ be equal to some constant value for all p_i. This is equivalent to saying that each conjunctive condition must be comprised of the same number of requirements.

To go from the abstract equations above to a decision tree, we must consider what a conjunctive condition represents. In general, a conjunctive condition defines a region in a space. Using Figure 1 as an example once more, consider the space formed by the variables x_1 = Rain and x_2 = Travel Time. The conjunctive condition that leads to output node 2 is x_1 == False AND $x_2 \leq 30$. This condition will define a rectangular region in the graph of (x_1, x_2) comprised of the area where x_1 is False, and the area where x_2 is less than 30. For more examples of regions formed by conjunctive conditions, see Figure 2 above. Now, when we have multiple conjunctive conditions, we have multiple regions in space. These regions will either be mutually exclusive, or they will overlap. It is crucial to note that any region defined by a set of overlapping conjunctive criteria can be expressed as a region defined by a set of mutually exclusive criteria. For instance, let $\{p\} = \{p_1, p_2\}$ be a region defined by a set of overlapping conjunctive conditions, p_1 and p_2. This region can be re-expressed

[9]Subset-conjunctive rules will be expressed as a special case of the formula for disjunctions-of-conjunctions.

as a set of mutually exclusive conjunctive conditions, $\{\tilde{p}\} = \{\tilde{p}_1, \tilde{p}_2\}$. One such re-expression is $\tilde{p}_1 = p_1$ and $\tilde{p}_2 = p_2 * p_1'$, where $\tilde{p}_2$ is read as "$\tilde{p}_2$ equals p_2 AND NOT p_1." Observations meeting the condition $\tilde{p}_2$ will therefore satisfy all the requirements of p_2, but they will not satisfy all the requirements of p_1.

With the possibility of re-expression in mind, recall that using disjunctions-of-conjunctions means making statements of the form "if at least one of some set of conjunctive conditions, $\{p\}$, is met, then y". As just noted, this statement can be reformulated as, "if at least one of some set of conjunctive conditions, $\{\tilde{p}\}$, is met, then y". Given the mutually exclusive conjunctive conditions of $\{\tilde{p}\}$, our reformulation can be expressed as a decision tree where each conjunctive condition in $\tilde{p}$ becomes an "if" statement in the tree with a corresponding "then y" statement. Note we will also need a final condition such as "if $\bigcap_{i=1}^{D} \tilde{p}_i'$ then y'," where $y' \neq y$. Here, the final condition ensures that the decision tree is comprised of a set of conditions that are both mutually exclusive and exhaustive. The condition $\bigcap_{i=1}^{D} \tilde{p}_i'$ is read as "NOT $\tilde{p}_1$ and NOT $\tilde{p}_2$ and ... and NOT $\tilde{p}_D$." Finally, we use y' as the outcome for the remaining conditions that are added to ensure exhaustiveness, e.g. $\bigcap_{i=1}^{D} \tilde{p}_i'$, simply because we assume that if there was any other condition that would result in y, then that condition would have been part of the original set of conditions, $\{p\}$.

4. A review of how non-compensatory protocols have been incorporated in discrete choice

In Section 3, we described conjunctive rules, disjunctive rules, subset-conjunctive rules, and disjunctions-of-conjunctions. However, researchers have gone beyond mere descriptions. These decision protocols have been incorporated into choice models and used to quantitatively study the concordance of non-compensatory processes with observed choices. In this section, we will review the ways that conjunctive rules, disjunctive rules, and their generalizations have been previously incorporated into discrete choice models. Afterwards, we will highlight drawbacks of the previous work that our paper seeks to address. Having said this, we state upfront that our review mainly focuses on the way that non-compensatory protocols have been used to model choice set generation as opposed to modeling the actual choice being made. The reason for our focus is that conjunctive rules, disjunctive rules, and their generalizations are (in general) not sufficient to uniquely choose a particular alternative. Multiple alternatives may meet an individual's non-compensatory rules, but (in our context) a decision strategy must still be employed to generate a single discrete choice. As a result, conjunctive rules, disjunctive rules, and their generalizations have almost exclusively been used in the discrete choice literature to winnow a decision maker's choice set before another strategy is used (if necessary) to make the final choice. In Subsection 4.1, we review this approach of choice set generation followed by compensatory choice amongst the considered alternatives, and we revisit this notion in Section 5.1.4 when we describe the decision tree variant known as "model trees." In Subsection 4.2, we will briefly review the few ways that observed choices have been directly[10] modeled with conjunctive rules, disjunctive rules, and their generalizations.

4.1. Choice-set generation via non-compensatory protocols

Across the literature, two main approaches have been used to incorporate conjunctive, disjunctive, and related protocols into discrete choice models. These two approaches differ primarily based on whether they explicitly model an individual's decision making using two-stages as prescribed by Manski (1977) or whether they use a single-stage model that implicitly performs choice-set generation. We will begin by first describing the single-stage models, also known as the "reduced-form" approach (Swait, 2001b).

Pioneered by Swait (2001b), single-stage models implement conjunctive and/or disjunctive rules by altering the systematic utility of an alternative. When representing strict non-compensatory behaviors, these models combine attribute values and attribute thresholds to set the systematic utility of an alternative to -/+ infinity, effectively removing an alternative from one's choice set or removing all other alternatives from one's choice set. Through the years, multiple single-stage models have been proposed, each with their own set of unique additions. Swait (2001b) allowed for non-strict non-compensatory behavior where violation of an attribute threshold was allowed but resulted in penalties to one's systematic utility. Elrod et al. (2004) estimated the attribute thresholds from choice data only, whereas Swait (2001b) required individuals to report their attribute thresholds. Moreover, Elrod et al. (2004) did not allow violation of one's attribute

[10] I.e., without estimating any rules or parameters that implicitly or explicitly determine one's choice set.

threshold and even penalized or rewarded the systematic utility when the value of an attribute approached that attribute's threshold, based on whether a conjunctive or disjunctive rule was being implemented. When allowing violation of one's attribute thresholds, Martínez et al. (2009) used non-linear penalty functions in contrast to the linear penalty functions of Swait (2001b). Most recently, Truong et al. (2015) proposed a novel way to estimate the attribute thresholds in the context of Swait's original (2001b) formulation. Common to all these implementations, however, is the fact that conjunctive or disjunctive behavior was operationalized through the systematic utility function.

The second approach used in the literature to represent conjunctive, disjunctive, and similar behaviors is the two-stage approach where one formally models the choice set generation process. To date, the vast majority of such two-stage models have relied on the Probabilistic Independent Availability Logit (PIAL) model (Swait, 1984, 2009). Here, the two-stage models use non-compensatory decision rules to determine whether each alternative will be present in an individual's choice set. The randomness underlying the probability that an alternative is in one's choice set is explained as coming from analyst uncertainty over the attribute thresholds used by each individual to evaluate the non-compensatory rules. Moreover, the probability of an alternative being in one's choice set is considered to be independent of the probability that any other alternative is in one's choice set, hence the name PIAL. Despite this independence assumption, PIAL models still suffer from the curse of dimensionality since they typically require one to enumerate all possible subsets of one's universal choice set. As a result, important differences can be seen in the way that various authors have dealt with this computational hardship. Some authors have used simulation techniques to avoid full enumeration of the various consideration sets, other authors have made no attempts at avoiding computational difficulties in estimating PIAL models, and still other authors have tried to minimize the number of possible consideration sets by collecting explicit consideration set information from decision makers. Our review below will be structured around these modeling differences.

To the best of our knowledge, the first paper to incorporate conjunctive and disjunctive rules into a two-stage model was the 2004 paper of Gilbride and Allenby. As described above, these authors parametrize the probability of an alternative being available as the probability of an alternative satisfying the conjunctive or disjunctive rules that are made up by the (unobserved) attribute thresholds for each attribute. To sidestep the computationally prohibitive step of enumerating each possible consideration set, Gilbride and Allenby use a bayesian estimation method. In particular, the authors use a MCMC sampling method to explore the space of possible thresholds, and each set of sampled thresholds induces a particular choice set that can be used in the second-stage choice process. While apparently successful in dealing with the curse of dimensionality, most models after Gilbride and Allenby (2004) take a different (i.e. a frequentist) approach.

For an example of this frequentist approach, we can look at the second paper on this topic, by Cantillo and de Dios Ortúzar (2005). These authors estimate a frequentist version of the Gilbride and Allenby model, using standard maximum likelihood estimation as opposed to a simulation-based optimization method. As a result, these authors are forced to enumerate all possible consideration sets, thereby incurring all estimation difficulties from the curse of dimensionality. On a positive note, however, Cantillo and Ortúzar are able to parameterize the attribute thresholds as a function of socioeconomic variables and choice conditions (e.g. trip purpose, time restrictions, etc.). This allows them to give greater behavioral interpretation to the estimated thresholds. Shortly thereafter, Jedidi and Kohli (2005) use a PIAL model where they allow for subset-conjunctive rules and for individual heterogeneity through the use of latent classes. To accommodate uncertainty in the number of requirements that need to be satisfied, Jedidi and Kohli estimate this parameter as well. Their approach amounts to full enumeration of all possible choice sets under each possible set of criteria and each possible number of requirements. Later, Swait (2009) returns to the issue of choice set generation with a two-stage choice model called a k-Mix model. This model is a PIAL model at its core, albeit with a couple of important differences. First, favorable conjunctive or disjunctive rules can be used to not only allow for consideration of alternatives but to place them in a "dominance" state wherein alternatives are preferred to all other alternatives that are not in a dominant state. Secondly, unfavorable non-compensatory rules can be used to place alternatives in a "rejection" state where alternatives are completely disregarded unless all other alternatives are also placed into the "rejection" state.

Finally, some authors have tried to retain a frequentist modeling framework while avoiding the curse of dimensionality that often plagues PIAL models. The approach taken by these authors has been to elicit information from individual decision makers that allows the analyst to specify the decision maker's choice set exactly. The underlying assumption that is made by these authors is that all alternatives that

meet the conjunctive or disjunctive criteria are deemed to be in an individual's consideration set. Given this assumption, the observation of the exact thresholds used by an individual permits one to specify an individual's consideration set with certainty. Prominent examples of models estimated in this vein include the series of papers by Kaplan et al. (2009; 2012; 2012). In addition to making use of the observed thresholds, Kaplan et al. model the choice of threshold, thereby allowing the model to be used for prediction with observations for whom thresholds have not been elicited. Another model that is estimated according to this approach is the model of Zolfaghari et al. (2013). Though similar to the Kaplan et al. models, Zolfaghari et al. allow for the possibility that individuals do not make use of all elicited attribute thresholds. As in the Jedidi and Kohli (2005) model, Zolfaghari et al. deal with the uncertainty over the number and composition of criteria being used by fully enumerating all possible combinations of number and sets of criteria. This leads to a formulation that is similar to that of a subset-conjunctive rule with uncertainty over the number of criteria that must to be met.

Across the aforementioned one-stage and two-stage models, there are two key issues that this paper seeks to address. The first issue is that the aforementioned models primarily represent only conjunctive or disjunctive rules. Only the model by Jedidi and Kohli (2005) allowed for subset-conjunctive rules, and none of the models allowed for disjunctions-of-conjunctions as described in Section 3. Secondly, the one-stage models described above suffer from theoretical issues due to their use of constraints to implement strict non-compensatory behavior. In particular, imagine that there are two attributes, x_1 and x_2, and that violating the threshold for attribute x_1 leads to a systematic utility of positive infinity while violating the threshold for attribute x_2 leads to negative infinity. Although none of the observations in one's original dataset may violate both of these estimated thresholds, there is no guarantee that these thresholds will not be simultaneously violated by one or more observations when making predictions. In a situation where both thresholds are simultaneously violated, it is not clear what value the systematic utility should be set to and how calculation of choice probabilities should proceed. The decision tree models described in Section 2 and 5 avoid this issue by using sets of conjunctive conditions that are all mutually exclusive, thus ensuring that no observation is ever described by more than one condition.

4.2. Direct choice modeling via non-compensatory protocols

As mentioned in the beginning of this section, few models have directly used conjunctive rules, disjunctive rules, or their generalizations to predict the probability of a given choice without estimating any rules or parameters that explicitly or implicitly determine an individual's choice set. To the best of our knowledge, there have only been two such modeling approaches: the cognitive process model of Zhu and Timmermans (2010) and the decision tree models of Arentze and Timmermans (2004, 2007). These will briefly be described below.

The cognitive process model first creates a new set of discrete features comprised of the originally discrete features and discretizations of the originally continuous features. The continuous features are discretized using estimated thresholds. Then, each alternative's set of discrete features are weighted using estimated weights, and a systematic utility for each alternative is created by summing the weighted, discretized features. Next, the systematic utilities are compared to estimated thresholds to determine the "state" that an alternative is determined to be in. In Zhu and Timmermans (2010), it is assumed that there is only a reject or accept state. Based on the estimated thresholds and estimated weights, conjunctive or disjunctive rules may be expressed, and some[11] disjunctions-of-conjunctions can also be expressed. A drawback of this model is that it is not clear how it works when there are more than two alternatives. In particular, it is not clear what would happen if two or more alternatives are placed into the "accept" state, and it is not clear what process would be used to determine a particular choice from the multiple acceptable alternatives.

