Machine-based Statistical Learning of Urban Metabolism

BY

DONGWOO LEE
B.S., University of Seoul, Korea, 2011
M.S., University of Seoul, Korea, 2013

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Civil engineering
in the Graduate college of the
University of Illinois at Chicago, 2019

Chicago, Illinois

Defense Committee:

Dr. Sybil Derrible, Chair and Advisor
Dr. Abolfazl Mohammadian
Dr. Bo Zou
Dr. Thomas L. Theis
Dr. Kazuya Kawamura, Urban Planning and Policy

To my grandmother, my parents, and my wife, Yeonjeong, I could not have done this without you. “Thank you, and I love you.”

ACKNOWLEDGEMENTS

I would like to whole-heartedly express my gratitude to my advisor and mentor, Professor Sybil Derrible, for the continuous support of my Ph.D. study, for his patience, motivation, and immense knowledge. I could not have imagined having a better advisor and mentor for the journey of my Ph.D. study.

Besides Prof. Derrible, I would like to thank my thesis committee, Prof. Abolfazl Mohammadian, Prof. Bo Zou, Prof. Thomas L. Theis, and Prof. Kazuya Kawamura, for their encouragement and support.

I must acknowledge Prof. Francisco Camara Pereira, who has introduced me to many new and fascinating concepts that are part of this thesis. I hope we can collaborate on exciting research in the future with Prof. Derrible.

I could not have done this without my parents. I deeply appreciate them for providing me with unfailing support and continuous encouragement throughout my academic journey.

Finally, I would like to whole-heartedly thank my wife. I cannot say any more words than "thank you, I love you". I would also like to leave a word to my future daughter, Yeondong, and to Emmy.

CONTRIBUTION OF AUTHORS

Chapter 2 represents a published article (Lee, D., Derrible, S., & Pereira, F., 2018 "Comparison of Four Types of Artificial Neural Networks and Multinomial Logit Model for Travel Mode Choice Modeling", Transportation Research Record, https://doi.org/10.1177/0361198118796971), written in collaboration with Sybil Derrible and Francisco Camara Pereira. The manuscript in chapter 3 (Lee and Derrible, "Predicting Residential Water Consumption: Modeling Techniques and Data Perspectives", Journal of Water Resources Management and Planning) is published, written with Sybil Derrible. The last manuscript in chapter 4 (Lee et al., Discrete Choice Modeling in Probabilistic Graphical Modeling Framework with Variational Inference) is under review, written with Sybil Derrible, Francisco Camara Pereira, and Filipe Rodrigues. Lastly, a published article (Lee, D., Mulrow, J., Haboucha, C. J., Derrible, S., & Shiftan, Y., 2019. "Attitudes on Autonomous Vehicle Adoption using Interpretable Gradient Boosting Machine." Transportation Research Record. https://doi.org/10.1177/0361198119857953) is partly presented in the Chapter 1.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Summary

This dissertation aims to gain a fundamental understanding of how machine-based statistical learning (ML) can contribute and be applied to the realm of urban metabolism. Along with substantial computational advances, a deluge of ML algorithms has been successfully applied to many domains—although most applications at the time of this writing have focused on the processing of images and sounds. In the realm of urban planning and urban engineering, ML approaches have often been viewed as "mystical." As acknowledged by many, the use of ML has received some reluctance in domains such as transportation due to their current limitations, often related to poor or lack of interpretability. In this sense, this dissertation fills knowledge gaps in the application of ML. In particular, the efforts are put into addressing interpretability (i.e., explainability), incorporating domain knowledge, and handling uncertainty to better capture patterns in resources use and human behaviors, providing more realistic conclusions in decision-making systems for urban metabolism. Furthermore, this dissertation also contributes to various domains in urban energy and resources consumption studies as these fields have not been exposed extensively to the realm of ML at the time of this writing.

Chapter 2 primarily focuses on investigating the capability and applicability of ML—that is, artificial neural networks (ANN) for the application of discrete choice models in the field of transportation. Four different types of ANN models are used to model, predict, and evaluate the behaviors, and compare modeling performances with traditional modeling methods such as MNL. These models are also used to interpret behavioral shifts based on sensitivity analysis. In particular, the series of analysis conducted in this chapter aimed to answer the question about how to learn and validate ML models for minimizing overfitting issues.

Summary (continued)

Chapter 3 aims to provide useful insights on several technical challenges, including a selection of modeling methods and data availability in modeling tasks, for the case of end-use water consumption. Specifically, 12 modeling methods grouped into two general categories—parametric models and non-parametric ML models—are adopted to model and predict household water use, based on two different data scenarios. The results reveal that the algorithmic properties of rule-based methods with boosting machines (i.e., gradient boosting machine (GBM)) are more suitable to analyze data that may include unobserved heterogeneity between users, partly thanks to their discriminative nature. Thus, this chapter provides useful technical insights into modeling techniques through a thorough review of modeling techniques and a technical discussion (see details in 3.4 and 3.7). The interpretability of ML techniques (e.g., GBM) is not discussed in this chapter, but it is studied in another article from the author of this dissertation—not included in this dissertation—that aimed to investigate attitudes and behaviors toward autonomous vehicles by using boosting machine. In general, ML models are interpreted based on the model-agnostic or model-specific method—e.g., feature importance, partial dependence.

Chapter 4 adopts different ways to enhance the interpretability of ML models not only to incorporate domain knowledge but also to handle uncertainties in modeling tasks. Specifically, the key concept of this chapter is to separate knowledge, model, and inference (e.g., probabilistic reasoning) in constructing decision-making systems (a.k.a., reasoning systems).

Summary (continued)

Toward this goal, this chapter uses probabilistic graphical models (PGM) with Bayesian inference (PGM-B) that coherently manipulate and quantify uncertainties through the use of probability theory and graph theory. The PGM-B can be adapted to any modeling technique to infer the full distributions of interest based on the PGM used. Specifically, variational inference (VI) is used to approximate probability densities, based on ML algorithms. To investigate applicability, this chapter also derives a way to develop a PGM-B to address travel mode choice behaviors based on several assumptions and under different specifications (i.e., level of pooling). In particular, these frameworks are derived to capture unobserved heterogeneity and quantify uncertainty by inferring the full posterior distributions. Prediction performances are also validated and compared with existing random utility models (RUM).

The last chapter concludes this dissertation and offers some future work directions. Overall, the chapters in this dissertation fill important theoretical and technical knowledge gaps in the realm of urban metabolism and urban modeling in the era of Artificial Intelligence (AI). In particular, the chapters focused not only on providing ways to take advantage of ML approaches, but also to address some of their limitations—e.g., interpretability, uncertainties. The principles and practical applications of Data Science can be further used to develop novel and creative approaches to gain a fundamental understanding of many other characteristics of cities that are not related to urban metabolism. Simultaneously, the ability to understand these other characteristics can shed lights into making cities more sustainable and resilient.

1 Introduction

"A set of things working together as part of a mechanism or an interconnecting network; a complex whole."

(Definition of System, Oxford Dictionary)

The term "system" is commonly used in the research realm on cities or urban areas. The definition of a system from the Oxford Dictionary contains two key terms for the research presented in this dissertation: an *interconnecting network* and a *complex whole*. Broadly, a city is treated here as a complex system, as the collection of interconnected components (i.e., parts) that are fundamentally utilized by people living in this system. In particular, examples of interconnected components include infrastructure (e.g., transport and water networks; parks; buildings) as well as resources (e.g., electricity, water, fuel, vehicles, electronic devices) that make use of infrastructure. In addtion to these components, human behaviors associated with the interconnected components are a key aspect of this dissertation.

To illustrate this point, smartphones have become ubiquitous and they even guide our lives in many respects such as by recommending where and how to make a trip. In fact, they require interconnected infrastructure systems that integrate telecommunication infrastructure with the power grid that themselves depend on other infrastructure systems to generate electricity such as transportation and water systems. Moreover, transportation also relies on the water system, since contingencies on water systems such as floods will make it impossible to use it. As seen in these examples, an urban system is formed by multiple interdependent infrastructure systems that often function simultaneously (Derrible 2019).

As a result, the focus of studying infrastructure includes the study of flows of resources in cities that characterize the functioning infrastructure and that are fundamentally associated with human behavior. Providing the necessary resources and energy for a growing population that also has a growing appetite for resources and energy represents a major challenge for today's engineers, planners, and policymakers. This is particularly relevant as cities are already the sites of tremendous flows of energy and materials (Beloin-Saint-Pierre et al. 2017). In addition, cities are continuously evolving, and over the half of the world's population now resides in cities and that number is expected to be increased to around 70% by 2050 ("2018 Revision of World Urbanization Prospects" 2018). Naturally, the production and consumption of resources by people to meet their own needs and desires is also interrelated and interconnected.

1.1 Urban Metabolism: Resources Consumption Behaviors and Flows

1.1.1 Definition of Urban Metabolism

The concept of urban metabolism has inspired new ways of thinking to understand how materials and energy are used to meet peoples' needs. Wolman (1965) conceived the metabolism of cities assessing the environmental impacts of urban development (Wolman 1965). In particular, Wolman estimated the footprints of interconnected urban infrastructure systems by quantifying the overall flow of resources and wastes for a hypothetical urban area of one million people. More recently and formally, Kennedy et al. (2007) define urban metabolism as "the sum total of the technical and socio-economic processes that occur in cities, resulting in growth, production of energy, and elimination of waste." In particular, the definition presented by Kennedy et al. (2007) encompasses most of the components and flows in urban systems that can be analyzed in urban studies. Thus, more practically, this definition suggests that urban metabolism can be seen as a series of models for individual or multiple components of cities to explore resource consumption behaviors and their flows in urban systems.

1.1.2 Interdisciplinary Perspectives of Urban Metabolism and Sustainability

Many studies have analyzed the interdisciplinary nature of urban metabolism (Barles 2010; Beloin-Saint-Pierre et al. 2017; Broto et al. 2012; Kennedy et al. 2011). Partly because they are from different fields (e.g., industrial ecology, economics), these studies used diverse modeling methods to measure various components of urban metabolism. Generally, studies disaggregate the entire urban systems into separate components (i.e., infrastructure; resources supply and demand) and develop individual models to address single or multiple issues across disciplines (Kennedy et al. 2011). Despite the differences in disciplines, they share common goals such as minimizing resources consumption that partly depends on human behavior.

The field of industrial ecology mostly focuses on describing and evaluating material resources flows and related environmental impacts (e.g., GHG emissions) on large systems, called Material Flow Analysis (MFA), later expanded to Material and Energy Flow Analysis (MEFA). MEFA is used for the integrated assessment of mass and energy flows in urban systems (Beloin-Saint-Pierre et al. 2017; Kennedy et al. 2011). From an analogous, although different, perspective, numerous analyses were conducted in economics (e.g., environmental, political, urban economics), using loosely similar approaches to industrial ecology, to explore the relationships between human behavior and related impacts on urban systems (e.g., negative externalities). The purpose of these studies is often virtually identical to the field of urban metabolism, such as to minimize social costs (e.g., resources consumption). Nonetheless, and using economic theory, these studies are generally not classified as urban metabolism studies (Beloin-Saint-Pierre et al. 2017). This is in part because they used different terminologies and perspectives while pushing themselves into their own boundaries of the economic realm (Broto et al. 2012).

By and large, and regardless of the discipline and theoretical background, urban metabolism studies tend to share common concerns, and among these concerns, resource consumption often comes first.

1.1.3 Challenges in Urban Resources Consumption Modeling

Accurately modeling resource consumption behaviors is key to determining how infrastructure systems are used and how they can perform better. This is particularly the case now as cities are expanding and as urban systems tend to be highly interdependent (Derrible 2016b). In addition, we are living in the era of Big Data, and enormous amounts of data are generated from virtually everything everywhere. When it comes to cities, millions of data points about most urban systems are simultaneously collected and stored. This collected information is often used to guide our everyday decisions, and it is also utilized to predict future activities. For example, smart metering devices for urban infrastructure resources (e.g., water, electricity, fuel) can not only provide information about resource consumption behaviors in real time, but they also suggest and guide users on how to consume less.

Nevertheless, this deluge of data is useless until relevant knowledge is extracted or unknown quantities are inferred (Robert 2014a).

1.1.4 Dilemmas in statistical learning

The field of statistical learning—that includes machine learning (ML)—has often focused on supplementing and tailoring algorithms for a certain problem or phenomenon to output accurate predictions. In general, there is a trade-off between modeling performance (e.g., prediction accuracy) and interpretability when selecting a learning algorithm and an approach. Some models such as parametric models are often considered more easily interpretable than nonparametric models that use complex algorithms (a.k.a., black box). In predicting and modeling discrete choice problems, for instance, parametric approaches (e.g., the family of logit models) have been predominantly used since they are intuitive and more easily interpretable based on strong theoretical backgrounds (e.g., random utility theory). Specifically, these approaches yield parameters that may be used to evaluate the impact of policies, economic changes, and technological adaptation. Nonetheless, the predetermined assumptions in traditional parametric models (e.g., conditional logit) can lead to biased estimations and misleading predictions. Moreover, interactions between the explanatory variables (e.g., the nonlinearity of attributes) are often neglected since it is difficult to represent them in a linear function (e.g., conditional and threshold effects). One way to take this issue into account is by introducing additional variables (e.g., polynomial and interaction terms) either by considering all possible combinations of explanatory variables or by measuring empirical relationships. Nonetheless, it is practically impossible to identify all the interactions between all the variables as well as to identify the necessary variables. Although flexible modeling methods (e.g., mixed logit) can have better modeling performances than others by relaxing IIA assumptions, predetermined structures and assumed linearity in the underlying functions still make it difficult to capture or infer high degrees of nonlinearity in a dataset.

Over the last decade, in particular, the application of ML has rapidly increased and proven to be successful in many real-world applications. This success was partly driven by the vast expansion of computational means, by the development of improved ML algorithms (e.g., deep neural network), and by the availability of extensive information (a.k.a., Big Data) (Bishop 2013). Besides, the convergence of multi-disciplinary efforts in the realm of modeling has contributed to enhancing the level performance and applicability of ML, and show relatively high performance to analyze a wide variety of modeling tasks compared to parametric approaches (Lee et al. 2018b; Wang and Ross 2018). In part, it is due to the fact that ML models possess fewer predetermined assumptions than the parametric models while adopting complex algorithms and myriads of sub-models (e.g., local nonparametric estimators) thanks to significant advancement in computational ability. Despite a high degree of predictive power, the general criticism about many ML techniques is their lack of interpretability due to their reliance on machine-based repetitive computation. This interpretable issue makes it difficult to incorporate domain knowledge. Future approaches should try to possess the advantages of both modeling approaches, especially for interpretability and prediction performances in parametric and ML model, respectively.

1.2 Machine-based Statistical Learning

1.2.1 Machine-based Statistical Learning

Due to the interdisciplinary nature and breadth of the general realm of machine learning, this dissertation purposedly uses machine-based statistical learning—the same acronym ML is used. Broady, statistical learning refers to a set of statistical tools for modeling and understanding data. Statistical learning can be classified into two main categories of learning: supervised and unsupervised. Supervised statistical learning involves developing statistical models to infer or predict output(s) based on a set of inputs. In contrast, unsupervised statistical learning is used to explore relationships and structure from data, without supervising output(s). In addition to these main categories, four types of learning methods exist that combine supervised and unsupervised learning features: (1) semi-supervised learning, (2) active learning, (3) reinforced learning, and (4) transfer learning. The first two learning methods can be applied to a particular case in which possible output values are limited or missing. Reinforced learning is also called goal-oriented learning, which output values are given in the form of reward for a set of actions in an environment. Lastly, transfer learning refers to a model developed for a certain problem (i.e., either the same or different domain) that can be used as a starting point for a model on a different by the dependent problem. Recently, these statistical learning methods have increasingly been adopted in many disciplines thanks to the addition of machine-based automation; i.e., ML.

ML is a set of self-adaptive methods that blend probabilistic and nonparametric features, that can automatically capture patterns in data, and that then use hidden machines (e.g., hidden layers in artificial neural networks) to infer or predict unknown quantities (Bishop 2006a; Robert 2014a). In particular, probabilistic features are used to explain uncertainty in modeling (Robert 2014a). For instance, it is possible that a statistical learning model returns a prediction for some unseen information that may lie outside of given data distribution. Naturally, this is an unreasonable prediction, which then becomes an error in the model. Nonetheless, the model may provide a different prediction while having the same information. This is but one example of *uncertainty* in modeling, and there are other forms of uncertainty, including: (1) noise in the observed data (e.g., measurement or sampling errors), (2) model parameters (e.g., the coefficient of regression models), and (3) model selection (e.g., modeling techniques or structure of model). As datasets about urban systems become more complex, and combined with human behaviors, addressing uncertainty issues in statistical modeling becomes fundamental.

In addition to probabilistic features, nonparametric features in ML are used to infer unknown information while making as few predetermined assumptions as possible (Friedman et al. 2001; Ghahramani 2015; Robert 2014a). Put differently, nonparametric approaches can be inherently adapted to the data without assuming the parametric function form such as $E(Y|X)$ is linear in the inputs (X) to predict the output (Y) or that the linear model reasonably fits along with a flat hyperplane (Hastie et al. 2009; James et al. 2013; Kuhn and Johnson 2013). This also means that ML models use infinite-dimension to infer unknown information (Friedman et al. 2001). For instance, ML models including nonparametric features (e.g., kernel density estimation) perform stochastic local optimization rather than single global optimization. Thus, they are likely to decrease biases while balancing variances (i.e., trade-off bias and variance), and they can provide more accurate predictions than traditional statistical models.

1.2.2 Limitation of Machine-based Statistical Learning

While the most relevant property of statistical modeling is to produce accurate predictions, the second most relevant may be how the model generates results. In other words, how *interpretable* a model is. When models adopt a more complex structure, their structures become less interpretable. This is particularly relevant for ML models that often face this issue because of their complex and "unknown" structures. For instance, deep neural networks (a.k.a deep learning) contain a set of automated features (e.g., hidden layers) that aim to capture and infer unknown quantities.

Interpretability refers to "the degree of which human can understand the cause of a decision" (Doshi-Velez and Kim 2017a; Miller 2017). Specifically, stating that a model has better interpretability than another model means that humans can more easily comprehend why certain decisions (inference or prediction) were made. That being said, a model does not need to be interpretable to make accurate predictions; however, only an accurate prediction may not be enough information that solves the given problem (Doshi-Velez and Kim 2017a; Miller 2017). In other words, predictions produced by a model should inform as much information (e.g., through causal latent variables) as possible to solve real-world problems, and it is fundamental to other important goals in the realm of statistical learning such as unbiasedness, reliability, causality, and usability (Doshi-Velez and Kim 2017a; Kim et al. 2014).

As for the application of ML to address real-world problems, researchers typically try to select a suitable method among existing ML techniques and map their problem onto it—often influenced by their knowledge and familiarity with a specific method. The selected method often requires some kind of modifications that correspondingly require in-depth technical understanding of the algorithm and its application. In this sense, methodological concepts and techniques that involve these tailored algorithms often become more complicated than already existing models, and researchers are left with a deluge of modeling algorithms. As mentioned earlier, the era of Big Data is creating unprecedented opportunities for researchers—especially in the ML community—to exploit the power of data-driven approaches, but few are able to apply and adapt ML algorithms properly. In the application of discrete choice models (DCM) in the field of transport, for instance, rarely do the models and inference techniques for DCM exactly correspond to some existing ML techniques. In addition, despite having been successful at its particular task, built models are often not applicable to other use cases. As a result, if a problem and a corresponding application change, the built model will have poor accuracy and it will have to be substantially modified.

1.3 Objectives

The main goal of this dissertation is to gain a fundamental understanding of how machine-based statistical learning (ML) can contribute and be applied to urban metabolism. As a by-product, this dissertation fills knowledge gaps in the application of ML. In particular, the efforts are put into addressing interpretability (i.e., explainability), incorporating domain knowledge, and handling uncertainty to better capture patterns in resources use and human behaviors, providing more realistic conclusions in decision-making systems for urban metabolism. Furthermore, this dissertation also contributes to various domains in urban behavior studies as these fields have not been exposed extensively to the realm of ML at the time of this writing. More specifically, the objectives of this dissertation are to:

- Investigate the **capability and applicability of ML methods** such as **ANNs** to predict mode choice behaviors and compare modeling performances with traditional modeling methods such as MNL, and also interpret behavioral shifts based on sensitivity analysis,
- Address **two technical dilemmas** of modeling resource use and behavior: (1) **data availability** and (2) **selection of modeling methods**, and provide useful technical insights into modeling techniques,
- Suggest a **modular probabilistic ML modeling framework** that can be **adapted with any algorithms** and enhance the **interpretability** of ML models not only to **incorporate domain knowledge** but also **handle uncertainties** in modeling tasks.

These objectives are achieved sequentially in chapters 2, 3, and 4 that are summarized below.

Chapter 2 investigates the capability and applicability of ML models to the field of transportation. Specifically, discrete mode choice modeling is a fundamental part of travel demand forecasting. To date, this field has been dominated by parametric approaches (e.g., logit models), but non-parametric approaches such as Artificial Neural Networks (ANN) possess much potential since choice problems can be assimilated to pattern recognition problems. In particular, ANN models are easily applicable with the higher capability to identify nonlinear relationships between inputs and designated outputs to predict choice behaviors. This chapter investigates the capability of four types of ANN models and compares their prediction performances with a conventional multinomial logit model (MNL) for mode choice problems. The four ANN models are: Backpropagation Neural Network (BPNN), Radial Basis Function Networks (RBFN), Probabilistic Neural Network (PNN), and Clustered PNN. To compare the modeling techniques, we present algorithmic differences of each ANN technique, and we assess their prediction accuracy with 10-fold cross-validation method. As for interpreting ANN models, we evaluate the contribution of explanatory variables by conducting sensitivity analyses on significant variables.

Chapter 3 aims to explore technical challenges when modeling resource consumption behaviors, with a focus on residential water consumption. Specifically, predicting residential water demand is challenging because of two technical questions: (1) which data and variables should be used and (2) which modeling technique is most appropriate for high prediction accuracy. To address these issues, this chapter investigates twelve statistical techniques, including parametric models and machine learning (ML) models, to predict daily household water use. In addition, two data scenarios are adopted: (1) one with only six variables, generally available to cities and water utilities (general scenario), and (2) one with all 19 variables available from the Residential End-Use 2016 database (REU 2016). The results for REU 2016 indicate that ML models outperform linear models. In particular, gradient boosting regression (GBR) regression performs best with an R_{adj}^2 of 0.69 compared to 0.54 for linear regression. The performance gap between ML and linear models becomes even wider for the general scenario with an R_{adj}^2 of 0.60 for GBR compared to 0.33 for linear regression. The finding in this chapter can be useful to researchers, municipalities, and utilities who are seeking novel modeling techniques that can provide consistent modeling performances—i.e., high prediction accuracy—depending on data availability. Future work could include the development of new measures to increase the interpretability of ML models to better understand causal relationships between the independent variables and daily household water use.

Chapter 4 provides a flexible probabilistic ML modeling framework for DCM in the field of transportation. Toward this goal, this chapter uses the concept of probabilistic graphical models (PGM) and Bayesian inference (PGM-B) that coherently tackle uncertainty issues through the use of probability theory. The framework can be adapted to any ML algorithm and can infer the full distributions of the model's parameters. In particular, PGM-B can be separated into three sub-processes: representation, modeling, and inference. Modeling tasks begin with considering all kinds of quantities governing the problem, and these are treated as random variables. Complex relationships between the random variables are intuitively and compactly represented as a graphical structure within the generative process of the algorithm. Once the PGM framework is constructed, the goal is to infer the full posterior distributions of interest through Bayesian inference. Specifically, this chapter uses variational inference (VI) that leverages techniques from ML to approximate probability densities. To investigate the applicability of PGM-B, this chapter mainly derives a way to develop a PGM-B to address travel mode choice behaviors. In particular, three different PGM-B frameworks are derived to represent mode choice behaviors based on our assumptions and under different specifications (i.e., level of pooling), which can capture unobserved heterogeneity and quantify uncertainty by inferring the full posterior distributions. In addition, prediction performances are validated and compared with existing random utility models (RUM).

The last chapter, “Conclusion and Future Work,” proposes future research opportunities for machine-based statistical learning of urban metabolism, which leverages unprecedented opportunities brought by the era of Big Data and Data Science.

2 Comparison of Four Types of Artificial Neural Networks and a Multinomial Logit Model for Travel Mode Choice Modeling

(previously published as Lee, D., Derrible, S., & Pereira, F., 2018 “Comparison of Four Types of Artificial Neural Networks and Multinomial Logit Model for Travel Mode Choice Modeling”, Transportation Research Record, doi.org/10.1177/0361198118796971)

2.1 INTRODUCTION

In travel mode choice modeling, and based on a set of given attributes, the objective is to model the choice processes and behaviors of travelers faced with several travel mode alternatives. Being able to model mode choice and predict future trends is critical to transportation planners and policy makers. Travel behaviors are typically examined by using statistical survey techniques such as Reveal Preference (RP) and Stated Preference (SP), in which respondents are asked to present their choices and preferable scenarios, respectively (De Carvalho et al. 1998). The surveys yield discrete choice data that can then be used to calibrate travel mode choice models that are fundamental components of disaggregate travel demand modeling approaches, e.g., for activity-based modeling (ABM).

As the main parametric modeling approach, random utility modeling has been the dominant technique used in the literature since the 1980s to investigate parametric relationships between mode choice and its possible determinants. In particular, logit models—one of the random utility models—gained popularity in the travel mode choice realm because they are based on simple mathematical formulations while accounting for unobserved variables (i.e., stochasticity). Logit models (e.g., multinomial logit (MNL)), however, assume that each choice is independent and identically distributed (IID), which can lead to biased estimations and misleading predictions. Moreover, interactions between the explanatory variables (e.g., the nonlinearity of attributes) are often neglected since it is difficult to represent them in the utility function (e.g., conditional and threshold effects). One way to take this issue into account in the linear utility function is by introducing additional variables (e.g., polynomial and interaction terms) either through considering all possible combinations of explanatory variables or by measuring empirical relationships (Bentz and Merunka 2000). Nonetheless, it is practically impossible to identify all the interactions between all the variables as well as to identify the necessary variables. MNL tries to overcome this issue by using domain knowledge as much as possible. Although the flexible Random Utility Model (RUM) methods such as mixed logit (Bhat 2001; Shen 2009; Train 2009) can have better modeling performance than MNL by relaxing IIA assumptions, predetermined structures and linear characteristics of underlying functions still make it difficult to capture or infer high degrees of nonlinearity in a dataset (James et al. 2013; Kuhn and Johnson 2013; Train 2009).

In contrast, algorithmic non-parametric approaches have proven to be incredibly powerful to study urban systems (Ahmad et al. 2016, 2017; Ahmad and Derrible 2015a; Cottrill and Derrible 2015; Derrible 2016a; b; Derrible and Ahmad 2015; Karduni et al. 2016). Specifically, Artificial Neural Networks (ANN) present higher adaptability in identifying nonlinear interactions (Bengio and Bengio 2000; Wong et al. 2017). Contrary to the logit model that deals with nonlinearity issue by reducing the “complexity” of the dataset, ANN can capture nonlinear properties more easily through additional units (e.g., hidden layers) without assuming predetermined functional forms (De Carvalho et al. 1998; Dougherty 1995; Sayed and Razavi 2000; Xie et al. 2003). Nonetheless, a commonly acknowledged disadvantage of ANN is their lack of interpretability, and they are often associated to black boxes. Another disadvantage is their relative incapability to use previously acquired knowledge (i.e., domain knowledge). Despite these challenges, ANN tends to outperform logit models (from the multinomial to nested logit models (Hensher and Ton 2000; Mohammadian and Miller 2002; Xie et al. 2003). Among different types of ANN models, several studies have used backpropagation neural networks (BPNN) to model mode choice (Golshani et al. 2017; Hensher and Ton 2000; Sayed and Razavi 2000). BPNN tend to achieve high prediction accuracy and are easy to apply. Nonetheless, many different types of ANN models exist, often superior to BPNN, combining both a strong statistical background with machine learning features. Probabilistic neural networks (PNN), for example, are derived by incorporating statistical features (e.g., Bayesian decision rule and kernel density estimation (KDE)) into the structure of the neural network. In addition, additional treatments on the dataset or network structure are applied to obtain more efficient and reliable models.

In this chapter, we investigate the feasibility and capability of four ANN to model discrete mode choice behaviors and compare their prediction performance with a MNL. Specifically, there have been a limited number of articles that conduct comprehensive analyses on different types of ANN algorithms to model discrete choice in the field of transportation. In this study, we focus on four types of ANN: Backpropagation Neural Network (BPNN), Radial Basis Function Networks (RBFN), Probabilistic Neural Network (PNN), and Clustered PNN (CPNN). As a typical travel survey, the Chicago Metropolitan Area for Planning (CMAP) Travel Tracker Survey dataset collected from 2007 to 2008 is used.

This study consists of six sections. After this introduction, section 2.2 presents the five methods for discrete choice modeling. Section 2.3 explains how the data was prepared. Section 2.4 provides the model specifications for MNL and ANN models. Section 2.5 presents the results and a discussion. Section 2.6 is the conclusion.

2.2 METHODOLOGIES: DISCRETE CHOICE PROBLEMS

2.2.1 Random utility model: Logit models

The logit model is the most popular type of random utility model derived from consumer economics theory, and it was initially developed by McFadden (Domencich and McFadden 1975; Train 2009). In utility maximization behavior, an individual i makes a decision to select one choice among discrete alternatives, by evaluating their associated attributes $\boldsymbol{X}$. The individual i chooses the alternative m that provides the largest utility:

$$U_{im} > U_{ik} \; \forall \, m \neq k \qquad (1)$$

In reality, researchers do not observe the complete utility of the individual. Thus, the utility can be classified into two parts: an observed utility V_{im} and an unobserved utility ε_{im}. The observed utility generally contains two sets of attributes: 1) covariates associated with both the individual and the alternative X_{im} and 2) decision maker characteristics, S_i (Train 2009). The observed (stated) utility (V) is a value determine from a linear combination of the attributes used, which captures the attractiveness of an alternative bounded to the given model specification as follows:

$$V_{im} = V(X_{im}, S_i) \qquad (2)$$

In contrast, the unobserved utility ε_{im} cannot be observed by researchers. This unobserved part mainly results from the specification of the observed utility V_{im}. In practice, it is impossible for statistical approaches to include all possible attributes. Therefore, researchers treat the unobserved terms as a stochastic element. Specifically, the logit model is derived by assuming that each unobserved terms, ε_{lm} are independently and identically distributed extreme values—i.e., Gumbel and type 1 extreme values. By combining two utilities, we can get the probability of individual i choosing alternative j by solving the mathematical formulation:

$$P_{ij} = \frac{e^{\beta' x_{im}}}{\sum_{k=1}^{M} e^{\beta' x_{ik}}} \tag{3}$$

where $\boldsymbol{X}_{ik}$ is a vector of observed explanatory variables to choose a given alternative, and β' is parameters for the observed utility. For more technical details about logit models, see (Train 2009).

2.2.2 Artificial Neural Networks

Artificial Neural Networks (ANN) essentially emulate decision-making processes occurring in the brain. As opposed to logit models that assume linearity in the estimation, ANN can algorithmically construct models that are based on nonlinear relationships between determinants. Figure 2-1 (b) shows multi-layer perceptron, MLP that has been widely adopted in most neural network models. In Figure 2-1 (a), the process of predicting output using a single neuron (i.e., a basis function) is similar to the logit model, since it is based on a linear function. ANN, however, are constructed by combining multiple neurons to identify nonlinear relationships between a choice and its associated explanatory variables.

