# A Rules-based Mode Choice Model using CHAID Decision Trees and Dynamic Transit Accessibility

by

Devin Feng

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Civil Engineering

Waterloo, Ontario, Canada, 2021

# Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

## Abstract

Transportation mode choice models typically represent user decision making using utility-based mode choice models. However, utility models assume that users make compensatory trade-offs between decision variables to maximize their expected utility. The decision process literature raises alternative, non-compensatory theories that suggest people employ simpler, cognitively frugal heuristics in their decision making. Non-compensatory models, including decision tree classifiers, present an opportunity to test the effects of transit accessibility variables on mode choices and improve descriptions of mode choice behaviour. Dynamic forms of transit accessibility, which measure variations in transit service over time, may better capture heuristic perceptions of transit service quality.

This research addresses the need to understand how dynamic transit accessibility (DTA) impacts mode choices, without compensatory decision process assumptions. First, this research develops DTA measures for the Region of Waterloo using General Transit Feed Specification (GTFS) transit schedule information to calculate travel impedance matrices for departures at every 5-minute interval of the day. Zonal mode shares are regressed against alternative DTA measures to analyze the effects of different destination types, time periods of aggregation, and statistical parameters of transit accessibility (i.e., mean and distribution over time). Based on the aggregate mode share predictive performance, a DTA metric is selected for analysis within a binary (transit and not transit) disaggregate mode choice model. Second, this research uses trip diary data to train and score a Chi-squared Automatic Interaction Detection (CHAID) decision tree classifier to represent and predict rules-based mode choice processes. Finally, the selected DTA metric is merged with the trip diary data and applied in another decision tree for comparison. The comparison between the two rules-based mode choice models is based on overall model accuracy, class recall, precision, and interpretability.

Results from the decision tree classifier reveal that users apply heuristics in their transportation mode decision making, including lexicographic and aspiration-level based decision rules. User choices depend primarily on transit pass ownership, and non-transit-pass users consider the trip's distance thereafter. Including DTA as an independent variable in the decision tree has a small but statistically significant effect: users only seem to consider DTA, a generalized location-based measure, if they do not own a transit pass and only after considering the trip-specific distance. Overall, the rules-based mode choice models report accuracies of roughly 84%; however, low precision in the transit predictions (i.e., many false positives) result in an overestimation of regional transit shares.

# Acknowledgements

I am fortunate to have had an abundance of support, challenges, and laughter during my time in graduate school.

Foremost, I thank my supervisors, Professor Chris Bachmann and Professor Dawn Parker, for their patience, guidance, and diverse perspectives on my work. Chris's keen and diverse interests gave me the flexibility to explore concepts in transportation, economics, and behaviour. His attentiveness, willingness to help, and faith in my ability have kept me motivated throughout this work. I will miss our conversations, ranging from the pursuit of acronyms to the inquiry of human behaviours that inspired this research. Dawn introduced me to complex systems thinking and always provided a critical review of my writing and knowledge of modelling concepts. She encouraged me to leverage my skills from my planning undergrad and steered me towards exploring the transit accessibility space. I would also like to thank Professor Jeff Casello for introducing me to the transportation world back when I was in the front row of his infrastructure class, years ago. He encouraged me to pursue quantitative work and wrote the email to get me into Chris's transportation engineering course. Thanks as well to Professor Liping Fu and Professor Chul Min Yeum for reviewing this thesis.

My family and friends have given me the support and security I needed to focus on my learning. To my academic colleagues and especially Jacob, Tina, Jess, and Anjie: thank you for enriching my time at this school with our many shared breaks, projects, bleak office spaces, and transpo-political frustrations. Your domain expertise and perseverance through this academic journey have always impressed me – I am proud to have made such smart friends in graduate school. To my partner, bike friends, friends from undergrad and even those as far back as high school: thanks for watching me grow and finding the interest to join me along the way. Your deep support and respect mean the world to me and have reinvigorated me several times over. Finally, I am grateful for my family for their financial support and patience with my learning.

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1 Introduction

## 1.1 Background

Transportation demand forecasts anticipate how people consume transportation facilities and services. Urbanization combined with ecological preservation pressures generally motivate forecasting as means to maintain accessibility within constrained urban space. However, the financial argument for transportation demand forecasting is perhaps more compelling. Decisions to invest in physical transportation infrastructure rely on accurate demand forecasts to understand infrastructure effects and justify their large direct costs. For example, the Province of Ontario's 2019-2020 economic outlook commits $8.5 billion to transportation infrastructure expenses, representing 58% of the Province's annual investment out of all sectors, including health, education, natural resource, social, and administration. Within the transportation budget, public transit expenses specifically account for $5.5 billion, or 37% of the provincial infrastructure budget (Ontario Ministry of Finance, 2020, p. 13). Governments develop transportation infrastructure seeking returns in the form of increased economic productivity through access to new markets, improved living standards through wage increases, and private sector investment due to reduced business costs (The Centre for Spatial Economics, 2017). The expectation of transit development is that increased transit adoption abates resource inefficiencies related to space, time or monetary costs, and environmental externalities notoriously exemplified by traffic congestion. Transit investment assumes that private automobile users will switch to transit in response to more attractive transit service: a causal relationship central to the investigation of mode choice modellers. Investing in transportation infrastructure requires an understanding of how user behaviours respond to changes in transportation ecology. Governments therefore depend on accurate and defensible mode choice models to reliably invest in sustainable and cost-effective infrastructures.

Mode choice is a component of transportation demand forecasting concerned with how people choose between transportation modes, including driving, cycling, walking, and transit. It is a critical process for understanding how personal choices respond to infrastructure-level changes. Modellers estimate choice outcomes using models that differ primarily by analysis units: aggregate or disaggregate. Aggregate models estimate macroscopic transportation phenomena as the "population demand" of people grouped into geographic zones. Aggregate models were popular in transportation analysis until the late 1970s, when disaggregate models grew popular because computing power increased and modellers realized that aggregate phenomena were the result of many individual decisions (Banister, 2002). Disaggregate models capture behaviour better by estimating decisions at the individual, household, or firm level, thereby eliminating aggregation errors (ecological fallacy) from the zonal homogeneity assumption (Ortúzar & Willumsen, 2011, p. 229). Since predictions based on disaggregate data may be aggregated at more flexible geographies, the approach maintains predictions at the macroscopic level, with which transportation planners are principally concerned. In refocusing analysis to the level at which behaviours occur, disaggregation allows the modeller to interpret decision maker sensitivities to a wider range of transport system and personal attributes.

Disaggregate mode choice modelling was among the first applications of discrete choice analysis (Ben-Akiva & Lerman, 1985, p. 3). Discrete choice analysis examines how individuals choose within a set of discrete and finite alternatives, as opposed to continuous and infinite alternatives, based on the microeconomic theory of consumer choice. Consumer choice theory suggests that individuals are economic agents who possess, observe, and act on objective (demand) functions to maximize self-interests (Ben-Akiva & Lerman, 1985; McFadden, 2000). Lancaster's (1966) interpretation of consumer choice suggests that individuals derive benefits (utility) from choices indirectly through the characteristics of the alternative and the individual making the decision, instead of the alternative in itself. Consumer choice is useful for transforming behavioural assumptions into models of choice because it supports arithmetic expressions of preferences as a function of choice set attributes (characteristics). Constant utility (or Lancaster's indirect utility) is an expression that represents a deterministic decision process consistent with consumer choice. In a simple bivariate example, individual $n$'s utility for alternative $i$ is $V_{ni} = \beta_n x_i + \alpha_n y_i$ where $x_i$ and $y_i$ are decision attributes (independent variables) linearly related to utility $V_{ni}$ (dependent variable) by parameters $\beta_n$ and $\alpha_n$. This deterministic (systematic) function of additive explanatory attributes (e.g., $x_i, y_i$) forms the basis for classical discrete choice models.

Although microeconomic assumptions support a model of behaviours, it is impossible for models to perfectly predict actual decisions. Modellers observe inconsistent choice outcomes between similar individuals and across trips. Probabilistic modelling approaches, including random utility models (RUMs), reason that choice discrepancies may be caused by inherently probabilistic human behaviour or analyst ignorance about the decision process (Ben-Akiva & Lerman, 1985, p. 49). The random utility concept, as opposed to deterministic utility, describes individual $n$'s utility from alternative $i$ in two parts: a systematic component ($V_{ni}$) made of observable independent variables ($x_i, y_i$) and a random error term ($\varepsilon_{ni}$). In its simplest form, the function used in random utility models (RUMs) is shown in [1.1].

$$U_{ni} = V_{ni} + \varepsilon_{ni} \ \forall i, \qquad V_{ni} = \beta_n x_i + \alpha_n y_i \qquad [1.1]$$

$\varepsilon_{ni}$ is a set random variables that are unobserved by $V_{ni}$ because of randomness from any of four sources: unobserved attributes, unobserved taste variation, measurement error, and instrumental/proxy variables (Manski, 1973).

Advancements in mode choice modelling, including disaggregate analysis and RUMs, reveal an underlying effort to better model human behaviour through an economic lens. Disaggregate analysis can correlate user- and trip-specific attributes with discrete choices, and random utility partly accounts for inconsistent choices; however, both maintain the economics-based assumption that utility is a representative decision rule. Aside from its conceptual and formulaic definitions, utility is a class of decision rules that assumes compensatory decision attributes (independent variables) to represent the attractiveness of an alternative using a scalar, linear index called the utility function ($U_{ni}$ in [1.1]) (Banister, 1978; Ben-Akiva & Lerman, 1985, p. 37). The compensatory characteristic appropriately describes scenarios when a user is willing to trade decreases in one good for increases in another. For example, a user may be willing to pay more fare, which reduces personal utility, in exchange for a decrease in travel time, which increases personal utility. The compensatory characteristic of utility models is central to the model's behavioural assumptions because it imposes requirements on how decision variables interact with each other (additively), how decision variables relate to the choice (linearly), and how the decision rule operates (user seeks maximum).

Transportation analysts apply utility models widely in mode choice analysis, accepting the utility decision rule and its behavioural assumptions as convention. The most popular utility model (and by extension, mode choice model) is the multinomial logit (MNL) due to its tractable (closed-form) structure, familiar microeconomic interpretation, and computational practicality (Domencich & McFadden, 1975; National Academies of Sciences Engineering and Medicine, 2018). Analysts apply MNL using travel diary survey data to test different specifications (i.e., combinations of cost, travel time, and other independent variables) and estimate the parameters ($\alpha_n$ and $\beta_n$ of utility function [1.1]) most likely to produce observed mode choices. Analysts use the calibrated utility functions to forecast future behaviours by logistic regression. However, unlike the decision variable (e.g., travel time, income, fare) coefficients $\alpha_n$ and $\beta_n$ embedded in the model, which must be estimated in new contexts, the representation of the decision process itself receives little critique in mode choice modelling practice: transportation users are assumed to behave by maximizing their utility.

Outside of mode choice, behavioural psychologists present alternative theories of decision processes that have limited reception in transportation modelling practice. For example, heuristic decision making (Tversky & Kahneman, 1974) is a cognitively-frugal decision rule that favours a rules-based model framework. Theories that reject the compensatory structure borrow methodological contributions from other disciplines, including machine learning, which presents an application barrier for some transportation analysts. However, if accurate behavioural representation is a concern of transportation analysts, transportation demand models ought to reflect actual decision processes in the mode choice context.

## 1.2 Problem Statement

Popular use of utility models has entrenched assumptions of compensatory decision-making within mode choice modelling practice. However, decision process literature raises alternative decision rules that challenge the utility rule's widespread application. Analysts should consider whether people truly perform compensatory trade-offs between decision variables as utility models suggest. To advance behavioural representation and improve the descriptive power of mode choice models, there is a need to examine the accuracy of decision process assumptions and test the predictive performance of alternative models in the mode choice space.

A non-prescriptive decision structure also presents an opportunity to test transit-related decision variables, investigating whether they influence choices at all, and the process by which users interpret them. Just as mode choice models endeavour for behavioural representation, the decision variables within them should respect decision-making psychology. Adapting decision variables to possible user considerations, including temporally aggregated transit service quality, may improve mode choice model accuracy and descriptive power. Despite the need to understand how transit system variables affect mode choice behaviours, no studies have applied transit accessibility as a decision variable within a non-compensatory mode choice model.

## 1.3 Research Objectives

The goal of this research is to understand how transit system attributes may support mode choice prediction as an explanatory variable in a non-compensatory decision process. The first objective is to build and test a rules-based mode choice model and the second objective is to develop a transit accessibility metric. Third, this research incorporates the transit accessibility metric as a decision variable within the rules-based mode choice model. Analyses of these components are focused on the objectives of urban transportation planning, which aim to identify transit-system-level characteristics that can shift user mode choices from private automobile to transit.

The application of a rules-based mode choice model can identify non-compensatory decision structures and the specific rules or heuristics that result in mode choices. This objective is consistent with a broader goal in transportation modelling, which is to express a description of human behaviour that supports prediction (Ben-Akiva & Lerman, 1985). Relaxing compensatory assumptions may better align mode choice models with theories of behaviour from psychological (decision process) literature. Since rules-based models are not frequently applied in practice, this research aims to produce an interpretable classification scheme using an existing, accessible data source. Increasing the applicability of rules-based models allows analysts to employ them during the demand modelling stages of infrastructure development.

Development of a transit accessibility metric aims quantify transit system variables relevant to mode choice. Specifically, the metric ought to combine the value of mobility offered by the transit system with the land-use-related activities that motivate it. Literature review informs different configurations of transit accessibility that this research tests to find the model that best describes actual user considerations. Since the value of travel varies spatially (where users go) and temporally (when users make the trip), a more representative metric should leverage available datasets and methods to likewise vary across space and time, taking a dynamic form.

Incorporating the transit accessibility metric within a rules-based mode choice model can reveal opportunities to encourage further transit adoption. In utility models, decision variables affect an outcome linearly and additively. Rules-based methods do not impose structural restrictions on the way decision variables affect an outcome and therefore more flexibly accommodate the influence of decision variables. The rules-based model is a useful test platform to understand the impact of transit accessibility on mode choice. Since transit accessibility is sensitive to the operational policies over which transportation planners have influence (e.g., service frequency, capacity, routing), understanding whether and how users interpret transit accessibility can guide government interventions towards those most likely to encourage transit adoption. More broadly, this research aims to support urban transportation system management by finding opportunities to encourage sustainable and efficient mass mobility through transit.

## 1.4 Research Scope

The Region of Waterloo (the "Region") is the application context of this research. The Region was chosen due to the availability of travel diary survey data (2016), transit schedule information, and its computationally manageable geographic size. Every day, the roughly 530,000 residents in the Region make approximately 1,009,897 intraregional trips and an additional 143,298 interregional trips to destinations outside of the Region. Of the intraregional trips, the mode shares are 7.5% transit, 70.2% drive alone, 14.5% passenger, 1.6% cycle, and 6.2% walk (Data Management Group, 2018). Overwhelmingly, trips related to the personal automobile, including drive alone and passenger trips, represent the largest share of trips at 84.7%. Based on the 2016 national census, 277,785 employed individuals lived in the Region (Region of Waterloo, 2016). Mode choice analyses that identify decision structures and account for transit service and land use variables may help planners select tools to reduce the disproportionate share of personal automobile users.

This research is focused on applying a non-compensatory, rules-based mode choice model, where transport users interpret decision variables and use decision rules within cognitive boundaries. The non-compensatory approach uses machine learning methods to empirically induce decision rules from travel diary survey responses. It does not pursue perfect representation of decision rules or structures because analysis is limited to observable variables in the dataset; rather, it searches for the most likely decision process from the training data. An explicit understanding of decision structures and rules may require qualitative decision process surveys (e.g., heuristics survey by Hannes et al. (2009)), which are beyond the scope of this thesis. Instead, this method relies on secondary data sources that are 1) more accessible to the analyst because they are regularly collected, and 2) specific to the estimation context (geographic region). Furthermore, this research does not offer a comparative analysis of rules-based model performance and utility-based models. Other papers, presented in the Literature Review, fill this niche with adequate context. Instead, this research endeavours to demonstrate the practical application of rules-based models in mode choice classification and address some limitations of compensatory models related to behaviour that rules-based models may overcome. Although this research analyzes decision process (behavioural) representation and assesses the predictive outputs of the mode choice model (see "Measures of Effectiveness" subsection), this research ignores some MOEs related to computational efficiency or applicability in practice, limiting its transferability in practice. Potentially relevant MOEs include speed (computational cost) and robustness (accuracy, given noisy or incomplete data) but remain outside the scope of this thesis.

## 1.5 Thesis Structure

Chapters in this thesis review and apply methods for constructing a non-compensatory mode choice model and transit accessibility metric. Chapter 1 describes the motivation for applying a non-compensatory mode choice model and transit accessibility as a decision variable. Chapter 2 presents a three-part literature review of decision process research, the predictive merit of machine learning classifiers, and transit accessibility measures, all within the context of mode choice research.

The next three chapters describe the methods and outputs of this research. Chapter 3 presents the public data sources and manipulations used throughout this research. Chapter 4 develops and analyzes different transit accessibility measures for predicting aggregate mode shares. Chapter 5 trains and tests a rules-based mode choice model using the Chi-squared Automatic Interaction Detection (CHAID) tree-growing algorithm. At the end of chapter 5, selected transit accessibility measures from chapter 4 are introduced as independent variables in the mode choice model. Each chapter discusses the strengths and limitations of the analysis in the context of decision process representation and choice prediction.

Chapter 6 concludes this thesis with a summary of goals and findings, providing direction for further research.

# Chapter 2 Literature Review

This literature review is divided among the two topics that this thesis combines: 1) decision process research (psychology) in mode choice, and 2) transit accessibility measurement. First, decision process theories are reviewed for their consistency with mode choice modelling conventions. Next, the machine learning research space is explored to support the application and testing of some behavioural theories. Finally, this review examines the different forms of transit accessibility, their applications, and their theoretical strengths and limitations related to transportation mode decision making.

## 2.1 Decision Processes

This section of the literature review focuses on theories and models of mode choice, considering cognitive and perceptual limits that suggest people do not pursue compensatory decision strategies. Instead, users may rely on non-compensatory, heuristic strategies. Exploring non-compensatory theories of choice can further the development of mode choice and transit accessibility models.

Decision process research is concerned with how choice outcomes (behaviours) are connected to psychological theory. Theorists in and outside of travel demand research raise ideas about how people make decisions, which make assumptions about the components of the choice process: problem definition, generation of alternatives, evaluation of alternatives' attributes, decision rule, and then implementation (Ben-Akiva & Lerman, 1985, pp. 31–32). Researchers often test conjectures about the decision rule component, given its implications on possible modelling structures and interesting insights about human psychology. Decision rule effectiveness seems to be sensitive to antecedent choice process components, including attribute evaluation. Thus, more complete theories of decision making describe assumptions about decision components that are guided or imposed by the decision rule component.

Four decision rule classes are raised in travel demand modelling literature: 1) dominance, 2) satisficing/aspiration levels, 3) lexicographic rules, and 4) utility (Ben-Akiva & Lerman, 1985, p. 35; Timmermans, 1983). Dominance suggests that users choose the alternative that is no worse across all attributes and is better in at least one attribute compared to other alternatives. Satisficing assumes some critical value (level of aspiration) for every attribute and that only an alternative that exceeds these levels can be selected. Lexicographic rules involve ranking alternatives' attributes by importance and selecting the best alternative for the most important attribute. If there is a tie, the next most important attribute is considered in sequence until an alternative is selected. Finally, utility involves forming a scalar objective function (index) of attractiveness for alternatives (the utility function, represented in [1.1]), and the user selects the alternative with the highest index value.

### *2.1.1 Utility Decision Rule*

Mode choice modellers often employ utility as the decision rule in travel demand modelling, specifically within logistic regression models (e.g., MNL, Mixed MNL, and nested logit). Utility is advantageous because of its familiar microeconomic basis, computational practicality, and elegant closed-form structure (Domencich & McFadden, 1975). Its objective function is easy to interpret, offering high explanatory power and the ability to derive elasticities. In application, logistic regression models exhibit consistently high predictive accuracy on testing datasets (samples held out during model calibration) but vary according to context. MNL model accuracy is the proportion of predicted choices that are correctly classified (true positives and true negatives). This research uses accuracy as one measure of effectiveness for model selection (see "Measures of Effectiveness" subsection). Applied MNL models often report overall accuracies around 60-70%, including 63.02% (Cheng et al., 2019), 64.7% (Zhao et al., 2020), 66.3% (G. Wets et al., 2000) and 70.5% (D. Lee et al., 2018).

Despite its predictive successes, some authors suggest the utility rule ought to be refuted because its representativeness is contingent on users optimizing decision process components related to information search and processing (i.e., defined goals, stable decision context, exhaustive alternatives sets, and exhaustive attribute evaluation) and adoption of the utility rule therefore assumes a rational person (Ewing, 1973; Klein, 2002, p. 109). Herbert Simon's (1955, 1972) bounded rationality theory submits that people are incapable of the collection and assimilation required for classically rational behaviour, which is the consistent and transitive ordering of preferences (Ben-Akiva & Lerman, 1985, p. 38). Simon cites mental computational constraints, complex environments, and incomplete information about alternatives for refuting models of classical rationality, like Luce's strict utility (Luce, 1959). Utility's widespread adoption nonetheless has led authors to criticize its imposition of rational decision processes and obsession with model specification over the basic theories of travel and decision making (Banister, 2002; Innocenti et al., 2013).

Although utility models would perform better if classical rationality were true, RUMs respond to bounded rationality's criticisms without altogether abandoning utility. RUMs, which include logistic regression models, are considered behavioural models because they introduce error terms (i.e., $\varepsilon_{ni}$ in [1.1]) to relax assumptions about user decision processes and account for choice discrepancies (irrational behaviours) (Ben-Akiva & Lerman, 1985). Representations of choice discrepancies stem from psychological theorist L.L. Thurstone's (1927) Law of Comparative Judgement, where stimuli (and its derivative benefit) is perceived with a normally distributed error that Thurstone calls discriminal deviation. Marschak (1960) adapts this psychometric approach to economic analyses, exploring what he calls RUMs, which Manski (1977) later formalizes (see Ben-Akiva & Lerman, 1985, pp. 55–57). RUMs are more consistent with bounded rationality compared with strict utility.

Unfortunately, the introduction of error terms may itself impose statistical assumptions (related to error distributions), and does not address the structural (compensatory) limitations of the deterministic component of utility (i.e., $V_{ni}$ in equation [1.1]). Timmermans and Golledge (1990) describe some assumptions of RUMs that can imperfectly represent decision processes:

- behaviour is compensatory,
- the independence from irrelevant alternatives (IIA) property for MNL,
- introducing new alternatives cannot increase choice probabilities for existing alternatives, and
- user indifference towards alternatives do not affect choices (indifference thresholds).

Some of these assumptions are logical consequences of others, and primarily stem from strict statistical assumptions (Wang & Ross, 2018). The next section discusses the behavioural consequences related to the statistical assumptions of RUMs, followed by assumptions of the compensatory model structure.

### *2.1.2 Statistical Assumptions*

RUMs can account for behaviours unexplained by systematic (deterministic) predictors; however, some RUMs introduce assumptions about the error distribution and error term placement within the utility function that can weaken behavioural representation. For example, MNL estimates choices between alternatives by assuming each error term ($\varepsilon_{ni}$) from equation [1.1] is identically and independently distributed (iid) Extreme Value Type I (Gumbel) (Domencich & McFadden, 1975). In developing MNL, McFadden (1974) found that logit is consistent with the RUM concept of choice probabilities if and only if the error terms are iid Extreme Value Type I. Notably, Luce (1959) originally derived the logit formula based on the assumption that discrete choices are characterized by the independence from irrelevant alternatives (IIA) property, whereas McFadden found IIA to be a resulting property of iid RUMs (Domencich & McFadden, 1975, p. 69; McFadden, 2000). The iid error assumption enables the probabilistic form of MNL: the probability of choosing alternative $M_i$ ($\Pr(M_i)$) depends on its random utility $U_{ni}$ relative to all other alternatives, or that $\Pr(U_{ni}) > \Pr(U_{nj})$ for all $M_j \in M$ (equation [2.1]).

$$Pr(M_{ni}) = \frac{e^{U_{ni}}}{\sum_{M_j \in M} e^{U_{nj}}} \quad [2.1]$$

A limitation of the MNL, and standard logit models in general, is that it represents *systematic* taste variation ($V_{ni}$ value varies by $\beta_n$ and $\alpha_n$) but does not account for *random* taste variation between households ($\beta_n$ and $\alpha_n$ parameters themselves would vary randomly) (Train, 2002, p. 44). To vary tastes randomly, additional random variables $\rho_i$ and $\eta_i$ would enter utility $U_{ni}$ grouped in a new error component $\tilde{\varepsilon}_{ni}$, such as in equation [2.2].

$$U_{ni} = \beta_n x_i + \alpha_n y_i + \tilde{\varepsilon}_{ni}, \qquad \tilde{\varepsilon}_{ni} = \rho_n x_i + \eta_n y_i + \varepsilon_{ni} \quad [2.2]$$

Introducing new random error terms $(\rho_i, \eta_i)$ within independent variable parameters would necessarily correlate the errors over alternatives in set $M$. As a result, random taste variation violates the iid assumption (Train, 2002, pp. 42–47).

The IIA property also imposes assumptions of proportional shifting in choice behaviour, a limitation famously exemplified by the red-bus-blue-bus paradox. The red-bus-blue-bus paradox describes a choice scenario where the proportion of choices between two equivalent alternatives such as driving, D, and taking a red bus, R (i.e., $\Pr(U_{nD}) = \Pr(U_{nR})$, predicting equal probability of choosing either mode), is maintained despite the addition of a blue bus, B, with identical characteristics to the red bus. As a result, choice probabilities are split evenly (33% for each mode $i$ because $\Pr(U_{nD}) = \Pr(U_{nR}) = \Pr\,(U_{nB})$ and $\sum_i Pr_n(i) = 1$, despite the blue bus being more likely to cannibalize red bus shares than automobile shares. The full mathematical explanation for this transitive property is widely available ((Ben-Akiva & Lerman, 1985, pp. 50–52; Luce, 1959; Train, 2002, p. 46)). Other models respond to these limitations, including mixed multinomial logit (MMNL) and nested logit, but receive less widespread application. MMNL can accommodate random taste variation in the $\beta$ parameters by discarding the assumption of iid errors. MMNL defines a more general class of models where MNL is a special case with no variance within the $\beta$ parameters (McFadden & Train, 2000). Nested configurations of MNL subdivide the alternatives set to address scenarios where proportional shifting between some alternatives are unrealistic.

Issues related to systematic taste variation (from distributional assumptions) and transitive substitution patterns (from the IIA property) require more advanced discrete choice models (e.g., MMNL, nested logit) to reconcile consumer choice theory with human behaviours. However, these advancements maintain the assumption that utility is a representative decision rule.

### *2.1.3 Structural Assumptions*

Utility is considered a compensatory rule because its additive combination of decision attributes imply that losses in one decision attribute may be compensated by gains in another decision attribute (Ewing, 1973). For example, the user may accept the alternative with a longer travel time if incentivized by a reduction in fare: cost savings in one attribute offset cost increases in another. Behaviourally, this is an important assumption of the utility rule because it precludes non-linear interactions between explanatory variables. Conditional (if-then) and threshold effects cannot be represented (D. Lee et al., 2018) and dominance effects are ignored. Heterogeneous decision rules and biases from habit formation (inelastic behaviours) also remain unexplained by compensatory rules (Innocenti et al., 2013). Compensatory assumptions also have important implications on policy. Foerster (1979) notes that if mode choices were compensatory, incentives in carpooling would be just as effective as disincentives to driving alone. If all people in all situations applied this strategy, as implied by the blanket application of utility in mode choice modelling, incremental policy interventions would effect changes in behaviour. No authors outright reject that the compensatory strategy is a plausible description of behaviour; however, that utility is adopted so widely has prompted authors to consider alternative decision rules and theories.

### *2.1.4 Alternative Theories of Decision Making*

Researchers in psychology have raised alternative decision process theories in response to utility's compensatory trade-offs, statistical assumptions, and sensitivity to irrational (i.e., contradictory, non-transitive) decisions. Simon's (1955, 1972) recognition of bounded rationality cast doubt on compensatory rules because of their consistency with rational decision making and the non-compensatory research paradigm emerged (Gigerenzer & Gaissmaier, 2011). Non-compensatory decision process theories do not allow trade-offs between the decision attributes of choice alternatives; instead, decisions are assumed to be made on "an attribute-by-attribute" basis because separate utilities (Lancaster's attribute-derived value) are not combined into a single value (Timmermans, 1983). Dominance, satisfaction, and lexicographic rules are employed within non-compensatory decision processes.

Heuristics describe cognitively frugal decision processes that involve search rules, stopping rules, and decision rules (Tversky & Kahneman, 1974; van der Pligt, 2015). Although different definitions of heuristics exist (e.g., heuristics assess targets by substituting another attribute; or heuristics describe all effort-reduction strategies, including compensatory strategies), this research defines heuristics as a class of decision process theories that employ non-compensatory rules (Jensen, 2016). Alternative terms such as procedural rationality, bounded rationality, and non-rational denote the same class of decision processes as heuristics (Gigerenzer & Gaissmaier, 2015). Generally, heuristics are associated with detrimental effects on decision quality (van der Pligt, 2015); however, some authors argue that heuristics can lead to economical decision making without a detailed deliberation in the decision process, especially for repeated decisions or habits (Aarts et al., 1997). Some specific heuristics include representativeness, availability, if-then-else (Hannes et al., 2009), take-the-best (dominance-based), and satisficing (aspiration levels based). In terms of decision process components, the representative and availability heuristics describe simple search and stop rules, whereas if-then-else, take-the-best, and satisficing heuristics are frugal decision rules.

Tversky and Kahneman popularized the representativeness (stereotyping/similarity) heuristic and availability (or frequency) heuristic in their studies of preference and classification under uncertainty and risk (1973, 1974). In contexts where the outcome of a decision is uncertain with respect to some criterion, the representativeness heuristic suggests that people evaluate the probability of an outcome by the degree to which it resembles another event/outcome from their experience. Therefore, if travel time delay is highly representative of transit travel from their own experience, the probability that delays originate from transit travel is judged high. Any subsequent decision rule would be biased against transit travel. The availability heuristic suggests that people evaluate the probability of an outcome based on how easily relevant instances come to mind. Since repetition strengthens association, a choice for which "good" outcomes are observed more frequently is evaluated more favourably. Within their experiments, authors presented subjects with information that triggered these heuristics and observed that subjects consistently neglected prior information (given). Heuristics bias the evaluation of alternatives in predictable directions. Tversky and Kahneman (1981) describe the concept of decision frames – the decision maker's conception of acts, outcomes, and contingencies (risks) related to choices – to relate perceptual biases from search processes to choices. Their experiments demonstrated shifts in choices for problems related to monetary risk or loss of human life after otherwise inconsequential changes in the descriptions (the "framing") of each problem.

Heuristics are difficult to transfer because they ought to be inferred for each choice context; that is, heuristics raised in experiments of monetary risk do not necessarily apply to transportation mode choice. Early behaviour-oriented transportation researchers reflected on the difficulty in testing the validity of non-compensatory strategies because of decision processes must be inferred from observations (Timmermans, 1983). Within mode choice research, heuristic decision making is often related to the formation of mode choice habits, resulting from adaptation to uncertainty over time. Mode "stickiness" or habituation causes observable inelastic behaviours that disregard incremental changes in alternatives' attributes that would have an effect within compensatory models (Aarts et al., 1997; Innocenti et al., 2013). More recently, supervised machine learning models have presented a means to infer how people may make non-compensatory choices.

## 2.2 Supervised Machine Learning

Supervised Machine Learning (ML) models demonstrate high prediction accuracy and embody structures consistent with heuristic decision process theory, offering mode choice modellers a promising methodological alternative to utility-based models. This section explores supervised ML literature for classification algorithms that may represent the non-compensatory decision processes proposed in the decision process literature. Within ML research, different terms describe the inputs, outputs, and processes of a model. Features refer to independent variables (e.g., age), and dimensions are the unique categories or values of a feature (e.g., 11-19, 20-30 years). Target variables refer to dependent variables (e.g., primary mode of a trip), and classes (e.g., transit, car, cycle, walk) are the categories of a target.

Supervised ML algorithms are model induction ("learning") processes using training data whose classes (e.g., chosen transportation mode) are known – hence, supervised – by the modeller, as opposed to unsupervised processes which lack target variable measurements (Hastie et al., 2008). Supervised ML methods can be segmented by their learning/induction tasks, which include regression and classification. Generally, problems with quantitative target variables are regression problems, whereas problems with qualitative (categorical, discrete) target variables are classification problems. Selection of a task and specific method depends on the problem at hand. For the classification task, model induction is the development of classification rules that can determine an observation's class (e.g., chosen transportation mode) from the dimensions of its features (Quinlan, 1986).

### *2.2.1 Decision Tree Classifiers*

One method of supervised ML is decision tree (DT) classification. DTs use splitting criteria, which are analogous to decision rules, to recursively partition the training data into mutually exclusive, increasingly homogeneous regions. The algorithm may be illustrated using the inverted tree analogy presented in Figure 1.

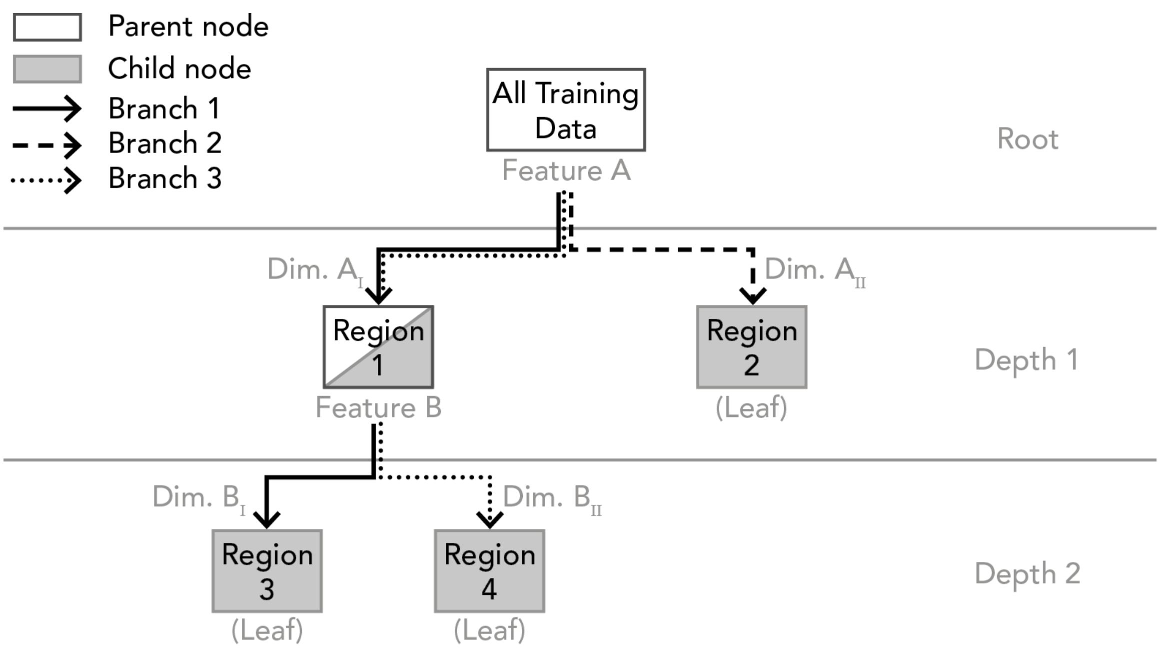



*Figure 1: Decision Tree Classifier Components (Generic, Binary Splits)*

James et al. (2013) describe a generic DT algorithm and its components using terminology from the inverted tree analogy. Each node along the tree is a region of the training data, characterized by the target classes therein and delimited by feature dimensions. The root node describes the entire predictor space, including all training data and their attributes (feature dimensions) used for tree induction. Tree depth relates to the number of splits performed on the data and provides an organizational structure for partitioned regions. Regions where splits occur are parent nodes, whereas regions resulting from splits are child nodes. DT classification algorithms use different splitting criteria to partition the predictor space. Creators of each algorithm describe splitting criteria in detail. Sutton (2005), Han (2011), and software documentation from IBM (2011) offer higher-level summaries. Splits form mutually exclusive pathways, called branches, which describe the sequence of if-then-else rules or heuristics leading to a terminal region. Terminal regions, or leaf nodes, describe child nodes when some stopping criterion is met and no further splits are allowed.

Generally, DTs are useful classifiers because they can handle numeric and categorial predictor variables, alongside missing values (assume a "missing" dimension for features where data are incomplete). In their ML textbook, Hastie et al. (2008) also credit DTs with their insensitivity to scaling issues (unaffected by monotonic variable transformations), immunity to outliers, and high predictive accuracy across disciplines. In their comparison of CHAID, MNL, and C4 performance in mode choice modelling, within transport studies, Wets et al. (2000) also discuss the insensitivity of DT algorithms to outliers in travel diary data. However, compared with regression techniques, DTs may not extrapolate beyond the range of training observations (Breiman et al., 1984). DTs are also dependent on comprehensive testing data (outside of the training dataset, used for prediction), since predictions using the tree structure are susceptible to missing elements (e.g., parent node links, variables). Since DTs grow by splitting the training dataset, trees with higher depths also increase the data requirements to populate representative classifications; overgrowing a tree could cause overfitting to the training data and reduce generalizability.

DTs are powerful tools for inducing and representing heuristic decision processes. DTs are non-parametric models because they do not assume a functional form or shape. In contrast, parametric models reduce the model estimation process to that of estimating a parameter set, such as by stepwise variable selection. The lack of functional form in DTs relaxes the decision process assumptions of the compensatory model, allowing the induction of non-compensatory rule sets. James et al. (2013, pp. 22–23) address this issue directly, referring to model flexibility as an advantage of non-parametric approaches for modelling phenomena where the functional form is unknown. Wets et al. (2000) argue that DT models are "theory-free" because they induce their structure from data. DTs are thus structurally consistent with ideas of non-compensatory, rules-based decision making.

This research reviews five common DT algorithms (i.e., CART, C4/4.5/5, QUEST, CHAID, Random Forests) for their advantages and disadvantages in mode choice classification, considering both theoretical limitations and constraints to application. The most straightforward DT algorithm is classification and regression trees (CART) (Breiman et al., 1984). CART refers to DTs that create binary splits, where parent nodes can only split into two child nodes, using measures of impurity as the splitting criteria: Gini impurity or entropy. Gini impurity indicates the goodness (rather, "badness") of a split by the proportion of a region that belongs to a class (e.g., transit) after a split. Each split minimizes the impurity (maximizes purity) within resulting child nodes, where a completely pure node includes training observations from only one class. Entropy employs a similar purity measure based on information gain (see Geert Wets et al. (2000) for a formulaic summary). CART features and classes can be numeric or categorical, but popular application libraries such as Scikit-learn first require that categorical variables be encoded into dummy (flag/interval) variables for CART to perform binary splits (Pedregosa et al., 2011). One limitation of CART is its interpretation: high-dimensionality features present many opportunities for splitting. Since CART only performs binary splits, these training data can induce deep, narrow trees that complicate interpretation. Deeper trees are also prone to overfitting on training data, and thus require pruning to reduce dimensionality and improve out-of-sample accuracy. CART uses "cost-complexity pruning", which eliminates nodes as a function of the misclassification (i.e., error) rate and the number of leaves in a given subtree through a bottom-up search (Breiman et al., 1984; Han et al., 2011).

Other popular DT classifiers improve computational performance using purity-based measures, like CART. One proprietary algorithm is Quinlan's Iterative Dichotomizer 3 (ID3) and its family of descendants C4, C4.5, and C5 (Quinlan, 1986, 1993). Quinlan's algorithms also use the information gain (entropy) splitting criterion, which partitions the training data using the feature that maximizes the information gain value (Quinlan, 1993). Han et al. (2011) provide a succinct summary of C4.5's pruning procedure. The more recent C5 algorithm focuses on computational improvements, but Quinlan's algorithms differ from CART primarily by allowing multicategory splits. Loh and Shih (1997) developed the Quick, Unbiased, Efficient, Statistical Tree (QUEST) algorithm which, as its authors demonstrate, has a lower computational cost and variable selection bias than CART. However, the QUEST algorithm also produces binary splits and reports similar accuracies and tree sizes to CART after pruning.

Chi-squared Automatic Interaction Detection (CHAID) is a DT algorithm that uses the chi-square statistic to determine splits (and resulting partitions) at decision nodes for categorical dependent variables (continuous dependent variables use the F-test). Kass (1980) developed CHAID to manage categorical variables using statistical tests to perform multiway splits. At every decision (parent) node, the CHAID algorithm applies the chi-squared test of independence, which finds the least significantly different (most homogenous) dimension(s) of a given feature with respect to the target classes. CHAID's statistically-significant splitting yields interpretable confidences and functions as an internal feature selection mechanism, requiring less data preprocessing (IBM Corporation, 2012). CHAID's splitting criterion also handles qualitative predictors without creating dummy variables, increasing application ease and interpretability (Kass, 1980). The confluence of multiway splits and categorical feature interpretation results in wider, more shallow trees than CART. Since this DT algorithm is chosen for application in this thesis, additional detail is available in the Splitting Criterion section.

Random Forests (RF) refers to an ensemble of CART trees, given its developmental history with CART as its basis (Breiman, 2001). RF represents a distinct category of models, called "ensemble" methods, where many decision trees (an ensemble) are generated and then aggregated to improve predictive performance. RF builds its trees using training samples that are drawn randomly with replacement (bootstrapping); that is, the same training observation may be drawn multiple times. As another source of randomness, RF also randomly selects a subset of features (independent variables) to train each tree. The splitting procedure for tree construction is therefore not the best split for all training data, but the best split for a random subset of the training data. For every tree grown, the withheld testing data are scored (applied to trained trees) and results are averaged to get the final prediction, resulting in high variance and low bias. The primary drawback of using ensemble methods is that the final prediction rules appreciate in complexity alongside the reduction of error (Sutton, 2005). Despite its complex interpretation, RF has been applied within mode choice literature with high predictive accuracy (Cheng et al., 2019).

### *2.2.2 Decision Trees and Mode Choice*

Researchers have successfully applied rules-based mode choice (RBMC) models using different machine learning algorithms, yielding similar predictive accuracies to logit models. In this research, rules-based models refer to the decision tree algorithms applied in mode choice to represent heuristic decision processes (i.e., if-then-else rules). Compared with conventional discrete choice methods (logistic regression), RBMC models have shown comparable and sometimes higher predictive accuracy. In their early application on activity diary data (full day travel diary, 3 modes) in Denmark for the ALBATROSS model system, Wets et al. (2000) found that overall prediction accuracies of CHAID (66.7%) and C4 (63.5%) were comparable to MNL (66.3%). Xie et al. (2003) modelled commuter mode choices (5 modes) in San Francisco with a C4.5 DT and compared it with a MNL model. On testing data, their MNL model had a 72.9% accuracy, whereas the DT had 76.8% accuracy. Lindner et al. (2017) modeled binary mode choices (transit and private automobile) in São Paulo using CART and found that it yielder greater (79.2%) accuracy than their logit model (73.9%) when applied to their test set. More recently, Zhao et al. (2020) compared the accuracy of ML models, including CART (59.3%) and RF (85.6%), with MNL (64.7%) and mixed logit (63.1%) for 8,141 testing samples. Notably, the Zhao et al. study's context has a uniquely low automobile share (14.89%) because it was sampled from a university campus (University of Michigan). Rasouli and Timmermans (2014) tested the impact of ensemble size in RF predictions of mode choice, reaching approx. 65% accuracy (35% error, stable after 20 trees). Cheng et al. (2019) also applied RF in mode choice and found a 85.36% accuracy (stability after 200 trees), compared with their MNL model's 63.02%. However, these authors continue to stress the difficulty of interpreting ensemble methods due to the inability to derive elasticities. The common adoption of RT suggests that some researchers are willing to trade off interpretability for higher prediction accuracies, sacrificing the descriptive power of their models.

## 2.3 Transit Accessibility

Accessibility is a performance indicator for the land-use transportation (LUT) system (Boisjoly & El-Geneidy, 2016). Originally defined by Hansen (1959), accessibility is a measure of the spatial distribution of relevant activities, adjusted for travel impediment. Although this research is specifically concerned with transit accessibility (i.e., travel impediment as experienced through use of public transportation), an understanding of general accessibility is useful. Accessibility measures include two major components: an attractiveness component, such as the number of activities (i.e., jobs, shops, and services) at a destination, and a travel impedance factor, which is often a function of distance or time (A. M. El-Geneidy & Levinson, 2006; Nassir et al., 2016). Since travel is uncertain in that users do not necessarily respond to destination attractiveness, probabilistic terms would define attractiveness as the number of opportunities (potential activities) at a destination and travel impedance as a decrease in the probability that the user is attracted to the destination. Accessibility reflects the core theory of transportation demand theory because attractiveness presents a motivation for trip-making, whereas the impedance factor reflects the inconvenience associated with travel (Handy & Niemeier, 1997). Other definitions of accessibility, including those concerned with barriers to persons with disabilities are not the focus of this thesis. Furthermore, accessibility is distinct from connectivity and mobility concepts. Connectivity refers to the existence of transit service between origins and destinations, ignoring the friction of travel. Mobility refers to the ease of making such movements but disregards the attractiveness of the trip. Accessibility incorporates the value of opportunities at destinations with mobility, capturing the value of a trip and its ease in a composite measure (Beimborn et al., 2003; Boisjoly & El-Geneidy, 2016). Accessibility is a useful concept to describe the large-scale interaction between urban land use and transportation systems, especially to non-experts (Straatemeier & Bertolini, 2019).