In contrast to the cognitive process model, which is quite different from the models described in this paper, the decision tree models of Arentze and Timmermans (2004, 2007) are highly related to our work. Using either decision trees by themselves or in combination with standard discrete choice models such as the MNL model, Arentze and Timmermans directly predict the probability of a given alternative. Though not heavily emphasized in the original works of Arentze and Timmermans (2004, 2007), these models do

[11] Note, we use the qualifier "some" because it is not clear to us that all disjunctions-of-conjunctions can be expressed using some combination of weights and thresholds in the cognitive process model.

permit the same microeconomic interpretations that we are describing in this paper. However, the models in Arentze and Timmermans (2007) were motivated mostly by an attempt to the estimate the effect of discrete variables on one's systematic utilities using a non-parametric function that is adept at detecting interactions. In particular, when a decision tree is combined with standard discrete choice models in Arentze and Timmermans (2007), the decision tree is estimated based only on the explanatory variables that are originally discrete, and then a dummy variable for each output node of the tree is added to the systematic utilities of the various alternatives. The coefficients of these dummy variables are then estimated along with the usual parameters of one's choice model. As we will explain in Section 5, the models of Arentze and Timmermans (2004, 2007) are actually special cases of the more general decision tree variant known as "model trees." Moreover, as we will further explain in Section 5, our paper is the first (as far as we know) to interpret model trees as operationalizing a type of non-compensatory, context-dependent preference heterogeneity.

5. Decision Tree Variants and Economic Considerations

In Section 4, we described the way that discrete choice models have incorporated conjunctive rules, disjunctive rules, and their generalizations, and in Section 3 we showed that these non-compensatory protocols can be expressed as decision trees. In this section, we concentrate on economic considerations that are likely to arise when choice modelers consider using decision trees in their own modeling activities. In particular, we will use Subsection 5.1 to focus on the ways that decision trees can (1) make probabilistic predictions, (2) represent heterogeneity in a population's non-compensatory rules, (3) represent estimation uncertainty, (4) represent context-dependent preference heterogeneity, and (5) satisfy monotonicity constraints. After this, we use Subsection 5.2 to discuss the ways that certain combinations of these considerations have been jointly accounted for by existing decision tree variants. Additionally, since choice modelers will likely need to account for all of these considerations simultaneously, we will end this section by pointing out the remaining methodological gaps that prevent these considerations from being addressed concurrently.

5.1. Major Considerations

5.1.1. Probabilistic predictions

Some readers may note that, thus far, all of our decision tree and disjunction-of-conjunction examples have involved deterministic outputs. However, people with the same values for their explanatory variables may nevertheless make different choices. As a result, models of individual decision making need to be capable of producing probabilistic predictions. Fortunately, decision trees can and often do make probabilistic predictions in their output nodes. Conditional on a particular output node, the probability of a given alternative is often predicted to be the fraction of observations in that output node who chose the alternative in question (Arentze and Timmermans, 2004; Strobl et al., 2009).

To economically motivate the move from deterministic outputs to the more general case of probabilistic outputs, we make two observations. First, we note that individuals may explicitly have probabilistic outputs in mind when they are using disjunctions-of-conjunctions. For instance, individuals may well say "if any of these conjunctive conditions are met, then it is highly likely that I will do y," where y is some outcome. In this case, the estimated decision tree will be estimating what "highly likely" means for this population. Secondly, it has long been noted that people violate their stated thresholds and attribute cutoffs when using non-compensatory protocols such as conjunctive and disjunctive rules (Green et al., 1988; Huber and Klein, 1991; Swait, 2001b). One implication of such cutoff violations is that even if an individual consciously operates as if satisfaction of some set of conjunctive conditions will result in a deterministic outcome y, there is still some probability that an individual in may choose another alternative y' because he or she is violating their own conditions. In either motivating case[12], a decision tree will estimate the probability that each alternative is chosen from a given set of options.

[12]We are aware that in random utility maximization models, probabilistic outputs are often motivated through the argument that an analyst is unable to observe all of the variables that lead to an individual's deterministic choice. We believe that a lack of analyst omniscience will also lead to probabilistic outputs for decision tree models, but this reasoning also begs the question of how decision tree models behave when important explanatory variables are omitted. Such an investigation is beyond the scope of this paper, so for ease of exposition, we assume analysts using decision tree techniques observe all relevant explanatory variables.

5.1.2. Heterogenous non-compensatory rules

When describing human behavior, it is often unreasonable to expect that all individuals in a population will use exactly the same non-compensatory rules. For example, imagine that the decision tree shown earlier in Figure 1 is generally accurate for two individuals: one who is fit and the other who is not fit. In this case, perhaps the fit individual believes commuting by bicycle for more than 45 minutes is unacceptable whereas the unfit individual thinks bicycling longer than 20 minutes is unacceptable. Here, the two individuals differ in the value that *Travel Time* is split on in the decision tree. We will refer to this heterogeneity in the split point for an explanatory variable as local heterogeneity. In contrast, we will use the term global heterogeneity to describe the situation where even the structure of the decision tree differs across individuals. For instance, perhaps the unfit individual does not consider bicycling if the topography is hilly, regardless of the travel time. This would be heterogeneity in the set of conjunctive conditions that must be met in order for the individuals to consider bicycling. Below, we will discuss how both local and global heterogeneity have been accounted for by existing decision tree variants.

To begin, we note that local heterogeneity is fully accounted for by "soft decision trees" (Quinlan, 1990; Villandré et al., 2012), also known as decision trees with "soft splits" (Kindermann and Paass, 1998) or "fuzzy decision trees" (Jang, 1994; Olaru and Wehenkel, 2003). These decision trees place a probability distribution over the splitting point of each continuous explanatory variable. Continuing the bicycle consideration example, these probability distributions enable soft decision trees to account for more realistic scenarios where 30 minutes is unacceptable to some people, 29 minutes is unacceptable to some other people, and yet still other people find 31 minutes to be acceptable. In these scenarios, the basic structure of the tree is correct, but individuals differ on the exact point at which their requirements are met. In order to account for this situation, one can make predictions as if a split point is known, and then one can use the given distributions to marginalize over the possible split points. When using this process, one eventually ends up still using formulas such as Equation 2, but now the probability of being in a given region (i.e. a given output node) will be some value between 0% and 100% instead of being deterministic.

Turning now to considerations of global heterogeneity, we find that this concern is accommodated by decision tree ensembles (Rokach, 2010). In particular, ensembles of decision trees such as random forests (Breiman, 2001) or boosted trees (Bühlmann and Hothorn, 2007) represent global heterogeneity in much the same way that ensembles of discrete choice models (i.e. latent class choice models) represent heterogeneity amongst the compensatory decision protocols being used by differing market segments in a population (Vij et al., 2013). The basic feature of tree ensembles is that many trees are estimated, and then predictions are made by averaging the predictions of each tree in the ensemble. However, a second feature of ensembles that we highlight is the ensemble's asymptotic behavior. What happens as the number of observations being used to estimate the trees goes to infinity[13] (Minka, 2002)? Asymptotically, decision tree ensembles such as bayesian decision trees and "bagging" (a portmanteau of "bootstrap aggregation") lead to the estimation of a single tree. We interpret these ensemble methods as catering for estimation uncertainty, so these methods will be described in Section 5.1.3. In contrast, global heterogeneity is represented by the ensemble methods that estimate multiple decision trees, even as the number of observations grows without bound. Analogously, as the number of observations tends to infinity, a latent class model still returns estimates for the different market segments in a population—it does not collapse to a choice model with one class.

Despite the similarities between latent class models and decision tree ensemble methods, there are some salient implementation differences between the two types of techniques. One of the most obvious differences is that latent class models often estimate a relatively small number of classes (Allenby and Rossi, 1999), but ensemble methods usually result in models with hundreds of decision trees. While perhaps initially disconcerting, we note that having many trees makes sense behaviorally. The disjunctions-of-conjunctions used by individuals can differ in many ways. Even the simple difference between how the fit and unfit cyclists processed topography information in our earlier example would lead to two separate decision trees. As a result, a population can be expected to have many different decision trees being used by different people.

[13]Note, this discussion is closely related to the notions of model averaging versus model combination (Minka, 2002). Asymptotically, ensembles that implement model averaging will reduce to the estimation of a single tree, while ensembles that implement model combination will still estimate multiple, distinct decision trees. Model averaging is therefore seen as way to reduce estimation uncertainty while model combination accounts for global heterogeneity.

5.1.3. Estimation uncertainty

In many statistical applications, quantifying one's inferential uncertainty is important. For models that depend on continuous parameters, uncertainty is often quantified by the sampling distribution of one's estimator. However, unlike traditional models that are indexed by continuous parameters, decision trees are made up of discrete parameters such as the depth of the decision tree, the variables that the tree is split on, the values of the variables that are being split on, etc. In such discrete settings, uncertainty is quantified by the probability of a given combination of parameters being the data-generating parameters. In other words, we need the probability of any given tree being the "correct tree." Unfortunately, as with estimation of the tree, one will have to make approximations since complete enumeration of the possible decision trees is typically prohibitive (Chipman et al., 1998, p. 960).

Here, as noted in Section 5.1.2, ensembles methods such as bayesian decision trees and bagging can provide a measure of estimation uncertainty. That bayesian decision trees provide the desired uncertainty quantification is due to the fact that bayesian methods explicitly estimate posterior probabilities of particular parameter values being true. The link between uncertainty quantification and bootstrap aggregation (i.e bagging) comes from the fact that the bootstrap is equivalent to a traditional bayesian analysis using a particular prior (Rubin, 1981; Newton and Raftery, 1994). In both cases, one would take the fraction of times a particular decision tree appears in the ensemble as being an estimate of the probability that the given decision tree is the "true" tree. These methods provide an approximate measure of the estimation uncertainty because there is no guarantee that these ensembles will contain all possible decision trees (Chipman et al., 1998, p. 960).

5.1.4. Context-dependent preference heterogeneity

In the discrete choice literature, and in the broader literature concerning human decision-making, it has long been acknowledged that "the context in which a decision is made is an important determinant of outcomes" (Swait et al., 2002). In particular, one's choice context may affect one's preferences or sensitivities to a given set of explanatory variables, and we use the term "context-dependent preference heterogeneity" to refer to this phenomenon. As an example, consider an individual making a choice of travel mode for his/her commute. When the cost of a given travel mode is low, perhaps the individual is most sensitive to that mode's travel time. However, when the cost of the travel mode is high, perhaps the individual becomes more sensitive to changes in travel cost than to changes in travel time. For such a simple scenario, a piecewise linear function for one's systematic utility may be sufficient. However, for scenarios where preferences are dependent on arbitrarily complex conditions, potentially involving multiple variables, we do not know of any accommodating methods within the traditional discrete choice literature.

Looking instead to the literature on decision tree methods, we note that decision tree variants known as "hybrid," "model," or "functional" trees (Zeilis et al., 2008; Rusch and Zeilis, 2013) are able to account for such notions of context-dependent preference heterogeneity. Model trees are decision trees where the output at a given output node is a statistical model (Chan and Loh, 2004; Landwehr et al., 2005; Zeilis et al., 2008; Yu et al., 2016). To make predictions, the decision tree is used to determine the output node that corresponds to the given observation, and then that output node's statistical model is used to provide the final outcome probabilities for the observation. In the specific case where discrete choice models are used in the output nodes, preference heterogeneity is represented by differing systematic utility functions in the models used in different nodes. Returning to our example from the previous paragraph, imagine that we had a decision tree that was split on the *Travel Cost* variable at a value that distinguished "low" versus "high" travel costs. The model at the low-travel-cost output node might have a systematic utility function that is linear-in-parameters with a coefficient β_{LowCost} being multiplied by the travel-cost variable. Conversely, the model at the high-travel-cost output node might also have a linear-in-parameters systematic utility function, with a coefficient β_{HighCost} being multiplied by the travel-cost variable, where $\beta_{\text{HighCost}} > \beta_{\text{LowCost}}$. Such a model tree would capture the notion that preferences (in this case, the travel-cost coefficients) are dependent on the context in which the choice is being made—a low travel cost context versus a high travel cost context.

Beyond the general description provided in the previous paragraph, we pause here to note that many decision tree methods and discrete choice methods can be seen as special cases of model trees. First, the standard decision tree described in Section 2 can be seen as a model tree where discrete choice models such as the MNL are used in each node, and each alternative's systematic utility is only comprised of an alternative specific constant (ASC). For decision trees with deterministic outputs, these constants are

either infinity or negative infinity. For decision trees with probabilistic outputs, the relative values of these constants can be determined by constraining a reference alternative's ASC to zero, and determining what ASCs of the other alternatives will lead to the decision tree's estimated choice probabilities. Secondly, other proposed models estimate a decision tree and then place a dummy variable for each output node into one's systematic utility functions in a discrete choice model. This methodology includes models such as the parametric-action decision tree (Arentze and Timmermans, 2007), the hybrid CART-logit model (Steinberg and Cardell, 1998), the tree-augmented logistic model (Su, 2007), and the two-stage MNL model(Kim, 2009; Kim and Kim, 2011). Such models can be seen as special cases of model trees that allow for context-dependent heterogeneity in the ASCs but enforce homogeneity on the remaining parameters in the choice models. Finally, the semi-compensatory models used in the discrete choice literature are also special cases of model trees. In these semi-compensatory models, described in Section 4, conjunctions, disjunctions, or disjunctions-of-conjunctions are used to screen alternatives and then a compensatory discrete choice model is used to select from any remaining alternatives. This can be seen as a model tree where the parameters of the systematic utility function for available alternatives are constrained to be equal across the various output nodes, and output nodes that result in a given alternative not being available simply set the systematic utility for that alternative to negative infinity.