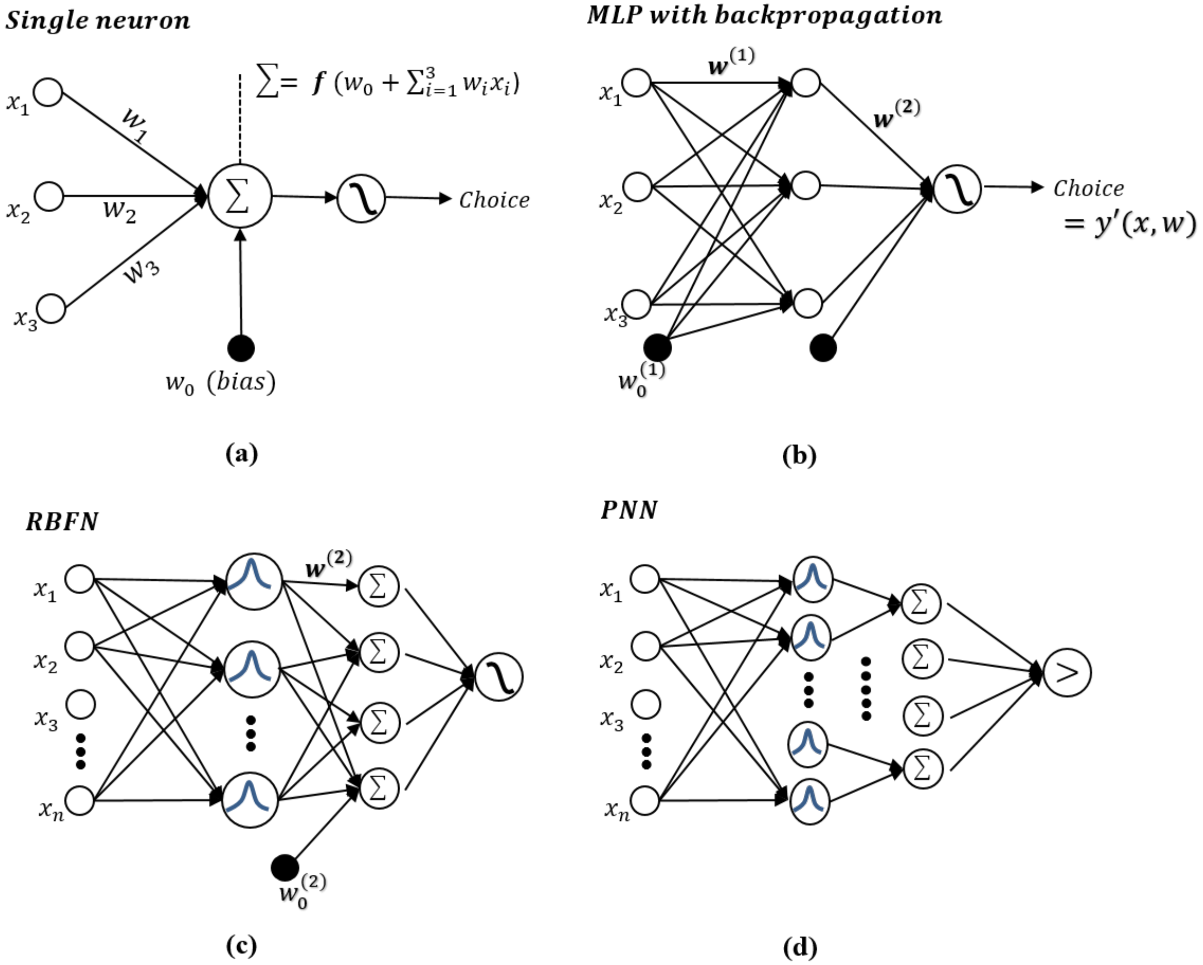


Figure 2-1. Structure of Artificial Neural Network – (a) single neuron (perceptron) network, (b) MLP with backpropagation process (a.k.a., backpropagation neural network, BPNN) (c) feed-forward MLP with radial basis function (a.k.a., radial basis function neural network, RBFN) (d) feed-forward MLP with radial basis function and Bayesian classifier (i.e., Parzen's window classifier) (a.k.a., probabilistic neural network, PNN)

Several types of neural networks exist, and the differences between them mainly come from the nature of the basis function for each model. In the following subsections, we review methodological details about four different types of neural networks, especially in their structural and mathematical differences. Table 2-1 compares these four ANN techniques including the main characteristics, advantages, and disadvantages of the four ANN models.

Backpropagation Neural Network (BPNN)

Backpropagation Neural Network (BPNN) is the most popular and simplest algorithm of MLP. The backpropagation process usually minimizes the error function (Equation 4) of the MLP by changing the value of the weights using a gradient descent method (GDM).

$$E = \sum_{i=1}^{N} \| y'_i - y_i \|^2 \qquad (4)$$

where E is the error, i is the number of training samples, y' is the predicted values through the training, and y is the actual values from the training dataset. Each neuron follows the same activation process as shown for a single neuron. Typical activation functions include the step, sigmoid, tanh, Rectified Linear Unit (ReLU), and the identity functions (Bishop 2006a).

The hidden neurons are "activated" by receiving the sum products of input vectors $\boldsymbol{x}$ and their associated weights $\boldsymbol{w}$. The output layer also obtains information from the hidden layer in the same manner. The final output, y' is:

$$y'(\boldsymbol{x}, \boldsymbol{w}) = \sigma \left(\sum_{j=1}^{3} w_j^{(2)} \cdot h \left(\sum_{i=1}^{3} x_i w_i^{(1)} + w_0^{(1)} \right) + w_0^{(2)} \right) \qquad (5)$$

Where $\boldsymbol{x}$ is input vector, $\boldsymbol{w}$ is the vector of associated weights, $w^{(1)}$ are the weights between the input layer and the hidden layer, $w^{(2)}$ is the weights between the hidden layer and the output layer, and the $h(\,)$ and $\sigma(\,)$ functions are the activation functions (see Figure 2-1 (b)).

The drawbacks of the backpropagation process notably come from the GDM that does not guarantee a global minimum is reached when local minima exist. Thus, the results usually vary each time the BPNN is trained. Second, no "optimal" rule exists to set the structural parameters of BPNN such as the number of hidden layers and their associated number of neurons; although it is generally believed that a larger number of neurons per hidden layer increases the accuracy of the model estimation up to a point (Bishop 2006a). Nonetheless, in general, one or few hidden layers are sufficient. Despite these limitations, it worth mentions that the BPNN is still the most dominant type of ANN used because it is simple to apply and it ensures relatively high accuracy.

Radial Basis Function Networks (RBFN)

A radial basis function network (RBFN) is also a feedforward network with three layers: an input, an RBF hidden, and an output layer (see Figure 2-1 (c)). In contrast to the BPNN, the RBFN is formulated by a linear combination of Gaussian basis functions g(x) as the outputs of the hidden layer:

$$\textit{Output Score} = \sum w\, h(\mathrm{g}(\mathrm{x})) \tag{6}$$

where $h(\,)$ is an activation function, w is a weight between a hidden neuron and an output neuron, and $Output\ score$ is an output for each given class. The basis function g(x) is conceptually obtained by calculating the distance between two vectors based on the Gaussian function whose outputs are inversely proportional to the distance from the mean:

$$\mathrm{g}(\mathrm{x}) = \frac{1}{\sigma(2\pi)^{\frac{1}{2}}}\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{c}_i\|^2}{2\sigma^2}\right) = \exp\left(-\beta\|\boldsymbol{x}-\boldsymbol{c}_i\|^2\right) \tag{7}$$

where $\boldsymbol{x}$ is input vector with n variables (vector to be classified) and $\boldsymbol{c}$ is a “prototype” vector. To be specific, this equation represents the function of the Euclidean distance between the new input vector and the prototype vector. Essentially, the term β determines the width of the Gaussian curve, which controls the decision boundaries. The prototype vector, $\boldsymbol{c_i}$, is prestored in each hidden neuron, and it is a preselected data point from the training set, whether randomly or not (discussed later). Each hidden neuron compares the input vector to its prototype vector to measure how “similar” they are.

The neurons in the output layer are the sum (linear combination) of the weighted hidden layer neurons (i.e., output score), and the final output, y' is determined by activation functions such as sigmoid function.

$$y' = f\left(\sum_{i=1}^{N} w_i \cdot \exp\left(-\beta \|\boldsymbol{x} - \boldsymbol{c_i}\|^2\right)\right) \tag{8}$$

To calibrate the weights, the GDM is also used. By using the outputs (i.e., activations) from each RBF hidden neuron as inputs, the gradient descent process is separately conducted for each output class. Although it tends to be superior to BPNN network structure, similar to BPNN, there is no guarantee that a global minimum in found when optimizing the weights in RBFN.

The performance of an RBFN depends on the number of neurons in the RBFN hidden layer and their prototype vectors. The simplest way to determine the k prototypes is to select them randomly. While it can improve the accuracy of the model, using more prototypes increases the computational costs of the algorithm, and it may generate overfitting issues as well. Another approach to select the prototypes is first to use a clustering algorithm (discussed later).

Probabilistic Neural Networks (PNN)

Probabilistic Neural Networks (PNN) add a Bayesian decision rule into the framework of RBFN (see Figure 2-1 (d)). Technically, PNN is derived from Bayes-Parzen classifiers (Masters 1995), which include Bayesian classification and classical estimators for probabilistic density function (PDF) (Specht 1988, 1990). Essentially, Parzen's method (Parzen 1962) is first used to estimate the PDF of each class (i.e., travel modes) from the provided training samples based on a certain type of kernel. These estimated PDF can then go through a Bayesian decision rule in order to find a final decision.

For any classification problem, a sample vector $\boldsymbol{x}$ is taken from a group of samples, and these samples are classified to the number of classes (i.e., alternatives). Specifically, we assume that the prior probability of a sample that belongs to class i is h_i, the loss associated with misclassifying that sample is c_i, and the PDF for each of the populations $f_k(x)$ is known. Bayesian decision rule for classifying an unknown sample into the ith class can be applied as:

$$h_i c_i f_i(x) > h_j c_j f_j(x) \quad \forall\, j \neq i \tag{9}$$

The PDF for class i, $f_i(x)$, represents the concentration of class i samples around the unknown sample. That is, the Bayesian decision rule chooses a class whether it has high density adjacent to the unseen (i.e., new) sample or misclassification cost or prior probability is high. However, in the Bayesian decision, the PDF for each class is usually known. Therefore, the Parzen's method (Parzen 1962) is usually used to estimate the PDF. The univariate KDE was developed by Parzen and it then was extended to the multivariate case by Cacoullos (Cacoullos 1966). The multivariate KDE for n random (independent) variables can be defined as:

$$\emptyset_i(\boldsymbol{x_n}) = \frac{1}{N\lambda}\sum_{k=1}^{N_i} W\left(\frac{x-x_{ik}}{\lambda}\right) \tag{10}$$

where $\boldsymbol{x}_n$ is the vector of independent variables with n numbers, and $\boldsymbol{x}_k$ is kth training vector choosing class i; N_i is the total number of observations in the training dataset for class i; λ represents all kernel width for the number of variables (n), λ = $[\sigma_1, \sigma_2, \cdots, \sigma_n]$; σ is a smoothing parameters representing kernel width (standard deviation; σ is important to obtain an optimal PNN); W is the weight function, and the most popular type of weight function is Gaussian Kernel, which is mainly derived from the concept of Euclidean distance as:

$$W\left({x - x_{ik}}/{\lambda}\right) = e^{\left(-{\|\boldsymbol{x}-\boldsymbol{x}_{ik}\|^2}/{2\sigma^2}\right)} \tag{11}$$

By employing a Gaussian weight function, the PDF for choosing class i in the given multivariate input vector $\boldsymbol{x}$ can be expressed as follows:

$$\emptyset_\mathrm{i}(\boldsymbol{x}) = \frac{1}{\mathrm{N_i}(2\pi)^{\frac{\mathrm{n}}{2}}\sigma^\mathrm{n}}\sum_{\mathrm{k}=1}^{\mathrm{N_i}} \exp\left(-{\|\boldsymbol{x} - x_{ik}\|^2}/{2\sigma^2}\right) \tag{12}$$

The final classification decision for the input vector $\boldsymbol{x}$ can be obtained by applying Bayes decision rule as follows:

$$C(x) = arg \max_{i=}(\emptyset_\mathrm{i}(\mathbf{x})) \tag{13}$$

Clustered PNN

The ANN using Gaussian KDE such as the PNN may encounter issues related to the size of the network. Specifically, without any pretreatment of the input dataset, the number of pattern neurons containing the prototype vectors is determined either by the number of training samples or by random selection during the training. Since each prototype vector plays a role in KDE by locating the center of the Gaussian distribution for measuring Euclidean distance, the number of pattern neuron can affect the model performance. As the number of pattern neurons gets larger, the structure of PNN will become more complicated and more computationally expensive. In previous studies, two feasible methods have been applied to the conventional PNN: 1) K-means clustering to determine the number of pattern neurons, and 2) self-adaptive learning for smoothing parameter (σ) (Kim et al. 2007; Monfort et al. 2015; Yi et al. 2016). Here, this study employs the K-means clustering method. Although determining σ is directly related to the decision boundaries, this study simply assigns a certain value according to the knowledge of dataset and values selected in other studies.

As an unsupervised learning technique, K-means clustering has been widely used to classify the data set into the number of clusters (Park and Jun 2009). Then, the centroid of a cluster (i.e., the mean value of the observations within the cluster) is defined by minimizing the distance between the observations belonging to a cluster and the center of the cluster (for details about K-means, see (Park and Jun 2009)). The centroids represent the centers of clusters in the Euclidean space, and they become new prototype vectors to measure a Euclidean distance in the pattern neurons. The K-means clustering process is separately applied to each choice alternatives.

2.3 Data

The data used to test the four models comes from the Chicago Metropolitan Agency for Planning (CMAP) Travel Tracker Survey that was collected in 2007 and 2008. The Reveal Preference (RP) survey includes travel diary information as well as detailed individual and household socio-demographic information, for 10,500 households. The original database consists of four interconnected datasets containing household, person, place, and vehicle information. The first three datasets are consolidated to make dataset.

Table 2-1. Description of Statistics

Variable	**Description**	**Mean**	**Std.**$^{1)}$
HHSIZ	Number of household members	2.80	1.43
HHWK	Number of workers in household	1.49	0.92
HHSTU	Number of students in household	0.87	1.14
FEMALE	1: if traveler is female, 0: otherwise	0.52	0.50
EMPLY	Employment status for traveler	1.59	1.05
EDUCA**	Education Level	3.77	1.91
STUDE	Student grade	2.64	0.74
HHVEH**	Number of vehicles in household	1.66	1.06
BIKES**	Number of bikes in household	1.48	1.76
CAPINCOME**	Capita income of household	0.75	0.37
WalkTT	Travel time for walk mode (hour)	4.76	3.92
BikeTT	Travel time for bike mode (hour)	1.32	1.53
AutoTT	Travel time for auto mode (hour)	0.88	0.91
TransitTT	Travel time for CTA mode (hour)	0.90	1.30
AUTO_COST	Total trip cost for auto (gas, toll, parking / $)	0.90	1.11
Transit_COST	Total trip cost for CTA ($)	0.96	1.12
AGE**	Traveler's age	44.02	21.20
ACT_DUR	Actual activity duration (hours)	3.48	3.58
WALK_ACC.	1: if the walking accessibility is within 0.30 miles, 0: otherwise	0.08	0.28

* Std.= Standard deviation
** Employed with either the type of dummy or nominal or categorical

To merge the datasets, the ESRI ArcGIS package is used to identify interrelated geographical and travel information such as the location of the origins (O) and the destinations (D), walk accessibility, and the actual distance between ODs. Since our focus is on the comparison of the five models, we only look at home-based trips. Other trips, such as the access and egress trips, transfer trips, and other nonhome-based trips are excluded in the dataset used for this study.

In addition, the alternatives for the home-based trip are classified into four classes: walk, bike, transit (CTA bus and train), and auto. The original database contains only 2 to 3% of transit observations, which is an issue since the proportion of transit observations in the training and testing datasets can be significantly different (e.g., one of the two could have 0 transit observations). To be able to fairly compare the model accuracies, we select to instead use a sub-sample of the original dataset to include more transit observations. Nonetheless, to assess the performance of each model to select poorly represented modes, we left the bike observations intact, which only account for about 4% of all observations. In the end, the dataset used includes 4,764 observations. It includes two sets of attributes as independent variables: (1) individual/household socio-demographic attributes and (2) travel attributes (e.g., travel time, cost, accessibility) (see Table 2-2).

Table 2-2. Comparison of Four Types of Neural Networks

	BPNN	RBFN	PNN	Clustered PNN
Application Scope	Classification/Regression	Classification/ Regression	Classification	Classification
Classification Process	Minimize sum squared errors by updating weights	RBF and BPNN process	Parzen PDF classifier (KDE and Bayesian decision rule)	Use clustering methods and PNN process
Advantages	• Simple application to predict the patterns • Does not require any statistical features in the learning process • Easy to identify the magnitude of attributes based on weights (relative importance) • A variety of applications are available	• Simpler format of Gaussian function enables to faster learning process than other Gaussian models • Radial basis function nodes can be substituted with different functional forms • Relatively performs well in both smaller and larger dataset	• Simple architecture (no backpropagation) • More way to manage the algorithm by determining the shape of bell curve, specifically width (σ) (more specific than RBFN) • Relatively good accuracy in classification problem • Insensitive to the noise points	• Smaller network size than ordinary PNN • Can avoid saturation of Parzen window that leads to misclassification • May be more applicable because it provides knowledge of relative importance between explanatory variables • Faster training time
Disadvantages	• Easily get stuck in local minima, resulting in suboptimal solution • Like a black box (not sure how to estimate the model) • Need sufficient observations • Prone to overfitting	• Difficult to determine the σ values • Constructing network architecture is complicating. • Long training time	• More computationally expensive than BPNN (prestored pattern neurons) • Saturated Gaussian function can lead some misclassification	• May not provide higher prediction accuracy than PNN for discrete choice data • Varied by number of clusters determined in K-means clustering

2.4 MODEL SPECIFICATION

2.4.1 Variable scaling

Discrete choice data contain a variety set of variables with varying scales and ranges. The different numerical properties among variables may result in estimation biases. Specifically, the PDF for the Gaussian kernel-based ANN (i.e., RBFN and PNN) cannot be estimated without any pretreatment of the input dataset. Moreover, it is preferable to normalize the input data when the sigmoid function is used in BPNN. Therefore, we normalize all values of attributes in the input dataset before training (estimating) the neural networks specifically; the non-normalized data was used for the MNL as is common in practice. The max-min normalization allows all attributes to be located ranging from 0 to 1:

$$x' = \frac{x - \min A}{\max A - \min A} \tag{14}$$

where x' is the new value of the attribute, x is the original value of the attribute; A is the set of all values for a variable from entire data set (training and testing), and $\min A$ and $\max A$ represent the minimum and maximum value of a variable respectively.

2.4.2 Cross-validation

Cross-validation is a technique used to evaluate model accuracy to prevent overfitting. To evaluate the performance, this study used both k-fold cross-validation method and holdout method. The holdout method consists of simply separating the dataset into a training (60%) and a testing (40%) set. Although the holdout method has been widely used for evaluating model performance, by its nature, it may lead to overfitting since only one dataset is used for training. To overcome this issue, the 10-fold cross-validation is also employed, which ensures reliable model performance while minimizing overfitting problems common in machine learning.

The 10-fold cross-validation process functions as follows. The data set is divided into ten subsets, and the ANN is trained 10 times. For each training (i.e., model estimation), one of the ten subsets is left for validating the trained model (a.k.a., test set), and the rest of nine subsets are used to form a training set. The cross-validation process is conducted for 10 times for 10 subsets, and the average accuracy is measured. In this chapter, the 10-fold cross-validation is applied to the four ANN, and the MNL used to measure the overall model accuracy.

Table 2-3. Multinomial logit model estimation for mode choice

Variable	**Coefficient**	**t-stat**
Alternative specific constant (Auto is base)		
Walk	2.442	3.48
Bike	-2.033	-11.41
Transit	-1.097	-3.85
Travel time × Auto	-0.92	-2.06
Travel time × Transit	-1.51	-3.71
Travel time × Bike	-0.05	13.87
Auto operating cost × Auto	-1.25	-12.931
Transit fare × Transit	-2.229	-16.71
Walk Accessibility × Transit	-0.755	-8.12
Walk Accessibility × Bike	0.424	2.74
Bikes in HH × Bike	1.794	13.18
Number of vehicles in HH × Walk	-0.941	-2.08
Number of vehicles in HH × Transit	-1.131	-1.98
Age over 75 × Walk	-0.358	-4.91
Age over 75 × Bike	-1.224	-11.94

Loglikelihood at constant: -4822.92
Loglikelihood at final convergence: -4121.61

2.4.3 Logit model

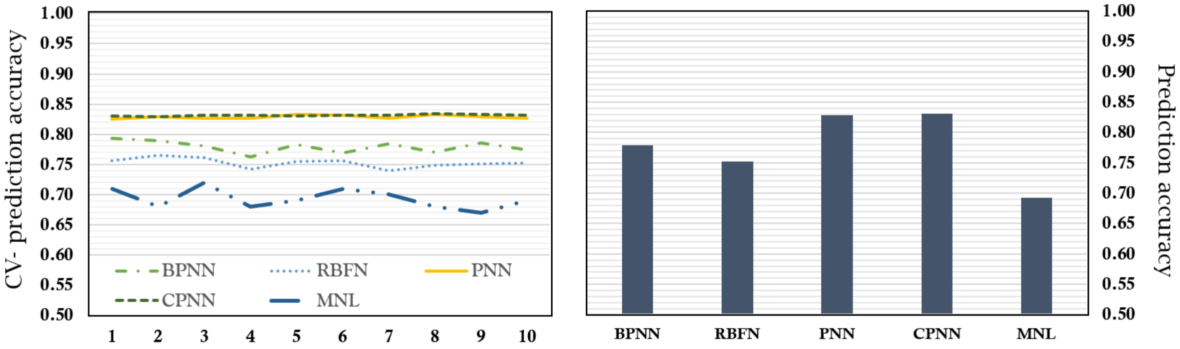
The same training dataset was used to calibrate the MNL, adding two extra dummy variables. The MNL is calibrated with SPSS and Biogeme (Bierlaire 2003a). The results suggest that household size, age, walk availability, vehicle fleets, travel time, and travel costs (i.e., gas price and transit fare) are statistically significant (Table 2-3).

2.4.4 ANN Model

The input dataset for the ANN consists of 14 input values (i.e., explanatory variables) and 4 target values (i.e., mode alternatives). The process of data transmission in the neural networks differs by ANN type. For the BPNN model, we adopt the sigmoid activation function. We also use one single hidden layer (using multiple hidden layers did not improve the model performance). The number of hidden neurons was determined through the training process, and we used 21 neurons in the hidden layer.

To run the kernel-based ANN (i.e., RBFN and PNN), the number of hidden neurons and activation function have to be determined. For the RBFN, the hidden neurons are activated by the Gaussian function, and the sigmoid function is applied to the output neurons. The number of hidden neurons is generally determined based on the size of data or in a random manner during the training process. In this study, the number of hidden neurons is optimally selected by the algorithm. We set the width of Gaussian function β to 0.2 based on the knowledge of the data and on previous values selected in other studies. For the PNN models, most modeling parameters are similar to RBFN, except for the activation function of the output neurons, which is not required in PNN because it is determined by the Bayesian decision rule. For the Clustered PNN, 10 clusters for each alternative are partitioned with K-means clustering.

The ANNs are trained using the following four Python open source packages: Scikit-learn (Pedregosa et al. 2011) for BPNN and 10-fold cross-validation, a combination of Theano (Al-Rfou et al. 2016) and Neupy (Shevchuk 2015) for PNN and RBFN, and a combination of Scikit-learn (Pedregosa et al. 2011) and Neupy (Shevchuk 2015) for Clustered PNN. In addition, all ANN models were run on a laptop computer with an average hardware specification (i.e., 6th generation inter processor with 8GB of RAM). Specifically, computational runtimes for ANN models were similar except for PNN that took 15% longer because of the KDE (specifically for the pattern layer). However, this extra time can be reduced with better hardware specification. In addition, it can be cut down by selecting a limited number of neurons in the pattern layer.



(a) 10-fold cross validation

(b) 4 ANN models and 1 MNL

Figure 2-2. (a) 10-fold cross-validation result for the models (b) comparison of accuracy between the models

2.5 RESULT AND DISCUSSION

The results of the 10-fold cross-validation for the ANN model are presented in Figure 2-2 (a). The results indicate that the CPNN model is relatively less sensible over each validation iteration than the BPNN and RBFN. This is because the BPNN and RBFN models are trained with GDM, which may find slightly different minimum values during cross-validation. Furthermore, we compare the average model accuracy between all five models in Figure 2-2 (b). The results show that the four ANN models achieve better prediction accuracies than the MNL model.

Table 2-4 presents the confusion matrix, in which each row and each column indicates the observed and predicted the number of travelers for each mode respectively. The overall model accuracy of the four ANN is around 80%, thus higher than the accuracy of the MNL with 70.5%. While the four ANN present similar accuracies, the prediction accuracies of each individual mode differ by ANN type. The PNN and CPNN notably show better prediction performance for poorly represented modes. Specifically, the CPNN has slightly better matching rates for the walk and bike modes, in part thanks to the preprocessing with K-means clustering.

In addition to the overall accuracy, this study identifies the sensitivity of mode choice decisions from the two most important explanatory attributes: transit costs and auto costs. Unlike traditional regression models, ANN models cannot estimate the impact of explanatory variables on an outcome variable. Thus, sensitivity analysis can be exploited as a sensible option for examining the impact of explanatory variables. Figure 2-3 (a) and (b) present the result of sensitivity analysis and compare the results by the models for two different variables: auto cost and transit cost. In particular, we can see that the BPNN and MNL model are relatively more sensitive to the variations in two variables. As expected, transit users are more likely to change their mode to auto when the gas price decreases. In contrast, auto users are relatively insensitive to increase in auto costs of 15% or lower.

Table 2-4. Confusion matrix for model accuracy

Test Dataset **(BPNN)**		**Predicted Choice**					
		Walk (114)	Bike (44)	Auto (958)	CTA (473)	Mode Accuracy	Overall Accuracy
Observed Choice	Walk (187)	**68**	13	73	33	36.4%	79.4%
	Bike (64)	22	**17**	11	14	26.6%	
	Auto (824)	7	3	**781**	33	94.8%	
	Transit (514)	17	11	93	**393**	76.5%	
Test Dataset **(RBFN)**		**Predicted Choice**					
		Walk (93)	Bike (27)	Auto (859)	CTA (610)	Mode Accuracy	Overall Accuracy
Observed Choice	Walk (187)	**64**	13	67	43	34.2%	78.4%
	Bike (64)	2	**10**	23	29	15.6%	
	Auto (824)	17	3	**719**	85	87.3%	
	Transit (514)	10	1	50	**453**	88.1%	
Test Dataset **(PNN)**		**Predicted Choice**					
		Walk (115)	Bike (45)	Auto (804)	CTA (625)	Mode Accuracy	Overall Accuracy
Observed Choice	Walk (187)	**94**	7	36	50	50.3%	82.9%
	Bike (64)	6	**27**	23	8	42.2%	
	Auto (824)	2	5	**723**	94	87.7%	
	Transit (514)	13	6	22	**473**	92.0%	
Test Dataset **(CPNN)**		**Predicted Choice**					
		Walk (140)	Bike (48)	Auto (819)	CTA (582)	Mode Accuracy	Overall Accuracy
Observed Choice	Walk (187)	**97**	11	45	34	51.9%	83.3%
	Bike (64)	9	**31**	10	14	48.4%	
	Auto (824)	16	4	**733**	71	89.0%	
	Transit (514)	18	2	31	**463**	90.1%	
Test Dataset **(MNL)**		**Predicted Choice**					
		Walk (108)	Bike (114)	Auto (990)	CTA (377)	Mode Accuracy	Overall Accuracy
Observed Choice	Walk (187)	**55**	41	67	24	29.4%	70.5%
	Bike (64)	2	**21**	33	8	32.8%	
	Auto (824)	28	13	**729**	54	88.5%	
	Transit (514)	11	27	160	**316**	61.5%	

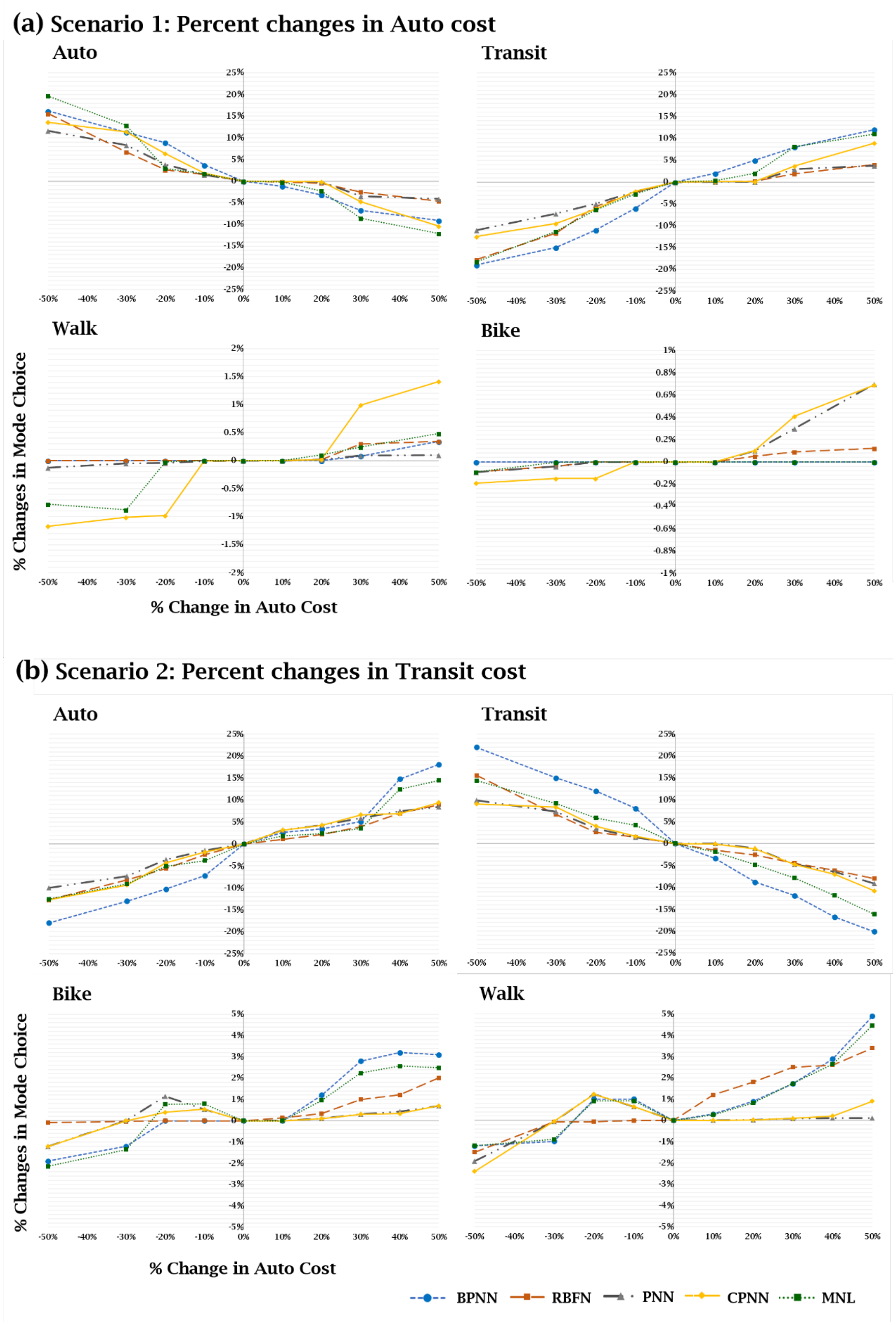


Figure 2-3. Result of sensitivity analysis - comparison between models based on two scenarios: (a) percent changes in auto cost and (b) percent changes in transit cost

2.6 CONCLUSION

This chapter investigated the feasibility, capability, and performance of four ANN methods in dealing with travel mode choice modeling. The prediction performance of ANN and MNL are evaluated by conducting cross-validation tests with the 10-fold cross-validation method. In addition, a confusion matrix was used to evaluate the matching accuracy between the models. The cross-validation results revealed that the four ANN achieve better prediction accuracies (around 80%) than the MNL (around 70%). In particular, the CPNN showed the highest performance among the ANN models in part thanks to the K-means clustering procedure used. The confusion matrix also showed that the matching rates for bike and walk modes were relatively poor since they were poorly represented in the dataset, although the PNN and CPNN performed best. We suggest that this phenomenon is linked to the fact that ANNs are better able to identify nonlinear relationships in a multivariate survey dataset. Nonetheless, although the model accuracy of the ANN outperformed the MNL in mode choice modeling, we must acknowledge that ANN suffers from a lack of interpretability. As a partial strategy to remediate this issue, we conducted a sensitivity analysis for two significant attributes (i.e., transit costs and auto costs) when choosing a mode.

