Chapter 1: Introduction

Timothy Brathwaite

At the age of fourteen, I[1] began commuting to school by bicycle. Unfortunately, at age sixteen, I was in a collision with a motor vehicle: I was "doored." In laymen's terms, while commuting with my best friend, a driver suddenly opened their car door into the lane we were traveling in, striking both of us and causing us to fly off our bicycles into the roadway. After crashing to the ground, my friend was able to pull me out of the roadway just before I was crushed by an oncoming van, thus saving my life. Unfortunately, my bicycle was not so lucky, and it laid mangled beneath the van. In the wake of the destruction of my bicycle and of my naive belief in the safety of modern streets, I left the scene of the crash with a bevy of questions. Foremost among these was the question: **"what makes people travel by bicycle?"** My thought, then and now, was that if we can figure out what makes people travel by bicycle, then we can design our cities and regions to maximally divert individuals from private automobiles to bicycles. Since private automobiles are the largest threat to cyclists, such a diversion would increase bicycle safety and reduce the chance of anyone suffering a similar or worse fate than I.

After the crash, I repeatedly searched the internet for what makes people cycle, and as a result, bicycling provided my first introduction to the fields of transportation planning and engineering. Eleven years later, this dissertation's introduction necessarily begins with the same motivating question. While the methods presented in the interior chapters of this dissertation are widely useful beyond the original setting of travel mode choice, to start with the methods would be putting the proverbial cart-before-the-horse. In other words, to understand why I made the methodological advances that I made, one must understand the problems that these methods were meant to solve.

Accordingly, this chapter discusses the substantive societal problem and historical practices that motivate this dissertation's methodological contributions. Focusing on the prediction of bicycle mode shares in the United States, I describe how my work is motivated by four particular issues with previous travel demand models. First, I detail issues with the input data to travel mode choice models that are used to predict bicycle mode shares. Then, I move on to highlight two overlooked, empirical observations of individuals making commute mode choices, paying special attention to the ways that traditional travel demand models might be enhanced by accounting for these observations. Finally, I comment on the technical requirements that are required for strategic bicycle planning efforts, focusing on the inferential gulf between these requirements and the intrinsic properties of traditional travel demand models.

1 Motivation

Across all levels of government in the United States (U.S.), transportation and planning agencies have prioritized encouraging bicycle use. For instance, as stated by the Federal Highway Administration, "it is federal transportation policy to promote the increased use and safety of bicycling and walking as transportation modes" (Federal Highway Administration, 2003). Locally, the California State Department of Transportation (Caltrans) shares the Federal governments cycling goals. As stated in Caltrans' 2002 California Blueprint for Bicycling and Walking, California had statewide goals of a 50% increase in the 0.8% bicycle commute mode share from the year 2000 to a 1.2% bicycle mode share by the year 2010 (California Department of Transportation, 2002). However, as of 2014–the most recent year for which U.S. Census data is available–California's bicycle commute mode share was 1.1% and the state's mode share goals for bicycling had still

[1] Note, Chapters 1, 2, and 6 use the pronoun "I" whereas Chapters 3, 4, and 5 use the pronoun "we." This discrepancy exists because Chapters 3-5 are based on work that was performed with collaborators, but Chapters 1, 2, and 6 represent writings that were produced without the vetting or consensus of collaborators. I thought it best not to wrongly associate any ideas that my collaborators might disagree with, so I have chosen to use "I" when I am not referring to joint work or opinions.

not been met (U.S. Census Bureau, 2014). Similarly, at the municipal level, San Francisco set a goal in 2010 to achieve a 20% bicycle mode share by 2020 (SFMTA, 2012). However, given the 2014 citywide, bicycle commute mode share of 3.8% (U.S. Census Bureau, 2014) San Francisco does not appear to be on track to meet its mode share goals either. While unfortunate, this pattern is not unique to California. There are agencies all across the U.S. and at all levels of U.S. government that are interested yet unsuccessful in meeting their bicycle commute mode share goals.

In analyzing how city and state transportation agencies might make progress towards their stated bicycle commuting goals, it is apparent that these agencies primarily affect the travel behavior and traveling conditions of their constituents through implemented infrastructure projects. To fully support their bicycling agenda, it is therefore necessary for government agencies to judge the extent to which each possible project will increase bicycle usage. However, despite this clear need, current travel demand models are often unable to assess the sensitivity of bicycle mode shares to key variables that agencies control (e.g., the layout of a city's bicycle network or the motor vehicle speeds on city roadways). Moreover, even if an agency's mode choice model is sensitive to the variables of interest, the agency's model should accurately predict the probability that an individual travels by bicycle so that the agency can be as best informed about the potential benefits of each project. In sum, U.S. transportation agencies need (1) mode choice models that are sensitive to variables of interest such as bicycle infrastructure and roadway conditions (vehicle travel speeds, roadway slopes, etc.) and (2) mode choice models that are as accurate as possible.

2 Research Objectives

To address the need for accurate and policy sensitive bicycle demand models, the research objectives of this dissertation are to:

1. incorporate roadway-level data (such as the location of bicycle lanes and roadway speed limits) directly into existing mode choice models

2. mitigate the negative impacts of relative class imbalance (i.e low rates of bicycling relative to other travel modes) on mode choice models

3. account for "alternative" (i.e. non-compensatory) decision protocols that may lead travelers to exclude bicycles from consideration as a travel mode

4. investigate how travel demand models might be altered or enriched in order to measure causal as opposed to associational relationships.

Overall, the four aforementioned research objectives serve three goals. First, the objectives aim to increase the practical usefulness of mode choice models by directly incorporating roadway-level variables into them. Such inclusions will allow mode choice models to answer questions about infrastructure investment decisions. Secondly, by accounting for overlooked behavioral features that characterize the bicycle commuting decision, the aforementioned research objectives aim to increase the accuracy of bicycle demand models. In particular, I reduce the negative impacts of low-cycling rates on mode choice models by allowing for differing willingnesses to adopt different travel modes, and I incorporate non-compensatory protocols into mode choice models to more realistically model both choice set construction and preference construction. Lastly, my research objectives aim to improve the usefulness of mode choice models for transportation planning by investigating how such models can be used to estimate causal as opposed to merely associational relationships. Such a change would allow practitioners to make more credible inferences about which intervention or set of interventions will lead to the greatest benefits for a given level of investment.