Figure 2 presents the urban system as a feedback cycle, where the transportation system varies the extent of accessibility to land uses across space. Land uses facilitate activities that generate transportation demand, thereby prompting further transportation service provision, and so forth (Guiliano & Agarwal, 2017; Meyer & Miller, 2001; Wegener, 1995).

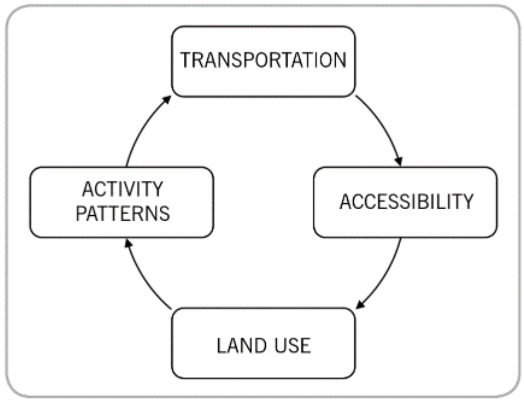



*Figure 2: "The urban system" adapted* (Guiliano & Agarwal, 2017)

Planners and governments can affect the LUT system by intervention to either the land use or transportation components. However, since land uses change more slowly, transportation agencies often hold land use changes constant and affect accessibility outcomes through interventions in the transportation system (Levinson & Krizek, 2005). For example, small operational changes to bus transit, including new branches extending from an existing route, can increase accessibility to destinations along the new corridor in the short term and induce new activity patterns in the longer term. Accessibility can explain the long-term evolution of urban structure, and it also influences a location's value. Recognizing the effects of transit accessibility is crucial towards improving life satisfaction through positive perceptions of transit accessibility, improving public health, and increasing engagement in social activities (Saif et al., 2019).

### *2.3.1 Classes of Accessibility*

In the transportation literature, four classes of measures attempt to quantify the value of accessibility: gravity-based, cumulative opportunity (isochronic), space-time (infrastructure-based), and utility-based (econometric) (A. M. El-Geneidy & Levinson, 2006; Handy & Niemeier, 1997). Based on the original formulation of accessibility, each measure describes the relative accessibility of an origin area to an activity type (e.g., employment) at a destination area. Thus, total accessibility to employment from an origin area is the sum of the accessibility to all individual destination areas around the origin (Hansen, 1959). The most common accessibility measures are the gravity-based and cumulative opportunity measures because of their interpretability and consistency with the original, location-based concept of accessibility. Space-time and utility-based accessibility measures are person-based. They address the socioeconomic dependencies and activity-based constraints that affect the directionality, scheduling, and ultimately, the value of accessibility. Person-based accessibility measures expand the definition of accessibility to include four components: land-use (attractiveness), transport (impedance), temporal, and individual components (Geurs & van Wee, 2004).

Gravity-based ("spatial interaction") measures reduce the weight of opportunities by the travel impedance. Alluding to the Law of Universal Gravitation, they define the attractiveness compelling travel between two physically separate points as a product of the weight of opportunities at the origin and destination locations (Roy & Thill, 2003). Using an exponential impedance function, the gravity measure can capture the influence of all opportunities in any given study area and the steep decrease in willingness to travel over longer distances. Gravity-based accessibility is a location-based measure because it represents the number of opportunities accessible from one location. Gravity-based measures are practical for application because their location-based outputs are potentially compatible with local travel forecasting models, which store information at the zonal level (Krizek et al., 2009). A limitation of the gravity measure is that it fails to recognize user-specific perceptions of attractiveness (taste heterogeneity), because it assumes that accessibility is the same for all users accessing a destination, although actual perceptions and travel contexts may vary (Cascetta et al., 2013; Hanson & Schwab, 1987).

$$Gravity\text{-}\ based\ Accessibility_i = \sum_{\forall j \neq i} \frac{O_i * D_j}{TT_{ij}^2} \qquad [2.3]$$

The travel impedance function in gravity-based accessibility often uses the squared travel time ($TT_{ij}^2$) between origin $i$ ($O_i$) and destination $j$ ($D_j$), following a denominator sometimes used in the trip distribution gravity model for demand forecasting (Niemeier, 1997). More recent travel forecasting models typically calibrate the travel impedance function (e.g., using power functions, $f(TT_{ij}) = TT_{ij}^{-b}$; exponential functions, $f(TT_{ij}) = e^{-bTT_{ij}}$; or gamma functions, $f(TT_{ij}) = (TT_{ij})^a e^{-bTT_{ij}}$, where parameters $a$ and $b$ are calibrated) using household travel diary data. However, transit accessibility studies continue to adopt the power-based impedance function, where $b = 2$. The calibration of travel impedance within transit accessibility measures is left for further research.

Cumulative opportunity measures, also described as count-based or isochronic measures, represent the simplest form of accessibility. Cumulative opportunity measures count the number of opportunities (at destination zones $j$, among set of all destination zones $J$) reachable from origin $i$ within a time threshold ($\alpha_{ij}$). Based on the gravity-based measure definition, the cumulative opportunity measure is a specific subset of gravity-based measures where the travel impedance has a value of 1 if the destination is reachable within a travel time threshold, $\beta$ (e.g., $\beta$ = 30 minutes), and 0 otherwise (Handy & Niemeier, 1997).

$$Cumulative\ Opportunity_i = \sum_j D_j * \alpha_{ij}\,, \qquad \alpha_{ij} = \begin{cases} 1, & if\ TT_{ij} \leq \beta\,; \\ 0, & otherwise \end{cases} \qquad [2.4]$$

Although the cumulative opportunity measure is easy to interpret, the measure uses discrete and arbitrary travel time thresholds ($\beta$) within which all reachable destinations are assumed to be equally attractive (Hasnine et al., 2019). In contrast to the gravity-based measure, the binary weighting function used in the cumulative opportunity measure does not accommodate continuously differentiated travel impedance.

Space-time, constrained-cumulative, or time-geography measures are constraint-based in that they delineate the countable opportunity space using activity availability and individual time budgets. Hägerstrand (1970) first developed the concept of a space-time prism to geometrically represent this opportunity space, where space is a two-dimensional plane (x and y axes) from which time extends along a perpendicular axis (z axis). Opportunities are fixed to the spatial plane but only occupy positions along the temporal axis when they are available, such as during working hours for a place of employment. Therefore, the boundary of accessible opportunities may follow an individual's schedule of activities throughout the day, only counting those within a travel budget bounded by the location and time of fixed, non-discretionary activities (called "space-time anchors"). Space-time measures can represent day-to-day variations in personal accessibility, based on individual activity schedules (Neutens et al., 2012). The temporal component of this measure overcomes the location-based accessibility assumption of static, unconstrained accessibility to all destinations. However, similar to the cumulative opportunity approach of delineating "accessible" spaces, the space-time prism also imposes a discrete boundary of inaccessible spaces outside of the prism interior (Miller, 2017). Fang et al. (2010, p. 4) present a space-time accessibility indicator for an activity type (e.g., shopping) that is assigned to a destination $i$ in equation [2.5] based on the availability of accessible activities. $k_s$ represents an activity at place $s$ (e.g., shopping at store "A"), within the activity place set $K$ ($K = \{k_s \mid \ s \in n\}$) of all places $n$. Elements of set $K$ are candidate activity places (available or unavailable). $K'(o, d)$ are the available activities within the space-time prism defined by space time anchors (first origin, $o$, and final destination, $d$) and activity place set $K$. $K'(o, d) = \ k'_j, \ k'_j \in K$, where place $j$ has an available activity among total available activities $m$ within the space-time prism. Thus, if all activity places in the prism are available, $K'(o, d) = K$. The available activity time of $k'_j$ ($k'_j \in K$) from origin $o$ to destination $i$ is $AAT(k'_j)$.

$$Space - Time\ Accessibility_i = \sum_{k'_j \in K'(o,j)} AAT\left(k'_j\right), K'(o, i) \in K \qquad [2.5]$$

Utility-based, accessibility-benefit, or econometric accessibility measures also constrain the accessibility space, but rather than count potential opportunities, accessibility assumes the form of a utility function with a value corresponding with the individual's maximum utility trip (Ben-Akiva & Lerman, 1985; Chorus & de Jong, 2011). The linear utility function takes the form of equation [1.1], where the dependent variable is accessibility, and the independent variables are socioeconomic and transportation service variables. Here, the accessibility envelope is a generated set of potential paths (perceived accessibility), which is assumed equivalent to the expected maximum utility path to a user (the logsum) (Nassir et al., 2016). The maximum utility path is chosen using a discrete choice model among a set of available activities, given the individual's preferences derived from socioeconomic characteristics (Hasnine et al., 2019). Although utility-based accessibility provides some measurement of perceived accessibility values and can thus capture preference heterogeneity, the unitless measure lacks interpretability of results and cannot be quantitatively compared across different areas because of its connection to individual characteristics (Cascetta et al., 2013). Instead, accessibility can be compared between schedules and activities as the difference in the individual's utility resulting from choosing an alternative, or opportunity cost.

### 2.3.2 Transit Accessibility Applications

This thesis is specifically concerned with transit accessibility, which describes the opportunities reachable using transit services from an origin location. For clarity, transit accessibility refers to the accessibility measurement of taking transit trips, or accessibility *by* transit, rather than accessibility *to* transit services, which is focused on access/egress trip components (often by walking). Applications of transit accessibility in the literature are focused on the differentiation of transit accessibility values across space to support planning and monitoring efforts. These applications include evaluations of infrastructure development or divestment impacts (Farber & Fu, 2016; J. Lee & Miller, 2018), service gap identification (Fayyaz, Liu, & Porter, 2017; Fransen et al., 2015), land value uplift (Higgins & Kanaroglou, 2018), and inequity between socioeconomic strata (A. El-Geneidy et al., 2016; Stępniak & Goliszek, 2017). A handful of studies have studied the impact of transit accessibility on transit mode shares (Moniruzzaman & Páez, 2012; Owen & Levinson, 2015; Papaioannou & Martinez, 2015).

The subject of transit accessibility measurement, and of non-automobile modes in general, has been subject to limited research despite the emergence of the accessibility concept roughly half a century ago (Krizek et al., 2009). Since transit trips have unique travel components and spatiotemporal constraints relative to other modes, methodological research in transit accessibility literature explores techniques to calculate transit travel impedance (i.e., total travel times) towards adapting accessibility measures (e.g., cumulative opportunity) to transit travel characteristics (Lei & Church, 2010). Like general accessibility measures, transit accessibility applications do not universally employ a specific measure (i.e., gravity-based, cumulative opportunity, space-time, or utility-based) because of the advantages and disadvantages associated with each measure class.

### *2.3.3 Temporal Component*

Historically, transit accessibility researchers faced difficulty with the accurate and comprehensive calculation of transit travel times. Compared with other modes, transit is uniquely constrained by its fixed departures over time, more limited accesses/egresses across space, and dependence on transfers (Nassir et al., 2016). Since transit accessibility evaluations are highly sensitive to these sources of measurement errors, transit accessibility researchers were motivated towards improving geographical models. Early computational approaches for transit travel time measurement in the late 1990s used Geographical Information Systems (GIS) to model transit trip components over space (Lei & Church, 2010). Authors often simplified travel times by dividing travel component distances with average speeds, or using distance as a proxy (Liu & Zhu, 2004). Without reference to transit schedules, these approaches would fail to represent variation in waiting, transfer, and departure times along alternative trip paths and between different times of day. O'Sullivan et al. (2000) were the first researchers to explicitly apply transit schedule data to create shortest-path travel time isochrones (lines of equal travel time) but still simplified the calculation. Specifically, they used scheduled route travel times to derive distance-weighted segment travel times and calculated transfer times as half of the departing route's scheduled headway. Although the timetable data were complete, network construction for representation of all possible trips and transfers was still too data intensive for modelling from even a single origin, let alone all origins in a network (O'Sullivan et al., 2000).

More recent improvements in computing capacity and GIS have allowed researchers to calculate transit accessibility at a near-continuous temporal resolution (Fayyaz, Liu, & Zhang, 2017). Google's introduction of the General Transit Feed Specification in 2005 enabled researchers and practitioners to model entire transit networks with detailed schedule information (see "Transit Schedule Information" subsection for detail). For example, Lei and Church (2010) developed the "Transit Accessibility Planning Analyst" GIS tool to calculate the shortest path by transit from an origin to every destination in the network. They produced isochrone-based maps using typical commuter arrival and departure times (i.e., 8:00 and 17:00). In a study of all non-motorized modes, Krizek et al. (2009) calculated gravity-based transit accessibility measures for eight time periods throughout the day, for seven entire transit networks (metropolitan areas), and for three different years (only the most recent year, 2005, of schedule information was digitally available, corresponding to GTFS). Boisjoly and El-Geneidy (2016) collect hourly departure times and aggregate them within five periods of the day to represent daily transit travel time fluctuation in their cumulative opportunity measure. Although these studies increased representation for multiple times of day, a temporal sampling error still exists because accessibility drops at transit stops immediately after vehicle departures. Since headways and other service parameters vary minute-by-minute, measures that can capture short-term transit service dynamics offer more descriptive assessments of transit accessibility.

Methodological research on transit accessibility has expanded to examine the descriptive and statistical merit of increasingly high-resolution temporal sampling. "Dynamic" accessibility modelling is an extension of accessibility measures concerned with how user trip-making, transportation system performance, and activities are distributed over time (Järv et al., 2018). Several studies use dynamic accessibility methods to represent intra-period variation in transit accessibility, including 15-minute (Stępniak & Goliszek, 2017), 10-minute (Fayyaz, Liu, & Porter, 2017), five-minute (Fransen et al., 2015), and even one-minute intervals (Farber & Fu, 2016). Stępniak et al. (2019) provide a comprehensive list of dynamic accessibility literature and then test temporal-sampling-frequency scenarios against a one-minute benchmark to understand losses in precision at lower frequencies. For all scenarios, they found that there is a negligible decrease in precision for sampling intervals less than five minutes (which have a mean absolute error (MAE) $<$ 1 min.). Furthermore, they recommend the use of 15-minute sampling intervals, trading a small decrease in precision (MAE $<$ 2 min.) for a large computation time reduction.

Murphy and Owen (2019) compare four different sampling strategies (simple random, systematic/regular interval, hybrid, and constrained-walk) alongside sampling frequency and found that all sampling strategies performed well at high frequencies (i.e., approximately 20 per hour, or every 4 minutes), but warned that systematic sampling is sensitive to harmonic error effects (i.e., when departure schedules are regularly coincident or discordant with the sampling time) transit schedules at 10-minute intervals. They suggest the use of two, three, and four-minute sampling intervals with the systematic sampling strategy to guarantee an average normalized root mean square error (NRMSE) below 2.5%. These findings consistently describe the trade-off between computation time and temporal sampling error, but that "best" sampling strategies are difficult to recommend without testing within specific application contexts. Thus, depending on computational resources, measurement complexity, and transit network characteristics, some authors continue to select a representative departure time for daily travel periods (J. Lee & Miller, 2018; Legrain et al., 2015).

### *2.3.4 Land Use Components*

In measuring transit accessibility, the selection of attraction components (land use interaction terms), and ultimately the type of measure, also depend on the application context and research objective. Generally, the land use component of accessibility describes the spatial distribution of activities from an economic, incentives-based perspective, where value reflects the user demand for accessibility to locations (Geurs & Ritsema van Eck, 2001). The incentives-based definition of attraction is especially accurate when used in studies about how transit accessibility affects land values. For example, Higgins and Kanaroglou (2018) relate the impact of increased transit accessibility (within a bundle of "transit oriented development" goods) to some measurable increase in nearby property values. Within their gravity-based transit accessibility measure, the land use interaction terms include total population at the origin zone ($Pop_i$) and total employment at the destination zone ($Emp_j$). Of course, people do not exclusively perform commuter trips. Studies interested in the assessing the value of transit accessibility outside of markets (e.g., user-derived value) ought to account for different trip purposes. Moya-Gómez et al. (2018) discuss the need for dynamic accessibility analyses that consider the location of the population throughout the day as a proxy for destination attractiveness, and Stępniak et al. (2019) note that the population uses accessibility to participate in activities that improve their well-being. Burns and Golob (1976) offered an early summary of interaction terms (attraction measures) based on generic activities:

- total population (general measure of attraction to activities corresponding to population centres: shopping, social, employment),
- total employment (employment opportunities),
- commercial building acres (shopping),
- retail employment (shopping), and
- industry (employment).

### *2.3.5 Transit Accessibility and Mode Choice*

Limited research exists on the relationship between transit accessibility and mode choices due to the difficulty in calculating complex transit travel times for individual users (Mavoa et al., 2012; Owen & Levinson, 2015). However, transit accessibility has a strong theoretical relationship with transit choices in urban dwellers because it can be sensitive to "valuable" land uses and transit service quality (Wegener & Fürst, 1999). Temporally dynamic measures of transit accessibility, or dynamic transit accessibility (DTA) measures, can reflect system characteristics to which individuals respond when making mode choices. DTA captures the accessibility impacts of transit frequency, often the most important attribute in mode choices (Cirillo et al., 2011; Eboli & Mazzulla, 2011; Legrain et al., 2015), and other statistically significant attributes of transit choice, including waiting time and in-vehicle travel time (Dell'Olio et al., 2011). DTA also presents an opportunity to reflect differences in transit demand over time. For example, Polzin et al. (2002) used a time-of-day based transit accessibility measure, where the proportion of daily trips in a period were applied as a weight factor for accessibility values calculated in that period. This incorporates the time sensitivity of travel demand which is related to the perceived value and availability of activity types at different times of day. Transit service to an employment location at the middle of the night is not as valuable as that same service during the morning peak. These studies reveal DTA's ability to capture pertinent mode choice factors and strengthen the theoretical motivation for this analysis.

A few researchers have attempted to quantify transit accessibility's relationship with mode choices (either aggregate or disaggregate). Studies of transit accessibility's impact on mode shares typically use coarse spatial or temporal analysis dimensions, but offer promising results relating transit accessibility to transit shares. Moniruzzaman and Páez (2012) apply a cumulative opportunity measure to jobs in a logistic mode share model and include a spatial filter to control for spatial autocorrelation. They find that transit accessibility contributes to higher transit use and remains a significant predictor even after applying the spatial filter. However, their measure of transit accessibility is atemporal (i.e., disregards distribution of services over time) because it assumes average travel times around stations to assess the accessibility provided by new transit routes. Papaioannou and Martinez (2015) also use an atemporal gravity-based transit accessibility measure in a structural equation model of mode choices, finding that accessibility must be good in general, but also high for a particular trip. However, the study uses theoretical impedance functions (estimated from stated preference surveys), rather than from network travel time samples (Miguel Martínez & Viegas, 2013). Although this kind of transit accessibility is not spatially derived, it may be more adapted to user attributes and perceptions (Papaioannou & Martinez, 2015).

Only one analysis increases the spatiotemporal resolution of transit accessibility to model transit shares. Owen and Levinson's (2015) binomial logistic regression predicting automobile and transit shares uses a transit accessibility metric in an application closely related to this thesis. They, like Moniruzzaman and Páez, measured the cumulative opportunity to jobs at a continuous temporal resolution (one-minute intervals) for the morning peak period (7 to 9 AM) and found that average transit accessibility is positively correlated with zonal transit shares for all time thresholds (isochrones). Their model was best fit using the 40-minute threshold, achieving a $\rho^2$ value of 0.597. Transit accessibility variation also made a small but statistically significant improvement to Owen and Levinson's model fit: variation is negatively correlated with zonal transit mode share (higher variation in service lowered transit shares). Other studies allude to the role of transit accessibility in mode choice but use alternative definitions of transit accessibility (accessibility *to* transit, rather than *by* transit) (e.g., Chow et al., 2006).

## 2.4 Research Gaps

### *2.4.1 Transit Accessibility*

Studies of transit accessibility attempt to improve the representation of transit accessibility value by manipulating spatiotemporal parameters (i.e., higher temporal sampling frequency, more disaggregate zones), introducing land use attraction components that are better connected to travel demand, and testing different classes of transit accessibility. Recently, developments in transit accessibility methodology have permeated mode choice analysis: researchers use higher spatiotemporal resolutions that are sensitive to transit system attributes found to influence transit choice. Dynamic transit accessibility measures can reflect changes in transit waiting time, in-vehicle travel time, and service frequency that are relevant to perceived transit value (Dell'Olio et al., 2011). Given the recency of dynamic accessibility measures, few studies have analyzed dynamic transit accessibility's impact on mode choices. Only Owen and Levinson (2015) have used dynamic transit accessibility to predict zonal mode shares, applying a cumulative opportunity measure within a binomial logit model. No consensus exists on the class or form of transit accessibility measure that is the best predictor for mode choices. Researchers that relate transit accessibility to mode choices also typically study the commute to jobs alone (Chow et al., 2006;

Moniruzzaman & Páez, 2012; Owen & Levinson, 2015), disregarding different activity demands and transit services over the span of a day.

A gap exists in evaluating the predictive abilities of different transit accessibility measures, including temporally dynamic measures, in the context of mode choice analysis. There is an opportunity to merge the different bodies of transit accessibility research related to computational methods, attraction (land-use) term specification, merit assessment of different measure classes (gravity-based, cumulative opportunity, etc.), and transportation mode decision making. Transportation mode choice analysis may be extended by testing transit accessibility's predictive power for:

- multiple accessibility measure classes,
- non-peak periods of the day, and
- land-use attraction terms specific to the time of day.

This research fills the gap by expanding the application domain of dynamic transit accessibility and evaluate its effects on transit mode shares. Testing different measure classes may provide insight on the appropriate measure form for mode share analysis. High-resolution temporal sampling can calculate these different measures dynamically and precisely across multiple time periods, which are connected to demand for different trip purposes throughout the day.

### *2.4.2 Rules-based Mode Choice*

Previous research has demonstrated that RBMC models can induce non-compensatory decision processes without sacrificing prediction accuracy. These models generally focus on reducing bias in the model induction process and increasing prediction accuracy, sometimes at the expense of interpretability. Like transit accessibility measures, there is no consensus on the appropriate model structure to induce mode choices. The methods presented in this thesis can strengthen the connection between decision process theory and RBMC application, improving descriptions of behaviour and encouraging further adoption of RBMC models in the field.

One avenue for improving the descriptive power of RBMC models is the analysis of how land-use transportation variables impact mode choices. Researchers have called for using more detailed land-use variables to improve forecasting because of its connection to travel demand. Within the RBMC space, Cheng et al. (2019) found that land use variables were important predictors for mode choice in their RF model and suggest that changing the land use strategy can shift travel demand. They emphasize the importance of using datasets from different sources: specifically, travel data measurements. Towards identifying decision rules that are consistent with travel demand theory, this thesis investigates dynamic transit accessibility's impact on mode choices in a less restrictive (non-compensatory) model structure.

# Chapter 3 Data

## 3.1 Travel Diary Survey

The Transportation Tomorrow Survey (TTS) includes stated trip diary information for the Greater Toronto and Hamilton Area (GTHA) and surrounding municipalities, including the Region of Waterloo. Funded by the 22 provincial and municipal government organizations, the TTS is the largest travel diary survey in Canada (Data Management Group, 2018). The TTS samples every five years, most recently in October 2016 when it achieved an overall sampling rate of 5.0% of all households in its study area. The sampling rate in the Region was 4.8% of all households, varying by ±0.5% between the Region's lower-tier municipalities (e.g., City of Waterloo achieved 5.4%). The TTS represents the most comprehensive set of trip information in the Region, recording trip, transit, person, and household attributes for each sample. Both the transit accessibility analysis and mode choice model retrieve trip and demographic information from the TTS.

Zonal aggregations of TTS data use the Population-Land-Use Model (PLUM) zone system, also known as PZ2165. PLUM zones divide the Region into 2165 zones, each with areas averaging 645,228 $m^2$ (0.645 $km^2$) and population and employment counts averaging 244.5 people and 147.9 jobs, respectively. This research uses PLUM zones for their higher degree of spatial disaggregation compared with the TTS's larger traffic analysis zones (TZ576), which divide the same area into 576 zones. Population figures from the TTS are derived from expanded person-level samples and are therefore aggregated to PLUM zones alongside all trip data. Employment figures expanded from person-level samples do not classify employment by sector and appear to underrepresent the total Regional employment (215,546 jobs) compared with a table of employment, also provided by the TTS, for the same area (317,539 jobs). No detail exists on why the discrepancy is so high; however, the employment data table is used in this thesis because it classifies employment into sectors (industrial, retail, office, service, primary, education, work-at-home), suggesting some manipulation occurred to match an external data source. Employment data from the sector-specific table is aggregated to TZ576 zones. For consistent use of PLUM zones in this research, employment counts were transformed into employment densities by sector for each TZ576 zone, spatially-joined to the PLUM zones contained therein, and then recalculated for PLUM zone areas, assuming uniform distribution across space.

Disaggregate trip data code to PLUM zone identities to spatially distribute observed trip attributes (modes used, trip times, purposes, and auxiliary information). Some trips recorded in the TTS are not attached to zone geography (no ID), possibly due to incomplete surveys or privacy reasons. Non-zone trips are still geographically connected to the Region, but their lack of zonal identification precludes them from this analysis, which identifies mode choices and trip departure times based on origin zones (i.e., for calculating location-based accessibility). Therefore, the sum of total trips by mode for all zones differ from Regional mode shares. Regional mode shares are still noted in the Calculating Mode Shares section because they include a larger sample of users from the Region and are therefore more representative of Regionally aggregate trip characteristics.

The rules-based mode choice model employs the same disaggregate trip data with additional personal and household information. Detailed information about preprocessing the independent variables for mode choice analysis are described in the Independent Variables section, which lists the independent variables, their categories, and the unit of organization used in the mode choice model.

## 3.2 Transit Schedule Information

Aside from Travel Diary (trip survey) data collection methods, transportation engineers and planners have system-level information related to measuring the performance of transportation infrastructure. The General Transit Feed Specification (GTFS) is a standard format for storing transit schedule information and associated geographic information (Google, 2020). Google created GTFS for transit agencies to distribute standardized schedule information for interpretation by the public, developers, and researchers. It is available in dynamic and static forms. The dynamic information is a feed of real-time transit vehicle location data, often used for live departure updates in transit information devices and mobile applications. The static information comprises files that collectively detail all planned trips within a schedule period (roughly, one month); that is, they represent the occurrence of every planned transit trip departure at any given stop in the system at a one-minute resolution. Static GTFS, at a minimum, includes five, comma-separated value (csv), files:

1. Agency: identity of the transit agency providing the service
2. Stops: locations and identification (numbers and names) for transit stops and stations
3. Routes: transit route identification and route type (vehicle technology)
4. Trips: trip identification, their associated route ID numbers, and dates of service (M-F, Sat, etc.)
5. Stop times: for stops along a particular trip, the trip ID number, stop ID number, departure time, stop sequence number within the trip

GTFS information is available publicly through online repositories for many transit systems around the world. Transit agencies regularly produce GTFS datasets and update them alongside changes in operating conditions. The Region's local transit agency, Grand River Transit (GRT), regularly publishes this information for their services. Thus, the schedule period was selected to align with the travel diary survey collection period in October 2016. An arbitrary day, Wednesday, October 19, 2016 was chosen because it is not a holiday, weekend, or otherwise eventful day.

## 3.3 Road Data

Road network elements are represented in Esri's shapefile format. Relevant attributes include road names, road classifications, and physical lengths (Region of Waterloo, 2019). The Region updated this data on January 4, 2019 and no historical repository exists. Within the shapefile, roads are represented by their centre lines, digitized as line elements. Road elements classified as freeways and expressways are restricted to pedestrians for the purposes of the transit accessibility analysis.

# Chapter 4 Impact of Transit Accessibility on Transit Share

This chapter explores the relationship between transit accessibility and aggregate mode shares. Based on findings from decision process literature related to the boundaries of cognition and heuristics (Gigerenzer & Gaissmaier, 2011; Simon, 1972; van der Pligt, 2015), the hypothesis of this chapter is that transit accessibility enters user decision frames through the heuristic recollection of generally positive or negative transit experiences. Thus, a representative measure of perceived transit accessibility ought to be a statistically significant predictor of higher observed transit shares. An interesting extension of this hypothesis is whether people also interpret the temporal variation in transit accessibility when making their decisions. Broadly, the objective of this chapter is to find a transit accessibility metric that is, by magnitude or its dispersion, a statistically significant explanatory variable of aggregate mode shares. The objectives of this chapter are:

- Develop alternative transit accessibility metrics
- Analyze possible linear relationships with mode shares using regression
- Compare different transit accessibility metrics by statistical merit
- Select a metric for use in subsequent mode choice analysis

Several assumptions underpin the research objectives. This research assumes that aggregate user behaviour responds to some appropriately specified form of transit accessibility, and that a statistically significant relationship with transit shares supports the subsequent implementation of the transit accessibility metric in a disaggregate mode choice model (see “DTA in a RBMC Model” subsection).

## 4.1 Dynamic Transit Accessibility Metric Development

Dynamic transit accessibility (DTA) is defined in this research as a statistical description of transit accessibility value over time. Generally, DTA refers to the transit-specific form of dynamic accessibility, which is a concept concerned with the distribution of transportation system services over time (Järv et al., 2018). Statistical measures of central tendency (e.g., mean) and variation (e.g., standard deviation) may describe the magnitude and distribution of transit accessibility values for multiple departures over time, thus representing DTA. Discussion in Measure Selection (DTA Magnitude and Dispersion subsection) addresses the statistical attributes of DTA used in this analysis.

Accessibility frameworks in literature, resource constraints (data or computational), and broader objectives of this thesis guided the development of the DTA metric. Handy and Niemeier (1997) created a development framework for transit accessibility measures from which this thesis partly borrows. First, the class of measure must be selected based on study objectives. Second, measure specification must address the degree and type (spatial or temporal) of disaggregation, the definition of origins and destinations, and how attractiveness and travel impedance are measured. The objectives of this chapter and broader objectives of this thesis inform the metric specification, interpretation, and application space. Objectives of DTA development can be organized by application requirements:

- Theory: apply a measure that is connected to transportation demand theory and decision process research (heuristics)
- Applicability and Transferability: use existing, publicly available datasets
- Interpretability: present a conceptually straightforward representation of accessibility value
- Robustness: statistically significant explanatory performance in a variety of scenarios
- Location-based: can be compared across space and provides opportunities for infrastructure/service intervention

### 4.1.1 *Selecting Measure Classes for Testing*

The original gravity-based measure of accessibility is attractive because of its simplicity over more complex measures (i.e., utility-based and space-time prisms). It is also useful for comparison across space because it is location-based. These attributes make it an attractive candidate for testing. An early criticism of gravity-based measures is presented by Kutter (1972), who argues that travel behaviour ought to be based on behavioural and attitudinal surveys, rather than the assumption that travel motivation is some function of origin and destination "masses," postulated by gravity measures (Banister, 2002). However, these measures may remain useful because the attraction components used to represent "masses" (e.g., population, employment) are proxies for underlying travel motivations and mode share dependencies. For example, measures using population magnitude may represent the activity size of an area, which is found to affect transit shares (Taylor et al., 2003). Population density is one of the 3D's (density, diversity, design), which are built-environment factors found to have a statistically significant affect on mode shares (Cervero, 2002). This research, therefore, tests gravity-based measure forms.

Countable accessibility measure classes (cumulative opportunity and space-time) offer an interpretable result in their expression of a magnitude of activities accessible within an isochrone. The cumulative opportunity measure is especially simple because the count is only constrained by the selected time threshold, within which travel is assumed to be equally valuable to users. However, a criticism of cumulative opportunity measures is that they are limited by their high sensitivity to the chosen time threshold value. Stępniak et al.'s (2019) study of temporal resolution in transit accessibility measurement found that cumulative opportunity measures are not suitable to explain accessibility changes over time because of discrete boundary effects and arbitration of the boundary itself. Cascetta et al. (2013) note an inability to calibrate a robust threshold from real data, and therefore the difficulty in reproducing user behaviours wherein highly attractive destinations overcome a high impedance of travel. Despite the weaknesses of cumulative opportunity measures, some forms have been applied successfully in mode choice modelling, producing statistically significant results when the threshold is set at 40 minutes (Owen & Levinson, 2015). Although it is technically a subset of the gravity-based measure, and therefore subject to the same weaknesses in behavioural representation, the single attraction component used within cumulative opportunity measures may be more consistent with travel demand theory than interaction-based gravity measures (i.e., where two attraction components exist) because trip makers may be compelled to travel based on the attraction of their destination alone. This research therefore tests cumulative opportunity measure forms.

Space-time prism measures introduce temporal constraints related to activity availability that could benefit this research. Temporal constraints on activities address the limitations of basic location-based accessibility measures related to user demand representation (i.e., changes in demand throughout the day to different activity locations). However, space-time measures are similarly at the mercy of the temporal resolution of analysis. Studies using space-time measures may still measure accessibility values sparsely throughout the day, which could distort actual accessibility envelopes. For example, Lee and Harvey (2018) generalize transit accessibility levels for four time periods using a single sample per period (8 a.m., 1 p.m., 6p.m., 9 p.m.) and discuss the value of using dynamic transit accessibility analysis in future work. Consideration towards activity availability also primarily affects trip generation rather than mode choices because activity availability is constant across all modes (i.e., the grocery store closes at the same time regardless of whether the user takes transit or car). Except for trips at the fringes of activity availability times, where travel time differentiation between modes are sufficiently large for users to "miss" an activity with some modes and not others, activity availability times are less relevant for mode choice analysis. In any case, no reliable activity availability data exists that are expressed in terms of employment counts for the Region. Information that does exist about business availability times are not connected to information about business size (i.e., number of employees, retail floor area). Activity availability times also vary within zones and thus require a user-/activity-based approach, where individual destinations and daily schedules are known, to provide meaningful accessibility values. Therefore, this research does not apply space-time accessibility measures.

Utility-based accessibility relies on economics-based, utility-maximizing axioms (Ben-Akiva & Lerman, 1979), from which this thesis departs in favour of other psychological theory. Utility-based accessibility measures are difficult to interpret, sensitive to existing user characteristics, and cannot be meaningfully compared between territorial areas (Cascetta et al., 2013). The inclusion of user characteristics is also problematic at this modelling stage because the transit accessibility metric developed in this chapter is intended for use in a mode choice model. User characteristics are considered later within a disaggregate decision process model. Including user characteristics within the transit accessibility metric risks covariance. This research does not apply a utility-based measure of accessibility.

Efforts to model perceived TA typically vary TA based on user socioeconomic attributes (Cascetta et al., 2013) and temporal accessibility components, which location-based measures ignore (Krizek et al., 2009). However, aggregate mode share models using dynamic, location-based transit accessibility measures have offered encouraging results (Owen & Levinson, 2015). This thesis modifies the standard gravity-based measure to increase sensitivity to user demand in two ways. First, by evaluating travel impedance at multiple departures, the metric is sensitive to changes in transit accessibility over time. Second, the attraction components are manipulated to reflect dominant trip purposes within each period of the day. This thesis tests these modifications on three types of location-based accessibility measures for their relationships with mode shares. A gravity-based measure with two attraction terms (origin and destination) is used alongside a cumulative opportunity measure with a 40-minute time threshold. A third, hybridized measure uses a single destination attraction term (e.g., employment only) and scales the weight of opportunities by the travel impedance. It is referred to in this research as the time-decayed opportunity measure. Other authors have described this measure as a gravity based measure (Alam et al., 2010). This research uses "Time-Decayed Opportunity" to avoid confusion with interaction-based (classical) gravity models.

### *4.1.2 Time-Decayed Opportunity*

The time-decayed opportunity measure differs from the classical, interaction-based gravity measure by reducing the number of interaction terms to 1, like the cumulative opportunity measure. Rather than weighing the number of people "gravitating" from an origin to a destination, individuals are assumed to be attracted to destinations alone, regardless of the number of other individuals in their origin zone. In contrast to the cumulative opportunity measure, no arbitrary time threshold is used. Instead, destination attractiveness decays based on the travel impedance so that far-away, high-attraction destinations are appropriately included. Continuous decay of opportunities can mitigate the strong boundary effects of cumulative opportunity measures. Time-decayed opportunity attempts to improve the behavioural representation of gravity-based measures by ignoring the "attractiveness" of the origin node, which may be irrelevant to user's individual decision making.

### *4.1.3 Spatiotemporal Disaggregation*

Increasing the temporal resolution of analysis aims to reduce sampling bias, include the influence of dynamic transit service parameters, and represent a period-level conception of transit's accessibility context that users may use during the mode choice process. Findings from transit accessibility's temporal sampling literature suggest that the 5-minute sampling interval is sufficiently accurate compared with the 1-minute baseline, while offering a five-fold reduction in calculation times (Stępniak et al., 2019). Although Stępniak et al. discuss the feasibility of a 15-minute sampling interval, this research attempts to limit the impact of harmonic departure times on travel time sampling error (Murphy & Owen, 2019). The TTS trip data are also reported at the 5-minute interval. This research therefore samples transit travel times (including waiting, in-vehicle travel time, and access/egress walking times) at the 5-minute interval.

Temporally dynamic transit accessibility can represent two categories of attributes affecting travel demand. First, intra-period variations in transit service parameters include departure frequency, transfer incidence/times, and overall travel time. Second, inter-period variations in travel purposes affect the relative importance of connectivity to some destination types over others. This research calculates all-day transit accessibility at a high temporal resolution to capture intra-period service attributes and varies attraction terms to capture inter-period changes in demand. Finally, this research postulates that users do not assess the value of discrete trips and instead inform their judgements using period-level temporal aggregations of service attributes. It is possible that users are more likely to ponder whether morning transit services would fit their morning trip-making needs rather than, for instance, the value of a specific trip at 8:36 AM. Therefore, dynamic transit accessibility is represented at the period level using statistical aggregations (i.e., measures of central tendency and measures of variance) of the 5-minute transit accessibility values.

GTFS integration with GIS software and general computing abilities enable the analysis of highly disaggregate spatial zones for increasingly large study areas. GTFS is interpretable by macroscopic transportation modelling software and other GIS software, which can digitize transit systems (stops and schedule data) and spatially relate them to other layers (e.g., land use, streets). Network-wide analyses can spatially attribute quantitative transit accessibility levels for comparison across an entire urban area. Spatial characteristics can also be joined to individuals, describing the individual's travel context with detail that may be missed in a travel demand survey. Within network accessibility analyses, higher levels of spatial disaggregation are preferable because larger zones suffer greater losses in accuracy from spatial aggregation bias (modifiable areal unit problem). Transit accessibility measures are sensitive to the resolution of spatial land use information because access and egress components rely on other modes (usually, walking) that may contribute greatly to total travel times in larger zones (Krizek, 2005). Therefore, this research uses the most highly disaggregated public dataset of trip information to the authors knowledge: PLUM zones. PLUM zones have a small average size (0.645 $km^2$) and populations and activities within each zone are assumed to be more homogeneous than other zones, based on Tobler's First Law of Geography (everything is related to everything else, but near things are more related than distant things).

### 4.1.4 Travel Impedance

The travel impedance function, $f\left(TT_{ij}\right)$, and attraction components ought to describe travel demand as a function of the cost to travel between zones $i$ and $j$. In this case, travel cost is represented in time units (minutes). This research applies a common travel impedance function shown in equation [4.1] (Fu, 2017, p. 54):

$$f\left(TT_{ij}\right) = TT_{ij}^{-2} \qquad [4.1]$$

where the $TT_{ij}$ represents the travel time between origin zone $i$ and destination zone $j$. This model of experienced travel impedance increasingly reduces the value of destinations to which it takes longer to travel. Other accessibility measures in the literature also apply this impedance function (see "Classes of Accessibility" subsection for references). This research assumes that all trip components are experienced equally: travel cost varies only by time, regardless of differences in comfort between trip components.

Transit and walking modes are both included in the travel impedance measure to include trips where walking is faster than transit, account for the influence of intrazonal destinations, and represent a baseline accessibility value that respects location. Performing transit trips requires walking trips for stop access and egress components. However, instances where waiting for the transit trip departure would be longer than the walking trip are completely replaced by walking trips. It is conventional to do this in the literature to avoid eliminating the value of short walking trips and to assess a baseline level of accessibility for time periods when no transit service is available (Owen & Levinson, 2015).

### 4.1.5 *Attraction Terms*

Three attraction terms describe the potential activities at each zone: population, employment, and retail. Attraction terms take the form of expanded counts from the TTS. Population values represent the total number of persons living in a PLUM zone. Employment values represent the total number of employed people working in the Primary, Education, Office, and Industrial & Warehousing sectors at a zone. Discretionary values are the total number of employed people working in Retail and Services sectors at a zone.

## 4.2 DTA Evaluation Method

Evaluating DTA involves creating a spatial system on which to run the analysis, including transit system information from GTFS; calculating travel time matrices on the network using geoprocessing tools across all departure times and zones; evaluating transit accessibility using attraction measures for each departure time, $n$ ($TA_n$); and finally, aggregating the results temporally for period $p$ ($DTA_p$). The method is shown in Figure 3.

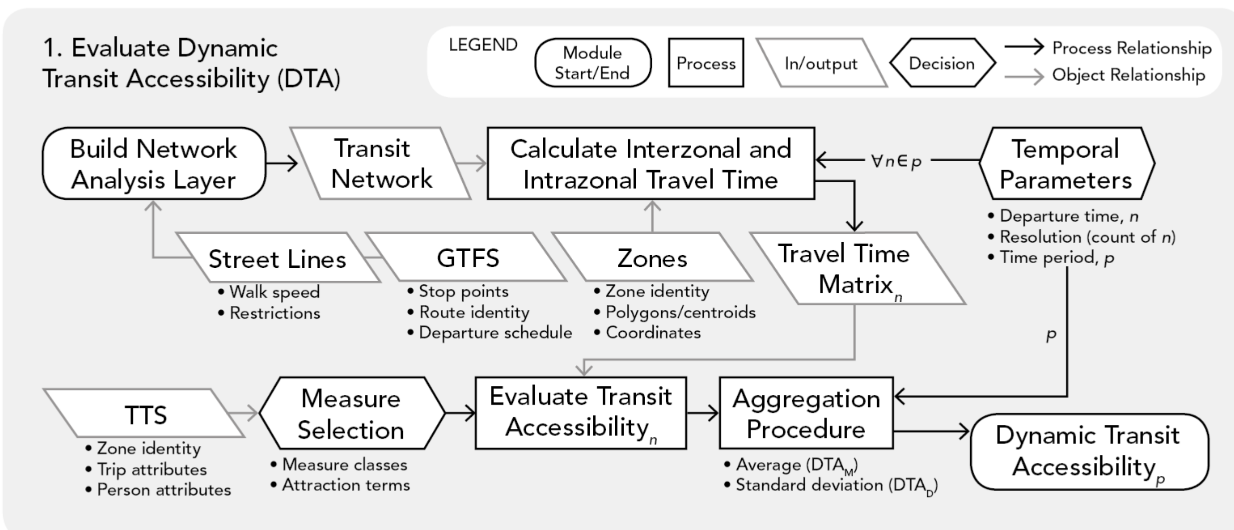



*Figure 3: Dynamic Transit Accessibility Evaluation Method*

### 4.2.1 *Building Transit and Pedestrian Network*

The spatial system was built on ArcGIS Pro version 2.6 using the network analysis toolbox, developed by Esri (2020). Inputs for the spatial system, at the minimum, include a geographical road network and the locations of transit stops (Lei & Church, 2010). Figure 4 shows the spatial elements that are used to model a user’s path through pedestrian and transit networks on GIS software.

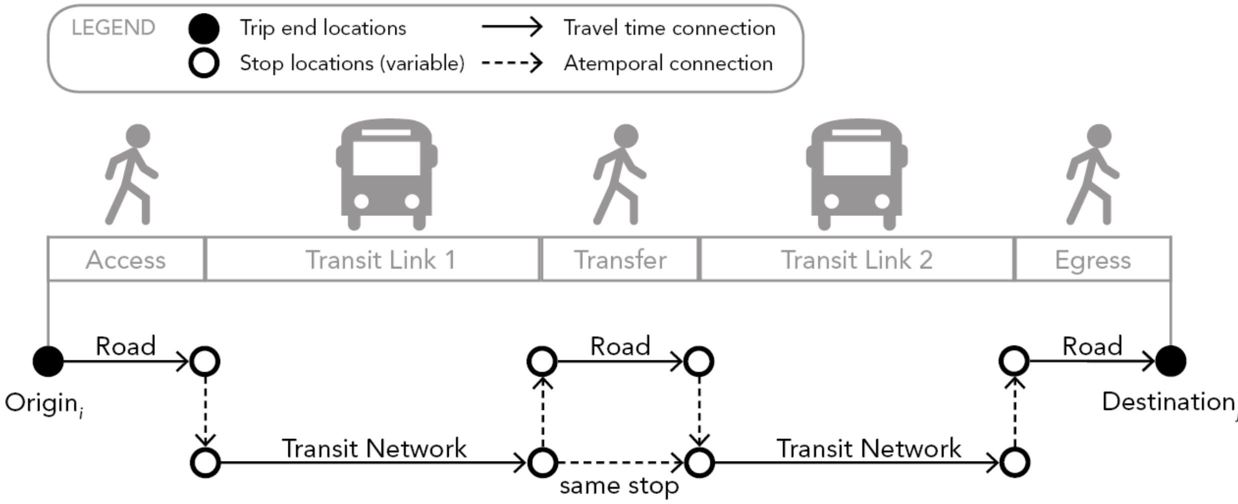



*Figure 4: Spatial network elements traversed during a transit trip*

ArcGIS's "GTFS to Network Dataset Transit Sources" tool plots stop locations as points using GTFS longitude/latitude data and connects successive stops along transit routes. Rather than representing the actual network paths taken by transit vehicles, Euclidean lines represent the transit routes that connect successive stops because their travel times do not depend on line geometry. Instead, ArcGIS parses GTFS schedules to attribute route lines with segment travel times for all trips between any two stops. This is used downstream during the network analysis that accumulates travel times along travel time connections.