Both the MNL and ANN have advantages and disadvantages. As a parametric approach, logit models are generally used for extracting parameters based on given behavioral patterns. However, when the behaviors being modeled are complex, it becomes difficult to design an accurate logit model. For instance, the MNL is susceptible to unobserved biases, and it is often impractical to identify the relationship and configuration of explanatory variables. In contrast, ANNs are able to capture nonlinearity and biases in the data by using additional information processing units (e.g., hidden layers). Although ANN is often assimilated to a "black box," its relatively easy applicability and its ability to capture complex patterns make it particularly powerful and promising for the future of mode choice modeling.

For future work, novel non-parametric models have emerged recently (Wong et al. 2017). For instance, deep learning (DL) generally require more complex algorithmic features, and they are computationally more expensive than ANN, but in theory, they should be able to even better capture complex and nonlinear relationships in a dataset. Advanced data mining methods such as DL may, therefore, possess a bright future in mode choice modeling in particular, and in travel demand forecasting in general.

3 Predicting Residential Water Consumption: Modeling Techniques and Data Perspectives

(will be published as Lee, D., Derrible, S., "Predicting Residential Water Demand with Machine-Based Statistical Learning", ASCE Journal: Journal of Water Resources Planning and Management)

3.1 INTRODUCTION

Being able to adequately model water demand is essential for municipalities and utilities to effectively meet their customers' demand while managing the available supply of water (House-Peters et al. 2010). In fact, modeling water demand has been integral not only for water resources planning but also for urban infrastructure planning and policy decision making. This is especially the case now as cities are expanding while simultaneously pursuing to consume less energy and fewer resources (Derrible 2016a; b, 2018). Specifically for the water realm, effort should be put into effectively modeling water demand of single-family households since they are the primary consumers of public supply water use in North America (DeOreo et al. 2016).

Determining the right modeling approach (i.e., modeling algorithm) to predict household water demand is a challenging task because water demand can be affected by numerous factors, including technological, demographic, social, economic, and climate characteristics, and public policies (Donkor et al. 2012; Fricke 2014; House-Peters and Chang 2011). A wide variety of statistical techniques exist and can be used to model household end-use water demand. In general, parametric statistical models (a.k.a. parametric models such as linear regression) have been dominantly applied to predict household water use. Recently, with the advent of Data Science and Big Data, the capabilities of available data mining techniques (a.k.a. machine-based statistical learning such as neural networks) seem virtually limitless (Ahmad et al. 2016, 2017; Ahmad and Derrible 2015b; Derrible and Ahmad 2015; Golshani et al. 2018; Lee et al. 2018), and they offer new opportunities to model household water use, especially as they can capture unobserved patterns and nonlinear relationships (Friedman et al. 2001; Lee et al. 2018b; "Pattern Recognition and Machine Learning" 2007).

Both families of algorithms (parametric and ML models) have their own technical characteristics and methodological advantages (e.g., interpretability versus the ability to capturing nonlinearities) that need to be leveraged based on the context around which they are applied. For instance, although attractively more intuitive and interpretable than other models, parametric models (e.g., linear regression) tend to generate more prediction errors than ML models. Furthermore, parametric models require a high degree of domain knowledge to construct and adjust a model's configuration, while taking into account the underlying relationships between factors (e.g., to avoid collinearity issues). In contrast, ML models show a high degree of predictive accuracy with most datasets thanks to their ability to capture nonlinear and complex characteristics in the data, but many ML models are less interpretable than parametric models due to their reliance on machine-based computation process. Thus, selecting the right modeling algorithms is not trivial. In this chapter, 12 statistical learning algorithms are tested, including 4 parametric statistical models and 8 ML models listed in Table 1. To validate the models, a 5-fold cross-validation process (5-fold CV) is applied (detailed later) (Friedman et al. 2001; Kohavi 1995; Zhou 2012), which is generally used to check the performance of ML models.

Beyond modeling, data availability can also be an issue. For example, in the US, most municipalities and utilities do not have access to detailed water use datasets such as longitudinal (i.e., multi-level) or time-series (e.g., smart metered data) datasets, although they have access to general individual or household level information from publicly accessible microdata. In this context, this chapter is purposely designed to investigate the role of data availability on modeling performance. For this, two data scenarios are elaborated: (1) one that intentionally only includes commonly available variables (i.e., general scenario), and (2) one that contains many variables from the Residential End-Use of water survey (i.e., REU 2016 scenario) carried out by the Water Research Foundation (WRF). Specifically, the general scenario only includes six variables on the demographic and economic characteristics of households and climate. These six variables were selected because they are mostly accessible to all US municipalities and utilities from micro-level public data sources or other city-level microdata sample. In contrast, the REU 2016 scenario includes 19 cross-sectional variables (i.e., for one single "typical" day) including variables related to water-saving behavior and detailed water-use patterns; technical details on how the REU 2016 dataset is used in this study is provided the data preprocessing section.

Overall, this study is useful to researchers, municipalities, and utilities who are seeking to model and forecast residential water use, especially when only limited data are available. The lessons from this study can be used for short and long term planning, especially in areas of rapid growth, and for routine operations by utilities. Moreover—although only partly done in this work—the models developed can also be used to infer the impact of individual variables on residential water use (e.g., determine which variables impact water use the most).

This study contains six sections. After this section, section 2 briefly reviews the literature on the two technical aspects central to this work. Section 3 details the data and the analysis process. Section 4 defines the 12 statistical learning algorithms and performance indicators selected for this chapter. In section 5, the overall findings of the study are presented and discussed, and several future tasks are suggested. Finally, section 6 concludes the chapter.

Table 3-1. Statistical methodologies used for two scenarios

Category	Regression methodologies
Parametric- statistical learning algorithm (Parametric models)	· Ordinary Least Squares regression (Linear regression) · Penalized: Ridge regression (Ridge) · Penalized: Lasso regression (Lasso) · Bayesian Ridge Regression (BRR)
Nonparametric- statistical learning algorithm (ML models)	· SVM with Radial Basis Function (RBF) kernel (RBF-SVM) · SVM with linear kernel (linear-SVM) · Kernel Ridge Regression (KRR) · Gradient Boosting Regression (GBR) · Random Forest regression (RF) · K-Nearest Neighbor regression (KNN) · Multi-Layer Perceptron regression (MLP) · Generalized Regression Neural Network (GRNN)

3.2 LITERATURE REVIEW

As mentioned, statistical algorithms can be broadly categorized into two families: parametric statistical models and ML models. Numerous studies focus on modeling household end-use water demand. In particular, parametric models have been dominantly applied to predict household water use since they tend to be more interpretable than ML models due to their strong predetermined (parametric) assumptions (Arbués et al. 2010; Arbues and Villanua 2006; Brentan et al. 2017; Donkor et al. 2012; Goodchild 2003; Guhathakurta and Gober 2007; House-Peters et al. 2010; House-Peters and Chang 2011; Kenney et al. 2008; Kontokosta and Jain 2015). While parametric models are theoretically intuitive and easy to interpret (i.e., since they yield parameters), they also possess serious statistical issues.

First, parametric models have predetermined structures (e.g., residuals are assumed to follow a normal distribution), and a hypothetical test is performed to statistically validate the relationship (Hastie et al. 2009). Furthermore, a single parametric equation is globally employed and is supposed to hold over the entire dataset (i.e., the same relationships are assumed to apply to everyone), while it is notoriously difficult for a linear parametric model to find a best-fitting mathematical function (Friedman et al. 2001; Hensher et al. 2005; Kuhn and Johnson 2013). To partially alleviate this issue, modeling algorithms incorporating clusters (i.e., generalized mixed-effect model) have been used by controlling detrimental effects (e.g., random, fixed) (House-Peters and Chang 2011; Wooldridge 2010). Nonetheless, these modeling algorithms are preferably applied to multi-level data (e.g., longitudinal) that may not be accessible to many cities. Furthermore, finding the best-fitted model specification and configuration while taking into account all possible interactions and relationships between variables (e.g., nonlinearity) is not trivial (Breiman et al. 1984; De'ath 2002; Elith et al. 2008).

As an alternative to parametric models, ML models also have been widely used in the urban infrastructure literature in general (Akbarzadeh et al. 2017; Derrible and Ahmad 2015; Golshani et al. 2018; Lee et al. 2018a; Wisetjindawat et al. 2018) and specifically for household water use (Adamowski et al. 2012; Altunkaynak and Nigussie 2017; Al-Zahrani and Abo-Monasar 2015; Bai et al. 2014; Donkor et al. 2012; Firat et al. 2009, 2010; House-Peters and Chang 2011; Vitter and Webber 2018; Yurdusev et al. 2010). In general, ML models have been shown to have high predictive performances on a wide range of modeling applications thanks to significant advances in computational ability. Specifically, ML models can recognize non-trivial patterns from a dataset that often result in high prediction accuracies.

In particular, Artificial Neural Network (ANN) models have been widely applied to predict or forecast water consumption (Altunkaynak and Nigussie 2017; Firat et al. 2009, 2010). For instance, Firat et al. (Firat et al. 2009, 2010) used six different ANN models to forecast monthly water consumption by using time-series data and found that Generalized Regression Neural Networks (GRNN) models performed best. Apart from ANN, Bai et al. (2014) used a step-wise support vector machine (SVM) regression to forecast daily water consumption by using time-series data, which is called a variable structure SVM. Instead of using ML to model water demand directly, ML models are also applied to facilitate water demand analysis. For example, Vitter and Webber (2018) used SVM classifier to classify specific water use events (e.g., shower, clothes wash) in households by incorporating electricity consumption information that correlates to water consumption. Numerous other ML models have been applied to predict other urban resources (e.g., electricity, energy), including kernel-based methods, boosting methods, and bagging methods (Bansal et al. 2015; Kusiak et al. 2010; Lozano and Gutiérrez 2008; Robinson et al. 2017; Tso and Yau 2007).

In general, ML models include nonparametric (e.g., kernel) and complex structure (e.g., network) models that can capture nonlinear or complex relationships between various factors and target values (e.g., household water use). Furthermore, they generally provide higher predictive performances than parametric models in resources demand modeling (Al-Zahrani and Abo-Monasar 2015; Firat et al. 2009, 2010; Robinson et al. 2017) since nonparametric features in the model are trained by machine-based repetitive computation. Due to their machine-based computation, however, ML models can face overfitting problems more easily, and they are also generally less interpretable than parametric models (e.g., neural networks are often described as "black-box" models). To address the interpretability issues in ML models, several useful statistical measures exist. For instance, rule-based ensemble methods (e.g., boosting and bagging) are able to examine the marginal effect of a given factor on the predicted values of a learned model (a.k.a., partial dependence plot) (Doshi-Velez and Kim 2017b; Friedman et al. 2001; Natekin and Knoll 2013).

In addition to modeling methodologies, the performance of water demand models largely depends on the quality of the data available that properly capture the relationship between water demand and the factors affecting the demand. Previous studies on household water demand modeling found that the most significant factors affecting water use include household demographic factors (e.g., size, income, type) (Arbués et al. 2010; Arbues and Villanua 2006; DeOreo et al. 2016; Domene and Saurí 2006; Grafton et al. 2011; House-Peters et al. 2010; Mayer et al. 1999; Mazzanti and Montini 2006; Schleich and Hillenbrand 2009), climate factors (e.g., precipitation, temperature) (Donkor et al. 2012; Froukh 2001; Goodchild 2003; Guhathakurta and Gober 2007; House-Peters et al. 2010; House-Peters and Chang 2011; Jentgen et al. 2007; Kontokosta and Jain 2015; Lee et al. 2010, 2015; Schleich and Hillenbrand 2009), price, and detailed water use and associated attitudinal information related to the households (Arbués et al. 2010; Arbues and Villanua 2006; Cominola et al. 2018; DeOreo et al. 2016; Fricke 2014; Ghimire et al. 2015; Grafton et al. 2011; House-Peters et al.

2010; Kontokosta and Jain 2015; Vitter and Webber 2018; Willis et al. 2011). The first two factors (i.e., household demographics and climate factors) are generally available to cities and water utilities in the United States, which is not the case for detailed information on water use and its behavioral characteristics that is rarely available.

3.3 RESEARCH DESIGN AND DATA PREPARATION

3.3.1 Research Design

This chapter is designed to examine two common technical issues and investigate the modeling performances of twelve techniques under two data scenarios; see the research framework in Figure 3-1. Before the main analysis, this study conducts a thorough descriptive analysis to detect the presence of statistical issues in the dataset, which is often the case for data that relate to resource consumption (e.g., water, electricity). Then, the main analysis is to train twelve statistical learning algorithms (see Table 3-1) on 70% of the data (i.e., train set) under two data scenarios (i.e., general and REU scenario). A 5-fold cross-validation (CV) process is also applied to the train set. The learned models are then validated on the remaining 30% of the data (i.e., test set). The next section offers details on the data and the two modeling scenarios.

3.3.2 Data preprocessing

This chapter uses the 2016 Residential End Use of water survey (REU 2016) database (DeOreo et al. 2016) released by the Water Research Foundation (WRF). The REU 2016 study contains extensive household water use information from 24 water utility companies across the United States and Canada. The REU database consists of four main datasets that come from two main sources: (1) household water use (e.g., 12 days of metered consumption) and billing information (e.g., annual water consumption) and (2) household survey responses (DeOreo et al. 2016). In particular, the metered consumption was originally measured every 10 seconds for two weeks, but it was subsequently aggregated by day for 12 days (DeOreo et al. 2016).

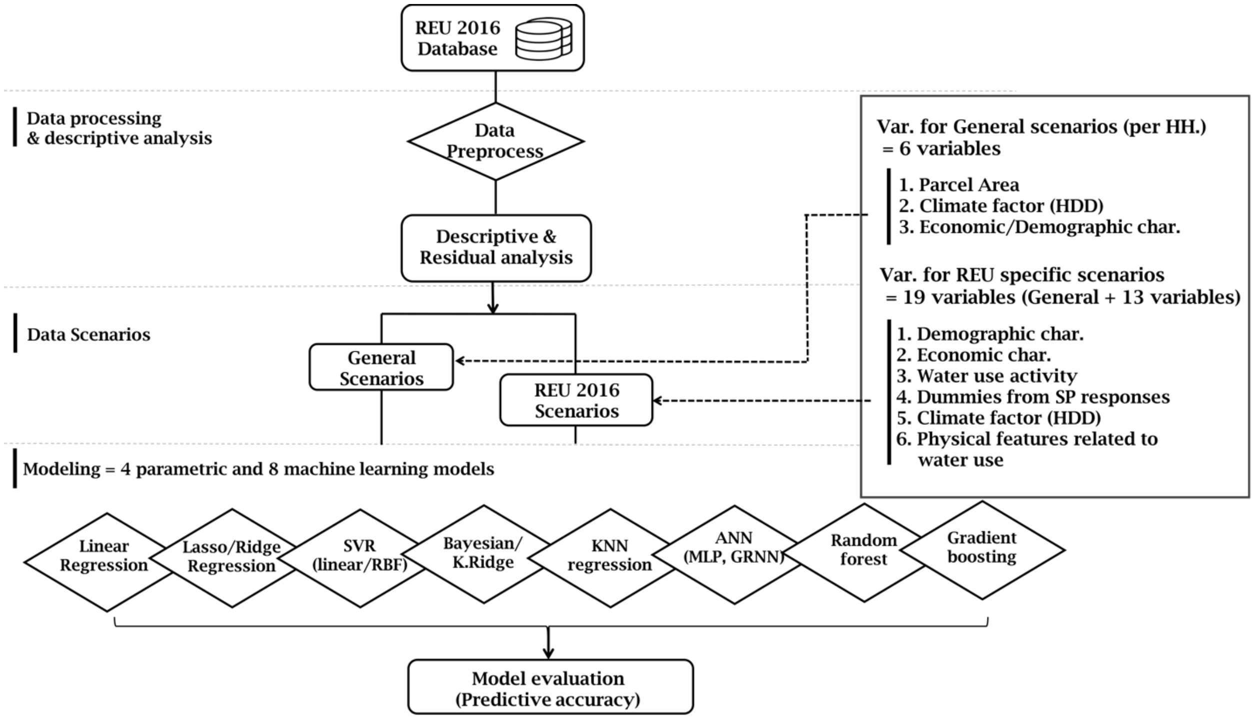

REU 2016 Database
Data processing & descriptive analysis
Data Preprocess
Descriptive & Residual analysis
Var. for General scenarios (per HH.) = 6 variables
1. Parcel Area
2. Climate factor (HDD)
3. Economic/Demographic char.
Var. for REU specific scenarios = 19 variables (General + 13 variables)
1. Demographic char.
2. Economic char.
3. Water use activity
4. Dummies from SP responses
5. Climate factor (HDD)
6. Physical features related to water use
Data Scenarios
General Scenarios
REU 2016 Scenarios
Modeling = 4 parametric and 8 machine learning models
Linear Regression
Lasso/Ridge Regression
SVR (linear/RBF)
Bayesian/ K.Ridge
KNN regression
ANN (MLP, GRNN)
Random forest
Gradient boosting
Model evaluation (Predictive accuracy)

Figure 3-1. Research Design

This study mainly uses two datasets from the REU 2016 database: (1) daily household water use ("REU2016_Daily_Use_Main") and (2) mailed household survey information about demographics and water consumption behaviors ("REU2016_End_Use_Sample"). The daily household water use dataset includes nine utilities for a total of 771 households over 12 days. In total, the number of observations available is around 9,300 (some households include more than 12 days). Furthermore, the mailed survey dataset contains detailed information on household demographic and economic characteristics as well as water consumption behaviors in the form of RP (revealed preference) and SP (stated preference).

Table 3-2. Variables used in two scenarios: General and REU 2016 specific

General Scenario: 6 variables				
Independent var. (X)	Description	N	Mean	Std.
Capita	Number of people in household	531	2.73	1.44
HDD	Heating degree days	531	4098.92	2432.28
Employed adults	Number of workers in household	531	1.32	0.9
Income	Household income (ten thousand dollars)	531	8.17	5.26
Parcel area	Size of parcel area (m^2)	531	809.08	496.0
Dummy outdoor	Existence of outdoor properties is 1; otherwise is 0 (e.g., garden, tree, lawn, pool)	531	0.6	0.49
REU 2016 Scenario: general scenario (6 variables) + 13 variables				
Independent var. (X)	Description	N	Mean	Std.
Bedrooms	Number of bedrooms in household	531	3.38	0.87
Outdoor Area	Size of outdoor area (m^2)	531	320.98	330.14
Pool Area	Size of pool area (m^2)	531	15.72	17
Homies	Person usually stay in the house	531	1.01	0.85
Vintage	Vintage of home	531	34.59	19.44
Fixed Charges	Fixed rates for water	531	17.6	9.57
Marginal Rate	Marginal rates for water	531	4.98	2.24
Dummy Treatment	treatment system in household is 1; otherwise 0 (e.g., water softener or reverse osmosis system)	531	0.12	0.33
Dummy pool	Household with pool (indoor or outdoor) is 1; otherwise is 0	531	0.11	0.32
Dummy toilet flush	Average toilet flush is less than 7.58 liters per flush is 1; otherwise is 0	531	0.45	0.5
Dummy shower flow	Average shower flow is less than 7.58 liters per min. 1; otherwise is 0	531	0.51	0.5
Dummy clothes load	Average washer load is less than 11 liters per load is 1; otherwise is 0	531	0.51	0.5
Dummy Hot water	Hot water wait in master bathroom is 1; otherwise is 0	531	0.45	0.5
Dependent var. (Y)		N	Mean	Std.
Trace Daily [1)]	Daily water consumption (liters per day, lpd)	531	714.98	428.72

1) Daily water consumption is transformed into $log_{10}Y$. See details in section, Descriptive analysis and Figure 3-3.

From this database, this study purposely creates cross-sectional data by combining average daily household water use with household survey information, based on the given identification codes (KEYCODES). To calculate the average daily water use, the 12 recorded days of household daily water use are averaged. In essence, this dataset is transformed into cross-sectional information to model one single "typical" day for the 771 households.

Subsequently, the combined data contained numerous missing values, and some variables contained redundant or inter-related information that can bias the results (i.e., collinearity issues). Therefore, multiple cleaning and variable selection processes were initially conducted. In particular, variables with a very low response rate (< 10~20 %) were eliminated and some variables having redundant or inter-related information were merged. In addition, household water demand also depends on climate conditions, which must be taken into account since all households are not located in the same geographic location. For this study, the number of heating degree days (HDD), the number of cooling degree days (CDD), and the climate zone (CZ) of each household was added to the dataset from www.degreeday.net and from maps provided by the American Society of Heating, Refrigeration, and Air-Conditioning Engineers (ASHRAE).

To further study the relationship between the variables, Figure 3-2 shows the correlation matrix between all independent variables. Specifically, no significant collinearity issues can be detected that can lead to biased estimation; i.e., Spearman coefficient = 1.0. Nonetheless, as expected, HDD, CDD, and climate zone are strongly correlated, and the pairs between outdoor properties (e.g., pool) and climate conditions also show some correlation. Moreover, variables related to household size also show some correlation, such as a number of toilets and bedrooms. In each case, only one of the variables that show some correlations was selected; for instance, only HDD was selected. In the end, the dataset contains 24 variables and 531 observations (i.e., single-family households). The full list of variables used is shown in Table 3-2.

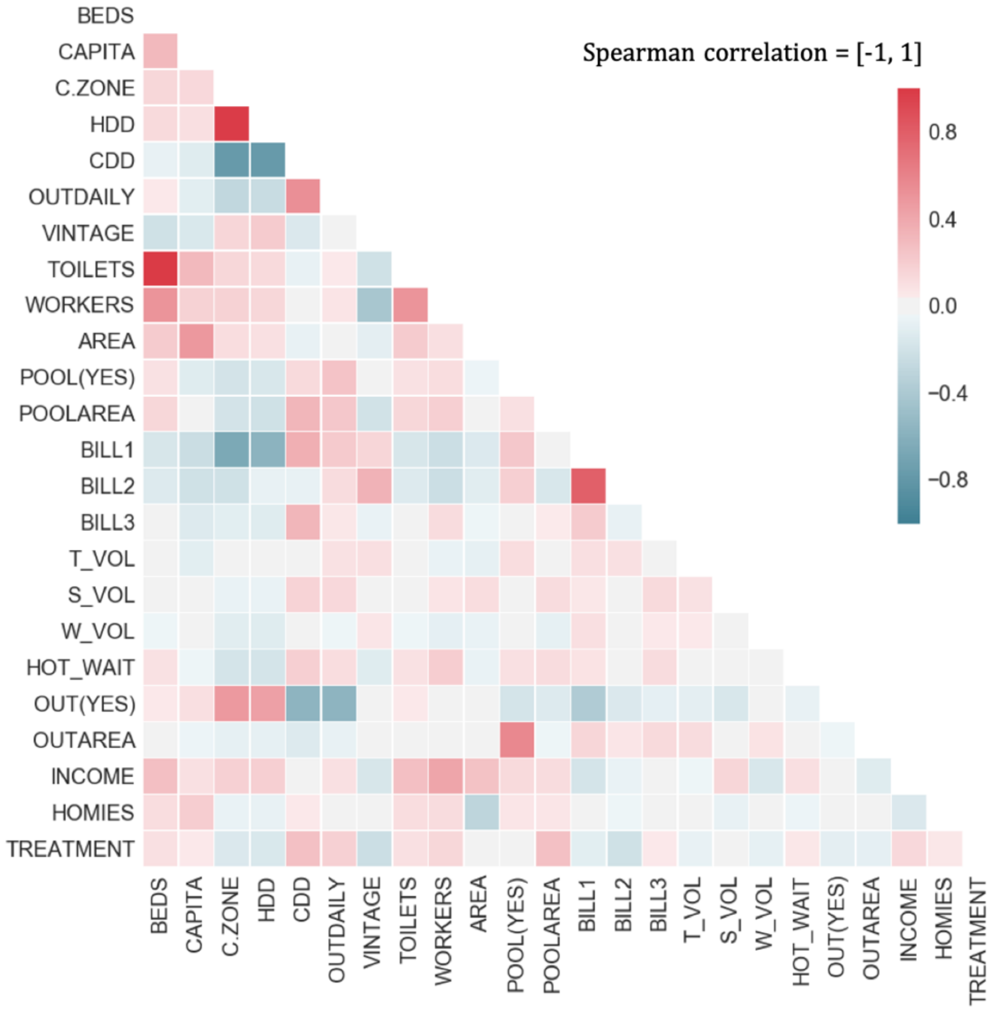


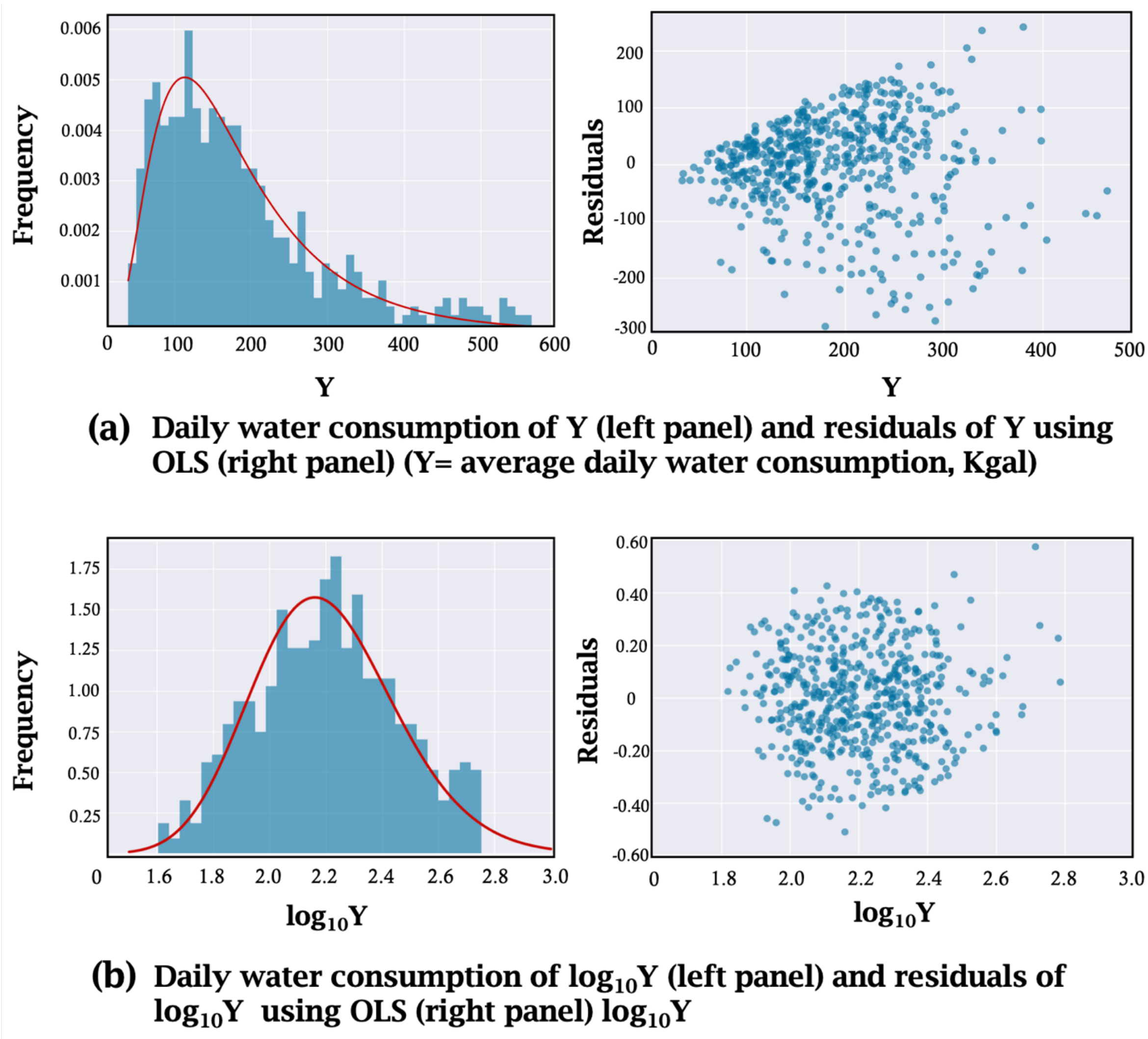
Figure 3-2. Correlation of independent variables (predictors)

Overall, the general scenario includes six variables: the number of workers, household size, income, type, areas, and HDD. In particular, in the United States, these variables are available from publicly accessible micro datasets such as the American Community Survey

(ACS), the Public Use Microdata Sample (PUMS), and from community-level household surveys. Although public micro datasets are mostly anonymous and only contain a limited number of samples, utilities and municipalities can use this information based on existing statistical approaches that are widely used in resources planning process—see details in Farooq et al. (2013), Guo and Bhat (2007), and Rosca et al. (2018). In contrast, the REU 2016 scenario includes all information in the general scenario and more detailed household level water use information that describe household water consumption from RP and SP responses. For both scenarios, daily total household water consumption in Liters per day is predicted, expressed as *TraceDaily*, that includes both indoor and outdoor water consumption (although only a limited number of households report outdoor properties (e.g., garden, lawn, pool, etc.)).

3.3.3 Descriptive Analysis

An early investigation of the distribution of household water use (i.e., *TraceDaily*) reveals that the variable is not normally distributed. Using Ordinary Least Squares (OLS) regression, Figure 3-3 (a) shows that the distribution of household water use (Y) is skewed to the right, which is common in lognormal distributions. Furthermore, the residual plots show the presence of a funnel-shaped pattern, suggesting a non-constant variance in the error terms—i.e., heteroscedasticity—which violates the predetermined assumption in linear models (e.g., linear regression). To solve this heteroscedasticity issue, the actual Y can be transformed using a concave function such as the logarithm function ($\log_{10} Y$) or the square root of the actual Y ($\sqrt{Y}$) (James et al. 2013; Kuhn and Johnson 2013). This transformation shrinks the responses, which can alleviate the heteroscedasticity issue. In the literature, the log-transformed is most commonly used (Keene 1995; Kuhn and Johnson 2013; Robinson et al. 2017). As shown in Figure 3-3 (b), taking the log-transformed of *TraceDaily* results in a normal distribution and randomly scattered residuals. As a result, $\log_{10} Y$ is used as the dependent variable instead of Y.