As noted earlier, while this dissertation's substantive motivation is the improving bicycle demand models in particular, the methods developed in the course of this research apply more broadly to the fields of choice modeling, statistics, machine learning, and causal inference. A clear example of this is the class-imbalance research mentioned above in point 2 and in Section 4. While useful in the context of mode choice models, choice modeling and statistics both benefit from our creation of flexible probability functions whose rates of increase and decrease from 50% (i.e. adoption and abandonment rates) can be estimated to fit one's data as best as possible. Likewise, this dissertation's research on non-compensatory decision making benefits choice modelling in general by more flexibly relaxing the assumption of rational consideration set formation than in previous choice models. At the same time, I contribute to the field of machine learning by creating a new decision tree variant to represent these non-compensatory rules, filling a missing rung in the hierarchy of advanced decision tree methods.

3 Omitted roadway-level variables

The current state of most mode choice models is that they focus on socio-demographic attributes of the decision makers and "level-of-service" variables such as travel times (e.g. in-vehicle travel time, waiting time, access time, egress time), travel cost, and travel distances of various modes (Singleton and Clifton, 2013). As noted above, variables that describe individual roadway segments and are of particular interest to policy makers and individual travelers, such as the presence of bicycle lanes or the speed limit on particular roadways, do not often appear in mode choice models. Such an exclusion affects transportation engineering in two major ways. First, the usefulness of current mode choice models for addressing policy questions, such as whether one proposed bicycle infrastructure plan will have a greater effect on bicycle demand than another possible plan, is greatly reduced when the relevant variables being altered do not even appear in one's model. Secondly, the ability of current models to accurately represent the choice processes of individuals may be reduced by omitted variable bias and the spatial autocorrelations that may occur due to the omission of these roadway-level variables (Goetzke, 2003). As noted in the Section 1, this omission indicates a clear need for methods of integrating roadway segment information into mode choice models.

Thus far, roadway-level variables have been incorporated into mode choice models in three main ways: through the use of buffer-based methods (as in geographic buffers around a point in space), through the use of pedestrian/bicycle environment factors, and through the use of route choice models (Guo et al., 2007; Replogle and Fund, 1995; Nassir et al., 2014). Each of these methods has their drawbacks. First, buffer-based methods require the arbitrary setting of a distance to use as the buffer radius around the origin and/or destination. Secondly, it is not necessarily clear that all of the area around a person's origin and/or destination is important to a traveler's mode choice decision, especially since an individual will be traveling in a particular direction and only the attributes of the built environment in that direction are expected to be relevant. Thirdly, if multiple spatial attributes are to be included in one's mode choice model, it is not clear whether or how those multiple variables should be combined for entry into the model.

Arbitrariness or the ad-hoc nature of methods for combining multiple attributes that one considers important is also a criticism of pedestrian/bicycle environment factors (Ewing and Cervero, 2001). Such environment factors are often hand-crafted indices that are thought to measure the "quality" of the built environment for walking and bicycling (Replogle and Fund, 1995), but the coefficients used to combine different variables into a single number are typically chosen through "expert-judgement" as opposed to being chosen through a systematic method. In contrast, by using route-choice models to measure the quality of various routes and then using the logsum in a mode choice model to measure the overall quality of the bicycling option, one avoids the pitfalls of using ad-hoc methods to quantify the quality of bicycling for an individual or of considering irrelevant geographies as with buffer based methods.

However, there are at least three problems with the use of combined route and mode choice models. First, transportation datasets that are used for mode choice models are typically travel diaries, and such datasets do not typically collect information on the precise routes that individuals, particularly cyclists, use. Since these datasets cannot be used to construct route choice models, the coefficients from route choice models developed on one set of individuals are then used to represent the sensitivities of different individuals in the mode choice dataset. Secondly, route choice models are only estimated on those for whom we have observed route choices–i.e. current cyclists. The sensitivities of current cyclists are then taken to be the same as the sensitivities of non-cyclists, and this is probably an inaccurate assumption. Nevertheless, the logsum measure based on cyclists' sensitivities are used to represent the overall quality of bicycling for non-cyclists. Lastly, it typically takes a long time to estimate and forecast with route choice models (Nassir et al., 2014). Both the initial estimation and the forecasting process of common route choice models depend on lengthy route generation processes that limit the practical usefulness of such models.

Given the issues with previous approaches seen in the literature, a new method of incorporating roadway-level variables into mode choice models is desired—a method that combines the virtues of the extant methods and avoids their problems. Ideally, this new methodology should

1. minimize the use of “arbitrary” parameters to be set by the analyst (such as the buffer distance in buffer-based methods or the weighting coefficients in bicycle environment factors)

2. avoid including “irrelevant” roadways when quantifying roadway conditions for bicycling

3. incorporate variables related to roadways *in between* an individuals origin and destination, not just *at* the origin or destination

4. avoid implying equality of sensitivities between disparate datasets without justification for such implications

5. avoid applying coefficients estimated solely based on bicyclists to samples of bicyclists and non-cyclists

6. minimize computing time so that the method can be practically useful.

In Chapter 2 I describe the “zone-of-likely travel” and decision-tree approach to incorporating roadway level variables into mode choice models. As will be shown, the new method meets all of the desiderata listed above.

4 Differential adoption and abandonment

Beyond the need for new methods of incorporating roadway-level variables into mode choice models, bicycle mode choice models should be as accurate as possible. Greater accuracy of one’s mode choice model enables agencies to more accurately: anticipate the future travel demands of their constituent populations, understand how the travel behavior of their populations are likely to change in response to changes in the transportation system, and quantitatively judge which projects will contribute the most towards their regions’ various goals (e.g. congestion management, air quality, health, sustainability, etc.). The perspective taken in this dissertation is that one will be most successful in improving the accuracy of bicycle demand models by taking into account the specific features that characterize choice modeling in the context of bicycle mode choice. To be concrete, two salient characteristics that are pertinent to bicycle mode choice modeling are that bicycle demand modeling in the U.S. often takes place in a significantly class imbalanced setting and that individuals rarely seem to go through a fully rational, compensatory, utility-maximizing procedure when choosing whether to commute by bicycle.

