# Conclusion

Timothy Brathwaite

March 3, 2018

## 1 Research Overview

In this dissertation, I made two broad types of advances. The first type of advances are substantive: I developed new bicycle demand models. These new models address three bicycle-specific concerns: incorporating important but frequently ignored roadway-level variables, allowing for differential rates of "abandonment" between bicycling and other travel modes, and reflecting the likely reality that individuals use non-compensatory or semi-compensatory decision rules to decide whether they will commute via bicycle.

The second type of advances are methodological. Here, I blended traditional discrete choice models with methods from statistics, machine learning, and causal inference. Specifically, I used statistical techniques to create a new class of finite-parameter, closed-form, asymmetric choice models. Next, I detailed a micro-economic framework for interpreting decision tree models from machine learning, and I created new, semi-compensatory choice models using the proposed framework. Lastly, through the lens of causal inference, I examined travel demand modeling overall and a bicycle demand modeling problem in particular. These examinations led to the recognition that, as a field, we discrete choice modellers still need to make significant methodological changes to our travel demand models (bicycle demand models being included as a subset). The validity of our work is questionable without such changes.

Together, these substantive and methodological advances have three main impacts. First, transportation analyst have more accurate and more realistic bicycle demand models for use in their jurisdictions. Second, discrete choice modellers have more accurate, flexible, and micro-economically diverse models for their various applications. And finally, discrete choice researchers have new frameworks for interacting with the fields of statistics, machine learning, and causal inference.

The remaining sections return to these points and will proceed as follows. First, I will provide a more detailed recap of the contributions I made in this dissertation. Next, I will note the limitations of my efforts. And finally, building off these limitations and looking at future opportunities, I will note some research directions that seem especially promising and/or needed.

## 2 Research Contributions

This dissertation is composed of four primary research efforts. Each of these efforts makes both substantive (i.e. bicycle-planning-focused) contributions and methodological ones. Below, I describe these contributions, one project at a time.

First, I developed a method for incorporating roadway-level variables (e.g. speed limits, slopes, the presence of bike lanes, "share the road arrows" (sharrows), etc.) into travel demand models. This methodology relies upon a novel concept termed the "zone-of-likely-travel" and decision trees from computer science. The basic idea of the zone-of-likely-travel is that it is a polygon that represents a geographic buffer around the shortest-path from one's origin to one's destination. However, this buffer is constrained to follow the roadway network instead of merely being laid over it, and the size of the buffer is such that an individual who cycles is likely to take a path that is within the buffer. For all roadways within the zone-of-likely-travel, the desired spatial variables are recorded, then these variables are aggregated to the level of the zone by taking averages and percentiles. Finally, a decision tree is built using a dependent variable of "bicycle or not," representing whether the individual commuted to work or school by bike. The nodes of the decision tree are used as descriptors of each zone-of-likely-travel.

Theoretically, this variable incorporation procedure blends existing "buffer-based" methods with objectively created "bicycle environment factors." Substantively, my new procedures avoids drawbacks of these earlier methods (and procedures based on route choice models). In particular, My zones are constructed to lie between one's origin and destination, so it excludes roadways that are not in the direction one would travel, and it includes roadways that may be left out by disjoint buffers around one's origin and destination. Moreover, it creates descriptors of the bicycling environment in an objective fashion that is tailored towards predicting bicycle demand. And finally, this method avoids the pitfall of route-choice models, which is that they are built solely on cyclists and then extrapolated to the entire population of cyclists and non-cyclists alike. Methodologically, I created a new method for improving the accuracy of bicycle demand models, and I provided another example of the successful and complementary combination of decision trees and discrete choice models.

Next, I created a class of asymmetric, multinomial discrete choice models with a finite number of parameters. This class of "logit-type" models has the substantive affect of allowing analysts to model the (a-priori likely) situation where individuals have differing rates of adoption or abandonment of depending on the travel mode being examined. For instance, from a state of indifference, an individual's probability of commuting by bicycle may fall quickly with respect to a change in its systematic utility. In contrast, an individual's probability of driving may fall less rapidly with an equal change in the systematic utility of driving. Empirically, this new class of models made more accurate predictions than standard, symmetric discrete choice models.

Methodologically, this extension makes a number of contributions to the discrete choice literature. First, such models provide a more flexible way to explain observed levels of class-imbalance as opposed to solely relying on one's alternative specific constants. Secondly, the new class of logit-type models generalizes a wide variety of existing, binary asymmetric models to the multinomial setting for the first time, making such models immediately useful to transportation, marketing, and other disciplines where individuals choose between multiple (i.e. more than two) alternatives. Third, I also introduced procedures for systematically creating new asymmetric choice models. This allows researchers to create new asymmetric models that are tailored to the specific desired qualities.