(a) **Daily water consumption of Y (left panel) and residuals of Y using OLS (right panel) (Y= average daily water consumption, Kgal)**

(b) **Daily water consumption of $\log_{10}$Y (left panel) and residuals of $\log_{10}$Y using OLS (right panel) $\log_{10}$Y**

Figure 3-3. Comparison of error plots for REU 2016 specific scenario (the horizontal is the logarithm of normalized actual water consumption ($\boldsymbol{\log_{10} Y}$) and the vertical axis is that of predicted consumption values)

3.4 METHODOLOGIES

Parametric linear models (i.e., linear and penalized regression) either assume that the regression function $E(Y|X)$ is linear in the inputs (X) to predict the output (Y) or that the linear model reasonably fits along with a flat hyperplane (Hastie et al. 2009; James et al. 2013; Kuhn and Johnson 2013). Thus, parametric linear models are simple and they can sometimes outperform nonlinear models, especially for small and sparse data (Hastie et al. 2009). In addition to the conventional linear regression technique (i.e., Ordinary Least Squares), several parametric linear models introduce additional information or statistical assumptions, such as partial least squares (PLS), two-staged least squares (2LS) and least squares with panalized terms such as lasso and ridge regression to decrease the level of biases while preserving the predetermined assumptions (i.e., linearity). In contrast, numerous nonparametric learning models or ML models exist that can be inherently adapted to the data without assuming a linear regression function in $E(Y|X)$ is linear. Due to differences between the two modeling categories, it is difficult to simply conclude which modeling technique is superior to the others. Specifically, it largely depends on the purpose of the research and the intrinsic characteristics of the data used in the model.

Regardless of the algorithm, all have several common features. In particular, most statistical models estimate the relationship between a set of independent variables X with a dependent variable y while minimizing a loss function. For example, many models minimize the sum of squared errors (SSE), and they are then evaluated by measuring how much they managed to minimize SSE; e.g., by using the mean of the squared errors (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{N}(y_i - f(\boldsymbol{x}_i))^2 \qquad (1)$$

where N is the total number of observations, $\boldsymbol{x}_i$ is a vector of an independent variable, and y_i a dependent variable for i_{th} observation. In fact, the MSE can be decomposed into three parts:

$$E[MSE] = \varepsilon^2 + (Bias)^2 + Variance \tag{2}$$

The first part (ε^2) consists of the unobserved errors that are impossible to eliminate in modeling. The "Bias" in the second term illustrates how well the estimated model can explain the relationship between x and y. The last term is the variance. Generally, the aim is to control the level of bias and variance when estimating a model. Specifically, more complex models (e.g., artificial neural networks) can have higher variances than models based on the linear assumption (e.g., linear regression), which can lead to overfitting. In contrast, simpler models can have lower variances, but they may not be able to fully infer the relationship between X and y, thus resulting in underfitting. This trade-off between the two families of techniques is often referred to as the variance-bias trade-off (James et al. 2013; Kuhn and Johnson 2013). The following sections detail the 12 statistical learning selected in this study.

3.4.1 Parametric statistical learning algorithms

3.4.1.1 Linear regression model

Linear regression aims to explain the relationship between a set of independent variables and a dependent variable (y) based on the linear function:

$$y = f(X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \tag{3}$$

where X_j is a vector for the j_{th} independent variable, and β_j and β_0 are unknown parameters (coefficients and an intercept), respectively. This linear combination is estimated by minimizing the sum of the squared errors (SSE) between X and y (see equation 1), and it is also known as the standard Ordinary Least Squares (OLS) regression.

3.4.1.2 Penalized models: Ridge and Lasso regression

Penalized models aim to mitigate problems related to model variance when the number of independent variables increases in the standard OLS regression. Specifically, it is possible that highly correlated variables (i.e., collinearity) can greatly increase the variance, and such variance issues can increase the overall MSE. Thus, the family of penalized models, including Ridge and Lasso regressions, regulate the estimation process by adding a penalty to the SSE. Ridge regression adds the L_2 penalty in the SSE, which controls the trade-off between the variance and the bias. Specifically, this penalty sacrifices some bias, and it can reduce the variance that provide a lower MSE:

$$SSE_{L_2} = \sum_{i=1}^{N}(y_i - f(\boldsymbol{x}_i))^2 + \lambda \sum_{j=1}^{P} \beta_j^2 \qquad (4)$$

where λ regulates the inflation of coefficient, and it is required to be calibrated through validation process.

In addition to the lasso regression, ridge regression has a L_1 penalty that substitutes the L_2 penalty in the ridge regression.

$$SSE_{L_1} = \sum_{i=1}^{N}(y_i - f(\boldsymbol{x}_i))^2 + \lambda \sum_{j=1}^{P} |\beta_j| \qquad (5)$$

3.4.1.3 Modified ridge regression: Kernel and Bayesian ridge regression

Ridge regression is the simplest algorithm that can be kernelized or combined with probabilistic features (e.g., Bayesian). Specifically, $\boldsymbol{x}$ in equation (5) is substituted with the kernel function, $\emptyset$:

$$SSE_{L_2} = \sum_{i=1}^{N}(y_i - f(\emptyset_i))^2 + \lambda \sum_{j=1}^{P} \beta_j^2 \qquad (6)$$

It is termed kernel ridge regression since it uses the same loss function used in ridge regression. Alternatively, the context of Bayesian statistics can also be applied to ridge regression. Specifically, the prior information and the posterior mean of a model for the parameter (β_j) follow:

$$\text{Prior: } \beta_j, \sim N(0,1/\lambda);\ \text{posterior: } \beta_j, \sim N(0,\sigma^2/\lambda),\ \text{for all } j \qquad (7)$$

All the modeling parameters are jointly estimated by maximizing the marginal log-likelihood (LL) function.

3.4.2 Non-parametric machine-based statistical learning algorithms

3.4.2.1 Rule-based ensemble models: Random Forest and Gradient Boosting Regression

Rule-based (a.k.a., tree-based) models estimate the relationship between X and y by partitioning the input based on specific rules. In particular, they provide a set of conditions and results that are highly interpretable, and they also easily include different types of variables without any assumptions and preprocessing. However, simple trees can have highly unstable performances and they tend to have higher variances than linear models (e.g., linear regression). Therefore, ensemble methods are generally preferred as they can reduce the variance (James et al. 2013; Kuhn and Johnson 2013). This chapter adopts two popular ensemble methods: (1) Random Forest (RF) and (2) Gradient Boosting Regression (GBR).

Bagging algorithms, also called bootstrap aggregation techniques, build a large number of de-correlated trees by using bootstrapping, and they then average them. Specifically, the bagging process in RF is as follows (Friedman et al. 2001):

(1) Draw bootstrapped samples (size N) from the original dataset
(2) Grow a regression tree for the bootstrapped samples and a subset of independent variables, and then recursively repeat tree growing process until the stopping criteria are reached (i.e., minimum node size).
(3) Average all regression tree (size N) while reducing the overall model variance, which is also called bagging.
(4) Random forest model predicts y given $\boldsymbol{x}_i$
$y\,(\boldsymbol{x}_i) = \hat{f}_{RF}^{N}(x_i) = \frac{1}{N}\sum_{b=1}^{N} T_b(x_i)$
where, x_i is a vector of independent variable. $T_b(x_i)$ represents a single regression tree grown by bootstrapped samples and a subset of variables. N represents the total number of regression trees.

GBR uses another tree ensemble technique, known as a boosting algorithm. Although bagging algorithms (i.e., RF) also uses multiple trees through sampling processes (e.g., bootstrapping), boosting algorithms sequentially grow the trees. Specifically, each tree is grown by using information (i.e., poorly fitted observations) from previously grown trees, and different weights are assigned at each step (James et al. 2013). The general boosting process for GBR is as follows:

(1) Initially set number of trees (estimators), N, and number of split (tree depth), D (stopping criteria)
(2) A target (dependent) variable, $\hat{f}(x) = 0$, is initially set as zero, and residual (r_i) and target (dependent) variable (y_i) are assumed to be identical for all observation (i).
(3) During the boosting process for each tree estimator (N number of trees), the following steps are repeatedly and sequentially conducted:

- Estimate a tree and compute the residual for each observation (computing negative gradient, *r*)
- Fit a regression tree $\hat{f}^b$ to the data $(\boldsymbol{x}, r)$; b denotes a single regression tree
- Compute a new target value, $\hat{f}$, by adding in a regularized new tree,
  $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$
- Update the residuals,
  $r_i \leftarrow r_i - \lambda \hat{f}^b(x_i)$

(4) Sum up sequential trees that predicts y given $\boldsymbol{x}$,
$f_B(\boldsymbol{x}) = \sum_{b=1}^{N} \lambda \hat{f}^b(x)$

3.4.2.2 Support Vector Machine

Support Vector Machine (SVM) is a kernel-based method to find the optimal generalization boundaries for fitting y based on X. In fact, when used for regression, SVM inherits some properties from the SVM algorithm used for classification. Specifically, SVM adopts different kernel functions ($\emptyset$) to capture the relationship between X and y:

$$y\,(\boldsymbol{x}) = \sum_{m=1}^{M} \beta_m \emptyset_m(x) + \beta_0 \qquad (8)$$

where $\emptyset$ is kernel function (also called basis function) with M numbers. To estimate parameters (β and β_0), the following kernel function is minimized:

$$\min\ \emptyset(\beta, \beta_0) = \sum_{i=1}^{N} V_\varepsilon^r + \frac{1}{2}\|\beta\|^2 = \sum_{i=1}^{N} V(y_i - f(x_i)) + \frac{\lambda}{2}\sum \beta_m^2 \qquad (9)$$

(where, $V_\epsilon^r = 0$ (if $|r| \leq \varepsilon$), $|r| - \varepsilon$, otherwise)

where V_ϵ^r measures the general errors from the support vectors selected by the model. The element ε is the threshold to manage the number of support vectors used for finding optimal bound, and λ is called the penalty parameter determining the flexibility of the model. Both ε and λ are required to be tuned to balance the variance-bias trade-off.

3.4.2.3 ANN models: multi-layer perceptron and generalized regression neural network

Artificial Neural Networks (ANN) algorithmically construct a simplified model of the human brain to explain or infer the relationship between X and y. The Multi-Layer Perceptron neural network (MLP) has been widely adopted and it consists of three layers: input, hidden, and output. It is also called the Backpropagation Neural Network (BPNN). In MLP, the hidden layer is able to capture nonlinear relationships between X (input layer) and y (output layer).

For regression problems, the MLP takes input data and computes an output result based on the value of inputs ($\boldsymbol{x}$) and the corresponding weights ($\boldsymbol{w}$) using an internal activation function (f). The activation function is used to transfer inputs to outputs according to the functional form of the activation function. The weights are scaled values associated with the connections between neurons. We can express that MLP predicts y given $\boldsymbol{x}$ as:

$$y\,(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \,f\left(\textstyle\sum_i^n w_i \emptyset_i\,(\boldsymbol{x})\right) = \,w_0 + \,f\,(\textstyle\sum_i^n w_i x_i) \qquad (10)$$

where $\boldsymbol{x}$ is a vector of input factors, $\boldsymbol{w}$ is the vector of associated weights, and the $\emptyset_x$ denotes basis functions. Here, the function (f) takes $\boldsymbol{x}$ as the basis function in the form of a linear combination. To estimate the weights, a backpropagation process is applied to minimize the loss function (SSE) for the MLP by generally using the gradient descent method (GDM).

Generalized regression neural network (GRNN) is a feed-forward network that is physically identical to the architecture of BPNN (i.e., MLP)—three layers consisting of an input layer, a hidden layer (radial basis function layer), and an output layer. In contrast to the BPNN model, GRNN is formulated by a linear combination of input ($\boldsymbol{x}$) and associated weights ($\boldsymbol{w}$) through radial basis function (RBF) such as Gaussian density function, g(x). For predicting y given $\boldsymbol{x}$, GRNN can be expressed as:

$$y\ (\boldsymbol{x}, \boldsymbol{w}) = w_0 + \ f\ (\textstyle\sum_i^n w_i \mathrm{g}_i\ (\boldsymbol{x})) \qquad (11)$$

where $\boldsymbol{w}$ is the vector of associated weights between output and hidden layers. The main difference between equations 10 and 11 is the basis function that is changed from a linear to a Gaussian basis function.

Specifically, the basis function g(x) is conceptually obtained by calculating the distance between two vectors based on the Gaussian function whose outputs are inversely proportional to the distance from the mean:

$$\mathrm{g(x)} = \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{c}_i\|^2}{2\sigma^2}\right) \sim \exp\left(-\beta\|\boldsymbol{x}-\boldsymbol{c}_i\|^2\right) \qquad (12)$$

where $\boldsymbol{x}$ indicates new input data samples to be classified with n variables and $\boldsymbol{c}$ is a mean of Gaussian distribution, which is also known as "prototype" vector. To be specific, this equation computes the geometric distance between the new input vector and the prototype vector (i.e., mean of Gaussian distribution), thus, the similarity of the input vector and prototype vector is measured.

3.5 MODEL SPECIFICATION AND EVALUATION

3.5.1 Variable scaling

A set of variables in the data set is recorded based on varying scales and ranges. These numerical differences among variables may result in the biased estimation, especially for some nonparametric models that are sensitive to scales (e.g., ANN, SVR). Therefore, scaled values (z_i) are preferred for both the independent and dependent variables in all models. Although variables do not need to be scaled for linear models, the same values are used in all models for consistency. In this study, the conventional min-max scaling technique is adopted; it is defined as:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{13}$$

where, z_i is the scaled value of the ith sample, x_i is the original value of the ith sample, and $\max(x)$ and $\min(x)$ represent the minimum and maximum value of x.

As mentioned above and as is common in data mining, all models are trained on 70% of the data and tested on the remaining 30% of the data. Furthermore, to detect any overfitting issues, a 5-fold cross-validation (CV) analysis is conducted. Namely, the train set (i.e., 70% of the data) is divided into 5 partitions, 4 of the 5 partitions are used for model training, and the 1 remaining partition is used for evaluating the model. This process gives us 5 trained models, and the average and standard deviation of the performance of each model are calculated. Each model is then trained again on the full train set and tested against the test set. Although this two-step process adds some redundancy, it offers a statistically robust method to validate the results.

3.5.2 Model evaluation metrics

To evaluate the models, three metrics are used: mean absolute error (MAE), mean squared error (MSE), and adjusted r-squared (R^2_{adj}). MAE and MSE are primarily used to measure the deviation between the actual and predicted water consumption values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{14}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{15}$$

where, $\hat{y}_i$ is the predicted value for i_{th} household. In addition, R^2_{adj} is calculated to see how close the predicted values are to a fitted line or curve to overcome the limitations of the traditional R^2 indicator. In particular, R^2 increases whenever more independent variables are added; thus, more variables may appear to better fit the data, while this is not necessarily the case. R^2 can also be affected by the "noise" in the data. The traditional R^2 is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{16}$$

where n is number of observations. In contrast, the R^2_{adj} is adjusted by the number of variables in the model, and it can control the increase of R^2 (Hastie et al. 2009). R^2_{adj} is therefore lower or equal to R^2, and it is defined as

$$R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-k-1} \tag{17}$$

where, k is number of independent variables, and $\bar{y}_i$ is the mean of actual values.

3.6 MODEL RESULT AND DISCUSSION

This section contains the model validation results for the REU 2016 scenario first and then for the general scenario.

3.6.1 REU 2016 Scenario: including 19 variables

The performance of the 12 statistical models using the 5-fold cross-validation for the REU 2016 scenario is shown in Table 3-3. The table includes the MAE, MSE, and R^2_{adj} values. The first value in the table is the average performance, and the uncertainty value after the '±'

is the standard deviation. All models use normalized data and predict the log-transformed of househodl water use. All are implemented with the Scikit-learn library built in Python (Pedregosa et al. 2011).

The results show that GBR outperforms the other models with an MAE of 0.098 and R^2_{adj} of 0.69. Furthermore, models containing probabilistic or nonparametric features such as KRR, RBF-SVR, RF, and GRNN achieve better performance than other models. Parametric (linear) models such as Ridge, linear-SVM, and linear regression also perform relatively well with a MAE of approximately 0.113 and R^2_{adj} of 0.55). On average, parametric models performed worse than ML models with a drop in R^2_{adj} of about 0.14. In addition to Table 3-3, Figure 3-4 shows error plots, comparing the predicted and the actual water consumption values. In particular, the predicted values for GBR are more closely scattered to the standard regression line than the other models. Moreover, the scattered values of the other models (especially parametric models) have a slightly lower tangent to the fitted line than GBR—i.e., most models overestimate low consumption values and underestimate high consumption values. For instance, the linear regression model tends to overestimate lower values (approximately bottom 30% in actual water consumption) and underestimate higher values (approximately ranging from the median to below the top 10%). This is partly because linear parametric models have strict assumptions of the error terms and the relations between independent variables. In general, they assume that covariance between independent variables are zero, and they also assume that the error terms follow normal distributions with a mean of zero. These assumptions are likely to partly account for the poor prediction, especially when the dimensionality of variables is high (i.e., a large number of independent variables).

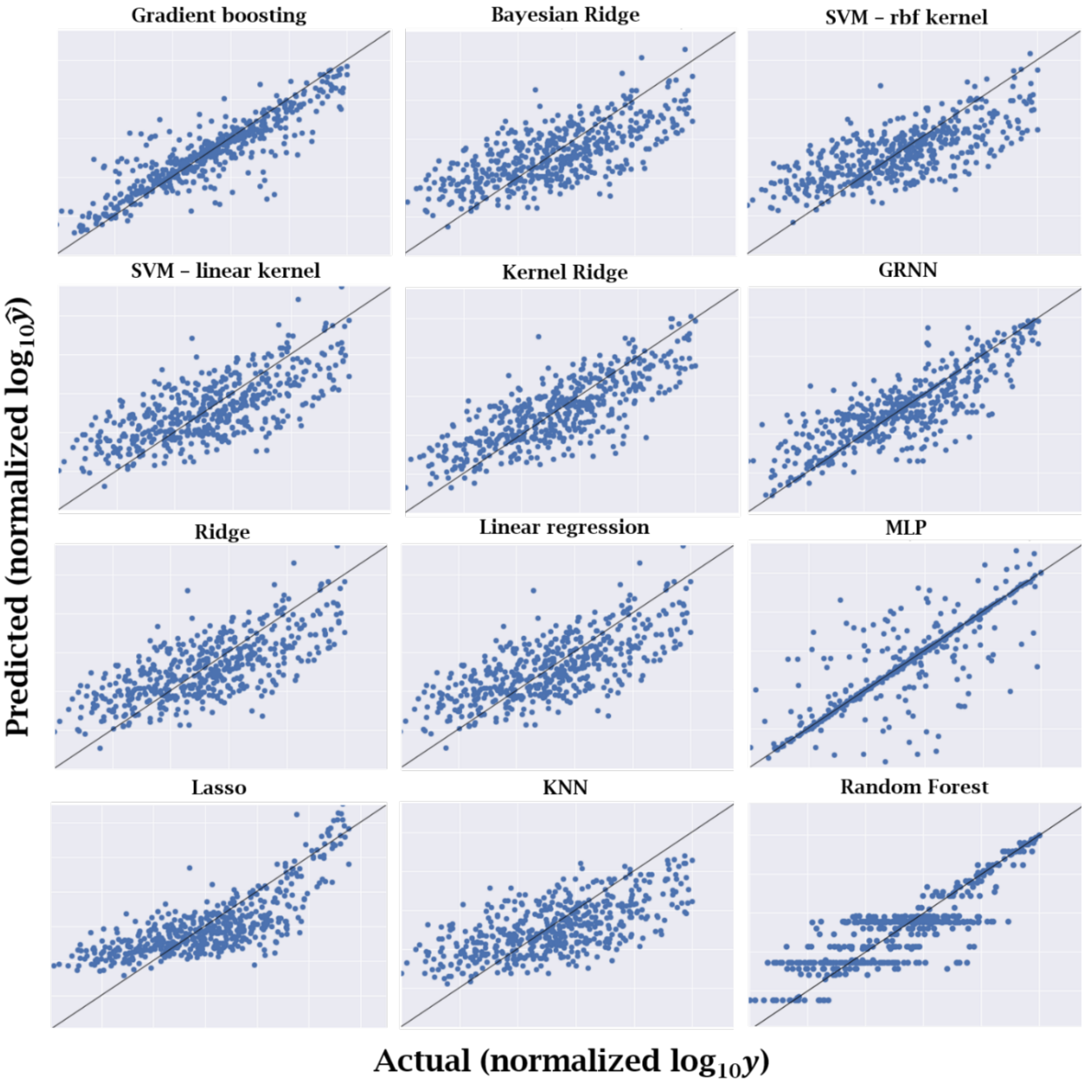


Figure 3-4. Comparison of error plots for REU 2016 specific scenario (the horizontal is the normalized $log10$ of actual water consumption and the vertical axis is that of predicted consumption values)

In addition, MLP shows comparatively poor predictive performance as many values are underestimated or overestimated across the entire range of consumption values. Although similar patterns are seen in other models, MLP shows to be more sensitive. It may due to the fact that MLP with a backpropagation process generally requires large datasets, preferably at least 10 times larger than the number of weights in the network structure (Anthony and

Bartlett 2009; Iyer and Rhinehart 1999). In this study, the MLP model only has 531 observations, while there are 19 independent variables. Furthermore, the MLP model shows the largest variation during the CV (standard deviation with ± 0.18 in R^2_{adj}), and it implies that it may not be optimized to the global minimum because of the algorithmic characteristics of GDM.

Table 3-3. Cross-validation results of all models using all variables available in the REU 2016

Statistical regression techniques	**MAE** ^{N*}	**MSE** ^{N}	$\mathbf{R^2_{adj}}$
GBM regression (GBR)	0.098 ± 0.01	0.017 ± 0.01	0.69 ± 0.09
Random Forest regression (RF)	0.099 ± 0.02	0.017 ± 0.01	0.64 ± 0.10
SVM with rbf kernel (RBF-SVM)	0.110 ± 0.02	0.018 ± 0.00	0.62 ± 0.08
Bayesian Ridge regression (BRR)	0.111 ± 0.02	0.018 ± 0.02	0.61 ± 0.10
GRNN regression (GRNN)	0.111 ± 0.01	0.019 ± 0.01	0.60 ± 0.11
Kernel Ridge regression (KRR)	0.112 ± 0.01	0.021 ± 0.00	0.58 ± 0.07
Ridge regression (Ridge)	0.113 ± 0.01	0.021 ± 0.01	0.55 ± 0.07
SVM with linear kernel(linear-SVM)	0.113 ± 0.02	0.021 ± 0.00	0.55 ± 0.07
Linear regression	0.113 ± 0.02	0.021 ± 0.01	0.54 ± 0.07
MLP regression (MLP)	0.115 ± 0.03	0.023 ± 0.02	0.52 ± 0.18
Lasso regression (Lasso)	0.116 ± 0.02	0.025 ± 0.01	0.49 ± 0.05
KNN regression (KNN)	0.119 ± 0.02	0.029 ± 0.01	0.41 ± 0.06

* MAE ^{N}, MSE ^{N}, and R^2 are calculated from the standardized data.

3.6.2 General Scenario: including 6 variables

Similar to Table 3-3 for the REU 2016 scenario, the performance of the 12 statistical models for the general scenario is shown in Table 3-4. The model performances are systematically lower for the general scenario compared to the REU 2016 scenario, which is expected since fewer variables are used. The results also show that the performance gaps between the ML and parametric models increased significantly. For instance, the R^2_{adj} values for the linear regression decreased from 0.54 to 0.33 (a 41% decrease) in contrast to the R^2_{adj} values for GBR that only decreased from 0.69 to 0.60 (a 13% decrease). In addition, the error plots in Figure 3-5 show that GBR shows similar patterns with Figure 3-4, and the number of under- and overestimated samples have only increased marginally. Consequently, these results suggest that the independent variables can still explain residential end-use water consumption behaviors relatively well. In particular, household type and size, and climate factors appear to significantly affect water consumption. The results also show that MLP now performs relatively well in the general scenario, with only a 0.04 decrease in R^2_{adj} and a 0.016 increase in MAE. This is partially because the general scenario includes fewer variables than the REU 2016 scenario, thus requiring the training of fewer weights (which is important as mentioned before).

Table 3-4. Cross validation result for the general scenario that only includes publicly available variables (7 variables)

Statistical regression techniques	**MAE** ^{N}	**MSE** ^{N}	R^2_{adj}
GBM regression	0.128 ± 0.01	0.026 ± 0.01	0.60 ± 0.13
Random Forest regression	0.130 ± 0.01	0.029 ± 0.01	0.51 ± 0.18
GRNN regression	0.131 ± 0.02	0.030 ± 0.01	0.50 ± 0.11
Kernel Ridge regression	0.131 ± 0.02	0.030 ± 0.01	0.49 ± 0.12
MLP regression	0.132 ± 0.02	0.031 ± 0.01	0.48 ± 0.14
SVM regression (linear-SVM)	0.134 ± 0.02	0.033 ± 0.01	0.44 ± 0.10
KNN regression	0.135 ± 0.02	0.034 ± 0.01	0.43 ± 0.10
Ridge regression	0.137 ± 0.02	0.036 ± 0.00	0.39 ± 0.11
Bayesian Ridge regression	0.137 ± 0.02	0.036 ± 0.00	0.37 ± 0.11
SVM regression (RBF-SVM)	0.139 ± 0.02	0.037 ± 0.00	0.34 ± 0.11
Linear regression	0.140 ± 0.02	0.037 ± 0.01	0.33 ± 0.11
Lasso regression	0.140 ± 0.02	0.037 ± 0.01	0.33 ± 0.08

Finally, in predictive modeling, it is generally desirable to gain an appreciation for the contribution of each variable; i.e., which independent variable contributes the most to explain the dependent variables? For instance, the top-ranked model, GBR, has the potential to measure the relative importance of each variable, generally referred to as the variable importance (VI). The results of VI can be explained as the predictive power of independent variables. Based on the VI in GBR, income, household size, parcel area, the existence of outdoor properties, and climate factor possess a higher contribution to predicting household water consumption. This also implies that the variables used in the general scenario are sufficient to provide an acceptable prediction performance.

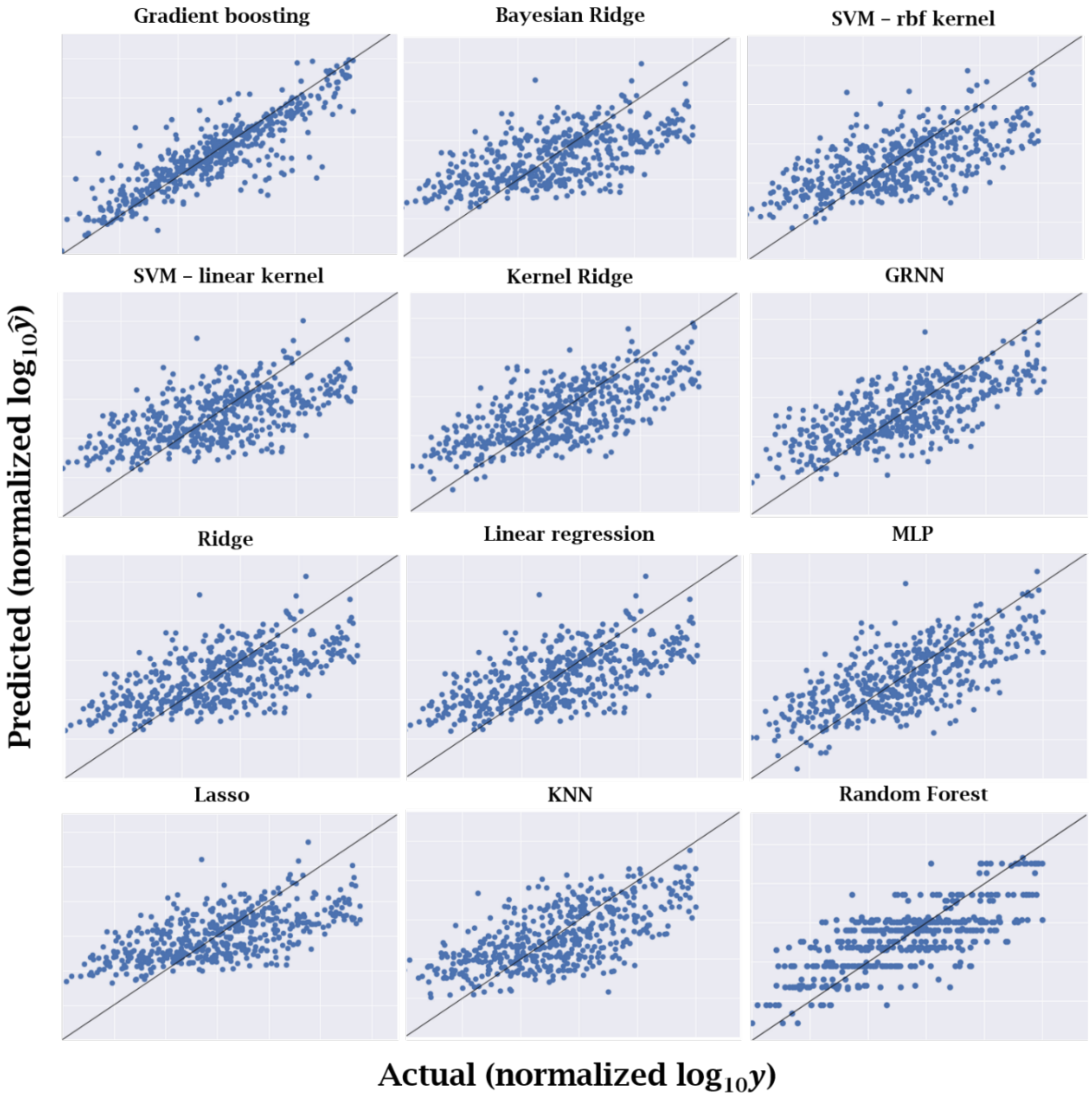


Figure 3-5. Comparison of error plots for general scenario

3.7 TECHNICAL DISCUSSION

Generally, the results demonstrate that ML models outperform parametric linear models such as linear regression, which is greatly thanks to the algorithmic differences between the two families of techniques. For instance, GBR, SVM, KRR, RF, and ANN models are primarily designed to capture nonlinear and complex relationships between variables in regression problems. For that, they perform stochastic local optimization (e.g., kernel, boosting, bagging) rather than single global optimization (Friedman et al. 2001). Thus, they are likely to decrease biases during the estimation process, and they can provide more accurate predictions than linear models.

Nonetheless, this local nonparametric learning process may generate overfitted models that can have high variances. For instance, ANN models using a high-dimensional input dataset tend to be overfitted since they have too many weights that need to be optimized. These overfitting issues can be seen in any ML models. The simplest way to mitigate them is to set an early stopping rule that is widely used to interrupt the repetitive learning of a machine (Robert 2014b). When it comes to algorithmic features, regularization and shrinkage methods are added to the ML models to avoid overfitting (Friedman et al. 2001). Specifically, a regularization term within some models (e.g., GBR, Lasso, Ridge, KRR) alleviates problems related to outliers (e.g., high biases) and high-dimensional inputs (e.g., high correlation) by introducing penalties, while balancing the trade-off between the variance and bias. In addition to the regularization, GBR and ANN models contain a shrinkage parameter that can also control variances by sacrificing some biases. For example, this is applied to hyperparameters (i.e., weights) that can control skewed variables and outliers. These algorithmic features partly substantiate the performance gaps between ML models and parametric models used in this study, and GBR, in particular, possesses all the features mentioned above. In other words, ML models can be more appropriate for predicting residential water use, especially for data such as REU 2016 that inherently exhibit high-variance (i.e., detrimental outliers).

In predictive modeling, the applicability of a model is also an important performance criterion; i.e., whether a model can easily be applied for other municipalities or utilities. The availability of data on daily water use can vary dramatically across geographical locations, however. For instance, some cities may have access to longer time-series daily water use than the REU 2016 datasets that are only available for 12 days. This longer data may include unobserved heterogeneity across households and other unobserved effects. To enhance future applicability, our modeling approaches are designed to alleviate these effects by controlling variables (a.k.a., covariates) that can incorporate seasonal and climate-related effects that affect water use. In addition, the top-ranked model, gradient boosting machine (GBM), possesses an algorithmic ability to handle heterogeneity issues, which is initially built in rule-based models (see details in "Methodology" section). Due to its algorithmic properties (e.g., nonparametric and rule-based properties), it is fundamentally suitable to handle multi-level data (i.e., panel data) to incorporate mixed effects (e.g., heterogeneity across observation) (Friedman 2001; Friedman et al. 2001). In addition to this algorithmic characteristic, GBM includes boosting machine processes that continuously "forgive" poorly learned observations or samples in a single estimator (e.g., tree) by using multiple estimators. This is one of the main technical reasons why GBM performs best among all modeling methods tested in this study. Therefore, the modeling performance will be consistent when the data even gets longer than what is used in this chapter.