Taking the first characteristic mentioned above, class-imbalance is defined in this dissertation as a condition where at least one discrete outcome is over- or under-represented compared to the other outcomes. In particular, if one is dealing with a cross-sectional dataset with J discrete outcomes ($J \geq 2$), class imbalance is the situation where one or more of the J outcomes is present in more (or less) than $\frac{N}{J}$ records in one’s dataset, where N is the total number of observations in the dataset. Given that bicycle mode shares in large U.S. cities are typically very low, one might correctly surmise that class-imbalance is typical of travel mode choice datasets that include bicycling. For instance, the highest bicycle commute mode share is approximately 6% in cities with between 300,000 and 1 million people; approximately 2% in cities with more than 1 million people; and even lower in the suburbs surrounding cities. (League of American Bicyclists, 2014). As a result, household travel surveys–the main data source for mode choice models–that follow a simple random sampling protocol will likely result in datasets where the percentage of bicycle commuters is much lower than the percentage of commuter of other modes (relative class imbalance) and where the absolute number of bicycle commuters is small (absolute class imbalance). This is evident, for example, in the travel demand model used by San Francisco County, SF-CHAMP. The survey used to estimate SF-CHAMP included 10,897 trips, of which only 100 trips were made by bicycle, leading to a bicycle mode share of approximately 0.92% in their dataset (Cambridge Systematics, 2002).

Far from being an innocuous quality, class imbalance is often mentioned as causing serious problems for analysts. Absolute class imbalance, where one simply has a low number of observations of one or more outcomes is claimed by travel demand modelers (Parsons Brinckerhoff Quade & Douglas et al., 2005), academic statisticians (Chen et al., 2004), and computer scientists (Weiss, 2004; He and Garcia, 2009) alike as making it hard to build common models and degrading the performance of such models. In general, absolute class imbalance is thought to lead to a situation where one cannot adequately "learn a concept" due to a scarcity of observations from which one builds a model (Weiss, 2004; He and Garcia, 2009). Likewise, relative class imbalance, where one has a relatively lower amount of observations from one or more outcomes as compared to other categories, is also thought to degrade an analyst's ability to model the probability of the "rare class(es)" (Cramer, 1999; King and Zeng, 2001; Japkowicz, 2000; Wallace et al., 2011). There are a number of reasons why this degradation is thought to take place, but in the context of standard discrete choice models, the explanations offered so far depend on whether the standard logit model is thought to be the correct probability model.

If the logit model is the correct probability model, then relative class imbalance is thought to cause problems by increasing the finite sample bias in one's estimated utility function coefficients, in such a way that one's probability estimates of the "rare" class are downward biased more than they would be in the case of balanced class observations (King and Zeng, 2001). In both this case, and in the case of absolute class imbalance, collecting choice-based samples and using relevant estimation techniques such as weighted-exogenous maximum likelihood can eliminate the problems caused by class imbalance (Manski and Lerman, 1977; Breslow et al., 1987; King and Zeng, 2001). However, since analysts often lack control over the data collection process, post-data-collection methods of mitigating the negative effects of class-imbalance are still useful. Additionally, there are many statisticians in fields such as marketing (Wang and Dey, 2010), finance (Calabrese and Osmetti, 2011), insurance (Bermúdez et al., 2008; Pérez-Sánchez et al., 2014), biology (Jiang et al., 2013), and medicine (Sáez-Castillo et al., 2010) who believe that relative class imbalance may be indicative of a situation where the true probability function is asymmetric[2], unlike the standard logit or probit model that are both symmetric (Chen et al., 1999). Such asymmetric models imply that individuals have rates of adoption and abandonment that differ by alternative. If the true probability function is choices in actually asymmetric, then using a logit or probit model (the typical transportation mode choice models) to predict the probability a particular outcome occurring will result in link-function mis-specification: a condition known to result in inconsistent coefficient estimates (Czado and Santner, 1992) and in probability estimates which can be far from the truth (Koenker and Yoon, 2009).

To guard against link function mis-specification, one would ideally estimate a *single-index model* where one performs a non-parametric estimation of the probability function, along with a parametric estimation of the coefficients of one's utility functions (Härdle et al., 1997; Horowitz, 2010). While ideal, such a procedure can be computationally intensive and may require a very large number of observations to obtain an adequate estimate of the probability function. As a compromise between doing nothing or performing a complete non-parametric estimation, some statisticians (Czado, 1994; Koenker and Yoon, 2009) have suggested "embedding" the standard logit and probit models in a wider class of "parametric link-functions" where a vector of parameters can be estimated according to one's data to allow varying types and degrees of asymmetry in one's probability function to fit one's data best.

Motivated by the aforementioned literature, a large number of asymmetric probability functions have been introduced for binary outcomes. For example, there have been numerous, distinct one parameter generalizations of the logit model: (Aranda-Ordaz, 1981; Guerrero and Johnson, 1982; Czado, 1992; Nagler, 1994; Chen et al., 1999; Masnadi-shirazi and Vasconcelos, 2010; Nakayama and Chikaraishi, 2015; Komori et al., 2015). Additionally, a number of two parameter generalizations of the binary logit model have been created: (Prentice, 1976; Pregibon, 1980; Stukel, 1988; Czado, 1994; Vijverberg, 2000; Vijverberg and Vijverberg, 2012). Beyond generalizations of the logit model, many other other binary probability functions have been put forth over the years based on distributions such as the student's t distribution, the cauchy distribution, the generalized extreme value distribution, the weibull distribution, the skew-normal distributions, and so on: (Kim, 2002; Liu, 2004; Castillo et al., 2008; Kim et al., 2008; Koenker and Yoon, 2009; Wang and Dey, 2010; Li, 2011; Jiang et al., 2013).

Despite the work mentioned above, to be useful in addressing relative class-imbalance problems in the context of bicycle mode choice modeling, the proposed parametric link-functions must be capable of handling multinomial outcomes such as "bicycle, walk, drive alone, bus, train" and so on. Unfortunately, besides the function's created in this dissertation, only three multinomial, parametric link-functions have been proposed and estimated so far: the link family originally introduced by Czado for binary regression (Das and Mukhopadhyay, 2014), the weibit model (Castillo et al., 2008; Fosgerau and Bierlaire, 2009), and the

[2] See Chapter 3 for a detailed definition of symmetric and asymmetric probability functions.

q-generalized logit model (Nakayama and Chikaraishi, 2015). These functions have never been applied in the context of bicycle travel, and they are either complex or too restrictive for general use. One reason for the paucity of multinomial, parametric link-functions may be the fact that no unified approach for the generation of these various probability functions has been provided in the literature so far. If the research on binary, parametric link-functions is to be built upon and be brought to bear on the problem of class imbalance in bicycle mode choice modeling, then we need (1) a systematic way of generating multinomial, parametric link-functions and (2) a demonstration of whether such functions can appreciably increase the accuracy of standard models of bicycle mode choice.