5.1.5. Monotonicity

Lastly, we note that models of human decision making are often subject to constraints based on economic theory. For instance, all else equal, as the price of a normal good increases, the probability that this good is chosen should decrease or, at worst, stay the same. This is a monotonicity constraint. In discrete choice models that use linear-in-parameters systematic utility functions, such monotonicity constraints are operationalized through constraints on the sign of the model coefficients. These sign constraints allow one to quickly check if one's estimated parameters comply with economic theory about the relationship between an explanatory variable and an outcome of interest. And as noted in the introduction, discrete choice modelers are highly unlikely to use a model that does not demonstrate compliance with economic theory.

Fortunately, decision tree variants that can incorporate monotonicity constraints have been created (Potharst and Feelders, 2002; Velikova and Daniels, 2004; Hu et al., 2012; Marsala and Petturiti, 2015; Pei et al., 2016). Such monotonic decision trees are constructed by altering the estimation process to ensure that the desired monotonicity constraints are not violated. By using monotonic decision trees, one can estimate the disjunctions-of-conjunctions that may be in use in one's population, while at the same time guaranteeing compliance with economic theory. The ability to ensure the monotonicity of key relationships should go a long way towards easing the concerns of choice modelers who are considering using decision trees in their analyses but want to make sure that their estimated trees "make sense."

5.2. Combining considerations

In Subsection 5.1, we sequentially detailed how various types of decision trees allow researchers to (1) make probabilistic predictions, (2) represent heterogeneity in a population's non-compensatory rules, (3) represent estimation uncertainty, (4) represent context-dependent preference heterogeneity, and (5) satisfy monotonicity constraints. However, in real applications, analysts may wish to simultaneously account for all of the considerations described above. In this subsection, we will briefly detail the ways that such goals can and cannot yet be met. Our discussion will point out advanced decision tree variants as well as point to methodological gaps that must be filled in order to make decision trees maximally useful to discrete choice researchers.

To begin, we first point out that all decision tree variants allow for the use of probabilistic predictions. Accordingly, we will focus our discussion on considerations (2) - (5), listed above. Next, we will make the point upfront that there are no decision tree variants that currently account for all four of the remaining considerations. The best that can be done with available methods is to account for combinations of two or three of considerations (2) - (5). Moving swiftly through such combinations, the only three considerations that have been combined are the representation of local heterogeneity, the representation of estimation uncertainty and the representation of context-dependent preference heterogeneity. These three concerns are simultaneously accounted for in the decision tree variant known as a bayesian hierarchical mixture-of-experts model (Bishop and Svensén, 2003). Such a model makes use of model trees with soft-splits and uses bayesian estimation techniques to account for estimation uncertainty. Moving to combinations of two of the four

considerations, only three of the six possible combinations have been accounted for in the literature. First, bayesian soft decision trees (Kindermann and Paass, 1998) and bagged soft decision trees (Yildiz et al., 2016) allow for estimation uncertainty and representations of local heterogeneity. Furthermore, soft tree ensembles such as a random forest of soft trees (Seyedhosseini and Tasdizen, 2015; Kumar et al., 2016) allow for representations of both local and global heterogeneity. Secondly, soft model trees known as mixtures of experts or hierarchical mixtures of experts (Jordan and Jacobs, 1994; Yuksel et al., 2012) allow for context-dependent preferences and local heterogeneity. Thirdly, global heterogeneity and monotonicity have been jointly represented by monotonic random forests (González et al., 2015).

To the best of our knowledge, no combination of considerations has been addressed beyond those detailed in the last paragraph. As a result, by developing decision tree models that account for the missing combinations of economic considerations, discrete choice researchers can help advance the fields of computer science and statistics while simultaneously catering for properties they wish to have in their own analyses. In Section 6, we illustrate such development by formulating and estimating what we believe is the first bayesian model tree. This allows us to account for estimation uncertainty and context-dependent preference heterogeneity. While not simultaneously addressing all of considerations (2) - (5) mentioned above, our model nevertheless fills a missing rung in the methodological ladder of existing decision trees.

6. Empirical Application

In the last section, we showed how common economic concerns can be addressed by existing variants of decision trees. Additionally, we pointed out gaps in existing decision tree methodologies that need to be filled in order to make decision trees most useful when modeling economic phenomena. In this section, we switch focus and review our paper's empirical application. Given the economic interpretation of decision trees representing disjunctions-of-conjunctions, we study whether such rules appear to be used by commuters in the San Francisco Bay Area. In particular, we model how disjunctions-of-conjunctions are used to choose whether or not bicycle would be considered as a travel mode and, if bicycle was considered, how the disjunctions-of-conjunctions affect the overall preference for bicycling when choosing between the considered travel modes. Moreover, we take pains to capture our uncertainty in the estimated disjunctions-of-conjunctions. As a result, our application contributes to the literature by creating the framework and estimation techniques for the first decision tree variant that accounts for both context-dependent preference heterogeneity and model uncertainty.

In the following subsections, we first review the motivation for our proposed semi-compensatory model (i.e. the combination of a decision tree with a standard mode choice model). Next, Section 6.2 reviews the details of how our proposed model works, and Section 6.3 details the proposed and implemented estimation techniques for our new model. In Section 6.4 we detail the model specification and data used in our application, and in Section 6.5 we present our results and discussion.

6.1. Motivation

As previously noted, our application concerns the choice of travel mode in the San Francisco Bay Area. Specifically, we are interested in whether people choose to commute by bicycle. Of critical importance are two phenomena. First, individuals may (for a variety of reasons) exclude bicycling from consideration, thereby removing all possibility that they will use a bicycle to commute to work/school. If such differences in consideration are not accounted for, then one will make incorrect inferences regarding the amount by which any project can be expected to increase the expected number of cyclists. Secondly, individuals may find themselves in situations that lead them to be more or less amenable to the idea of commuting by bicycle. If an individual has a very low general preference for bicycling, then policies to increase bicycling rates may only have a minor impact on this individual's probability of bicycling. In other words, before judging the ability of an intervention to increase the probability that the individual actually bikes, one must be sure that an individual is considering bicycling as a commuting option, and one should attempt to judge an individual's general preference for bicycling.

In previous discrete choice research that allowed for heterogeneous consideration sets, mode choice models have been operationalized based on assumptions regarding: the existence of latent market segments that each have their own consideration sets and utility coefficients (Vij et al., 2013; Vij and Walker, 2014), the existence of individuals that have either complete choice sets or who irrationally only consider a single travel

mode (Swait and Ben-Akiva, 1987b), or whether alternatives are independently chosen for inclusion in one's consideration set (Swait and Ben-Akiva, 1987a; Swait, 2001a, 2009). With these formulations, researchers have already found support for the hypothesis that, beyond deterministic differences in the travel modes which are available to a given person, individuals differ in whether they consider bicycling as a commuting option and in how much they generally prefer cycling (Swait, 2009; Vij et al., 2013; Vij and Walker, 2014; Mahmoud et al., 2016).

In all the modeling efforts just described, the probability of an individual considering a particular mode was always based on a compensatory model. These models are curious in light of the fact that when asked about why they don't commute by bicycle, individuals do not state that the issues which make them avoid bicycling to work can be compensated for by other commonly used variables in mode choice models. Individuals commonly state that they live too far away to commute by bicycle, that roadway conditions are too dangerous for them to commute by bike, that cycling would require too much physical exertion, that they have to transport children to some place, and so on (Goldsmith, 1992; Cleland and Walton, 2004). It is not clear *a-priori* that these type of concerns can be incrementally compensated for by changes in sociodemographic variables or level-of-service variables for the various travel modes. As a result, it is reasonable to think that non-compensatory models of consideration set formation may be better able to emulate the actual decision making process of individuals. Our goal for this application was to develop a policy analysis tool for bicycling that could capture the effect of non-compensatory protocols on choice set formation and on the general preference for bicycling. We used disjunctions-of-conjunctions as our non-compensatory protocol in order to account for the "if-then" nature of people's stated reasons for not bicycling. Beyond using decision trees to model the consideration of the bicycle alternative, we wanted to be sure to account for the effect of the attributes of the non-bicycle alternatives. As a result, we follow the lead of the semi-compensatory models reviewed in Section 4 by using a compensatory model to predict the final choice between any alternatives that are considered.

6.2. Model Framework

In the last subsection's discussion, we reviewed why we desire a semi-compensatory model that combines decision trees and discrete choice models. In this subsection, we will review our desired model in more detail so readers are clear about how it works and so that readers of Section 6.3 have enough context to understand why we chose the estimation methods that we chose.

First, as described in Section 5.1.4, our proposed type of model is known in the decision tree literature as a model tree. Model trees are decision trees that use statistical models in their output nodes to predict the outcome of interest. Here, the statistical models in the output nodes typically differ from one another. In our application, the model tree will function as follows. There will be a decision tree with mode choice models in the output nodes. The tree will be used to winnow the bicycle from an individual's choice set, and across the different situations where bicycle is considered, the general preference for bicycling will be allowed to differ. This results in differing bicycle ASCs in the mode choice models of the different output nodes of the decision tree. For simplicity, we have constrained the other parameters in our choice model to remain constant across the various output nodes. In other words, accounting for context-dependent preference heterogeneity in the parameters other than the bicycle ASC is left for future research, as is accounting for global and local heterogeneity in the estimated disjunctions-of-conjunctions or accounting for a-priori monotonicity constraints.

Second, we go beyond the mere use of model trees as they have already been implemented. Instead, we contribute to the literature of decision tree methodologies by developing a bayesian model tree. By using bayesian estimation techniques, we can account for estimation uncertainty about which model tree is the "true" tree. These various candidate trees, denoted by m, represent different non-compensatory decision protocols, and we are using the bayesian estimation to compute the probabilities of these different protocols being the one used in our population. In addition, as is always done when estimating bayesian choice models, the bayesian estimation also accounts for the estimation uncertainty in the choice model parameters.

Now, because we are estimating a model tree, we can partition the model parameters into those that describe the tree and the parameters that describe the choice models at the output nodes of the tree. We will start with the tree parameters. Using the notation from Section 3.2, a decision tree is uniquely identified by three sets of parameters. The first parameter is how many conjunctive conditions (i.e. output nodes) are in the tree. We denote this as D^m. The second set of parameters is how many requirements are in each

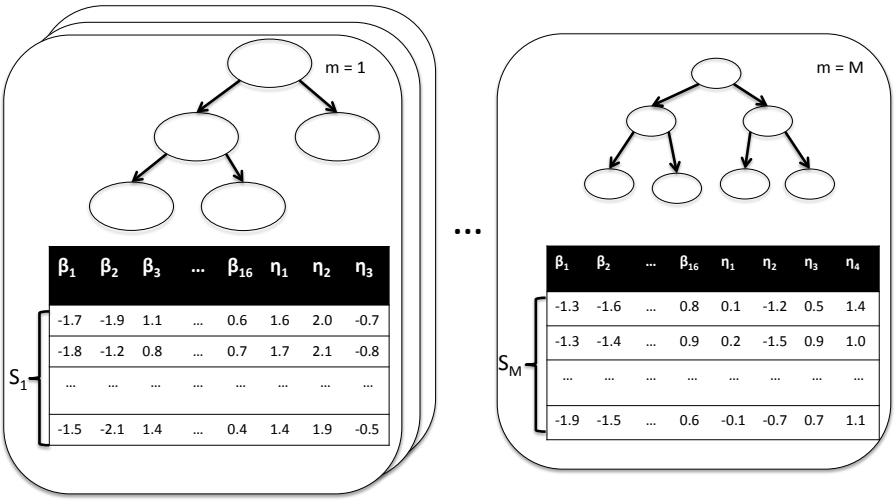

Figure 3: Procedural diagram of bayesian model trees

conjunctive condition. We denote these parameters as $\mid p_i^m \mid$, where $i \in \{1, 2, ..., D^m\}$. Lastly, the third set of parameters is the primitive boolean conditions that make up each requirement. We denote these parameters as $b_j^{i,m}$ where $j \in \{1, 2, ..., \mid p_i^m \mid\}$.

Next, we will move onto the parameters of the choice models at the output nodes of the tree. We denote these parameters as γ^m, and we note that in our application, we are only allowing the bicycle ASC to differ across output nodes. As a result, we can further partition the parameters that describe the choice models at the output nodes. Conditional on a tree (m), there will be one parameter per output node (i), and these parameters will determine the bicycle ASC for the given node. We will denote these node-varying parameters by η_i^m. Additionally, there will be the remaining choice model parameters that do not change from one output node to the next. We will denote these parameters by β. All together, we have $\gamma^m = (\eta_i^m, \beta)$. Combining this paragraph with the last, the parameters to be estimated are $\theta = \left(D^m, \mid p_i^m \mid, b_j^{i,m}, \eta_i^m, \beta\right)$ for all $i \in \{1, 2, ..., D^m\}$ and for all $j \in \{1, 2, ..., \mid p_i^m \mid\}$.

Due to the bayesian estimation techniques, our estimation results will now be a posterior distribution that reflects our uncertainty in the "true tree" and in the "true" parameters of the choice models in that tree's output nodes. Moreover, since we do not have a closed-form expression for this posterior distribution, it will be represented by a sample from this joint distribution of trees and choice model parameters. Each sampled element (s) will be a decision tree (m) and the parameters of the choice models at that tree's output nodes (γ_s^m). We denote the number of sampled elements containing tree m as S_m. Next, we can use the fraction of times that a specific tree appears in the posterior sample to estimate the posterior probability of a given tree $(P_{\text{Post}}(m) = \frac{S_m}{\sum_\ell S_\ell})$. Finally, in a bayesian model tree setting, we calculate the predicted probability of outcome Y given explanatory variables X using the following formula:

$$\begin{aligned}
\hat{P}(Y \mid X) &= \sum_{m=1}^{M} P_{\text{Post}}(Y \mid X, m) P_{\text{Post}}(m) \\
&= \sum_{m=1}^{M} \left[\frac{1}{S_m} \sum_{s=1}^{S_m} P(Y \mid X, \gamma_s^m, m) \right] P_{\text{Post}}(m) \qquad (6) \\
\text{where } M &= \text{The total number of unique trees in one's sample.} \\
P(Y \mid X, \gamma_s^m, m) &= \text{The choice model probability of } Y \text{ given } X, \gamma_s^m, \text{ and tree } m.
\end{aligned}$$

For a graphical depiction of this process, see the diagram in Figure 3.