Despite a high degree of prediction performances, the acknowledged drawback of many machine-based algorithms is their lack of interpretability, unlike parametric models in which a domain expert can validate the parameters estimated. To address this interpretable issue in ML models, several statistical measures exist (Doshi-Velez and Kim 2017b; Samek et al. 2017). Among numerous measures, a model-agnostic approach (i.e., not specific to a particular algorithm) is generally preferred since it can be applied to any predictive modeling process using ML models (Friedman et al. 2001; Lundberg and Lee 2017; Molnar 2019; Ribeiro et al. 2016). For example, and as provided in this chapter, GBR is able to identify the magnitude of the contribution of each variable by measuring the reduction in the overall error (i.e., bias and variance), called Variable Importance (VI). Nonetheless, VI cannot represent the sensitivity of variables on the dependent variables. In contrast, the marginal effect of an independent variable on the predicted values of a learned model can also be examined with ML models and is generally referred to as "partial dependence" (Friedman et al. 2001; Natekin and Knoll 2013; Semanjski and Gautama 2015). Although out of the scope of this chapter, these interpretable features are straightforward. Not only do they provide valuable insights into the performance of a model, but they can also help determine effective policies by assessing the contribution of individual variables and therefore help municipalities and utilities make better long term and short-term decisions.

3.8 CONCLUSION

This chapter aimed to test two technical challenges in water demand modeling: (1) which modeling technique is most appropriate and (2) how much data and what variables are required for learning an acceptable model. Specifically, the performance of 12 statistical learning algorithms including parametric and nonparametric models was investigated to model household end-use water demand (i.e., the cross-section of average daily water use), while taking into account two data scenarios. For the general scenario, only selected 6 variables were intentionally kept because they are commonly available in public micro databases, thus accessible to all cities and water utilities.

The results for the REU 2016 scenario indicate that nonparametric machine learning models (ML models) perform better than parametric linear models; specifically, GBR performed best. Furthermore, MLP showed relatively poor accuracy and the largest variation during the CV, although it reportedly performed well in previous studies. This is likely due to the fact that GDM may have issues finding optimal solutions while minimizing loss functions when the dimensionality of the input data (i.e., the number of variables) is small. In the general scenario, ML models perform adequately as well despite the data constraints (i.e., 6 variables), and GBR, here again, performed best. In contrast to the REU 2016 scenario, the performance gaps between the ML models and the linear models were even wider. In addition, linear models less accurately predicted under- and overestimated samples compared to the REU 2016 scenario.

The findings in this study can fill important technical knowledge gaps in predicting household water demand. Moreover, this study can be useful to municipalities and utilities that can adopt the same techniques (e.g., gradient boosting) on their own dataset to predict water demand and to infer the importance of individual variables in their area. In order to further improve water demand prediction accuracy, future work can focus on simulating datasets that can provide more information with utilities to better capture household and individual water consumption behavior. In addition, as mentioned in the technical discussion, a single metric, such as predictive accuracy, is often not enough to be able to develop effective policies. Instead, learned models should have both predictive power and be interpretable to fully take advantage of the usability and adaptability of models in the future.

4 Modeling Discrete Choice Behavior in the Framework of Probabilistic Graphical Models

4.1 INTRODUCTION

Choice behavior is statistically random in modeling, containing a significant amount of uncertainty that is a consequence of several factors (Kahneman and Tversky 2013; Koller and Friedman 2009; Tversky and Kahneman 1974). The real-world problems can be rarely determined with certainty based on our limited information since the ability to observe problems and world is limited. Uncertainties are therefore prevalent in an individual decision-making process, and these become worse from the viewpoint of researchers when modeling choice behaviors for population or sub-population. Uncertainty is required to be investigated in modeling tasks not only to build a more realistic model but also to obtain more meaning conclusion—e.g., confidence about the models' predictions. To handle uncertainty issues in modeling tasks, recent machine learning (ML) techniques, in particular, are adopting the concept of probability theories to quantifying, handling, and manipulating uncertainty. In this sense, this chapter aims to provide a modular probabilistic modeling framework that can be complied with any modeling techniques, which is also being able to not only capture uncertainty related unobserved factors (e.g., heterogeneity between individuals) but also quantify uncertainty in behaviors (e.g., level of confidence for conclusions). Toward this goal, we adopt the concept of probabilistic graphical models (PGM) and Bayesian inference that leverage potential opportunities by combining data-driven likelihood in the era of Big data, prior beliefs, and machine-based repetitive computation.

The core interest of modeling tasks has been the development of modeling techniques, including algorithmic and technical components, to handle and minimize uncertainty (e.g., heterogeneity) about given problems with higher flexibility (Bishop 2013; Train 2009). In this sense, methodological concepts and techniques that involve these tailored algorithms often become more complicated than already existing models, and researchers are left with a deluge of modeling algorithms, as well as various nomenclature. As a consequence, the traditional paradigm in modeling begins with selecting an appropriate method, then adapting the selected method to a given problem and relevant data to learn patterns (Bishop 2013; Olson et al. 2017). Selecting an algorithm is therefore far from trivial and even challenging for researchers, especially for whom have few backgrounds in learning approaches such as parametric and nonparametric (e.g., ML techniques). Furthermore, the selection of algorithm becomes more challenging since dilemmas are prevalent among the modeling techniques in both approaches—e.g., interpretability and prediction accuracy (see details in Lee et al., 2018). In these circumstances, it is useful to have a modular framework that enables to accommodate various algorithms in modeling design to effectively not only address uncertainty issues but also alleviate the existing dilemmas for researchers.

In addition, the estimation process applied to a family of random utility model (RUM) has been dominantly based on the likelihood principle that can be certainly fallen with the probabilistic paradigm, but it is not completely probabilistic approaches for the application of probabilistic reasoning to some extent (Jordan 2003; Murphy 2012). In particular, the inference based on maximum likelihood estimation (MLE) and maximum a posteriori (MAP) is a point estimate (a.k.a., frequentist statistics) instead of returning interval estimation of unknown quantities—i.e., quantifying uncertainty. Although the frequentist statistics are conceptually and computationally appealing and have proven to be successful in several applications, it is somehow limited in measuring uncertainty about any outcomes from the estimation. Probabilistic approaches based on Bayesian perspectives primarily aims to quantify uncertainty in modeling through the estimation of probability distribution by adopting probability theory (Murphy 2012). Specifically, Bayesian approaches are to estimate a joint probability distribution over a set of random variables (Friedman et al. 2001; Ng and Jordan 2002). In Bayesian settings, all unknown quantities are treated as random variables (Jordan 2003). Accordingly, probabilistic approaches inherently include the stochasticity of each variable, which can be powerful to measure uncertainty. In real-world cases, however, it is often impossible to explicitly represent all possible relations when the given behavior information is high-dimensional. Put differently, although one may be able to specify the joint distribution through intense computing, the process would be inefficient since many relations are trivial and not considered by people contemplating to make decisions (Koller and Friedman 2009). Instead, to efficiently estimate the joint distribution for high-dimensional data, declarative representations will be useful for identifying some probabilistic relationships between random variables as well as other modeling components (e.g., Markov properties) to reduce the number of possible combinations with the application of probability theory (Jordan 2004; Koller and Friedman 2009).

One way to do this is by representing complex probabilistic models and their distributions compactly under a graphical structure and effectively utilizing this structure to answer queries. This declarative and diagrammatic representation of modeling tasks is commonly called probabilistic graphical models (PGM) (Bishop 2006b; Jordan 2003, 2004; Koller and Friedman 2009). PGM uses a graph-based representation to compactly and intuitively express the given modeling problems that combines probability theory and graph theory, and the PGM has highly advantageous to represent complex distributions over high-dimensional information and spaces. Specifically, the fundamental idea of PGM enables to build a versatile module for dealing with two common problems that often occur during the real-world application of complex tasks—uncertainty and complexity (i.e., a high-dimensional data). PGM framework (e.g., directed acyclic graph, DAG), in particular, describes the probabilistic relationships between random variables, and this graph is useful to define a high-dimensional joint distribution as a product of distributions over smaller subset of random variables—i.e., factorized joint distribution using probability theory such as product and sum rules (Bishop 2006b; Jordan 1998). The factorized joint distribution using probability theory can yield any probabilistic distributions by taking advantage of Bayesian statistics.

The aim of this chapter is to provide a flexible modeling framework that takes advantages of the power of data-driven approaches by using probability theory (i.e., Bayesian statistics) and declarative graphical representation (i.e., PGM) to address uncertainty issues in modeling. Thus, this chapter mainly adopts the PGM framework with Bayesian inference (PGM-B), which can be classified as model-based approach to machine learning (MBML). Modeling in PGM-B is antithetical to the traditional modeling paradigm, where it generally starts with the dilemma of selecting an algorithm when learning and recognizing patterns revealed from the data (Olson et al. 2017). Specifically, the modeling in PGM framework can provide modeling palette that represents our given information and knowledge, which can be separated from algorithms used in inference. Thanks to this separation, we can even include any algorithms to make inference in this framework such as the RUM and ML techniques.

Furthermore, we can improve this modeling palette for a specific domain without modification of probabilistic reasoning algorithms continuously. In this chapter, we mainly focus on the applicability of the PGM framework with Bayesian inference to discrete choice model (DCM) based random utility maximization theory. Although Bayesian statistics have adopted in few studies, the combination of PGM and Bayesian inferences is very limited in the realm of transportation. To infer intractable probability distribution (i.e., posterior distribution of interest), we adopt variational inference (VI) algorithm that leverages ML for approximating probability distributions (Jordan et al. 1999). To demonstrate the applicability of PGM, we theoretically review popular RUMs and convert these models into compatible PGM framework. These converted models are further used to infer travel mode choice behaviors and also validated through cross-validation method.

After this introduction, the theoretical and technical motivation of our studies are discussed, and the key theoretical components in PGM and Bayesian inference will be reviewed in section 3. Also, we will review five RUMs in section 4. In section 5, the application of PGM framework for RUM will be discussed. Section 6 will show the results of the Bayesian inference of PGM. Lastly, some discussions and conclusion will be discussed.

4.2 THEORETICAL MOTIVATION

4.2.1 Uncertainty in modeling

Most modeling tasks aim to learn users' (can be also agents', decision-makers') behaviors from the observable information (i.e., partial evidence), and algorithms (e.g., statistical learning including machine) are used to investigate reasons for a certain specific case. The built model from the evidences and complied algorithms is further used to predict the corresponding actions and conclusions that a person or a system possibly make. This series of modeling tasks can be called “reasoning” or “reasoning system”. When reasoning behaviors, *uncertainty* is a significant issue that has placed burdens on the researchers. Uncertainty has been widely acknowledged to the fundamental issue that is required to be addressed somehow since it may lead us to make false conclusions. There are different forms of uncertainty, including (1) noise in the observed data (i.e., errors and unobserved information), (2) model parameters, and (3) model structure (e.g., relationships between variables), which are inevitable in modeling real-world problems. The formal uncertainty can be alleviated by accommodating more knowledge (e.g., prior domain beliefs and assumptions) and data as well as tailoring modeling techniques. The latter two uncertainties can be addressed by adopting the concept of probability theory—i.e., infer the distribution of parameters based on the assumed structure. As problems and given data that are combined with human behaviors become more complex (e.g., high-dimensional), addressing uncertainty issues in modeling becomes critical. Thus, the history of modeling development can also be viewed as a series of efforts overcoming the amount of uncertainty issues, which aim to build more realistic models.

4.2.2 Dilemmas in modeling

In general, a reasoning system based on statistical learning is required to be built with flexible model structure and algorithms to address uncertainties—i.e., demystify users’ behaviors from partial evidence. In particular, heterogeneity in users’ behavior is one of the imperative issues in field of DCM, which is usually resulted from variations among individual preferences, attitudes, and other sentimental factors. For instance, if heterogeneity in users' behavior is ignored in behavior model, it can cause biased estimation and misleading conclusions (e.g., policy impacts) (Yuan et al. 2015). Even this issue is far from trivial, since these behavioral variations among users are not easily noticeable and handled. In this context, researchers have been focused on developing and supplementing modeling techniques to better investigate hidden patterns by enhancing the flexibility of the models and their estimation.

In the realm of transportation field, for instance, uncertainty about choices such as inter-individual taste heterogeneity has been a key concern in modeling tasks (Bhat 1998; Train 2009; Vij and Krueger 2017), which has been dominantly addressed by the family of RUM (i.e., parametric approaches). The early models in RUM had some limitations, which generally simplify the real-world problems for mathematical convenience (e.g., constraints). Thus, it was not flexible enough to reason real-world problems (Train 2009). Although some limitations resulted from mathematical simplifications were already recognized at the time, it is either not feasible or possible to overcome them—e.g., approximating intractable likelihood. Thanks to the advancement in computational ability, it is becoming increasingly feasible to learn more information from given data by addressing unapproachable components before such as the family of mixed logit and probit. In particular, these algorithmic and technical advancements alleviate some aspects of uncertainties not only about unobserved information from the given information (e.g., latent structures) but also about stochastic properties in models' predictions (e.g., random coefficients). Although these parametric approaches are intuitive and more easily interpretable based on strong theoretical backgrounds—i.e., random utility maximization, it is limited to identify hidden interactions between variable and assumptions for capturing high degrees of complexity (e.g., nonlinearity) in the given information.

In the era of Data science and Big data, many opportunities are available for us to exploit that relatively hassle-free to demystify problems from given information (i.e., data), and their applications have proven to be successful in many real-world problems. Specifically, machine learning (ML), mostly nonparametric approaches, have recently adopted in many domains to handle uncertainty issues since the power of data-driven approaches enables to possess flexible algorithmic properties (e.g., fewer predetermined assumptions) than the parametric models and myriads of local optimization processes thanks to their machine-based repetitive computation process. Many ML models, however, are less interpretable than parametric approaches such as RUM due to their reliance on repetitive computation process, and it is also difficult to incorporate domain knowledge. In these circumstances, researchers face a variety of technical dilemmas when selecting modeling techniques. The dilemma has been the most pressing issues to the researchers since the traditional paradigm in modeling tasks generally starts with selecting a technique.

4.2.3 Probabilistic modeling

Probabilistic approaches based on Bayesian statistics aims to address uncertainty issues in modeling by the coherent use of probability theory through the estimation of probability distribution (Murphy 2012). The model estimation with likelihood principle that has been dominantly used in the field of transportation can be certainly categorized into the probabilistic paradigm, but it is not completely probabilistic reasoning (Jordan 2003; Murphy

2012). In particular, the inference in the likelihood principle such as maximum likelihood estimation (MLE) and maximum a posteriori (MAP) is a point estimate of an unknown quantity by computing fixed parameters (e.g., fixed mode and variance in random coefficients distribution) for the predetermined distribution (a.k.a., frequentist statistics). These ways of inference are somehow limited in representing and manipulating uncertainty about outcomes—i.e., full posterior distributions. This limitation may sometimes result in overfitting issues that make our predictive distribution to be overconfident (Murphy 2012).

As mentioned above, uncertainty has been acknowledged to an inevitable aspect of modeling tasks, and it is a consequence of various factors. Nonetheless, it is practically impossible to handle all uncertain aspects in modeling tasks, since the real-world problems can be rarely determined with certainty based on our limited information. Therefore, we need reasoning systems that must not only handle prevalent uncertainty from unobserved factors, but also measure the amount of uncertainty to be able to obtain more meaningful conclusion such as probable outcomes (Koller and Friedman 2009). This is due to the facts that the true answers of the world and related phenomena are inherently not deterministic, and it is not simply provided with discriminative values. In other words, models of real-world must not only handle prevalent uncertainty from unobserved factors, but also measure the amount of uncertainty to be able to obtain more meaningful conclusion—not only what is possible, but also what is probable. (Koller and Friedman 2009).

To handle any uncertainty in the application, Bayesian statistics has been widely adopted in many fields, however, the majority of applications for travel choice modeling are derived from frequentist statistics (Daziano et al. 2013). Although there are few excellent applications of Bayesian approaches (Daziano and Bolduc 2013), some studies applied pseudo-Bayesian statistics. It may due to the fact that the goal of frequentist approaches prefers to have an "objective" statistical outcome for the estimation, thus, they avoid to using Bayes rule by simultaneously considering prior probabilities and likelihood to calculate posterior (Jordan 2003). Put it differently, probabilistic reasoning for inferential problems (i.e., probabilistic approaches) can be completely fulfilled by Bayesian statistics involving the notion of "subjective". This concept of subjective probabilities, however, is loosely analogous to the behavioral assumption of RUM that assumes choices and associated utilities are random (Train 2009) (Daziano 2013). Put it differently, if the goal of a model does not merely focus on the discriminative outcome (e.g., predicted choice), Bayesian statistics are preferred to fully measure uncertainty issues in travel choice behaviors because it bridges frequentists and Bayesian through Bayes rule (Jordan 2003). In particular, the joint distribution with Bayesian statistics provides consistent quantification of uncertainty by coherently updating prior distribution and evidence (i.e., likelihood from given data), and evaluating the posterior distribution. Conversely, Bayesian statistics can also be manipulated and calculate the quantities that are the interest of frequentist.

During the estimation of the joint distribution with Bayesian statistics, however, it is often challenging to explicitly represent all possible probabilistic relationships between variables and estimate associated distributions. This challenge becomes exacerbated when the given behavior information is high-dimensional.

4.2.4 Proposed modeling approach

This chapter aims to provide a flexible probabilistic modeling framework that is able to accommodate a wide range of existing algorithms, in which uncertainty is intuitively represented as well as addressed, and quantified. In addition, this framework can give useful insights into modeling DCM in the field of transportation, while tackling the dilemmas in the era of Big Data and artificial intelligence (AI). Toward this goal, we use the combination of two solid probabilistic modeling concepts: (1) probabilistic graphical models (PGM) and (2) Bayesian inference (PGM-B). This combination is also called model-based approach to machine learning (MBML) that is antithetical to the traditional modeling paradigm, where it generally starts with the dilemma of selecting an algorithm when learning and recognizing patterns in the data (Olson et al. 2017). The following sections will discuss methodological backgrounds related to PGM framework and Bayesian inference in this framework.

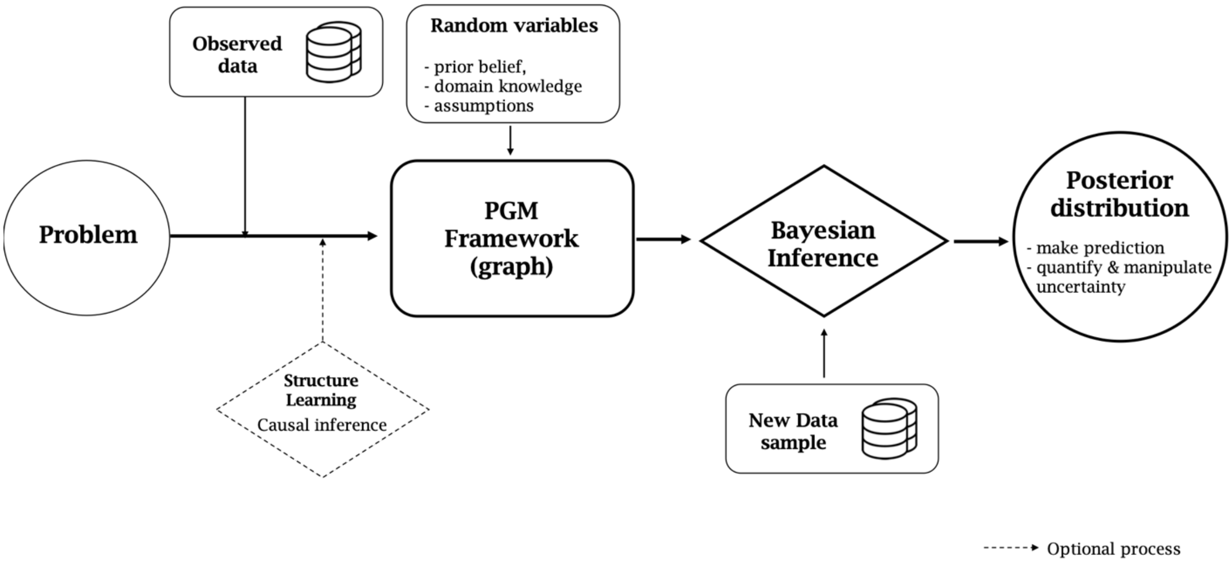


Figure 4-1. General modeling process of PGM-B

4.3 METHODOLOGICAL BACKGROUND OF PGM-B

PGM-B can be separated into two sub-process: models and inference (a.k.a., reasoning systems) (Koller and Friedman 2009) that can be fulfilled by the framework of PGM. For given decision-making problems, for instance, a process in constructing models begins with considering all kinds of quantities governing the data and possible modeling components based on our belief and domain knowledge (i.e., priors, clusters, hierarchies), and these are treated as random variables. Then, these variables are intuitively represented as a declarative graphical structure and our beliefs (e.g., assumptions and domain knowledge) about how the observed data generated is also explicitly described along with the graphical structure—i.e., generative process. Once the models (i.e., PGM framework) are constructed, now the goal is to answer any probabilistic queries of interests (e.g., parameters). Specifically, we infer the posterior distribution of model parameters (unknown quantities of our interests) after observing the new evidence (e.g., observed responses, y) through the notion of Bayesian statistics. Figure 4-1 presents the overall modeling process of PGM-B. Specifically, the optional process, structure learning, is used to learn the causality of a set of variables by using algorithms with domain knowledge. This will not be covered in this chapter, however, it will be discussed in the later sections (i.e., see details in section 7).

In this section, we will discuss the theoretical background of PGM framework: (1) representation, (2) factorization of joint distribution, and (3) generative process. As for the inference process, (4) variational inference (VI) method (i.e., mean-field approximation) will be discussed, used to approximate posterior distribution.

4.3.1 PGM framework

4.3.1.1 Representation in PGM

PGM is a modular modeling framework that represents statistical modeling problems that connect graph theory to probability theory and that provide effective ways to handle uncertainty with Bayesian inference (Jordan 1998, 2003; Pearl 2014). The PGM framework provides an intuitive and compact way of representing the structure of a probabilistic model, which give us insights about the properties of the model. With the PGM, for instance, researchers can clearly articulate what kinds of quantities (e.g., variables, errors, hidden structures) are governing the data and embody complex modeling process from simpler components such as hierarchies, clusters, sequences, and others. The PGM can have many different types of graph structures, however, and this chapter focuses on the Directed Acyclic Graph (DAG) structure that represents conditional dependencies (i.e., causal relations) between variables with directional edges (also called a Bayesian network). PGM (G) consists of a set of nodes (V) and a corresponding edge sets (E), that can be expressed as follows:

$$G = (V, E) \tag{1}$$

In the graph, the nodes (large circles) represent random variables, and the corresponding directional edges depict probabilistic dependencies between the variables, which correspond to conditional probability distributions. In Figure 4-2 (b), shaded nodes (black circles) indicates observed variables, whereas unshaded nodes (white circles) represent unobserved (a.k.a., latent) variables. Moreover, the outer box (harnessing x_n and z_n) marked with N in the lower corner of the plate defines the repeated applications that depict the relationship between two random variables for N times (i.e., from 1 to N). Lastly, small black circles describe the pre-fixed parameter (a.k.a., hyperparameters), which is optional to be presented in the graph to concise visualization.

4.3.1.2 Factorized joint probability distribution

Theoretically, PGM enables the compact representation of joint probability distributions between random variables by means of network structure in a factorized way, while taking advantages of independence properties (Koller and Friedman 2009; Sucar 2015). Accordingly, the graph structure gives us an intuitive and comprehensive framework to solve the modeling problems at hand, and probability theory is employed to quantify the uncertainty (e.g., inherent stochasticity) that comes with the problems themselves and relevant data. Based on the graph G, probability theory (i.e., sum and product rules) and linear algebra are required to specify a joint distribution and infer different probabilistic queries (i.e., marginal or conditional probabilities) based on the structure for a graph (Bishop 2006b; Sucar 2015). For example, a joint probability for a certain behavior problem given a set of variables is presented in Figure 4-2 (a) and it can be expressed as the follows:

$$P\ (z_1, z_2, z_3, z_4, x_5, x_6, x_7) = p(z_1)p(z_2)p(z_3)p(z_4|z_1, z_2, z_3) \times$$

$$p(x_5|z_1, z_3)p(x_6|z_4)p(x_7|z_4, x_5) \qquad (2)$$

Equation (2) factorizes variables according to the given graph structure in Figure 4-2 (a). In particular, it gives intuitive and effective ways to answer any specific queries within this graph such as conditional probabilities (e.g., $p(y|x_3, x_4)$) and marginal probabilities (e.g., $p(y)$) by determining the probability of any given assignment to the set of variables (Bishop

2006b; Koller and Friedman 2009; Sucar 2015). The general case of the joint distribution with N number of nodes can be factorized as follows:

$$P\,(x_1, \ldots, x_N) = \prod_{n=1}^{N} P(x_n | pa(x_n)) \tag{3}$$

where, x_i is a variable and $pa(x_i)$ are the parents of x_i. Specifically, each variable is a probabilistic (i.e., stochastic) function of its parents, which is inherently encoded with a generative process. Specifically, the value (i.e., distribution) for each variable is determined by using a distribution that depends only on its parents.

4.3.1.3 Generative process

The generative process is to specify how data might have been generated from the model. The combination of factorized joint distribution encoded in the PGM, and the generative process is intuitively way to understand the given problem and data while incorporating our domain knowledge.

PGM representation in Figure 4-2 (b) describes a simplified Bayesian Gaussian mixture model as an example, which is defined by N numbers of observations, $n \in \{1, \ldots, N\}$ that are distributed over K mixture components (i.e., clusters), $k \in \{1, \ldots, K\}$. Each cluster is characterized by the center of Gaussian distribution μ_k with the variance of prior on the clusters σ_0^2. The latent variable z_n controls the allocation of each observation (i.e., $z_n = k$), thus, z_n is a realization (a.k.a., mixture assignment in mixture models) from a multinomial distribution with parameter π (a.k.a., mixture proportions in mixture models). In particular, π is assumed to be fixed in this case but it can be also characterized by a certain distribution and related parameters (e.g., Dirichlet process and concentration parameter) (Blei et al. 2003). The generative process of this model is succinctly expressed as follows:

$$\mu_k \sim \mathcal{N}(0, \sigma_0^2) \qquad k = 1, \ldots, K, \tag{4}$$

$$z_n \sim categorical(\pi) \qquad n = 1, \ldots, N, \tag{5}$$

$$x_n \sim \mathcal{N}(\mu_{z_n}, \sigma^2) \qquad n = 1, \ldots, N. \tag{6}$$

where σ_0^2, σ^2, and π are assumed to be fixed. This process with PGM also helps us to identify variables that operate globally and locally. For instance, the cluster center (μ_k) and mixture proportion π is globally governing this model, whereas cluster assignment (z_n) only indicates the assignment of each observation, which is locally governing the model.

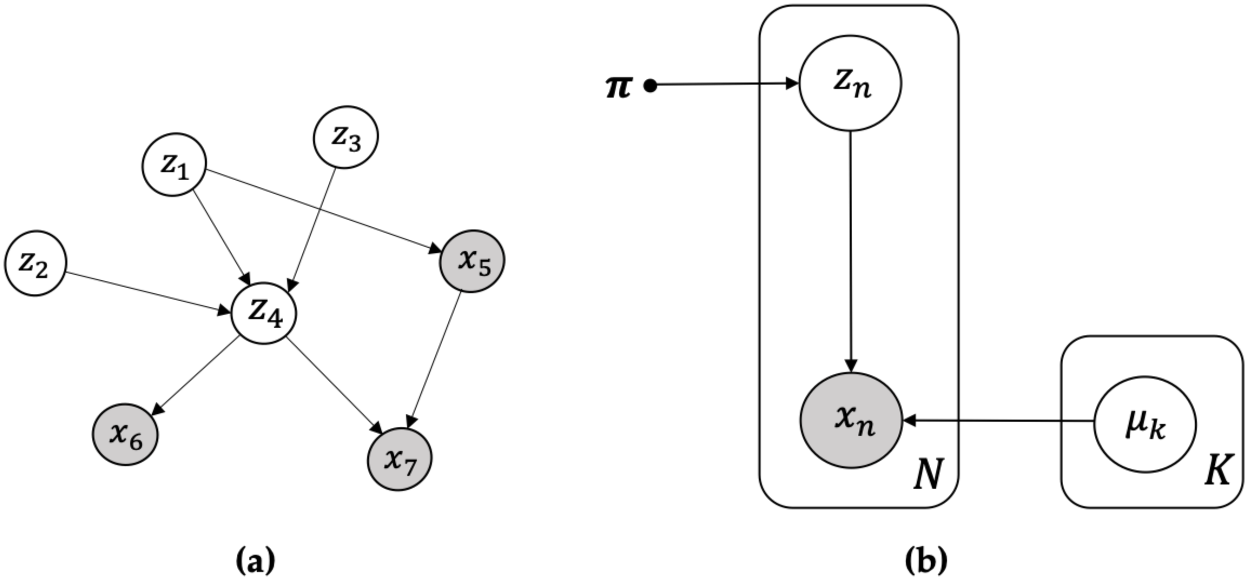


Figure 4-2 (a) graphical representation of variable (b) graphical representation of Bayesian Gaussian mixture model in PGM (shaded circles indicate observed variables; unshaded circles are unobserved (latent) variables and the square represents repetitive applications—i.e., a plate)

4.3.2 Inference in PGM framework

4.3.2.1 Bayesian inference

The goal of the model in the example of Figure 4-2 (b) is to compute the posterior distribution over μ and z, based on the joint distribution of all of given variables, assumptions, and parameters. This computation process is also known as Bayesian inference in modeling on PGM framework. In addition to quantifying uncertainty (mentioned in section *2.2.1*), Bayesian inference is technically preferred than the classical procedure such as frequentist approaches (e.g., MLE and MAP). In particular, Bayesian inference is not necessary to maximize of any function (Bishop 2006b; Train 2009). For instance, maximum simulated likelihood (MSL) function is often difficult to be computed in numerically since the maximum values are prone to be affected by starting values and can be reached to local maxima instead of global maximum. Without using frequentist approaches, Bayesian inference can provide inference results that can be examined and interpreted in the viewpoint of frequentist. This dual interpretation can be conducted by Bayes' rule:

$$p(z|x) = \frac{p(x,z)}{p(x)} = \frac{p(z)p(x|z)}{p(x)} \tag{7}$$

where the $p(z)$ is prior distribution of our interest (z) and $p(x|z)$ is likelihood distribution that is explained by data. The nominator in this equation is typically easy to evaluate for any configuration of the latent variable z. The model evidence $p(x)$ is normalization constant, which is the sum of the nominator over all possible configurations of the latent variable z. This can be also expressed:

$$p(x) = \int_z \; p(x,z) \tag{8}$$

This is the marginal likelihood of x. For example, if we specify the marginal likelihood of the example in Figure 4-2 (b) by taking advantage of the factorization, the marginal likelihood $p(x)$ is:

$$p(x) = \int_{\mu} \left(\prod_{k=1}^{K} p(\mu_k)\right) \prod_{n=1}^{N} \sum_{z_n} p(z_n|\pi) p(\, x_n|z_n, \boldsymbol{\mu}) d\mu \qquad (9)$$

To fully make inference on posterior distribution, the marginal likelihood is required to be evaluated. Specifically, the marginal likelihood is an integral over all possible values of model parameters that we are interested in, and this integral is often intractable (i.e., non-closed form).

4.3.2.2 Approximate inference: variational inference

Posterior distribution of our interest may not have closed-forms and need to be approximated by approximation inference methods. There are various approximation inference methods to solve non-closed form integration, in particular, Markov Chain Monte Carlo (MCMC) or Variational Inference (VI) are widely adopted in Bayesian inference. For instance, one of the core research questions of DCM is to approximate intractable probability densities (e.g., random parameters) and predict future choice behaviors. In PGM-B, this approximation problem becomes especially important, which modules all inference about unknown quantities through posterior estimation.