In Chapter 3, I present methods for generating new asymmetric choice models, I show that these new models can fit one's data much better than traditional mode choice models, and I demonstrate how these asymmetric models lead to new insights that are missing from traditional methods.

5 Non-compensatory bicycle mode choices

Returning to the second characteristic of bicycle commuting mentioned above, it is not known *a-priori* that individuals make completely rational choices of whether to commute by bicycle. Indeed, it is not clear that individuals perform a comprehensive, weighted comparison of all of the attributes of each of their possible commuting alternatives and then select the alternative that maximizes their individual utility. It is possible that individuals may use heuristic decision rules to shrink the set of alternatives that they subject to a utility maximizing procedure, i.e. before using utility maximization to make a final choice from their consideration sets, individuals may use other decision rules to first narrow their consideration set from their overall choice sets. Critically, 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 one does not account for the fact that such individuals do not even consider bicycling, then one will make incorrect inferences regarding the amount by which any project can be expected to increase the expected number of cyclists. In other words, one must be sure that an individual is considering bicycling as a commuting option before trying to judge the ability of an intervention to increase the probability that the individual actually bikes.

To allow for heterogenous consideration sets in a population, 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 are either completely rational 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 (Swait, 2009; Vij and Walker, 2014; Mahmoud et al., 2015).

In all the modeling efforts just described, the probability of an individual considering a particular mode was always based on a compensatory model where one variable with a positive effect on a mode's probability of being considered could make up for a variable with a negative effect on a mode's probability of being considered. The compensatory nature of the aforementioned models is 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 drop-off children 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, thereby achieving greater model accuracy, more realistic model forecasts, and/or qualitatively different model forecasts than the compensatory models which have been used in the mode choice setting thus far.

In contexts other than travel mode choice, multiple models that combine non-compensatory decision rules for consideration set construction with compensatory utility-maximization for choosing from within that consideration set have already been designed and estimated. Non-compensatory decision rules that have been used include conjunctive rules (Swait, 2001b; Elrod et al., 2004; Gilbride and Allenby, 2004), subset conjunctive rules (Jedidi and Kohli, 2005), disjunctive rules (Swait, 2001b; Elrod et al., 2004; Gilbride and Allenby, 2004), "disjunctions of conjunctions" (Hauser et al., 2010), dominance (Cascetta and Papola, 2009), elimination by aspects (Gilbride and Allenby, 2006), economic screening (Gilbride and Allenby, 2006), lexicographic rules (Kohli and Jedidi, 2007), and satisficing (Stüttgen et al., 2012) to name a few. Given this existing literature, what is now needed is an application of the various models based on non-compensatory decision rules to the context of consideration set formation in bicycle mode choice models, and a comparison with the prevailing compensatory approaches to choice set formation.

Beyond the purely pragmatic concerns over which type of model will most accurately represent the bicycle mode choice process, there are at least two theoretical contributions to be made by combining non-compensatory models of consideration set formation with compensatory models of the conditional choice made by each individual. First, a wide variety of methods have been used to estimate the various non-compensatory decision rules for different datasets. However, despite the eclectic set of estimation methods, and despite the fact that a subset of these rules–namely conjunctive rules, subset conjunctive rules, disjunctive rules, and "disjunctions of conjunctions"–can be represented as decision trees (Hauser et al., 2010), tree-induction algorithms from computer science and statistics have not been used to estimate these non-compensatory decision rules. Similar to how McFadden (1972) used random utility maximization to bring economic meaning and theory to the long-used logistic regression model, I use non-compensatory decision rules to bring economic meaning and theory to classification trees.

Additionally, disjunctions-of-conjunctions and its various special cases can be thought of as describing distinct situations. These situations are then used to characterize whether an alternative is available to a decision maker or not. Typically, the choice is then modeled using a probability model whose coefficients do not differ according to the situation associated with the decision maker. This procedure implicitly throws away information since we know the shares of each alternative chosen by individuals in each alternative, and we could use this information to infer individual preferences in each situation. Hybrid decision-tree logit models (Steinberg and Cardell, 1998) or similar types of "model trees" present one way to formalize such influence, but we still lack methods to perform joint (as opposed to sequential), probabilistically motivated estimations for models that (1) use tree-like models to model an individual's consideration set and (2) use information from the shares of each alternative in a given situation to infer individual preferences in the compensatory, utility maximizing choice model used to make the final choice given one's consideration set. The development of such methods would allow analysts to make greater use of the data that they already have in order to more accurately predict the probability that an individual considers and then chooses bicycling as a commute mode.

In Chapter 4, I present the general framework for interpreting decision trees through the lens of microeconomic theories, and I present methods for jointly estimating bayesian model trees. These are new hybrids of decision trees and discrete choice models that allow for non-compensatory consideration set formation, analyst uncertainty over the non-compensatory rules, and context-dependent preference heterogeneity. By applying such models to disaggregate bicycle mode choice data in the San Francisco Bay area, we are able to form much better fitting models of bicycle mode choice, we generate predictions that are far more plausible than traditional mode choice models, and we gain qualitative insights into the way individuals are likely to react to bicycle infrastructure investments.

6 Causes versus associations

In an ideal setting, transportation professionals would look at the budget they have available to spend on bicycle infrastructure projects, and then they would proactively solve an optimization problem to determine where new infrastructure should be installed to maximize the bicycle mode share in their jurisdiction. However, in order to have such optimization problems be meaningful, the mode choice model that is used to predict the bicycle mode shares must be measuring *causal* as opposed to *associational* relationships. The distinction between the two types of relationships is that causal inferences will be valid under external policy interventions (such as a transportation agency installing new bicycle lanes), whereas an associational inference will not necessarily be accurate under external intervention.

The academic discipline of causal inference is exclusively devoted to determining when and how such causal relationships can be estimated. However, travel demand modeling and causal inference research have largely remained separate from each other. Though travel demand modelers expect that their models will be accurate under external intervention, the relevant techniques from the causal inference literature are almost never used to verify and guide the creation of traditional travel demand models.

