The Holy Trinity:
Blending Statistics, Machine Learning and Discrete Choice,
with Applications to Strategic Bicycle Planning

by

Timothy Brathwaite

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Civil and Environmental Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Joan Walker, Chair
Assistant Professor William Fithian
Assistant Professor Alexei Pozdnoukhov

Spring 2018

The Holy Trinity:
Blending Statistics, Machine Learning and Discrete Choice,
with Applications to Strategic Bicycle Planning

Abstract

The Holy Trinity:
Blending Statistics, Machine Learning and Discrete Choice,
with Applications to Strategic Bicycle Planning

by

Timothy Brathwaite

Doctor of Philosophy in Engineering - Civil and Environmental Engineering

University of California, Berkeley

Professor Joan Walker, Chair

Every day, decision-makers make choices among finite and discrete sets of alternatives. For example, people decide whether to walk, bike, take transit, or drive to work; shoppers decide which of the available brands of toothpaste to buy; and firms decide which vacant buildings they will rent for office space. Across these disparate domains, discrete choice models mathematically represent the procedures that analysts believe decision-makers are using to make such choices.

Historically, the field of discrete choice modeling grew mainly out of economics, and this lineage has had long-lasting methodological ramifications. In particular, despite the great mathematical similarity between discrete choice models and models in statistics, machine learning, and causal inference, discrete choice research remains mostly siloed, seldom drawing from or contributing to methods in these related disciplines.

In this dissertation, we help demolish the methodological silo around discrete choice research. Drawing from recent techniques in statistics, machine learning, and causal inference, we remove substantive limitations on the decision-making processes that could be be represented and predicted with previously available discrete choice methods. At the same time, by addressing concerns of discrete choice modelers, we make methodological contributions to the fields of statistics and machine learning, and we identify future research areas where discrete choice modelers are well suited to advancing the state of the art in causal inference.

Importantly, the methodological advances described above were not divorced from today's societal concerns. Given that more and more government agencies are (unsuccessfully) attempting to raise bicycle commuting rates in their jurisdictions, we guide our interactions with the statistics, machine learning, and causal inference literatures by trying to more accurately model an individual's choice of commuting by bicycle. In particular, we use parametric link functions from statistics to better model the adoption and abandonment of bicycling. From machine learning, we use decision trees to represent the non-compensatory decision protocols that individuals appear to follow when deciding whether to commute by bicycle,

and we use diagrams from the causal inference literature to gain insight into how we can better model the effects of bike lane investments on bicycle commute mode shares. All together, we not only makes methodological contributions to the fields of discrete choice, statistics, machine learning, and causal inference, but we contribute to the efforts of transportation planners and modelers who are trying to make our cities and regions more sustainable and environmentally friendly. The methods developed in this dissertation have applications to strategic bicycle planning, helping analysts understand when certain interventions are not enough to cause people to abandon non-bicycle modes of travel at the desired rates and what alternative interventions might be more effective.

In total, the specific contributions of this dissertation are the following:

1. We create a new spatial unit of analysis (the zone of likely travel) for the incorporation of roadway-level variables such as presence and type of bicycle infrastructure, roadway slopes, and traffic speeds into mode choice models.

2. We propose and demonstrate the novel use decision-tree methods for directly including roadway-level variables in mode choice models.

3. We create a new class of closed-form, finite-parameter, multinomial choice models that avoid an undesirable symmetry property that we describe in Chapter 3.

4. We create a procedure for using this new class of models to extend many existing binary choice models to the multinomial setting for the first time.

5. We create methods for creating new, symmetric and asymmetric, binary choice models.

6. We provide a microeconomic framework for interpreting decision trees by showing that decision trees represent a non-compensatory decision rule known as disjunctions-of-conjunctions and that such rules generalize many of the non-compensatory rules used in the discrete choice literature so far.

7. We propose and estimate the first bayesian model tree, thereby combining decision trees and discrete choice models in the first two-stage, semi-compensatory model that jointly:

    a) uses disjunctions-of-conjunctions for the choice-set generation stage,

    b) allows for context-dependent preference heterogeneity in the choice stage, and

    c) quantifies analyst uncertainty in the estimated disjunctions-of-conjunctions

8. We identify techniques such as the use of causal diagrams that can be borrowed from the causal inference literature to improve the ability of discrete choice modelers to predict outcomes under external changes or policy interventions such as investing in on-street bicycle lanes.

9. We identify areas of the causal inference literature that can be improved through the incorporation of techniques from discrete choice or through the application of causal inference techniques that are very relevant to discrete choice modellers yet only infrequently researched by traditional causal inference researchers.

Through this dissertation, we empirically demonstrate most of our contributions using commute mode choice data from the San Francisco Bay Area. In every case, we found that the new models developed as part of this dissertation fit our data better than traditional discrete choice models. These results were stable across all measures of fit that were used, whether the measures were in-sample or out-of-sample, frequentist or bayesian. Beyond fit, all of our new models also proved to be qualitatively different than traditional discrete choice methods. Our new models provided insights and forecasts that both made more sense and were more accurate than their traditional counterparts. Finally, our contributions related to causal inference are the only items from the list above without empirical demonstrations. Instead, these contributions are bolstered by substantial literature review, discussion, and thought exercises that show the (general and bicycle specific) benefits of merging discrete choice and causal inference techniques.