6.3. Estimation Methods

The previous subsection reviewed the overall framework, mechanics, and parameters of our proposed bayesian model tree. In this subsection, we detail our estimation techniques. These details are discussed at length because we found estimation of this new model to be a nontrivial challenge, and we want other researchers to be able to replicate and build off our work. Readers who would like to immediately get to the results and 'big-picture' discussion may feel free to skip ahead to Section 6.5.

Subsection 6.2's formulation of θ shows that the total number of parameters being estimated depends on the decision tree. In particular, as we change from tree to tree, the number of conjunctive conditions (D^m) will vary, and as a result, the dimensionality of the parameter vector will vary. Unfortunately, such changing dimensionality necessitates the use of specialized estimation techniques (Das and Bhattacharya, 2017, p.5). Of these, the reversible-jump algorithm (Green, 1995) is the most common bayesian estimation technique for problems of varying dimensionality, both overall (Sisson, 2005) and for decision trees in particular (Denison et al., 1998; Wu et al., 2007; Mohammadi and Kaptein, 2016).

As noted by Fan and Sisson (2011, p.72-73), efficient reversible-jump algorithms require a way for one to propose parameter values of high posterior probability while switching between parameter spaces of varying dimensions, and creating such proposal mechanisms is not straightforward. Our initial attempts at using a reversible-jump algorithm to estimate our model tree failed because we were unable to devise a good way to propose new choice model parameters when switching from one tree to another. How should we propose new bicycle ASCs when the groups of individuals in each output node are completely different? The creation of an efficient, reversible-jump proposal mechanism for bayesian model trees remains an open problem, and it is one that we would be happy to collaborate with others on.

Given our difficulties with the reversible-jump algorithm, we instead sought an alternative estimation strategy. The approach we settled on was to split the problem into two sub-problems, each of which was more easily solved than the original problem. Specifically, as we noted in Section 2.2, there are existing methods for performing a bayesian estimation of decision trees. Additionally, one can estimate the parameters of a given choice model using almost all existing bayesian estimation techniques for fixed-dimensional problems. In light of these two facts, we sought to break our model tree estimation into a first step where we estimate the decision trees by themselves and a second step where, conditional on a given decision tree, we estimate the choice models that belong in each output node of the tree. Finally, some procedure would be needed to tie these two estimation tasks together.

To implement this divide-and-conquer approach, our original (and idealized) plan was as follows. First, we would use the techniques of Letham et al. (2015) to perform a bayesian estimation of the decision trees. Then, conditional on each tree, we would use the techniques of Braun and Damien (2016) to estimate our mode choice model with varying bicycle ASCs. And lastly, we would use importance sampling to adjust the original posterior distribution of decision trees in light of the information provided by the choice models at the output nodes at each tree. Below, we briefly justify each of these choices.

Beginning with the estimation of the decision trees, we chose to use the techniques of Letham et al. (2015) for two reasons. First, their methods were implemented in freely available python scripts, so we would not have to re-invent their techniques. Secondly, their approach requires researchers to specify the possible requirements that can be used in the conjunctive conditions that comprise the decision tree. This specification gives researchers the ability to check for sensible relationships between the explanatory variables and the outcomes of interest. For example, by specifying the regions of parameter space that the travel distance is split into, the researcher can empirically check whether the fraction of individuals bicycling decreases as one moves from the region where travel distance is between 2 and 3 miles to the region where travel distance is between 3 and 4 miles.

Moving to the choice model estimation, we (again) had two reasons for choosing the techniques of Braun and Damien (2016). First, unlike typical MCMC procedures that only generate dependent samples from the posterior distribution of one's choice model parameters, the techniques of Braun and Damien (2016) generate independent samples, resulting in higher effective sample sizes per unit of computational time. Secondly, the methods of Braun and Damien (2016) automatically provide accurate estimates of the total probability of the data given one's decision tree (i.e. after marginalizing over the parameters in the choice model). This probability is needed for our last step: importance sampling.

After the initial estimation of the decision trees and the mode choice models, conditional on the decision trees, we need to link these two estimation procedures. In particular, we want a sample from the joint

distribution of decision trees and their accompanying choice models. However, our original sample of decision trees was produced without using any information from the choice models at the output nodes. As a result, our original sample of decision trees is (in general) drawn from an incorrect distribution. We use importance sampling (Gelman, 1992; Hesterberg, 1995) to weight our original sample of decision trees such that the weighted sample comes from our desired distribution. Since our original sample was drawn from a distribution $P\left(\text{tree without choice models} \mid \text{data}\right)$ instead of $P\left(\text{tree with choice models} \mid \text{data}\right)$, we will weight each tree by the ratio $\frac{P(\text{tree with choice models}|\text{data})}{P(\text{tree without choice models}|\text{data})}$. The probabilities in the numerator and denominator are computed up to a constant of proportionality using Bayes rule, and then the importance weights are normalized such that they sum to one across all the trees in our sample. At this point, estimation is complete and the weighted sample is then available for prediction or further inference tasks.

As just described, this three step procedure is our current, ideal method for estimating bayesian model trees. However, this procedure is computationally expensive. For example, our initial sample of trees contained more than 5,000 unique decision trees. On average, for a single decision tree, it took approximately 2 hours to perform the bayesian estimation of the choice models at the output nodes. The total estimation time would have taken more than a week for our dataset and choice model specification (described in Section 6.4). Given our current computing resources (a single laptop), we deemed this estimation time unreasonable, so we made further approximations to speed up the estimation process. In particular, we selected a subset of 10 decision trees from the total set of unique trees so that the total estimation time would be less than a day. Then, we then estimated the choice models at the output nodes of these trees, and we proceeded as if these ten trees were the complete set of possible trees for our data. As far as we know, it is impossible to account for the existence of the other trees without performing the estimation of those trees' choice models, which is exactly what we wished to avoid. While numerous ways of choosing the ten trees are possible, we tried to follow the intuition of Breiman (2001) who noted that the accuracy of a set of trees "depends on the strength of the individual tree classifiers and a measure of the dependence between them." Specifically, we chose the ten trees as follows. We chose top three trees in terms of their (approximate) log-posterior from step 1, and we also chose the top three trees in terms of their (approximate) log-likelihood from step 1[14]. We chose the final four trees by first selecting the trees that had approximately the posterior mean number of output nodes (D^m) and then, from the selected trees, choosing the 4 trees with the highest log-posterior. This procedure closely follows the recommendation of Letham et al. (2015) for selecting a single decision tree to be a point estimate for the posterior distribution of trees. In the end, our selected trees were all "strong" in some way, whether that be high log-likelihoods or high log-posterior values, and across the three selection criteria, the trees were quite different from one another. We will refer to the procedure described in this paragraph as our "actual" estimation methodology, whereas the procedures described in the paragraphs above are our "ideal" estimation methodology. As we will see in Section 6.5, despite our radical simplifications, our actual estimation methodology still produces a model that provides quantitatively more accurate and qualitatively more reasonable inferences than the traditional MNL model.

6.4. Data and model specification

In the previous subsections, we described our model framework and estimation methods. In this section we describe the data used in our application and the precise specification (i.e. model priors and choice model specification) of our bayesian model trees.

6.4.1. Data

Starting with the data, we are using 1,015 observations from the California Household Travel Survey (2013). Each individual in our sample lives in Oakland, Berkeley, or San Francisco, CA, and the observations represent home to work or school commute tours. For level-of-service variables (such as travel time, cost, and distance) we use estimates provided by the San Francisco Metropolitan Transportation Commission (MTC) (2012). Basing our set of possible alternatives on the alternatives used by MTC, we classify observations as having traveled via one of eight travel modes. There were three driving modes, each differentiated by the number of passengers: drive-alone, shared-ride with two passengers, and shared-ride with three or more

[14]Note, we use the term 'approximate' because the log-posterior values and log-likelihood values of step 1 do not take into account the choice models at the output nodes of the tree.

passengers. There were also three transit modes, each differentiated by their access and egress modes: walk-transit-walk (where walking is used for access and egress), drive-transit-walk, and walk-transit-drive. Finally, there were two non-motorized modes: walking and bicycling. For each tour, the travel mode that was used for the longest distance was used as the "chosen travel mode" for that tour.

Importantly, one of our uses for non-compensatory rules is to determine whether or not an individual considers bicycling as a travel mode. Accordingly, our decision trees are based on spatial variables and socio-demographics that have been mentioned in reasons why individuals did not consider bicycling. In particular, the trees are based on spatial variables such as distance, roadway slopes, elevation, on-street bicycle infrastructure, speed limits, and socio-demographics such as the number of children. Post-processing of the raw spatial data was done using a novel concept called the zone of likely travel. The main idea is, for each individual, to form a buffer around the shortest path between one's home and work or school. This buffer is constrained to follow the roadway network instead of merely being laid atop of a map, and the buffer is constructed so its perimeter is based on each user's likely, maximum deviation from the shortest path. In other words, the zone should contain the roadways over which one is likely to travel. All spatial variables are then calculated over the roadways in one's zone of likely travel. In general, the details of this post-processing procedure are not related to the main purpose of this paper, so we will not review them any further. However, we instead encourage interested readers to review the details of this processing in Brathwaite (2018).

6.4.2. Model Specification

Traditionally, when discrete choice modelers talk about model specification, they mean the specification of one's utility functions. However, in a bayesian paradigm, one also needs to specify his/her model priors. These priors are probability distributions that encapsulate the modeler's prior beliefs about the true value of the model parameters. Together with the utility specifications and likelihood function, these specification choices allow for model estimation. Below, we will note each our specifications in turn, starting with the choice model.

Specifically, we specify the systematic utility functions in our choice model as follows:

$$
\begin{aligned}
V_{DA} &= \beta_{\text{travel-time-auto}}\text{TravelTime}_{\text{DA}} + \beta_{\text{autos-per-driver}}\text{AutosPerDriver} \\
V_{SR2} &= \text{ASC}_{\text{shared-ride-2}} + \beta_{\text{travel-time-auto}}\text{TravelTime}_{\text{SR2}} + \beta_{\text{autos-per-driver}}\text{AutosPerDriver} \\
&\quad + \beta_{\text{cross-bay}}\text{CrossBay} + \beta_{\text{num-kids}}\text{NumberKids} + \beta_{\text{household-size}}\text{HouseholdSize} \\
V_{SR3} &= \text{ASC}_{\text{shared-ride-3}} + \beta_{\text{travel-time-auto}}\text{TravelTime}_{\text{SR3}} + \beta_{\text{autos-per-driver}}\text{AutosPerDriver} \\
&\quad + \beta_{\text{cross-bay}}\text{CrossBay} + \beta_{\text{num-kids}}\text{NumberKids} + \beta_{\text{household-size}}\text{HouseholdSize} \\
V_{WTW} &= \text{ASC}_{\text{walk-transit-walk}} + \beta_{\text{travel-time-transit}}\text{TravelTime}_{\text{WTW}} + \beta_{\text{travel-cost-transit}}\text{TravelCost}_{\text{WTW}} \\
V_{WTD} &= \text{ASC}_{\text{walk-transit-drive}} + \beta_{\text{travel-time-transit}}\text{TravelTime}_{\text{WTD}} + \beta_{\text{travel-cost-transit}}\text{TravelCost}_{\text{WTD}} \\
V_{DTW} &= \text{ASC}_{\text{drive-transit-walk}} + \beta_{\text{travel-time-transit}}\text{TravelTime}_{\text{DTW}} + \beta_{\text{travel-cost-transit}}\text{TravelCost}_{\text{DTW}} \\
V_{walk} &= \text{ASC}_{\text{walk}} + \beta_{\text{distance-walk}}\text{TravelDistance}_{\text{walk}} \\
V_{bike} &= \text{ASC}_{\text{bike}} + \beta_{\text{distance-walk}}\text{TravelDistance}_{\text{bike}}
\end{aligned}
\tag{7}
$$

In the systematic utility equations above, DA means "drive alone," SR2 means "shared-ride with two passengers," SR3 means "shared-ride with three or more passengers," WTW means "walk-transit-walk," WTD means "walk-transit-drive," and DTW means "drive-transit-walk." Though not indicated using subscripts on the variables, all of these systematic utility equations are specific to a given individual.

Next, we note that our specifications above were not made arbitrarily. Travel cost was excluded from the driving alternatives because it was too collinear with the travel time variable to permit estimates that had the correct sign. This is to be expected since MTC calculates both its travel cost and travel time estimates for driving modes as a function of travel distance. Secondly, income and gender are not present in our specifications because it was missing for numerous individuals in our dataset.