In Chapter 5, I examine the similarities and disconnects between the fields of causal inference and travel demand modeling. Though the two fields share similar objectives, I explore the striking differences between the methods employed by each discipline. Moreover, at a more general level, I step back to review the expected benefits that will likely come from a cross-pollination of techniques between travel demand modelers and causal inference researchers. Using bicycling as a case study, I examine how travel demand modeling practices might change in order to incorporate insights from causal inference research, and I detail challenges to such a methodological merger, i.e. the kind of merger that is needed for transportation planners and engineers to make strategic investments in bicycle infrastructure.

7 Contributions

Synthesizing the information presented above, this dissertation makes contributions related to each of the four research objectives described in Section 2.

First, I develop a new decision-tree based method for incorporating roadway-level variables into discrete choice models. This new method uses a novel concept called the "zone of likely travel." In doing so, I build upon previous GPS-based research that characterizes the way cyclists travel. By combining the zone of likely travel with decision trees, my new method combines the strengths of buffer-based approaches and bicycle environment factors. Importantly, my new method avoids many drawbacks of these earlier methods and of route-choice based methods for incorporating roadway level variables into mode choice models.

Secondly, motivated by the fact that bicycle demand models are almost always estimated on class-imbalanced datasets, I make three methodological contributions related to the issue of class imbalance. First, I introduce a new class of closed-form, finite-parameter multinomial choice models. This new class generalizes many existing models from discrete choice and statistics. Additionally, these new models can capture differential rates of adoption and abandonment between alternatives. It is this asymmetry in adoption and abandonment rates that enables these models to better explain why class-imbalance is occurring.

Beyond the mere creation of a new class of models. I develop two procedures for creating new models within this class. These procedures fill two existing gaps in the statistical and discrete choice literature. The first procedure provides a way for researchers to create new binary choice models, either symmetric or asymmetric. In contrast to the hitherto ad-hoc (and often undocumented) methods used to create new binary choice models, this new procedure allows researchers to systematically derive new choice models with particular properties such as whether the model is symmetric or not. The second procedure provides a systematic way to extend binary choice models to the multinomial setting using my newly created class of multinomial choice models. Together with the first procedure, this advancement allows for the standardized creation of new, possibly asymmetric, multinomial choice models. Furthermore, the second procedure allows for the creation of multinomial extensions to existing statistical models, theereby making these models vastly more useful to researchers in fields such as marketing, transportation, etc.

Thirdly, I make two contributions related to the incorporation of non-compensatory decision rules into discrete choice and bicycle demand models. First, I develop a microeconomic framework for the interpretation of decision trees and their many generalizations. In doing so, I show (1) how decision trees represent a class of non-compensatory decision rules known as disjunctions-of-conjunctions and (2) how decision trees increase the flexibility of the non-compensatory behaviors that can be empirically estimated in discrete choice settings. Secondly, I make methodological contributions to both machine learning and the existing literature of two-stage, semi-compensatory choice models. By developing the first bayesian model tree, I allow for estimation uncertainty in the estimated decision tree, and I allow for the joint estimation of the trees and discrete choice models at the output nodes of the decision trees. Such a model represents a notion of context-dependent preference heterogeneity where the parameters of one's choice model can differ across output nodes, thereby making the choice model parameters a function of the variables used to construct the decision tree.

Finally, because strategic bicycle planning would be best supported by mode choice models that measure causal relationships, I make two contributions that try to bridge the gap between the fields of travel demand modeling and causal inference. First, I identify causal inference techniques that can be used by travel demand modelers to improve our ability to make causally valid inferences. Secondly, I identify areas of the causal inference literature that can benefit (either methodologically or in application) from the engagement of travel demand modelers.

Together, these contributions represent significant methodological advances in discrete choice, statistics, machine learning. Moreover, these contributions solve practical problems related to the use of discrete choice models for bicycle planning. And lastly, these contributions forge critical connections with the field of causal inference, thereby paving the way towards more useful demand models—for bicyclists and for all.

8 Dissertation Outline

This dissertation is structured as follows. Chapter 2 incorporates roadway level variables into mode choice models using decision trees and a novel concept called the zone of likely travel. Chapter 3 uses asymmetric mode choice models to capture differential rates of adoption and abandonment of alternative travel modes, thereby allowing the models to better explain the observed levels of class imbalance (e.g. the relatively low levels of cycling). In Chapter 4, I use bayesian model trees to account for "if-then" methods of decision making that are likely to be used by individuals that are deciding whether or not to commute by bike. Next, in Chapter 5, I reflect on the state of the union between causal inference and travel demand modeling. I review the benefits to each community that may come from a stronger link between the two fields; I examine why causal inference techniques have not been incorporated into travel demand models so far; and I take first steps at sketching how one might incorporate causal inference techniques into travel demand modeling overall and bicycle demand modeling in particular. Finally in Chapter 6, I recap what has been accomplished, what next steps are ready to be taken now, and what issues remain to be addressed in the indeterminate future.

References

Francisco J. Aranda-Ordaz. On two families of transformations to additivity for binary response data. *Biometrika*, 68(2):357–363, August 1981. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/68.2.357. URL `http://biomet.oxfordjournals.org/content/68/2/357`.

Ll. Bermúdez, J. M. Pérez, M. Ayuso, E. Gómez, and F. J. Vázquez. A Bayesian dichotomous model with asymmetric link for fraud in insurance. *Insurance: Mathematics and Economics*, 42(2):779–786, April 2008. ISSN 0167-6687. doi: 10.1016/j.insmatheco.2007.08.002. URL `http://www.sciencedirect.com/science/article/pii/S0167668707000947`.

Norman E. Breslow, Nicholas E. Day, and others. *Statistical methods in cancer research*, volume 2. International Agency for Research on Cancer Lyon, 1987. URL `http://w2.iarc.fr/en/publications/pdfs-online/stat/sp32/SP32_vol1-0.pdf`.

Raffaella Calabrese and Silvia Osmetti. Generalized Extreme Value Regression for Binary Rare Events Data: an Application to Credit Defaults. Working Paper 201120, Geary Institute, University College Dublin, September 2011. URL `http://econpapers.repec.org/paper/ucdwpaper/201120.htm`.

California Department of Transportation. CALIFORNIA BLUEPRINT FOR BICYCLING AND WALKING : REPORT TO THE LEGISLATURE. Technical report, California Department of Transportation, 2002. URL `http://www.dot.ca.gov/hq/LocalPrograms/bike/CABlueprintRpt.pdf`.

Cambridge Systematics. San Francisco Travel Demand Forecasting Model Development: Mode Choice Models. Technical report, San Francisco County Transportation Authority, October 2002. URL `http://www.sfcta.org/modeling-and-travel-forecasting`.