Road network lines (see "3.3Road Data" subsection) represent all the possible paths taken by pedestrians either on their way to transit stops or directly to final destinations. Pedestrian restrictions exclude road network lines categorized as "Freeways" or "Highways" from this analysis. Transit stop locations are the only points by which pedestrians traversing the road network may access/egress transit services. For different times of day and for different destinations, the "nearest" stop location may vary because of trip directionality and alternative route availability. Since stops are not placed along street centre lines, modelling pedestrian movement between the road and transit networks requires virtual connective elements between transit stops and road lines. The "Connect Network Dataset Transit Sources to Streets" tool connects stop locations with a line perpendicular to the nearest road element using a search radius of 500m. Connections between stops and roads are atemporal (frictionless), incurring no travel impedance. Travel impedance only accumulates along streets using an average walk speed (1.4 m/s or 5 km/h) and along transit route lines using scheduled travel times. Based on this connectivity policy, the network is built using the ArcGIS's "Build Network" tool. Figure 5 geographically illustrates a part of the resulting network.

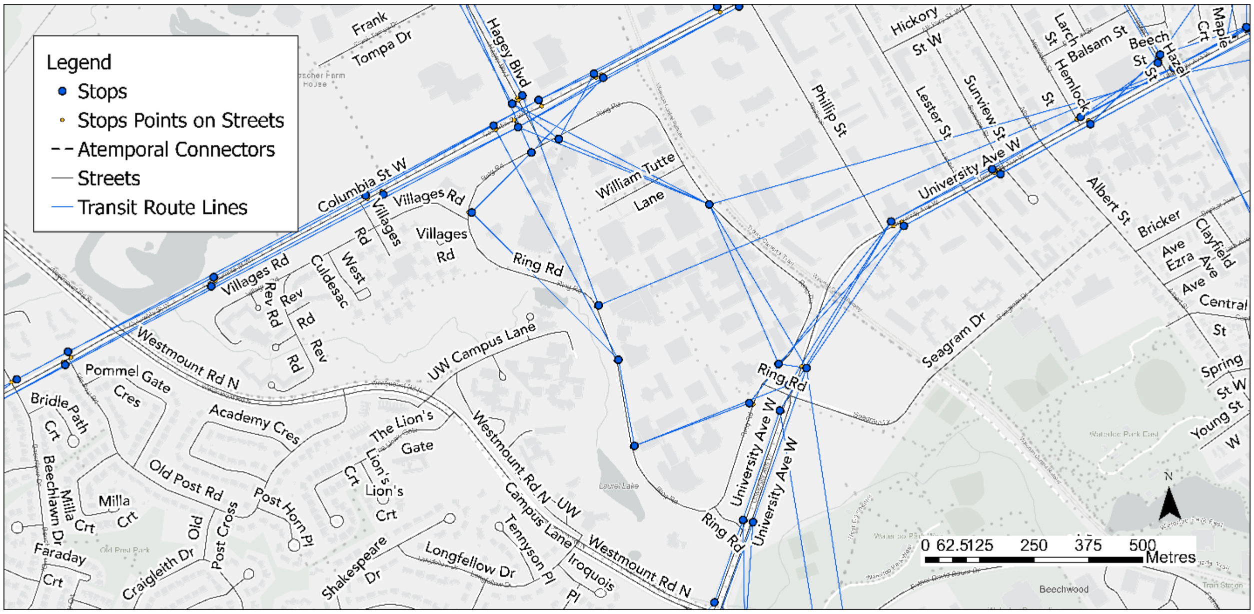



*Figure 5: Built Transit Network Excerpt from the University of Waterloo campus*

#### *4.2.2 Interzonal Travel Time Matrices*

Calculating an interzonal travel time matrix requires specification of the origin and destination points, departure time, an evaluator that can accumulate walk and transit travel time components of a trip, and a shortest path algorithm. The 2145 PLUM zones (polygons) were transformed into points at their geometric centroids to represent the trip end locations (points $i$ and $j$). These centroids are connected to the nearest perpendicular road segment using a generous 1km search radius to ensure that larger zones or zones with sparse road lines are connected to the transportation network. Given the spatial network of pedestrian-traversable roads and transit elements, shortest path travel times are calculated between every zone in the Region, $i$, to every other zone, $j$, using ArcGIS's Network Analyst toolbox ($i \in I; j \in I \neq i$). First, an empty origin-destination (OD) cost matrix was created using the "Make OD Cost Matrix Analysis Layer". Second, origin and destination points are loaded onto the layer using the "Add Origins/Destination Locations" tool and a single departure time is specified (e.g., 15:05 on October 19, 2016). The matrix is populated using the "public transit evaluator" within ArcGIS Pro, which interprets transit travel times using GTFS transit schedule information and pedestrian walking times as a function of the distance travelled. Pedestrian speed is set at 1.4 m/s to calculate transit access/egress and transfer times. Finally, travel times between every origin and destination zone are calculated using the Djikstra shortest-path algorithm built into ArcGIS Pro (Esri, 2020).

DTA requires the travel time analysis of multiple departure times to measure the changes in transit accessibility over time. Since the cost matrix analysis layer accepts only a single departure time argument, the matrix calculation must reiterate over multiple departure times. Figure 6 shows the structure by which interzonal transit travel times ($TT_{ijn}$) are calculated between origin $i$ and destination $j$ for every departure time $n$ and period $p$. $N_p$ is the set of departures in a period $p$. Periods include the morning (AM), mid-day (MD), afternoon (PM), early evening (EE), and night (NT). This analysis also aggregates travel times across all periods for an all-day (AD) analysis ($p = [AM, MD, PM, EE, NT, AD]$).

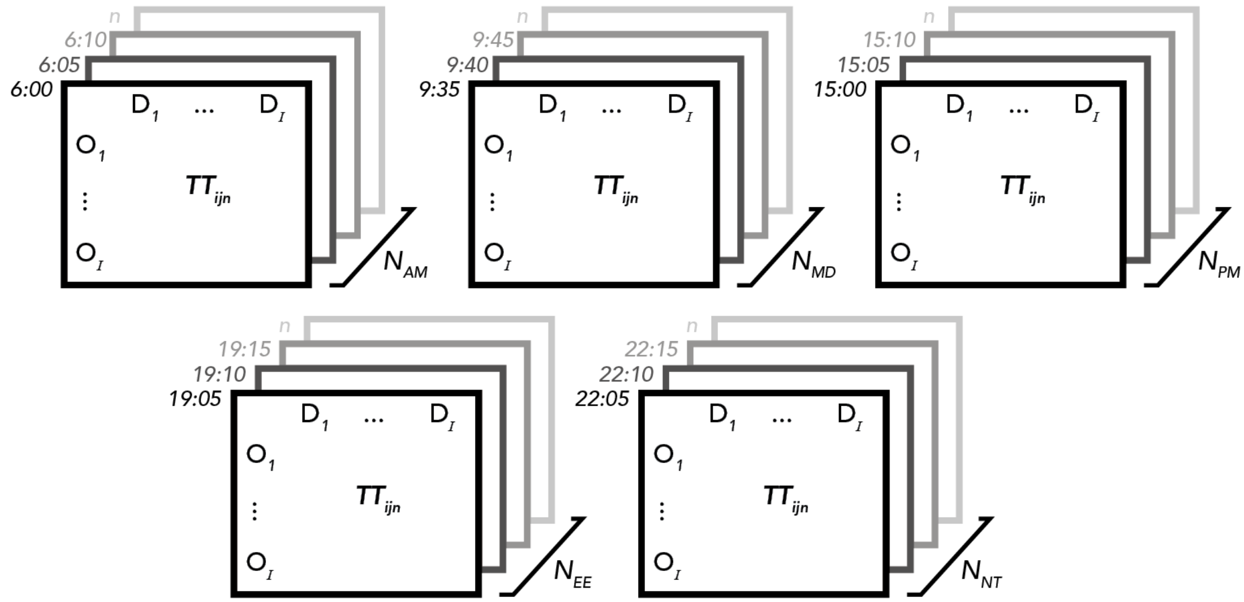



*Figure 6: Travel Time ($TT_{ijn}$) Matrix Structure for Time Periods (p)*

Python scripts using ArcPy, ArcGIS's python module, automated the $TT_{ijn}$ matrix iteration over multiple departure times ($n$). This research aggregates transit accessibility by the period of day, but uses travel time samples at $n$ ($TT_{ijn}$) to calculate the transit accessibility between each OD pair before aggregating to the origin zone at the period level.

### 4.2.3 Intrazonal Walking Trips

Intrazonal trips are assumed to be completed using walking due to high spatial disaggregation of the PLUM zone system. Any transit trips that may occur within zones would require discrete origin-destination pairs to model, which would be cumbersome to randomly populate for travel time evaluation and unlikely to compete with walking-only times. The impedance function (equation [4.1]) produces a value of 0 for intrazonal trips because the origin and destination points are the same zonal centroid. This undervalues accessibility for users making trips originating from high-activity zones and overvalues the relative accessibility of users in low-activity zones that happen to be adjacent to high-activity zones. Imposing a small travel impedance relative to the size of each zone would improve modelling accuracy.

Calculating average intrazonal walking times requires a set of assumptions to reflect differences between zonal sizes. First, arbitrary zonal polygons (based on geopolitical boundaries) are assumed to be perfect squares proportional to zone areas. Second, the walking trips within the square zones are assumed to be randomly distributed (i.e., pairs of points are randomly placed) because there is no information to infer the spatial distribution of intrazonal activities. Third, a Manhattan (grid) road network is assumed within each zone. Based on findings related to square line picking, the average Euclidian (straight-line/ "as the crow flies") distance within a square is equal to $0.5214 * x$, where $x$ is the length (metres) of the square (Weisstein, n.d.). Within a Manhattan road network, an additional penalty applies to travel distance incurred by having to navigate right angles. For any two, randomly distributed points within a rectangular zone, the travel distance is $\frac{1}{3}(x + y)$, where the $x$ and $y$ represent two adjacent sides of the rectangle (Larson & Odoni, 1981). Given the assumption of square zones, the average distance of random travel within a zone with an area of 90,000m$^2$ would be:

$$\frac{1}{3}(\sqrt{90{,}000m^2} + \sqrt{90{,}000m^2}) = 200m$$

Average walking speed is consistent with the interzonal travel time calculation: 83.33 m/min (5km/h). Thus, the intrazonal travel impedance would be approximately 3.6 minutes. This procedure is applied to intrazonal trips within the travel time impedance matrix (Figure 6).

### 4.2.4 Intra-period Travel Time Variation

Figure 7 shows the variations in transit trip components and total travel times for multiple hypothetical "transit trips" occurring at three departures times: $n$, $n-1$, and $n+1$.

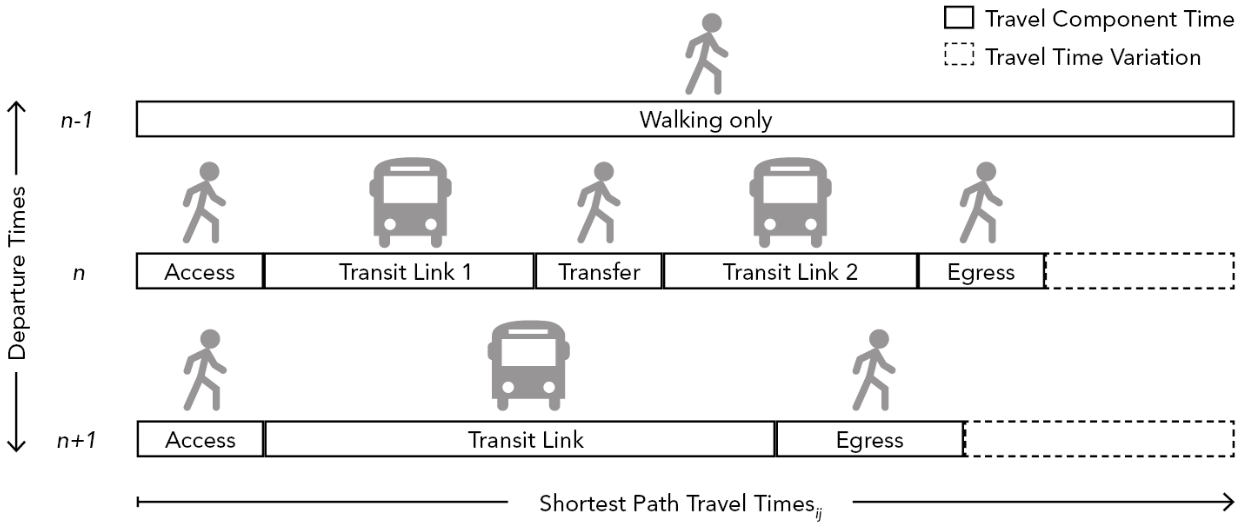



*Figure 7: Temporal Sensitivity of Transit Trips*

These trips are possible shortest path travel times between origin and destination points, which vary depending on the availability of specific transit services at time of departure, $n$. For example, some departure times may miss the "best" transit route (departure n vs departure n+1) and later require a transfer, resulting in a longer total travel time. Departure n also reveals that transit trip components are temporally interdependent. The most inflexible component is the first transit vehicle of the trip. Depending on the departure time and in-vehicle travel time of the first vehicle, users may miss departures at sequential stops, compounding with trip complexity (i.e., number of transfers). Sometimes, walking trips are shorter than the transit alternative (n-1). In all cases, the walking travel time between any two zones also represents a baseline travel impediment, or "worst-case" accessibility value that still weighs land use impacts for the purposes of calculating a DTA measure. Thus, many urban zones nearby high-activity zones have transit accessibility values that do not vary dramatically with transit trip departures because walking consistently affords a high value of accessibility. DTA (with 5-minute temporal sampling) measures the intra-period travel time variations that result from these phenomena, offering better representation of transit's modal attributes.

### 4.2.5 Measure Selection

Six different measures of dynamic transit accessibility are selected based on the different measure classes (gravity, time-decayed opportunity, cumulative opportunity) and attraction terms (population, employment, discretionary). Measures ending in "2" include discretionary trip purposes as attraction terms. All measures include destination attraction terms, but only the gravity-based measures include origin attraction terms. Table 1 describes the six transit accessibility measures by their classes, attraction term components, and temporal aggregation.

*Table 1: DTA Measure Components*

| Category | Measure Dimensions and Components | Measures | | | | | |
|---|---|---|---|---|---|---|---|
| | | G1 | G2 | D1 | D2 | C1 | C2 |
| Accessibility Measure | Gravity-based | • | • | | | | |
| | Time-decayed Opportunity | | | • | • | | |
| | Cumulative Opportunity | | | | | • | • |
| Attraction Components | Origin Attraction Terms | • | • | | | | |
| | Destination Attraction Terms | • | • | • | • | • | • |
| | Discretionary Trips Included | | • | | • | | • |
| Temporal Aggregation | All Day | • | • | • | • | • | • |
| | Time Periods (AM, MD, PM, EE, NT) | • | • | • | • | • | • |

The Gravity-based transit accessibility ($G_{in}$) for origin zone $i$ at departure time $n$ is defined by:

$$G_{in} = \sum_{\forall j \neq i} \frac{O_{ip} D_{jp}}{TT_{ijn}^2} \quad [4.2]$$

where $O_{ip}$ is the attraction term (frequency) for origin zone $i$ during period $p$, $D_{jp}$ is the attraction term for destination zone $j$ during period $p$, and $TT_{ijn}$ is the total travel time by transit and walking between $i$ and $j$ at departure time $n$. $n \in N_p$ represents every 5-minute increment of set $N_p$, which is the frequency of 5-minute increments in period $p$. If $p$ represents a different period, different interaction terms (e.g., employment at $i$ to population at $j$) apply (see Table 3).

The Time-decayed Opportunity ($D_{in}$) transit accessibility for origin zone $i$ at departure time $n$ is defined by:

$$D_{in} = \sum_{\forall j \neq i} \frac{D_{jp}}{TT_{ijn}^2} \quad [4.3]$$

where $D_{jp}$ is the attraction term (frequency) for destination zone $j$ during period $p$.

The Cumulative Opportunity transit accessibility ($C_{in}$) for origin zone $i$ at departure time $n$ is defined by:

$$C_{in} = \sum_{\forall j \neq i} D_{jp} \quad [4.4]$$

where $D_{jp}$ is the attraction term for destination zone $j$ during period $p$, and all possible destinations are simply counted. Based on cumulative opportunities studies in the past, the best performing threshold for mode share analysis of 40 minutes is used (Owen & Levinson, 2015).

### 4.2.6 DTA Magnitude and Dispersion

DTA measures should be described using statistical measures of transit accessibility because DTA represents a collection of transit accessibility values over a time period. This research chooses two statistics, mean and standard deviation, to represent DTA based on the hypothesis that high magnitude and low variation transit accessibility encourages transit use. These statistics describe aggregations of transit accessibility (TA) values because DTA characterizes an entire period of TA. Since "mean DTA" may be confused for the average DTA between multiple periods ($p$) and "DTA standard deviation" would likewise refer to the variation between periods, this research refers to DTA "magnitude" and "dispersion" to describe temporally aggregated TA values:

- DTA magnitude ($DTA_m$) is the mean transit accessibility of an origin zone, describing its transit accessibility to all other zones in the Region, across all departure times in a period ($n \in N_p$)
- DTA dispersion ($DTA_s$) is the standard deviation of an origin zone's transit accessibility values throughout a period ($p$), measured at every departure time $n$.

The DTA measures used in this research are expressed in terms of $DTA_m$ and $DTA_s$ in Table 2.

*Table 2: DTA Measures Selected for Analysis*

| $Transit\ Accessibility_{in}$ | DTA Magnitude ($DTA_m$)[1] | DTA Dispersion ($DTA_s$)[1] |
|---|---|---|
| $G_{in} = \sum_{\forall j \neq i} \frac{O_{ip} D_{jp}}{TT_{ijn}^2}$ | $G_{m,ip} = \frac{\sum_{n=1} \sum_{\forall j \neq i} \frac{O_{ip} D_{jp}}{TT_{ijn}^2}}{N_p}$ | $G_{s,ip} = \sqrt{\frac{\sum_{n=1} \left(G_{in} - G_{m,ip}\right)^2}{N_p}}$ |
| $D_{in} = \sum_{\forall j \neq i} \frac{D_{jp}}{TT_{ijn}^2}$ | $D_{m,ip} = \frac{\sum_{n=1} \sum_{\forall j \neq i} \frac{D_{jp}}{TT_{ijn}^2}}{N_p}$ | $D_{s,ip} = \sqrt{\frac{\sum_{n=1} \left(D_{in} - D_{m,ip}\right)^2}{N_p}}$ |
| $C_{in} = \sum_{\forall j \neq i} D_{jp}$ | $C_{m,ip} = \frac{\sum_{n=1} \sum_{\forall j \neq i} D_{jp}}{N_p}$ | $C_{s,ip} = \sqrt{\frac{\sum_{n=1} \left(C_{in} - C_{m,ip}\right)^2}{N_p}}$ |

[1] These DTA measures describe a single origin zone, $i$.

*Note: log transformations may apply to these measures in the form ln(DTA) (see "Transformations for Linearity" subsection).*

$DTA_m$ allows quick comparison between origin zones: urban zones generally have higher $DTA_m$ than suburban or rural zones due to the proximity of concentrated activities. However, $DTA_m$ is not unilaterally representative of transit service value: extreme high- or low-accessibility departure instances (at time $n$) may skew the mean. Since transit accessibility is temporally dynamic, a measure of DTA's dispersion is useful to capture inconsistent transit services. $DTA_s$ captures service changes in a way that is sensitive to the destinations that origin zone $i$ can reach via transit by including location-specific attraction terms. Variation between transit accessibility values in a time series can be interpreted in terms of transit trip departures.

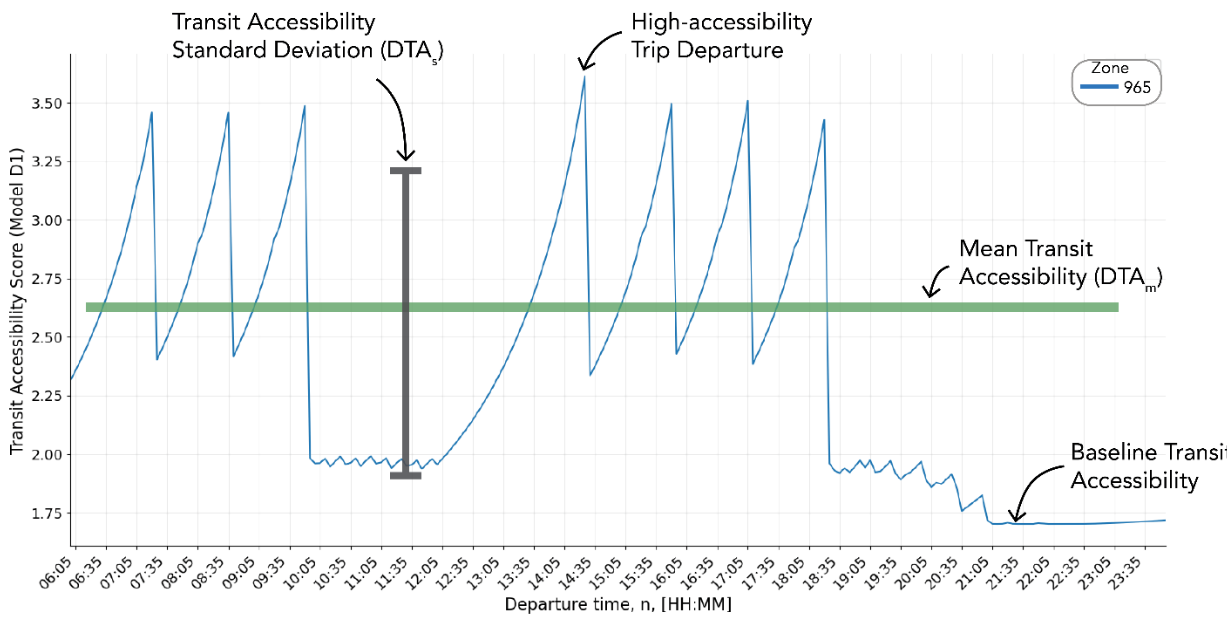



*Figure 8: Statistical Description of Transit Accessibility over Time (Zone 965)*

Figure 8 plots the AD transit accessibility values of origin zone 965 for each departure time, $n$, using measure D1. Zone 965 was chosen to exemplify "peaky" transit service characteristics. The baseline represents times during which there is no transit service or when walking is consistently better than transit. This is equivalent to the walking accessibility, or "baseline" transit accessibility: the lowest transit accessibility value for that origin. More urban areas generally have higher baseline accessibility values due to the concentration of urban activities. Local minima are departure instances with either the longest waiting time for transit services, or the baseline walking time – whichever is higher. Local maxima are the transit departure times, or shortest waiting time before departure (always <5 mins, the temporal sampling resolution) – whichever is shortest. Upward trends in transit accessibility over time represent lowered waiting times for a departure, where either the waiting time for the transit trip is shorter the baseline walking time or the waiting time for the transit trip is shorter than alternative transit paths – whichever is shortest.

### *4.2.7 Attraction Term Selection*

This research uses different origin and destination terms to measure transit accessibility depending on the time period of analysis. Choice of attraction term(s) for each period and DTA measure type (G, D, C) specifically references the Region's predominant trip patterns, plotted throughout the day in Figure 9.

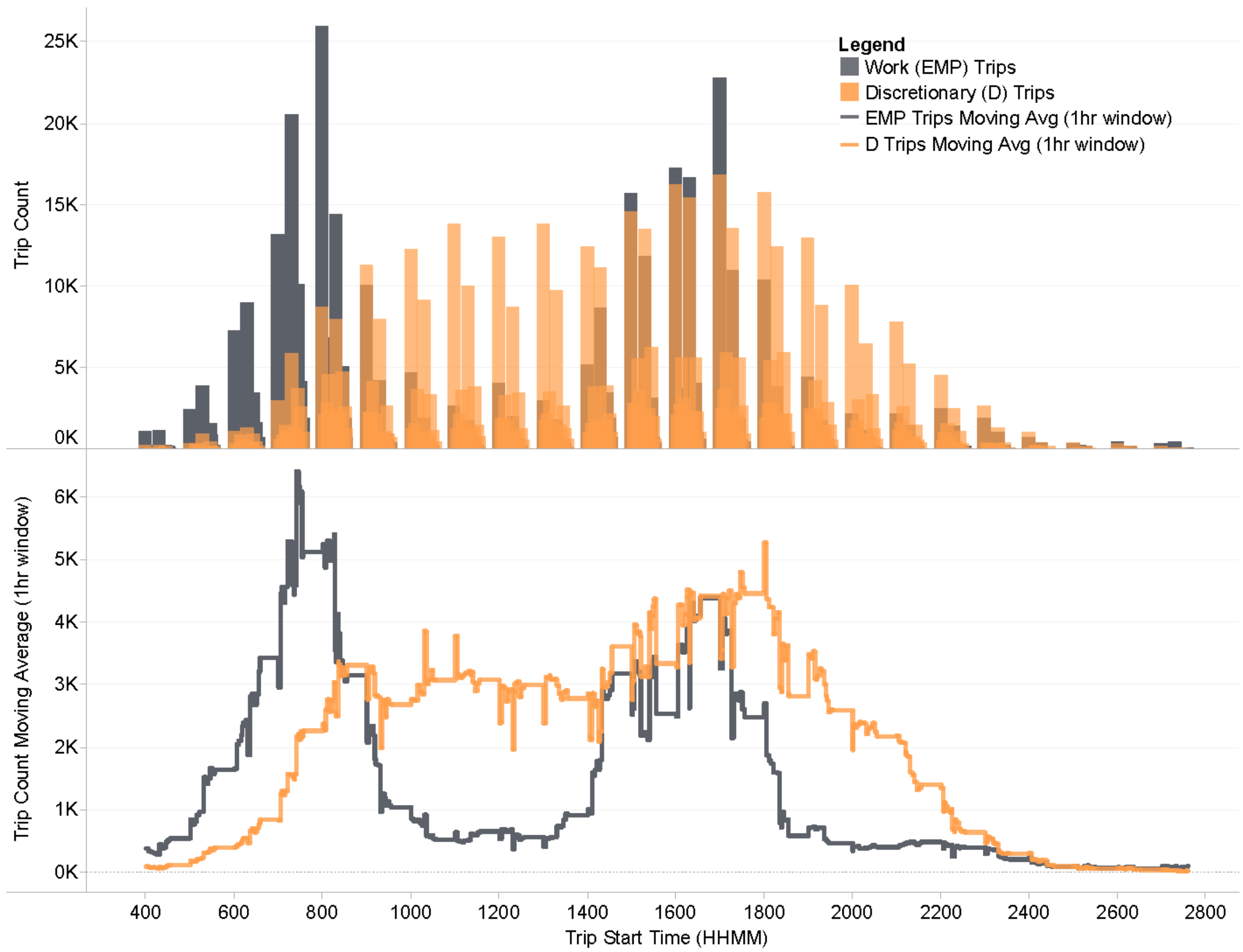



*Figure 9: Trip Start Time Distribution by Trip Purpose, Region of Waterloo* (Data Management Group, 2018)

The trip purposes are groups of trip purpose categories surveyed in the TTS. Work trips include "Home-based Work" and "Home-based School" trips (i.e., home to work, work to home, home to school, and school to home). Discretionary trips include "Home-based Discretionary" and "Non Home-based" trips (i.e., discretionary to discretionary, work to discretionary, home to discretionary, discretionary to home). The bottom section of Figure 9 shows a moving average (one hour around every point, 30 minutes before and 30 minutes after) for legibility and response bias reduction (e.g., people may tend to report departures at 8:30 rather than 8:38). The peaks and valleys of this time-series data is not unique to the Region. Home-based work trips spike during morning and afternoon peak periods and discretionary trips are spread throughout the day. These data informed judgement in selecting the attraction terms that are consistent with the motivations for travel demand. These are listed in Table 3.

*Table 3: Time Period Attraction Component Terms (Origin/Destination) and Sample Frequency*

<table>
<tr><th rowspan="2">Period ($p$)</th><th rowspan="2">Time Span (inclusive)</th><th rowspan="2">Temporal Freq. ($N_p$)</th><th colspan="6">Measures</th></tr>
<tr><th>G1</th><th>G2</th><th>D1</th><th>D2</th><th>C1</th><th>C2</th></tr>
<tr><td>AM</td><td>6:00-9:30</td><td>43</td><td rowspan="2">Pop/ Emp</td><td>Pop/ (Emp+D)</td><td rowspan="2">-- /Emp</td><td>--/ (Emp+D)</td><td rowspan="2">-- /Emp</td><td>--/ (Emp+D)</td></tr>
<tr><td>MD</td><td>9:35-14:55</td><td>65</td><td>Emp/D</td><td>--/D</td><td>--/D</td></tr>
<tr><td>PM</td><td>15:00-19:00</td><td>49</td><td rowspan="3">Emp/ Pop</td><td>Emp/ (Pop+D)</td><td rowspan="3">-- /Pop</td><td rowspan="2">--/(Pop+D)</td><td rowspan="3">-- /Pop</td><td rowspan="2">--/ (Pop+D)</td></tr>
<tr><td>EE</td><td>19:05-22:00</td><td>36</td><td rowspan="2">(Emp+D)/ Pop</td></tr>
<tr><td>NT</td><td>22:05-23:55</td><td>23</td><td>--/Pop</td><td>--/Pop</td></tr>
<tr><td>All day</td><td>6:00-23:55</td><td>216</td><td>Pop/ Emp</td><td>Pop/ (Emp+D)</td><td>-- /Emp</td><td>--/ (Emp+D)</td><td>-- /Emp</td><td>--/ (Emp+D)</td></tr>
</table>

*Periods: morning peak (AM), midday (MD), afternoon peak (PM), early evening (EE), night (NT).*
*Attraction terms: Nothing (--), Home (Pop), Work/School (Emp), Retail+Service Discretionary (D).*

It is important to note that these attraction terms limit the number of zones (data points) used in the regression analysis with mode shares. This is unavoidable because zones (and specifically, zoning by-laws) separate and unevenly distribute activities across space. Since transit accessibility measures in this thesis use the number of attractions in a zone to represent the demand for travel to/from that zone, zones without incidences of these attraction terms are excluded from analysis. For example, the AM analysis of measure G1 does not include origin zones with zero population. The number of zones analyzed are therefore limited to origins with recorded populations (1091 zones). In the PM, measure G1's sample of zones is limited to origin zones with employment (2042 zones). Table 6 tabulates the final number of zonal samples used in the regression analysis.

## 4.3 Calculating Mode Shares

Mode shares in each PLUM zone represent the percentage of trips originating from that zone for which the analysis mode is the "primary mode of trip" recorded in the TTS. The 13 possible modes surveyed in the TTS are grouped into five analysis modes: transit, driver, passenger, bicycle, walk, other. Although the school bus mode may be similar to transit's user experience and operations, it is excluded from the transit category since GTFS schedule information does not include its services when calculating transit accessibility impedance (i.e., travel times). Likewise, users of GO Rail that do not use local public transit on any part of the trip are also excluded from the transit category because they facilitate trips to destinations beyond the scope of this thesis. Table 4 lists the original classes of modes from the survey included in each analysis mode and their respective all-day shares (6:00 – 23:55) including interregional trips.

*Table 4: Analysis Mode Shares for the Region of Waterloo*

| **Analysis Modes** | **Specific Survey Modes Included (TTS index)** | **Mode Shares** |
|---|---|---|
| Transit (excl. GO transit) | Public Transit excl. GO (B)<br>Joint GO Rail and Public Transit (J) | 4.37% |
| Non-local Transit | GO Rail (G) | 0.01% |
| Drive alone | Auto Driver (D)<br>Motorcycle (M) | 71.45% |
| Passenger | Auto Passenger (P)<br>Taxi (T)<br>Paid Rideshare (U) | 14.21% |
| Cycle | Bicycle (C) | 1.45% |
| Walk | Walk (W) | 5.65% |
| Other | School bus (S)<br>Other (O)<br>Unknown (9) | 2.86% |

Trip departure times can further divide mode shares into each analysis period, $p$, to correspond with the period-level DTA metrics. Temporal parity between trip departure times and DTA constructions may increase predictive accuracy within the mode choice model because mode choices are connected to activity types and activity types depend on the time of day. Table 5 shows the average transit shares across all PLUM zones for each period used in this analysis. “Total trips” and “average trips per non-zero zone” values are provided for reference. Total trips refer to the trips taken using each mode that depart within the given period. Trips in non-zero zones refer to the number of transit trips taken when and if a transit trip is sampled at all (transit share > 0). Excluding the weight of zero-share zones provides some indication of the concentration of transit trips within zones and periods where transit trips have been observed. Regional mode shares within each period are useful for comparison because they represent the weighted average mode shares (average zonal mode shares are unweighted, disregarding each zone’s relative number of trips). Regional shares also include trips that are missing origin identification, which represent a small share of total trips.

*Table 5: Period Mode Shares for Transit and Walking Trips in the Region of Waterloo*

| | $p$ | Average Mode Shares per Zone (%) | Average Trip Count per Non-Zero Zone | Regional Mode Shares (%) | Total Expanded Trips by Mode (all zones) | Percent of Trips without Origin ID |
|---|---|---|---|---|---|---|
| Transit Only | AM | 4.60 | 46.9 | 4.49 | 15,101.10 | 4.94 |
| | MD | 3.84 | 50.35 | 4.90 | 14,400.94 | 2.85 |
| | PM | 2.18 | 72.87 | 4.06 | 17,853.48 | 6.2 |
| | EE | 1.19 | 39.82 | 1.71 | 2,269.99 | 2.73 |
| | NT | 3.23 | 34.88 | 3.94 | 976.67 | 1.22 |
| | All Day | 4.20 | 53.94 | 4.37 | 50,602.18 | 4.62 |
| Walk Only | AM | 6.62 | 57.49 | 6.77 | 22,764.91 | 0.69 |
| | MD | 4.13 | 59.56 | 5.15 | 15,128.37 | 1.11 |
| | PM | 3.41 | 74.23 | 4.94 | 21,748.55 | 0.42 |
| | EE | 2.27 | 33.67 | 2.00 | 2,659.77 | 1.40 |
| | NT | 1.95 | 35.08 | 2.41 | 596.41 | 0 |
| | All Day | 3.29 | 60.54 | 5.65 | 62,898.01 | 0.72 |
| Transit and Walk Combined | AM | 11.22 | 52.74 | 11.26 | 37,866.01 | 2.38 |
| | MD | 7.97 | 54.68 | 10.05 | 29,529.31 | 1.96 |
| | PM | 5.59 | 73.61 | 9.00 | 39,602.03 | 3.03 |
| | EE | 3.46 | 36.25 | 3.70 | 4,929.76 | 2.01 |
| | NT | 5.18 | 34.96 | 6.35 | 1,573.08 | 0.75 |
| | All Day | 7.49 | 57.41 | 10.02 | 113,500.19 | 2.46 |

*All values refer to or are derived from expanded trip counts from the TTS.*

Dividing TTS trip samples between periods (defined in Table 3) excludes some trips based on the time of departure and whether spatial information for trip origins were sampled. Trips sampled before 6:00 or after 23:55 are excluded from this analysis but only account for a small portion of all trips. Many TTS trips are excluded because of incomplete spatial information (origin ID not attributed). 553 out of 2145 zones do not record any trips at all within any analysis period and within the 1592 zones that have at least one logged trip throughout the day, some periods have no trip information. The number of zones with a minimum of 1 sampled trip are noted for every period in Figure 10. Although the disaggregation of zonal mode share data to period-levels may better define the context for many trips, zones and periods from which fewer trips originate suffer more from sampling error. Therefore, the average trip count for zones that have non-zero trips (by the respective mode) are noted for each period as an indicator of the typical sample by which zonal mode shares are calculated. Mode shares for sampled trips without origin ID attributes are included for reference in Table 4 and counted within mode shares at the Regional aggregation. However, since transit accessibility is spatially explicit, aspatial trips are not considered when seeking zonal correlations with mode shares in the analysis.

The final number of zones regressed in each period are only those that have non-zero attraction terms and, of those, zones with non-zero trips within that period. Table 6 notes the number of zones that have data in each measure and period, representing the number of samples in regression results. D- and C-measures have identical sample zones because both DTA measures only have destination-based attraction terms; that is, an origin zone’s DTA is measured based on the destinations it can access. Since there is no interaction term describing the origin (i.e., an origin with 0 population would be excluded), these measures always have a non-zero accessibility value.

*Table 6: Spatial Sample Size (zone count) by DTA Measure and Analysis Period*

| Analysis Period ($p$) | G1/G2 | D1/D2/C1/C2 |
|---|---|---|
| AM | 1001 | 1221 |
| MD | 949 | 1341 |
| PM | 926 | 1369 |
| EE | 708 | 993 |
| NT | 274 | 415 |
| AD | 1071 | 1592 |

For all regression models predicting mode shares in this thesis, the combined transit and walk modes shares are used as the “transit” mode shares for each zone. Figure 10 shows the spatial distribution of the combined transit and walking mode shares. Zones with higher combined shares are expectedly concentrated around the three downtown areas, with lower transit shares and fewer trips in general during later periods of the day (zones without trips by any mode are removed).

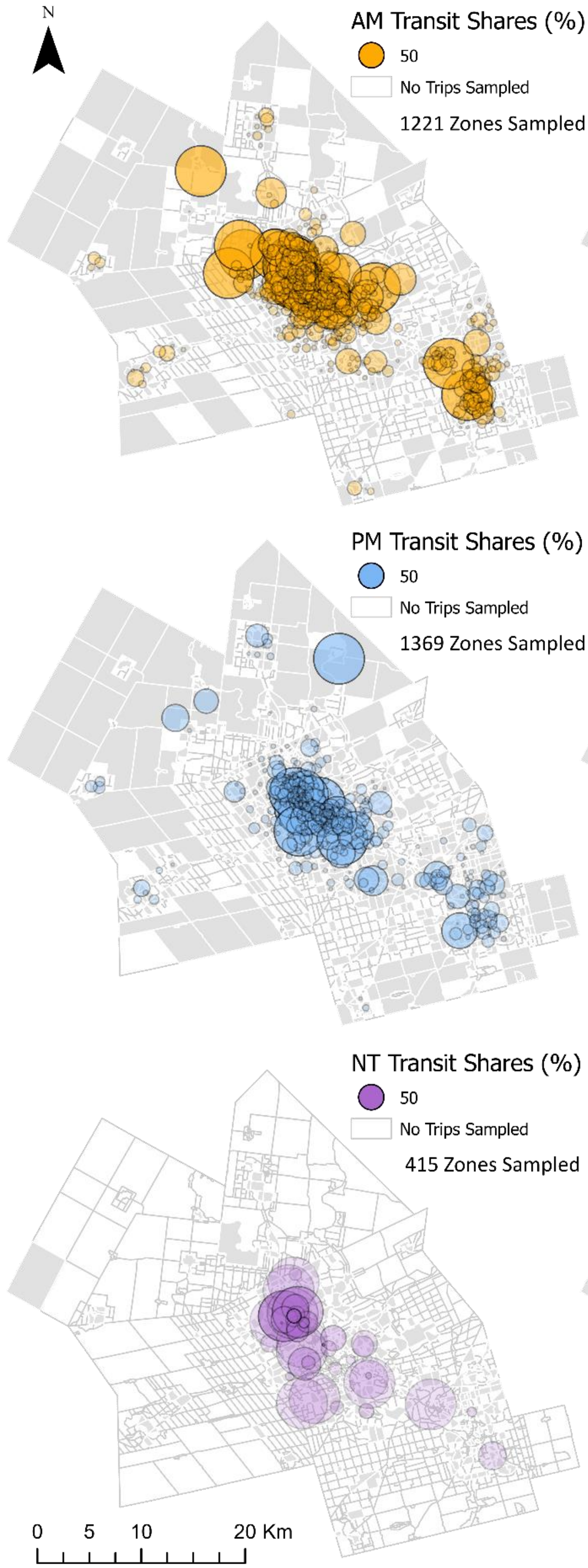



*Figure 10: Zonal Combined Transit & Walk Mode Shares (proportional sizes) by Analysis Period*

## 4.4 Measures of Effectiveness

DTA's effectiveness as an explanatory variable for transit mode shares is measured through an ordinary least squares (OLS) regression analysis to predict zonal mode shares using two independent variables related to the magnitude and variation of transit accessibility (see "DTA Magnitude and Dispersion" subsection): $DTA_m$ and $DTA_s$. Regression outputs, including the $p$-value, coefficient of determination ($R^2$), and sign of the fitted model are used to determine the statistical significance, strength of correlation, and confirm or reject the hypothesis.

### *4.4.1 OLS Regression*

Ordinary least squares regression is used to measure the relationship between zonal transit shares and DTA. This process is repeated for every period of analysis and measure of DTA, testing both DTA magnitude and dispersion. If the resulting $p$-value is less than 0.05 (alpha value), the DTA measure is statistically significant at the 95% confidence level. If DTA has a statistically significant relationship with zonally aggregated transit shares, it is possible that people also perceive and react to transit accessibility values within their individual decision processes. Goodness-of-fit is measured by the coefficient of determination, $R^2$, which has a value between 0 and 1. A higher value indicates a better fit because $R^2$ is the variation of mode shares that are explained by DTA divided by the total variation in mode shares. An $R^2$ value of 1 means that all mode share values can be explained by the DTA measure. Finally, the sign of the fitted model (positive or negative) indicates whether people respond positively to $DTA_m$ and negatively to $DTA_s$, as hypothesized. Regressing mode shares with DTA across different time periods, measure classes, and attraction terms (Table 3) may justify the use of $DTA_m$ and/or $DTA_s$ measures as explanatory variables in a mode choice model.

### *4.4.2 Transformations for Linearity*

The use of a linear regression model requires some assumptions about the relationship between zonal DTA measures and mode shares. The assumptions are that 1) the relationship is linearly structured, 2) errors are homoscedastic (variance of dependent variable $y$ is the same for any value of independent variable $x$), 3) errors are normally distributed, and 4) observations are independent from each other (Box & Cox, 1964). Sometimes, the terms of the original observations do not meet these assumptions (specifically, homoscedasticity) so a transformation of either the dependent or independent variable may be necessary. Since the true relationship between DTA (magnitude and dispersion) and mode shares is unknown, this research uses the Box-Cox transformation to estimate the appropriate non-linear transformation for an approximately normal error distribution of $x$ (DTA magnitude and dispersion). The Box-Cox transformation defines a family of non-linear transformations $x^{(\lambda)}$ as follows:

$$x^{(\lambda)} = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \log(x), & \lambda = 0 \end{cases} \qquad [4.5]$$

The Box-Cox transformation is a scaled form of Tukey's ladder of transformations. For non-zero values of $\lambda$, the Box-Cox applies the same power transformations as Tukey's ladder because an analysis of variance (ANOVA) is unaffected by linear transformations (-1 in numerator and division by $\lambda$ are both scalar transformations of $x^{\lambda}$) (Box & Cox, 1964, p. 214). Thus, for the purposes of conducting an ANOVA, equation [4.5] is equivalent to Tukey's ladder of transformations in Table 7.

Unique to the Box-Cox transformation is that it is continuous at $\lambda = 0$. Since $x^{\lambda}$ is undefined at $\lambda = 0$, the formula can be rewritten in indeterminate form (as $\lambda \to 0$) and simplified to log(x) (Scott, n.d.).

$$x' = \frac{e^{\lambda \log(x)} - 1}{\lambda} \cong \frac{\left(1 + \lambda \log(x) + \frac{1}{2}\lambda^2 \log(x)^2 + \cdots\right) - 1}{\lambda} = \log(x) \quad [4.6]$$

Transformations applied to $x$ are indexed to values of $\lambda$ in Table 7, which provides a legend for common, non-linear transformations that result in $x'$. The natural logarithm ($ln(x)$) is used for the logarithmic transformation of $x$ at $\lambda = 0$, following the use of ln-transformed gravity-based transit accessibility metrics in literature (Higgins & Kanaroglou, 2018).

*Table 7: Box Cox Power Transformation per Index Value ($\lambda$)*

| $\lambda$ | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| $x'$ | $x^{-3}$ | $x^{-2}$ | $x^{-1}$ | $\ln(x)$ | $x$ | $x^2$ | $x^3$ |

This research uses a Python software package from the SciPy library to apply the Box-Cox transformation (scipy.stats.boxcox). The package estimates the optimal value of $\lambda$ by maximum likelihood, which searches for a value of $\lambda$ ($-5 \leq \lambda \leq 5$) that minimizes the residual sum of squares. Optimal values of $\lambda$ (in grey text) were rounded to the nearest integer to simplify the transformation for interpretation and ease of reverse transformation (transformed values ($x'$) must be transformed back to original scale ($x$) for prediction). Table 8 notes the rounded values used for transformation of every DTA measure and the unrounded $\lambda$ estimations. Untransformed values ($\lambda = 1$) are shaded in grey for reference. Most transformations are $\lambda = 0$ (i.e., $x' = \ln(x)$, where $x'$ are transformed DTA measures).