Finally, the systematic utility specifications shown in Equation 7 are common across both the MNL model and the bayesian model trees used in this paper. There are only two differences between the systematic utility specification of the MNL and the bayesian model trees. First, as mentioned above, the ASC_{bike} is allowed to differ from output node to output node. In other words, the bayesian model trees replace ASC_{bike} with

$\sum_{i=1}^{D^m} \delta_i ASC_{\text{bike},i}$ where i denotes a particular output node of decision tree m and δ_i is a dummy variable that indicates whether or not an individual is in output node i. Briefly, we note that we do not directly estimate the parameters $ASC_{\text{bike},i}\ \forall i \in \{1, 2, ..., D^m\}$. This would correspond to using a *no-pooling* estimator that treats the output nodes as being completely different from one another. Instead, we would rather estimate how different the nodes are from one another. To do this, we use a hierarchical logit estimator (i.e. a *partial-pooling* estimator) (Bafumi and Gelman, 2006; Gelman, 2006; Gelman et al., 2014) that combines (i.e. pools) information about the overall bicycle preference across output nodes. The $ASC_{\text{bike},i}$ parameters are conceptualized as instances from an overall, normal distribution of bicycle ASCs with mean ASC_{bike} and variance σ^2_{bike}. Here, the mean and variance parameters are estimated along with the individual $ASC_{\text{bike},i}$ parameters[15]. As $\sigma^2_{\text{bike}} \to \infty$, we are increasingly certain that the output nodes are completely different from one another, and as $\sigma^2_{\text{bike}} \to 0$ we are increasingly confident that the general preference for bicycling is actually the same across output nodes.

The second difference is, as noted in Section 6.4.1, that we use spatial variables in the construction of the decision trees. In order to fairly compare the MNL and the bayesian model tree, we include the spatial variables in the MNL model by placing these variables in the bicycle systematic utility. In particular, the bicycle utility of the MNL model is expanded to include the following variables and their coefficients: shortest path length, median slope, average speed limit, proportion of roadway miles on the shortest path with speed limits of 25 mile per hour or less, proportion of roadway miles with bicycle lanes, and the proportion of roadway miles with "share the road" markings (also known as "sharrows"). These variables are excluded from the bicycle utility of the choice models in the bayesian model tree as they are already used when constructing the decision trees.

Next we state our model priors. In a bayesian setting, priors must be specified for all parameters that are being estimated. We start with the choice model parameters. For all choice model parameters, excluding $ASC_{\text{bike},i}\ \forall i \in \{1, 2, ..., D^m\}$ and σ^2_{bike}, we assumed independent priors of $\mathcal{N}(0, 4)$ where 4 is the variance of the normal distribution. This prior was chosen to reflect the fact that we think it is highly unlikely for a 1-unit change of any of our variables to cause a change of 4 in our systematic utility functions. Such changes would greatly increase or decrease the probability of choosing a given alternative, and we don't expect a 1 minute change in travel time, a 1 dollar change in travel cost, a change in 1 mile of travel distance, etc. to cause drastic changes in the probability of a given mode. For, the $ASC_{\text{bike},i}$ parameters, we use the prior distribution mentioned above. That is, the prior distribution of $ASC_{\text{bike},i}$ is $\mathcal{N}\left(\text{ASC}_{\text{bike}}, \sigma^2_{\text{bike}}\right)$. Here, we again use a $\mathcal{N}(0, 4)$ for the hyperprior on ASC_{bike}. The hyperprior for the variance is specified as $\ln[\mathcal{N}(0, 4)]$, i.e. log-normal with a location parameter of zero and a scale parameter of 2 ($\sqrt{4} = 2$).

Moving to our priors for the parameters of the model trees, we need a prior distribution for $\left(D^m, \mid p_i^m \mid, b_j^{i,m}\right)$ for all $i \in \{1, 2, ..., D^m\}$ and for all $j \in \{1, 2, ..., \mid p_i^m \mid\}$. To construct our prior, we precisely follow the methodology described in Section 2 of Letham et al. (2015). Unfortunately, this methodology took Letham et al. nearly four pages and much mathematical notation to describe. Additional pages would be needed to relate their original description to the characterization of decision trees that we have given in Sections 2 and 3. Since reviewing the techniques of Letham et al. (2015) is not a primary focus of our article, we state upfront that the following description of our prior distribution of decision trees will be necessarily brief and will likely require a reader to consult Letham et al. (2015) for full understanding. For readers who prefer reading code to reading verbal descriptions of our procedures, all scripts used in this application are available upon request.

Now, we begin with D^m, the number of output nodes (or conjunctive conditions) in our decision tree. Given that one of the arguments for non-compensatory rules is that humans are boundedly rational and only spend but so much mental effort making decisions, we do not think individuals are using overly complex rules. Our prior for D^m was therefore specified as a truncated Poisson distribution with a rate parameter of 5, reflecting our prior belief that the expected number of output nodes in one's decision tree is approximately

[15] To tie this paragraph back to Section 6.2 where we first discussed our model parameters, we define $ASC_{\text{bike},i} = ASC_{\text{bike}} + \eta_i$. In our application we actually estimate η_i and ASC_{bike} instead of $ASC_{\text{bike},i}$ and ASC_{bike}. In the statistical literature, this choice is referred to as the use of a non-centered parametrization (Papaspiliopoulos et al., 2007). We used the non-centered parametrization instead of the traditional approach of directly estimating $ASC_{\text{bike},i}$ because this method led to faster estimation times.

five. A truncated (as opposed to standard) Poisson distribution was used because the support of the standard Poisson distribution extends to positive infinity whereas the number of possible conjunctive conditions for the trees is limited by the finite number of possible requirements from which the conjunctive rules can be composed. See Letham et al. (2015, p. 1355) for the specific form of the truncated Poisson distribution and for more details on this prior specification.

Continuing to the next parameter, we have to specify a prior for $\mid p_i^m \mid$: the number of requirements in each conjunctive condition. We will not delve into the details here, but the methods of Letham et al. (2015) use a slightly different representation of decision trees than have been described in this paper. In their formulation, output nodes are evaluated sequentially, and $\mid p_i^m \mid$ represents the number of requirements in output node i, conditional on the requirements of the previous nodes not being met. Given this set up, and given the assumption that people are using relatively simple rules to make their decisions, we specify our prior for $\mid p_i^m \mid$ as a truncated Poisson distribution with a rate parameter of 2. In other words, besides the requirement of not meeting the conditions specified by the previous output nodes, we expect that a given output node will be described by approximately two requirements.

Next, we need prior distributions for the requirements ($b_j^{i,m}$) that make up each conjunctive condition. We pause here to note that such prior distributions implicitly define a prior on the conjunctive conditions (p_i) that correspond to each output node. Alternatively, placing a prior directly on the conjunctive conditions (p_i) will implicitly define a prior on the requirements ($b_j^{i,m}$) that make up these conditions. Following the procedures in Letham et al. (2015), we use a three-stage procedure to place a prior directly on the conjunctive conditions (p_i). First, we specify the possible requirements that a conjunctive condition can be composed of. These requirements are formed by discretizing the explanatory variables into various ranges (e.g. minimum distance greater than 4 miles). Second, we specify which of the possible combinations of requirements will be allowed as possible conjunctive conditions. And finally, we specify a prior distribution over the possible conjunctive conditions. We will discuss each of these three steps below.

To specify the possible requirements from which a conjunctive condition could be composed, our strategy[16] was to subdivide the explanatory variables used to construct the decision tree into as many equal sized groups as possible. The only constraint was that the partition had to maintain the expected relationships between the groups and the outcome of bicycling or not. For example, the variable denoting the number of children was split into three groups: $[0, 1]$, $(1, 2]$, and $(2, \infty)$. For these categories of the number of children, the percentages of individuals in each category that owned a bicycle and actually bicycle to work or to school were approximately 16%, 13% and 0%. Such trends follow our a-priori expectations that the probability of bicycling decreases as one has more children. Sub-dividing the number of children variable into 4 or more categories led to relationships that were deemed to be spurious since they did not match our a-priori beliefs about the relationship between number of children and the probability of bicycling commuting. All together, the possible requirements used to construct the decision tree were as follows (with numbers rounded to two decimal places, or more when necessary):

- Number of Kids: $[0, 1]$, $[2]$, and $[3, \infty)$
- Minimum distance (miles): $[0, 1.17]$, $(1.17, 1.92]$, $(1.92, 3.00]$, $(3.00, 4.37]$, $(4.37, \infty)$
- Average Speed Limit (miles per hour): $[23.01, 25.15]$, $(25.15, 25.78]$, $(25.78, \infty)$
- Median Slope (meters per foot): $[0, 0.01]$, $(0.01, 0.02]$, $(0.02, 0.03]$, $(0.03, 0.04]$, $(0.04, \infty)$
- Proportion of roadway miles along one's shortest path with speed limits < 25 miles per hour: $[0, 0.66]$, $(0.66, 0.83]$, $(0.83, 0.95]$, $(0.95, 0.9984]$, $(0.9984, 0.9986]$, $(0.9986, \infty)$
- Proportion of roadway miles with bicycle lanes: $[0, 0.04]$, $(0.04, 0.11]$, $(0.11, \infty)$
- Proportion of roadway miles with "share the road" markings: $[0, 0.08]$, $(0.08, 0.14]$, $(0.14, \infty)$

[16]Note, we are aware that other strategies could have been used to discretize our variables in order to create requirements for use in the decision tree. Future researchers are free to use any such discretization strategies they prefer and to make a case for such strategies. We chose to follow the procedures of Letham et al. (2015) who manually discretized their variables according to their a-priori beliefs.

These requirements are all binary boolean conditions that are to be read as "*variable* in *range*." For instance, "minimum distance in [0, 1.17]."

Given the possible requirements specified above, the next task is to specify the combinations of these requirements that will be allowed as possible conjunctive conditions in our decision trees. Going along with the notion of non-compensatory rules are at least partially motivated by a desire to minimize cognitive effort, we hypothesize that no individual conjunctive condition will be made up of a large number of requirements. As a result, we specify the maximum number of requirements in a conjunctive condition to be 2. Moreover, since we are trying to estimate a decision tree that (by assumption) is used by our entire population, we limit the possible conjunctive conditions to those conjunctions that apply to a large percentage of the population. In particular, we only consider those conjunctive conditions that apply to (1) 10% or more of those who bicycle or (2) 10% or more of those who did not bicycle. As is done in Letham et al. (2015), we use the FP-growth algorithm implemented by Borgelt (2005) to enumerate these conjunctive conditions.

Now, given the possible conjunctive conditions, our remaining task is to assign a prior probability to each of these conjunctive rule. Because we do not have any prior information about whether one conjunctive condition would be used more than any other, we use a uniform distribution as our prior. In particular, we follow Equation 2.2 of Letham et al. (2015) and place a uniform distribution prior over all conjunctive conditions that (1) have $\mid p_i^m \mid$ requirements and (2) have not already been used in the decision tree.

Lastly, in order to use the methods of Letham et al. (2015), we initially use the decision trees to predict the choice of bicycling or not, for those individuals who own a bicycle. This bicycle focused prediction is performed for two reasons. First, we are being agnostic (initially) about the presence of a more general choice model at the output nodes of the decision tree, and unlike our mode choice models, our trees are not constructed with the relevant variables for predicting all travel modes. Secondly, we focus on the choice of bicycling because the tree is will ultimately be used specifically to determine whether bicycle is considered or not and to what extent bicycling is generally preferred when it is considered. Either way, to make predictions about whether or not an individual bicycles to work or to school, the methods of Letham et al. (2015) require us to specify a prior for the probability that an individual chooses to bicycle. For a fully unknown person, we chose our prior to express maximal ignorance about his/her probability of bicycling. Our prior for the probability that an individual commutes by bicycle is Beta $(1, 1)$: a uniform distribution over the range $(0, 1)$.

For further clarification of how we initially sampled the decision trees, see Letham et al. (2015).

6.5. Results and Discussion

In this subsection, we discuss the results of our empirical application. Specifically, we compare the MNL model with our proposed bayesian model trees in four ways. We first quantitatively compare these two models in terms of in-sample fit. Note that we do not compare the two models in terms of out-of-sample predictive ability simply because our long estimation times and small sample size made both cross-validation and the use of a holdout sample unappealing. Moreover, it is well known that while frequentist estimation techniques such as maximum likelihood are prone to over-fitting, bayesian estimation techniques are much less likely to overfit, and bayesian model selection techniques automatically penalize model complexity (Dawid, 2002; Wagenmakers et al., 2008, Section 3.2). Secondly, we qualitatively compare the two models in terms of their forecasted relationship between public investments in bicycle lanes and expected bicycle mode shares. Thirdly, in an attempt to uncover the model differences that lead to the divergent forecasts, we compare the estimation results of those coefficients that are common to the two models. Finally, we conclude this subsection by discussing the behavioral differences that lead to greater plausibility of our bayesian model tree forecasts as compared to the forecasts of the MNL model.