Ennio Cascetta and Andrea Papola. Dominance among alternatives in random utility models. *Transportation Research Part A: Policy and Practice*, 43(2):170–179, February 2009. ISSN 0965-8564. doi: 10.1016/j.tra.2008.10.003. URL `http://www.sciencedirect.com/science/article/pii/S0965856408001894`.

Enrique Castillo, José María Menéndez, Pilar Jiménez, and Ana Rivas. Closed form expressions for choice probabilities in the Weibull case. *Transportation Research Part B: Methodological*, 42(4):373–380, May 2008. ISSN 0191-2615. doi: 10.1016/j.trb.2007.08.002. URL `http://www.sciencedirect.com/science/article/pii/S019126150700077X`.

Chao Chen, Andy Liaw, and Leo Breiman. Using random forest to learn imbalanced data. Technical report, University of California, Berkeley, 2004. URL `http://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf`.

Ming-Hui Chen, Dipak K. Dey, and Qi-Man Shao. A New Skewed Link Model for Dichotomous Quantal Response Data. *Journal of the American Statistical Association*, 94(448):1172–1186, December 1999. ISSN 0162-1459. doi: 10.2307/2669933. URL `http://www.jstor.org/stable/2669933`.

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

J. S. Cramer. Predictive Performance of the Binary Logit Model in Unbalanced Samples. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 48(1):85–94, April 1999. ISSN 1467-9884. doi: 10.1111/1467-9884.00173. URL `http://onlinelibrary.wiley.com/doi/10.1111/1467-9884.00173/abstract`.

C. Czado. Parametric link modification of both tails in binary regression. *Statistical Papers*, 35(1):189–201, December 1994. ISSN 0932-5026, 1613-9798. doi: 10.1007/BF02926413. URL `http://link.springer.com/article/10.1007/BF02926413`.

Claudia Czado. On Link Selection in Generalized Linear Models. In Ludwig Fahrmeir, Brian Francis, Robert Gilchrist, and Gerhard Tutz, editors, *Advances in GLIM and Statistical Modelling*, number 78 in Lecture Notes in Statistics, pages 60–65. Springer New York, 1992. ISBN 978-0-387-97873-4 978-1-4612-2952-0. URL `http://link.springer.com/chapter/10.1007/978-1-4612-2952-0_10`.

Claudia Czado and Thomas J. Santner. The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference*, 33(2):213–231, November 1992. ISSN 0378-3758. doi: 10.1016/0378-3758(92)90069-5. URL `http://www.sciencedirect.com/science/article/pii/0378375892900695`.

I. Das and S. Mukhopadhyay. On generalized multinomial models and joint percentile estimation. *Journal of Statistical Planning and Inference*, 145:190–203, February 2014. ISSN 0378-3758. doi: 10.1016/j.jspi.2013.08.015. URL `http://www.sciencedirect.com/science/article/pii/S0378375813002103`.

Terry Elrod, Richard D. Johnson, and Joan White. A new integrated model of noncompensatory and compensatory decision strategies. *Organizational Behavior and Human Decision Processes*, 95(1):1–19, September 2004. ISSN 0749-5978. doi: 10.1016/j.obhdp.2004.06.002. URL `http://www.sciencedirect.com/science/article/pii/S0749597804000573`.

Reid Ewing and Robert Cervero. Travel and the Built Environment: A Synthesis. *Transportation Research Record: Journal of the Transportation Research Board*, 1780:87–114, January 2001. ISSN 0361-1981. doi: 10.3141/1780-10. URL `http://trrjournalonline.trb.org/doi/abs/10.3141/1780-10`.

Federal Highway Administration. Bicycle and Pedestrian Transportation Planning Guidance - Guidance - Bicycle and Pedestrian Program - Environment - FHWA, August 2003. URL `http://www.fhwa.dot.gov/environment/bicycle_pedestrian/guidance/inter.cfm#i2`.

M. Fosgerau and M. Bierlaire. Discrete choice models with multiplicative error terms. *Transportation Research Part B: Methodological*, 43(5):494–505, June 2009. ISSN 0191-2615. doi: 10.1016/j.trb.2008.10.004. URL `http://www.sciencedirect.com/science/article/pii/S0191261508001215`.

Timothy J. Gilbride and Greg M. Allenby. A Choice Model with Conjunctive, Disjunctive, and Compensatory Screening Rules. *Marketing Science*, 23(3):391–406, 2004. ISSN 0732-2399. URL `http://www.jstor.org/stable/30036705`.

Timothy J. Gilbride and Greg M. Allenby. Estimating Heterogeneous EBA and Economic Screening Rule Choice Models. *Marketing Science*, 25(5):494–509, 2006. ISSN 0732-2399. URL `http://www.jstor.org/stable/40057038`.

Frank Goetzke. *Are Travel Demand Forecasting Models Biased because of Uncorrected Spatial Autocorrelation? By.* 2003.

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

Victor M. Guerrero and Richard A. Johnson. Use of the Box-Cox transformation with binary response models. *Biometrika*, 69(2):309–314, August 1982. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/69.2.309. URL `http://biomet.oxfordjournals.org/content/69/2/309`.

Jessica Guo, Chandra Bhat, and Rachel Copperman. Effect of the Built Environment on Motorized and Nonmotorized Trip Making: Substitutive, Complementary, or Synergistic? *Transportation Research Record: Journal of the Transportation Research Board*, 2010:1–11, January 2007. ISSN 0361-1981. doi: 10.3141/2010-01. URL `http://trrjournalonline.trb.org/doi/10.3141/2010-01`.

Wolfgang Härdle, V. Spokoiny, Stefan Sperlich, and others. Semiparametric single index versus fixed link function modelling. *The Annals of Statistics*, 25(1):212–243, 1997. URL `http://projecteuclid.org/euclid.aos/1034276627`.

John R Hauser, Olivier Toubia, Theodoros Evgeniou, Rene Befurt, and Daria Dzyabura. Disjunctions of Conjunctions, Cognitive Simplicity, and Consideration Sets. *Journal of Marketing Research*, 47(3):485–496, June 2010. ISSN 0022-2437. doi: 10.1509/jmkr.47.3.485. URL `http://journals.ama.org/doi/abs/10.1509/jmkr.47.3.485`.

Haibo He and E.A Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, September 2009. ISSN 1041-4347. doi: 10.1109/TKDE.2008.239.