*Table 8: Box-Cox Transformation Index Values ($\lambda$) for DTA Measures*

| | G1 | | G2 | | D1 | | D2 | | C1 | | C2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mag. | Disp. | Mag. | Disp. | Mag. | Disp. | Mag. | Disp. | Mag. | Disp. | Mag. | Disp. |
| AM | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 0.17 | 0.20 | 0.18 | 0.20 | -0.01 | 0.37 | 0.01 | 0.40 | 0.47 | 0.47 | 0.48 | 0.46 |
| MD | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| | 0.18 | 0.21 | 0.04 | 0.05 | -0.03 | 0.29 | -0.02 | 0.14 | 0.47 | 0.46 | 0.50 | 0.51 |
| PM | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| | 0.08 | 0.08 | 0.07 | 0.08 | 0.05 | 0.60 | 0.04 | 0.58 | 0.45 | 0.45 | 0.45 | 0.45 |
| EE | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| | 0.09 | 0.09 | 0.14 | 0.14 | 0.14 | 0.54 | 0.13 | 0.54 | 0.50 | 0.45 | 0.51 | 0.45 |
| NT | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| | 0.07 | 0.12 | 0.12 | 0.18 | 0.04 | 0.72 | 0.04 | 0.72 | 0.50 | 0.44 | 0.50 | 0.44 |
| AD | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 0.15 | 0.18 | 0.16 | 0.18 | -0.03 | 0.38 | -0.02 | 0.40 | 0.39 | 0.40 | 0.40 | 0.40 |

## 4.5 Results and Discussion: DTA

All measures of DTA are regressed with the mode shares of trips departing from a given zone during the respective analysis period, $p$, for temporal parity between transit accessibility and activities. Note that comparison between the different measure times (i.e., gravity-based (G), time-decayed opportunity (D), and cumulative opportunity (C)) is limited because of some transformations applied to the unitless values. DTA values for the same measure types also cannot be compared between periods that use different attraction components. Therefore, the different DTA evaluations will be scrutinized based on their respective statistical relationships with zonal mode shares, and only slope signs (positive or negative) are relevant for describing effects.

### 4.5.1 Gravity-based Measures

G1 measures are correlated with period $N$ mode shares in Figure 10. Regression results are in Table 9. Gravity-based measure G1 includes $POP_i$ and $EMP_j$ interaction terms for every period of analysis. Results are plotted by period to visualize the $DTA_m$ (blue fitted lines) and $DTA_s$ (orange fitted lines) relationships with mode shares.

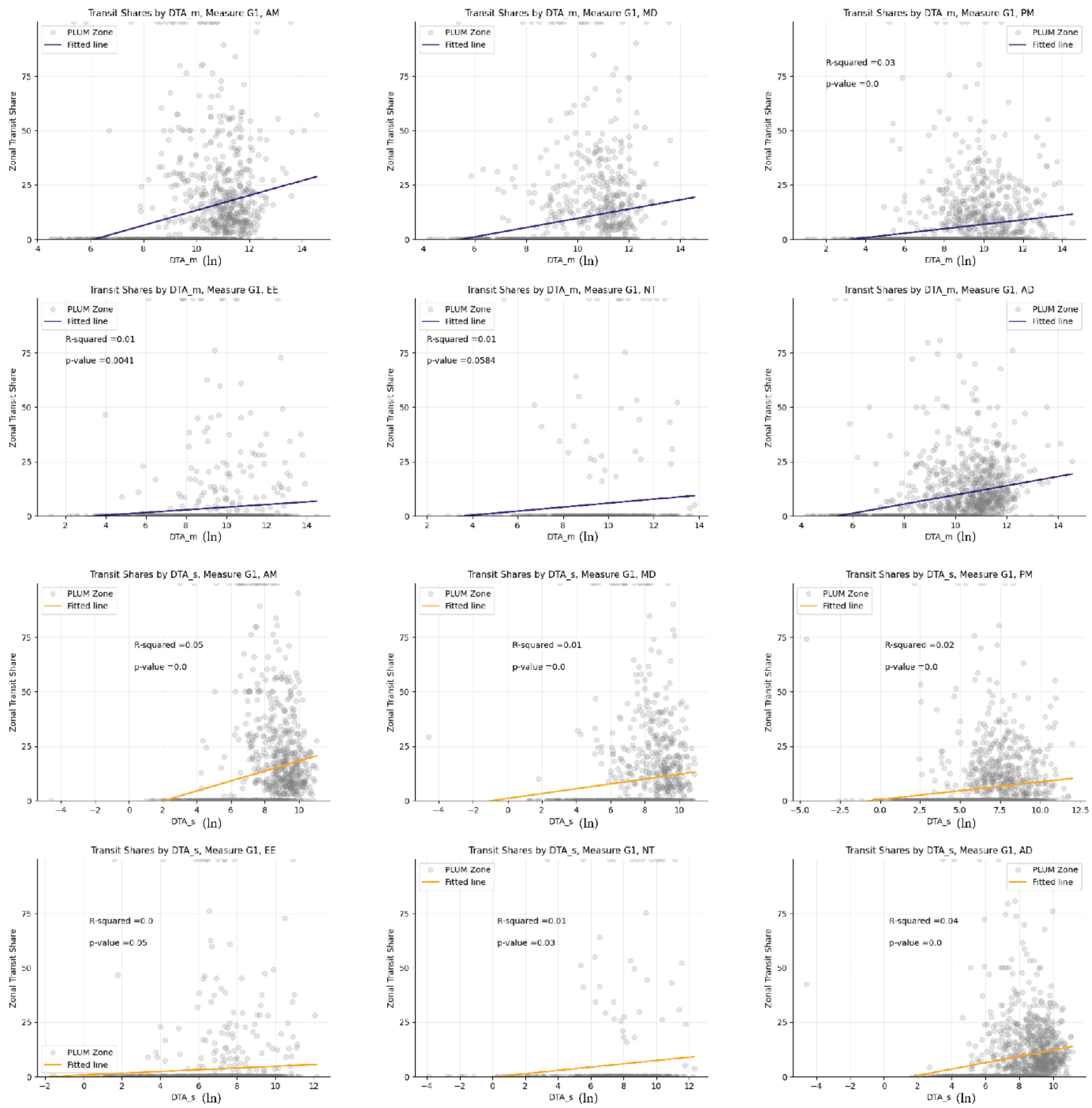


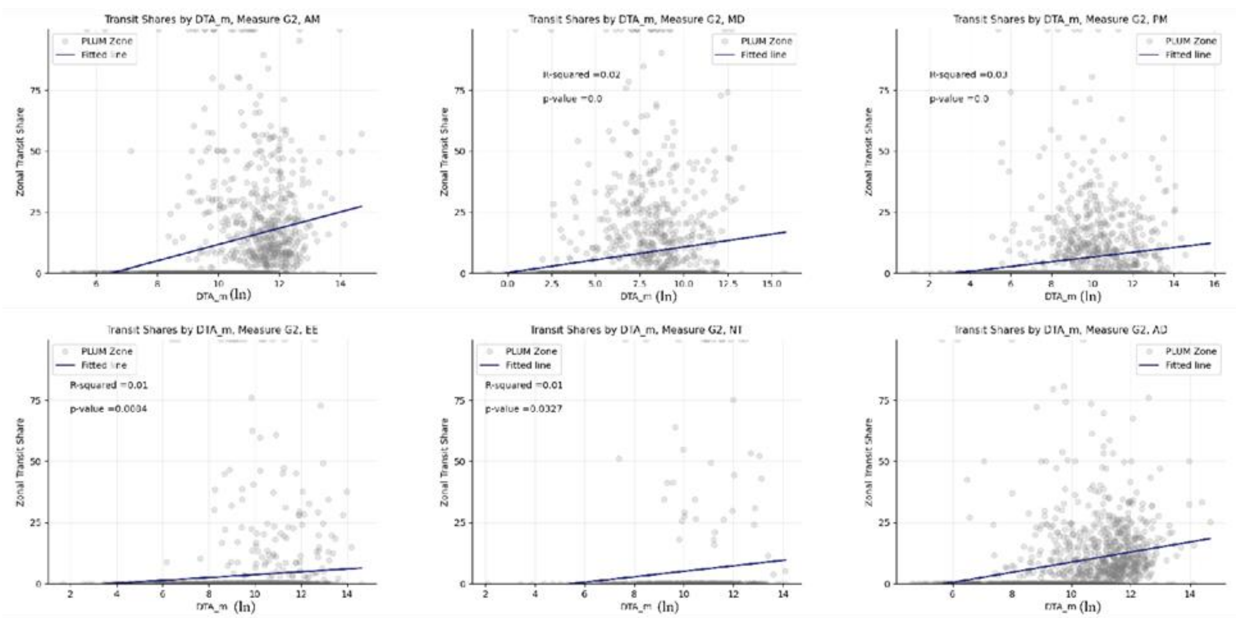

*Figure 11: Plots of Zonal Transit Shares by DTA and Period, G1*

*Table 9: G1 Regression Analysis Results*

| DTA Measure | Period | $DTA_m$ p-value | $DTA_m$ $R^2$ | $DTA_m$ slope | $DTA_s$ p-value | $DTA_s$ $R^2$ | $DTA_s$ slope |
|---|---|---|---|---|---|---|---|
| G1 | AM | 0.000 | 0.08 | 3.4 | 0.000 | 0.05 | 2.3 |
| G1 | MD | 0.000 | 0.04 | 2.1 | 0.000 | 0.01 | 1.1 |
| G1 | PM | 0.000 | 0.03 | 1.0 | 0.000 | 0.02 | 0.8 |
| G1 | EE | 0.004 | 0.01 | 0.6 | 0.046 | 0.00 | 0.4 |
| G1 | NT | ***0.058*** | 0.01 | 0.9 | 0.031 | 0.01 | 0.8 |
| G1 | AD | 0.000 | 0.09 | 2.1 | 0.000 | 0.04 | 1.5 |

$DTA_m$ using measure G1 is statistically significant for all periods outside of the night (22:00-23:55), perhaps due to the relatively fewer samples collected (274 zones). Slope signs for $DTA_m$ are also consistent with the hypothesis that transit shares positively correlate with $DTA_m$; however, slope signs reveal and unexpected positive correlation between $DTA_s$ and zonal mode shares. The explanatory power across all periods is consistently low for both $DTA_m$ ($R^2 \leq 0.09$) and $DTA_s$ ($R^2 \leq 0.05$). Whether this result is consistent across measures provides additional insight.

G2 measures are correlated with period $N$ mode shares in Figure 10. Regression results are in Table 10.



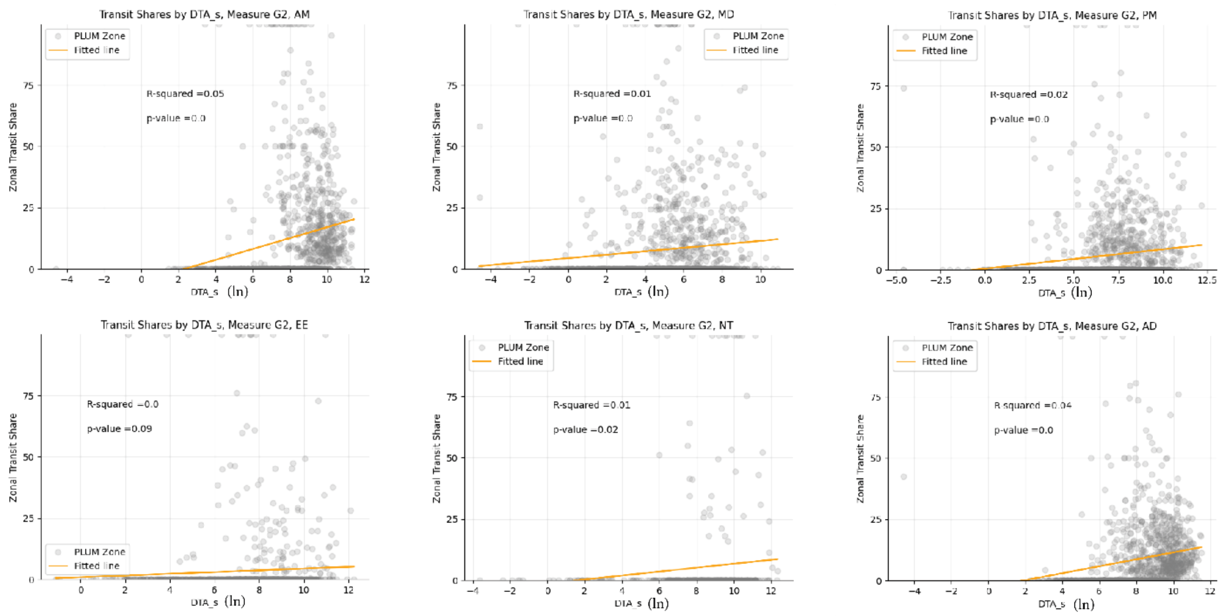



*Figure 12: Plots of Zonal Transit Shares by DTA and Period, G2*

*Table 10: G2 Regression Analysis Results*

| DTA Measure | Period | $DTA_m$ p-value | $DTA_m$ $R^2$ | $DTA_m$ slope | $DTA_s$ p-value | $DTA_s$ $R^2$ | $DTA_s$ slope |
|---|---|---|---|---|---|---|---|
| G2 | AM | 0.000 | 0.08 | 3.3 | 0.000 | 0.05 | 2.2 |
| G2 | MD | 0.000 | 0.02 | 1.1 | 0.000 | 0.01 | 0.7 |
| G2 | PM | 0.000 | 0.03 | 1.0 | 0.000 | 0.02 | 0.8 |
| G2 | EE | 0.008 | 0.01 | 0.6 | ***0.089*** | 0.00 | 0.4 |
| G2 | NT | 0.033 | 0.01 | 1.1 | 0.024 | 0.01 | 0.8 |
| G2 | AD | 0.000 | 0.08 | 2.1 | 0.000 | 0.04 | 1.4 |

G2 includes discretionary trip attraction terms ($DIS$) throughout the day, differing with G1 in only this regard. While most results are the same, early evening $DTA_s$ no longer has a statistically significant relationship with zonal mode shares. EE's attraction terms are $EMP_i + DIS_i$ at the origin and $POP_j$ at the destination, meaning that the change in the origin attraction term affected the mode share relationship. In contrast, adding the $DIS$ term to the destination attraction term during the PM period did not; $DTA_m$ during the PM remained significant, but no better a predictor.

### 4.5.2 Time-Decayed Opportunity

D1 measures are correlated with period $N$ mode shares in Figure 10. Regression results are in Table 11. Time-decayed opportunity measure D1 includes only destination-based attraction terms. Results are plotted by period to visualize the $DTA_m$ (blue fitted lines) and $DTA_s$ (orange fitted lines) relationships with mode shares.

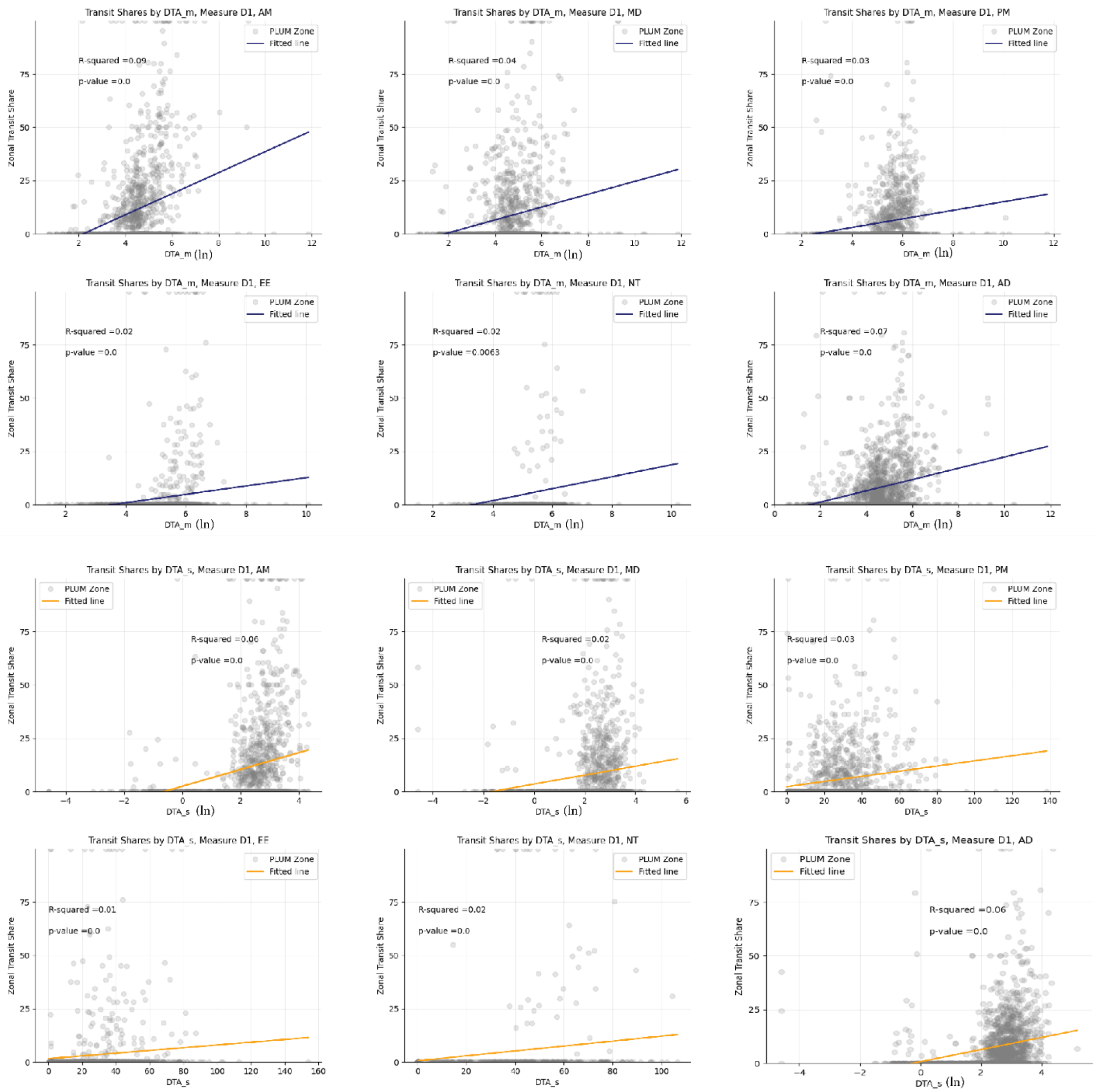



*Figure 13: Plots of Zonal Transit Shares by DTA and Period, D1*

*Table 11: D1 Regression Analysis Results*

| DTA Measure | Period | $DTA_m$ p-value | $DTA_m$ $R^2$ | $DTA_m$ slope | $DTA_s$ p-value | $DTA_s$ $R^2$ | $DTA_s$ slope |
|---|---|---|---|---|---|---|---|
| D1 | AM | 0.000 | 0.09 | 4.9 | 0.000 | 0.06 | 3.9 |
| D1 | MD | 0.000 | 0.04 | 3.0 | 0.000 | 0.02 | 2.1 |
| D1 | PM | 0.000 | 0.03 | 2.0 | 0[1] | 0.027[1] | 0.121[1] |
| D1 | EE | 0.000 | 0.02 | 2.0 | 0.0036[1] | 0.009[1] | 0.064[1] |
| D1 | NT | 0.006 | 0.02 | 2.8 | 0.0031[1] | 0.021[1] | 0.115[1] |
| D1 | AD | 0.000 | 0.07 | 2.6 | 0.000 | 0.06 | 2.8 |

[1] These shaded values are untransformed (no log transformation) as per Figure 8.

D1 performs similarly to measure G1 despite the complete exclusion of the origin attraction terms. D1's magnitude ($DTA_m$) is significant for all periods including NT, in contrast to G1. Since the AM attraction component only includes destination employment ($EMP_j$), this suggests that users who are travelling to employment are unconcerned with the population characteristics of the zone from which they originate. D1 $DTA_s$ slope signs are also positive, corroborating the positive relationship between mode shares and transit service dispersion apparent in G- measures.

*D2 measures are correlated with period N mode shares in Figure 10. Regression results are in Table 12.*

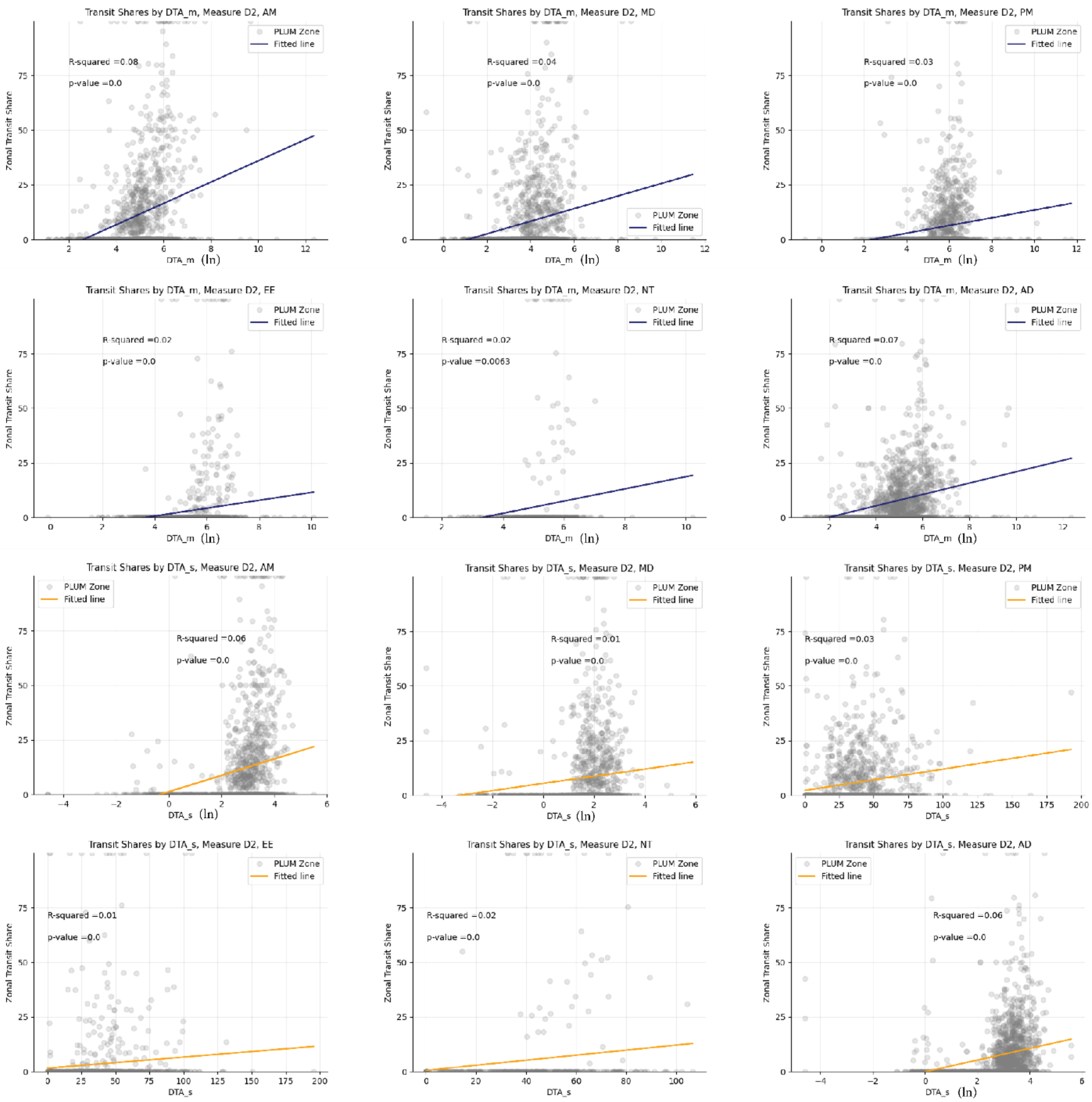



*Figure 14: Plots of Zonal Transit Shares by DTA and period, D2*

*Table 12: D2 Regression Analysis Results*

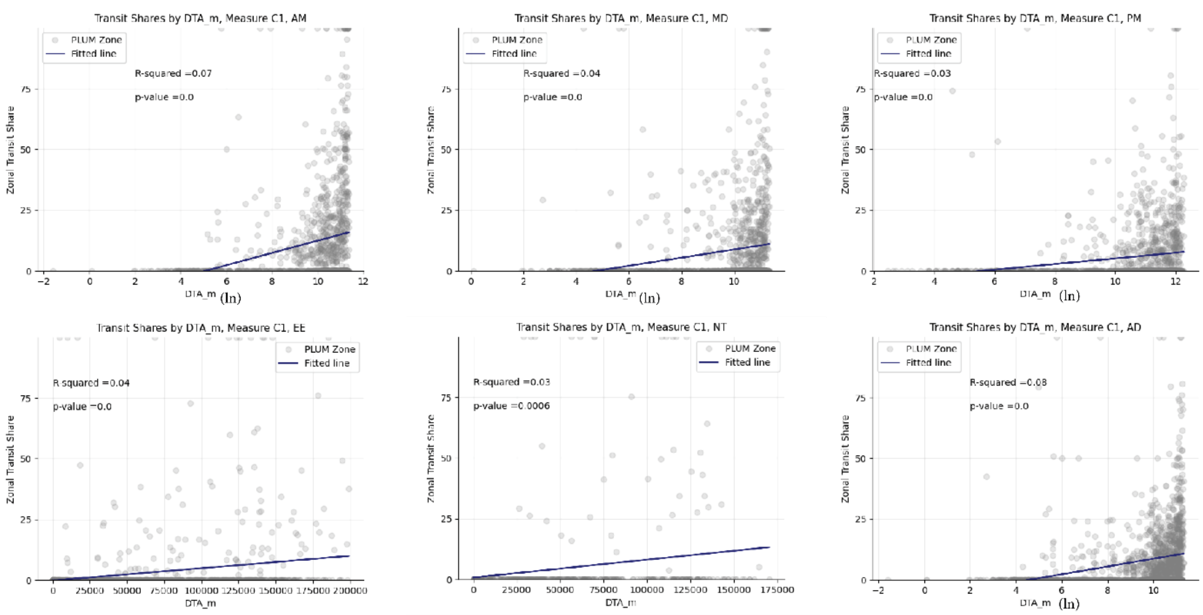
| DTA Measure | Period | $DTA_m$ p-value | $DTA_m$ $R^2$ | $DTA_m$ slope | $DTA_s$ p-value | $DTA_s$ $R^2$ | $DTA_s$ slope |
|---|---|---|---|---|---|---|---|
| D2 | AM | 0.000 | 0.08 | 4.9 | 0.000 | 0.06 | 3.8 |
| D2 | MD | 0.000 | 0.04 | 2.9 | 0.000 | 0.01 | 1.6 |
| D2 | PM | 0.000 | 0.03 | 1.8 | 0[1] | 0.03[1] | 0.098[1] |
| D2 | EE | 0.000 | 0.02 | 1.8 | 0.003[1] | 0.009[1] | 0.051[1] |
| D2 | NT | 0.006 | 0.02 | 2.8 | 0.0031[1] | 0.021[1] | 0.115[1] |
| D2 | AD | 0.000 | 0.07 | 2.6 | 0.000 | 0.06 | 2.7 |

[1] These shaded values are untransformed (no log transformation) as per Figure 8.

D2 differs from D1 by including discretionary attraction terms in all periods, except for NT (D2's NT is the same as D1's NT). Specifically, the $DIS$ values are added to AM and PM destinations, and $DIS$ is the sole attraction term during the MD. D2 has similar results to D1, which may be attributed to the spatial correlation between discretionary and employment trip destinations. Compared with the G2 measure, the removal of the $(EMP_i + DIS_i)$ origin term in D- class measures results in a statistically significant relationship with mode shares in the EE. This finding supports the motivation for using the D-class measures (i.e., ignoring the origin node trip production terms).

### 4.5.3 Cumulative Opportunity

C1 measures are correlated with period $N$ mode shares in Figure 10. Regression results are in Table 13. C1's $DTA_m$ values are untransformed for the EE and NT periods, where the x-axis is the sum of opportunities within 40 minutes of each zone. Results are plotted by period to visualize the $DTA_m$ (blue fitted lines) and $DTA_s$ (orange fitted lines) relationships with mode shares.



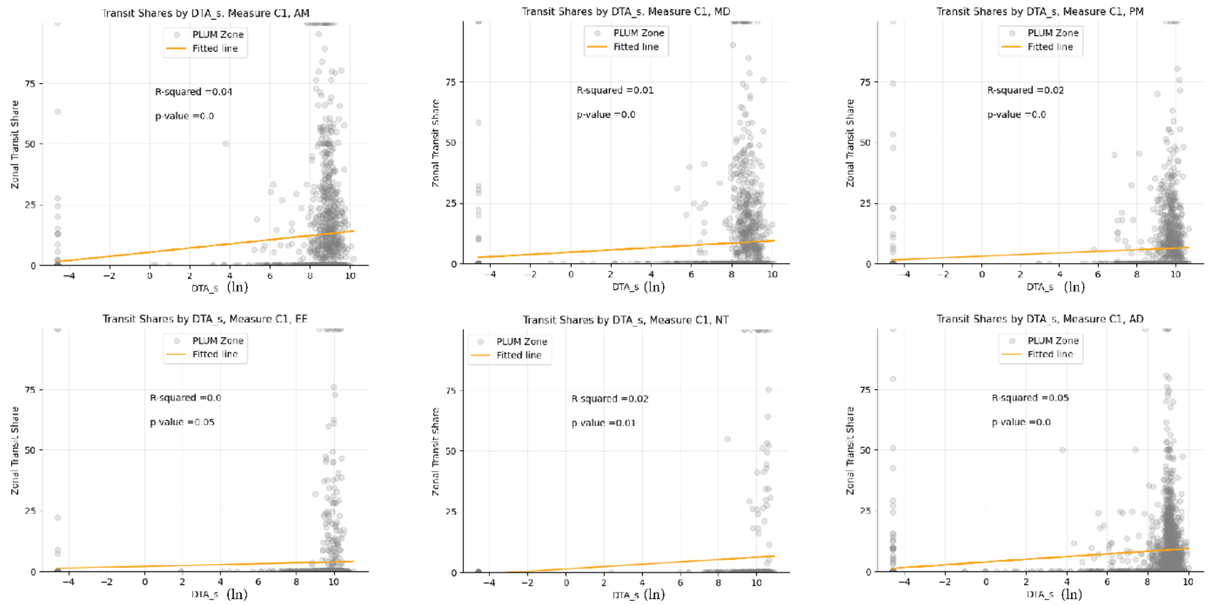



*Figure 15: Plots of Zonal Transit Shares by DTA and period, C1*

*Table 13: C1 Regression Analysis Results*

| DTA Measure | Period | $DTA_m$ p-value | $DTA_m$ $R^2$ | $DTA_m$ slope | $DTA_s$ p-value | $DTA_s$ $R^2$ | $DTA_s$ slope |
|---|---|---|---|---|---|---|---|
| C1 | AM | 0.000 | 0.07 | 2.5 | 0.000 | 0.04 | 0.9 |
| C1 | MD | 0.000 | 0.04 | 1.7 | 0.000 | 0.01 | 0.5 |
| C1 | PM | 0.000 | 0.03 | 1.2 | 0.000 | 0.02 | 0.3 |
| C1 | EE | 0[1] | 0.037[1] | 0[1] | 0.047 | 0.00 | 0.2 |
| C1 | NT | 0.0006[1] | 0.028[1] | 0[1] | 0.012 | 0.02 | 0.5 |
| C1 | AD | 0.000 | 0.08 | 2.5 | 0.000 | 0.05 | 0.9 |

[1] These shaded values are untransformed (no log transformation) as per Figure 8.

C1 $DTA_m$ has a statistically significant relationship with mode shares for all periods. This is consistent with findings from the time-decayed opportunity measures (D1, D2) that the attraction of the origin zone may be irrelevant to transit choices. That C1 is a significant measure also suggests that the 40-minute time boundary captures destinations relevant to the common transit user.

*C2 measures are correlated with period N mode shares in Figure 10. Regression results are in Table 14.*

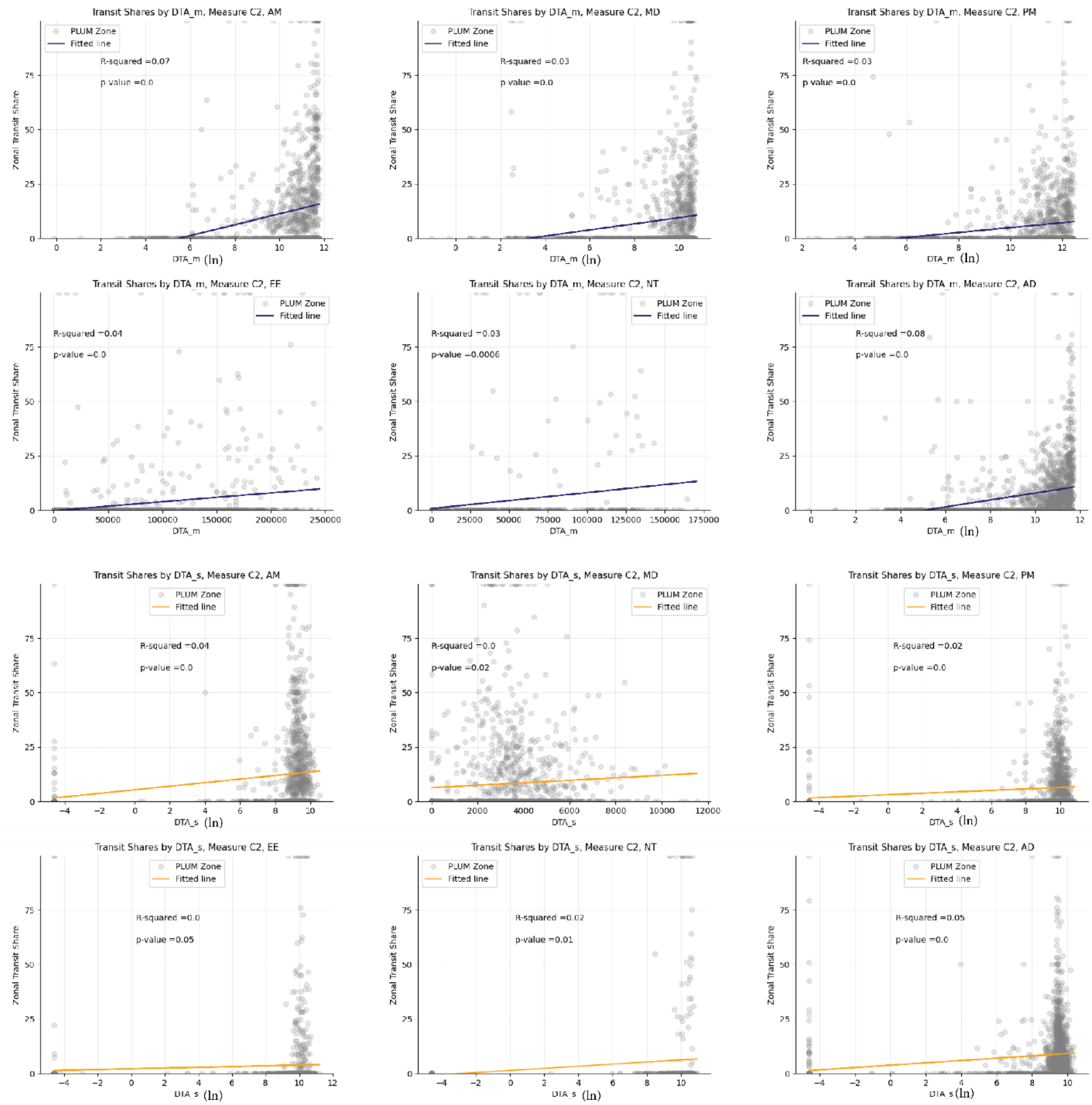



*Figure 16: Plots of Zonal Transit Shares by DTA and period, C2*

*Table 14: C2 Regression Analysis Results*

| DTA Measure | Period | $DTA_m$ p-value | $DTA_m$ $R^2$ | $DTA_m$ slope | $DTA_s$ p-value | $DTA_s$ $R^2$ | $DTA_s$ slope |
|---|---|---|---|---|---|---|---|
| C2 | AM | 0.000 | 0.07 | 1.6 | 0.000 | 0.04 | 0.6 |
| C2 | MD | 0.000 | 0.03 | 2.5 | 0.0168[1] | 0.004[1] | 0.8[1] |
| C2 | PM | 0.000 | 0.03 | 1.4 | 0.000 | 0.02 | 0.001 |
| C2 | EE | 0[1] | 0.036[1] | 1.1[1] | ***0.054*** | 0.00 | 0.3 |
| C2 | NT | 0.0006[1] | 0.028[1] | 0[1] | 0.012 | 0.02 | 0.2 |
| C2 | AD | 0.000 | 0.08 | 0 | 0.000 | 0.05 | 0.5 |

[1] These shaded values are untransformed (no log transformation) as per Figure 8.

C2 results are very similar to C1 despite the use of discretionary trip attraction terms. This may be related to the concentration and/or spatial correlation of employment and retail/service sector destinations or the use of a generous time threshold ($\alpha = 40$ minutes). Although $DTA_m$ values are statistically significant, it is possible that the transit travel time threshold of 40 minutes is a large boundary for travel within the Region. If there is no continuous differentiation in destination attractiveness based on a time- or distance-decay measure, C1 and C2 measures may not adequately differentiate between zonal behaviours. Finally, the addition of $DIS_j$ to C1's $POP_j$ in the EE period seems to make the $DTA_s$ lose significance; however, the D2 dispersion measure is significant despite the inclusion of $DIS_j$. This suggests that shorter trips (≤ 40 minutes) may not be as impacted by transit accessibility dispersion as longer trips during the EE.

### 4.5.4 Results Summary

Table 15 summarizes linear regression analysis results for each DTA measure and period. Keeping measure classes constant, DTA yields similar predictive power and signs of slope across all periods. For all the different periods and attraction terms studied, statistical outcomes describe the relationship between aggregate mode choices and each type of DTA ($DTA_m$, $DTA_s$).

*Table 15: Regression Analysis Results Summary*

| | | $DTA_m$ | | | | $DTA_s$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| DTA Measure | Period | p-value | $R^2$ | Slope | Zones (n) | p-value | $R^2$ | Slope | Zones (n) |
| G1 | AM | 0.000 | 0.08 | 3.4 | 1002 | 0.000 | 0.05 | 2.3 | 1002 |
| G1 | MD | 0.000 | 0.04 | 2.1 | 950 | 0.000 | 0.01 | 1.1 | 950 |
| G1 | PM | 0.000 | 0.03 | 1.0 | 1296 | 0.000 | 0.02 | 0.8 | 1296 |
| G1 | EE | 0.004 | 0.01 | 0.6 | 939 | 0.046 | 0.00 | 0.4 | 939 |
| G1 | NT | ***0.058*** | 0.01 | 0.9 | 399 | 0.031 | 0.01 | 0.8 | 399 |
| G1 | AD | 0.000 | 0.09 | 2.1 | 1072 | 0.000 | 0.04 | 1.5 | 1072 |
| G2 | AM | 0.000 | 0.08 | 3.3 | 1002 | 0.000 | 0.05 | 2.2 | 1002 |
| G2 | MD | 0.000 | 0.02 | 1.1 | 1268 | 0.000 | 0.01 | 0.7 | 1268 |
| G2 | PM | 0.000 | 0.03 | 1.0 | 1298 | 0.000 | 0.02 | 0.8 | 1298 |
| G2 | EE | 0.008 | 0.01 | 0.6 | 981 | ***0.089*** | 0.00 | 0.4 | 981 |
| G2 | NT | 0.033 | 0.01 | 1.1 | 413 | 0.024 | 0.01 | 0.8 | 413 |
| G2 | AD | 0.000 | 0.08 | 2.1 | 1072 | 0.000 | 0.04 | 1.4 | 1072 |
| D1 | AM | 0.000 | 0.09 | 4.9 | 1221 | 0.000 | 0.06 | 3.9 | 1221 |
| D1 | MD | 0.000 | 0.04 | 3.0 | 1341 | 0.000 | 0.02 | 2.1 | 1341 |
| D1 | PM | 0.000 | 0.03 | 2.0 | 1367 | 0[1] | 0.027[1] | 0.121[1] | 1367 |
| D1 | EE | 0.000 | 0.02 | 2.0 | 992 | 0.0036[1] | 0.009[1] | 0.064[1] | 992 |
| D1 | NT | 0.006 | 0.02 | 2.8 | 413 | 0.0031[1] | 0.021[1] | 0.115[1] | 413 |
| D1 | AD | 0.000 | 0.07 | 2.6 | 1592 | 0.000 | 0.06 | 2.8 | 1592 |
| D2 | AM | 0.000 | 0.08 | 4.9 | 1221 | 0.000 | 0.06 | 3.8 | 1221 |
| D2 | MD | 0.000 | 0.04 | 2.9 | 1341 | 0.000 | 0.01 | 1.6 | 1341 |
| D2 | PM | 0.000 | 0.03 | 1.8 | 1369 | 0[1] | 0.03[1] | 0.098[1] | 1369 |
| D2 | EE | 0.000 | 0.02 | 1.8 | 993 | 0.003[1] | 0.009[1] | 0.051[1] | 993 |
| D2 | NT | 0.006 | 0.02 | 2.8 | 413 | 0.0031[1] | 0.021[1] | 0.115[1] | 413 |
| D2 | AD | 0.000 | 0.07 | 2.6 | 1592 | 0.000 | 0.06 | 2.7 | 1592 |
| C1 | AM | 0.000 | 0.07 | 2.5 | 1221 | 0.000 | 0.04 | 0.9 | 1221 |
| C1 | MD | 0.000 | 0.04 | 1.7 | 1341 | 0.000 | 0.01 | 0.5 | 1341 |
| C1 | PM | 0.000 | 0.03 | 1.2 | 1364 | 0.000 | 0.02 | 0.3 | 1364 |

| DTA Measure | Period | $DTA_m$ p-value | $R^2$ | Slope | Zones (n) | $DTA_s$ p-value | $R^2$ | Slope | Zones (n) |
|---|---|---|---|---|---|---|---|---|---|
| C1 | EE | 0[1] | 0.037[1] | 0[1] | 991 | 0.047 | 0.00 | 0.2 | 991 |
| C1 | NT | 0.0006[1] | 0.028[1] | 0[1] | 412 | 0.012 | 0.02 | 0.5 | 412 |
| C1 | AD | 0.000 | 0.08 | 2.5 | 1592 | 0.000 | 0.05 | 0.9 | 1592 |
| C2 | AM | 0.000 | 0.07 | 1.6 | 1221 | 0.000 | 0.04 | 0.6 | 1221 |
| C2 | MD | 0.000 | 0.03 | 2.5 | 1333 | 0.0168[1] | 0.004[1] | 0.8[1] | 1333 |
| C2 | PM | 0.000 | 0.03 | 1.4 | 1369 | 0.000 | 0.02 | 0.001 | 1369 |
| C2 | EE | 0[1] | 0.036[1] | 1.1[1] | 993 | ***0.054*** | 0.00 | 0.3 | 993 |
| C2 | NT | 0.0006[1] | 0.028[1] | 0[1] | 412 | 0.012 | 0.02 | 0.2 | 412 |
| C2 | AD | 0.000 | 0.08 | 0 | 1592 | 0.000 | 0.05 | 0.5 | 1592 |

[1] These shaded values are untransformed (no log transformation) as per Figure 8.

Overall, all measures have low explanatory power across all periods. The low explanatory power is not surprising because DTA only constitutes a single possible variable in mode choice decisions. DTA only represents some transit system (modal) and land use factors (related to trip purpose) but excludes many other environmental, personal, and modal characteristics that are normally considered in modal split analysis.

Statistical outcomes only slightly differed between the measures ending in 1 (no discretionary trips) and measures ending in 2 (discretionary trip attractors included). The similar results between measures ending in 1 and 2 suggests that DIS attraction terms at the destination do not generally improve predictive ability and that employment-destined trips during the AM are most sensitive to average transit accessibility values. One exception seems to be the addition of origin attractor $DIS_i$ in G2 at NT, which makes $DTA_m$ significant (G1 $DTA_m$ is not significant). This supports the inclusion of discretionary trip origins for gravity-based formulations of DTA, especially because $DIS$ trips constitute most EE trips (see Figure 9).

Class D- measures are useful for understanding the influence of destination trip attractors because they exclude origin-based attraction terms from the measurement. D1 performs similarly to G1 despite completely removing the origin attraction term from the measure. During the NT period, D1 $DTA_m$ ($\rightarrow POP_j$) is significant although G1 is not. Since D1 uses the same destination $j$ attraction terms as G1, similar performance supports the original intent of applying this measure class, which is to understand whether origin zone "attractiveness" is relevant to user decision making (see "Selecting Measure Classes for Testing" subsection). The finding that D- measures offer comparable performance to G- measures agrees with literature that criticizes the inclusion of origin-based attraction terms in transit accessibility measurement because of its disconnect from travel motivations (Kutter, 1972).

The best performing (highest correlation coefficient) periods for all measures are the AD and AM, which share the same attraction terms, generally to $EMP_j$. This suggests that aggregate transit shares may not be as sensitive to the trip times as much as it is to trip purpose (attraction terms). All-day analysis is also coarser than the period-level analyses because it uses constant attraction terms throughout the day. Despite its insensitivity to varying trip purpose demands over time, all-day analysis allows comparison of transit services across periods. Thus, all-day $DTA_m$ and $DTA_s$ describe zonal mode share's relationship with inter-period service changes. Compared with period-level $DTA_m$, AD is statistically significant using all measure classes and attraction terms. This suggests that users may have temporally large decision frames with which they discern transit accessibility's value towards making mode choices.