To begin, we start with the in-sample results. In a frequentist setting, models are commonly compared using log-likelihood ratios. For non-nested models, one might use the Vuong test (a generalization of the likelihood ratio test) to test which of two models is closer in terms of Kullback-Leibler divergence to the true data generating process (Vuong, 1989). In a bayesian setting, the same interpretation can be given to the posterior probability of a model. Simply, the posterior probability of a model is the probability that, out of one's set of models, a given model is closest in terms of Kullback-Leibler divergence to the true data generating process (Walker, 2013). Because we use the scalable rejection sampling algorithm of Braun and Damien (2016), we automatically get an estimate of $P(Y \mid X, m)$ for each of our ten model trees, and because we have the prior of each tree m, we can calculate $P(Y \mid X)$ given our bayesian model tree. Likewise, the scalable rejection sampling algorithm provides an estimate of $P(Y \mid X)$ for our MNL model. Combining

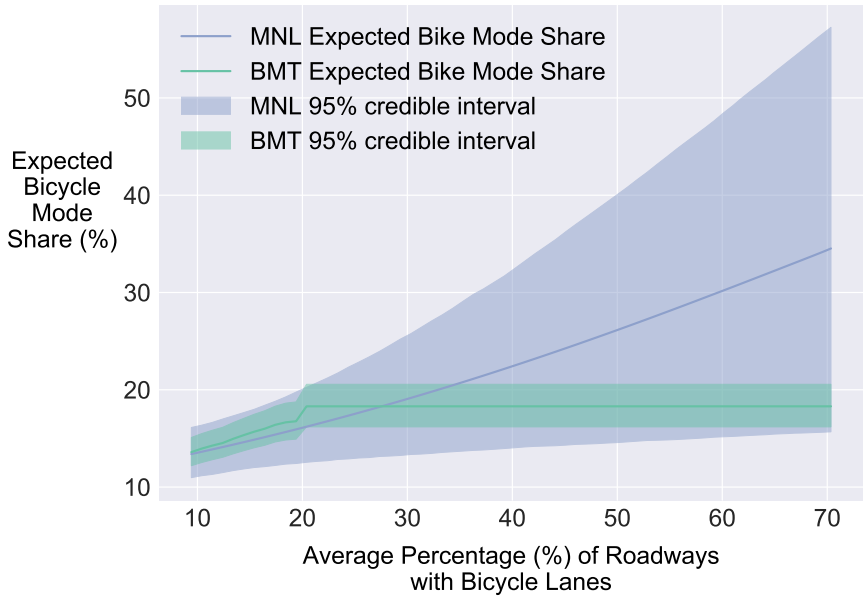

Figure 4: Expected Bicycle Mode Share versus Mean Percentage of Bicycle Lanes

these probabilities with prior probabilities of $\frac{1}{2}$ for each model (to reflect maximal uncertainty about which model is closest to the data generating process), we find that the posterior probability of our bayesian model trees is 99.91% compared to the posterior probability of 0.09% for the MNL model. In other words, based on our data, the bayesian model tree is overwhelmingly more likely to be closer to the true data generating process than the MNL model.

Given that the bayesian model trees are likely to be a better representation of the true data generating process, we now turn to the question of whether this model leads to different policy implications as compared to the MNL model. For our policy application, we considered the effect of increasing the proportion of bicycle lanes for the individuals in our sample. In particular, we raised the proportion of bicycle lanes for each individual in our sample, one percentage point at a time, until each individual's proportion of bicycle lanes was approximately 70% (the maximum value observed in our estimation dataset). After each incremental increase in the proportion of roadways with bicycle lanes, we predicted the average bicycle mode share across the dataset. In Figure 4 we plot the expected bicycle mode shares, as predicted by the bayesian model trees and the MNL model, along with their associated credible intervals[17]. Note that in this plot, we use the acronym "BMT" to refer to the bayesian model trees. Naming aside, Figure 4 shows that the two models lead to very different forecasts. In particular, as one begins to install bicycle lanes, both models show an increase in the expected bicycle mode share, but eventually, the bayesian model trees predicts that the expected bicycle mode share will flatline. In contrast, the MNL model predicts that the expected bicycle mode share will increase continually. A-priori, the predictions of the bayesian model trees appear more plausible than the predictions of the MNL model. In particular, we expect diminishing returns from increasing the proportion of roadways with bicycle lanes since individuals will eventually come to feel safe on public roadways but will still reject the bicycling alternative due to other factors such as time pressures due to childcare obligations, concerns about sweating, etc.

To corroborate our a-priori expectations, we note that in the United States (U.S.) and internationally, solely having many bicycle lanes does not lead to the huge bicycle mode shares predicted by the MNL model. For instance, take the case of Davis, California. Davis has lead the U.S. in the installation of on-street bicycle infrastructure. The first on-street bicycle lanes, the first bicycle traffic signal, and the first 'protected intersection' for bicyclists were all installed in Davis (Caltrans, 2017). Accordingly, out of all cities in the U.S. with populations of greater than 20,000 individuals, Davis has the highest bicycle commuting

[17]Note that credible intervals are the bayesian analog of confidence intervals

mode share. Depending on one's source, Davis' bicycle commuting mode share is 17% - 19% (McKenzie, 2014; McLeod, 2016), a value that precisely matches the predictions of our bayesian model trees. Looking internationally, we can observe countries such as the Netherlands that lead the world in on-street bicycle infrastructure investments. Here, we are quick to note that bicycle infrastructure in the Netherlands is often of much higher quality than in the U.S. Bike lanes in the Netherlands are often 'protected' in the sense that they are physically-separated from motor vehicles (Pucher and Buehler, 2008). Additionally, the entire travel context in the Netherlands is more supportive of bicycling: fuel and automobile-ownership costs are much higher than in the U.S., more downtown areas are designated as automobile-free, local roads are often 'traffic-calmed,' and travel by bicycle is often more direct than by automobile (Pucher and Buehler, 2008). Even with all of these advantages, only about 36% percent of all trips are taken by bicycle in the Netherlands (TNS Opinion & Social, 2015). We are immediately sceptical of any model, such as our MNL model, that predicts a similar level of bicycling based solely on the installation of 'unprotected' bicycle lanes (the only type of bicycle lane present in our study area at the time the data was collected).

Now, to investigate the causes of the differing forecasts shown in Figure 4, we start with the estimated choice model coefficients (β) of the MNL model and our bayesian model trees. A summary of the posterior distribution of the choice model parameters for both the MNL and bayesian model trees is given in Table 1. Briefly, Table 1 shows the posterior mean, the 2.5th percentile, and the 97.5th percentile of the posterior samples for each choice model parameter. To calculate the posterior summary for the bayesian model trees, we calculated a weighted posterior mean and weighted percentiles where the posterior samples of each choice model parameter, for each decision tree, were weighted by the posterior probability of that tree. To make the display manageable, Table 1 only displays the parameters estimated in the MNL model. The parameters that are specific to the model trees, such as the bicycle ASCs that are specific to each output-node ($\text{ASC}_{\text{bike},i}$), are not shown. Now, the main finding from Table 1 is that the estimation results of parameters that are common to both models are very similar. The only parameter whose posterior mean shows large differences between the two models is ASC_{bike}, and this is because the ASC_{bike} in the bayesian model tree plays a different role than it does in the MNL model. Recall that in the bayesian model tree, ASC_{bike} is just the group mean that the output-node specific $\text{ASC}_{\text{bike},i}$ are centered around. Details aside, knowing that the estimated choice models are the largely the same means that we should look towards the decision trees themselves to find out why the two models have such differing forecasts. This line of investigation is pursued below.

Behaviorally, we believe that four qualities of our bayesian model trees lead to its differing forecasts from the MNL model. The first quality we note is that the bicycle lane variable is almost never included in the conjunctive condition that splits the root node. In other words, the bicycle lane variable is almost never in the conditions at the top of our decision trees. In nine of our ten decision trees, there were nodes that filtered out individuals before they could get to an output node that depended on bicycle lanes. Furthermore, the one tree that had a bicycle lane requirement for the first output node actually had a low probability (0.2%) of being the tree that is closest to representing the true data generating process. The behavioral interpretation of this finding is that bicycle lanes are not the most important variable in an individual's decision making process about whether or not to commute by bike. Variables that appear to take precedence over bicycle lanes when deciding whether or not to commute by bike include topography, the number of children an individual has, and the minimum distance between an individual's home and work/school. As a result, the impact of installing bicycle lanes will be moderated by these other variables.

The second quality that we note about our bayesian model trees is that the bicycle lane variable never appeared by itself. In particular, when the proportion of roadways containing bicycle lanes was present in a conjunctive condition, it always appeared alongside another variable. For instance, in seven out of ten decision trees, the bicycle lane variable appeared in the following conjunctive condition: 'proportion of roadways with bicycle lanes is greater than 0.11 and the number of children is 0 or 1.' The posterior probability that the true data generating process was most closely represented by one of these seven trees was over 99%. Such a finding emphasizes that the proportion of roadways containing bicycle lanes is not always important. In particular, if a person has 2 or more children, then bicycle lanes are unlikely to affect whether the individual commutes by bicycle. Presumably, childcare pressures will be a bigger determining factor of those individuals' choice of travel mode.

Table 1: Posterior Summaries of Choice Model Parameters

	MNL			Bayesian Model Trees		
Variables	Mean	2.5%	97.5%	Mean	2.5%	97.5%
Alternative Specific Constants						
Shared Ride: 2	-1.746*	-2.282	-1.231	-1.787*	-2.322	-1.259
Shared Ride: 3+	-1.865*	-2.378	-1.345	-1.909*	-2.453	-1.378
Walk-Transit-Walk	1.070*	0.495	1.623	1.078*	0.497	1.666
Drive-Transit-Walk	-2.183*	-2.969	-1.431	-2.223*	-3.060	-1.418
Walk-Transit-Drive	-2.677*	-3.537	-1.886	-2.734*	-3.602	-1.910
Walk	2.428*	1.826	3.042	2.504*	1.885	3.149
Bike	1.087	-2.573	4.854	0.171	-3.419	3.738
Travel Time, units:0.1min						
All Auto Modes	-1.113*	-1.358	-0.875	-1.129*	-1.383	-0.894
All Transit Modes	-0.374*	-0.479	-0.273	-0.378*	-0.486	-0.276
Travel Cost, units:$						
All Transit Modes	-0.173*	-0.300	-0.053	-0.173*	-0.301	-0.048
Travel Distance, units:mi						
Walk	-1.125*	-1.299	-0.960	-1.151*	-1.337	-0.979
Bike	-0.242	-0.503	0.052	-0.356*	-0.469	-0.245
Systematic Heterogeneity						
Autos per licensed drivers (All Auto Modes)	1.181*	0.784	1.582	1.221*	0.822	1.621
Cross-Bay Tour (Shared Ride 2 & 3+)	-0.517	-1.264	0.167	-0.518	-1.286	0.201
Household Size (Shared Ride 2 & 3+)	0.108	-0.097	0.311	0.127	-0.079	0.330
Number of Kids in Household (Shared Ride 2 & 3+)	0.662*	0.414	0.903	0.634*	0.394	0.892
Spatial Variables						
Minimum Distance units:mi (Bike)	-0.232	-0.822	0.294	_	_	_
Median Slope units:m/ft (Bike)	-1.433	-5.046	2.304	_	_	_
Mean Speed Limit units:mph (Bike)	-0.064	-0.206	0.072	_	_	_
Proportion of Shortest Path Roads slower than 25mph (Bike)	0.484	-0.701	1.791	_	_	_
Proportion of Roadways with Bike Lanes (Bike)	2.218*	0.363	4.048	_	_	_
Proportion of Roadways with Bicycle Chevrons (Bike)	-0.765	-2.647	1.108	_	_	_

Note: * means the equal-tailed 95% credible interval excludes zero.
Additionally, bicycle chevrons are another name for "share the road" arrows.

Third, we point out that decision trees, by their very nature, incorporate a notion of threshold effects. These threshold effects can be seen in our application by the fact that all of our decision trees with posterior probabilities of greater than 0.2% all feature the requirement that the 'proportion of roadways with bicycle lanes is greater than 0.11.' Undoubtedly, the presence of this sharp discontinuity at 0.11 is partly an artifact of our discretization methods. However, as described in Section 5.1.2, even when using soft decision trees that don't implement such "hard" thresholds, the interpretation is that individuals do use hard thresholds, but we are merely uncertain about what those hard thresholds are. Either way, the presence of these threshold effects leads to two features of our forecasts. First, the threshold effects lead to the discrete jump in the expected bicycle mode share when the average percentage of all roadways containing bicycle lanes is about 20%. At this point, almost everyone's proportion of roadways with bicycle lanes rises above 0.11, so all individuals now belong to output nodes with the highest chance of bicycling. Secondly, the threshold effects also cause the flatline in expected bicycle mode share. Because further increases in the proportion of roadways with bicycle lanes do not cause any more changes in the output node's of an individual, there are no more changes in the probability that an individual chooses to bike.

Finally, the last major difference between the forecasts of the bayesian model trees and the MNL model is that as the average percentage of roadways with bicycle lanes increases, the variance in the expected bicycle mode share increases for the MNL model but not for the bayesian model trees. This finding is perhaps best explained mathematically. Whether operating in a frequentist or bayesian setting, the parameters of one's choice model will have an associated probability distribution. In a frequentist setting, this will be the sampling distribution of $\hat{\beta}$ and in a bayesian setting, this will be the posterior distribution of β. For ease of exposition, we will continue our discussion from a bayesian perspective, but our explanation is equally valid from a frequentist perspective. Since the MNL model multiplies the proportion of roadways with bicycle lanes ($X_{\text{bike-lanes}}$) by $\beta_{\text{bike-lanes}}$, we can calculate the variance of the product as $\text{Var}\left[X_{\text{bike-lanes}}\beta_{\text{bike-lanes}}\right] = X^2_{\text{bike-lanes}}\text{Var}\left[\beta_{\text{bike-lanes}}\right]$. This means that as $X_{\text{bike-lanes}}$ increases, the variance of $X_{\text{bike-lanes}}\beta_{\text{bike-lanes}}$ increases. Because the MNL's probability that an individual commutes by bicycle is dependent on $X_{\text{bike-lanes}}\beta_{\text{bike-lanes}}$, the increase in the variance of $X_{\text{bike-lanes}}\beta_{\text{bike-lanes}}$ leads to an increase in the variance of the probability that an individual commutes by bicycle. Aggregated over all individuals, the increases in the variances of the bicycle probabilities lead to an increase in the variance of the expected bicycle mode share.