For my third project, I detailed a micro-economic framework for the interpretation of decision tree models from computer science. At a practical level, motivated by this new framework, I created a bayesian model tree that places a discrete choice model inside a decision tree, and performs a bayesian estimation of this entire system. Substantively, these expansions allow one to capture non-compensatory and context-dependent semi-compensatory decision making amongst individuals. Based on qualitative and anecdotal reports, these effects are expected to be significant for the choice of bicycle commuting. This intuition was supported in the empirical application of our bayesian model trees. The bayesian model tree model had dramatically better levels of fit than a standard MNL model, and the insights and forecasts of the bayesian model trees were far more plausible than those offered by the MNL model.

Methodologically, I merged an existing class of machine learning models (decision trees) with economic theory. This should be greatly helpful in promoting the use and advancement of tree-based models by discrete choice modelers and other researchers that come from a background in econometrics. By relying on tree-based models, I provided a model for very flexible non-compensatory decision rules that generalize those used in the discrete-choice literature so far. At the same time, tree-based models avoid theoretical drawbacks of previous research methods such as the ability to make predictions for all future scenarios. Lastly, I methodologically contribute to the literature on tree-based models by developing the first model that simultaneously allows for estimation uncertainty (through the bayesian estimation) and context-dependent preference heterogeneity (through the model trees).

For this dissertation's final project, I examined travel demand modeling procedures through the perspective of the causal inference literature. Because these two fields have different methods but largely overlapping goals, I was able to use causal inference results to identify inferential practices within the travel demand modeling practices that are likely to be invalid. Substantively, this project resulted in the recognition of the overwhelming need for bicycle demand models to address the issue of unmeasured confounding. Without doing so, analyses and plans based on bicycle demand models that are estimated from observational data should be viewed with suspicion.

Methodologically, this project pointed toward a number of causal inference practices that could immediately benefit the field of travel demand modeling. Such practices include the unambiguous stating of causal assumptions for a given study. These statements should include, but not be limited to, drawing causal diagrams. Moreover, helpful causal inference practices would also include techniques such as making greater use of models with fewer assumptions (e.g. non-parametric models) and checking the validity of one's model predictions using techniques such as before-after studies, natural experiments,difference-in-difference studies, and randomized controlled trials. At the same time, this project uncovered ways that travel demand modelers could contribute to the causal inference literature. Such ways include taking the lead on studies of causal transportability, studies including feedback processes, and the representation of analytic uncertainty in all its forms. These are all topics where transportation researchers have either done much work or they are working conditions that travel demand modelers are forced to face, for better or for worse. By tackling the causality issues inherent in such studies, travel demand modelers can increase the validity of their own work and improve the state of causal inference knowledge overall.

## 3 Research Limitations

Overall, my research efforts in this dissertation suffer from two main drawbacks. First, many of the new techniques require ad-hoc decisions on the part of the researcher. For instance, the zone-of-likely-travel requires a researcher to choose a percentile between 0 and 100 to approximate the distribution of how far cyclists are willing to deviate from their shortest paths. To develop new asymmetric choice models, my procedures require an individual to choose a loss function from which the new model will be derived. There are few principled guidelines for choosing a specific loss function. As a final example, my proposed estimation procedures for bayesian model trees required the choice of a small number of individual decision trees to represent the full posterior distribution of trees. In this dissertation, I chose the trees in an ad-hoc manner that, while practically successful in the current research setting, does not provide guiding principles that are proven to be successful in more general settings.

Secondly, the methods developed in this dissertation require a substantial effort for use in new settings. For instance, an analyst seeking to use my zone-of-likely-travel ideas may have to program their own routines to construct the zones. The routines used in this research were "minimally-viable-products" that worked for my contexts (i.e. the cities being used in my research, the origin-destination pairs observed in my datasets, etc.). However, my routines are neither robust nor computationally efficient. Likewise, for any asymmetric choice models developed using this dissertation's techniques, beyond the four that I have already created, one will have to program the models without being able to rely on existing implementations. Given that "the devil is in the details," the way that a new researcher implements the estimation code for the new asymmetric models may have a significant effect on the apparent success of the new model. Finally, the bayesian model trees that I created in this dissertation require numerous choices of priors, of choice model "kernel" (I used an MNL model, but one can use any discrete choice model more generally), and of the type of decision tree being used. Since these different choices can lead to vastly different code to implement such models, new researchers will have to perform such implementations themselves. Such implementation may seem daunting, and may therefore discourage the most appropriate and flexible use of these new techniques.