Joel L. Horowitz. *Semiparametric and Nonparametric Methods in Econometrics*. Springer Science & Business Media, July 2010. ISBN 978-0-387-92870-8.

Nathalie Japkowicz. The Class Imbalance Problem: Significance and Strategies. In *In Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI*, pages 111–117, 2000.

Kamel Jedidi and Rajeev Kohli. Probabilistic Subset-Conjunctive Models for Heterogeneous Consumers. *Journal of Marketing Research*, 42(4):483–494, November 2005. ISSN 0022-2437. doi: 10.1509/jmkr.2005.42.4.483. URL `http://journals.ama.org/doi/abs/10.1509/jmkr.2005.42.4.483`.

Xun Jiang, Dipak K. Dey, Rachel Prunier, Adam M. Wilson, and Kent E. Holsinger. A new class of flexible link functions with application to species co-occurrence in cape floristic region. *The Annals of Applied Statistics*, 7(4):2180–2204, December 2013. ISSN 1932-6157. doi: 10.1214/13-AOAS663. URL `http://arxiv.org/abs/1401.1915`. arXiv: 1401.1915.

Hea-Jung Kim. Binary Regression with a Class of Skewed t-Link Models. *Communications in Statistics - Theory and Methods*, 31(10):1863–1886, January 2002. ISSN 0361-0926. doi: 10.1081/STA-120014917. URL `http://dx.doi.org/10.1081/STA-120014917`.

Sungduk Kim, Ming-Hui Chen, and Dipak K. Dey. Flexible generalized t-link models for binary response data. *Biometrika*, 95(1):93–106, March 2008. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/asm079. URL `http://biomet.oxfordjournals.org/content/95/1/93`.

Gary King and Langche Zeng. Logistic Regression in Rare Events Data. *Political Analysis*, 9(2):137–163, January 2001. ISSN 1047-1987, 1476-4989. URL `http://pan.oxfordjournals.org/content/9/2/137`.

Roger Koenker and Jungmo Yoon. Parametric links for binary choice models: A Fisherian–Bayesian colloquy. *Journal of Econometrics*, 152(2):120–130, October 2009. ISSN 0304-4076. doi: 10.1016/j.jeconom.2009.01.009. URL `http://www.sciencedirect.com/science/article/pii/S0304407609000207`.

Rajeev Kohli and Kamel Jedidi. Representation and Inference of Lexicographic Preference Models and Their Variants. *Marketing Science*, 26(3):380–399, May 2007. ISSN 0732-2399. doi: 10.1287/mksc.1060.0241. URL `http://pubsonline.informs.org/doi/abs/10.1287/mksc.1060.0241`.

Osamu Komori, Shinto Eguchi, Shiro Ikeda, Hiroshi Okamura, Momoko Ichinokawa, and Shinichiro Nakayama. An asymmetric logistic regression model for ecological data. *Methods in Ecology and Evolution*, pages n/a–n/a, October 2015. ISSN 2041-210X. doi: 10.1111/2041-210X.12473. URL `http://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12473/abstract`.

League of American Bicyclists. Where We Ride: Analysis of bicycling in American cities. Technical report, The League of American Bicyclists, September 2014. URL `http://bikeleague.org/sites/default/files/ACS_report_2014_forweb.pdf`.

Baibing Li. The multinomial logit model revisited: A semi-parametric approach in discrete choice analysis. *Transportation Research Part B: Methodological*, 45(3):461–473, March 2011. ISSN 0191-2615. doi: 10.1016/j.trb.2010.09.007. URL `http://www.sciencedirect.com/science/article/pii/S0191261510001190`.

Chuanhai Liu. Robit regression: a simple robust alternative to logistic and probit regression. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, pages 227–238, 2004. URL `http://www.stat.purdue.edu/~chuanhai/teaching/Stat598A/robit.pdf`.

Mohamed Salah Mahmoud, Adam Weiss, and Khandker Nurul Habib. Latent Captivation or Mode Culture? Investigation into Mode Choice Preference Structures in Competitive Modal Arrangements. 2015. URL `http://trid.trb.org/view.aspx?id=1337392`.

Charles F. Manski and Steven R. Lerman. The Estimation of Choice Probabilities from Choice Based Samples. *Econometrica*, 45(8):1977–1988, November 1977. ISSN 0012-9682. doi: 10.2307/1914121. URL `http://www.jstor.org/stable/1914121`.

Hamed Masnadi-shirazi and Nuno Vasconcelos. Variable margin losses for classifier design. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1576–1584. Curran Associates, Inc., 2010. URL `http://papers.nips.cc/paper/4024-variable-margin-losses-for-classifier-design.pdf`.

D. McFadden. CONDITIONAL LOGIT ANALYSIS OF QUALITATIVE CHOICE BEHAVIOR. *WORKING PAPER INSTITUTE OF URBAN AND REGIONAL*, (199/), 1972. URL `http://trid.trb.org/view.aspx?id=235187`.

Jonathan Nagler. Scobit: An Alternative Estimator to Logit and Probit. *American Journal of Political Science*, 38(1):230–255, February 1994. ISSN 0092-5853. doi: 10.2307/2111343. URL `http://www.jstor.org/stable/2111343`.

Shoichiro Nakayama and Makoto Chikaraishi. A Unified Closed-form Expression of Logit and Weibit and its Application to a Transportation Network Equilibrium Assignment. *Transportation Research Procedia*, 7:59–74, 2015. ISSN 2352-1465. doi: 10.1016/j.trpro.2015.06.004. URL `http://www.sciencedirect.com/science/article/pii/S2352146515000721`.

Neema Nassir, Jennifer Ziebarth, Elizabeth Sall, and Lisa Zorn. A Choice Set Generation Algorithm Suitable for Measuring Route Choice Accessibility. 2014. URL `http://trid.trb.org/view/2014/C/1289516`.

Parsons Brinckerhoff Quade & Douglas, PB Consult, AECOM Consult, Urbitran Associates, Urbanomics, Alex Anas & Associates, NuStats International, and George Hoyte & Associates. Transportation Models and Data Initiative: General Final Report–New York Best Practice Model. Technical report, New York Metropolitan Transportation Council, January 2005. URL `http://www.nymtc.org/project/bpm/model/bpm_finalrpt.pdf`.