In contrast to the process described above, changing the value of $X_{\text{bike-lanes}}$ when forecasting with the bayesian model trees only changes what output node one falls into for a given decision tree. The structure of the trees remains unchanged, the posterior probabilities across the trees remains unchanged, and the variance of the $\text{ASC}_{\text{bike},i}$ remain mostly constant across output nodes that have bicycle available as an alternative. As a result, the variance of the expected bicycle mode share remains mostly constant as $X_{\text{bike-lanes}}$ is increased for the various individuals in our dataset. Behaviorally, these differences in forecast uncertainty can be attributed to the fact that when using a compensatory model, one is uncertain about the extent to which the proportion of roadways with bicycle lanes compensates for the other variables that affect one's probability of bicycling. In other words, one is uncertain about the value of $\hat{\beta}_{\text{bike-lanes}}$ or $\beta_{\text{bike-lanes}}$. Since $X_{\text{bike-lanes}}$, only appears in the non-compensatory portion of our bayesian model trees (i.e. in the decision trees as opposed to the choice model), we are always certain that the bike lane proportion does not compensate for other variables. I.e. our bayesian model trees are based on the assumption that $\beta_{\text{bike-lanes}} = 0$. This constant level of uncertainty with respect to the compensatory nature of $X_{\text{bike-lanes}}$ leads directly to the constant level of forecast uncertainty for the bayesian model trees in Figure 4.

7. Conclusion

In this paper, we have made three contributions to the literature. First, we have provided a micro-economic framework for interpreting a class of machine learning models known as decision trees. In particular, we reviewed how decision trees are used and estimated (Section 2), we showed how decision trees represent a non-compensatory decision protocol known as disjunctions-of-conjunctions (Section 3), and we discussed how existing decision tree variants can account for economic considerations that discrete choice modelers are likely to have (Section 5).

Secondly, we contributed to both the discrete choice and decision tree literatures by formulating and estimating the first bayesian model tree: a semi-compensatory, two-stage model of human decision making.

Our model uses a non-compensatory, disjunctions-of-conjunctions protocol to determine one's choice set, and conditional on a given choice set, it uses a compensatory discrete choice model (e.g. an MNL model) to make a final selection if more than one alternative is available. Beyond one's choice set, our bayesian model tree allows the non-compensatory rules to influence one's preferences, as embodied in the choice model parameters, and our model allows for quantification of one's uncertainty over which set of disjunctions-of-conjunctions are actually being used in a population. To the best of our knowledge, this is the first time a bayesian model tree has ever been proposed and estimated.

Finally, beyond the mere proposition of the bayesian model tree, our paper carried out an empirical application of this model. We made three major findings. First, our proposed bayesian model tree is more than 1,000-times more likely ($\frac{99.01}{0.09} \approx 1,100$) to be closer to our application's true data-generating process than the MNL model. Second, our bayesian model trees provide forecasts that are consistent with observed bicycle mode shares in areas with abundant bike lanes such as Davis, CA and the Netherlands. In comparison, the forecasts of the MNL model were overly optimistic. Third, our bayesian model trees provide insights that are qualitatively different than the MNL model. Specifically, our bayesian model trees suggest that (1) investments in on-street bicycle lanes will eventually suffer from diminishing returns and (2) that factors such as travel distance, child-related pressures, and topography may all prevent individuals from bicycling even if there are many bicycle lanes. These insights are missing from the more traditional (and compensatory) MNL model.

Moving forward, we note that in the decades after McFadden revealed the economic implications of the conditional logit model, discrete choice modelers moved swiftly to create needed extensions. As a result, we can now avoid many of the troubling assumptions and properties of the conditional logit model, leading to more accurate analyses and more sensible policy implications. Analogously, by linking decision trees to economics, our paper brings decision trees to a similar, infantile stage. As noted in Section 5.2, there remain a number of economic concerns (or more specifically, combinations of concerns) that must be confronted before decision trees will be maximally useful in policy settings. By detailing the microeconomic implications of decision trees, we aim to draw the attention of choice modelers. Hopefully, our paper will encourage the use and extension of current decision tree methodologies, thereby increasing the accuracy and usefulness of such models for policy analyses.

8. Acknowledgements

We thank Paul Waddell, the MacArthur Foundation, and UCCONNECT's Dissertation Grant for funding this research. We also thank Feras El Zarwi, Sreeta Gorripaty, and Madeleine Sheehan for their helpful discussions in the beginning stages of this research endeavor.

References

References

Allenby, G.M., Rossi, P.E., 1999. Marketing models of consumer heterogeneity. Journal of Econometrics 89, 57–78.

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., Rudin, C., 2017. Learning certifiably optimal rule lists, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM. pp. 35–44.

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

Arentze, T.A., Timmermans, H.J., 2007. Parametric action decision trees: Incorporating continuous attribute variables into rule-based models of discrete choice. Transportation Research Part B: Methodological 41, 772–783.

Bafumi, J., Gelman, A., 2006. Fitting multilevel models when predictors and group effects correlate. Technical Report. URL: `http://dx.doi.org/10.2139/ssrn.1010095`.

Bajari, P., Nekipelov, D., Ryan, S.P., Yang, M., 2015a. Demand estimation with machine learning and model combination. Technical Report. National Bureau of Economic Research.

Bajari, P., Nekipelov, D., Ryan, S.P., Yang, M., 2015b. Machine learning methods for demand estimation. The American Economic Review 105, 481–485.

Barros, R.C., Basgalupp, M.P., De Carvalho, A.C., Freitas, A.A., 2012. A survey of evolutionary algorithms for decision-tree induction. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42, 291–312.

Ben-Akiva, M.E., 1973. Structure of passenger travel demand models. Ph.D. Thesis, Dept. of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA.

Bhat, C.R., 1998. Accommodating variations in responsiveness to level-of-service measures in travel mode choice modeling. Transportation Research Part A: Policy and Practice 32, 495–507.

Bhat, C.R., 2015. A comprehensive dwelling unit choice model accommodating psychological constructs within a search strategy for consideration set formation. Transportation Research Part B: Methodological 79, 161–188.

Bishop, C.M., 2006. Pattern recognition and machine learning. Springer.

Bishop, C.M., Svensén, M., 2003. Bayesian Hierarchical Mixtures of Experts, in: Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc.. pp. 57–64.

Borgelt, C., 2005. An implementation of the FP-growth algorithm, in: OSDM'05 Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations, ACM, New York. pp. 1–5.

Brathwaite, T., 2018. The Holy Trinity: Blending Statistics, Machine Learning and Discrete Choice with Applications to Strategic Bicycle Planning. Ph.D. Thesis, Dept. of Civil Engineering, University of California, Berkeley, Berkeley, CA.

Braun, M., Damien, P., 2016. Scalable rejection sampling for bayesian hierarchical models. Marketing Science 35, 427–444.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Bronner, A., 1982. Decision styles in transport mode choice. Journal of Economic Psychology 2, 81–101.

Bühlmann, P., Hothorn, T., 2007. Boosting Algorithms: Regularization, Prediction and Model Fitting (with discussion). Statistical Science 22, 477–505.

California Department of Transportation, 2013. 2010-2012 California Household Travel Survey Final Report. Technical Report. URL: http://www.dot.ca.gov/hq/tsip/FinalReport.pdf.

Caltrans, 2017. Toward An Active California: State Bicycle + Pedestrian Plan. Technical Report. California State Department of Transportation. URL: http://www.dot.ca.gov/activecalifornia/documents/Hi-Res_Final_ActiveCA.pdf.

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

Cantillo, V., Heydecker, B., de Dios Ortúzar, J., 2006. A discrete choice model incorporating thresholds for perception in attribute values. Transportation Research Part B: Methodological 40, 807–825.

Cascetta, E., Papola, A., 2009. Dominance among alternatives in random utility models. Transportation Research Part A: Policy and Practice 43, 170–179.

Chan, K., Loh, W., 2004. LOTUS: An Algorithm for Building Accurate and Comprehensible Logistic Regression Trees. Journal of Computational and Graphical Statistics 13, 826–852.

Chipman, H.A., George, E.I., McCulloch, R.E., 1998. Bayesian CART model search (with discussion). Journal of the American Statistical Association 93, 935–960.

Chorus, C., 2012. Random regret minimization: an overview of model properties and empirical evidence. Transport Reviews 32, 75–92.

Cleland, B.S., Walton, D., 2004. Why don't people walk and cycle. Technical Report 528007. Central Laboratories. New Zealand. URL: http://can.org.nz/system/files/Why%20dont%20people%20walk%20and%20cycle.pdf.

Commission, S.F.M.T., Brinckerhoff, P., 2012. Travel Model Development: Calibration and Validation. Technical Report. San Francisco Metropolitan Transportaton Commission. URL: http://mtcgis.mtc.ca.gov/foswiki/pub/Main/Documents/2012_05_18_RELEASE_DRAFT_Calibration_and_Validation.pdf.

Conlisk, J., 1996. Why bounded rationality? Journal of Economic Literature 34, 669–700.

Coombs, C.H., 1951. Mathematical models in psychological scaling. Journal of the American Statistical Association 46, 480–489.

Cox, D.R., 1966. Some procedures connected with the logistic qualitative response curve, in: David, F.N. (Ed.), Research Papers in Statistics, John Wiley & Sons, New York. pp. 55–71.

Das, M., Bhattacharya, S., 2017. Transdimensional Transformation based Markov Chain Monte Carlo. arXiv preprint arXiv:1403.5207v5 .

Dawes, R.M., 1964. Social selection based on multidimensional criteria. The Journal of Abnormal and Social Psychology 68, 104.

Dawid, A.P., 2002. Comment on bayesian measures of complexity and fit. Journal of the Royal Statistics Society: Series B (Methodological) 64, 583–639.

Denison, D.G., Mallick, B.K., Smith, A.F., 1998. A bayesian cart algorithm. Biometrika 85, 363–377.

Einav, L., Levin, J., 2014. The data revolution and economic analysis. Innovation Policy and the Economy 14, 1–24.

Elrod, T., Johnson, R.D., White, J., 2004. A new integrated model of noncompensatory and compensatory decision strategies. Organizational Behavior and Human Decision Processes 95, 1–19.

Esposito, F., Malerba, D., Semeraro, G., Kay, J., 1997. A comparative analysis of methods for pruning decision trees. IEEE transactions on pattern analysis and machine intelligence 19, 476–491.

Fan, Y., Sisson, S.A., 2011. Reversible Jump MCMC, in: Brooks, S., Gelman, A., Jones, G.L., Meng, X.L. (Eds.), Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC, New York, pp. 67–91.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research 15, 3133–3181.

Foerster, J.F., 1979. Mode choice decision process models: a comparison of compensatory and non-compensatory structures. Transportation Research Part A: General 13, 17–28.

Friedman, J., Hastie, T., Tibshirani, R., 2008. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2 ed., Springer Series in Statistics.

Gelman, A., 1992. Iterative and non-iterative simulation algorithms. Computing Science and Statistics 24, 433–438.

Gelman, A., 2006. Multilevel (hierarchical) modeling: what it can and cannot do. Technometrics 48, 432–435.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2014. Bayesian data analysis. volume 3. CRC press Boca Raton, FL.

Gigerenzer, G., Goldstein, D.G., 1996. Reasoning the fast and frugal way: models of bounded rationality. Psychological review 103, 650.

Gilbride, T.J., Allenby, G.M., 2004. A choice model with conjunctive, disjunctive, and compensatory screening rules. Marketing Science 23, 391–406.

Goldsmith, S.A., 1992. Reasons why bicycling and walking are and are not being used more extensively as travel modes. 1, Federal Highway Administration. URL: `http://safety.fhwa.dot.gov/ped_bike/docs/case1.pdf`.

González, S., Herrera, F., García, S., 2015. Monotonic Random Forest with an Ensemble Pruning Mechanism based on the Degree of Monotonicity. New Generation Computing 33, 367–388.

Gordon, L., Olshen, R.A., 1980. Consistent nonparametric regression from recursive partitioning schemes. Journal of Multivariate Analysis 10, 611–627.

Gordon, L., Olshen, R.A., 1984. Almost surely consistent nonparametric regression from recursive partitioning schemes. Journal of Multivariate Analysis 15, 147–163.

Green, P.E., Krieger, A.M., Bansal, P., 1988. Completely unacceptable levels in conjoint analysis: A cautionary note. Journal of Marketing Research , 293–300.

Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732.

Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50, 1029–1054.

Hauser, J.R., Toubia, O., Evgeniou, T., Befurt, R., Dzyabura, D., 2010. Disjunctions of conjunctions, cognitive simplicity, and consideration sets. Journal of Marketing Research 47, 485–496.

Hess, S., Stathopoulos, A., Daly, A., 2012. Allowing for heterogeneous decision rules in discrete choice models: an approach and four case studies. Transportation 39, 565–591.

Hesterberg, T., 1995. Weighted average importance sampling and defensive mixture distributions. Technometrics 37, 185–194.

Hu, Q., Che, X., Zhang, L., Zhang, D., Guo, M., Yu, D., 2012. Rank Entropy Based Decision Trees for Monotonic Classification. IEEE Transactions on Knowledge and Data Engineering 24, 2052–2064.

Huber, J., Klein, N.M., 1991. Adapting cutoffs to the choice environment: the effects of attribute correlation and reliability. Journal of Consumer Research 18, 346–357.

Ittner, A., Schlosser, M., 1996. Non-linear decision trees-ndt, in: ICML, Citeseer. pp. 252–257.

Jang, J.S., 1994. Structure determination in fuzzy modeling: a fuzzy cart approach, in: Fuzzy Systems, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the Third IEEE Conference on, IEEE. pp. 480–485.