## 4.6 Limitations: DTA in Aggregate Modal Analysis

Analysis of DTA's impact on mode shares is constrained by its data, methodological, and theoretical limitations. The first data limitation relates to the "Calculating Mode Shares" subsection of this research. Many zones exhibit extreme values for zonal transit shares (100% or 0%), possibly due to sampling issues inherited from the TTS dataset. A validation procedure could remove the outliers if a secondary, more reliable dataset were spatially joined and compared at the PLUM zone level. The 2016 Census data was considered but dismissed due to incongruent boundaries, larger zones (755 Census Dissemination Areas vs 2145 PLUM zones in the Region), unreported non-commuter trips, and limited observation periods (4:59-11:59 a.m. only). Travel diary surveys from previous years (e.g., 2011) have the same geographic resolution and temporal span as the current dataset but suffer similar sampling errors and merging the data risks sample duplication. Since no secondary dataset are available to systematically remove outliers, the original mode share data were kept because the extent of disaggregation in PLUM zones is very high and therefore unlikely to have dramatic impacts on the results. Data limitations also cause inaccurate representation of non-work travel demands and bar the use of alternative frameworks for transit accessibility measurement. Travel demand to discretionary activities are particularly weak in this analysis, resulting in inconsistent results between measures ending in 1 and 2 (e.g., between AM D1 and PM D2). This research uses TTS employment counts for the Retail and Service sectors to approximate the degree of attraction for non-work destinations. However, the data are not sufficiently specific or expansive enough to delineate high-value activities or capture the range of trip types that people make. The "Retail and Service" sector cannot differentiate the relative attraction of retail, dining, and leisure activities (e.g., gym, entertainment) classified therein. It also ignores other discretionary destinations, including trips to the park, groceries, or recreational programming. Unfortunately, attraction terms are difficult to specify because attractiveness may not scale with measurable terms. For example, the value of grocery stores may not scale with the number of employees, nor the value of parks with their sizes. Handy and Niemeier (1997, p. 1180) suggest measuring attractiveness using the activity's physical or economic size, the price of products, the quality of services, or the mere existence of an opportunity itself. The attractiveness of a location for discretionary activity is especially difficult to measure because discretionary activity value is up to individual arbitration. It is perhaps more related to the frequency and necessity of a trip type (user based) than to economic land use variables. Qualitative study could better reflect perceived attraction term values and improve travel demand representation in a location-based transit accessibility measure.

Methodological limitations related to the development and evaluation of the DTA metric can misrepresent transit travel behaviours. The development of the DTA metric necessitates the inclusion of walking trips between zones to create a baseline accessibility level and represent trips where walking is faster than transit; however, walking is generally perceived has a higher-friction travel mode due to the discomfort of physical exertion. This thesis does not distinguish the difference in travel by walking or transit, and therefore may overrepresent the value of zones accessible with walking. Underrepresentation of walking's higher burden is compounded by the travel time evaluation method. This thesis uses ArcGIS Pro 2.6's to accumulate total travel times along a road (sidewalk) and transit network using GTFS schedule data. The built-in network evaluator encounters two problems: it cannot limit the total number of transfers, and it cannot limit the walking time component of a transit trip separately from the total travel time (Esri, 2020). Therefore, some instances of unrealistic travel times may be included in the travel time matrices used for DTA measurement. Finally, the period-level DTA analysis uses discrete boundaries to aggregate transit accessibility measurements sampled at 5-minute intervals. Transit service parameters do not fit neatly into time periods because there are transitionary step up/down periods of service. Since the periods are aggregated and represented as a constant, service provision characteristics are not fully captured in this research. Further research could explore the effects of inter-period, transitionary transit services (e.g., between AM and MD) on mode shares.

Theoretical weaknesses of this research are related to the linear regression analysis performed, the forms of dynamic transit accessibility (i.e., transit accessibility mean, standard deviation, or otherwise), and the possibility of confounding variables underlying DTA measures. Linear regression may be an inappropriate classification method for mode share prediction using DTA because people may make transit decisions using discrete thresholds of DTA value, or if-then-else rules (e.g., choose transit if DTA is above a threshold, DTA is relevant if another requirement is met). Alternative measures of transit accessibility dispersion also warrant further examination because transit accessibility standard deviation offers different interpretations. Specifically, low $DTA_s$ values do not necessarily indicate that the transit *service* is more consistent at an origin location – only that transit *accessibility scores* are more consistent. Consistent transit services are just as undispersed as a complete lack of transit services because transit accessibility still has value at times when transit service is not available (baseline accessibility). A zone with no transit service would have uniform transit accessibility values (at the baseline) and thus zero dispersion. It would be reasonable to hypothesize a negative mode share response to high-dispersion transit accessibility in this scenario because of the complete lack of transit service. Inversely, users may also react positively (higher mode shares) to high dispersion in scenarios where they are dependent on a few high-accessibility trips. For example, a zone with a low baseline but very good transit service (e.g., express buses into a city centre) would have a high standard deviation. $DTA_s$ and $DTA_m$ values should be interpreted together to ensure the consistency of transit services is not merely the result of unavailable transit services; however, the specification of such an interaction term is left for further research.

Finally, testing mode share's sensitivity to transit accessibility in a regression model does not isolate the impacts of transit accessibility nor imply causality. The regression analysis describes a statistical relationship between DTA and transit shares, but DTA's formulation indexes many underlying land-use and transit service variables while ignoring many other mode choice factors. Further investigation in Chapter 5 may identify confounding land-use, user, or mode-specific variables embedded in DTA values that may better predict transit choices. The regression analysis from this chapter only supports the selection of a singular DTA metric (i.e., between gravity-based, cumulative opportunity, and time-decayed opportunity) for use in the mode choice model.

# Chapter 5 Rule-based Mode Choice Model

Towards a theory of behaviour that is descriptive rather than prescriptive of human behaviour, this research supports that use of non-compensatory mode choice modelling methods as an alternative to the compensatory approach because non-compensatory models are capable of representing lexicographic, dominance, and satisficing decision rules where compensatory models cannot. Synthesis of alternative decision-making theories from the review of psychological literature (see “Alternative Theories of Decision Making” section) supports this approach. The first objective of this chapter is to apply a rules-based mode choice (RBMC) model to the Region of Waterloo to demonstrate success in prediction. A secondary objective is to induce (train) a decision tree structure that yields interpretable choice processes, providing insight into the mode choice process. Third, this chapter takes one of the transit accessibility metrics developed in Chapter 4 and uses it as an additional feature to train the RBMC model.

## 5.1 RBMC Model Development: Training Decision Trees

This chapter’s methods (algorithms and decisions) are guided by the Literature Review on Alternative Theories of Decision Making and the predictive ability of supervised ML models previously applied to mode choice. Development of the RBMC involves two major steps: learning a decision tree using a training dataset and applying it to the testing (holdout) dataset. Figure 17 shows the DT training method in this chapter.

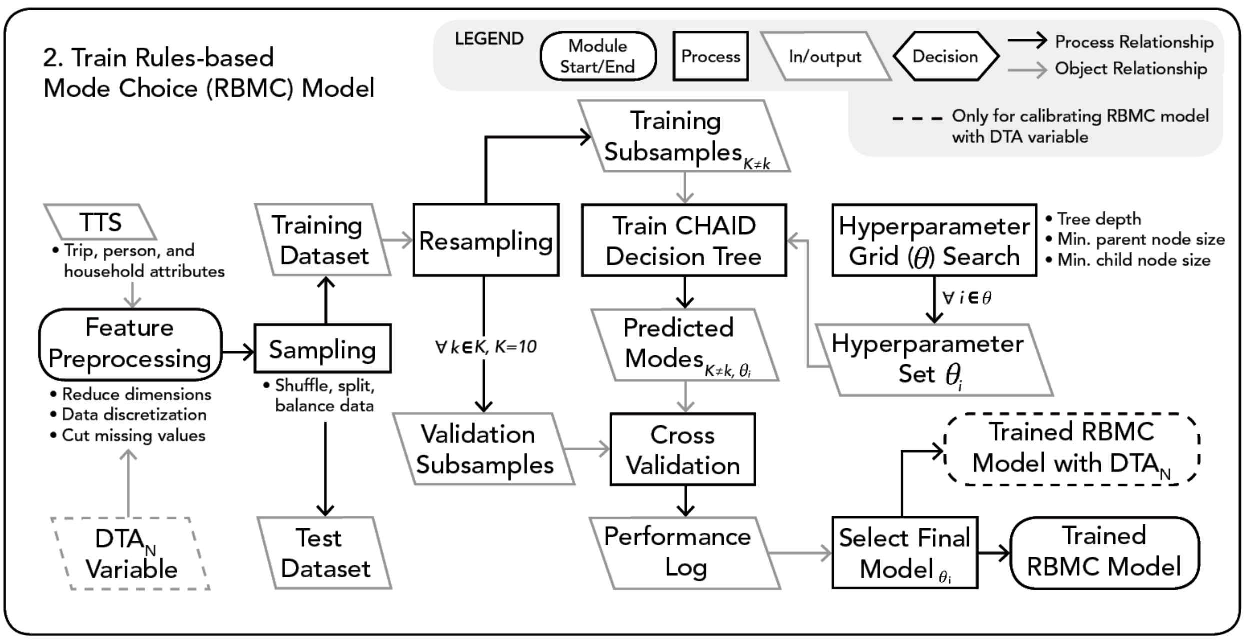



*Figure 17: Rules-based Mode Choice Model Training Method*

Broadly, the first step in modelling a decision tree is the training (tree induction) and the second step is post-processing. The initial training step involves setting the independent and dependent variables to the model and separating them into training and testing (holdout) datasets. Selecting the independent variables for DT training requires merging TTS data (i.e., household- and person-level attributes to trip attributes), creating equal-sized (percentile-based bins), and expanding the data to ensure representativeness of the population. Sampling the data into training and testing datasets requires definition of the split size, randomly shuffling the data, and then balancing the target classes (dependent variables) to ensure no bias in DT splits towards high-frequency classes. The post-processing stage occurs after training DTs with different configurations of training parameters (hyperparameters). Through cross-validation (resampling the training dataset), tree predictions are compared based on the measures of effectiveness and a final, trained tree is scored based on the thus far held out testing sample.

## 5.2 Independent Variables

The independent (explanatory) variables used in the analysis are noted in Table 16. These variables, referred to as features, each have dimensions referring to the number of unique values or categories that delimit the sample space. Figure 19 illustrates how the terminology used in DT analyses define the predictor space.

*Table 16: Independent Variables (Features)*

| Features (Label) | Dimensions | Bins | Unit |
|---|---|---|---|
| *Sociodemographic* | | | |
| Age (age) | (11-19]; (19-28]; (28-35]; (35-41]; (41-47]; (47-53]; (53-58]; (58-65]; (65-72]; (72-99] | 10 | PER |
| Sex (sex) | F: female; M: male | 2 | PER |
| Income (hh_income) | <15k; 15k-39k; 40k-59k; 60k-99k; 100k-124k; >124k; Unknown | 7 | HH |
| Employment (emp_stat) | Employed; Not_employed; Work_at_Home; 9: Unknown | 4 | PER |
| Student (stu_sat) | Student; Not_student | 2 | PER |
| Occupation (occupation) | Retail&Service; General_Office; Manufacturing; Not_employed; 9: Unknown | 5 | PER |
| Licensed (driver_lic) | Y: yes; N: no; 9: Unknown | 2 | PER |
| Vehicle ownership (hh_n_vehs) | 0; 1; 2; more than 2 | 4 | HH |
| Transit pass (tran_pass) | Y: yes; N: no; Other agency (non-GRT pass); 9: Unknown | 2 | PER |
| Free parking at work(free_park) | Y: yes; N: no; NA: not applicable; 9: Unknown | 4 | PER |
| Household size (hh_size) | (0.9-1]; (1-2]; (2-3]; (3-4]; (4-9] | 5 | HH |
| Dwelling type (hh_dwell_type) | House; Townhouse, Apartment | 3 | HH |
| *Trip-related* | | | |
| Trip Purpose | Home-based work; home-based discretionary; non-home-based | 3 | TRP |
| Trip Time | Peak: [6:00-9:35], [15:00-17:00]; Off_Peak | 2 | TRP |
| Manhattan Trip distance (trip_dist) | (0-1]; (1-2]; (2-3]; (3-4]; (4-5]; (5-6]; (6-8]; (8-11]; (11-16]; (16-56] | 10 | TRP |
| Daily trip count (n_pers_trip) | [1-2]; (2-3]; (3-4]; (4-6]; (6-18] | 5 | PER |

*Note: all numeric variables are discrete (numeric categories)*

## 5.3 Feature Preprocessing

### *5.3.1 Merging Personal Attributes to Trips*

TTS data are available in four .csv files: person, trip, household, and transit, which are merged for DT classification. Unique trips constitute the unit of analysis for the mode choice model. All non-trip attributes are therefore merged with the trip table to relate trip attributes with trip maker attributes, exemplified in Figure 18.

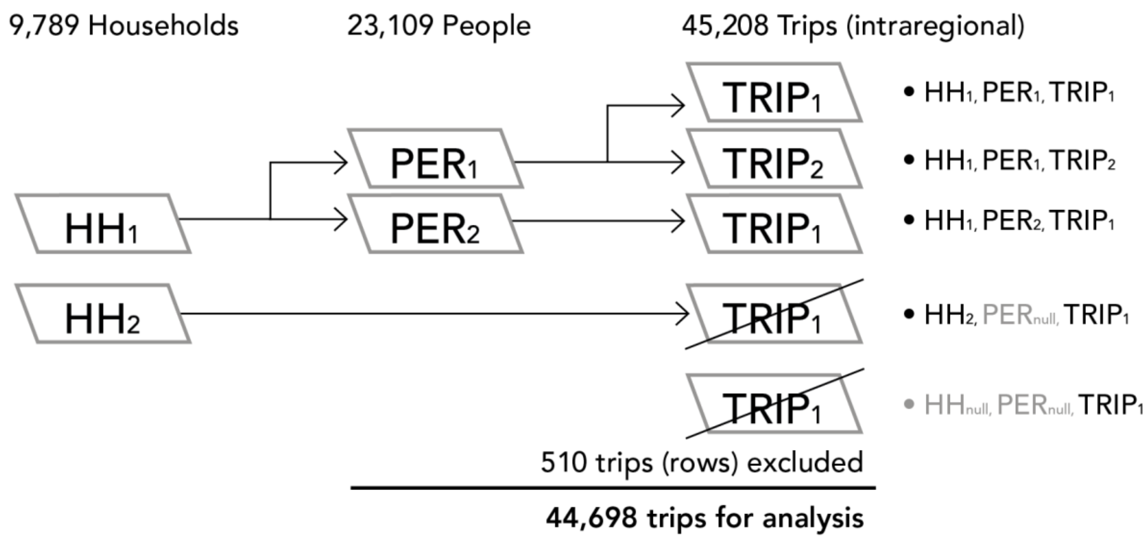



*Figure 18: Merging TTS Datasets and Removing Missing (Null) Values*

Unfortunately, some trip samples have incomplete attributes (missing values) and the problem is exacerbated by the merging process: since TTS data are split into four files, merging the tables into a single dataset may attribute an otherwise complete observation (e.g., complete trip attributes) with missing values (e.g., incomplete personal attributes). One strategy to manage incomplete datasets is to simply remove rows (observations) with incomplete data (as in Figure 18); however, this results in the loss of potentially useful data. Of the original 55,958 trips that begin or end in the Region, only 45,208 trips are intraregional (both trip ends in Region). After merging the household and person tables to trips, 44,698 trips remain for analysis.

Note that this research distinguishes "Unknown" values (given a universal label "9" for each feature) where the travel diary survey notes an unwillingness to respond. Response hesitancy may, in itself, yield some insight. The CHAID DT manages these unknown values as separate categories that should not support prediction, assuming the unknown values are indeed randomly distributed. Alternative strategies for handling missing/unknown values involve imputing (infer values using available data) values through some statistical measure (most frequent categorical variable, mean/median of numerical variables) or machine learning algorithm (Pedregosa et al., 2011). However, data imputation effects are left out of scope for this research.

### 5.3.2 Percentile-based Bins

CHAID DTs benefit from discrete bins of continuous variables because each unique value is interpreted by the algorithm as its own "category." Larger categories (bins) may better describe trends within feature dimensions and equal-sized bins reduces training bias towards frequently observed categories. Continuous variables in this research include age, household size, number of vehicles in household, travel distance, and (later) DTA. Percentile-based bin sizes are used to discretize the continuous data. As opposed to equal-interval, which segments based on equal ranges of a variable's value, percentile-based bins have equal sample sizes. Percentile-based discretization is advantageous for decision tree building because DTs require high sample sizes and model training may be biased by unbalanced dimensions. A limitation of using percentile-based discretization is that the tree may not discern useful splits if the bins are not meaningful. For example, if observations for the number of vehicles in a household are skewed right and the "most useful split" (see "Splitting Criterion" subsection) occurs between the lower two quartiles at the 2.5 vehicle boundary ($0 \leq Q1 < 2.5, 2.5 \leq Q2 < 3$), then the split provides less information about the variable's influence on mode class outcomes.

Departure times were grouped according to whether they were peak or off-peak periods. Peak and off-peak definitions are consistent with those used in the "Dynamic Transit Accessibility Metric Development" section (see Table 3), where AM and PM peak periods are 6:00-9:35 and 15:00-17:00, respectively. This follows methodology from (Xie et al., 2003), who used binary peak and off-peak dimensions for departure time classification in their mode choice DT model.

### *5.3.3 Data Expansion*

TTS trip samples include expansion factors (expf) to improve representation of population and demographic characteristics within zones (see "Travel Diary Survey" subsection for more detail). This research uses the expanded TTS trip samples for modelling, deferring to the expansion process adopted by the TTS. Expansion factors are numerical attributes that describe each sampled (unexpanded) trip in the TTS trip dataset. DTs require data to be structured such that each observation occupies its own row. Thus, a Python script was written to iterate through unexpanded trips and copy each sample (including all trip, person, and household attributes) according to the trip expansion factor. Expansion factors are applied after rounding to the nearest whole number, so if a sampled trip has an expansion factor of 3.7, the trip is copied to occupy four (4) observations in the dataset used in the DT.

## 5.4 Dependent Variable

The dependent variable (target variable) of this analysis is individual (disaggregate) mode choices, where a transportation mode is a target class among the assumed choice set of modal classes. Since this research is specifically concerned with the prediction of transit choices, the dependent variable is binary: transit or non-transit. In this thesis, the transit class includes local transit system (GRT) and walking modes because of the inclusion of walking accessibility within the transit accessibility metric (see "Travel Impedance" subsection). Since this mode choice model is intended to include the transit accessibility metric as a feature and baseline transit accessibility is measured using the walking trip between zones, the target class also includes walking trips. Since the characteristics and decision processes leading to walking trips may be different than for transit trips and not all transit trips are replaceable with walking trips, this is an important limitation of including location-based transit accessibility within a mode choice analysis. Furthermore, the transit class excludes the GO Transit interregional system. The not-transit class includes the other modes defined in the "Calculating Mode Shares" subsection of the DTA chapter: Non-local Transit, Drive Alone, Passenger, Cycle, Other.

## 5.5 Data Subsets: Training and Testing

### *5.5.1 Split Size*

Splitting the balanced dataset into subsets is a component of the validation procedure used in DT analysis. The data are split into a random training group (80%) and a random testing group (20%). A portion of the training data is required for data-intensive post-processing steps, including hyperparameter tuning through cross-validation (DT post-processing). Therefore, the size of the training group is much larger than the testing group, but the exact proportions chosen for the split are arbitrary but comparable to other studies (e.g., Wets et al. (2000) used a 75% training dataset, extracted from a 4,810-trip sample).

### *5.5.2 Shuffling Data*

A pseudorandom number generator in Python ("train_test_split" module) is used to randomly shuffle the data during the splitting procedure. The pseudorandom number generator produces a deterministic sequence of numbers that depends on an initial input number used to seed the generator. Given the same seed (non-negative integer), the module's "random_state" parameter produces the same random subset and is therefore reproducible (Pedregosa et al., 2011).

The splitting procedure maintains representation of the different target classes within the resulting subsets (training and test) through the "stratify" parameter. Stratified sampling splits each target class between training and test data subsets proportionally; that is, if 10% of observations belong to the Not-transit target class, 10% of the training dataset will belong to the Not-transit target class and 10% of the test dataset will belong to the Not-transit target class.

### *5.5.3 Class Balancing*

Balancing the training dataset between the different target classes (i.e., transit, not-transit) is necessary to prevent a training bias towards the dominant classes (in this case, driving). Performing class balancing requires either sampling an equal number of samples per class (e.g., undersampling, which reduces sample size for dominant classes) or normalizing the sum of sample weights to the same value for each class. Normalizing the sum of sample weights is preferable because it does not lose feature information that could be useful in prediction, but still scales an observation's impact. This research applies the latter method, so the tree pruning criteria (hyperparameter tuning criteria) become weight based. For example, the "minimum *sample* per split" becomes the "minimum *weight* per split," so leaf nodes must contain a minimum fraction of the overall sample weights. Class balancing is performed on expanded trip sample outputs from the Data Expansion step that belong to the training data subset created in the Data Subsets: Training and Testing step.

This research uses the scikit-learn "compute_class_weight" module to automate the class balancing procedure. The class balancing process involves taking all target classes (transportation modes) and applying a weight to every class that increases the weight of infrequently sampled classes (e.g., transit), and decreases the weight of frequently sampled classes (e.g., non-transit) for tree learning. The module applies equation [5.1].

$$weight_m = \frac{n_{samples}}{n_m * n_{bins}(y)} \qquad [5.1]$$

where $weight_m$ is the weight of the target class $m$, relative to all other target classes of the dependent variable, $y$. $n_{samples}$ is the total number of samples in the training dataset, including all target classes. $n_m$ is the number of samples belonging to class $m$, and $n_{bins}$ is the number of bins (different classes) in dependent variable, $y$. For example, since there are two bins (Transit and Not-Transit), a dataset of size $n_{samples} = 10$, where 2 belong to $n_{transit}$ would have a $weight_{not-transit}$ of 0.625 and $weight_{transit}$ of 2.5. Weights are recorded within the attributes of respective observations in a new column.

Final class weights are calculated as 4.641 for the Transit class and 0.56 for the Not-Transit class. Alternatively, every Transit observation has a weight of 8.2875 and every Not-Transit observation has a weight of 1. SPSS Modeler Subscription Version 18.2 can interpret the weight variable column when it trains CHAID DTs.

## 5.6 Splitting Criterion

A DT's splitting criterion influences how the tree grows during the training process. Splitting criteria partition the predictor space (i.e., training observations) into mutually exclusive regions based on the boundaries between each independent variable's values/categories (i.e., feature dimensions). Although splitting criteria and resulting DT algorithms differ, they generally aim to improve the homogeneity of the dependent variable categories (i.e., target classes) within partitioned regions. A tree "grows" through the recursive partitioning of the predictor space until some stopping criteria are met. Figure 19 illustrates the splitting operation using a fictional example.

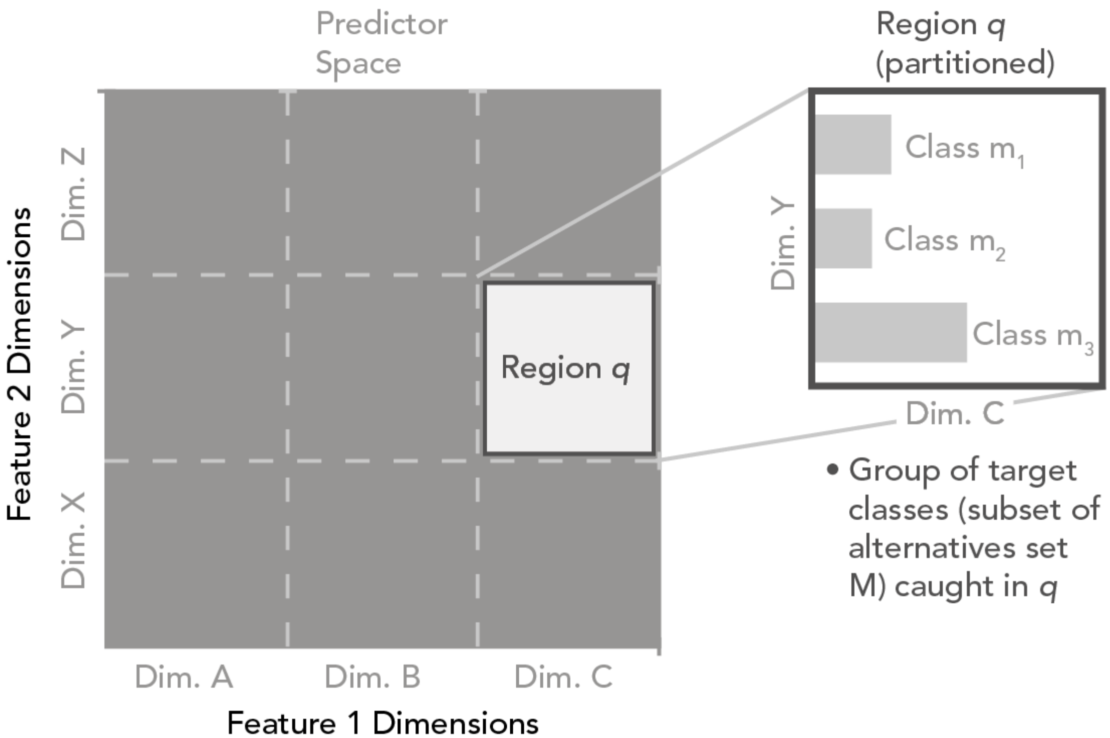



*Figure 19: Splitting Along Dimensions of the Target Class Predictor Space: Visual Example*

Figure 19 exemplifies a predictor space for a classification problem with two features: Feature 1 and Feature 2 (independent variables). Feature dimensions (1:[A, B, C]; 2:[X,Y,Z]) divide the predictor space into regions, which include Region $q$. Region $q$ defined as the set of training observations that fall into feature dimensions Y and C. If dimension C is the age group 50-60 years old and dimension Y is the $60,000-$80,000 per year income group (i.e., Feature 1 is "Age" and Feature 2 is "Income"), then Region $q$ includes all the observations from people who are 50-60 years old and making between $60 and $80 thousand per year. Observations within Region $q$ may include any number of target classes (unique values of the dependent variable) within the alternatives set M, but only the observations that fall under dimensions Y and C. The fictional partition of Region $q$ reveals an uneven split between the classes, where class $m_3$ is more frequently observed than $m_1$ and $m_2$. Therefore, Region $q$ may be chosen for a split if it compares favourably with other possible regions based on the splitting criterion.

### *5.6.1 CHAID*

This research uses the CHAID algorithm, which applies the chi-square test of independence as the splitting criterion. Compared with the common CART algorithms, which apply binary splits at decision nodes using purity measures, CHAID can create multiway splits of categorical features using a statistical test (Kass, 1980). Multiway splits create wider trees because all possible feature dimensions descend from the parent node, grouping dimensions (categories of independent variables) by their feature (independent variables). Wider trees are advantageous over deeper trees because all the dimensions of a feature are presented together, as opposed to further down the tree, to compare class proportions across all dimensions for a sample region and thus represent aspiration-level based decision rules. High-dimension features increase tree complexity and reduce interpretation; however, binary splitting algorithms likewise suffer from high dimensionality.

The chi-square statistic ($\chi^2$ score) was introduced by Karl Pearson to test the independence of categorical variables (Franke et al., 2011). Equation [5.2] arithmetically describes the chi-square statistic.

$$Chi - square\ Statistic\ (\chi^2) = \sum_{i=1} \sum_{m=1} \frac{(q_{im} - \hat{q}_{im})^2}{\hat{q}_{im}} \qquad [5.2]$$

where $q_{im}$ is the frequency of observations in class $m$ ($m \in M$) and dimension $i$ ($i \in I$), and $\hat{q}_{ij}$ is the expected frequency of observations in class $m$ and dimension $i$ under the null hypothesis. $\chi^2$ is calculated for each feature.

The $\chi^2$ test of independence uses the chi-square statistic to test the null hypothesis that observed classes are independent of a feature. For a given feature, the test involves cross tabulating the target classes (rows) with each of the feature's dimensions (columns). The expected frequencies of class $m$ in dimension $i$ ($\hat{q}_{im}$) represent the null hypothesis: if class outcomes are independent of features, the proportion of class $m$ in dimension $i$ would be the same as the proportion of total observations that are in dimension $i$, across all classes $M$. Equation [5.3] is used to calculate expected frequency ($\hat{q}_{ij}$),

$$\hat{q}_{im} = \sum_{i=1} q_{im} * \left(\frac{\sum_{m=1} q_{im}}{n}\right) \qquad [5.3]$$

where $q_{im}$ is the frequency of observations in dimension $i$ and class $m$, and $n$ is the total number of observations. In the cross tabulation, the sum of $q_{im}$ across all dimensions $i$ is the row total of the class and the sum of $q_{im}$ across all classes $m$ is the column total of the dimension. The sum of all cell values in the entire cross tabulation is the $\chi^2$ score for a feature. The $\chi^2$ is corresponds to a $p$-value based on the chi-square distribution with degrees of freedom $df = (M - 1) - (I - 1)$. A $p$-value lower than the selected significance level ($\alpha$ value) rejects the null hypothesis and the result is statistically significant.

Kass (1980) describes the original CHAID method and more recent implementations provide documentation on the process (IBM, 2011). This research uses IBM's SPSS Modeler Subscription Version 18.2 implementation of CHAID due to its ability to interpret (call) trained DTs for validation and testing. Three general steps comprise IBM's implementation of CHAID: merging, splitting, and stopping. The steps are paraphrased from the algorithms guide (IBM, 2011):

Merging

1. If the feature only has 1 dimension, set the $p$-value to 1.
2. If the feature has 2 dimensions, go to splitting step.
3. Otherwise, search for the pair of dimensions in feature X that are the least significantly different (most similar) with respect to the dependent variable using the $\chi^2$ test of independence (largest $p$-value).
4. If $p$-value > $\alpha$, merge the two dimensions into a new dimension.
5. (Optional) if any dimension has too few observations ($n < min.parent\ node\ size$), the dimension is merged with the category with the largest $p$-value
6. Compute the Bonferroni adjusted $p$-value for the merged dimensions.

Splitting

7. Select the feature with the lowest $p$-value.
8. If the $p$-value is less than $\alpha$, split the node using this feature. This is a parent node.
9. If the $p$-value is greater than $\alpha$, do not split the node. This is a terminal (leaf) node.

Stopping

10. Stop if all observations in the node belong to the same class (pure)
11. Stop if tree depth hyperparameter limit is reached
12. Stop if minimum parent node size limit is reached

13. If a child node size is below the minimum child node size limit, merge with another child node with the largest $p$-value. If no other child nodes, stop.

## 5.7 Measures of Effectiveness

This analysis uses three measures of effectiveness (MOEs) to assess the trained decision tree model: overall accuracy, sensitivity (true positive rate), and specificity (true negative rate). Mode choice models that apply machine learning classification techniques commonly report these three MOEs for comparison (see the "Decision Trees and Mode Choice" subsection). The hyperparameter tuning process uses cross validation to find and select the best performing model with respect to these MOEs. After selecting the best performing CHAID RBMC model, the held-out testing sample is applied to the trained model and scored with respect to the same MOEs.

### *5.7.1 Model Accuracy*

Han et al. (2011, p. 365) describe common DT evaluation measures using the classification terminology of a confusion (or misclassification) matrix, adapted in Table 17. ML classifiers applied in mode choice prediction also use confusion matrices for comparative analyses (Xie et al., 2003).

*Table 17: Binary confusion matrix for classifier evaluation, adapted (Han et al., 2011, p. 366)*

| **Actual Class** | **Predicted Class** | | Total | Recall (%) |
|---|---|---|---|---|
| | Transit | Not_Transit | | |
| Transit | True Positives ($TP$) | False Negatives ($FN$) | $TP + FN = P$ | $Sensit. = \frac{TP}{P}$ |
| Not_transit | False Positives ($FP$) | True Negatives ($TN$) | $FP + TN = N$ | $Specif. = \frac{TN}{N}$ |
| Total | $TP + FP = P'$ | $FN + TN = N'$ | $P + N = Total$<br>$P' + N' = Total$ | $\frac{TP + TN}{Total} = Acc.$ |
| Precision | $\frac{TP}{P'}$ | $\frac{TN}{N'}$ | | |

The confusion matrix describes the potential outcomes of a classifier, listed below, in terms of prediction correctness with respect to observations from the test set.

- True positives ($TP$): the number of positive outcomes that the classifier labeled correctly
- True negatives ($TN$): the number of negative outcomes that the classifier labeled correctly
- False positives ($FP$): the number of positive outcomes that the classifier labeled incorrectly
- False negatives ($FN$): the number of negative outcomes that the classifier labelled incorrectly

Each predicted (labelled) outcome is the class with the highest probability at the terminal (leaf or child node) into which an observation falls. This assumption is built into SPSS Modeller's model scoring process and is therefore used to calculate MOEs for the hyperparameter tuning process; however, assuming choice based on highest proportion is an imprecise interpretation of odds. During the Model Predictions: Proportional Enumeration step, when test data are applied to the trained DT, predictions are represented by the shares of the leaf node into which an observation falls.

The model accuracy (equation [5.4]), or recognition rate, is the number of correctly predicted outcomes ($TP + TN$) out of the total test set. The inverse of accuracy is the risk, or error rate, which describes the number of misclassified observations.

$$Accuracy_{model} = \frac{TP + TN}{P + N} \quad [5.4]$$

$$Risk_{model} = \frac{FP + FN}{P + N} \quad [5.5]$$

### 5.7.2 Recall: Sensitivity and Specificity

Recall is a class-specific measure of accuracy. Sensitivity is the proportion of positive outcomes that were classified correctly. Inversely, specificity is the proportion of negative outcomes that were classified correctly. In this binary study, transit is a positive outcome and not-transit is a negative outcome. Sensitivity and specificity resist data imbalance effects because they are isolated measures concerning each outcome class. Since class imbalances in the training data can cause the model to more likely predict the more frequent class, these measures are considered the “true positive/negative rate” (Han et al., 2011, p. 367).

$$Sensitivity_{class} = \frac{TP}{P} \quad [5.6]$$

$$Specificity_{class} = \frac{TN}{N} \quad [5.7]$$

### 5.7.3 Precision

Precision (equation [5.8]) is another class-specific measure that is concerned with the exactness of the prediction. While recall measures the proportion of outcomes (observations) that are labelled correctly, precision measures the proportion of a class’s predictions that match the observations (Han et al., 2011). A perfect precision score (1.0) for the transit class means that every “transit” prediction made by the DT indeed belongs to the transit class. However, it does not reveal the number of observations that the DT mislabelled. Conversely, a perfect recall score does not reveal the number of predictions that were incorrect. Together, recall and precision give a more complete picture of the model’s predictive ability.

$$Precision_{class} = \frac{TP}{TP + FP} \quad [5.8]$$

### 5.7.4 Interpretability

Interpretability generally refers to a subjective degree of insight that a classifier or its predicted output provides (Han et al., 2011). Pursuant the objectives of this thesis, interpretability is the extent to which the model can coherently represent non-compensatory decision processes and respect behavioural limits. Therefore, discussion related to model interpretability considers model outputs as follows:

- Ease to deduce the model structure, components, and mechanics (i.e., splitting criteria, features, dimensions, hierarchy),
- Whether the model can be read in terms of discrete non-compensatory decision rules, and
- Whether features have plausible directions of association (variable sign).

## 5.8 Model Hyperparameter Tuning

Decision tree development requires that the modeller define parameters that create boundaries to the model learning process. These parameters are called “hyperparameters” in machine learning literature because they refer to parameters of the model induction process (Wang & Ross, 2018). Hyperparameters affect model outcomes (i.e., predictive ability, structure) and the resources required to learn a model. Decision tree hyperparameters include tree depth, number of samples required for a split (at the parent node), and number of samples required for a leaf (child) node. The tree depth describes the number of splits produced by the model. More shallow trees can lead to more interpretable trees but reduce prediction accuracy, whereas deeper trees are prone to overfitting. Minimum samples for a split refers to the number of samples that must exist within a region (e.g., region $q$, defined by feature dimensions $X$ and $C$) for a split to be considered. Low minimum samples for split may also overfit the data (enables creation of many insignificant splits), whereas a high minimum sample may prevent the tree form learning (insufficient samples for splitting).

Hyperparameter tuning involves searching the hyperparameter space to improve model prediction (i.e., optimize a measure of performance). Modellers must "tune" hyperparameters carefully because the unrestricted optimization of hyperparameters may lead to overfit, uninterpretable, or inaccurate decision trees. For example, increasing the tree depth hyperparameter through an unbounded optimization routine would produce a model that is perfectly fit to the training data (i.e., each leaf/terminal node is a discrete observation) but unusable for interpretation or out-of-sample prediction. Therefore, hyperparameters are tuned through the model validation process.

### *5.8.1 Model Validation and Selection*

Model validation is an evaluation process concerned with reducing model bias, abating overfitting to training data, and affirming model flexibility (i.e., generalized performance). The validation step tests the trained model using "unseen" data that are excluded from the learning process, thereby mitigating sampling bias. Resampling methods, which redraw subsamples from a training set and refit a model, are the primary tools for model validation (James et al., 2013, p. 177).

This research uses SPSS Modeler for model validation. In the context of the *model selection* step in ML post-processing, a validation procedure is applied to part of the training dataset, further partitioning the original training dataset into secondary training and validation subsamples (Hastie et al., 2008, pp. 241–254). These subsamples serve a distinct purpose from the original training/testing datasets. Whereas the testing dataset is used to evaluate the unbiased performance of a model (*model assessment*), the training/validation subsamples are drawn from the training dataset to tune the hyperparameters of a final model with less bias (*model selection*). If hyperparameters were tuned to improve performance on the test set alone, knowledge from the test set could "leak" into the tuned model and reduce generalizability (Pedregosa et al., 2011). The testing dataset is still held out at this modelling stage because it is used to measure the trained model's performance (i.e., after hyperparameters are selected) on unseen data. Figure 20 offers a 4-subsample (fold) example of the data partitioning process.

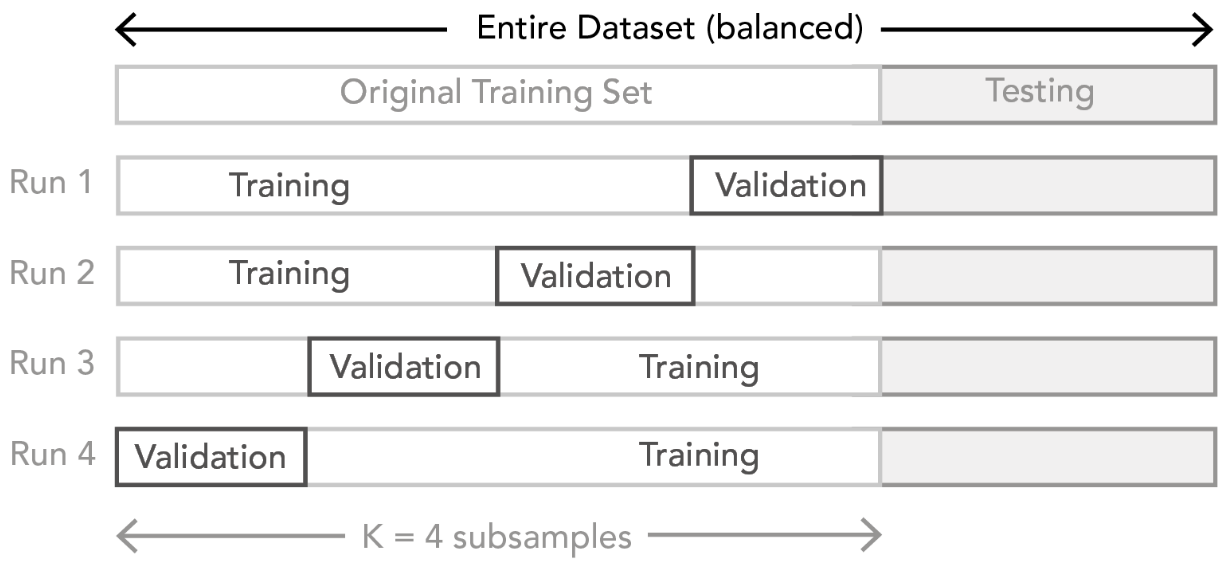



*Figure 20: Decision Tree Cross-validation Subsampling (example)*

This research uses K-fold cross validation, adapted from Bishop, C.M. (2011, p. 33), to tune the tree depth hyperparameter from the range of 1 level to 5 levels. K-fold cross validation is a popular resampling method where the data is randomly partitioned (resampled) into K number of subsamples of equal size ("folds") and then successively held out for validation. A common value for $K$ is 10 (i.e., 10 folds), where the original training set is resampled ten times without replacement. K-fold cross validation is advantageous over hold-out validation (single hold out sample) because K-fold uses the entire dataset for both training and validation over multiple runs (Figure 20 uses four). Hold-out validation is subject to greater bias from the delineation of a single validation subsample. One disadvantage of K-fold cross validation is its relatively higher computational cost because a higher number of folds ($K$) increases the number of times the model must be trained on the $K \neq k$ folds; however, modern computers overcome this barrier with relative ease.

This research uses 10-fold cross validation for hyperparameter tuning. Each decision tree, characterized by a unique hyperparameter set ($\theta_i$), trains on the $K \neq k$ subsamples and tests on fold $k$. Note that $i$ is a list with notation in the format [depth, parent node size, child node size] for all hyperparameter combinations within the grid search range, shown in Table 18. For example, $\theta_{3,10,5}$ is the tree trained using max depth = 3 levels, parent node size = 10% of the training sample, and child node size = 5% of the training sample.

*Table 18: Hyperparameter Search Details*

| **Hyperparameter** | **Range** | **Search Increment** |
|---|---|---|
| Max Tree Depth (levels) | [1,5] | 1 |
| Min. parent node size (% training samples) | [0.1, 50] | 5% |
| Min. child node size (% training samples) | [0.05, 25] | 5% |

The range of grid search values ($\theta_{1,1,0.5}, \theta_{1,5,0.5}, \ldots, \theta_i \in \Theta$) occupy extreme values for each hyperparameter. Maximum tree depth begins at 1, producing a tree with a single split. A single-split tree would have low prediction accuracy and insufficiently describe mode choice behaviours. The maximum of 5 is chosen to avoid overfitting and because tree depth can quickly increase model complexity: the number of leaf nodes is a product of the number of dimensions for features selected at each level. Parent and child node sizes are proportions of the training data sample ($n$ = 807,916 expanded trips) and must be greater than 0. At the lower boundary, a minimum of 1% of the total training sample ($n$ = 8,079) is used for a split (at any parent node). Since the sample includes expanded trips, this is a low number for a parent node. Notably, the CHAID tree is less sensitive to low parent node sizes because the $\chi^2$ test only allows statistically significant splits. At the higher boundary for parent node size, 50% of the training sample must exist within a parent node to allow a split ($n = 403{,}958$), which would decrease tree complexity and limit prediction accuracy. Since child nodes must result from splits at parent nodes, child node values are capped at 25% of the training dataset ($n$ = 201,979).

Five MOEs are calculated for each $K \neq k$ tree using data from fold $k$ (the fold excluded from training): 1) overall model accuracy (equation [5.4]), 2) sensitivity (equation [5.6]), 3) specificity (equation [5.7]), 4) precision of transit (equation [5.8]), and 5) precision of not-transit (equation [5.8]). Since each unique combination of hyperparameters ($\theta_i$) defines a tree that is cross-validated 10-fold, reported measures of effectiveness describe average results across the 10 folds. Table 19 reports the average recall and precision measures across 10-folds for the best performing parent and child node combination by tree depth.

*Table 19: Hyperparameter Search Results Scored on Validation Subsamples*

| Max. Tree Depth | 1 | 2 | 3 | **4** | 5 |
|---|---|---|---|---|---|
| Overall Model Accuracy | 89% | 82.5% | 79.2% | **84.3%** | 83.9% |
| Transit Recall (Sensitivity) | 42.8% | 79.7% | 88.7% | **84%** | 87.6% |
| Not-Transit Recall (Specificity) | 94.6% | 82.8% | 78% | **86.9%** | 83.5% |
| Transit Precision | 49% | 35.8% | 32.7% | **39.6%** | 39% |
| Not-Transit Precision | 93.2% | 97.1% | 98.3% | **98.2%** | 98.2% |

Among the model hyperparameters (i.e., tree depth, minimum parent node size, minimum child node size), Table 19 only reports changes across tree depth. Node sizes are effectively unrestricted at 1% and 0.5% of samples for minimum parent and child node sizes, respectively. The hyperparameter grid search indicated that restrictions on node sizes can quickly reduce the tree to lower depths without improving model accuracy. Confusion matrices report that the trees of parent-node-restricted models $\theta_{2,25,0.5}$, $\theta_{3,25,0.5}$, $\theta_{4,25,0.5}$, and $\theta_{5,25,0.5}$ are all the same with an 82.4% overall accuracy. Since the child node remains unrestricted, the search indicates that prohibiting splits using <25% of training data do not allow trees to grow above a maximum depth of 2. Similarly, prohibiting splits using <10% of data at the parent node do not allow trees to grow beyond a depth of 3 and prohibiting splits using <5% of data at the parent node do not allow depths beyond 4 (models $\theta_{4,5,0.5}$ and $\theta_{5,5,0.5}$ produced identical predictions). At the lower node size restrictions, the primary inhibitor to growth is the tree depth. This is inferred from models with adjacent parent node size restrictions producing the same predictions, holding tree depth and child node sizes constant. For example, models $\theta_{1,25,0.5}$, $\theta_{1,10,0.5}$, $\theta_{1,5,0.5}$, and $\theta_{1,1,0.5}$ produce identical confusion matrices on validation data.

Based on the hyperparameter tuning process, this research selects the RBMC model trained on hyperparameter set $\theta_{4,1,0.5}$. This combination of unrestricted node sizes and higher tree depth produced a high overall accuracy (84%), high recall values (84-86%), and the highest transit precision (39.6%) above a tree depth of 1. Increasing the tree depth to 5 had the effect of reducing accuracy and recall for the not-transit class (transit specificity). The trained tree is plotted in Figure 21 with some reduced detail (not every dimension is labeled at higher depth) to support interpretation.