J.m. Pérez-Sánchez, M.a. Negrín-Hernández, C. García-García, and E. Gómez-Déniz. BAYESIAN ASYMMETRIC LOGIT MODEL FOR DETECTING RISK FACTORS IN MOTOR RATEMAKING. *ASTIN Bulletin*, 44(02):445–457, May 2014. ISSN 1783-1350. doi: 10.1017/asb.2013.32. URL `http://journals.cambridge.org/article_S0515036113000329`.

Daryl Pregibon. Goodness of Link Tests for Generalized Linear Models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(1):15–14, January 1980. ISSN 0035-9254. doi: 10.2307/2346405. URL `http://www.jstor.org/stable/2346405`.

Ross L. Prentice. A Generalization of the Probit and Logit Methods for Dose Response Curves. *Biometrics*, 32(4):761–768, December 1976. ISSN 0006-341X. doi: 10.2307/2529262. URL `http://www.jstor.org/stable/2529262`.

Michael Replogle and Environmental Defense Fund. *Integrating Pedestrian and Bicycle Factors Into Regional Transportation Planning Models: Summary of the State-of-the-art and Suggested Steps Forward*. Environmental Defense Fund Washington, DC, 1995. URL `http://media.tmiponline.org/clearinghouse/udes/replogle.pdf`.

Antonio José Sáez-Castillo, María José Olmo-Jiménez, José María Pérez Sánchez, Miguel Ángel Negrín Hernández, Ángel Arcos-Navarro, and Juan Díaz-Oller. Bayesian Analysis of Nosocomial Infection Risk and Length of Stay in a Department of General and Digestive Surgery. *Value in Health*, 13(4):431–439, 2010. ISSN 1524-4733. doi: 10.1111/j.1524-4733.2009.00680.x. URL `http://onlinelibrary.wiley.com/doi/10.1111/j.1524-4733.2009.00680.x/abstract`.

SFMTA. 2012 San Francisco State of Cycling Report. Technical report, San Francisco Municipal Transportation Agency, September 2012. URL `http://archives.sfmta.com/cms/rbikes/documents/2012StateofCyclingReport8_9_12.pdf`.

Patrick A. Singleton and Kelly J. Clifton. Pedestrians in Regional Travel Demand Forecasting Models: State of the Practice. 2013. URL http://trid.trb.org/view.aspx?id=1242847.

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

Thérèse A. Stukel. Generalized Logistic Models. *Journal of the American Statistical Association*, 83 (402):426–431, June 1988. ISSN 0162-1459. doi: 10.1080/01621459.1988.10478613. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1988.10478613.

Peter Stüttgen, Peter Boatwright, and Robert T. Monroe. A Satisficing Choice Model. *Marketing Science*, 31(6):878–899, September 2012. ISSN 0732-2399. doi: 10.1287/mksc.1120.0732. URL http://pubsonline.informs.org/doi/abs/10.1287/mksc.1120.0732.

Joffre Swait. Choice set generation within the generalized extreme value family of discrete choice models. *Transportation Research Part B: Methodological*, 35(7):643–666, 2001a. URL http://www.sciencedirect.com/science/article/pii/S0191261500000291.

Joffre Swait. A non-compensatory choice model incorporating attribute cutoffs. *Transportation Research Part B: Methodological*, 35(10):903–928, November 2001b. ISSN 0191-2615. doi: 10.1016/S0191-2615(00)00030-8. URL http://www.sciencedirect.com/science/article/pii/S0191261500000308.

Joffre Swait. Choice models based on mixed discrete/continuous PDFs. *Transportation Research Part B: Methodological*, 43(7):766–783, August 2009. ISSN 0191-2615. doi: 10.1016/j.trb.2009.02.003. URL http://www.sciencedirect.com/science/article/pii/S0191261509000216.

Joffre Swait and Moshe Ben-Akiva. Empirical test of a constrained choice discrete model: Mode choice in São Paulo, Brazil. *Transportation Research Part B: Methodological*, 21(2):103–115, April 1987a. ISSN 0191-2615. doi: 10.1016/0191-2615(87)90010-5. URL http://www.sciencedirect.com/science/article/pii/0191261587900105.

Joffre Swait and Moshe Ben-Akiva. Incorporating random constraints in discrete models of choice set generation. *Transportation Research Part B: Methodological*, 21(2):91–102, April 1987b. ISSN 0191-2615. doi: 10.1016/0191-2615(87)90009-9. URL http://www.sciencedirect.com/science/article/pii/0191261587900099.

U.S. Census Bureau. B08006: Sex of Workers by Means of Transportation to Work. Technical report, 2014. URL http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_14_5YR_B08006&prodType=table.

Akshay Vij and Joan L. Walker. Preference endogeneity in discrete choice models. *Transportation Research Part B: Methodological*, 64:90–105, June 2014. ISSN 0191-2615. doi: 10.1016/j.trb.2014.02.008. URL http://www.sciencedirect.com/science/article/pii/S0191261514000344.

Akshay Vij, André Carrel, and Joan L. Walker. Incorporating the influence of latent modal preferences on travel mode choice behavior. *Transportation Research Part A: Policy and Practice*, 54:164–178, August 2013. ISSN 0965-8564. doi: 10.1016/j.tra.2013.07.008. URL http://www.sciencedirect.com/science/article/pii/S0965856413001304.

Chu-Ping C. Vijverberg and Wim P. M. Vijverberg. Pregibit: A Family of Discrete Choice Models. SSRN Scholarly Paper ID 2010974, Social Science Research Network, Rochester, NY, February 2012. URL http://papers.ssrn.com/abstract=2010974.

Wim P. M. Vijverberg. Betit: A Family That Nests Probit and Logit. SSRN Scholarly Paper ID 264789, Social Science Research Network, Rochester, NY, December 2000. URL http://papers.ssrn.com/abstract=264789.

Byron C. Wallace, Kevin Small, Carla E. Brodley, and Thomas A. Trikalinos. Class imbalance, redux. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 754–763. IEEE, 2011. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6137280`.

Xia Wang and Dipak K. Dey. Generalized Extreme Value Regression for Binary Response Data: An Application to B@B Electronic Payments System Adoption. *The Annals of Applied Statistics*, 4(4):2000–2023, December 2010. ISSN 1932-6157. URL `http://www.jstor.org/stable/23362457`.

Gary M. Weiss. Mining with Rarity: A Unifying Framework. *SIGKDD Explor. Newsl.*, 6(1):7–19, June 2004. ISSN 1931-0145. doi: 10.1145/1007730.1007734. URL `http://doi.acm.org/10.1145/1007730.1007734`.