This chapter mainly uses VI, which leverages techniques from the machine learning to approximate probability densities for approximating probability densities (Wainwright and Jordan 2008). As an alternative strategy to MCMC sampling, VI has been widely used to large-scale data analysis since VI is more computationally efficient than MCMC, which is particularly important when data sets become large. As a loose analog to VI, the expectation-maximization (EM) algorithm has been widely used for approximation problems in DCM, which provides a point estimate rather than a probability density (i.e., posterior distribution)—which is the interest of Bayesian statistics (see details in this section later). Despite the difference in results between EM and VI, the fundamental approximation method is almost similar by minimizing the distance between the true values and approximated values (i.e., the closeness between two distributions)—e.g., Kullback-Leibler ($\mathbb{KL}$) divergence (Kullback and Leibler 1951).

VI is generally adapted to approximate the (conditional) probability densities of latent variables (z) given observed data (x), $p(z|x)$, using a simpler distribution. The main idea is to introduce a tractable variational distribution $q(z|v)$, with variational parameters (v), and to find the values of (v) that make $q(z|v)$ as close as possible to the true posterior $p(z|x)$. The goal of VI is to find the best approximation of $q(z|v)$, the one closest in $\mathbb{KL}$ divergence to $p(z|x)$. Specifically, $\mathbb{KL}$ divergence is asymmetrical measure of proximity between two distributions:

$$\mathbb{KL}\,(q\|p) \neq \mathbb{KL}\,(p\|q) \tag{10}$$

that is minimized when $q(\cdot) \approx p(\cdot)$. Based on the criteria of $\mathbb{KL}$ divergence, posterior inference can be expressed as:

$$q^{*}(z)= \arg \min_{q(z)} \mathbb{KL}\,(q(z)\|p(z|x)) \tag{11}$$

Unfortunately, the above equation cannot be directly minimized due to the model evidence term—i.e., the denominator in Bayes' theorem that requires marginalizing over the latent variables (i.e., unknown parameters of interests) to be computed (Wainwright and Jordan 2008). In particular, the $\mathbb{KL}$ is represented by:

$$\mathbb{KL}\ (q(z)\|p(z|x)) = \mathbb{E}[\log q(z)] - \mathbb{E}[\log p(z|x)] \tag{12}$$

where all the expectations are with respect to $q(z)$. Making use of the conditional probability formula, we can rewrite equation (12) as:

$$\mathbb{KL}(q(z)\|p(z|x)) = \mathbb{E}[\log q(z)] - \mathbb{E}[\log p(z, x)] + \log p(x) \tag{13}$$

where the model evidence term $\log p(x)$ becomes constant with respect to q. Therefore, instead of minimizing the $\mathbb{KL}$ (intractable), an alternative (tractable) function can be minimized:

$$\text{ELBO}\ (q) = \mathbb{E}[\log p(z, x)] - \mathbb{E}[\log q(z)] \tag{14}$$

This is the negative $\mathbb{KL}$ without an additive constant $\log p(x)$—i.e., the (log) evidence, which is called the evidence lower bound (ELBO). As its name suggests, ELBO indicates the lower-bound of the (log) evidence (see Figure 4-3), i.e. $\log p(x) \geq$ ELBO (q) for any $q(z)$. With the ELBO, we can geometrically interpret the relations between $\mathbb{KL}$, ELBO, and the (log) evidence. The ELBO is getting tight when $q(\cdot) \approx p(\cdot)$, in which case ELBO $\approx \log p(\cdot)$.

As mentioned earlier, EM and VI are loosely analogous to each other. In particular, the first term in equation (14) is the expected log-likelihood in the EM algorithm that is designed to solve maximum likelihood estimates for point estimation with latent features (Blei et al. 2016). In contrast to the VI, EM assumes that $\mathbb{E}[p(\log p(z|x)]$ can be calculated and uses it to further parameter estimation problems.

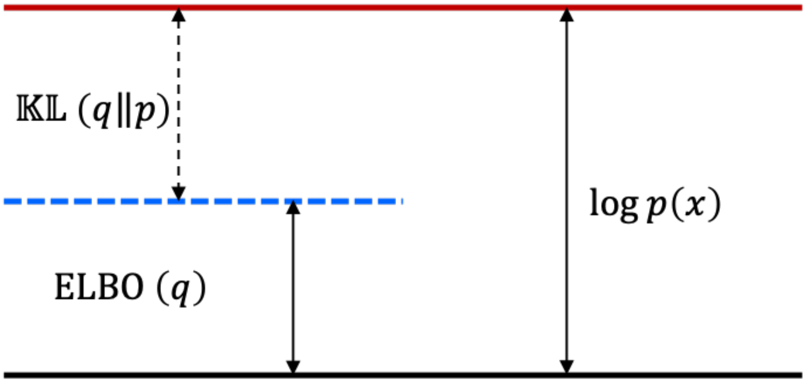


Figure 4-3. Geometric relationships between 𝕂𝕃, ELBO, and the (log) evidence (see details in Bishop 2006)

4.4 REVIEW OF MODELING IN DISCRETE CHOICE BEHAVIOR

In the previous sections, the theoretical background of key components for PGM-B is discussed. This section will investigate how does the framework of PGM-B modularize discrete choice problems, especially for travel choice behaviors. Although various kinds of modeling techniques are available to address discrete choice behaviors—e.g., parametric and nonparametric approaches, this chapter primarily derives PGM-B as for the application of random utility models (RUM) that are derived from the assumption of random utility maximization behavior. Specifically, modeling components (e.g., coefficients) in PGM-B are somewhat similar to the RUM, but the ways to define variables and their structures (e.g., relations) are different from each other. Besides, there are some discrepancies in the estimation process and results—i.e., frequentist and Bayesian. Thus, the following section will briefly review methodological backgrounds of DCM, especially for the family of RUM, and closely modularize those into PGM-B.

4.4.1 Discrete choice modeling with random utility maximization

Discrete choice models (DCM) fundamentally describe users' choice behaviors among available alternatives (i.e., set of options) (Train 2009), thus, the primary goal is to learn the users' decision-making process from the data ($\boldsymbol{x}$) that leads to the user's choice (y). Although detailed specification and terminology are different from the modeling purpose and associated data, behavior models can be expressed as the following function:

$$y = f(\boldsymbol{x}) + \ \varepsilon \qquad (15)$$

where $\boldsymbol{x}$ is a vector of observed variables (i.e., attributes, factors). Moreover, in practice, data itself is not perfect, and there exists unobserved data and some noise, which is presented by the error term, ε, in the model.

In DCM, observed and unobserved factors relate to the user's choice through the function expressed in equation (15). Due to the existence of the error term, ε, the user's choice cannot be predicted exactly (i.e., the choice is not deterministic). Instead of an exact prediction, DCM derives the probability of available choice by considering the error term as random with specific density. This derivation of choice probability can be expressed in a more useable form by entailing an indicator function, $I[\cdot]$:

$$P\ (y|x) = \int I[y] f(\varepsilon) d\varepsilon \qquad (16)$$

where the indicator function, $I[\cdot]$, returns binary values, either 0 and 1. For instance, $I[\cdot]$ takes 1 if the value of x and ε induces the user to choose a specific outcome, y, and 0 otherwise. Then, the conditional probability that the user chooses y becomes the expected value of this indicator function over all possible values of the error terms.

4.4.1.1 Random utility Theory

Discrete choice modeling under random utility theory (a.k.a., random utility model, RUM) is one of the dominant methodologies applied to address discrete choice behaviors, and it is derived from the assumption of utility-maximizing behavior by the decision-maker (Marschak 1950; Train 2009). In utility-maximization theory, a decision-maker, n, chooses an alternative, i, among J alternatives since the alternative i provides the greatest level of utility U with a decision-maker (Train 2009). In equation form:

$$U_{ni} > U_{nj} \; \forall \, j \neq i \tag{17}$$

Although RUMs are derived from the utility-maximization theory, the representation of models is analogous to any other models that describe users' behavioral process. Therefore, RUMs can also be represented by relating explanatory variables to the outcome (see equation (15)). Based on equation (16), this utility function, U, can be also classified into two parts: an observed and an unobserved utility. Similar to the form of equation (15), the utility function can be expressed as follows:

$$U_{ni} = V_{ni}(\boldsymbol{x}, \boldsymbol{\beta}) + \varepsilon_{ni} \tag{18}$$

where the observed (stated) utility, V_{ni}, is a value determined by a linear combination of the observed variables, which include covariates associated with both a decision-maker and the alternative presented to the decision maker. The unobserved utility (i.e., errors) for all alternatives, ε_n, on the other hand, cannot be observed by researchers but by decision-makers since data is incomplete, and there inherently exists the level of stochasticity— i.e., uncertainty. Therefore, researchers treat the unobserved terms as a random variable that follows an assumed probability density function (e.g., independently, identically distributed with extreme value).

For this random variable, the joint density of this unobserved utility for the n^{th} decision-maker (ε_n) can be denoted as $f(\varepsilon_n)$. This joint density function is also known as a "mixing distribution" in transportation discrete choice modeling studies. Technically, however, there are some misconceptions about the mixing distribution (see details later). In general, the probability that decision-maker n choose alternative i can be expressed as:

$$P_{ni} = \int \left[I\left(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \; \forall j \neq i\right)\right] f(\varepsilon_{ni}) \, d\varepsilon_{ni} \tag{19}$$

This is a marginalization over the unobserved utility (ε), which is the way of calculating the weighted average over different ε for each alternative. Similar to the unobserved utility, some observed utilities can be treated as random variables $f(\beta)$, which is known as random coefficients in a way of addressing random tastes heterogeneity. In this sense, the different types of RUMs largely depend on how these random variables and corresponding mixing distribution is specified and estimated (i.e., marginalization).

Each RUM has different assumptions about the mixing distribution— i.e., $f(\varepsilon)$, $f(\beta)$. Put differently, these assumptions are generally used to overcome the two modeling issues, but it is also made to achieve mathematical convenience for the calculation of integral. Logit and generalized extreme value (GEV) models (e.g., nested logit) intentionally make assumptions on the unobserved terms, which enables the integral in equation (19) to have a closed-form solution, so that the model parameters can be determined using maximum likelihood estimation (MLE). On the other hand, probit and mixed-logit have more flexible assumptions than the two previous models. For instance, probit assumes that $f(\varepsilon)$ takes a normal distribution. Mixed logit is derived by classifying the unobserved term ε into two components and assumes that one represents random tastes variation that follows a certain distribution determined by researchers, while the other follows an independently, identically distributed (iid) extreme value distribution.

The following sections discuss the different types of RUMs and their specifications. In particular, the family of mixed logit that addresses random tastes heterogeneity is discussed based on the types of mixing distribution.

4.4.1.2 Standard logit (conditional logit)

The simplest and easiest way to address a discrete choice problem is to adopt a logit model, which was developed to achieve tractable calculation. In particular, the logit model is derived by assuming a specific density distribution for the unobserved term (ε). This specific iid extreme value distribution is also called Gumbel and type I extreme value (i.e., zero mean and scale one). The density (f) and cumulative distribution (F) of the unobserved term in equation (19) are assumed to be:

$$f\left(\varepsilon_{nj}\right) = e^{-\varepsilon_{nj}} e^{-e^{-\varepsilon_{nj}}} \tag{20}$$

$$F\left(\varepsilon_{nj}\right) = e^{-e^{-\varepsilon_{nj}}} \tag{21}$$

With the iid extreme value on the unobserved term, the equation (19) can be expressed as:

$$P_{ni} = \int \left[\prod_{j\neq\, i} e^{-e^{-(\varepsilon_{nj}+ V_{ni}- V_{nj})}}\right] e^{-\varepsilon_{ni}} e^{-e^{-\varepsilon_{ni}}}\, d\varepsilon_{ni} \tag{22}$$

This logit choice probability of a decision-maker (n) choosing an alternative (i) results in a closed-form expression by using some algebraic manipulation:

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_{j=1}^{J} e^{V_{nj}}} \tag{23}$$

Since this choice probability is limited to the specific case when unobserved terms are independent for repeated choice situations over time, it is generally not flexible enough to address real-world situations such as when unobserved factors are correlated. To alleviate correlation issues in logit specification, GEV models such as nested logit can be used. Furthermore, if the tastes of decision-makers are random over the population (i.e., heterogeneity between individuals), it is difficult for logit models to incorporate random taste variation under the specification of fixed coefficients. To enhance the flexibility of RUMs, a mixed logit or a probit can be adopted to address random tastes heterogeneity.

4.4.1.3 Mixed logit

Mixed logit models make relatively more flexible assumptions than other RUMs, and they are designed to relax random tastes heterogeneity, substitution patterns, and other correlation issues (i.e., correlation with choices over time) (McFadden and Train 2000; Train 2009). Thanks to its more flexible nature—i.e., relaxing constraints—mixed logit is also widely adopted not only to analyze cross-sectional data, but also multi-level data sets (e.g., panel data). Technically, mixed logit can be approximated to any RUMs (e.g., GEV, probit, logit), based on how observed and unobserved terms are specified, interacted, and assumed (McFadden and Train 2000; Train 2009). Since mixed logit models have widely adopted to the multi-level data (i.e., panel data), the specification of the model in the below considers choice situations for each individual traveler. The utility of person n can be expressed from alternative j in choice situation t, which is the extension of equation (18) by considering repetitive choice situations for decision-maker:

$$U_{njt} = \beta_n^T x_{njt} + \varepsilon_{njt} \tag{24}$$

where β_n^T is a vector of person's coefficients within the population, x_{njt} is a vector of variables of alternative j over choice situation t for person n, and ε_{njt} is the stochastic component, assumed to be iid extreme values across all persons, choice situations, and alternatives. As presented above (equation 23), the conditional probability of standard logit that person n chooses alternative j can be derived based on utility-maximization behavior:

$$L_{ni}(y_{nt}|\beta_n) = \left(\frac{e^{\beta_n^T x_{njt}}}{\sum_{j' \in C_{nt}} e^{\beta' x_{nj'}}}\right) \tag{25}$$

where C_{nt} indicates the choice set of given choice situations and person. For the iterative choice of alternatives over choice situations (i.e., $y_n = [y_{n1}, \cdots, y_{nT}]$), the conditional probability of a vector of choices y_n for person n can be expressed as follows:

$$P(y_n|\, \beta_n) = \prod_{t=1}^{T} L_{nt}(y_{nt}|\beta_n) \tag{26}$$

A vector of person's coefficients β_n is not known, which can be represented by a certain density distribution (i.e., mixing distribution). In particular, this mixing distribution is the key element that determines the ways of specification and estimation process in addressing heterogeneity issues in mixed logit.

In this section, for instance, a parametric probability density distribution is assumed, and the marginal probability of the person's iterative choices can be derived by the integration of $P(y_n|\, \beta_n)$ over the assumed distribution of β_n:

$$P(y_n|x_n, \theta) = f_y(y_n|x_n, \theta) = \int f_y(y_n|x_n, \beta_n) f_\beta(\beta_n|\theta) d\beta_n \quad (27)$$

This probability can be simulated by estimating the unknown parameters θ (a.k.a., hyperparameters) that assumes the function form of β_n. The most common approach for parametric continuous mixing distribution is the maximum simulated likelihood since this marginalization over β_n cannot be performed in closed-form. The following section discuss random tastes heterogeneity issues and mixing distributions that capture random tastes heterogeneity.

The most common type of mixed logit assumes a predetermined parametric continuous distribution such as normal and log-normal, which generally has a single mode (i.e., univariate Gaussian). Although there are many different kinds of mixing distributions that can be assumed in principle, only a limited number of mixing distributions have been used in empirical applications. This is mainly due to the facts that there exist several technical and algorithmic dilemmas when determining different types of mixing distributions—e.g., parametric and nonparametric; continuous and discrete (see details in (Train 2016; Yuan et al. 2015)).

4.5 Application of PGM framework in travel discrete choice behavior

In several centuries, progressions in modeling—e.g., the family of RUM and ML, have often focused on supplementing and tailoring algorithms for a certain (or specific) problem and phenomenon to not only output accurate predictions but also examined associated behaviors. As a consequence, a variety of different modeling specifications are developed to more realistic models while addressing some challenges in data, estimation process, and predetermined assumptions—e.g., the family of RUMs ranging from MNL to mixed models. In this sense, methodological concepts and techniques that involve these tailored algorithms often become more complicated than already existing models, and researchers are left with a deluge of modeling algorithms, as well as various nomenclature. When it comes to modeling the given real-world problems, therefore, researchers typically try to select a suitable method among existing modeling techniques and map their problem onto it—often influenced by their knowledge and familiarity with a specific method (Bishop 2013; Olson et al. 2017). As the era of Big Data and Data Science is creating unprecedented opportunities for researchers, selecting a suitable algorithm becomes more challenging than before, especially for those without a strong background in a certain modeling technique. In addition, despite having been selecting appropriate techniques and successful at its particular task, built models are often not applicable to other use cases. As a result, if a problem and a corresponding application change, the built model will have a poor accuracy and it will have to be substantially modified.

PGM framework in this section aims to address discrete mode choice behaviors to alleviate the existing dilemmas and challenges (e.g., data availability, familiarity of modeling techniques) for researchers, which is antithetical to the traditional paradigm that begins with selecting an appropriate modeling technique based on given conditions. PGM framework, for instance, begins with focusing on the given problems (see details in the previous section 3.1) while accommodating various modeling methods with less technical background in the algorithms selected. In this section, the major specifications of RUMs ranging from MNL to mixed logit models will be modularized into PGM framework with the level of pooling/shrinkage (a.k.a., hierarchical modeling). Specifically, pooling/shrinkage aims to handle unobserved detrimental effects in modeling behaviors (e.g., heterogeneity across individuals) while leveraging useful information for data structure (e.g., information across individuals or groups). Generative process is also presented which are powerful to handle missing or inadequate information required to model through providing stories about how the data has been generated.

In addition, the observations can be different from the types of data such as cross-sectional and panel data. To provide generalized modeling specification, the observation (n) indicates each sampled individual (k) whose making repeated choice occasions (t). Complete pooling, in particular, observation is considered as identical to individual since complete pooling model assumes that a single parameter vector is shared by all individuals.

4.5.1 PGM-complete pooling (PGM-CP)

PGM-complete pooling (PGM-CP) globally shares a single parameter vector within each choice alternative ($\boldsymbol{\beta}_c$), where i indicates the choice alternative. The distribution of parameter vector for each choice alternative ($\boldsymbol{\beta}_c$) is linearly combined with the distribution of variable vector ($\boldsymbol{x}$), and this combination enters the utility function that is converted to the exponential distribution family (i.e., categorical) through a softmax function. In general, this modeling specification can also belong to the class of a generalized linear models (GLM).

Figure 4-4 presents a graphical representation that compactly depicts the relationships between random variables ($\boldsymbol{X}$, $\mathbf{y}$, $\boldsymbol{\beta}$) in the form of PGM-CP specification. The generative process of PGM-CP is also presented in Figure 4-4 (on the right panel), defined by N observations, $n \in \{1, \dots, N\}$ and C choice alternatives, $i \in \{1, \dots, C\}$. Based on the PGM framework in Figure 4, the joint probability distribution of given $\boldsymbol{y}$ and $\boldsymbol{\beta}$ can be factorized as:

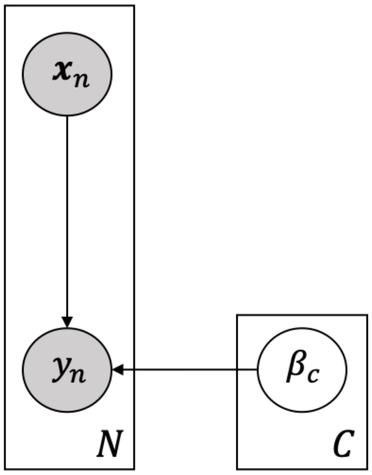
$$p\ (\mathbf{y}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_C | \boldsymbol{X}, \lambda) = \left(\prod_{i=1}^{C} p(\boldsymbol{\beta}_c | \mathbf{0},\ \lambda\mathbf{I})\right) \times$$

$$\left(\prod_{n=1}^{N} p(\ y_n | \boldsymbol{\beta}_1^T \boldsymbol{x}_n, \dots,\ \boldsymbol{\beta}_C^T \boldsymbol{x}_n)\right) \qquad (28)$$

where $\boldsymbol{\beta}_c$ indicates a column vector of parameters for an alternative c, $\mathbf{y}$ is a column vector of response (i.e., choice), $\boldsymbol{X}$ is a matrix of variables, and $\mathbf{0}$ and $\mathbf{I}$ indicates a zero vector and an identity matrix, respectively. Moreover, λ is scale parameter for the variance of random parameter distributions, which often assumed to be constant under this setting (i.e., same variation for every mode). To find $\boldsymbol{\beta}_c$, the equation (28) can be expressed by Bayes' rule:

$$p\ (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_C | \mathbf{y}, \boldsymbol{x}, \lambda) = \frac{\left(\prod_{c=1}^{C} p(\boldsymbol{\beta}_c | \mathbf{0}, \lambda\mathbf{I})\right) \times \left(\prod_{n=1}^{N} cat(y_n | softmax(x_n, \beta_1, \dots, \beta_C)\right)}{\int_{\boldsymbol{\beta}_C} \left(\prod_{c=1}^{C} p(\boldsymbol{\beta}_C)\right) \times \left(\prod_{n=1}^{N} p(y_n |, \beta_1, \dots, \beta_C)\right)} \qquad (29)$$

The last factorization term in the nominator is the likelihood that is the main interest of MNL from the perspectives of frequentist approaches. Thus, this term gives deterministic values for the coefficients (i.e., point estimate) by MLE. Without maximizing the likelihood function, Bayesian inference can make predictions by using prior, posterior, and normalizing constant in the denominator. The estimation of posterior distribution in equation (29) is not closed-form due to the denominator, thus, approximate inference methods are required, and this article primarily uses VI, as described in the previous section.



```
Generative process:
For given $X, C, \lambda$,
  1) for each choice alternative $c = 1, \dots, C$:
     - draw coefficients
         $\boldsymbol{\beta}_c \sim \mathcal{N}(0, \lambda\mathbf{I})$.
  2) for each observation $n = 1, \dots, N$:
     -draw choice
         $y_n \sim cat\left(softmax(\boldsymbol{\beta}_1^T\boldsymbol{x}_n, \dots, \boldsymbol{\beta}_C^T\boldsymbol{x}_n)\right)$.
```

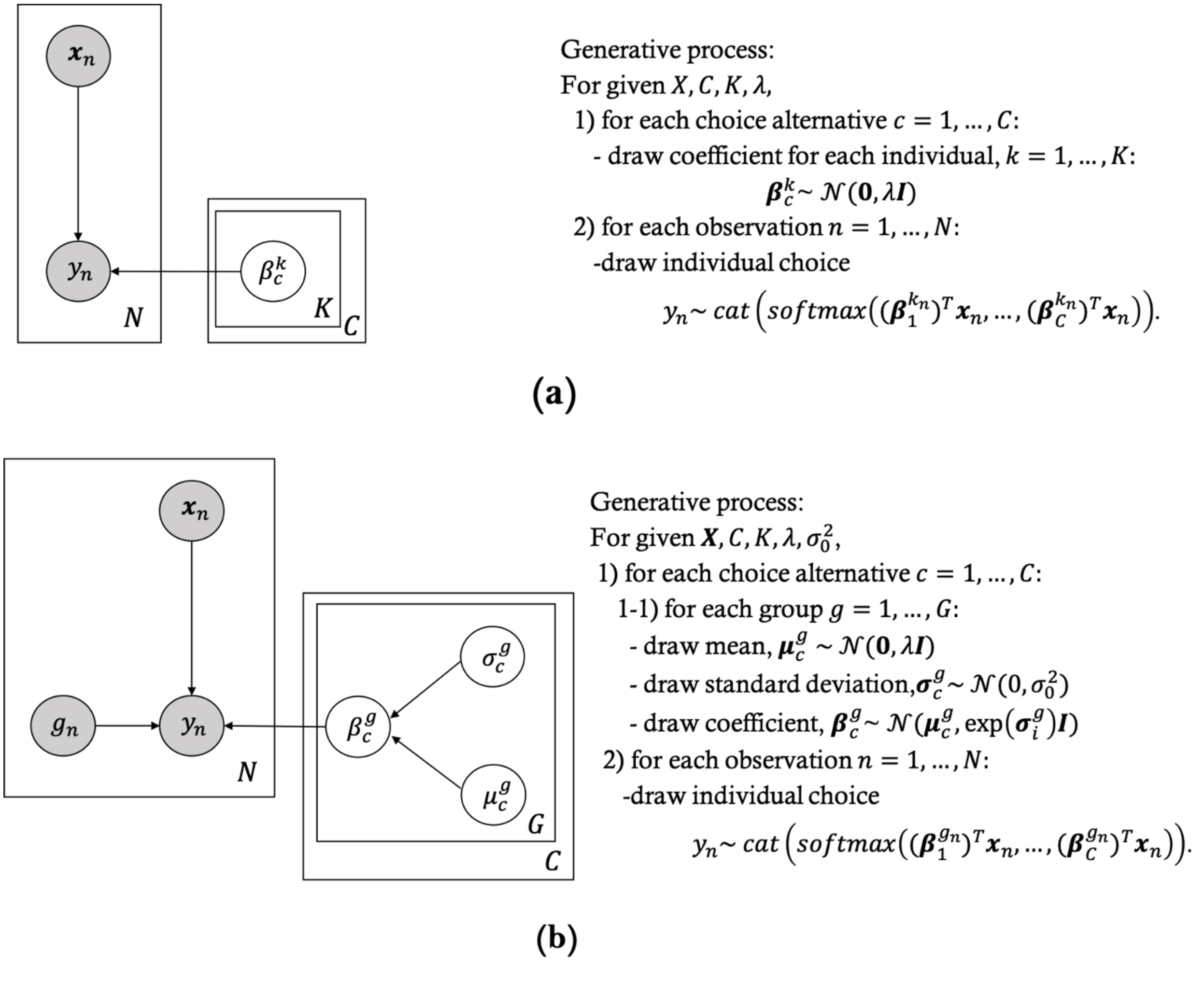
Figure 4-4 Graphical representation and generative process for PGM-CP

4.5.2 PGM-no pooling (PGM-NP)

The previous model in Figure 4-4 is that all individuals share a single set of parameter vector for each class ($\boldsymbol{\beta}_c$). In other word, the model is strongly pooled (i.e., complete pooling), and it is difficult to accommodate unobserved heterogeneity issues (e.g., heterogeneity between individuals or level of sub-groups) in choice behaviors (Wooldridge, 2010). To capture unobserved fixed and random effects, the previous model can be revised by constructing more hierarchies, which is loosely generalized to the modularization of mixed logit models with different mixing distribution—e.g., parametric and nonparametric mixing distribution (Guo et al., 2018; McFadden and Train, 2000; Train, 2016, 2009). In PGM, hierarchical modeling structures can be designed to address these unobserved random effects (Robert, 2014; I and Jordan, 2010). For instance, mode choice behavior may be different from each observation or the number of distinct sub-groups (a.k.a., no pooling and partial pooling, respectively) since different individuals or sub-groups could have different preferences.

$$p(y, \boldsymbol{\beta}_1^1, \dots, \boldsymbol{\beta}_C^1, \dots, \boldsymbol{\beta}_1^K, \dots, \boldsymbol{\beta}_C^K | \boldsymbol{X}, \lambda) \; = \left(\prod_{i=1}^{C} \prod_{k=1}^{K} p(\beta_C^k | \mathbf{0}, \lambda \mathbf{I})\right) \times \prod_{n=1}^{N} p(\, y_n |\, x_n, k_n, \boldsymbol{\beta}_1^1, \dots, \boldsymbol{\beta}_C^1, \dots, \boldsymbol{\beta}_1^K, \dots, \beta_C^K) \qquad (30)$$

where n is each trip made by a certain individual k. $\left(\prod_{i=1}^{C} \prod_{k=1}^{K} p(\beta_C^k | \mathbf{0}, \lambda \mathbf{I})\right)$ is the hierarchical prior and $\prod_{n=1}^{N} p(\, y_n |\, x_n, k_n, \boldsymbol{\beta}_1^1, \dots, \boldsymbol{\beta}_C^1, \dots, \boldsymbol{\beta}_1^K, \dots, \beta_C^K)$ is the likelihood term.



```
Generative process:
For given $X, C, K, \lambda$,
  1) for each choice alternative $c = 1, \dots, C$:
    - draw coefficient for each individual, $k = 1, \dots, K$:
          $\boldsymbol{\beta}_c^k \sim \mathcal{N}(\mathbf{0}, \lambda \boldsymbol{I})$
  2) for each observation $n = 1, \dots, N$:
    -draw individual choice
          $y_n \sim cat\left(softmax\left((\boldsymbol{\beta}_1^{k_n})^T \boldsymbol{x}_n, \dots, (\boldsymbol{\beta}_C^{k_n})^T \boldsymbol{x}_n\right)\right)$.
```

```
Generative process:
For given $\boldsymbol{X}, C, K, \lambda, \sigma_0^2$,
  1) for each choice alternative $c = 1, \dots, C$:
    1-1) for each group $g = 1, \dots, G$:
      - draw mean, $\boldsymbol{\mu}_c^g \sim \mathcal{N}(\mathbf{0}, \lambda \boldsymbol{I})$
      - draw standard deviation,$\boldsymbol{\sigma}_c^g \sim \mathcal{N}(0, \sigma_0^2)$
      - draw coefficient, $\boldsymbol{\beta}_c^g \sim \mathcal{N}(\boldsymbol{\mu}_c^g, \exp(\boldsymbol{\sigma}_i^g) \boldsymbol{I})$
  2) for each observation $n = 1, \dots, N$:
    -draw individual choice
          $y_n \sim cat\left(softmax\left((\boldsymbol{\beta}_1^{g_n})^T \boldsymbol{x}_n, \dots, (\boldsymbol{\beta}_C^{g_n})^T \boldsymbol{x}_n\right)\right)$.
```

Figure 4-5 Graphical representation and generative process for (a) PGM-NP where every individual has own parameter vector (b) PGM-NP with group allocation (PGM-NPG) model where each grouped individuals has own parameter vector

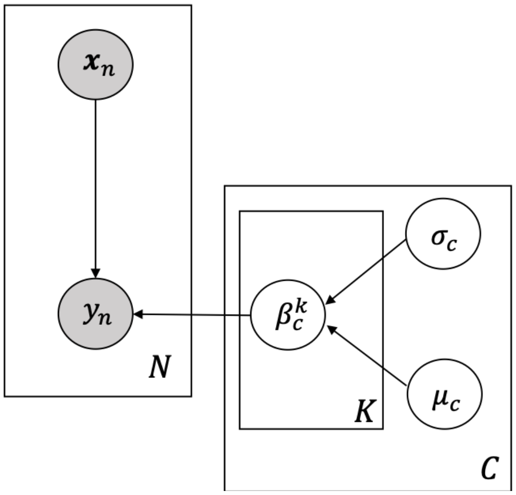
Individuals in PGM-NP models can be categorized into the groups or collections of individuals that shares similar characteristics (i.e., constant over choice situations or time). By introducing group allocation variables g_n, PGM-NP with group allocation (PGM-NPG) can capture inter-group taste heterogeneity, while each group has homogenous tastes (i.e., own set of parameters). The segmentation of individuals can be determined by the given evidence (information) from the samples. Otherwise, it can be treated as unobserved random variables whereas it is predetermined from observed variable (e.g., categorical information).

In particular, individuals are probabilistically assigned to finite- or infinite number of groups, and the number of groups (a.k.a., mixing components) can be inferred by heuristically. This is widely known as the family of latent class (mixed) multinomial logit. For example, latent class model using Dirichlet process priors is one of a popular way to approximate the number of unobserved groups, which can be also viewed as the combination of mixture modeling with logistic regression under the PGM framework. This article is not focusing on methodological details about the latent class cases, thus, more details can be found in (Blei et al., 2016; Teh and Jordan, 2010).