Jedidi, K., Kohli, R., 2005. Probabilistic subset-conjunctive models for heterogeneous consumers. Journal of Marketing Research 42, 483–494.

Jordan, M.I., Jacobs, R.A., 1994. Hierarchical Mixtures of Experts. Neural Computation 6, 181–214.

Kamakura, W.A., Kim, B.D., Lee, J., 1996. Modeling preference and structural heterogeneity in consumer choice. Marketing Science 15, 152–172.

Kaplan, S., Bekhor, S., Shiftan, Y., 2009. Two-stage model for jointly revealing determinants of noncompensatory conjunctive choice set formation and compensatory choice. Transportation Research Record: Journal of the Transportation Research Board , 153–163.

Kaplan, S., Prato, C.G., 2012. Closing the gap between behavior and models in route choice: The role of spatiotemporal constraints and latent traits in choice set formation. Transportation Research Part F: traffic psychology and behaviour 15, 9–24.

Kaplan, S., Shiftan, Y., Bekhor, S., 2012. Development and estimation of a semi-compensatory model with a flexible error structure. Transportation Research Part B: Methodological 46, 291–304.

Kim, J.H., Kim, M., 2011. Two-Stage Multinomial Logit Model. Expert Systems with Applications 38, 6439–6446.

Kim, M., 2009. Two-Stage Logistic Regression Model. Expert Systems with Applications 36, 6727–6734.

Kindermann, J., Paass, G., 1998. Model switching for bayesian classification trees with soft splits. Principles of Data Mining and Knowledge Discovery , 148–157.

Kohli, R., Jedidi, K., 2007. Representation and inference of lexicographic preference models and their variants. Marketing Science 26, 380–399.

Kumar, G.K., Viswanath, P., Rao, A.A., 2016. Ensemble of randomized soft decision trees for robust classification. Sadhana 41, 273–282.

Landwehr, N., Hall, M., Frank, E., 2005. Logistic Model Trees. Machine Learning 59, 161–205.

Lemon, S.C., Roy, J., Clark, M.A., Friedmann, P.D., Rakowski, W., 2003. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. Annals of behavioral medicine 26, 172–181.

Leong, W., Hensher, D.A., 2012. Embedding decision heuristics in discrete choice models: A review. Transport Reviews 32, 313–331.

Letham, B., Rudin, C., McCormick, T.H., Madigan, D., 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. The Annals of Applied Statistics 9, 1350–1371.

Loh, W.Y., 2011. Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1, 14–23.

Loh, W.Y., 2014. Fifty years of classification and regression trees. International Statistical Review 82, 329–348.

Lomax, S., Vadera, S., 2013. A survey of cost-sensitive decision tree induction algorithms. ACM Computing Surveys (CSUR) 45, 16.

Mahmoud, M.S., Weiss, A., Habib, K.N., 2016. Myopic choice or rational decision making? An investigation into mode choice preference structures in competitive modal arrangements in a multimodal urban area, the City of Toronto. Canadian Journal of Civil Engineering 43, 420–428.

Manski, C.F., 1977. The structure of random utility models. Theory and decision 8, 229–254.

Manski, C.F., 2001. Daniel mcfadden and the econometric analysis of discrete choice. The Scandinavian Journal of Economics 103, 217–229.

Marsala, C., Petturiti, D., 2015. Rank discrimination measures for enforcing monotonicity in decision tree induction. Information Sciences 291, 143–171.

Marschak, J., 1960. Binary-choice constraints and random utility indicators, in: Stanford Symposium on Mathematical Methods in the Social Sciences, Stanford University Press, Stanford, California.

Martínez, F., Aguila, F., Hurtubia, R., 2009. The constrained multinomial logit: A semi-compensatory choice model. Transportation Research Part B: Methodological 43, 365–377.

McFadden, D., 2001. Economic choices. The American Economic Review 91, 351–378.

McKenzie, B., 2014. Modes less traveled—Bicycling and walking to work in the United States: 2008-2012. Technical Report. United States Census Bureau. Suitland, MD.

McLeod, K., 2016. Where We Ride: Analysis of bicycle commuting in American Cities. Technical Report. League of American Bicyclists.

Meila, M., Jordan, M.I., 2000. Learning with mixtures of trees. Journal of Machine Learning Research 1, 1–48.

Mingers, J., 1989. An empirical comparison of pruning methods for decision tree induction. Machine Learning 4, 227–243.

Minka, T.P., 2002. Bayesian model averaging is not model combination. URL: `https://pdfs.semanticscholar.org/e30a/1d14dd097608583d6c200b43fada35dac444.pdf`.

Mohammadi, A., Kaptein, M., 2016. Comment on "Efficient Metropolis–Hastings proposal mechanisms for Bayesian regression tree models". Bayesian analysis 11, 938–940.

Murphy, K.P., 2012. Machine learning: a probabilistic perspective. MIT press.

Murthy, S.K., 1998. Automatic construction of decision trees from data: A multi-disciplinary survey. Data mining and knowledge discovery 2, 345–389.

Murthy, S.K., Kasif, S., Salzberg, S., 1994. A system for induction of oblique decision trees. Journal of Artificial Intelligence Research 2, 1–32.

Newton, M.A., Raftery, A.E., 1994. Approximate bayesian inference with the weighted likelihood bootstrap. Journal of the Royal Statistical Society. Series B (Methodological) , 3–48.

Olaru, C., Wehenkel, L., 2003. A complete fuzzy decision tree technique. Fuzzy sets and systems 138, 221–254.

Papaspiliopoulos, O., Robers, G.O., Sköld, M., 2007. A General Framework for the Parametrization of Hierarchical Models. Statistical Science 22, 59–73.

Pei, S., Hu, Q., Chen, C., 2016. Multivariate decision trees with monotonicity constraints. Knowledge-Based Systems 112, 14–25.

Potharst, R., Feelders, A., 2002. Classification trees for problems with monotonicity constraints. SIGKDD Explorations Newsletter 4, 1–10.

Pratola, M.T., 2016. Efficient metropolis–hastings proposal mechanisms for bayesian regression tree models. Bayesian Analysis 11, 885–911.

Pucher, J., Buehler, R., 2008. Making cycling irresistible: lessons from the netherlands, denmark and germany. Transport reviews 28, 495–528.

Quinlan, J.R., 1990. Probabilistic decision trees. Machine Learning: an artificial intelligence approach 3, 140–152.

Rivest, R.L., 1987. Learning decision lists. Machine Learning 2, 229–246.

Rokach, L., 2010. Ensemble-based classifiers. Artificial Intelligence Review 33, 1–39.

Rokach, L., Maimon, O., 2005. Top-down induction of decision trees classifiers-a survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 35, 476–487.

Rokach, L., Maimon, O., 2014. Data mining with decision trees: theory and applications. World Scientific.

Rubin, D.B., 1981. The bayesian bootstrap. The Annals of Statistics 9, 130–134.

Ruggieri, S., 2017. Enumerating distinct decision trees, in: Precup, D., Teh, Y.W. (Eds.), Proceedings of the 34th International Conference on Machine Learning, PMLR, International Convention Centre, Sydney, Australia. pp. 2960–2968. URL: http://proceedings.mlr.press/v70/ruggieri17a.html.

Rusch, T., Zeilis, A., 2013. Gaining insight with recursive partitioning of generalized linear models. Journal of Statistical Computation and Simulation 83, 1301–1315.

Seyedhosseini, M., Tasdizen, T., 2015. Disjunctive normal random forests. Pattern Recognition 48, 976–983.

Simon, H.A., 1955. A behavioral model of rational choice. The Quarterly Journal of Economics 69, 99–118.

Sisson, S.A., 2005. Transdimensional markov chains. Journal of the American Statistical Association 100, 1077–1089.

Steinberg, D., Cardell, N.S., 1998. The hybrid CART-Logit model in classification and data mining. Salford Systems White Paper URL: http://media.salford-systems.com/pdf/the-hybrid-cart-logit-model-in-classification-and-data%20mining-1998.pdf.

Strobl, C., Malley, J., Tutz, G., 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychological methods 14, 323.

Stüttgen, P., Boatwright, P., Monroe, R.T., 2012. A satisficing choice model. Marketing Science 31, 878–899.

Su, X., 2007. Tree-based model checking for logistic regression. Statistics in medicine 26, 2154–2169.

Swait, J., 1984. Probabilistic choice set formation in transportation demand models. Unpublished Ph.D. Thesis, Dept. of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA.

Swait, J., 2001a. Choice set generation within the generalized extreme value family of discrete choice models. Transportation Research Part B: Methodological 35, 643–666. URL: http://www.sciencedirect.com/science/article/pii/S0191261500000291.

Swait, J., 2001b. A non-compensatory choice model incorporating attribute cutoffs. Transportation Research Part B: Methodological 35, 903–928.

Swait, J., 2009. Choice models based on mixed discrete/continuous pdfs. Transportation Research Part B: Methodological 43, 766–783.

Swait, J., Adamowicz, W., Hanemann, M., Diederich, A., Krosnick, J., Layton, D., Provencher, W., Schkade, D., Tourangeau, R., 2002. Context Dependence and Aggregation in Disaggregate Choice Analysis. Marketing Letters 13, 195–205.

Swait, J., Ben-Akiva, M., 1987a. Empirical test of a constrained choice discrete model: mode choice in sao paulo, brazil. Transportation Research Part B: Methodological 21, 103–115.

Swait, J., Ben-Akiva, M., 1987b. Incorporating random constraints in discrete models of choice set generation. Transportation Research Part B: Methodological 21, 91–102.

Tibshirani, R., Knight, K., 1999. Model search by bootstrap bumping. Journal of Computational and Graphical Statistics 8, 671–686.

TNS Opinion & Social, 2015. Quality of transport. Technical Report. Directorate-General for Mobility and Transport (European Commission). URL: https://publications.europa.eu/en/publication-detail/-/publication/fb18e0ed-52d0-41cd-be6c-61aad310fb53/language-en.

Toth, D., Eltinge, J.L., 2011. Building consistent regression trees from complex sample data. Journal of the American Statistical Association 106, 1626–1636.

Train, K., 2009. Discrete Choice Methods With Simulation. 2 ed., Cambridge University Press, New York, NY, USA.

Truong, T.D., Wiktor, L., Boxall, P.C., 2015. Modeling non-compensatory preferences in environmental valuation. Resource and Energy Economics 39, 89–107.

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

Tversky, A., Kahneman, D., 1986. Rational choice and the framing of decisions. The Journal of Business 59, S251S278.

University of California at Berkeley, 2000. 10.11.00 - daniel l. mcfadden wins nobel prize in economics. URL: http://www.berkeley.edu/news/features/2000/nobel/.

Velikova, M., Daniels, H., 2004. Decision trees for monotone price models. Computational Management Science 1, 231–244.

Vij, A., Carrel, A., Walker, J.L., 2013. Incorporating the influence of latent modal preferences on travel mode choice behavior. Transportation Research Part A: General 54, 164–178.

Vij, A., Walker, J.L., 2014. Preference endogeneity in discrete choice models. Transportation Research Part B: Methodological 64, 90–105. URL: http://www.sciencedirect.com/science/article/pii/S0191261514000344, doi:10.1016/j.trb.2014.02.008.

Villandré, L., Rich, B., Ciampi, A., 2012. Soft classification trees. Communications in Statistics-Theory and Methods 41, 3244–3258.

Vuong, Q.H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica: Journal of the Econometric Society 57, 307–333.

Wagenmakers, E.J., Lee, M., Lodewyckx, T., Iverson, G.J., 2008. Bayesian versus frequentist inference, in: Hoijtink, H., Klugkist, I., Boelen, P. (Eds.), Bayesian Evaluation of Informative Hypotheses. Springer, pp. 181–207.

Wainer, J., 2016. Comparison of 14 different families of classification algorithms on 115 binary datasets. arXiv preprint arXiv:1606.00930 .

Walker, S.G., 2013. Bayesian inference with misspecified models. Journal of Statistical Planning and Inference 143, 1621–1633.

Wu, Y., Tjelmeland, H., West, M., 2007. Bayesian CART: Prior Specification and Posterior Simulation. Journal of Computational and Graphical Statistics 16, 44–66.

Yildiz, O.T., Irsoy, O., Alpaydin, E., 2016. Bagging Soft Decision Trees, in: Holzinger, A. (Ed.), Machine Learning for Health Informatics. Springer, pp. 25–36.

Young, W., 1984. A non-tradeoff decision making model of residential location choice. Transportation Research Part A: General 18, 1–11.

Yu, P.L., Lee, P.H., Cheung, S., Lau, E.Y., Mok, D.S., Hui, Harry, C., 2016. Logit tree models for discrete choice data with application to advice-seeking preferences among Chinese Christians. Computational Statistics 31, 799–827.

Yuksel, S.E., Wilson, J.N., Gader, P.D., 2012. Twenty Years of Mixture of Experts. IEEE Transactions on Neural Networks and Learning Systems 23, 1177–1193.

Zeilis, A., Hothorn, T., Hornik, K., 2008. Model-Based Recursive Partitioning. Journal of Computational and Graphical Statistics 17, 492–514.

Zhu, W., Timmermans, H., 2010. Cognitive process model of individual choice behaviour incorporating principles of bounded rationality and heterogeneous decision heuristics. Environment and Planning B: Planning and Design 37, 59–74.

Zolfaghari, A., Sivakumar, A., Polak, J., 2013. Simplified probabilistic choice set formation models in a residential location choice context. Journal of choice modelling 9, 3–13.