### 5.8.2 Tree: RBMC Model

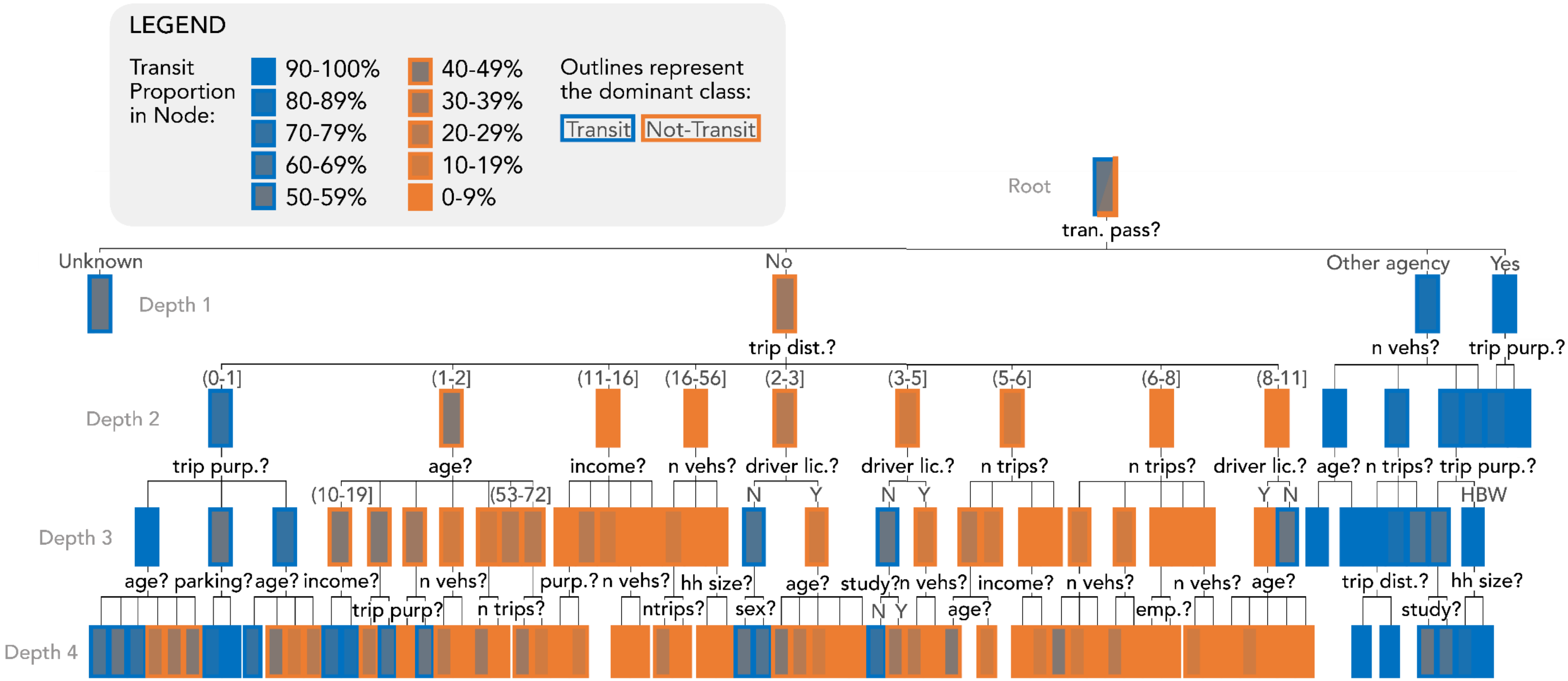



Figure 21: Trained Decision Tree Diagram (RBMC model)

The raw, unedited output of the DT is in Appendix B with bar charts representing the (unbalanced) proportion of each node that belongs to each class. Note that the raw output's class proportions do not represent the balanced class values on which the DT trains (see "Class Balancing" subsection). Therefore, the raw output graphic underestimates the training samples belonging to transit within each node by a factor of approximately 8.29 relative to not-transit samples, which Figure 21 corrects. A full list of tree nodes, including the balanced proportion of each class per node, is in Appendix A.

CHAID produces nodes that are selected based on their probabilities ($p - value < 0.05$). The selected splits (parent nodes) for the trained DT with hyperparameter set $\theta_{4,1,0.5}$ are noted in Table 20. All parent nodes between the root node and leaf nodes satisfy the significance criterion.

*Table 20: CHAID Split (Parent Nodes) Statistics: Trained Model*

| Depth | Node | Feature (var. name) | n split | df | $\chi^2$ | p-value |
|---|---|---|---|---|---|---|
| 0 | Root | Transit pass (tran_pass) | 807,916 | 3 | 135,503.4 | <0.000 |
| 1 | 2 | Trip distance (trip_man_km) | 731,870 | 8 | 99,565.1 | <0.000 |
| 1 | 3 | Vehicle ownership (hh_n_vehs) | 61,679 | 3 | 8,091.8 | <0.000 |
| 1 | 4 | Trip purpose (trip_purp) | 9,165 | 1 | 525.9 | <0.000 |
| 2 | 5 | Trip purpose (trip_purp) | 117,540 | 2 | 29,546.2 | <0.000 |
| 2 | 6 | Age (age) | 80,930 | 6 | 8,723.0 | <0.000 |
| 2 | 7 | Income (hh_income) | 67,094 | 4 | 2,399.7 | <0.000 |
| 2 | 8 | Vehicle ownership (hh_n_vehs) | 74,040 | 2 | 4,577.8 | <0.000 |
| 2 | 9 | Licensing (driver_lic) | 72,225 | 1 | 4,312.5 | <0.000 |
| 2 | 10 | Licensing (driver_lic) | 120,063 | 1 | 5,390.3 | <0.000 |
| 2 | 11 | Daily trips (n_pers_trip) | 50,120 | 3 | 2,938.7 | <0.000 |
| 2 | 12 | Daily trips (n_pers_trip) | 79,228 | 4 | 2,125.3 | <0.000 |
| 2 | 13 | Licensing (driver_lic) | 70,630 | 1 | 4,118.7 | <0.000 |
| 2 | 14 | Age (age) | 14,046 | 1 | 845.733 | <0.000 |
| 2 | 15 | Daily trips (n_pers_trip) | 21,597 | 2 | 2,463.4 | <0.000 |
| 2 | 16 | Trip purpose (trip_purp) | 19,532 | 1 | 3,135.6 | <0.000 |
| 3 | 20 | Age (age) | 60,119 | 5 | 2,940.2 | <0.000 |
| 3 | 21 | Free parking (free_park) | 34,913 | 1 | 4,623.6 | <0.000 |
| 3 | 22 | Age (age) | 22,508 | 3 | 2,112.6 | <0.000 |
| 3 | 23 | Income (hh_income) | 12,747 | 1 | 149.5 | <0.000 |
| 3 | 24 | Trip purpose (trip_purp) | 14,977 | 1 | 1,032.4 | <0.000 |
| 3 | 25 | Trip purpose (trip_purp) | 9,448 | 1 | 602.0 | <0.000 |
| 3 | 26 | Vehicle ownership (hh_n_vehs) | 12,716 | 1 | 193.7 | <0.000 |
| 3 | 27 | Daily trips (n_pers_trip) | 10,168 | 1 | 1,649.7 | <0.000 |
| 3 | 29 | Daily trips (n_pers_trip) | 13,893 | 1 | 324.2 | <0.000 |
| 3 | 30 | Trip purpose (trip_purp) | 9,786 | 1 | 386.8 | <0.000 |
| 3 | 33 | Vehicle ownership (hh_n_vehs) | 32,761 | 1 | 157.8 | <0.000 |
| 3 | 35 | Daily trips (n_pers_trip) | 14,893 | 1 | 321.4 | <0.000 |
| 3 | 37 | Household size (hh_size) | 23,002 | 1 | 207.5 | <0.000 |
| 3 | 38 | Sex (sex) | 9,976 | 1 | 64.9 | <0.000 |
| 3 | 39 | Age (age) | 62,249 | 4 | 1,819.4 | <0.000 |
| 3 | 40 | Student status (stu_stat) | 15,174 | 1 | 1,525.5 | <0.000 |
| 3 | 41 | Vehicle ownership (hh_n_vehs) | 104,889 | 1 | 1,790.7 | <0.000 |
| 3 | 42 | Age (age) | 15,489 | 1 | 611.4 | <0.000 |
| 3 | 44 | Income (hh_income) | 11,062 | 1 | 357.1 | <0.000 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 3 | 46 | Vehicle ownership (hh_n_vehs) | 27,668 | 2 | 2,914.2 | <0.000 |
| 3 | 47 | Vehicle ownership (hh_n_vehs) | 8,619 | 1 | 804.9 | <0.000 |
| 3 | 48 | Employment status (emp_stat) | 18,286 | 1 | 1,350.4 | <0.000 |
| 3 | 50 | Vehicle ownership (hh_n_vehs) | 9,439 | 1 | 53.8 | <0.000 |
| 3 | 51 | Age (age) | 66,398 | 4 | 1,254.2 | <0.000 |
| 3 | 55 | Trip distance (trip_man_km) | 9,453 | 1 | 615.1 | <0.000 |
| 3 | 58 | Student status (stu_stat) | 8,226 | 1 | 26.3 | <0.000 |
| 3 | 59 | Household size (hh_size) | 11,306 | 1 | 373.7 | <0.000 |

*Note: n split refers to observations in the training dataset that belong to each node. Since only parent nodes are reported, n split of subsequent levels may be lower than antecedent levels whenever stopping criteria are reached along any intermediate child branches.*

## 5.9 Results and Discussion: RBMC Model

### *5.9.1 Learned Decision Rules*

A review of the specific decision processes induced by the DT, following the tree diagram presented in Figure 21, reveals interesting relationships between the independent variables and transit choice. At the top of the tree, the best predictor at the root node is transit pass ownership. This discussion is organized by the branches resulting from the root node split (transit pass owners, non transit pass owners, and unknowns). Proportions at each node are specifically listed in Appendix A, while the rough value is represented by colour in Figure 21.

#### *5.9.1.1 Transit Pass Owners*

People who reply “Yes” own a local transit pass with GRT, and already 91% of users in this category are predicted to take transit. Descending from this, trip purpose shows that HBW trips are slightly more likely to use transit (94% vs 86%). No further splits result from this branch at depth 2, meaning no other predictors reveal statistically significant differences within these regions.

Respondents who have transit passes for other agencies (e.g., regional transit services) also tend to favour transit (90%); however, their choices depend on the number of vehicles in the household (depth 2 split). Households without vehicles (personal automobiles) are predicted to choose transit 98% of the time; for households with a non-GRT transit pass (“other agency”) and only one car, that probability drops to 88% and further splitting (depth 3) occurs based on the number of trips per person. People who take more trips are less likely to take transit. This reflects the phenomenon where households compete for access to not-transit modes, including driving and passenger modes, and that more frequent trip users may find transit less appealing. Households with 2 cars will still probably take transit if they own a transit pass but will consider the trip purpose (depth 3), where commuter (HBW) trips are more likely (91%) to use transit than HBD and Non-HB trips (61%). It is unclear whether “other agency” passes in the TTS include University-distributed transit passes (e.g., the University of Waterloo’s U-Pass program), which could affect this interpretation.

#### *5.9.1.2 No Transit Pass*

Users who do not own a transit pass are far more likely to not take transit, constituting the entire vertical midsection of the tree. People who do not have transit passes may still choose to use transit based primarily on the trip distance (depth 2). Measured in Manhattan distances (grid- or network-based distances), trips less than 1km have a 76% chance of taking transit, and longer trips less than 2 or 3km quickly decrease that probability to 45% and 28%, respectively. Since walking trips are included in the transit share, findings related to shorter trips more likely represent important decision variables for walking rather than transit. Short trips less than 1km split based on trip purposes (depth 3), where HBW trips are, again, transit oriented. Depth 3 splits for the (1-2km] range show that people more likely to take transit if they are adolescents (<19 years). People who do not have a transit pass (depth 1) and travel 2-5km (depth 2) may make transit decisions based on whether they have a driver's license (depth 3). This split is very polarizing because people who do not have driver's licenses are predicted to take transit 53%-63% of the time, compared with 11-17% if they do have a license. The most surprising finding from this model is that users who take 3-5 km trips tend to take transit (77%) if they are not students (depth 4) and tend not to take transit if they are students (34%). However, student attributes and preferences may be underrepresented in these data since they seldom have permanent addresses in the Region and are less likely to receive invitations to participate in the travel diary survey.

Users without transit passes who are travelling longer distances (>5 km) take Not-Transit modes almost universally; however, a segment (node 13, depth 3) of users are somewhat likely to take transit (52%) for 8-11km trips if they do not have a driver's license. Licensed users that make 8-11km trips are further split (depth 4) by age, where younger people (11-28 years) have a slim chance of taking transit (12%) and older people (41+ years) have virtually none (<1%). Non-transit pass, long-distance trips have statistically significant splits at depth 3 by income, number of vehicles, and number of trips. None of these variables dramatically impact the proportions of transit and not-transit observations between child feature dimensions. Survey respondents whose transit pass status is unknown reveal an even class distribution (presumably random) and do not produce further splits.

### 5.9.2 Model Scoring

Scoring a model (extrasample) involves running testing data through a trained DT to evaluate the DT's out-of-sample performance. Extrasample scoring is used for model assessment, whereas intrasample scoring is used for model selection during the hyperparameter tuning stage. During the hyperparameter tuning stage (cross-validation procedure), DTs trained on $K \neq k$ data are scored by running $k$ subsamples (folds) through the DTs to derive predictions, respectively. Although each validation fold is successively excluded from the training sample during CV, the final trained tree uses the entire training dataset ($K$) to derive its rules. During the extrasample scoring stage, the DT is trained with all the training data ($K$) and is scored using thus far held-out testing data to produce an unbiased assessment.

During the scoring process, rules produced by the trained DT (with hyperparameter set $\theta_{4,1,0.5}$) are applied to the testing dataset's attributes so that each testing observation falls into a leaf node. This research uses SPSS Modeler to score the trained DT based on the highest probability class at a leaf node.

### 5.9.3 Model Predictions: Highest Probability

Using SPSS Modeler's scoring process, the predicted mode for each testing observation is the highest probability class of the leaf into which the testing observation fell. For example, if the validation sample landed in a leaf where 55% of training observations were in the transit class (45% in the not-transit class), the predicted choice would be transit. Recall that class weights were applied during the DT training process and that proportions within each leaf therefore represent the number of balanced, expanded training observations that fall into each node. Column totals in Table 21 show the resulting predictions for the testing dataset based on the highest probability at leaves.

*Table 21: Confusion Matrix: Testing Data Applied to Trained Model*

| | Predicted Class | | | |
|---|---|---|---|---|
| **Actual Class** | Transit | Not_Transit | Total | Recall (%) |
| Transit | 19,060 | 2,702 | 21,762 | 87.58% |
| Not_Transit | 29,340 | 150,877 | 180,217 | 83.72% |
| Total | 48,400 | 153,579 | 201,979 | 84.14% |
| Precision | 39.38% | 98.24% | | |

Scoring the tree based on the testing data, resulting recall values are comparable to the results from the cross-validation. The RBMC model accuracy of 84.14% is lower than the average accuracy from the cross validation (86.9%). Model sensitivity (transit recall) is higher at 87.58% compared with 84.0%, and specificity (not-transit recall) is lower at 83.72% compared with 86.9%. The precision of not_transit predictions is high, whereas the precision of the transit class is low. Less than half of the model's transit predictions actually belonged to transit, even though over 87% of testing observations were predicted correctly. The model tends to overestimate transit choices. One possible cause of transit overestimation is the rounding error produced by taking the highest probability class at a leaf. Therefore, alternative sample enumeration methods may improve precision.

### 5.9.4 Model Predictions: Proportional Enumeration

For a given testing observation's prediction (highest probability class), prediction *confidence* refers to the proportion of training data (weighted) that belongs to the highest-proportion class in that node. For a "Not-Transit" prediction in a leaf where 60% of training data are Not-Transit and 40% are transit, the confidence is 60%. Likewise, a "Transit" prediction at a leaf with 75% Transit and 25% Not-Transit has a confidence of 75%. SPSS Modeler also defines *propensity* to describe the confidence of one specific target class. This research calculates the propensity relative to the Transit class. Therefore, the propensity of transit predictions is the same as the confidence value, and the propensity of not-transit predictions is equal to 1-confidence. This research uses the propensity to enumerate the popular mode shares using a simple Python script that interprets SPSS Modeler prediction outputs, shown in Figure 22.

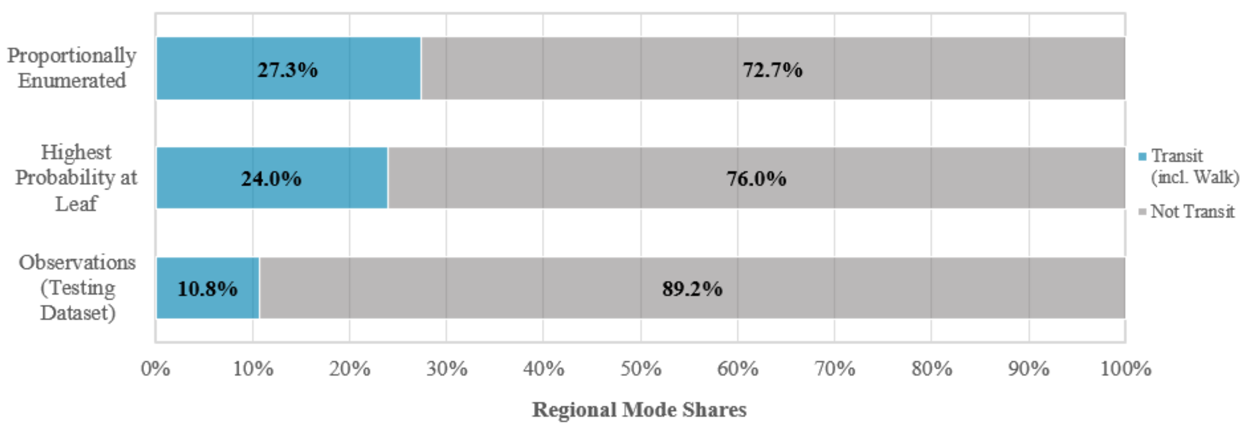



*Figure 22: Predicted Mode Shares by Scoring Method Compared with Observed Outcomes*

Confusion matrices cannot represent results from proportional enumeration because predicted choices are not discrete. Therefore, the class-specific MOEs, including class-specific recall and precision, are only reported for the "highest probability at leaf" method.

Figure 22 compares the observed modes in the testing dataset with the predicted modes for highest-probability and proportional enumeration methods. The proportional enumeration approach seems to overestimate transit (27.3%) more than the highest probability at leaf method (24%). This suggests that transit overestimation in the DT is negatively related to rounding at leaf nodes. Since more leaves result in Not-Transit predictions (i.e., nodes with orange outline in Figure 21), rounding to the highest probability class may have obscured the influence of Transit observations in Not-Transit-dominant leaves more than the influence of Not-Transit observations in Transit-dominant leaves. Few authors explore the trade off between accuracy and precision in mode choice literature, since RBMC analyses (see "Decision Trees and Mode Choice" subsection) tend to emphasize recall values and overall model accuracy. In general ML textbooks, hyperparameter tuning allows modellers to trade off DT precision and accuracy (Han et al., 2011), but alternative tree depth hyperparameters for this specific model appear to suffer worse precision values. This suggests an issue related to the representativeness of the expanded TTS data for transit choice, and further research ought to explore methods to improve representativeness through either bootstrap aggregation (as opposed to cross validation used in this research), data imputation, and investigating data expansion factors (see "Further Research" subsection).

### *5.9.5 Model Interpretability*

The CHAID decision tree algorithm is specifically chosen in this research for its ability to produce a visually interpretable (wide) and statistically-defensible description of possible mode choice decision processes. DT width is important for interpretability because it can reveal differences between the choices of users that belong to different dimensions of the same feature (e.g., people 20-30 vs 30-40 years old) at the same parent node. Some DTs, including the popular CART alternative, encode categorical variables into dummy variables because they perform binary splits, and therefore do not allow comparison of choices within the bins (dimensions) of the same feature (only whether a categorical dimension is true).

The tree depth hyperparameter is analogous to the user's psychological stopping criteria in that it can reflect limits to cognitive effort (e.g., user does not consider more than the first two decision attributes) or the number of factors the user is willing to consider. A higher tree depth represents the potentially more complex user decision rules, but the hyperparameter tuning process showed that past a tree depth of 4, the increase in accuracy is almost negligible, and precision begins to drop. In the mode choice context, this suggests that users are unwilling to consider more than 4 decision elements within their decision process, and that many people will consider even fewer variables than that (i.e., where leaves emerge at depths less than 4).

The DT can represent heterogeneous decision processes because each branch represents a different set of considerations (features) and each node represents a viable stopping point for a user to make a choice. Decisions are not restricted to occurring at leaf nodes (terminals). The user may decide, even partway along a branch (i.e., at an intermediate parent node), that the information gathered and assimilated so far is sufficient for a choice. At this node, the model represents their choice probability through proportions of each class, and a backwards trace towards the root of the DT represents the user's decision process in terms of complexity (length of if-then sequences) and factors (features and their dimensions).

### *5.9.6 Interpreting Non-Compensatory Decision Processes*

The RBMC model structure is consistent with non-compensatory decision processes raised in literature, including lexicographic and aspiration-level based decision rules. The induced model explicitly represents a lexicographic decision process, which relies on ordering attributes (features) by importance. The user selects the best class based on the most important feature, and if no choice is made, the user considers the next most important attribute until a class is selected. The DT branch structure represents this with sequences of hierarchical rules: higher order (lower depth) nodes represent more important attributes and when users proceed along a branch (no choice is made), their probability of choosing a class depends on the sequence of nodes that descend.

The RBMC model's wide, multidimension splitting procedure represents features based on their "degrees" of value (i.e., for ordinal dimensions). Reading the tree across the child nodes of a split can reveal the aspiration levels that users consider when choosing one class over another. Differentiated class proportions across dimensions reveal the direction of association and the aspiration levels to which these users respond. For example, splits along the trip distance (trip_dist) variable (depth 2) show that the aspiration level for a not-transit trip is roughly >1km for most users (inversely, the level for transit choice is <1km for those same users).

### *5.9.7 Interpreting Compensatory Decision Processes*

The RBMC model allows non-compensatory decision processes to emerge without necessarily restricting the compensatory processes. If a compensatory decision process was learned by the DT algorithm, users in all regions would have considered the same set of features. Thus, each level of the decision tree would be occupied by the same feature, regardless of the preceding dimension from which the branch grew. This condition is necessary for the observation of compensatory decision processes because compensatory trade-offs must be observable between the dimensions of two or more features. In an ordinal example, decreases in transit fare costs would result in more transit choice but regardless of (or at least, less sensitively react to) increases in trip distance. One of these features must precede the other, and all regions resulting from the dimensions of the higher-order (lower depth) feature must then split on the same successive feature.

Unfortunately, this decision tree structure cannot deny the possibility of compensatory decision processes. Although the DT does not reveal a compensatory structure, underlying compensatory behaviours may exist because the full set of significant features at a given region is hidden by the DT algorithm's choice of "best" splits (i.e., most significant feature at a given region). This limitation is discussed further in the "Limitations: RBMC Model" subsection. Given this limitation, it is still worth noting that the algorithm's selection criteria "learned" non-compensatory decision rules.

## 5.10 DTA in a RBMC Model

This section tests whether the developed DTA metric may supplement the RBMC model. In a decision process that uses the availability heuristic, DTA is a theoretically relevant variable because it describes temporally aggregate transit service quality that users experience locally and frequently. Availability suggests that users evaluate an outcome's probability based on the ease with which they recall relevant instances. Therefore, if transit services are consistently (over time) poor, users will evaluate negative transit service outcomes as likely and perhaps less likely take transit.

The RBMC model is used for testing DTA because of its non-parametric form. If the model's framework imposes assumptions on the decision process, explanatory variables specified within that framework are limited in how they affect the decision. In the case of the utility function, the assumption of a compensatory strategy forces decision variables to influence a decision by the linear utility functions that relate variable quantities with their coefficients. A RBMC model presents an opportunity to flexibly accommodate and test a decision variable's impact on mode choice. Figure 23 presents the method applied in this section.

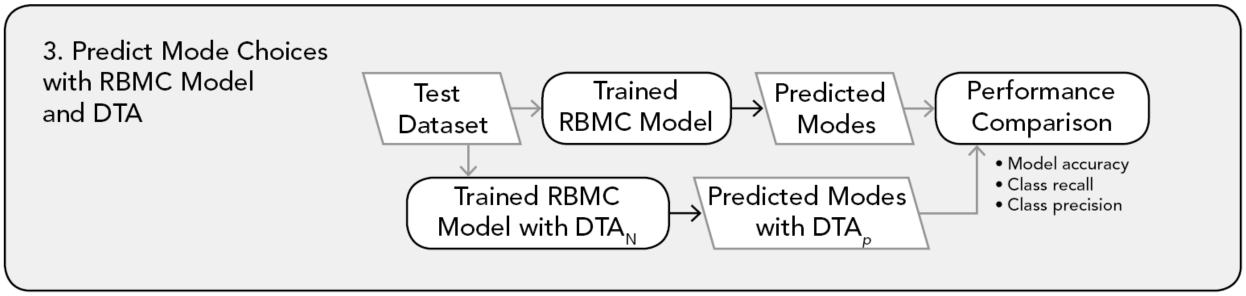



*Figure 23: DTA Variable in RBMC Model*

### *5.10.1 DTA Metric Selection for Mode Choice Model*

"DTA Metric" refers to the DTA measure selected for disaggregate mode choice prediction (feature in training RBMC model). Selection of the appropriate DTA measure for mode choice modelling is based on the aggregate relationships with zonal (aggregate) mode shares observed in Impact of Transit Accessibility  and whether a metric has met the DTA development objectives, repeated here for reference: the DTA metric ought to be connected to transportation demand theory and decision process research; use existing. publicly available datasets; present a straightforward representation of accessibility value; is statistically significant across scenarios; and is location based for spatial comparison.

The all-day, D1 measure of $DTA_m$ for zone $i$, shown in equation [5.9], is most appropriate for use as the DTA metric within a mode choice analysis because of its statistical significance across analysis periods, consistency with travel demand theory, and large sample size given its simple trip attraction terms.

$$D1_{i,AD,m} = ln \frac{\sum_{n \in AD} \sum_{\forall j \neq i} \frac{EMP_j}{TT_{ijn}^2}}{N_{AD}} \tag{5.9}$$

Class D- DTA measures has a stronger theoretical basis than the other location-based measures. The conventional gravity-based measure (class G-) is disconnected from travel demand decision processes because it supposes that users consider the characteristics of their origin zone when making trips. The cumulative opportunity measure (class C-) creates a discrete time boundary, within which destinations are considered accessible, but risks excluding high-attraction zones and misrepresenting accessibility perception along these boundaries (e.g., a destination that is 41 minutes away may be just as attractive as one that is 40 minutes away). D- class measures address these limitations by excluding the origin-based attraction terms and dissolving the boundary of travel, hybridizing the gravity-base and cumulative opportunity measures. Note that some researchers have referred to the distance-decay measure as simply a gravity-based measure (Alam et al., 2010); however, this research distinguishes the two types to understand origin term exclusion effects. GTFS data are almost universally collected from transit agencies and population variables from the TTS (e.g., zonal employment, population, mode shares) are often collected via travel diary surveys. The interpretability of the D1 measure is slightly less than that of the simpler cumulative opportunity measures, which use unweighted activity counts. However, all DTA measures are moderately challenging to interpret due to the temporal resolution of analysis and use of transformations for normality (often log transformed).

Regression results suggest that the first hypothesis about mean transit accessibility's (DTA magnitude, or $DTA_m$) relationship with mode shares is stronger than the second, regarding transit accessibility variation (DTA dispersion, or $DTA_s$). The D1 measure of $DTA_m$ consistently rejects the null hypothesis across all periods and provides some explanatory power for mode shares ($R^2 \leq 0.09$). Nonetheless, despite the stronger relationship (higher $R^2$) between $DTA_m$ and mode shares, $DTA_s$ is a significant predictor (at $\alpha < 0.05$) and is therefore included for DT training. Finally, since the AD analysis encompasses a larger span of time, it is preferable for mode choice analysis because more observations can be used for model estimation.

In addition to the independent variables in Table 16, Table 22 presents the DTA variables this research uses to train the RBMC model. The log-transformed DTA variables' dimensions represent data quantiles to ensure they are balanced.

*Table 22: DTA Variables for Training (Features)*

| Features (Label) | Dimensions | Bins | Unit |
|---|---|---|---|
| *All Day DTA Magnitude (DTA_m)* | (0-3.96], (3.96-4.36], (4.36-4.68], (4.68-5.24], (5.24-11.86] | 5 | ZONE |
| *All Day DTA Dispersion (DTA_s)* | (0-2.39], (2.39-2.72], (2.72-2.94], (2.94-3.25], (3.25-5.18] | 5 | ZONE |

### *5.10.2 Hyperparameter Tuning with DTA*

The hyperparameter search yielded very similar results (Table 23) to the base RBMC model in terms of the quantitative MOEs. The model resulting from a tree depth limit of 4 is chosen (Figure 23).

*Table 23: Hyperparameter Search Details (with DTA)*

| Max. Tree Depth | 1 | 2 | 3 | **4** | 5 |
|---|---|---|---|---|---|
| Overall Model Accuracy | 89% | 82.4% | 79.1% | **84.2%** | 83.9% |
| Transit Recall (Sensitivity) | 42.9% | 79.7% | 88.7% | **87%** | 87.9% |
| Not-Transit Recall (Specificity) | 93.2% | 82.7% | 78% | **83.8%** | 83.3% |
| Transit Precision | 49% | 35.8% | 32.7% | **39.4%** | 39% |
| Not-Transit Precision | 93.2% | 97.1% | 98.3% | **98.2%** | 98.3% |

### 5.10.3 Tree: RBMC Model with DTA

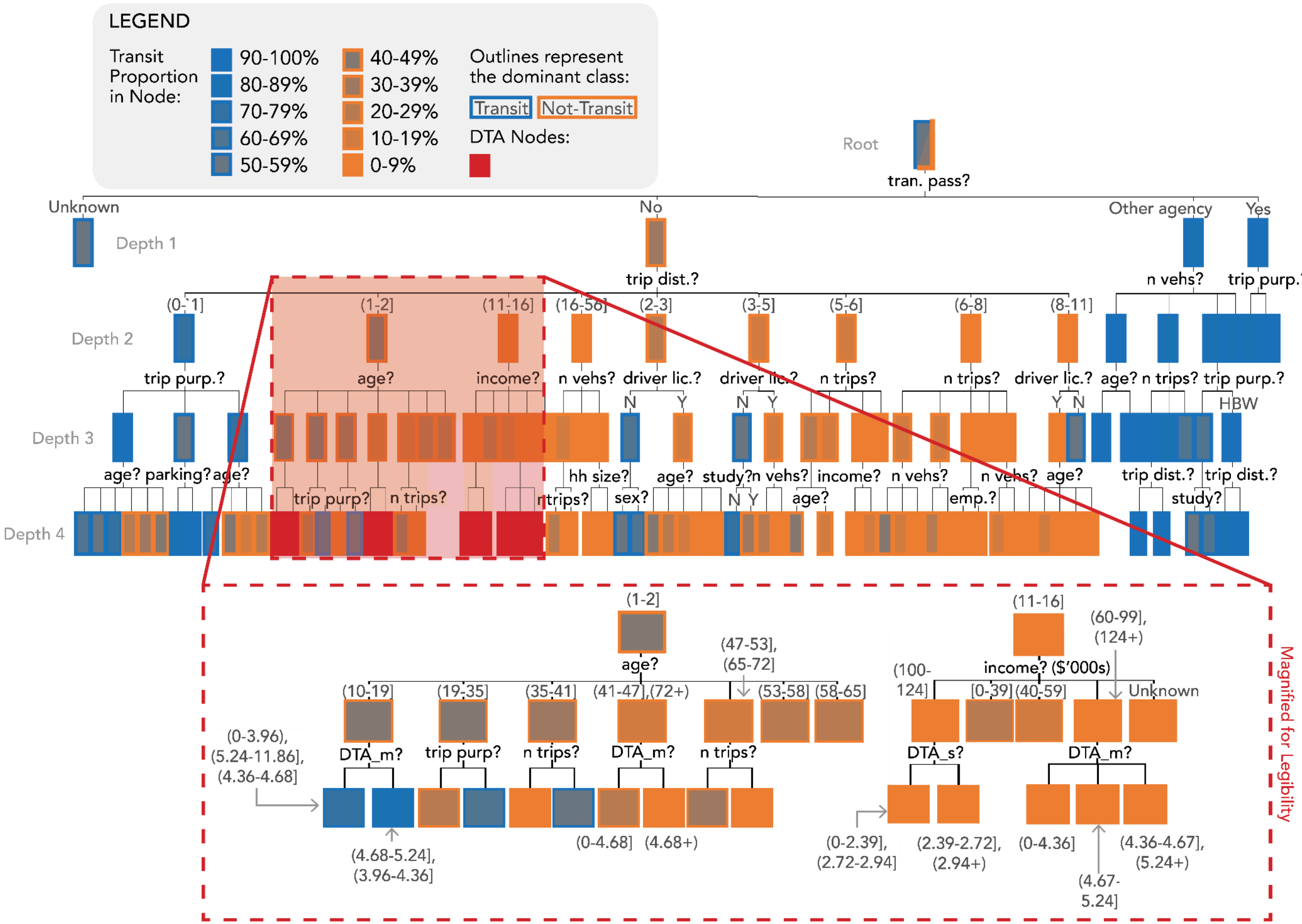



*Figure 24: Trained Decision Tree Diagram (RBMC model with DTA)*

## 5.11 Results and Discussion: RBMC Model with DTA

The raw, unedited output of the RBMC model DT with DTA is in Appendix C with bar charts representing the (unbalanced) proportion of each node that belongs to each class.

This section retrains the decision tree classifier on the same set of observations, now attributed with the $DTA_m$ and $DTA_s$ variables. First, new decision rules induced by including DTA as a feature within the RBMC model are discussed. Then, this analysis evaluates the same MOEs used for the RBMC model analysis (without DTA) to compare model performance. The training and testing datasets are constant between the RBMC model (or "base DT") and the RBMC model with DTA.

### *5.11.1 Learned Decision Rules with DTA*

The base DT remains largely unaffected by the inclusion of DTA variables. Only splits at depth 3 changed in the model, where $DTA_m$ created a split at two places (nodes 23 and 33) and $DTA_s$ produced a split at one (node 30). All of these nodes descend from the branch of users without transit passes. The first $DTA_m$ split (node 23) followed people who travelled a distance between 1 and 2km and are either adolescent (10-19 years) or retired (72-99 years). In the case of the older group, who were already predisposed to not-transit modes (15% transit in the parent node), the split occurred between $DTA_m$values less than 107 and $DTA_m$ values greater than 107 (unitless). Users with lower $DTA_m$ were unlikely to use transit (5%) whereas users with higher $DTA_m$ had a relatively higher probability of using transit (27%). DTA's relationship with transit use is inconsistent for adolescents making short trips: users with $DTA_m$ values in the range of 0-53, 78-107, and >187 had only slightly lower transit use (72%) compared with the groups in between (84%). Medium-distance trips (11-16km) at depth 2 were influenced by both DTA variables ($DTA_m$ and $DTA_s$). Medium-distance trip makers split based on household income (depth 3), where the income group making \$100k-\$120k per year split on $DTA_s$ (node 30) and the group making \$60k-\$99k or >\$124k split on $DTA_m$ (node 33). Unfortunately, these parent nodes already have very low proportions (7% and 1%, respectively) and therefore did not produce very impactful child nodes.

Even though DTA is a statistically significant predictor in the model, its nonetheless small impact on predictions reflects the lexicographic decision process represented in the model. The CHAID would not allow a different split at an earlier depth unless $DTA_m$ or $DTA_s$ were the most statistically significant feature for an earlier region of the training dataset. This emulates lexicographic rules, where no other attributes are considered until after the user evaluates the first, most important attribute. Ensemble methods (discussed in the "Decision Tree Classifiers" subsection of the literature review) present an alternative method to reduce the consistency of the learned ruleset (introduce heterogeneity) by using random subsets of both the data (bootstrapping) and features to induce many decision trees. However, an ensemble method would limit interpretability because model scoring relies on averaging results across many DTs.

One inference from this RBMC application of DTA is related to the comparison between the two travel impedance variables in in the mode choice model: Manhattan distance of trip, and overall transit accessibility magnitude. In contrast with the Manhattan distance of trip (in km) variable, $DTA_m$ describes location-based accessibility, whereas the Manhattan distance measures mobility for a specific trip. Evaluations (splits in the tree) using the trip distance variable are more closely associated with the behavioural response to friction caused by a specific trip (the one observed in the TTS survey). Inclusion of both variables suggests that both general accessibility and accessibility of a specific trip may be important in mode choice decisions. This is consistent with findings in literature that transit choice depends on the combination of good overall transit accessibility and good trip-specific connectivity (Papaioannou & Martinez, 2015). Since trip distance is a measure of mobility (or "goodness of connectivity") for a specific trip, this research expands on the topic by showing the dominant effect of trip-specific impedance compared with overall transit accessibility. Specifically, the RBMC model reveals a common decision process sequence: the user considers trip-specific impedance first, then the distance of the trip, and then overall DTA.

### 5.11.2 Model Predictions with DTA: Highest Probability

Table 24 is the confusion matrix resulting from scoring the RBMC model with DTA on the testing dataset.

*Table 24: Confusion Matrix (DTA): Testing Data Applied to Trained Model*

| | Predicted Class | | | |
|---|---|---|---|---|
| **Actual Class** | Transit | Not_Transit | Total | Recall (%) |
| Transit | 19,019 | 2,743 | 21,762 | 87.4% |
| Not_Transit | 29,188 | 151,029 | 180,217 | 83.8% |
| Total | 48,207 | 153,772 | 201,979 | 84.19% |
| Precision | 39.45% | 98.22% | | |

Scored test data produced recall and precision values similar to the cross validated predictions. Like the base RBMC model results, the column totals of Table 24 show the resulting predictions for the testing dataset based on the highest probability at leaves. This analysis applies the same proportional sample enumeration process explained in the "Model Predictions: Proportional Enumeration" subsection.

Compared with the RBMC model without DTA, there were very minor changes in MOE values, resulting again in the selection of a model with tree depth 4. After including the DTA variables, any changes to overall model accuracy were less than 0.1%. For the recall values at tree depth 4 (no differences at other depths), inclusion of DTA increased the ability to recall transit observations at the expense of not-transit recall. Specifically, transit recall increased by 3% from 84% to 87% and not-transit recall decreased by 3% from 86.9% to 83.8%. There are no differences in precision, throughout.

### 5.11.3 Model Predictions with DTA: Proportional Enumeration

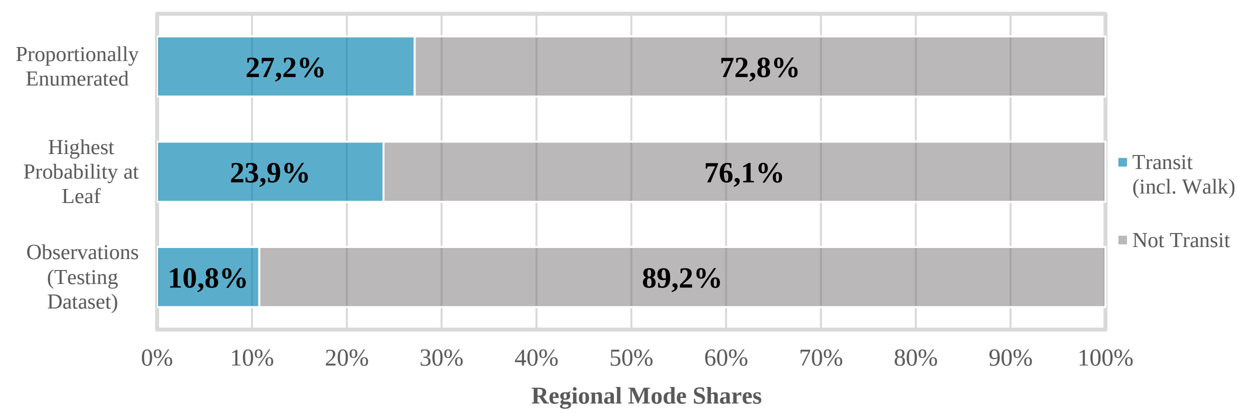



*Figure 25: Predicted Mode Shares by Scoring Method Compared with Observed Outcomes (with DTA)*

The proportional enumeration scoring process at DT leaves also produces very similar results to the base RBMC model. Compared with the non-DTA RBMC model, there is only a 0.1% difference in mode shares, regardless of enumeration method.

## 5.12 Limitations: RBMC Model

This section discusses the limitations of the RBMC modelling method, data, and inclusion of the DTA variable as an independent variable. The Limitations: DTA in Aggregate Modal Analysis section discusses limitations related to the DTA metric development and regression analysis with zonal mode shares.

The RBMC model's induction (tree learning) process is limited by its dependence on sufficiently large, representative, and balanced datasets. First, training on datasets with unbalanced classes (i.e., different sample sizes between modes) tends to bias more frequently sampled classes (Hastie et al., 2008). This research's aggregation of trips into only two classes (Transit and Not-Transit) may reduce this effect; however, Not-Transit observations are still much more frequent in the sample. The class balancing process (see "Class Balancing" subsection) increased the weight of every transit observation (oversampling) by 8.28 to train the decision tree. Unfortunately, increasing the influence of transit observations may have overinflated the representativeness of existing transit samples and biased the induced rulesets, contributing to the high transit prediction during extra-sample scoring. Although class balancing improved representativeness of the Transit class, sample sizes would ideally be sufficiently large and balanced to induce robust rules across target classes without balancing or aggregating modes. DT algorithms also favour high-dimensionality features for their potential to increase purity within resulting child nodes. Hastie et al. (2008) warn that although partitioning algorithms favour features with many dimensions, such variables can result in "severe overfitting" and should therefore be avoided. Unfortunately, dimensionality reduction may also reduce prediction precision and interpretation of results because the behaviours of small groups within some features (e.g., people making $50-55 thousand a year) may be obfuscated within the lower-dimensionality aggregations (e.g., a group making $40-80 thousand a year). Sample representativeness is an important limitation in RBMC modelling because DTs are insensitive to (unable to classify) feature values beyond those observed in the training data. Due to the partitioning process, DTs can neither interpret values beyond input feature dimensions (no extrapolation) nor differentiate effects within dimension boundaries (interpolate) (Zhao et al., 2020). Without representative training datasets that cover the full extent of possible values, scoring the RBMC model using outside samples (i.e., possibly introducing new feature dimensions) may cause classification errors.

Applying the DTA metric to the RBMC model can introduce limitations from transit accessibility measurement (see "Limitations: DTA in Aggregate Modal Analysis" subsection) to the mode choice model. Since transit accessibility measurement determines a baseline accessibility value (accessibility does not drop to 0), researchers must aggregate transit and walking classes to form the dependent variable in mode share analyses (Owen & Levinson, 2015). Disentangling the base transit accessibility value (the walking accessibility) from transit accessibility would allow the RBMC model to distinguish separate heuristic rules for each mode and induce more powerful transit choice rules.

The CHAID algorithm also limits detailed analysis of the DTA metric in a RBMC model structure. CHAID operates by selecting the "best" splitting features at every decision (parent) node. The inability to prune or force the selection of specific features for splitting at each decision node limits feature-oriented analysis efforts (i.e., of specific independent variables), in favour of class-oriented analyses. A region of data may have a set of statistically significant features where CHAID would choose the most significant (i.e., highest $\chi^2$ statistic, lowest p-value) feature and ignore other, also statistically valid splits. For example, the algorithm may have found that in the first region (entire training dataset), four variables – including transit pass ownership, Manhattan distance of trip, age, and DTA – would produce statistically significant splits. However, the algorithm only selects transit pass ownership because it happens to be the "most significant" and ignores the effects of other features until after the first partition. Since the training dataset is split, the other three variables may not have a significant relationship with subsequent regions. The analyst therefore cannot identify alternative regions on which the DTA variable may have significant influence.

# Chapter 6 Conclusion

This thesis seeks to understand whether and how transit accessibility to land uses impacts the mode choice decision process. The DTA metric development aims to reflect the user's heuristic recollection of generally positive or negative transit experiences over time. The application of a RBMC model aims to improve descriptions of mode choice behaviour by departing from utility models and their compensatory assumptions without losses in accuracy. Putting them together, the RBMC model provided a non-compensatory framework to investigate DTA's effect size on mode choices and the DTA variable introduced travel impedance-weighted land use attraction terms to the mode choice model.

The regression analysis in Chapter 4 identified a transit accessibility metric for application in the RBMC model. This research compared three classes of transit accessibility, including gravity-based, time-decayed opportunity, and cumulative opportunity measures, based on their correlations with aggregate (zonal) mode shares. Regression analyses considered measure sensitivity to periods of the day (AM, MD, PM, EE, NT, and AD) and attraction terms (discretionary, employment, and population). Based on its explanatory power ($R^2 = 0.07$), statistical significance, interpretability, and large sample size (1592 zones), an all-day time-decayed opportunity measure, D1 (see equation [5.9]), was selected as a feature in the RBMC model.