In our case, the information of group (i.e., degree of pooling) is pre-determined by the given data and domain knowledge—i.e., observed random variables (shaded circle). For example, individual mode choice behavior can be significantly affected (clustered) by geographical categories, trip purpose, and other categorical features, researchers therefore use these observed features as a group. Under this specification, the joint posterior distribution in Figure 4-5 (b) can be expressed:

$$p(y, \boldsymbol{\beta}^G_{1\ldots C},\ \boldsymbol{\mu}^G_{1\ldots C}, \boldsymbol{\sigma}^G_{1\ldots C} \big| \boldsymbol{X}, \boldsymbol{I}) = \left(\prod_{c=1}^{C} \prod_{g=1}^{G} p(\boldsymbol{\mu}_c^g) p(\boldsymbol{\sigma}_c^g)\ \prod_{c=1}^{C} \prod_{g=1}^{G} p(\beta_c^g | \boldsymbol{\mu}_c^g, \boldsymbol{\sigma}_i^g)\right) \times \prod_{n=1}^{N} p(\, y_n |\, x_n, g_n, \boldsymbol{\beta}_1^1, \ldots, \boldsymbol{\beta}_C^1, \ldots, \boldsymbol{\beta}_1^G, \ldots, \boldsymbol{\beta}_C^G) \quad (31)$$

where g_n indicates the allocation of observation n to a certain group g.



Generative process:
For given $X, C, K, \lambda, \sigma_0^2$,
1) for each choice alternative $i = 1, \ldots, C$:
- draw mean, $\mu_i \sim \mathcal{N}(\boldsymbol{0}, \lambda \boldsymbol{I})$
- draw standard deviation, $\sigma_i \sim \mathcal{N}(0, \sigma_0^2)$
- draw coefficient for each individual, $k = 1, \ldots, K$:
$\boldsymbol{\beta}_i^k \sim \mathcal{N}(\mu_i, \exp(\sigma_i)\boldsymbol{I})$
2) for each observation $n = 1, \ldots, N$:
-draw individual choice
$y_n \sim cat\left(softmax\big((\boldsymbol{\beta}_1^{k_n})^T \boldsymbol{x}_n, \ldots, (\boldsymbol{\beta}_C^{k_n})^T \boldsymbol{x}_n\big)\right).$

Figure 4-6. Two level hierarchical no-pooling model (HML-NP^2) where all parameters are assumed to be drawn from the same parametric distribution

4.5.3 PGM-partial pooling (PGM-PP)

The previous two models indicate two extreme cases of pooling. Specifically, one extreme, PGM-CP, assumes that all the observations globally share a single set of parameters (i.e., complete pooling). On the other extreme, PGM-NP and -NPG assumes that each individual/group has its own set of parameters (i.e., no pooling). For example, PGM-NP and -NPG models are likely to have overfitting issue because the longitudinal data set contains fewer number of observations (trips) per individual than the number of variables. This overfitting issue might be exacerbated when accommodating ML algorithms in the PGM framework. To alleviate this overfitting, another level of a pooling/hierarchies can be specified.

As for compromising two extreme cases, PGM-PP introduces two level hierarchies for the parameters of our interests. Specifically, PGM-PP shared global prior (i.e., hyper-prior) ties together the parameters of each individual (see Figure 4-6). The joint posterior distribution can be expressed as:

$$p\left(y, \boldsymbol{\beta}_c^g, \boldsymbol{\mu}_c, \sigma_c \middle| \boldsymbol{X}, \boldsymbol{I}\right) = \left(\prod_{c=1}^{C} p(\boldsymbol{\mu}_c) p(\sigma_c) \prod_{g=1}^{G} p(\boldsymbol{\beta}_c^g | \boldsymbol{\mu}_c, \sigma_c)\right) \times \prod_{n=1}^{N} p(y_n | \, x_n, g_n, \boldsymbol{\beta}_1^1, \dots, \boldsymbol{\beta}_C^1, \dots, \boldsymbol{\beta}_1^G, \dots, \boldsymbol{\beta}_C^G) \qquad (32)$$

where n is each observation.$\left(\prod_{c=1}^{C} p(\boldsymbol{\mu}_c) p(\sigma_c) \prod_{g=1}^{G} p(\boldsymbol{\beta}_c^g | \boldsymbol{\mu}_c, \sigma_c)\right)$ is the hierarchical prior and $\prod_{n=1}^{N} p(y_n | \, x_n, g_n, \boldsymbol{\beta}_1^1, \dots, \boldsymbol{\beta}_C^1, \dots, \boldsymbol{\beta}_1^G, \dots, \boldsymbol{\beta}_C^G)$ is the likelihood term. Specifically, $\boldsymbol{\mu}_C$ and σ_C are globally drawn by the assumed prior parametric distribution (i.e., hyper-prior) for each class, then we draw coefficients ($\boldsymbol{\beta}$) for each individual/group. Although σ_c is theoretically follows the conjugate prior (i.e., Inverse Gamma) instead of exp(σ_c) with Gaussian distribution, we applied the exp(σ_c) with Gaussian since the differences between two specifications are marginal when testing two difference specifications.

4.6 CASE STUDIES: TRAVEL MODE CHOICE BEHAVIOR

4.6.1 Data

4.6.1.1 National Household Travel Survey 2017

This chapter mainly uses the 2017 National Household Travel Survey (NHTS) database to investigate the applicability of PGM-B. The NHTS data sets contain the nation's travel diary information across all 50 US states and the District of Columbia ("2017 NHTS Data User Guide" 2018). Specifically, survey respondents are collected directly from a geographically stratified random sample of U.S. households (i.e., 129,112 households), which includes a national sample of 26,000 households and 103,112 additional samples from state departments of transportation (see details in ("2017 NHTS Data User Guide" 2018)). Based on the samples, the NHTS database mainly provides daily travel for all members of households linked to individual personal and household characteristics including demographics, vehicle, and other attitudinal information.

Table 4-1. Descriptive statistics of variables

Variables	**Description**	**mean**	**Std.**	**min**	**max**
Dependent choice alternatives					
CHOICE	1: AUTO 2: TRANSIT 3: BIKE 4: WALK				
Independent variable					
TRPMILES	Travel miles	5.05	5.11	0.2	20.941
TRACCTM	Transit access time	2.66	5.17	0	30
R_AGE	Respondent ages	45.20	19.26	5	92
VEH_DUMMY	Existence of private vehicle			0	1
BIGPOPDEN	Household located in high population density (>10000 per square miles)			0	1
EDUC	Education level of respondent			1	5
TRIPPURP1	Identifier for trip purpose (home-based work, shopping, others, recreation)				
STATES	Identifier for U.S. States				

To obtain all of this information and variables from the NHTS database (i.e., 4 different data sets), the different data sets are combined based on each household, person, and trip identifier (i.e., HOUSEID, PERSONID, VEHICLEID, TRIPID). After obtaining a merged data set, variables with very low response rate (< 10%), some variables containing and redundant and missing values (e.g., "not as certain", "avoid to answering") are eliminated, and inter-related information is merged into smaller discretized values (i.e., dummy). Moreover, the original database contains only 2 to 3 % of transit (e.g., bus, subway) since the database is collected from random U.S. samples. To be able to fairly investigate travel mode choice behaviors, we only select regions that are accessible to transit and use only these selected observations in our analysis.

4.6.1.2 Balancing in minority modes

This case study is to investigate the applicability and feasibility of PGM-B, thus, the original datasets are statistically modified to test our PGM-B models. For instance, samples for transit are significantly fewer than other modes, accounting for only 1,248 observations compared to 93,192 for auto, and observations using transit are generally sampled from the urban areas. This imbalanced samples may cause biased estimation and provide unreasonable modeling predictions. To address this imbalanced issue, we intentionally balance the original observations by using under-sampling methods. Specifically, the nearest neighbor method and random sampling methods (Lindenbaum et al. 2004) are used as an under-sampling method to obtain well-balanced observation throughout the States of U.S. After both sampling processes and removing some inadequate information, we obtain 258 samples for each choice alternatives (i.e., auto, transit, walk, bike).

4.6.1.3 Selection of variables

Table 4-1 presents the descriptive statistics of the selected variables used in this case study. These variables are selected through the variable importance (VI) that shows the relative contribution of variables (i.e., reduction in errors) during the estimation of the rule-based model such as gradient boosting machine (Friedman 2001). Variables in Table 4-1 are most influential for individual mode choice. Thus, this analysis reduces the dimensionality of the original NHTS data. In addition to the information from VI, we intentionally add some variables based on our domain knowledge of mode choice behavior (Golshani et al. 2018; Lee et al. 2018a).

4.6.2 Model specification

4.6.2.1 PGM-B models

PGM-CP, NP, and PP are used to model discrete mode choice behaviors. Specifically, PGM-NP specifies each individual has its own set of parameters. In addition, PGM-PP specifies that each group (i.e., the State of U.S.) have its own set of parameters that globally drawn by the assumed prior parametric distribution (i.e., hyper-prior) for each choice alternative. For instance, individuals are grouped into each region to capture inter-regional taste variations, while each group has homogenous tastes (i.e., own set of parameters).

In addition, variables in data contain varying scales and ranges. The different numerical scales may cause biased estimation since ML models are generally susceptible to the scales and ranges of values. Therefore, we scale all values of variables ranging from 0 to 1 by using the min-max normalization.

$$x' = \frac{x - \min A}{\max A - \min A} \tag{33}$$

4.6.2.2 RUM models

To make a comparison of modeling performances, multinomial logit and mixed logit models are also estimated. The baseline modeling specification for MNL is as follows:

$$U_{nj} = \alpha_{nj} + \beta_{nj}\, x_{nj} + \varepsilon_{nj} \tag{34}$$

The utility maximization behavior of person n is assumed to follow the equation above. α_{nj} denotes alternative-specific constant. β_{nj} are a vector of coefficients of alternative j for person n, and x_{nj} is a vector of variables. ε_{nj} is the stochastic error component that is assumed to be iid extreme values (i.e., Gumbel) across all persons and alternatives.

To address heterogeneity issues across entities (e.g., individuals), MNL model is further specified to incorporate fixed effects and random parameters on mode choice behaviors. Specifically, variables representing individual characteristics are constant among different choice situations within an individual, thus, parameters for these variables are assumed to be fixed. On the other hand, time-variant variables (e.g., alternative-specific constant and variables) are assumed to have random parameters to incorporate heterogeneity issues across individuals. The modeling specification of mixed logit is as follows:

$$U_{njt} = \beta_{n,invariant}\, x_{nj} + \beta_{n,variant}\, x_{njt} + \varepsilon_{njt} \tag{35}$$

$$\varepsilon_{njt} = \alpha_{nj} + \mu_{njt} \tag{36}$$

where $\beta_{n,invariant}$ is a vector of fixed parameters representing individual-specific information such as age, gender, and education level that are constant among choice situations (t). $\beta_{n,variant}$ include time-variant information minimize implausible values for parameters such as travel time, we also tried lognormal distributions for parameters related travel time and cost instead of normal distributions. In addition, ε_{njt} , are specified into two random components. The first component, α_{nj} , is fixed over choice situations but varied by individuals for each choice alternative to incorporate unobserved detrimental effects that are correlated with other variables. The second term, μ_{njt}, is stochastic error component with Gumbel distribution.

The estimation results of MNL and mixed logit is presented in Table 4-2. Both models are estimated with Python Biogeme (Bierlaire 2003b). The prediction accuracy in Table 4-3 and 4-4 are evaluated by the test dataset (40%).

Table 4-2 Estimation of multinomial and mixed logit models

Variable	**Multinomial logit**	**Mixed logit** b)	
		Coefficients	Std.
Auto a)			
constant	-	-	1.226
Auto travel time	-1.312 (-3.98)	-0.267 (-1.92)	1.464 (1.74)
Transit			
constant	-4.033 (3.05)	-0.293 (-3.32)	3.567 (1.91)
Transit access/waiting time	-4.581 (-3.79)	-1.267 (-2.04)	0.629 (4.02)
Vehicle dummy [0, 1]	-5.364 (-14.74)	-5.422 (-11.56)	-
High population density	2.403 (3.07)	1.762 (2.48)	-
Urban indicator [0, 1]	1.268 (3.15)	1.385 (11.44)	-
Education level	2.306 (2.81)	0.417 (3.84)	-
Gender of respondent	0.412 (1.45)	0.013 (1.01)	-
Age of respondent	-1.567 (-3.08)	-2.356 (-4.71)	-
Walk			
Constant	-2.925 (13.78)	0.767 (5.22)	2.660 (2.21)
Walking time	-8.617 (-3.09)	-5.946 (-20.82)	1.001 (4.13)
Vehicle dummy [0, 1]	-4.664 (-10.82)	-4.477 (-12.81)	-
High population density	2.554 (2.91)	2.130 (8.52)	-
Urban indicator [0, 1]	-0.916 (-3.67)	1.130 (5.16)	-
Education level	1.489 (2.32)	2.065 (3.21)	-
Gender of respondent	-0.423 (-2.03)	0.149 (2.41)	-
Age of respondent	-1.443 (-3.06)	-2.035 (-4.37)	-
Bike			
constant	-5.562 (3.89)	-0.372 (-4.68)	2.010 (5.01)
Biking travel time	-6.021 (-2.03)	-12.818 (-12.11)	0.477 (6.91)
Vehicle dummy [0, 1]	-4.715 (-9.29)	-3.605 (-7.21)	-
High population density	1.396 (2.20)	0.170 (3.86)	-
Urban indicator [0, 1]	0.445 (1.79)	-0.048 (-2.81)	-
Education level	1.940 (3.57)	-0.066 (-1.13)	-
Gender of respondent	-1.572 (-1.67)	-3.797 (-4.32)	-
Age of respondent	-2.674 (-3.51)	-3.663 (-8.41)	-
Null log-likelihood at constant	-8,533	-5,712	
Final log-likelihood at convergence	-5,131	-3,153	

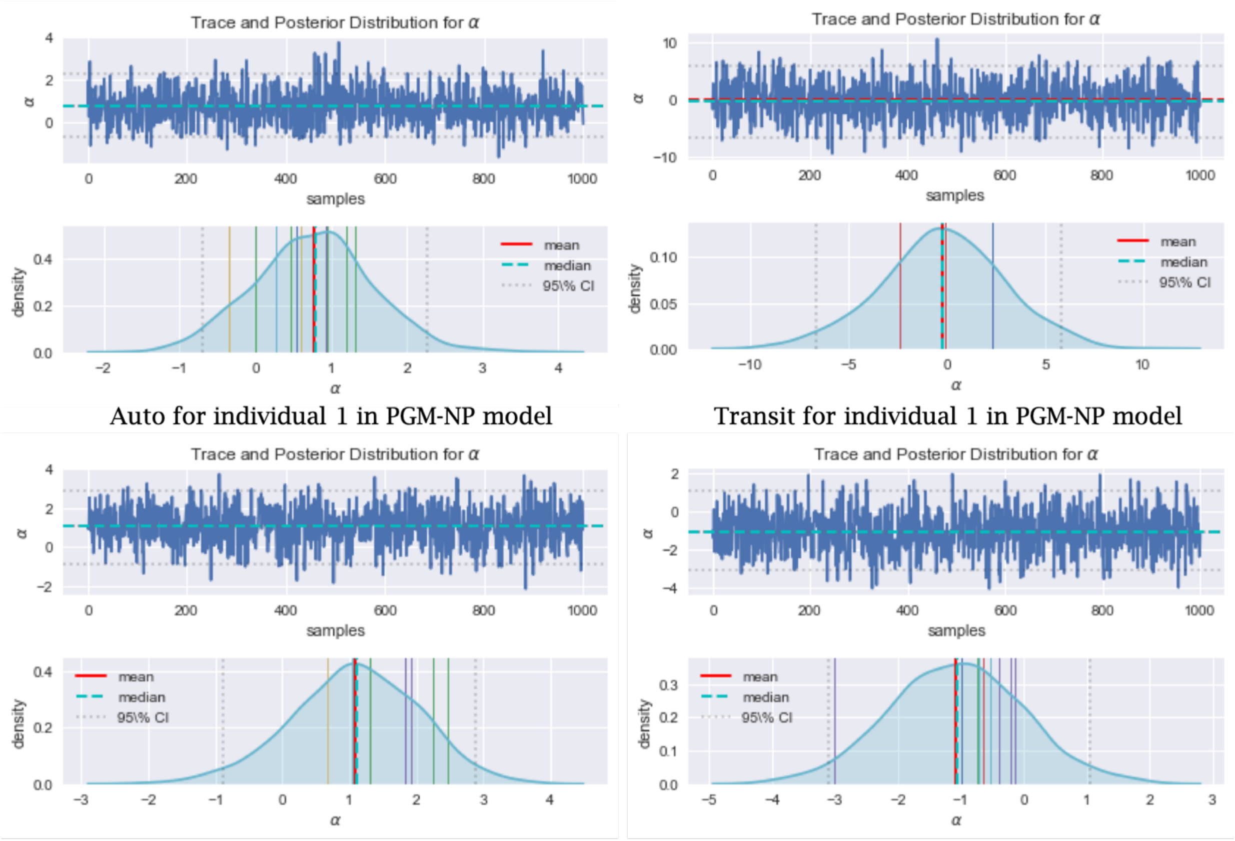
a) Auto is set as a reference alternative.

b) time-invariant variables (e.g., socio-demographic) are adopted as fixed parameters. Random effects are applied to alternative-specific constant and alternative-specific variables that vary by each trip.

4.6.3 Posterior inference

Given the PGM-B frameworks for discrete choice modeling problems, this chapter computes full posterior distributions of our interest based on Bayesian perspectives. As mentioned above, the posterior distributions in Bayesian inference are generally intractable due to the denominator (i.e., normalizing constant). Thus, approximate Bayesian inference methods are used to infer the intractable posterior distributions in Bayesian settings. In particular, this chapter adopts VI that is designed to find a tractable proxy distribution (e.g., exponential family) for the true posterior distributions. All of PGM-B are implemented in Pystan that is probabilistic programming tools built in Python. Each PGM-B is manually customized to accommodate the specific structure of PGM and specified priors.

Two specifications of PGM including PGM-NP and PGM-PP (see section 4.2) are primarily used to infer the posterior of parameters. Figure 4-7 presents the trace distribution of posterior inference process and confidence intervals for the approximated posterior for each choice alternative (auto; transit; bike; walk) within a single individual under PGM-NP specification. Figure 4-8 presents the example of random parameters (i.e., random biases) for a certain regional group (i.e., U.S. State) under PGM-PP specification. Specifically, PGM-PP is designed to group individuals into distinct groups (i.e., U.S. States), and groups share hierarchical prior of mean and variance for the random biases and parameters to alleviate overfitting issue—i.e., two level hierarchies. Posterior distributions of parameters for choice alternatives in region 1, for example are separately evaluated by coherently updating prior distribution and likelihood from the given observed data, and there are some discrepancies in posterior distributions (i.e., mean and variance) between the choice alternatives. In addition, the posterior distributions in Bayesian inference, in contrast to frequentist approaches that give point estimates of parameters, can provide the quantification of uncertainty in choice behaviors.



Auto for individual 1 in PGM-NP model

Transit for individual 1 in PGM-NP model

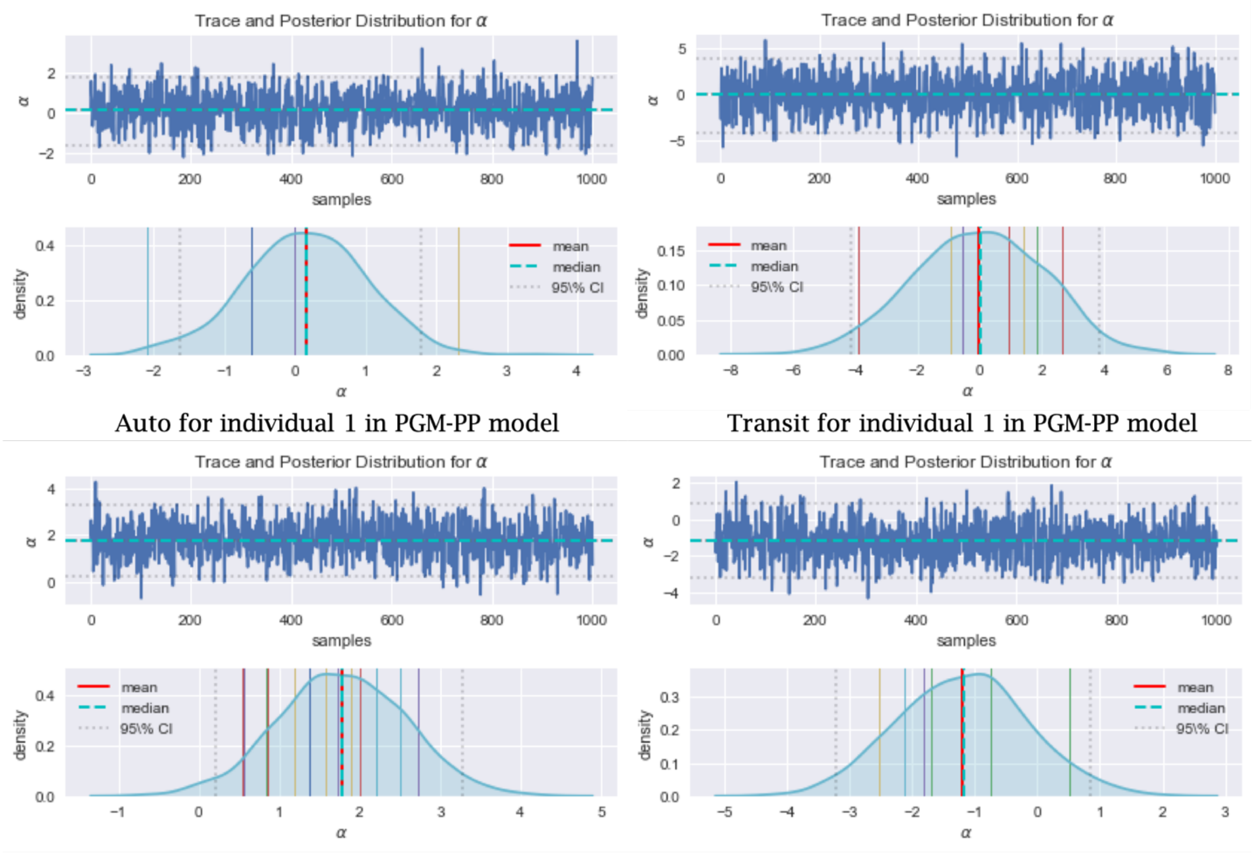
Bike for individual 1 in PGM-NP model

Walk for individual 1 in PGM-NP model

Figure 4-7 Examples of posterior distribution and trace of random biases through the posterior inference process using VI and confidence intervals for the approximated posterior distribution for each choice alternative (auto; transit; bike; walk) within a single individual (PGM-NP)

The estimation of full posterior distributions can explain uncertainty in modeling (Robert, 2014). For instance, it is possible that a model in particular statistical learning returns a prediction for the unseen information that may lie outside of the given data distribution. Naturally, this is an unreasonable prediction, which then becomes an error in the model.

Nonetheless, the model may provide a different prediction for the identical information. In this context, Bayesian approaches in modeling travel behaviors have the potentials to be useful in addressing complex transportation behaviors in the future, while providing uncertainties in analyzing behaviors.



Auto for individual 1 in PGM-PP model

Transit for individual 1 in PGM-PP model

Bike for individual 1 in PGM-PP model

Walk for individual 1 in PGM-PP model

Figure 4-8 Examples of posterior distribution and trace of random biases through the posterior inference process using VI and confidence intervals for the approximated posterior distribution for each choice alternative (auto; transit; bike; walk) within a group (PGM-PP)

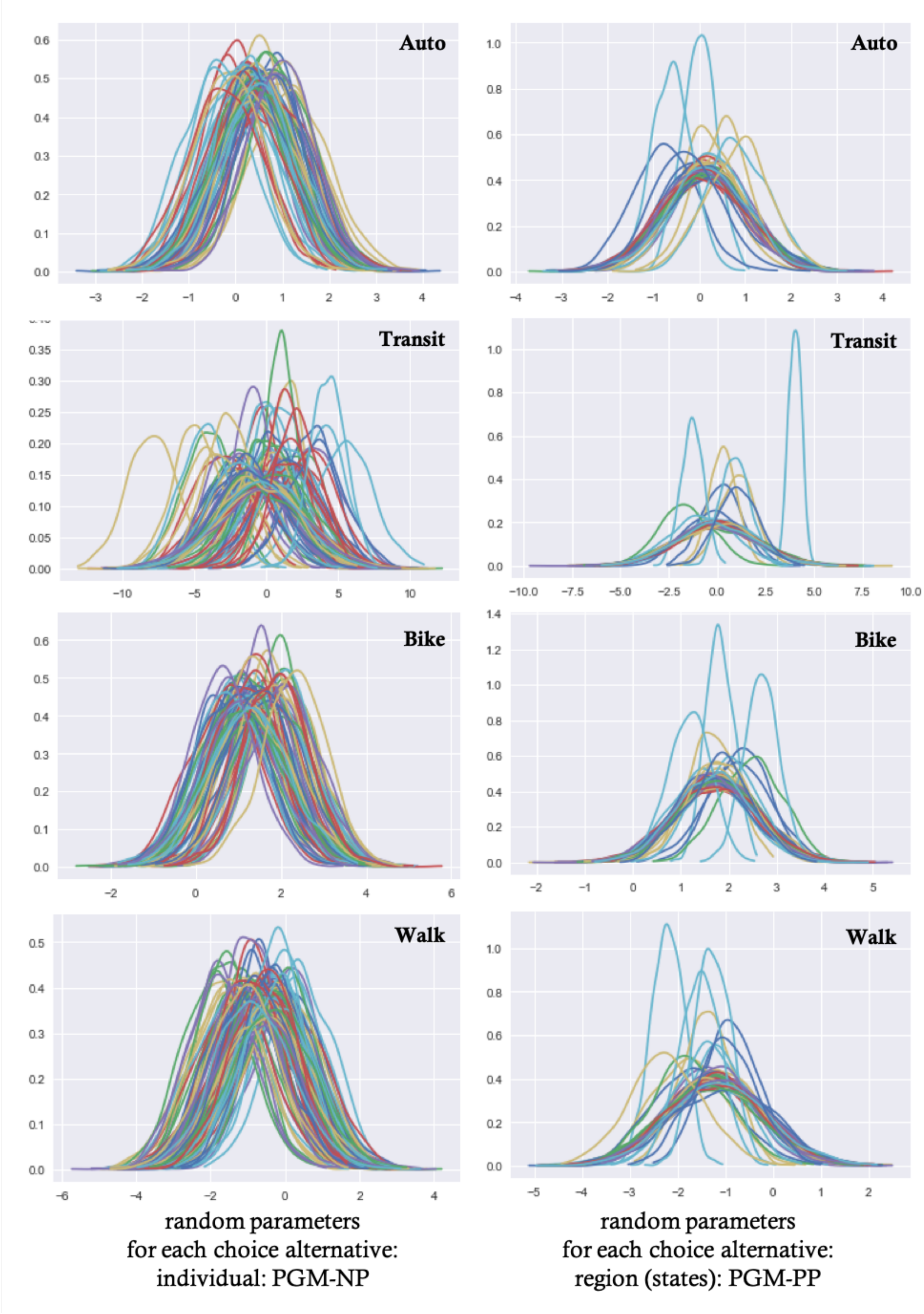


Figure 4-9 Comparison of posterior distributions of random biases for each choice alternative in two models PGM-NP and PGM-PP

Figure 4-9 presents the comparison of of posterior distributions of random biases for each choice alternative in two models PGM-NP and PGM-PP. The series of figures in the left panel depicts the random biases that capture inter-individual heterogeneity, and each distribution within a figure indicates the distribution for a single individual. Thus, a single individual has distinct biases (i.e., intercepts), and transit mode shows the large discrepancies among individuals compared to other modes. The one in the right panel shows the random biases for every region (i.e., State of U.S.) to capture inter-regional heterogeneity of choice behavior within alternative, and each distribution depicts the distribution for a specific State of U.S. In particular, there are large discrepancies (i.e., heterogeneities) in choice behaviors between regions for every mode.

4.6.4 Prediction performance

To evaluate the feasibility of PGM-B for DCM in mode choice, we predict mode choice behaviors by using the inferred posterior distribution of parameters $\boldsymbol{\beta}$. The data sets are divided into two subsets: train (60%) and test (40%) sets. train set is used to fit each model and test set is used to obtain predictions based the posterior estimations from the fitted model.

Table 4-2 and Table 4-3 present the confusion matrix, in which each row and each column indicates the observed and predicted individuals for each mode respectively. When it comes to the performances between PGM-CP and PGM-NP, unobserved heterogeneity issues for inter-individuals or groups are somewhat captured in PGM-NP model compared to PGM-CP model. The PGM-PP show the highest prediction accuracy among PGM models. It is mainly due to the fact that PGM-PP model is non only capture heterogeneous preferences and behaviors across groups/individuals but also alleviate overfitting issue that may cause high-variance estimation by specifying additional hierarchies (i.e., hyper-priors).

Table 4-3. Confusion matrix for HML-CP, HML-NP, HML-PP

Test Dataset (HML-CP)		**Predicted Choice**					
		Auto (72)	Transit (76)	Bike (32)	Walk (78)	Mode Accuracy	Overall Accuracy
Observed Choice	Auto (71)	**51**	12	3	5	71.8%	71.7%
	Transit (71)	9	**56**	4	2	78.9%	
	Bike (46)	6	3	**22**	15	47.8%	
	Walk (70)	6	5	3	**56**	80.0%	
Test Dataset (HML-NP^2)		**Predicted Choice**					
		Auto (70)	Transit (79)	Bike (37)	Walk (73)	Mode Accuracy	Overall Accuracy
Observed Choice	Auto (71)	**56**	8	2	5	78.9%	73.7%
	Transit (71)	7	**58**	4	2	81.7%	
	Bike (46)	4	5	**24**	13	52.2%	
	Walk (70)	3	8	7	**53**	74.6%	
Test Dataset (HML-PP)		**Predicted Choice**					
		Auto (67)	Transit (68)	Bike (41)	Walk (82)	Mode Accuracy	Overall Accuracy
Observed Choice	Auto (71)	**59**	6	1	5	83.1%	77.5%
	Transit (71)	3	**58**	7	3	81.7%	
	Bike (46)	1	4	**25**	16	54.3%	
	Walk (70)	4	0	8	**58**	82.9%	

In addition, the capability of PGM-B with RUM models are also investigated. Although the discrepancies in prediction accuracy among mixed logit, PGM- NP, and PGM-PP are marginal, PGM-PP shows the best performance. In this sense, we may believe that Bayesian inference using VI shows compatible or better performances to simulate choice behaviors compared to the maximum simulated likelihood (MSL) estimation.