Despite the points above, one could argue that the obligation of researchers is to produce new ideas, not necessarily to produce commercial-grade implementations of these ideas. In this sense, the aforementioned issues are unfortunate realities of my work, as opposed to limitations. However, my opinion is that since researchers are often funded by the public (through government grants), researchers should strive to make their work as useful as possible to the public. This means making research as easy to use by others as possible. Accordingly, the next section begins with this goal, recommending future research directions that (1) address the limitations mentioned in the preceding paragraphs and (2) leverage the work in this dissertation to address problems that I believe are of the utmost importance for discrete choice modelling in general, and for bicycle demand modelling in particular.

## 4 Recommendations for Future Research

As noted above, this section will make two general types of recommendations for future research. The first type of recommendation will focus on how to improve the techniques created in earlier chapters. The second type of recommendation will focus on new issues, beyond those that were investigated in this dissertation.

Lastly, in addition to the recommendations below, each of the dissertation chapters contains its own set of further research directions (many of which are unique from those described below).

Now, beginning with the bicycle focused research, it is likely that some analysts will encounter difficulties when constructing zones-of-likely-travel due to computational speed or due to the complexities of real, digital roadway network files. Future research efforts should therefore focus on creating computationally efficient and robust[1] procedures for creating the zones-of-likely-travel. This would enable individuals to reap all the theoretical benefits of this method's approach to incorporating roadway-level variables into travel demand models.

Next, future research should investigate the relationship between intrinsic features of one's dataset and the type of shapes that one's asymmetric choice model should have in order to generate the most accurate predictions. Such research would guild the selection of loss functions from which one can derive the desired choice model, and it would help analysts who, reasonably, may not be familiar with the various types of loss functions in existence. Similarly, future research should investigate how tools such as automatic differentiation and computational graphs can be used to unify and streamline the implementation of such asymmetric choice models.

Finally, my last recommendation for improving this dissertation's research is that future efforts should create numerically robust, conceptually streamlined, and computationally fast estimation methods for bayesian model trees. Ideally, such estimation methods will be 'one-step' methods as opposed to the three-step methodology used in this dissertation. An estimation technique of this kind would greatly reduce the trepidation that some modellers may experience when thinking of using a new choice model.

To conclude, I will three highlight research recommendations that I believe are more 'forward-facing,' as opposed to being primarily focused on improving this dissertation's efforts. First, one of the most exciting research topics that was not explored in this dissertation is the effect of using asymmetric choice models as the class-membership models in a latent-class framework. Here, the asymmetric choice models would provide information on how difficult it is to switch members of the population from one latent-class to another. If the latent-classes capture behavioral heterogeneities such as differing choice-sets and taste-parameters, then such asymmetric latent-class models may entail drastically different policy implications than our current models. For instance, we may find that radically more aggressive policy changes are needed to switch individuals into a latent market-segment that considers 'active' transportation modes. Alternatively, we may discover that certain latent classes are easily affected by policy changes, thus warranting a greater focus on these groups than would have been given when using a traditional latent-class model.

Secondly, as noted in the chapter on causal inference, unmeasured confounders greatly increase the difficulty of drawing causal inferences from observational data. To date, researchers (both in and outside of discrete choice) have not figured out how to mitigate the effects of unmeasured confounding in general settings. To make matters worse, the presence of unobserved confounding is likely the rule, as opposed to the exception, in observational studies. Given this state of affairs, it behooves discrete choice modelers to attack the problem of modeling in the presence of unobserved confounding. Without progress on this front, it is unlikely that our profession has any hope of performing defensible policy analyses based on observational data.

Lastly, my final recommendation is emblematic of this dissertation overall, as it is both methodological in nature and substantively focused on the issue of bicycle planning. Thus far, most transportation planning agencies act in a "reactive" manner. In the (unfortunately) unlikely scenario that a planning agency has a bicycle mode share goal, the agency will likely first come up with a set of plans for bicycle infrastructure investment. Only afterwards will the agency (at best) check to see whether the plan is expected to increase bicycle demand levels to the desired amounts. I believe that this order of operations should be reversed. Given a budget constraint, agencies should proactively optimize the placement of bicycle infrastructure to maximize the expected bicycle mode share. Methodologically, this requires (1) causally valid bicycle demand models, and (2) optimization techniques that treat demand as the objective function while being able to handle huge amounts of boolean decision variables (e.g. bike lane on this street or not). Such a collaboration between optimization experts and discrete choice modellers has yet to be realized, but to best manage public funds, this type of joint research should be a priority for future efforts.

[1] Note that I mean robust to different roadway network configurations.