In Chapter 5, a rules-based, binary mode choice model using the CHAID algorithm yielded high accuracy, low precision, and interpretable heuristics for mode choice processes. When scored on testing data, the RBMC's overall model accuracy of 84.14% was similar to reported DT accuracies in the literature. The DT produced high recall rates for testing observations, correctly predicting 87.58% of transit observations and 83.72% of not transit observations. However, transit predictions were imprecise: only 39.38% of the model's transit predictions are truly transit observations whereas 98.24% of not transit predictions are correct. Two different sample enumeration strategies, higher probability at leaf and proportional enumeration, produced regional mode splits that overestimated the transit mode share (24% and 27.3%, respectively) compared with the 10.8% transit share observed in the testing data. Model interpretability is high due to the hierarchical, wide structure of the CHAID decision tree. Prior to including DTA, results suggested that users evaluate their transit pass ownership and trip distance with the greatest importance when choosing transit. DTA's inclusion in the RBMC model showed that DTA magnitude (represented as the average D1 transit accessibility throughout the day) is a significant predictor of mode choice for non-transit pass users who are travelling short distances (>1km and $\leq$ 2km). However, this accounted for a small subsection of the population and no higher-order (i.e., lower tree depth) regions of the trained DT changed. Predictions of the testing data scored similarly (84.19% overall accuracy) due to DTA's small effect size. DTA only affected a few predictions at the fourth tree depth.

The learned decision trees provide valuable descriptive insight into heuristic decision process structures and the features that users judge important. The RBMC model's hierarchical structure is consistent with a lexicographic decision process, where users evaluate features ordinally based on feature importance. Based on the first tree splits, the most important factors for transit choice are transit pass ownership and the trip distance, where transit pass ownership has a positive association with transit use. Among users without transit passes, trips with low to medium distances still have high probabilities of taking transit. Within the non-transit-pass subset, which represents most of the training dataset that does not take transit, conditional transit users still occupy some regions: users making trips shorter than 1km have a 76% chance of taking transit, and users who make 2-5km trips have a 53-63% chance of taking transit, if they also do not have driver's licenses. Including DTA in the rule induction process revealed that general (location-based) accessibility is not as important in transit choice as the impedance related to a specific trip. For some users, both high transit accessibility magnitude and high trip-specific mobility (i.e., low trip distance, related to high accessibility) is necessary to take transit.

## 6.1 Further Research

Further research can extend the understanding of accessibility impacts on mode choice and improve the representativeness of the RBMC model.

This research makes a case for the further development of DTA variables. Given transit accessibility is an important predictor of mode choice here and elsewhere (see "Transit Accessibility and Mode Choice" subsection) the accuracy of the specification presented in this research may be improved. Towards approximating a more detailed understanding of general accessibility's impact on mode choices, further analysis ought to extend the testing of dynamic accessibility using different specifications and datasets. The positive, statistically significant correlation between DTA and mode shares found in this research suggests that users respond to the period-level aggregations of transit accessibility when choosing modes. Future research should investigate user perception, rules-based or otherwise, of accessibility for different modes and trip types (i.e., via new attraction terms). Whether people evaluate the accessibility of standalone modes in their decision process or the accessibility of all modes before applying decision criteria is unknown. Applying network skims (i.e., travel impedance matrices) for different modes, such as those required in logit model predictions, in RBMC models may support model prediction. Within transit accessibility research, analysis of different measures of DTA dispersion ($DTA_s$) may also discern more relevant predictors of transit choice. DTA dispersion's weaker and unexpected positive correlation with mode choices is difficult to interpret because of baseline transit accessibility effects and dependency on the accessibility magnitude ($DTA_m$) (see "Limitations: DTA in Aggregate Modal Analysis" subsection). One avenue of further research is to examine the interaction effects of $DTA_m$ and $DTA_s$ as a combined index, respecting that low dispersion transit accessibility is not necessarily an indicator of "good" or "poor" services.

This thesis selected CHAID, a single-tree DT algorithm, to produce a visually interpretable diagram of discrete user decision processes (i.e., if-then-else sequences). Then, a K-fold cross-validation approach was used to reduce sample bias rather than popular ensemble methods in the supervised classifier space. Bootstrap aggregation (bagging) methods, including random forests or CHAID forests, ought to be explored in further research to improve prediction accuracy and precision at the expense of computation time and interpretability. Although other researchers have applied random forests in the mode choice space, none have included the influence of DTA in their models or applied CHAID forests in mode choice (Van Middelkoop et al., 2003; Wang & Ross, 2018; Zhao et al., 2020).

Predictive performance of the RBMC model in terms of accuracy is high but its poor precision results in the dramatic overestimation of regional transit shares. Future research may improve RBMC model precision through increasing the representativeness of infrequent class observations (e.g., of transit). This may be possible through exploring data imputation methods that can reduce the number of missing values (i.e., reduce DT limitations related to data hungriness), or by approximating a minimum sample size in travel diary surveys for transit to support transit choice analysis (i.e., build a more representative transit user profile). A more qualitative research direction would involve supplementing the travel diary data with stated preference surveys of mode choice decision factors, integrating methods such as Hannes et al. (2009); however, such surveys would dramatically increase collection and interpretation difficulty while constraining model transferability. Finally, this research makes the argument that future research should continue to apply supervised ML algorithms to model non-compensatory decision processes in transportation. The CHAID algorithm is easy to implement given adequate modeller knowledge of machine learning processes. However, its popularity seems to be hindered by the required knowledge of software and domain to select an appropriate ML implementation. As further work continues to apply supervised ML algorithms to model non-compensatory decision processes, the transferability of this research's methodology and understanding of heuristic behaviours may improve.

# Bibliography


Aarts, H., Verplanken, B., & Van Knippenberg, A. (1997). Habit and information use in travel mode choices. *Acta Psychologica*, *96*(1–2), 1–14. https://doi.org/10.1016/s0001-6918(97)00008-5

Alam, B. M., Thompson, G. L., & Brown, J. R. (2010). Estimating transit accessibility with an alternative method. *Transportation Research Record: Journal of the Transportation Research Board*, *2144*, 62–71. https://doi.org/10.3141/2144-08

Banister, D. (1978). The influence of habit formation on modal choice -- A heuristic model. *Transportation*, *7*, 19–33.

Banister, D. (2002). *Transport Plannning* (2nd ed.). Spon Press.

Beimborn, E. A., Greenwald, M. J., & Jin, X. (2003). Accessibility, Connectivity, and Captivity Impacts on Transit Choice. *Transportation Research Record*, *03*, 1–9.

Ben-Akiva, M., & Lerman, S. R. (1979). Disaggregate travel and mobility choice models and measures of accessibility. In D. A. Hensher & P. R. Stopher (Eds.), *Behavioural travel modelling* (pp. 654–679).

Ben-Akiva, M., & Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand* (M. L. Manheim (ed.)). The MIT Press.

Bishop, C. M. (2011). *Pattern Recognition and Machine Learning* (M. Jordan, J. Kleinberg, & B. Schölkopf (eds.)). Springer. https://doi.org/10.13109/9783666604409.185

Boisjoly, G., & El-Geneidy, A. (2016). Daily fluctuations in transit and job availability: A comparative assessment of time-sensitive accessibility measures. *Journal of Transport Geography*, *52*. https://doi.org/10.1016/j.jtrangeo.2016.03.004

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211–252.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classfication and Regression Trees*. Chapman & Hall/CRC.

Burns, L. D., & Golob, T. F. (1976). The role of accessibility in basic transportation choice behaviour. *Transportation*, *5*, 175–198. https://link-springer-com.proxy.lib.uwaterloo.ca/content/pdf/10.1007%2FBF00167272.pdf

Cascetta, E., Armando, C., & Montanino, M. (2013). A new measure of accessibility based on perceived opportunities. *Procedia - Social and Behavioural Sciences, 87*, 117–132. https://reader.elsevier.com/reader/sd/pii/S1877042813040433?token=DB8DFFC46319ABF4916B62DEC6C36F13CABC7E8B924E51B65D3AE746B17C592D7133EEF8E8B0696A1EFA6446ED78040A

Cervero, R. (2002). Built environments and mode choice: Toward a normative framework. *Transportation Research Part D: Transport and Environment*, *7*(4), 265–284. https://doi.org/10.2307/4416969

Cheng, L., Chen, X., De Vos, J., Lai, X., & Witlox, F. (2019). Applying a random forest method approach to model travel mode choice behavior. *Travel Behaviour and Society, 14*, 1–10. https://doi.org/10.1016/j.tbs.2018.09.002

Chorus, C. G., & de Jong, G. C. (2011). Modeling experienced accessibility for utility-maximizers and regret-minimizers. *Journal of Transport Geography*, *19*(6), 1155–1162. https://doi.org/10.1016/j.jtrangeo.2011.02.009

Chow, L.-F., Zhao, F., Liu, X., Li, M.-T., Ubaka, I., Chow, L.-F., Zhao, F., & Liu, X. (2006). Transit Ridership Model Based on Geographically Weighted Regression. *Transportation Research Record: Journal of the Transportation Research Board*, *1972*(1), 105–114.

Cirillo, C., Eboli, L., & Mazzulla, G. (2011). On the Asymmetric User Perception of Transit Service Quality. *International Journal of Sustainable Transportation*, *5*(4), 216–232. https://doi.org/https://doi.org/10.1080/15568318.2010.494231

Data Management Group. (2018). *Transportation Tomorrow Survey*. www.dmg.utoronto.ca

Dell'Olio, L., Ibeas, A., & Cecin, P. (2011). The quality of service desired by public transport users. *Transport Policy*, *18*(1), 217–227. https://doi.org/10.1016/j.tranpol.2010.08.005

Domencich, T., & McFadden, D. L. (1975). *Urban Travel Demand: A Behavioural Analysis*. North-Holland Publishing Co. https://eml.berkeley.edu/~mcfadden/travel.html

Eboli, L., & Mazzulla, G. (2011). A methodology for evaluating transit service quality based on subjective and objective measures from the passenger's point of view. *Transport Policy*, *18*(1), 172–181. https://doi.org/10.1016/j.tranpol.2010.07.007

El-Geneidy, A., Buliung, R., Diab, E., Van Lierop, D., Langlois, M., & Legrain, A. (2016). Non-stop equity: Assessing daily intersections between transit accessibility and social disparity across the Greater Toronto and Hamilton Area (GTHA). *Planning and Design*, *43*(3), 540–560. https://doi.org/10.1177/0265813515617659

El-Geneidy, A. M., & Levinson, D. M. (2006). Access to Destinations: Development of Accessibility Measures. In *Technical Report*.

Esri. (2020). *Create and use a network dataset with public transit data*. Network Analyst Tutorials. https://pro.arcgis.com/en/pro-app/help/analysis/networks/create-and-use-a-network-dataset-with-public-transit-data.htm

Ewing, R. H. (1973). Psychological Theory Applied to Mode Choice Prediction. *Transportation*, *2*, 391–392.

Fang, Z., Shaw, S.-L., & Li, Q. (2010). A sensitive indicator of regional space–time accessibility. *Annals of GIS*, *16*(3), 155–164. https://doi.org/10.1080/19475683.2010.513151

Farber, S., & Fu, L. (2016). Dynamic public transit accessibility using travel time cubes: Comparing the effects of infrastructure (dis)investments over time. *Computers, Environment and Urban Systems*, *62*, 30–40. https://doi.org/10.1016/j.compenvurbsys.2016.10.005

Fayyaz, S. K., Liu, X. C., & Porter, R. J. (2017). Dynamic transit accessibility and transit gap causality analysis. *Journal of Transport Geography*, *59*, 27–39. https://doi.org/10.1016/j.jtrangeo.2017.01.006

Fayyaz, S. K., Liu, X. C., & Zhang, G. (2017). An efficient General Transit Feed Specification (GTFS) enabled algorithm for dynamic transit accessibility analysis. *PLoS ONE*, *12*(10), e0185333. https://doi.org/10.1371/journal.pone.0185333

Foerster, J. F. (1979). Mode choice decision process models: A comparison of compensatory and non-compensatory structures. *Transportation Research Part A: General*, *13*(1), 17–28. https://doi.org/10.1016/0191-2607(79)90083-9

Franke, T. M., Ho, T., & Christie, C. A. (2011). The Chi-Square Test: Often used and more often misinterpreted. *American Journal of Evaluation*, *33*(3), 448–458. https://doi.org/10.1177/1098214011426594

Fransen, K., Neutens, T., Farber, S., De Maeyer, P., Deruyter, G., & Witlox, F. (2015). Identifying public transport gaps using time-dependent accessibility levels. *Journal of Transport Geography*, *48*, 176–

187. https://doi.org/10.1016/j.jtrangeo.2015.09.008

Fu, L. (2017). *Modelling Travel Demand for Urban Transportation Planning*.

Geurs, K. T., & Ritsema van Eck, J. R. (2001). *Accessibility measures: review and applications* (Vol. 787).

Geurs, K. T., & van Wee, B. (2004). Accessibility evaluation of land-use and transport strategies: Review and research directions. *Journal of Transport Geography*, *12*(2), 127–140. https://doi.org/10.1016/j.jtrangeo.2003.10.005

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, *62*(1), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Gigerenzer, G., & Gaissmaier, W. (2015). Decision Making: Nonrational Theories. In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition* (Second Edi, Vol. 5). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.26017-0

Google. (2020). *Static Transit Reference*. GTFS Extensions. https://developers.google.com/transit/gtfs/reference

Guiliano, G., & Agarwal, A. (2017). Land Use Impacts of Transportation Investments. In G. Guiliano & S. Hanson (Eds.), *The Geography of Urban Transportation* (Fourth, pp. 218–246).

Hägerstrand, T. (1970). What about people in regional science? *Papers of the Regional Science Association*, *24*, 7–21.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)* (Third). Morgan Kaufmann Publishers. http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

Handy, S. L., & Niemeier, D. A. (1997). Measuring accessibility: An exploration of issues and alternatives. *Environment and Planning A*, *29*(7), 1175–1194.

Hannes, E., Janssens, D., & Wets, G. (2009). Does space matter? Travel mode scripts in daily activity travel. *Environment and Behavior*, *41*(1), 75–100. https://doi.org/10.1177/0013916507311033

Hansen, W. G. (1959). How Accessibility Shapes Land Use. *Journal of the American Planning Association*, *25*(2), 73–76. https://doi.org/10.1080/01944365908978307

Hanson, S., & Schwab, M. (1987). Accessibility and intraurban travel. *Environment and Planning A*, *19*, 735–748.

Hasnine, M. S., Graovac, A., Camargo, F., & Habib, K. N. (2019). A random utility maximization (RUM) based measure of accessibility to transit: Accurate capturing of the first-mile issue in urban transit. *Journal of Transport Geography*, *74*, 313–320. https://doi.org/10.1016/J.JTRANGEO.2018.12.007

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning Data Mining, Inference, and Prediction* (Second Edi). Springer. https://web.stanford.edu/~hastie/Papers/ESLII.pdf

Higgins, C. D., & Kanaroglou, P. S. (2018). Rapid transit, transit-oriented development, and the contextual sensitivity of land value uplift in Toronto. *Urban Studies*, *55*(10), 2197–2225. https://doi.org/10.1177/0042098017712680

IBM. (2011). CHAID and Exhaustive CHAID Algorithms. In *Index of SPSS Support Statistics* (Issue 13). ftp://ftp.software.ibm.com/software/analytics/spss/support/Stats/Docs/Statistics/Algorithms/13.0/

IBM Corporation. (2012). *IBM SPSS Decision Trees 21*. http://www.ibm.com/spss

Innocenti, A., Lattarulo, P., & Pazienza, M. G. (2013). Car stickiness: Heuristics and biases in travel choice. *Transport Policy*, *25*, 158–168. https://doi.org/10.1016/j.tranpol.2012.11.004

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*.

Järv, O., Tenkanen, H., Salonen, M., Ahas, R., & Toivonen, T. (2018). Dynamic cities: Location-based accessibility modelling as a function of time. *Applied Geography*, *95*, 101–110. https://doi.org/10.1016/j.apgeog.2018.04.009

Jensen, A. F. (2016). Bounded rational choice behaviour: applications in transport. *Transport Reviews*.

Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, *29*(2), 119–127. https://www-jstor-org.proxy.lib.uwaterloo.ca/stable/pdf/2986296.pdf?ab_segments=0%2Fbasic_SYC-5187_SYC-5188%2Fcontrol&refreqid=fastly-default%3Abf0981590493bd5a44a9ada51015e496

Klein, G. (2002). The Fiction of Optimization. In G. Gigerenzer & R. Selten (Eds.), *Bounded Rationality: The Adaptive Toolbox*. The MIT Press.

Krizek, K. J. (2005). Perspectives on Accessibility and Travel. In D. M. Levinson & K. J. Krizek (Eds.), *Access to Destinations* (1st ed., pp. 109–130). Elsevier Ltd.

Krizek, K. J., Iacono, M., El-Geneidy, A., Liao, C. F., & Johns, R. (2009). *Access to Destinations: Application of Accessibility Measures for Non-Auto Travel Modes*. http://www.lrrb.org/PDF/200924.pdf

Kutter, E. (1972). *Demographische Determinanten des Städtlischen Personenverkehrs*.

Lancaster, K. J. (1966). A New Approach to Consumer Theory. *Journal of Political Economy*, *74*(2), 132–157. https://www-jstor-org.proxy.lib.uwaterloo.ca/stable/pdf/1828835.pdf?ab_segments=0%2Fbasic_SYC-5152%2Fcontrol&refreqid=search%3A9d7abe94057ecd2531be98706c2d685c

Larson, R. C., & Odoni, A. R. (1981). Ratio of Right Angle to Euclidean Distance Metrics. In *Urban Operations Research*. Prentice-Hall. https://web.mit.edu/urban_or_book/www/book/chapter3/3.1.html

Lee, D., Derrible, S., & Pereira, F. C. (2018). Comparison of four types of artificial neural network and a multinomial logit model for travel mode choice modeling. *Transportation Research Record*, *2672*(49), 101–112. https://doi.org/10.1177/0361198118796971

Lee, J., & Miller, H. J. (2018). Measuring the impacts of new public transit services on space-time accessibility: An analysis of transit system redesign and new bus rapid transit in Columbus, Ohio, USA. *Applied Geography*, *93*(September 2017), 47–63. https://doi.org/10.1016/j.apgeog.2018.02.012

Legrain, A., Buliung, R., & El-Geneidy, A. M. (2015). Who, What, When, and Where: Revisiting the influences of transit mode share. *Transportation Research Record*, *2537*, 42–51. https://doi.org/10.3141/2537-05

Lei, T. L., & Church, R. L. (2010). Mapping transit-based access: integrating GIS, routes and schedules. *International Journal of Geographical Information Science*, *24*(2), 283–304. https://doi.org/10.1080/13658810902835404

Levinson, D. M., & Krizek, K. J. (Eds.). (2005). *Access to Destinations* (1st ed.). Elsevier Ltd.

Lindner, A., Souza Pitombo, C., & Luiz Cunha, A. (2017). Estimating motorized travel mode choice using classifiers: An application for high-dimensional multicollinear data. *Travel Behaviour and*

*Society*, *6*, 100–109. https://doi.org/10.1016/j.tbs.2016.08.003

Liu, S., & Zhu, X. (2004). Accessibility Analyst: An integrated GIS tool for accessibility analysis in urban transportation planning. *Environment and Planning B: Planning and Design*, *31*(1), 105–124. https://doi.org/10.1068/b305

Loh, W. Y., & Shih, Y. S. (1997). Split selection methods for classification trees. *Statistica Sinica*, *7*(4), 815–840. https://www.researchgate.net/publication/2393893_Split_Selection_Methods_for_Classification_Trees

Luce, R. D. (1959). *Individual Choice Behaviour: A Theoretical Analysis*. Wiley.

Manski, C. F. (1973). *The Analysis of Qualitative Choice* [Massachusetts Institute of Technology]. https://dspace.mit.edu/bitstream/handle/1721.1/13927/24931753-MIT.pdf?sequence=2&isAllowed=y

Manski, C. F. (1977). The Structure of Random Utility Models. *Theory and Decision*, *8*, 229–254. https://link-springer-com.proxy.lib.uwaterloo.ca/content/pdf/10.1007/BF00133443.pdf

Marschak, J. (1960). Binary Choice Constraints on Random Utility Indicators. In K. J. Arrow (Ed.), *Stanford Symposium on Mathematical Methods in the Social Sciences*. Stanford University Press. https://link-springer-com.proxy.lib.uwaterloo.ca/content/pdf/10.1007%2F978-94-010-9276-0_9.pdf

Mavoa, S., Witten, K., McCreanor, T., & O'Sullivan, D. (2012). GIS based destination accessibility via public transit and walking in Auckland, New Zealand. *Journal of Transport Geography*, *20*(1), 15–22. https://doi.org/10.1016/j.jtrangeo.2011.10.001

McFadden, D. L. (1974). Conditional Logit Analysis of Qualitative Choice Behaviour. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 105–142). Academic Press. https://eml.berkeley.edu/reprints/mcfadden/zarembka.pdf

McFadden, D. L. (2000). *Economic Choices: Prize Lecture* (pp. 330–366). https://www.nobelprize.org/prizes/economic-sciences/2000/mcfadden/lecture/

McFadden, D. L., & Train, K. (2000). Mixed MNL models for discrete response. *Journal of Applied Econometrics*, *15*(5), 447–470. https://doi.org/10.1002/1099-1255(200009/10)15:5<447::aid-jae570>3.3.co;2-t

Meyer, M. D., & Miller, E. J. (2001). *Urban Transportation Planning: A Decision-oriented Approach*. McGraw-Hill.

Miguel Martínez, L., & Viegas, J. M. (2013). A new approach to modelling distance-decay functions for accessibility assessment in transport studies. *Journal of Transport Geography*, *26*, 87–96. https://doi.org/10.1016/j.jtrangeo.2012.08.018

Miller, H. J. (2017). Time geography and space-time prism. In D. Richardson, N. Castree, M. F. Goodchild, A. Kobayashi, W. Liu, & R. A. Marston (Eds.), *The International Encyclopedia of Geography*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118786352.wbieg0431

Moniruzzaman, M., & Páez, A. (2012). Accessibility to transit, by transit, and mode share: application of a logistic model with spatial filters. *Journal of Transport Geography*. https://doi.org/10.1016/j.jtrangeo.2012.02.006

Moya-Gómez, B., Salas-Olmedo, M. H., García-Palomares, J. C., & Gutiérrez, J. (2018). Dynamic Accessibility using Big Data: The Role of the Changing Conditions of Network Congestion and Destination Attractiveness. *Networks and Spatial Economics*, *18*(2), 273–290. https://doi.org/10.1007/s11067-017-9348-z

Murphy, B., & Owen, A. (2019). Temporal sampling and service frequency harmonics in transit accessibility evaluation. *Journal of Transport and Land Use*, *12*(1), 893–913. https://doi.org/10.5198/jtlu.2019.1379

Nassir, N., Hickman, M., Malekzadeh, A., & Irannezhad, E. (2016). A utility-based travel impedance measure for public transit network accessibility. *Transportation Research Part A: Policy and Practice*, *88*, 26–39. https://doi.org/10.1016/j.tra.2016.03.007

National Academies of Sciences Engineering and Medicine. (2018). *Integrated Transportation and Land Use Models: A Synthesis of Highway Practice*. National Academies Press. https://doi.org/10.17226/25194

Neutens, T., Delafontaine, M., Scott, D. M., & De Maeyer, P. (2012). An analysis of day-to-day variations in individual space-time accessibility. *Journal of Transport Geography*, *23*, 81–91. https://doi.org/10.1016/j.jtrangeo.2012.04.001

Niemeier, D. A. (1997). Accessibility: An evaluation using consumer welfare. *Transportation*, *24*(4), 377–396. https://doi.org/10.1023/A:1004914803019

O'Sullivan, D., Morrison, A., & Shearer, J. (2000). Using desktop GIS for the investigation of accessibility by public transport: an isochrone approach. *International Journal of Geographical Information Science*, *14*(1), 85–104. https://doi.org/10.1080/136588100240976

Ontario Ministry of Finance. (2020). *2019–20 Third Quarter Finances*. https://www.fin.gov.on.ca/en/budget/finances/2019/ofin19_3.pdf

Ortúzar, J. de D., & Willumsen, L. G. (2011). Modelling Transport. In *Modelling Transport* (Fourth Edi). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119993308

Owen, A., & Levinson, D. M. (2015). Modeling the commute mode share of transit using continuous accessibility to jobs. *Transportation Research Part A*, *74*, 110–122. https://doi.org/10.1016/j.tra.2015.02.002

Papaioannou, D., & Martinez, L. M. (2015). The role of accessibility and connectivity in mode choice. A structural equation modeling approach. *Transportation Research Procedia*, *10*(July), 831–839. https://doi.org/10.1016/j.trpro.2015.09.036

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Polzin, S. E., Pendyala, R. M., & Navari, S. (2002). Development of time-of-day-based transit accessibility analysis tool. *Transportation Research Record*, *1799*, 35–41. https://doi.org/10.3141/1799-05

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, *1*, 81–106.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning* (Vol. 5, Issue 3). Morgan Kaufmann Publishers Inc. https://doi.org/10.1007/BF00993309

Rasouli, S., & Timmermans, H. J. P. (2014). Using ensembles of decision trees to predict transport mode choice decisions: Effects on predictive success and uncertainty estimates. *European Journal of Transport and Infrastructure Research*, *14*(14), 515–522.

Region of Waterloo. (2016). *Census Bulletin 2016 Employment and Education*. https://www.regionofwaterloo.ca/en/regional-government/resources/Census/Census-Bulletin-9-Employment-and-Education-ACCESS.pdf

Region of Waterloo. (2019). *Region of Waterloo Open Data*. Roads. https://rowopendata-rmw.opendata.arcgis.com/

Roy, J. R., & Thill, J. C. (2003). Spatial interaction modelling. In *Papers in Regional Science* (Vol. 83, Issue 1, pp. 339–361). https://doi.org/10.1007/s10110-003-0189-4

Saif, M. A., Zefreh, M. M., & Torok, A. (2019). Public transport accessibility: A literature review. *Periodica Polytechnica Transportation Engineering*, *47*(1), 36–43. https://doi.org/10.3311/PPtr.12072

Scott, D. (n.d.). *Box-Cox Transformations*. OnlineStatBook. Retrieved November 30, 2020, from http://onlinestatbook.com/2/transformations/box-cox.html

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99. https://doi.org/10.2307/1884852

Simon, H. A. (1972). Theories of Bounded Rationality. In *Decision and Organization: Vol. Chapter 8* (pp. 161–176).

Stępniak, M., & Goliszek, S. (2017). Spatio-temporal variation of accessibility by public transport—The equity perspective. In *The Rise of Big Spatial Data* (Issue October, pp. 241–261). Springer. https://doi.org/10.1007/978-3-319-45123-7

Stępniak, M., Pritchard, J. P., Geurs, K. T., & Goliszek, S. (2019). The impact of temporal resolution on public transport accessibility measurement: Review and case study in Poland. *Journal of Transport Geography*, *75*(July 2018), 8–24. https://doi.org/10.1016/j.jtrangeo.2019.01.007

Straatemeier, T., & Bertolini, L. (2019). How can planning for accessibility lead to more integrated transport and land-use strategies? Two examples from the Netherlands. *European Planning Studies*. https://doi.org/10.1080/09654313.2019.1612326

Sutton, C. D. (2005). Classification and Regression Trees, Bagging, and Boosting. In *Handbook of Statistics* (Vol. 24, pp. 303–329). Elsevier B.V. https://doi.org/10.1016/S0169-7161(04)24011-1

Taylor, B. D., Miller, D., Iseki, H., & Fink, C. (2003). Analyzing the Determinants of Transit Ridership Using a Two-Stage Least Squares Regression on a National Sample of Urbanized Areas. *2004 Annual Meeting of the Transportation Research Board*.

The Centre for Spatial Economics. (2017). *The Economic Benefits of Public Infrastructure Spending in Ontario*. http://www.c4se.com/documents/Ontario Public Infrastructure Final Report.pdf

Thurstone, L. L. (1927). A Law of Comparative Judgement. *Psychological Review*, *34*, 273–286. https://search-proquest-com.proxy.lib.uwaterloo.ca/docview/614335671?OpenUrlRefId=info:xri/sid:primo&accountid=14906

Timmermans, H. (1983). Non-compensatory decision rules and consumer spatial choice behaviour: A test of predictive ability. *The Professional Geographer*, *35*(4), 449–455. https://doi.org/10.1111/j.0033-0124.1983.00449.x

Timmermans, H., & Golledge, R. G. (1990). Applications of behavioural research on spatial problems II: preference and choice. *Progress in Human Geography*, *14*(3), 311-354[1] Prog. Hum. Geogr., vol. 14, no. 3, p.

Train, K. (2002). *Discrete Choice Methods with Simulation* (Second). Cambridge University Press. https://doi.org/10.1017/cbo9780511805271.006

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458. https://science.sciencemag.org/content/211/4481/453

van der Pligt, J. (2015). Decision Making, Psychology of. In *International Encyclopedia of the Social & Behavioral Sciences* (pp. 917–922). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.24014-2

Van Middelkoop, M., Borgers, A., & Timmermans, H. (2003). Inducing heuristic principles of tourist choice of travel mode: A rule-based approach. *Journal of Travel Research*, *42*(1), 75–83. https://doi.org/10.1177/0047287503254116

Wang, F., & Ross, C. L. (2018). Machine Learning Travel Mode Choices: Comparing the Performance of an Extreme Gradient Boosting Model with a Multinomial Logit Model. *Transportation Research Record*, *2672*(47), 35–45. https://doi.org/10.1177/0361198118773556

Wegener, M. (1995). Current and future land-use models. In G. A. Shunk, P. L. Bass, C. A. Weatherby, & L. J. Engelke (Eds.), *Travel Model Improvement Program Land-use Modellng Conference Proceedings* (pp. 13–40).

Wegener, M., & Fürst, F. (1999). Land-Use Transport Interaction: State of the Art. In *Berichte aus dem Institut für Raumplanung* (Issue 46). https://doi.org/10.2139/ssrn.1434678

Weisstein, E. W. (n.d.). *Square Line Picking*. From MathWorld--A Wolfram Web Resource.; Wolfram Research, Inc. Retrieved October 26, 2020, from https://mathworld.wolfram.com/SquareLinePicking.html

Wets, G., Vanhoof, K., Arentze, T., & Timmermans, H. (2000). Identifying decision structures underlying activity patterns: An exploration of data mining algorithms. *Transportation Research Record*, *1718*, 1–9. https://doi.org/10.3141/1718-01

Wets, Geert, Vanhoof, K., Arentze, T., & Timmermans, H. (2000). Identifying Decision Structures Underlying Activity Patterns An Exploration of Data Mining Algorithms. *Transportation Research Record*, *1718*(1), 1–9. https://journals.sagepub.com/doi/10.3141/1718-01

Xie, C., Lu, J., & Parkany, E. (2003). Work travel mode choice modeling with data mining: Decision trees and neural networks. *Transportation Research Record: Journal of the Transportation Research Board*, *1854*(1), 50–61.

Zhao, X., Yan, X., Yu, A., Van Hentenryck, P., & Stewart, H. M. (2020). Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel Behaviour and Society*, *20*, 22–35. https://doi.org/10.1016/j.tbs.2020.02.003

# Appendices

## Appendix A: Trained RBMC Nodes Table

| | | | |
|---|---|---|---|
| **Branch and Leaf Node Values** | | | |
| (Node values are calculated using weight factor on n_Transit) | | | |
| WEIGHT: | 8.2875 Transit trips for every Not-Transit trip | | |
| **Node ID** | **transit_proportion** | **nottransit_proportion** | **predicted mode (majority)** |
| 1 | 0.549 | 0.451 | Transit |
| 2 | 0.377 | 0.623 | N/A |
| 3 | 0.896 | 0.104 | N/A |
| 4 | 0.914 | 0.086 | N/A |
| 5 | 0.757 | 0.243 | N/A |
| 6 | 0.451 | 0.549 | N/A |
| 7 | 0.071 | 0.929 | N/A |
| 8 | 0.044 | 0.956 | N/A |
| 9 | 0.280 | 0.720 | N/A |
| 10 | 0.196 | 0.804 | N/A |
| 11 | 0.120 | 0.880 | N/A |
| 12 | 0.102 | 0.898 | N/A |
| 13 | 0.089 | 0.911 | N/A |
| 14 | 0.977 | 0.023 | N/A |
| 15 | 0.876 | 0.124 | N/A |
| 16 | 0.839 | 0.161 | N/A |
| 17 | 0.810 | 0.190 | Transit |
| 18 | 0.861 | 0.139 | Transit |
| 19 | 0.943 | 0.057 | Transit |
| 20 | 0.552 | 0.448 | N/A |
| 21 | 0.930 | 0.070 | N/A |
| 22 | 0.536 | 0.464 | N/A |
| 23 | 0.780 | 0.220 | N/A |
| 24 | 0.476 | 0.524 | N/A |
| 25 | 0.372 | 0.628 | N/A |
| 26 | 0.150 | 0.850 | N/A |
| 27 | 0.197 | 0.803 | N/A |
| 28 | 0.295 | 0.705 | Not_Transit |
| 29 | 0.232 | 0.768 | N/A |
| 30 | 0.074 | 0.926 | N/A |
| 31 | 0.255 | 0.745 | Not_Transit |
| 32 | 0.192 | 0.808 | Not_Transit |

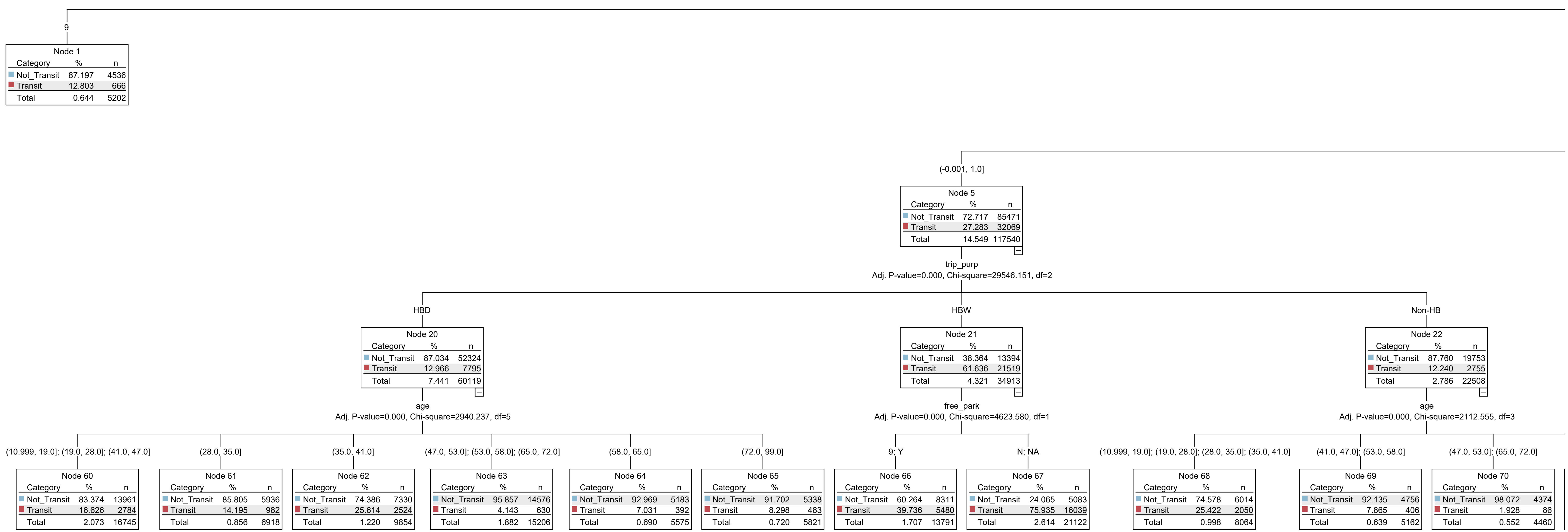
| | | | |
|---|---|---|---|
| 33 | 0.007 | 0.993 | N/A |
| 34 | 0.000 | 1.000 | Not_Transit |
| 35 | 0.181 | 0.819 | N/A |
| 36 | 0.000 | 1.000 | Not_Transit |
| 37 | 0.009 | 0.991 | N/A |
| 38 | 0.631 | 0.369 | N/A |
| 39 | 0.173 | 0.827 | N/A |
| 40 | 0.534 | 0.466 | N/A |
| 41 | 0.113 | 0.887 | N/A |
| 42 | 0.252 | 0.748 | N/A |
| 43 | 0.128 | 0.872 | Not_Transit |
| 44 | 0.081 | 0.919 | N/A |
| 45 | 0.000 | 1.000 | Not_Transit |
| 46 | 0.192 | 0.808 | N/A |
| 47 | 0.136 | 0.864 | N/A |
| 48 | 0.053 | 0.947 | N/A |
| 49 | 0.000 | 1.000 | Not_Transit |
| 50 | 0.009 | 0.991 | N/A |
| 51 | 0.041 | 0.959 | N/A |
| 52 | 0.517 | 0.483 | Transit |
| 53 | 0.958 | 0.042 | Transit |
| 54 | 0.989 | 0.011 | Transit |
| 55 | 0.935 | 0.065 | N/A |
| 56 | 0.848 | 0.152 | Transit |
| 57 | 0.683 | 0.317 | Transit |
| 58 | 0.609 | 0.391 | N/A |
| 59 | 0.911 | 0.089 | N/A |
| 60 | 0.623 | 0.377 | Transit |
| 61 | 0.578 | 0.422 | Transit |
| 62 | 0.740 | 0.260 | Transit |
| 63 | 0.264 | 0.736 | Not_Transit |
| 64 | 0.385 | 0.615 | Not_Transit |
| 65 | 0.428 | 0.572 | Not_Transit |
| 66 | 0.845 | 0.155 | Transit |
| 67 | 0.963 | 0.037 | Transit |
| 68 | 0.738 | 0.262 | Transit |
| 69 | 0.414 | 0.586 | Not_Transit |
| 70 | 0.140 | 0.860 | Not_Transit |
| 71 | 0.277 | 0.723 | Not_Transit |
| 72 | 0.719 | 0.281 | Transit |
| 73 | 0.809 | 0.191 | Transit |

| | | | |
|---|---|---|---|
| 74 | 0.281 | 0.719 | Not_Transit |
| 75 | 0.692 | 0.308 | Transit |
| 76 | 0.077 | 0.923 | Not_Transit |
| 77 | 0.567 | 0.433 | Transit |
| 78 | 0.256 | 0.744 | Not_Transit |
| 79 | 0.036 | 0.964 | Not_Transit |
| 80 | 0.363 | 0.637 | Not_Transit |
| 81 | 0.000 | 1.000 | Not_Transit |
| 82 | 0.389 | 0.611 | Not_Transit |
| 83 | 0.101 | 0.899 | Not_Transit |
| 84 | 0.000 | 1.000 | Not_Transit |
| 85 | 0.133 | 0.867 | Not_Transit |
| 86 | 0.015 | 0.985 | Not_Transit |
| 87 | 0.000 | 1.000 | Not_Transit |
| 88 | 0.241 | 0.759 | Not_Transit |
| 89 | 0.000 | 1.000 | Not_Transit |
| 90 | 0.000 | 1.000 | Not_Transit |
| 91 | 0.025 | 0.975 | Not_Transit |
| 92 | 0.580 | 0.420 | Transit |
| 93 | 0.680 | 0.320 | Transit |
| 94 | 0.428 | 0.572 | Not_Transit |
| 95 | 0.201 | 0.799 | Not_Transit |
| 96 | 0.152 | 0.848 | Not_Transit |
| 97 | 0.068 | 0.932 | Not_Transit |
| 98 | 0.029 | 0.971 | Not_Transit |
| 99 | 0.772 | 0.228 | Transit |
| 100 | 0.337 | 0.663 | Not_Transit |
| 101 | 0.252 | 0.748 | Not_Transit |
| 102 | 0.036 | 0.964 | Not_Transit |
| 103 | 0.475 | 0.525 | Not_Transit |
| 104 | 0.106 | 0.894 | Not_Transit |
| 105 | 0.000 | 1.000 | Not_Transit |
| 106 | 0.129 | 0.871 | Not_Transit |
| 107 | 0.445 | 0.555 | Not_Transit |
| 108 | 0.112 | 0.888 | Not_Transit |
| 109 | 0.000 | 1.000 | Not_Transit |
| 110 | 0.252 | 0.748 | Not_Transit |
| 111 | 0.000 | 1.000 | Not_Transit |
| 112 | 0.000 | 1.000 | Not_Transit |
| 113 | 0.162 | 0.838 | Not_Transit |
| 114 | 0.000 | 1.000 | Not_Transit |

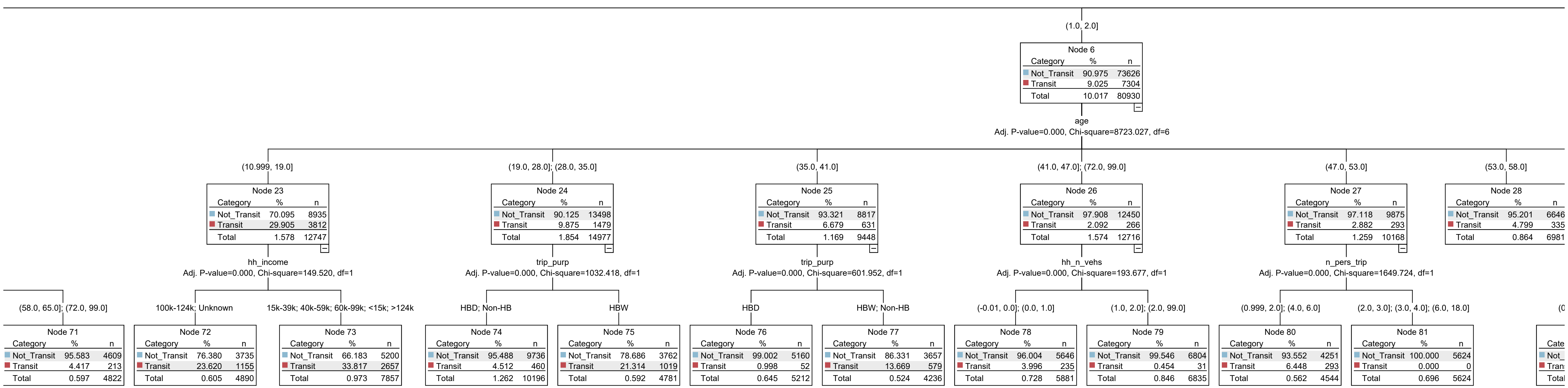
| | | | |
|---|---|---|---|
| 115 | 0.019 | 0.981 | Not_Transit |
| 116 | 0.123 | 0.877 | Not_Transit |
| 117 | 0.036 | 0.964 | Not_Transit |
| 118 | 0.062 | 0.938 | Not_Transit |
| 119 | 0.000 | 1.000 | Not_Transit |
| 120 | 0.022 | 0.978 | Not_Transit |
| 121 | 0.965 | 0.035 | Transit |
| 122 | 0.900 | 0.100 | Transit |
| 123 | 0.568 | 0.432 | Transit |
| 124 | 0.643 | 0.357 | Transit |
| 125 | 0.868 | 0.132 | Transit |
| 126 | 0.933 | 0.067 | Transit |

# Appendix B: Trained Decision Tree Outputs, SPSS

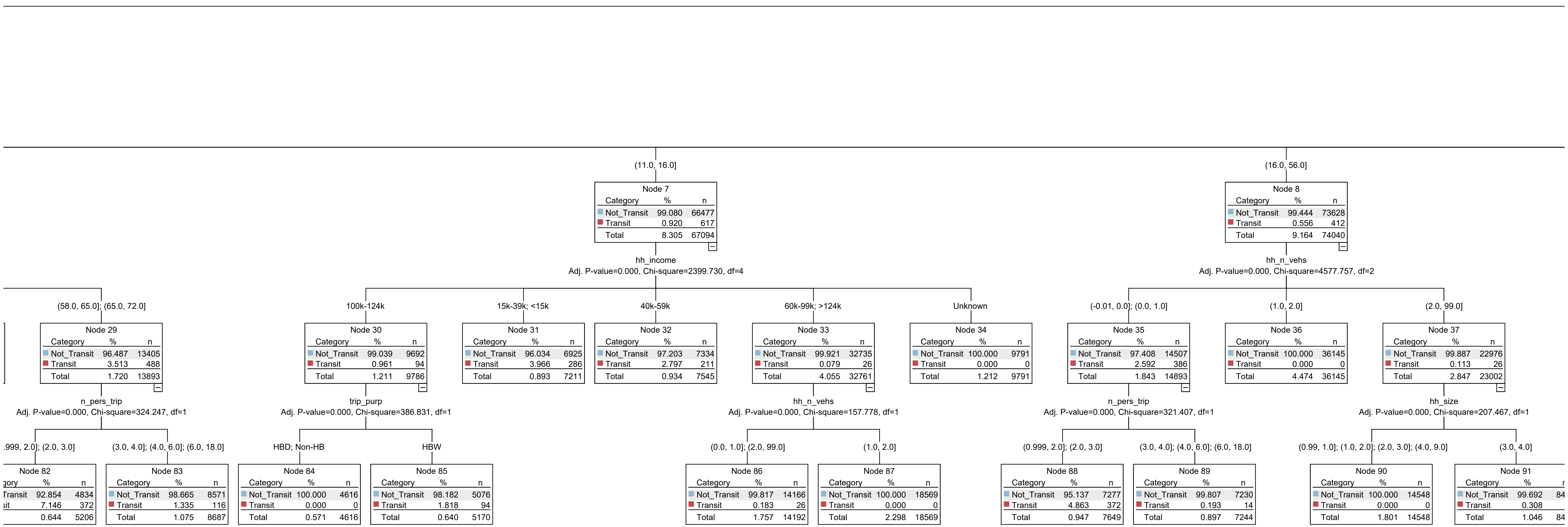
mode_prime


9
Node 1
Category % n
Not_Transit 87.197 4536
Transit 12.803 666
Total 0.644 5202
(-0.001, 1.0]
Node 5
Category % n
Not_Transit 72.717 85471
Transit 27.283 32069
Total 14.549 117540
trip_purp
Adj. P-value=0.000, Chi-square=29546.151, df=2
HBD
Node 20
Category % n
Not_Transit 87.034 52324
Transit 12.966 7795
Total 7.441 60119
age
Adj. P-value=0.000, Chi-square=2940.237, df=5
HBW
Node 21
Category % n
Not_Transit 38.364 13394
Transit 61.636 21519
Total 4.321 34913
free_park
Adj. P-value=0.000, Chi-square=4623.580, df=1
Non-HB
Node 22
Category % n
Not_Transit 87.760 19753
Transit 12.240 2755
Total 2.786 22508
age
Adj. P-value=0.000, Chi-square=2112.555, df=3
(10.999, 19.0]; (19.0, 28.0]; (41.0, 47.0]
Node 60
Category % n
Not_Transit 83.374 13961
Transit 16.626 2784
Total 2.073 16745
(28.0, 35.0]
Node 61
Category % n
Not_Transit 85.805 5936
Transit 14.195 982
Total 0.856 6918
(35.0, 41.0]
Node 62
Category % n
Not_Transit 74.386 7330
Transit 25.614 2524
Total 1.220 9854
(47.0, 53.0]; (53.0, 58.0]; (65.0, 72.0]
Node 63
Category % n
Not_Transit 95.857 14576
Transit 4.143 630
Total 1.882 15206
(58.0, 65.0]
Node 64
Category % n
Not_Transit 92.969 5183
Transit 7.031 392
Total 0.690 5575
(72.0, 99.0]
Node 65
Category % n
Not_Transit 91.702 5338
Transit 8.298 483
Total 0.720 5821
9; Y
Node 66
Category % n
Not_Transit 60.264 8311
Transit 39.736 5480
Total 1.707 13791
N; NA
Node 67
Category % n
Not_Transit 24.065 5083
Transit 75.935 16039
Total 2.614 21122
(10.999, 19.0]; (19.0, 28.0]; (28.0, 35.0]; (35.0, 41.0]
Node 68
Category % n
Not_Transit 74.578 6014
Transit 25.422 2050
Total 0.998 8064
(41.0, 47.0]; (53.0, 58.0]
Node 69
Category % n
Not_Transit 92.135 4756
Transit 7.865 406
Total 0.639 5162
(47.0, 53.0]; (65.0, 72.0]
Node 70
Category % n
Not_Transit 98.072 4374
Transit 1.928 86
Total 0.552 4460

mode_prime

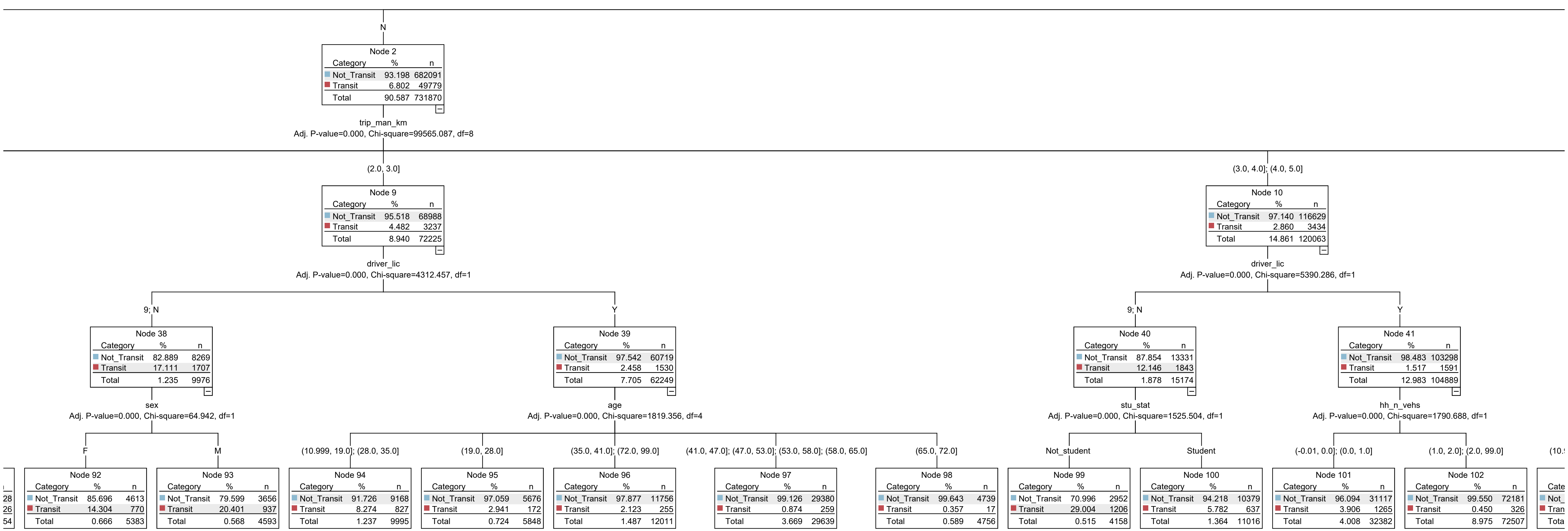

(1.0, 2.0]
Node 6
Category % n
Not_Transit 90.975 73626
Transit 9.025 7304
Total 10.017 80930
age
Adj. P-value=0.000, Chi-square=8723.027, df=6
(10.999, 19.0]
Node 23
Category % n
Not_Transit 70.095 8935
Transit 29.905 3812
Total 1.578 12747
hh_income
Adj. P-value=0.000, Chi-square=149.520, df=1
(19.0, 28.0]; (28.0, 35.0]
Node 24
Category % n
Not_Transit 90.125 13498
Transit 9.875 1479
Total 1.854 14977
trip_purp
Adj. P-value=0.000, Chi-square=1032.418, df=1
(35.0, 41.0]
Node 25
Category % n
Not_Transit 93.321 8817
Transit 6.679 631
Total 1.169 9448
trip_purp
Adj. P-value=0.000, Chi-square=601.952, df=1
(41.0, 47.0]; (72.0, 99.0]
Node 26
Category % n
Not_Transit 97.908 12450
Transit 2.092 266
Total 1.574 12716
hh_n_vehs
Adj. P-value=0.000, Chi-square=193.677, df=1
(47.0, 53.0]
Node 27
Category % n
Not_Transit 97.118 9875
Transit 2.882 293
Total 1.259 10168
n_pers_trip
Adj. P-value=0.000, Chi-square=1649.724, df=1
(53.0, 58.0]
Node 28
Category % n
Not_Transit 95.201 6646
Transit 4.799 335
Total 0.864 6981
(58.0, 65.0]; (72.0, 99.0]
Node 71
Category % n
Not_Transit 95.583 4609
Transit 4.417 213
Total 0.597 4822
100k-124k; Unknown
Node 72
Category % n
Not_Transit 76.380 3735
Transit 23.620 1155
Total 0.605 4890
15k-39k; 40k-59k; 60k-99k; <15k; >124k
Node 73
Category % n
Not_Transit 66.183 5200
Transit 33.817 2657
Total 0.973 7857
HBD; Non-HB
Node 74
Category % n
Not_Transit 95.488 9736
Transit 4.512 460
Total 1.262 10196
HBW
Node 75
Category % n
Not_Transit 78.686 3762
Transit 21.314 1019
Total 0.592 4781
HBD
Node 76
Category % n
Not_Transit 99.002 5160
Transit 0.998 52
Total 0.645 5212
HBW; Non-HB
Node 77
Category % n
Not_Transit 86.331 3657
Transit 13.669 579
Total 0.524 4236
(-0.01, 0.0]; (0.0, 1.0]
Node 78
Category % n
Not_Transit 96.004 5646
Transit 3.996 235
Total 0.728 5881
(1.0, 2.0]; (2.0, 99.0]
Node 79
Category % n
Not_Transit 99.546 6804
Transit 0.454 31
Total 0.846 6835
(0.999, 2.0]; (4.0, 6.0]
Node 80
Category % n
Not_Transit 93.552 4251
Transit 6.448 293
Total 0.562 4544
(2.0, 3.0]; (3.0, 4.0]; (6.0, 18.0]
Node 81
Category % n
Not_Transit 100.000 5624
Transit 0.000 0
Total 0.696 5624

mode_prime

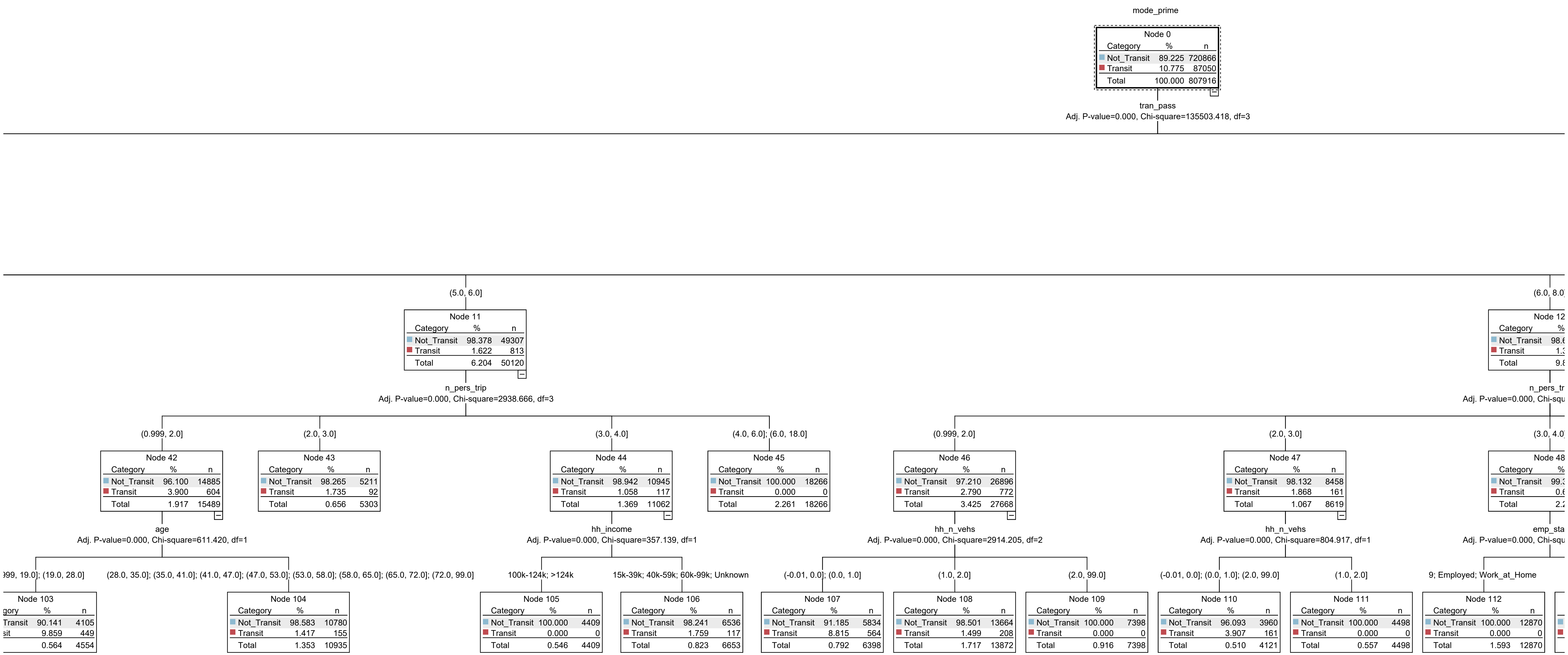


(11.0, 16.0]
Node 7
Category % n
Not_Transit 99.080 66477
Transit 0.920 617
Total 8.305 67094
hh_income
Adj. P-value=0.000, Chi-square=2399.730, df=4
(16.0, 56.0]
Node 8
Category % n
Not_Transit 99.444 73628
Transit 0.556 412
Total 9.164 74040
hh_n_vehs
Adj. P-value=0.000, Chi-square=4577.757, df=2
(58.0, 65.0]; (65.0, 72.0]
Node 29
Category % n
Not_Transit 96.487 13405
Transit 3.513 488
Total 1.720 13893
n_pers_trip
Adj. P-value=0.000, Chi-square=324.247, df=1
100k-124k
Node 30
Category % n
Not_Transit 99.039 9692
Transit 0.961 94
Total 1.211 9786
trip_purp
Adj. P-value=0.000, Chi-square=386.831, df=1
15k-39k; <15k
Node 31
Category % n
Not_Transit 96.034 6925
Transit 3.966 286
Total 0.893 7211
40k-59k
Node 32
Category % n
Not_Transit 97.203 7334
Transit 2.797 211
Total 0.934 7545
60k-99k; >124k
Node 33
Category % n
Not_Transit 99.921 32735
Transit 0.079 26
Total 4.055 32761
hh_n_vehs
Adj. P-value=0.000, Chi-square=157.778, df=1
Unknown
Node 34
Category % n
Not_Transit 100.000 9791
Transit 0.000 0
Total 1.212 9791
(-0.01, 0.0]; (0.0, 1.0]
Node 35
Category % n
Not_Transit 97.408 14507
Transit 2.592 386
Total 1.843 14893
n_pers_trip
Adj. P-value=0.000, Chi-square=321.407, df=1
(1.0, 2.0]
Node 36
Category % n
Not_Transit 100.000 36145
Transit 0.000 0
Total 4.474 36145
(2.0, 99.0]
Node 37
Category % n
Not_Transit 99.887 22976
Transit 0.113 26
Total 2.847 23002
hh_size
Adj. P-value=0.000, Chi-square=207.467, df=1
.999, 2.0]; (2.0, 3.0]
Node 82
% n
92.854 4834
7.146 372
0.644 5206
(3.0, 4.0]; (4.0, 6.0]; (6.0, 18.0]
Node 83
Category % n
Not_Transit 98.665 8571
Transit 1.335 116
Total 1.075 8687
HBD; Non-HB
Node 84
Category % n
Not_Transit 100.000 4616
Transit 0.000 0
Total 0.571 4616
HBW
Node 85
Category % n
Not_Transit 98.182 5076
Transit 1.818 94
Total 0.640 5170
(0.0, 1.0]; (2.0, 99.0]
Node 86
Category % n
Not_Transit 99.817 14166
Transit 0.183 26
Total 1.757 14192
(1.0, 2.0]
Node 87
Category % n
Not_Transit 100.000 18569
Transit 0.000 0
Total 2.298 18569
(0.999, 2.0]; (2.0, 3.0]
Node 88
Category % n
Not_Transit 95.137 7277
Transit 4.863 372
Total 0.947 7649
(3.0, 4.0]; (4.0, 6.0]; (6.0, 18.0]
Node 89
Category % n
Not_Transit 99.807 7230
Transit 0.193 14
Total 0.897 7244
(0.99, 1.0]; (1.0, 2.0]; (2.0, 3.0]; (4.0, 9.0]
Node 90
Category % n
Not_Transit 100.000 14548
Transit 0.000 0
Total 1.801 14548
(3.0, 4.0]
Node 91
Category %
Not_Transit 99.692
Transit 0.308
Total 1.046

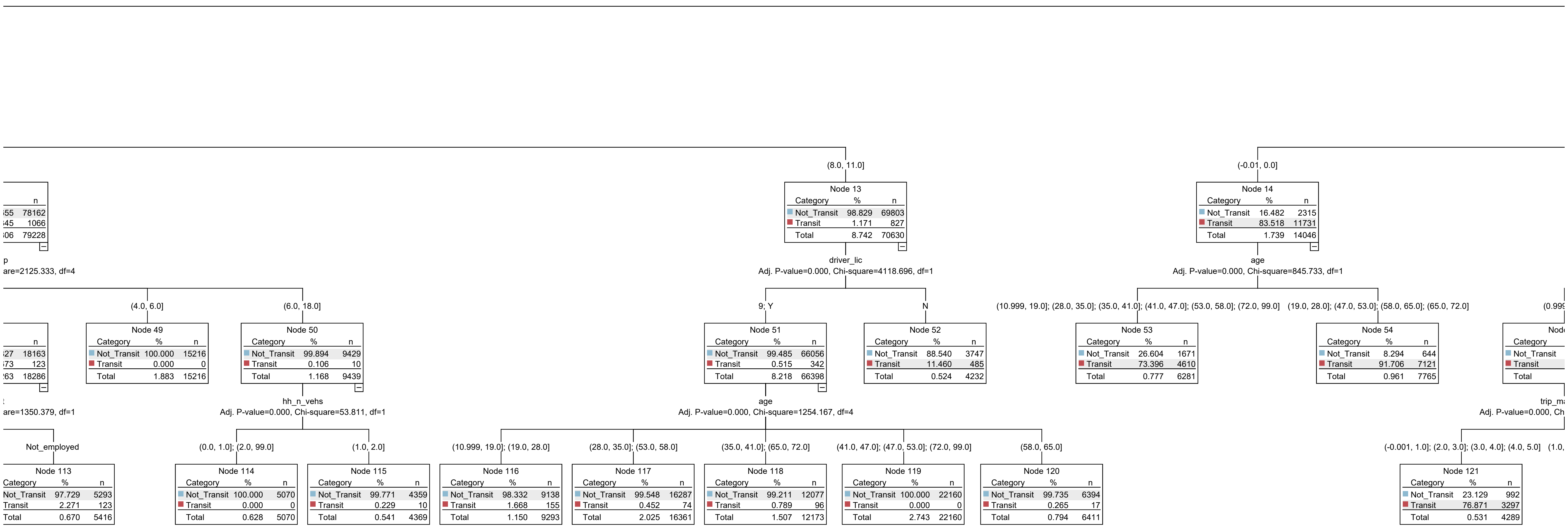
mode_prime


N
Node 2
Category % n
Not_Transit 93.198 682091
Transit 6.802 49779
Total 90.587 731870
trip_man_km
Adj. P-value=0.000, Chi-square=99565.087, df=8
(2.0, 3.0]
(3.0, 4.0]; (4.0, 5.0]
Node 9
Category % n
Not_Transit 95.518 68988
Transit 4.482 3237
Total 8.940 72225
driver_lic
Adj. P-value=0.000, Chi-square=4312.457, df=1
Node 10
Category % n
Not_Transit 97.140 116629
Transit 2.860 3434
Total 14.861 120063
driver_lic
Adj. P-value=0.000, Chi-square=5390.286, df=1
9; N
Y
9; N
Y
Node 38
Category % n
Not_Transit 82.889 8269
Transit 17.111 1707
Total 1.235 9976
sex
Adj. P-value=0.000, Chi-square=64.942, df=1
Node 39
Category % n
Not_Transit 97.542 60719
Transit 2.458 1530
Total 7.705 62249
age
Adj. P-value=0.000, Chi-square=1819.356, df=4
Node 40
Category % n
Not_Transit 87.854 13331
Transit 12.146 1843
Total 1.878 15174
stu_stat
Adj. P-value=0.000, Chi-square=1525.504, df=1
Node 41
Category % n
Not_Transit 98.483 103298
Transit 1.517 1591
Total 12.983 104889
hh_n_vehs
Adj. P-value=0.000, Chi-square=1790.688, df=1
F
M
(10.999, 19.0]; (28.0, 35.0]
(19.0, 28.0]
(35.0, 41.0]; (72.0, 99.0]
(41.0, 47.0]; (47.0, 53.0]; (53.0, 58.0]; (58.0, 65.0]
(65.0, 72.0]
Not_student
Student
(-0.01, 0.0]; (0.0, 1.0]
(1.0, 2.0]; (2.0, 99.0]
Node 92
Category % n
Not_Transit 85.696 4613
Transit 14.304 770
Total 0.666 5383
Node 93
Category % n
Not_Transit 79.599 3656
Transit 20.401 937
Total 0.568 4593
Node 94
Category % n
Not_Transit 91.726 9168
Transit 8.274 827
Total 1.237 9995
Node 95
Category % n
Not_Transit 97.059 5676
Transit 2.941 172
Total 0.724 5848
Node 96
Category % n
Not_Transit 97.877 11756
Transit 2.123 255
Total 1.487 12011
Node 97
Category % n
Not_Transit 99.126 29380
Transit 0.874 259
Total 3.669 29639
Node 98
Category % n
Not_Transit 99.643 4739
Transit 0.357 17
Total 0.589 4756
Node 99
Category % n
Not_Transit 70.996 2952
Transit 29.004 1206
Total 0.515 4158
Node 100
Category % n
Not_Transit 94.218 10379
Transit 5.782 637
Total 1.364 11016
Node 101
Category % n
Not_Transit 96.094 31117
Transit 3.906 1265
Total 4.008 32382
Node 102
Category % n
Not_Transit 99.550 72181
Transit 0.450 326
Total 8.975 72507

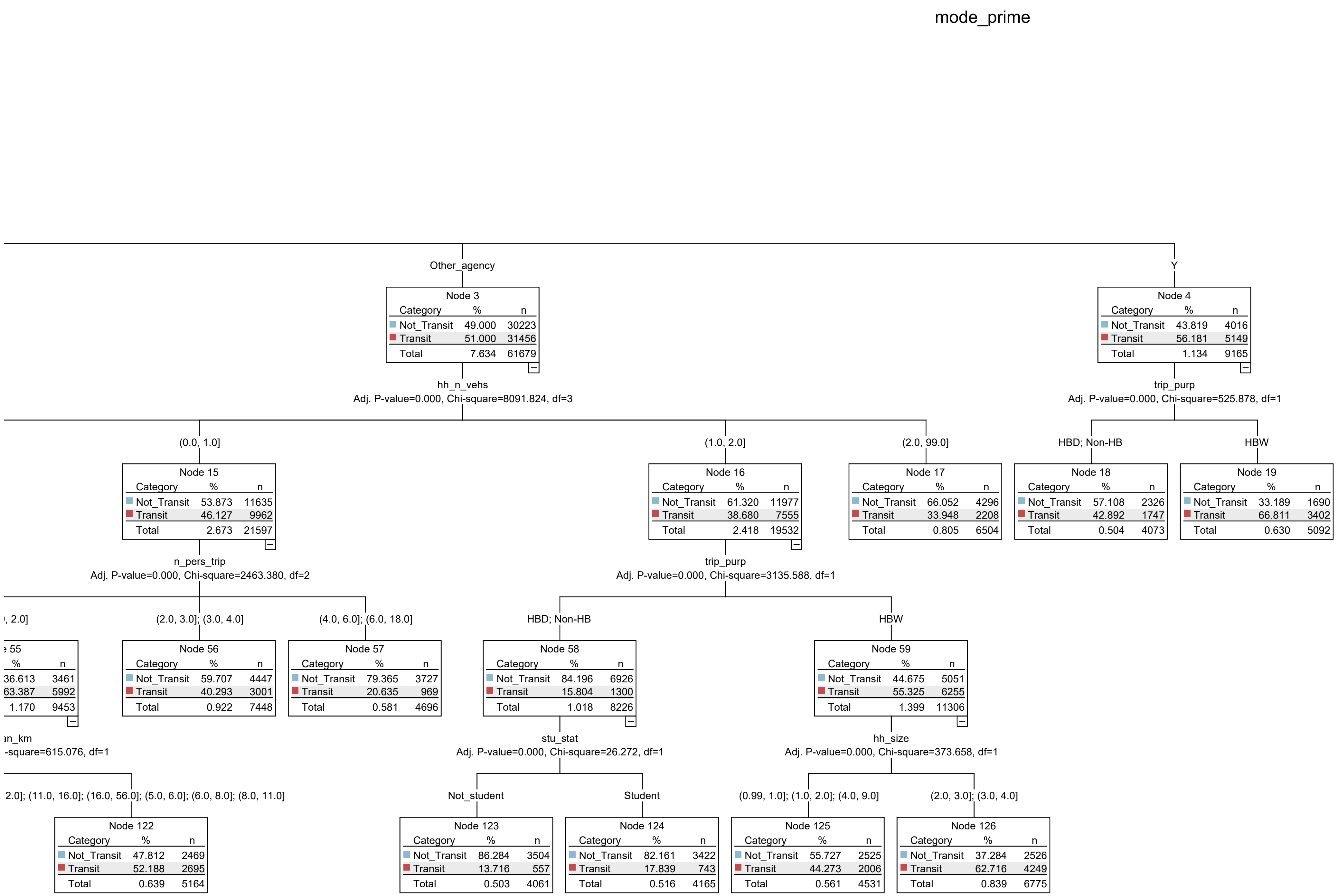
mode_prime


mode_prime
Node 0
Category % n
Not_Transit 89.225 720866
Transit 10.775 87050
Total 100.000 807916
tran_pass
Adj. P-value=0.000, Chi-square=135503.418, df=3
(5.0, 6.0]
Node 11
Category % n
Not_Transit 98.378 49307
Transit 1.622 813
Total 6.204 50120
n_pers_trip
Adj. P-value=0.000, Chi-square=2938.666, df=3
(0.999, 2.0]
Node 42
Category % n
Not_Transit 96.100 14885
Transit 3.900 604
Total 1.917 15489
age
Adj. P-value=0.000, Chi-square=611.420, df=1
(2.0, 3.0]
Node 43
Category % n
Not_Transit 98.265 5211
Transit 1.735 92
Total 0.656 5303
(3.0, 4.0]
Node 44
Category % n
Not_Transit 98.942 10945
Transit 1.058 117
Total 1.369 11062
hh_income
Adj. P-value=0.000, Chi-square=357.139, df=1
(4.0, 6.0]; (6.0, 18.0]
Node 45
Category % n
Not_Transit 100.000 18266
Transit 0.000 0
Total 2.261 18266
(0.999, 2.0]
Node 46
Category % n
Not_Transit 97.210 26896
Transit 2.790 772
Total 3.425 27668
hh_n_vehs
Adj. P-value=0.000, Chi-square=2914.205, df=2
(2.0, 3.0]
Node 47
Category % n
Not_Transit 98.132 8458
Transit 1.868 161
Total 1.067 8619
hh_n_vehs
Adj. P-value=0.000, Chi-square=804.917, df=1
Node 103
Total 0.564 4554
(28.0, 35.0]; (35.0, 41.0]; (41.0, 47.0]; (47.0, 53.0]; (53.0, 58.0]; (58.0, 65.0]; (65.0, 72.0]; (72.0, 99.0]
Node 104
Category % n
Not_Transit 98.583 10780
Transit 1.417 155
Total 1.353 10935
100k-124k; >124k
Node 105
Category % n
Not_Transit 100.000 4409
Transit 0.000 0
Total 0.546 4409
15k-39k; 40k-59k; 60k-99k; Unknown
Node 106
Category % n
Not_Transit 98.241 6536
Transit 1.759 117
Total 0.823 6653
(-0.01, 0.0]; (0.0, 1.0]
Node 107
Category % n
Not_Transit 91.185 5834
Transit 8.815 564
Total 0.792 6398
(1.0, 2.0]
Node 108
Category % n
Not_Transit 98.501 13664
Transit 1.499 208
Total 1.717 13872
(2.0, 99.0]
Node 109
Category % n
Not_Transit 100.000 7398
Transit 0.000 0
Total 0.916 7398
(-0.01, 0.0]; (0.0, 1.0]; (2.0, 99.0]
Node 110
Category % n
Not_Transit 96.093 3960
Transit 3.907 161
Total 0.510 4121
(1.0, 2.0]
Node 111
Category % n
Not_Transit 100.000 4498
Transit 0.000 0
Total 0.557 4498
Node 112
Category % n
Not_Transit 100.000 12870
Transit 0.000 0
Total 1.593 12870

mode_prime

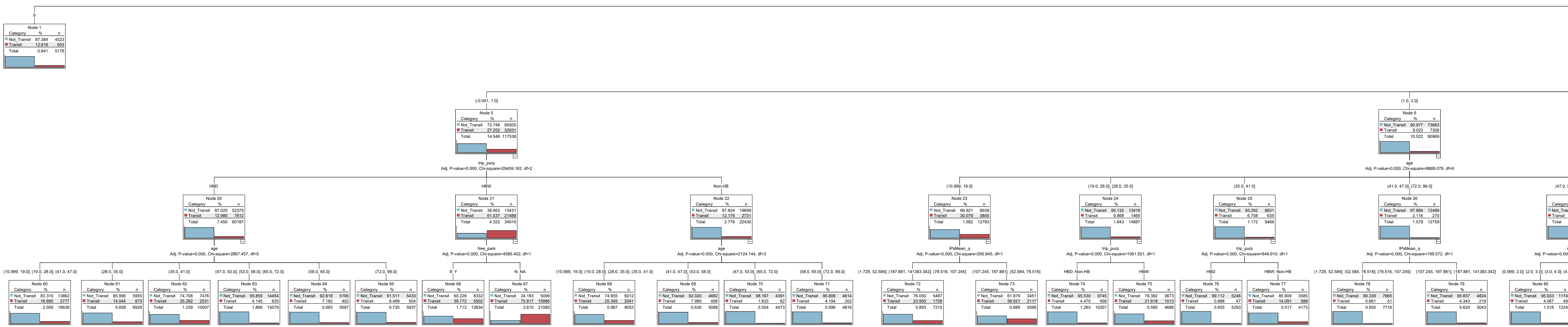


(8.0, 11.0]
Node 13
Category % n
Not_Transit 98.829 69803
Transit 1.171 827
Total 8.742 70630
driver_lic
Adj. P-value=0.000, Chi-square=4118.696, df=1
(-0.01, 0.0]
Node 14
Category % n
Not_Transit 16.482 2315
Transit 83.518 11731
Total 1.739 14046
age
Adj. P-value=0.000, Chi-square=845.733, df=1
are=2125.333, df=4
(4.0, 6.0]
Node 49
Category % n
Not_Transit 100.000 15216
Transit 0.000 0
Total 1.883 15216
(6.0, 18.0]
Node 50
Category % n
Not_Transit 99.894 9429
Transit 0.106 10
Total 1.168 9439
hh_n_vehs
Adj. P-value=0.000, Chi-square=53.811, df=1
9; Y
Node 51
Category % n
Not_Transit 99.485 66056
Transit 0.515 342
Total 8.218 66398
age
Adj. P-value=0.000, Chi-square=1254.167, df=4
N
Node 52
Category % n
Not_Transit 88.540 3747
Transit 11.460 485
Total 0.524 4232
(10.999, 19.0]; (28.0, 35.0]; (35.0, 41.0]; (41.0, 47.0]; (53.0, 58.0]; (72.0, 99.0]
Node 53
Category % n
Not_Transit 26.604 1671
Transit 73.396 4610
Total 0.777 6281
(19.0, 28.0]; (47.0, 53.0]; (58.0, 65.0]; (65.0, 72.0]
Node 54
Category % n
Not_Transit 8.294 644
Transit 91.706 7121
Total 0.961 7765
are=1350.379, df=1
Not_employed
Node 113
Category % n
Not_Transit 97.729 5293
Transit 2.271 123
Total 0.670 5416
(0.0, 1.0]; (2.0, 99.0]
Node 114
Category % n
Not_Transit 100.000 5070
Transit 0.000 0
Total 0.628 5070
(1.0, 2.0]
Node 115
Category % n
Not_Transit 99.771 4359
Transit 0.229 10
Total 0.541 4369
(10.999, 19.0]; (19.0, 28.0]
Node 116
Category % n
Not_Transit 98.332 9138
Transit 1.668 155
Total 1.150 9293
(28.0, 35.0]; (53.0, 58.0]
Node 117
Category % n
Not_Transit 99.548 16287
Transit 0.452 74
Total 2.025 16361
(35.0, 41.0]; (65.0, 72.0]
Node 118
Category % n
Not_Transit 99.211 12077
Transit 0.789 96
Total 1.507 12173
(41.0, 47.0]; (47.0, 53.0]; (72.0, 99.0]
Node 119
Category % n
Not_Transit 100.000 22160
Transit 0.000 0
Total 2.743 22160
(58.0, 65.0]
Node 120
Category % n
Not_Transit 99.735 6394
Transit 0.265 17
Total 0.794 6411
(-0.001, 1.0]; (2.0, 3.0]; (3.0, 4.0]; (4.0, 5.0]
Node 121
Category % n
Not_Transit 23.129 992
Transit 76.871 3297
Total 0.531 4289

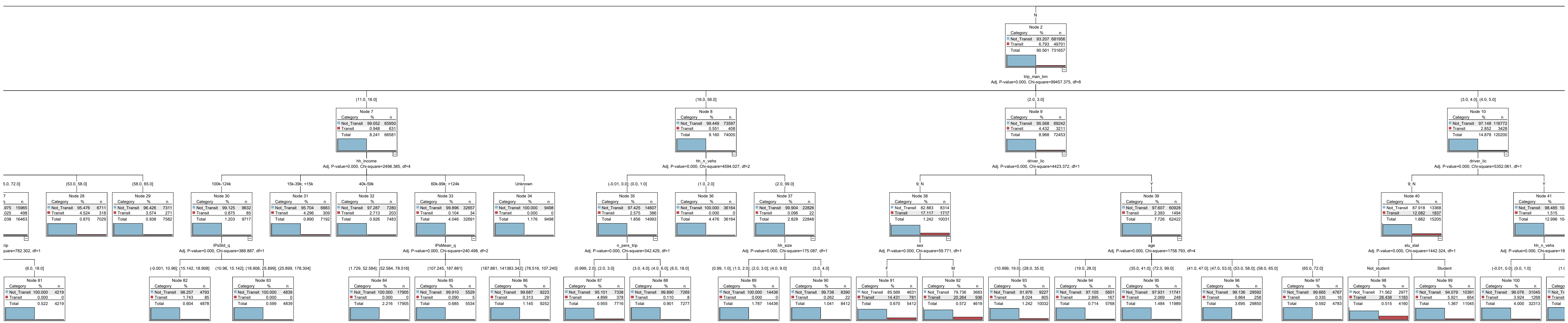

mode_prime
Other_agency
Y
Node 3
Category % n
Not_Transit 49.000 30223
Transit 51.000 31456
Total 7.634 61679
hh_n_vehs
Adj. P-value=0.000, Chi-square=8091.824, df=3
Node 4
Category % n
Not_Transit 43.819 4016
Transit 56.181 5149
Total 1.134 9165
trip_purp
Adj. P-value=0.000, Chi-square=525.878, df=1
(0.0, 1.0]
(1.0, 2.0]
(2.0, 99.0]
HBD; Non-HB
HBW
Node 15
Category % n
Not_Transit 53.873 11635
Transit 46.127 9962
Total 2.673 21597
Node 16
Category % n
Not_Transit 61.320 11977
Transit 38.680 7555
Total 2.418 19532
Node 17
Category % n
Not_Transit 66.052 4296
Transit 33.948 2208
Total 0.805 6504
Node 18
Category % n
Not_Transit 57.108 2326
Transit 42.892 1747
Total 0.504 4073
Node 19
Category % n
Not_Transit 33.189 1690
Transit 66.811 3402
Total 0.630 5092
n_pers_trip
Adj. P-value=0.000, Chi-square=2463.380, df=2
trip_purp
Adj. P-value=0.000, Chi-square=3135.588, df=1
, 2.0]
(2.0, 3.0]; (3.0, 4.0]
(4.0, 6.0]; (6.0, 18.0]
HBD; Non-HB
HBW
e 55
% n
36.613 3461
63.387 5992
1.170 9453
Node 56
Category % n
Not_Transit 59.707 4447
Transit 40.293 3001
Total 0.922 7448
Node 57
Category % n
Not_Transit 79.365 3727
Transit 20.635 969
Total 0.581 4696
Node 58
Category % n
Not_Transit 84.196 6926
Transit 15.804 1300
Total 1.018 8226
Node 59
Category % n
Not_Transit 44.675 5051
Transit 55.325 6255
Total 1.399 11306
an_km
-square=615.076, df=1
stu_stat
Adj. P-value=0.000, Chi-square=26.272, df=1
hh_size
Adj. P-value=0.000, Chi-square=373.658, df=1
2.0]; (11.0, 16.0]; (16.0, 56.0]; (5.0, 6.0]; (6.0, 8.0]; (8.0, 11.0]
Not_student
Student
(0.99, 1.0]; (1.0, 2.0]; (4.0, 9.0]
(2.0, 3.0]; (3.0, 4.0]
Node 122
Category % n
Not_Transit 47.812 2469
Transit 52.188 2695
Total 0.639 5164
Node 123
Category % n
Not_Transit 86.284 3504
Transit 13.716 557
Total 0.503 4061
Node 124
Category % n
Not_Transit 82.161 3422
Transit 17.839 743
Total 0.516 4165
Node 125
Category % n
Not_Transit 55.727 2525
Transit 44.273 2006
Total 0.561 4531
Node 126
Category % n
Not_Transit 37.284 2526
Transit 62.716 4249
Total 0.839 6775

# Appendix C: Trained Decision Tree Outputs (DTA), SPSS

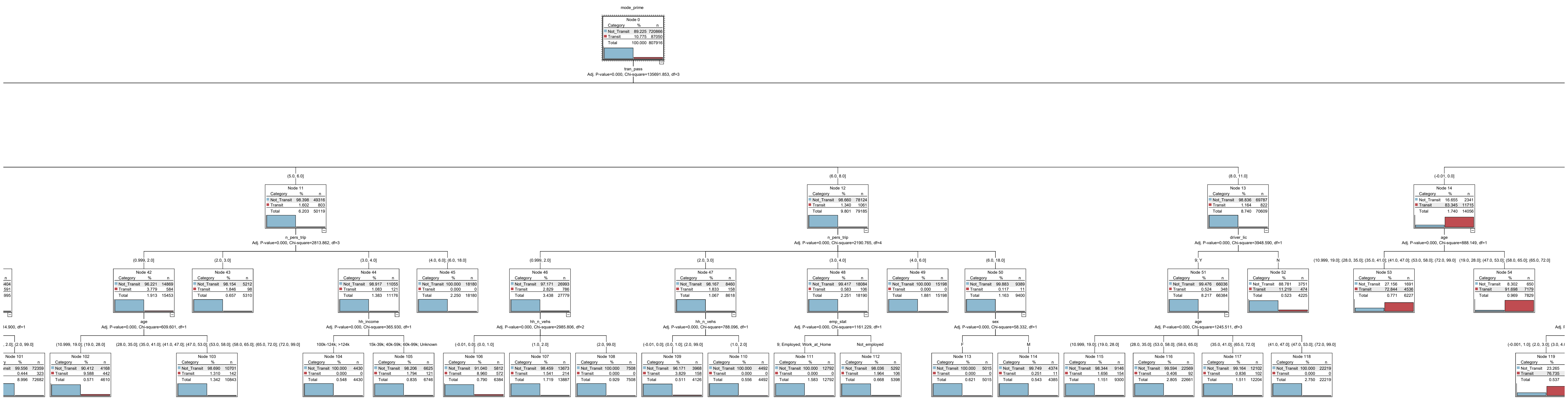
mode_prime



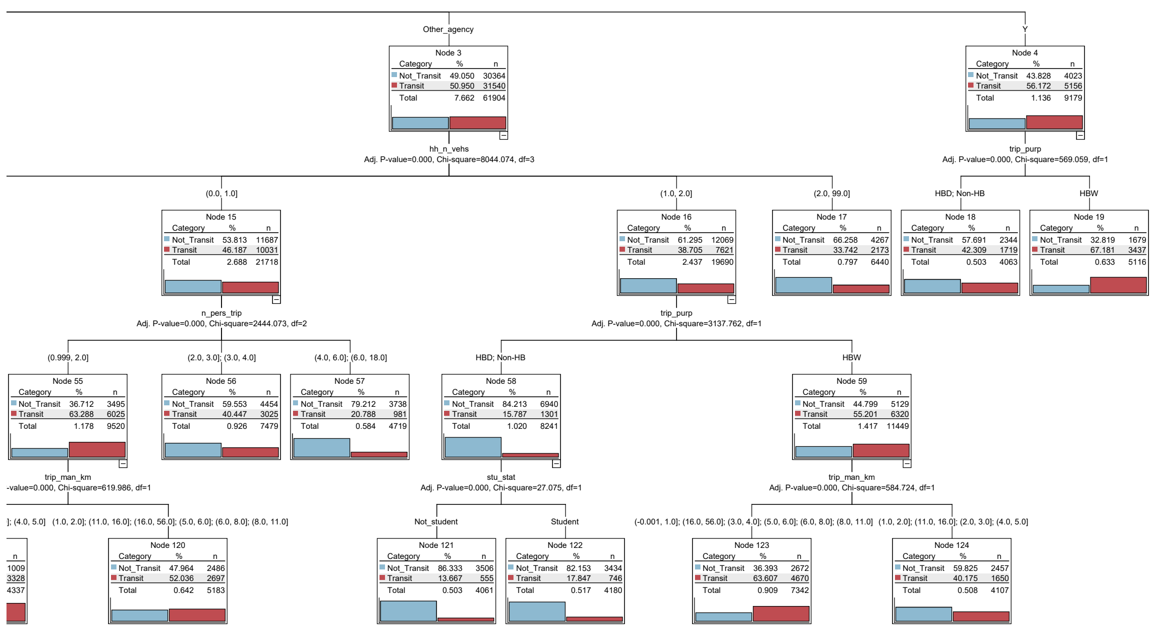
mode_prime



# mode_prime



mode_prime


Other_agency
Node 3
Category % n
Not_Transit 49.050 30364
Transit 50.950 31540
Total 7.662 61904
hh_n_vehs
Adj. P-value=0.000, Chi-square=8044.074, df=3
Y
Node 4
Category % n
Not_Transit 43.828 4023
Transit 56.172 5156
Total 1.136 9179
trip_purp
Adj. P-value=0.000, Chi-square=569.059, df=1
(0.0, 1.0]
Node 15
Category % n
Not_Transit 53.813 11687
Transit 46.187 10031
Total 2.688 21718
n_pers_trip
Adj. P-value=0.000, Chi-square=2444.073, df=2
(1.0, 2.0]
Node 16
Category % n
Not_Transit 61.295 12069
Transit 38.705 7621
Total 2.437 19690
trip_purp
Adj. P-value=0.000, Chi-square=3137.762, df=1
(2.0, 99.0]
Node 17
Category % n
Not_Transit 66.258 4267
Transit 33.742 2173
Total 0.797 6440
HBD; Non-HB
Node 18
Category % n
Not_Transit 57.691 2344
Transit 42.309 1719
Total 0.503 4063
HBW
Node 19
Category % n
Not_Transit 32.819 1679
Transit 67.181 3437
Total 0.633 5116
(0.999, 2.0]
Node 55
Category % n
Not_Transit 36.712 3495
Transit 63.288 6025
Total 1.178 9520
trip_man_km
-value=0.000, Chi-square=619.986, df=1
(2.0, 3.0]; (3.0, 4.0]
Node 56
Category % n
Not_Transit 59.553 4454
Transit 40.447 3025
Total 0.926 7479
(4.0, 6.0]; (6.0, 18.0]
Node 57
Category % n
Not_Transit 79.212 3738
Transit 20.788 981
Total 0.584 4719
HBD; Non-HB
Node 58
Category % n
Not_Transit 84.213 6940
Transit 15.787 1301
Total 1.020 8241
stu_stat
Adj. P-value=0.000, Chi-square=27.075, df=1
HBW
Node 59
Category % n
Not_Transit 44.799 5129
Transit 55.201 6320
Total 1.417 11449
trip_man_km
Adj. P-value=0.000, Chi-square=584.724, df=1
]; (4.0, 5.0]
n
1009
3328
4337
(1.0, 2.0]; (11.0, 16.0]; (16.0, 56.0]; (5.0, 6.0]; (6.0, 8.0]; (8.0, 11.0]
Node 120
Category % n
Not_Transit 47.964 2486
Transit 52.036 2697
Total 0.642 5183
Not_student
Node 121
Category % n
Not_Transit 86.333 3506
Transit 13.667 555
Total 0.503 4061
Student
Node 122
Category % n
Not_Transit 82.153 3434
Transit 17.847 746
Total 0.517 4180
(-0.001, 1.0]; (16.0, 56.0]; (3.0, 4.0]; (5.0, 6.0]; (6.0, 8.0]; (8.0, 11.0]
Node 123
Category % n
Not_Transit 36.393 2672
Transit 63.607 4670
Total 0.909 7342
(1.0, 2.0]; (11.0, 16.0]; (2.0, 3.0]; (4.0, 5.0]
Node 124
Category % n
Not_Transit 59.825 2457
Transit 40.175 1650
Total 0.508 4107

# Glossary

| | |
|---|---|
| Alternative (Class): | One mode of transportation among a set of modes generated by the user (choice/alternative set) within the decision process. |
| Decision attributes (Features): | Independent/explanatory variables in mode choice models, or variables presumed relevant to the individual making the choice. |
| Decision process: | A sequence of components associated with making a choice. |
| Decision rule: | Mechanism to assess decision attributes within decision structures, resulting in a choice. Decision rules influence the form of decision structures. |
| Decision structures: | The decision process framework; it defines the interaction of alternatives generation, decision attribute search, and decision rule mechanisms. |
| Decision Trees (DT): | Recursive partitioning algorithms that divide the predictor space (along feature dimensions) to improve node homogeneity. Includes the CHAID algorithm. |
| DTA dispersion: | The standard deviation of transit accessibility values over a period of time. |
| DTA magnitude: | The mean transit accessibility over a period of time. |
| DTA measures: | Dynamic Transit Accessibility (DTA) measures. Different forms of quantitatively representing time-series transit accessibility. Includes DTA dispersion and DTA magnitude. |
| Feature dimensions: | the unique categories or bins of values associated with each feature. |
| Transit accessibility: | a measure of the spatial distribution of activities, adjusted for transit travel impedance. |
| User: | an individual who chooses a mode for transportation. |