Table 4-4. Confusion matrix for the multinomial logit model and mixed logit model

Test Dataset (multinomial logit)		Predicted Choice					
		Auto (77)	Transit (63)	Bike (35)	Walk (83)	Mode Accuracy	Overall Accuracy
Observed Choice	Auto (71)	**55**	6	4	6	77.5%	70.5%
	Transit (71)	11	**54**	4	2	76.1%	
	Bike (46)	6	3	**15**	22	32.6%	
	Walk (70)	5	4	3	**58**	82.9%	
Test Dataset (mixed logit)		**Predicted Choice**					
		Auto (77)	Transit (63)	Bike (35)	Walk (83)	Mode Accuracy	Overall Accuracy
Observed Choice	Auto (71)	**55**	8	3	5	77.5%	74.4%
	Transit (71)	18	**49**	2	2	69.0%	
	Bike (46)	3	1	**27**	15	58.7%	
	Walk (70)	1	5	3	**61**	87.1%	

4.7 CONCLUSION

This article aimed to suggest a way to apply probabilistic modeling approaches to transportation behavior through a flexible modeling framework that can be compiled with any algorithms. Toward this goal, we used PGM framework to compactly represent all of modeling components governing the given problem—e.g., the (causal) structure of data, assumptions, and other prior belief based on domain knowledge. This framework is combined with Bayesian inference to coherently address uncertainty issues through the use of probability theory and infer the full distributions of model's parameters. In particular, PGM and Bayesian inference (PGM-B) can be separated into three sub-processes: representation, modeling, and inference. Using a PGM-B framework, modeling tasks begin with considering all modeling components that are treated as random variables, which is antithetical to traditional paradigm that starts with selecting an algorithm. Then, these high-dimensional variables are intuitively and compactly represented as a graphical structure with generative process that clarifies our initial beliefs about the problem. Once the PGM framework is constructed, now the goal is to infer the full posterior distributions of our interests through Bayesian inference. To infer high-dimensional and intractable posterior distributions, this article uses variational inference (VI) that a leverages techniques from ML to approximate probability densities for approximating probability densities.

This article derives three PGM frameworks with the different specifications of priors and relationships between modeling components, which include PGM-CP, PGM-NP, PGM-NPG, and PGM-PP. The posterior distribution of random parameters are approximated through the posterior inference process using VI, shown in Figure 8-1, 8-2, and Figure 9. To check the applicability of PGM-B, the prediction accuracy of PGM-B models and RUM including MNL and mixed logit are evaluated by hold-out method. The results are presented in the Table 2-1 and 2-2 by using a confusion matrix. In particular, PGM-PP showed the highest performance among other models in part thanks to their specification that is being able to non only better capture uncertainties related to heterogeneity but also alleviate overfitting issues.

For future work, the combination of PGM framework and Bayesian inference suggested in this article is a way to design probabilistic reasoning system (Spirtes et al., 2000). In probabilistic reasoning, understanding causal relations among related factors can be also useful to not only better understand decision-making behaviors but also evaluate possible subjunctive scenarios such as policy impacts. In particular, structure learning presented in Figure 1 aim to explore and evaluate relations among all variables and their probabilistic relations from the given problem and its data—i.e., markov properties, $P(x_i|pa(x_i))$. This probabilistic dependency over a set of random variables can contribute to construct more realistic decision-making system that provides the highly-taliored system based on causal hypothesis (i.e., subjunctive scenarios). Traditional Functional causal models (FCM) methods, however, mostly perform well with discrete information (Pearl, 2009; Rohekar et al., 2018; Zhang et al., 2017), and they may not take into account the full information from the high-dimensional and observational data (Goudet et al., 2017; Zhang et al., 2017), which is built under strong assumptions (e.g., linearity and no additive random noise in the relationships between two variables). To overcome several limitations present in traditional algorithms, we aim to utilize the power of artificial intelligence that combines the traditional causal search algorithm with ML techniques such as a generative neural network are highly adaptable to both continuous and discrete data.

In addition, novel non-parametric models have widely used in many domains (Wong et al., 2017). For instance, deep learning (DL) generally require more complex algorithmic features, and they are computationally more expensive than other modeling algorithms, but in theory, they should be able to even better capture complex and nonlinear relationships in a dataset. Nonetheless, they have low interpretability and are limited to incorporate prior and domain knowledge. By using PGM framework and Bayesian approaches, it is possible to solve problems in the field of transportation by simultaneously taking the advantages of different modeling approaches such RUM and ML. PGM-B therefore possess a bright future in the realm of transportation modeling during the era of AI and Big data.

5 CONCLUSION AND FUTURE WORKS

5.1 CONCLUSION

The main goal of this dissertation was to gain a fundamental understanding of how machine-based statistical learning (ML) can contribute and be applied to urban metabolism. Recently, the application of ML approaches has been incredibly successful in many domains. Deep learning (DL), for instance, has had unprecedented success in the process of images and sounds, but it has not been exposed extensively to resources consumption and users' behaviors. As acknowledged by many, the use of ML has received some reluctance in domains such as transportation due to their current limitations often related to poor or lack of interpretability. These interpretability issues can lead to difficulties in explaining behaviors, incorporating domain knowledge, and providing the level of statistical confidence for the results. In this sense, this dissertation not only aimed to contribute to the applicability of ML approaches within the urban context but also to address issues resulting from low interpretability of ML models.

As previously mentioned in the introduction (chapter 1), chapters 2 to 4 aimed to address limitations regarding the application of ML approaches for urban modeling purposes—i.e., applicability, capability, interpretability, flexibility, and uncertainty. Chapter 2 primarily focused on investigating the capability and applicability of ML—that is, artificial neural networks (ANN) for the application of discrete choice models in the field of transportation. In particular, four different types of ANN models are used to model, predict, and evaluate travel mode choice behaviors—all ANN models are algorithmically different from each other. For instance, clustered probabilistic neural network (CPNN) combines that the concept of kernel-based approach (i.e., k-means clustering) and probabilistic decision theory (i.e., Bayesian Parzen window classifier). In contrast, backpropagation neural network (BPNN) are based on traditional optimization methods (i.e., gradient descent) with the assistance of machine-based repetitive computation. In addition, the learned (estimated) models are evaluated and validated using k-fold cross-validation, and the prediction accuracies of ANN models outperform a multinomial logit model (a traditional modeling approach). These models are also used to interpret user behaviors through sensitivity analysis.

Chapter 3 aimed to provide useful insights on several technical challenges, including a selection of modeling methods and data availability in modeling tasks, for the case of end-use water consumption and behaviors. Specifically, 12 modeling methods grouped into two general categories—parametric models and non-parametric ML models—were adopted to model and predict household water use, based on two different data scenarios. The results revealed that nonparametric ML methods such as gradient boosting machine (GBM) perform best thanks to their algorithmic properties. Specifically, the algorithmic properties of rule-based methods with the boosted machine are more suitable to analyze data that may include unobserved heterogeneity between users, partly thanks to their discriminative nature. Although the interpretation of GBM is not discussed in this chapter, rule-based methods are generally more interpretable than other ML techniques as well. The ways to interpret GBM are discussed in another article from this author of this dissertation—not included in this dissertation—that aimed to investigate attitudes and behaviors toward autonomous vehicles by using boosting machine (Lee et al. 2019). In general, ML models are interpreted based on the model-agnostic or model-specific method—e.g., feature importance, partial dependence (Friedman et al. 2001; Molnar 2019).

Chapter 4 adopted different ways to enhance the interpretability of ML models not only to incorporate domain knowledge but also to handle uncertainties in modeling tasks. Specifically, the key concept of this chapter is to separate knowledge, model, and inference (e.g., probabilistic reasoning) in constructing decision-making systems (a.k.a., reasoning systems). Toward this goal, this chapter used probabilistic graphical models (PGM) and Bayesian inference that coherently manipulate and quantify uncertainties through the use of probability and graph theories. This modular probabilistic modeling framework, PGM-B, can be separated into representation and inference processes. In the representation process, we begin with considering all kinds of quantities governing the problem, which is antithetical to the traditional modeling paradigm that might struggle with selecting algorithms. All quantities, treated as random variables, are intuitively and compactly represented as a graphical structure based on their relationships between each other. During this representation, in particular, we can incorporate domain knowledge—e.g., personal beliefs, priors. Then, the PGM-B can be adapted to any modeling techniques to infer the full distributions of our interests based on the PGM. Specifically, this chapter used variational inference (VI) that leverages techniques from ML to approximate probability densities. To investigate applicability, this chapter also derives a way to develop a PGM-B to address travel mode choice behaviors based on our assumptions and under different specifications (i.e., level of pooling). In particular, these frameworks are derived to capture unobserved heterogeneity and quantify uncertainty by inferring the full posterior distributions. Prediction performances are also validated and compared with existing random utility models (RUM).

The chapters in this dissertation fill important theoretical and technical knowledge gaps in the realm of urban metabolism and urban modeling in the era of Artificial Intelligence

(AI). In particular, the chapters focused not only on providing ways to take advantage of ML approaches, but also to address some of their limitations—e.g., interpretability, uncertainties. The principles and practical applications of Data Science can be further used to develop novel and creative approaches to gain a fundamental understanding of many other characteristics of cities that are not related to urban metabolism. Simultaneously, the ability to understand these other characteristics can shed lights into making cities more sustainable and resilient.

5.2 FUTURE WORKS

Cities are shaped by the interconnections between their infrastructure systems and are operated by the spatial and temporal interactions between humans and infrastructure systems. In particular, over 50% of the world's population now resides together in cities, and that number is expected to be increased to 68% by 2050. As a consequence, cities are the sites of tremendous flows of energy and materials. The fact that cities account for over 75% of primary energy use worldwide is one example of the enormous responsibility that rests on them. Most of this energy is used to provide the necessary services required from infrastructure systems that are fundamentally used to meet the needs of people.

Recently, many of the significant efforts to make urban systems smarter, more sustainable, and more resilient have been driven by a wave of advances in technologies. For instance, the Internet-of-Things (IoT) global phenomenon that includes smartphones, social media, web applications, and crowdsourced environments are changing human behaviors and lifestyles. The convergence of technologies, infrastructure, and human behavior make it possible to generate a large amount of complex and multi-sourced data in spatiotemporal dimensions. Specifically, the recently generated data that were not previously available contain highly interrelated information about infrastructure and resources use, as well as human behaviors, that could very well capture important urban dynamics. These phenomenal changes in data-generating systems provide unprecedented opportunities that enable more accurate modeling of urban metabolism through the understanding of these urban dynamics—e.g., real-time interdependencies and fluctuations of behaviors. Current stand-alone sectoral methodologies, however, are limited to explore urban dynamics due to their restrictions on multi-dimensional spatiotemporal datasets—e.g., predetermined assumptions. A shift in modeling paradigm is required to effectively explore information, and this dissertation suggests some novel concepts and methodologies by leveraging the power of data-driven modeling, closer to the theme of this dissertation.

5.2.1 An integrated urban metabolism modeling platform (Digital Twin)

A modularized and integrated modeling platform is required through the use of extensive data sources (i.e., Big Data), treated with additional machine-oriented computation (e.g., data science). This platform offers promising opportunities in the field of urban systems analysis such as the concept of Digital Twins. Digital Twins are virtual modeling environments that bridge the virtual and physical worlds using iterative feedback workflows between data captured by the physical world and behaviors and prediction acquired from the virtual world. This pairing of the two worlds allows us to explore, visualize, model, predict, and monitor systems with the aim to design them better and head off problems before they occur.

Toward this goal, the concept of ML can be a promising option that provides a modular probabilistic framework, which combines probability theory, logical graph structure, and Bayesian inference. Specifically, MBML can be coupled with any modeling algorithms and coherently handle uncertainties that are inevitable in the modeling of real-world problems as it is impossible to fully observe the world. Digital Twins would directly contribute to the general body of work on urban metabolism, not only to more realistically replicate actual urban systems and behaviors, but also to evaluate what-if scenarios in terms of future changes in the overall urban systems, user behaviors, and enforcement of policies.

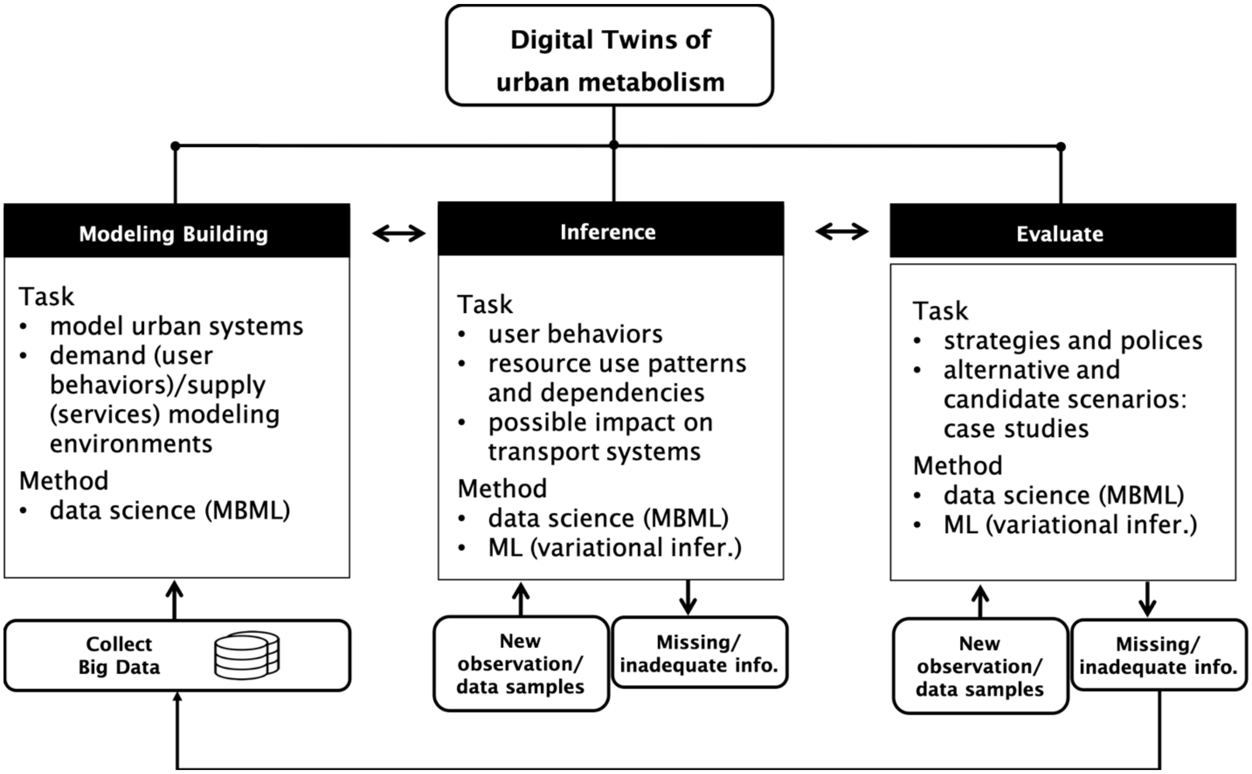

Digital Twins of urban metabolism
Modeling Building
Task
• model urban systems
• demand (user behaviors)/supply (services) modeling environments
Method
• data science (MBML)
Inference
Task
• user behaviors
• resource use patterns and dependencies
• possible impact on transport systems
Method
• data science (MBML)
• ML (variational infer.)
Evaluate
Task
• strategies and polices
• alternative and candidate scenarios: case studies
Method
• data science (MBML)
• ML (variational infer.)
Collect Big Data
New observation/ data samples
Missing/ inadequate info.
New observation/ data samples
Missing/ inadequate info.

Figure 5-1 Flow of Digital Twins of urban metabolism

Digital twins of urban metabolism enable an accurate:

- **Model** urban systems including demand (user behaviors) and supply (services) modeling environments to evaluate policy impacts, enabled by Big Data and existing evidence (data samples) and leveraged by machine learning and powerful computation means

- **Infer (or predict)** shifts in resource use patterns and user behaviors, and possible impacts on the urban systems—missing or inadequate information required to model to enhance prediction accuracy and interpretability of user behaviors (e.g., socio-demographic information about users)
- **Evaluate** the strategies and policies geared toward the development of future urban systems (e.g., smart cities) and services to design sustainable and resilient cities

5.2.2 Causal inference in PGM

As mentioned above, the consequences of digital revolution and relevant shifts in human behaviors can increase the interdependencies between urban infrastructure, humans, and technologies. Simultaneously, data generated in cities is by nature multi-domain, multi-dimensional, and spatiotemporal. The increased complexity and diversity of cities and their data may further increase uncertainties that would make it even more difficult to manage cities. Although reasoning systems-based PGM and Bayesian inference can manipulate and quantify uncertainties by using probability theory and by incorporating domain knowledge, it is limited to explore fundamentals of interdependencies. In other words, we require mathematical language to better express the interdependencies in urban systems and user behaviors. Toward this goal, the concept of causality can be adopted to measure the interrelationships among the components in urban metabolism, and it can also be viewed as the flow of information (i.e., belief propagation) among the components. Causal analysis aims to examine the dependencies between variables. It mainly explores conditional independence between variables and identifies d-separation segments. At the time of this writing, three possible methods exist to learn the structure of dependencies (a.k.a., structure learning): (1) domain knowledge, (2) hypothesis testing, and (3) algorithmic learning. In particular, algorithmic learning can be conducted by traditional functional causal models (FCM) (e.g., constraint-based method, hybrid method); however, these traditional models mostly perform well with discrete information. Thus, they may not take into account the full information from high-dimensional and observational data (Goudet et al. 2017; Zhang et al. 2017), which is built under strong assumptions. Instead, we can utilize Data Science techniques such as DL methods that combine the hybrid causal search algorithm with a generative neural network.

Overall, the methodologies and approaches presented in this dissertation further contributed to laying the foundations for a paradigm of modeling urban metabolism, particularly leveraging the unprecedented power of Big Data and ML. Much work remains to be accomplished, however, but there is no doubt that machine-based statistical learning will be a major role to assess and model the urban metabolism of neighborhoods and cities over the world.

6 References

"2017 NHTS Data User Guide." (2018). Federal Highway Administration.

"2018 Revision of World Urbanization Prospects." (2018). United Nations.

Adamowski, J., Fung Chan, H., Prasher, S. O., Ozga-Zielinski, B., and Sliusarieva, A. (2012). "Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada." *Water Resources Research*, 48(1).

Ahmad, N., and Derrible, S. (2015a). "Evolution of public supply water withdrawal in the USA: A network approach." *Journal of Industrial Ecology*, 19(2), 321–330.

Ahmad, N., and Derrible, S. (2015b). "Evolution of Public Supply Water Withdrawal in the USA: A Network Approach: Water Withdrawal in the USA: A Network Approach." *Journal of Industrial Ecology*, 19(2), 321–330.

Ahmad, N., Derrible, S., and Cabezas, H. (2017). "Using Fisher information to assess stability in the performance of public transportation systems." *Royal Society open science*, 4(4), 160920.

Ahmad, N., Derrible, S., Eason, T., and Cabezas, H. (2016). "Using Fisher information to track stability in multivariate systems." *Royal Society Open Science*, 3(11), 160582.

Akbarzadeh, M., Memarmontazerin, S., Derrible, S., and Salehi Reihani, S. F. (2017). "The role of travel demand and network centrality on the connectivity and resilience of an urban street system." *Transportation*.

Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., Bastien, F., Bayer, J., Belikov, A., and Belopolsky, A. (2016). "Theano: A Python framework for fast computation of mathematical expressions." *arXiv preprint*.

Altunkaynak, A., and Nigussie, T. A. (2017). "Monthly water consumption prediction using season algorithm and wavelet transform–based models." *Journal of Water Resources Planning and Management*, 143(6), 04017011.

Al-Zahrani, M. A., and Abo-Monasar, A. (2015). "Urban Residential Water Demand Prediction Based on Artificial Neural Networks and Time Series Models." *Water Resources Management*, 29(10), 3651–3662.

Anthony, M., and Bartlett, P. L. (2009). *Neural network learning: Theoretical foundations*. cambridge university press.

Arbues, F., and Villanua, I. (2006). "Potential for pricing policies in water resource management: estimation of urban residential water demand in Zaragoza, Spain." *Urban Studies*, 43(13), 2421–2442.

Arbués, F., Villanúa, I., and Barberán, R. (2010). "Household size and residential water demand: an empirical approach." *Australian Journal of Agricultural and Resource Economics*, 54(1), 61–80.

Bai, Y., Wang, P., Li, C., Xie, J., and Wang, Y. (2014). "Dynamic forecast of daily urban water consumption using a variable-structure support vector regression model." *Journal of Water Resources Planning and Management*, 141(3), 04014058.

Bansal, A., Rompikuntla, S. K., Gopinadhan, J., Kaur, A., and Kazi, Z. A. (2015). "Energy Consumption Forecasting for Smart Meters." *arXiv preprint arXiv:1512.05979.*

Barles, S. (2010). “Society, energy and materials: the contribution of urban metabolism studies to sustainable urban development issues.” *Journal of Environmental Planning and Management*, 53(4), 439–455.

Beloin-Saint-Pierre, D., Rugani, B., Lasvaux, S., Mailhac, A., Popovici, E., Sibiude, G., Benetto, E., and Schiopu, N. (2017). “A review of urban metabolism studies to identify key methodological choices for future harmonization and implementation.” *Journal of Cleaner Production*, 163, S223–S240.

Bengio, Y., and Bengio, S. (2000). “Modeling high-dimensional discrete data with multi-layer neural networks.” 400–406.

Bentz, Y., and Merunka, D. (2000). “Neural networks and the multinomial logit for brand choice modelling: a hybrid approach.” *Journal of Forecasting*, 19(3), 177–200.

Bhat, C. R. (1998). “Accommodating variations in responsiveness to level-of-service measures in travel mode choice modeling.” *Transportation Research Part A: Policy and Practice*, 32(7), 495–507.

Bhat, C. R. (2001). “Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model.” *Transportation Research Part B: Methodological*, 35(7), 677–693.

Bierlaire, M. (2003). “BIOGEME: a free package for the estimation of discrete choice models.”

Bishop, C. M. (2006a). *Pattern recognition and machine learning*. springer.

Bishop, C. M. (2006b). *Pattern recognition and machine learning*. springer.

Bishop, C. M. (2013). “Model-based machine learning.” *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 371(1984), 1–17.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2016). "Variational Inference: A Review for Statisticians."

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). "Latent dirichlet allocation." *Journal of machine Learning research*, 3(Jan), 993–1022.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Brentan, B. M., Luvizotto Jr, E., Herrera, M., Izquierdo, J., and Pérez-García, R. (2017). "Hybrid regression model for near real-time urban water demand forecasting." *Journal of Computational and Applied Mathematics*, 309, 532–541.

Broto, V. C., Allen, A., and Rapoport, E. (2012). "Interdisciplinary Perspectives on Urban Metabolism: Interdisciplinary Perspectives on Urban Metabolism." *Journal of Industrial Ecology*, 16(6), 851–861.

Cacoullos, T. (1966). "Estimation of a multivariate density." *Annals of the Institute of Statistical Mathematics*, 18(1), 179–189.

Cominola, A., Spang, E. S., Giuliani, M., Castelletti, A., Lund, J. R., and Loge, F. J. (2018). "Segmentation analysis of residential water-electricity demand for customized demand-side management programs." *Journal of Cleaner Production*, 172, 1607–1619.

Cottrill, C. D., and Derrible, S. (2015). "Leveraging big data for the development of transport sustainability indicators." *Journal of Urban Technology*, 22(1), 45–64.

Croissant, Y., and Croissant, M. Y. (2019). "Package 'mlogit.'"

Daziano, R. A., and Bolduc, D. (2013). "Incorporating pro-environmental preferences towards green automobile technologies through a Bayesian hybrid choice model." *Transportmetrica A: Transport Science*, 9(1), 74–106.

Daziano, R. A., Miranda-Moreno, L., and Heydari, S. (2013). “Computational Bayesian statistics in transportation modeling: from road safety analysis to discrete choice.” *Transport reviews*, 33(5), 570–592.

De Carvalho, M., Dougherty, M., Fowkes, A., and Wardman, M. (1998). “Forecasting travel demand: a comparison of logit and artificial neural network methods.” *Journal of the Operational Research Society*, 717–722.

De’ath, G. (2002). “MULTIVARIATE REGRESSION TREES: A NEW TECHNIQUE FOR MODELING SPECIES–ENVIRONMENT RELATIONSHIPS.” *Ecology*, 83(4), 1105–1117.

DeOreo, W. B., Mayer, P. W., Dziegielewski, B., and Kiefer, J. (2016). *Residential end uses of water, version 2.* Water Research Foundation.

Derrible, S. (2016a). “Urban infrastructure is not a tree: Integrating and decentralizing urban infrastructure systems.” *Environment and Planning B: Planning and Design*.

Derrible, S. (2016b). “Complexity in future cities: the rise of networked infrastructure.” *International Journal of Urban Sciences*, 1–19.

Derrible, S. (2018). “An approach to designing sustainable urban infrastructure.” *MRS Energy & Sustainability*, 5.

Derrible, S. (2019). *URBAN ENGINEERING FOR SUSTAINABILITY.* MIT PRESS, S.l.

Derrible, S., and Ahmad, N. (2015). “Network-based and binless frequency analyses.” *PloS one*, 10(11), e0142108.

Domencich, T. A., and McFadden, D. (1975). *Urban travel demand-a behavioral analysis*.

Domene, E., and Saurí, D. (2006). “Urbanisation and water consumption: Influencing factors in the metropolitan region of Barcelona.” *Urban Studies*, 43(9), 1605–1623.

Donkor, E. A., Mazzuchi, T. A., Soyer, R., and Alan Roberson, J. (2012). "Urban water demand forecasting: review of methods and models." *Journal of Water Resources Planning and Management*, 140(2), 146–159.

Doshi-Velez, F., and Kim, B. (2017a). "Towards a rigorous science of interpretable machine learning."

Doshi-Velez, F., and Kim, B. (2017b). "Towards a rigorous science of interpretable machine learning."

Dougherty, M. (1995). "A review of neural networks applied to transport." *Transportation Research Part C: Emerging Technologies*, 3(4), 247–260.

Elith, J., Leathwick, J. R., and Hastie, T. (2008). "A working guide to boosted regression trees." *Journal of Animal Ecology*, 77(4), 802–813.

Farooq, B., Bierlaire, M., Hurtubia, R., and Flötteröd, G. (2013). "Simulation based population synthesis." *Transportation Research Part B: Methodological*, 58, 243–263.

Firat, M., Turan, M. E., and Yurdusev, M. A. (2010). "Comparative analysis of neural network techniques for predicting water consumption time series." *Journal of hydrology*, 384(1–2), 46–51.

Firat, M., Yurdusev, M. A., and Turan, M. E. (2009). "Evaluation of artificial neural network techniques for municipal water consumption modeling." *Water resources management*, 23(4), 617–632.

Fricke, K. (2014). *Analysis and Modelling of Water Supply and Demand Under Climate Change, Land Use Transformation and Socio-Economic Development*. Springer Theses, Springer International Publishing, Cham.

Friedman, J. (2001). "Greedy function approximation: a gradient boosting machine." *Annals of statistics*, 1189–1232.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*. Springer series in statistics New York.

Froukh, M. L. (2001). "Decision-support system for domestic water demand forecasting and management." *Water Resources Management*, 15(6), 363–382.

Ghahramani, Z. (2015). "Probabilistic machine learning and artificial intelligence." *Nature*, 521(7553), 452.

Ghimire, M., Boyer, T. A., Chung, C., and Moss, J. Q. (2015). "Estimation of residential water demand under uniform volumetric water pricing." *Journal of Water Resources Planning and Management*, 142(2), 04015054.

Golshani, N., Shabanpour, R., Mahmoudifard, S. M., Derrible, S., and Mohammadian, A. (2017). *Comparison of artificial neural networks and statistical Copula-based joint models*.

Golshani, N., Shabanpour, R., Mahmoudifard, S. M., Derrible, S., and Mohammadian, A. (2018). "Modeling travel mode and timing decisions: Comparison of artificial neural networks and copula-based joint model." *Travel Behaviour and Society*, 10, 21–32.

Goodchild, C. (2003). "Modelling the impact of climate change on domestic water demand." *Water and environment Journal*, 17(1), 8–12.

Goudet, O., Kalainathan, D., Caillou, P., Lopez-Paz, D., Guyon, I., Sebag, M., Tritas, A., and Tubaro, P. (2017). "Learning functional causal models with generative neural networks." *arXiv preprint arXiv:1709.05321*.

Grafton, R. Q., Ward, M. B., To, H., and Kompas, T. (2011). "Determinants of residential water consumption: Evidence and analysis from a 10-country household survey:

DETERMINANTS OF RESIDENTIAL WATER CONSUMPTION." *Water Resources Research*, 47(8).

Guhathakurta, S., and Gober, P. (2007). "The Impact of the Phoenix Urban Heat Island on Residential Water Use." *Journal of the American Planning Association*, 73(3), 317–329.

Guo, J., Feng, T., and Timmermans, H. (2018). *Modeling Co-dependent Choice of Workplace, Residence and Commuting Mode Using an Error Component Mixed Logit Model*.

Guo, J. Y., and Bhat, C. R. (2007). "Population Synthesis for Microsimulating Travel Behavior." *Transportation Research Record*, 2014(1), 92–101.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). "Linear Methods for Regression." *The Elements of Statistical Learning*, Springer New York, New York, NY, 1–57.

Hensher, D. A., Rose, J. M., and Greene, W. H. (2005). *Applied choice analysis: a primer*. Cambridge University Press.

Hensher, D. A., and Ton, T. T. (2000). "A comparison of the predictive potential of artificial neural networks and nested logit models for commuter mode choice." *Transportation Research Part E: Logistics and Transportation Review*, 36(3), 155–172.

House-Peters, L., and Chang, H. (2011). "Urban water demand modeling: Review of concepts, methods, and organizing principles." *Water Resources Research*, 47(5).

House-Peters, L., Pratt, B., and Chang, H. (2010). "Effects of urban spatial structure, sociodemographics, and climate on residential water consumption in Hillsboro, Oregon." *JAWRA Journal of the American Water Resources Association*, 46(3), 461–472.

Iyer, M. S., and Rhinehart, R. R. (1999). "A method to determine the required number of neural-network training repetitions." *IEEE Transactions on Neural Networks*, 10(2), 427–432.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Jentgen, L., Kidder, H., Hill, R., and Conrad, S. (2007). "Energy management strategies use short-term water consumption forecasting to minimize cost of pumping operations." *American Water Works Association. Journal*, 99(6), 86.

Jordan, M. I. (1998). *Learning in graphical models*. Springer Science & Business Media.

Jordan, M. I. (2003). "An introduction to probabilistic graphical models."

Jordan, M. I. (2004). "Graphical models." *Statistical Science*, 19(1), 140–155.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). "An introduction to variational methods for graphical models." *Machine learning*, 37(2), 183–233.

Kahneman, D., and Tversky, A. (2013). "Choices, values, and frames." *Handbook of the Fundamentals of Financial Decision Making: Part I*, World Scientific, 269–278.

Karduni, A., Kermanshah, A., and Derrible, S. (2016). "A protocol to convert spatial polyline data to network formats and applications to world urban road networks." *Scientific data*, 3, 160046.

Keene, O. N. (1995). "The log transformation is special." *Statistics in medicine*, 14(8), 811–819.

Kennedy, C., Pincetl, S., and Bunje, P. (2011). "The study of urban metabolism and its applications to urban planning and design." *Environmental Pollution*, 159(8–9), 1965–1973.

Kenney, D. S., Goemans, C., Klein, R., Lowrey, J., and Reidy, K. (2008). "Residential water demand management: lessons from Aurora, Colorado." *JAWRA Journal of the American Water Resources Association*, 44(1), 192–207.

Kim, B., Rudin, C., and Shah, J. A. (2014). "The Bayesian Case Model: A Generative Approach for Case-Based Reasoning and Prototype Classification." *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds., Curran Associates, Inc., 1952–1960.

Kim, D. K., Kim, D. H., Chang, S. K., and Chang, S. K. (2007). "Modified probabilistic neural network considering heterogeneous probabilistic density functions in the design of breakwater." *KSCE Journal of Civil Engineering*, 11(2), 65–71.

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection." Stanford, CA, 1137–1145.

Koller, D., and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Kontokosta, C. E., and Jain, R. K. (2015). "Modeling the determinants of large-scale building water use: Implications for data-driven urban sustainability policy." *Sustainable Cities and Society*, 18, 44–55.

Kuhn, M., and Johnson, K. (2013). *Applied Predictive Modeling*. Springer New York, New York, NY.

Kullback, S., and Leibler, R. A. (1951). "On information and sufficiency." *The annals of mathematical statistics*, 22(1), 79–86.

Kusiak, A., Li, M., and Zhang, Z. (2010). "A data-driven approach for steam load prediction in buildings." *Applied Energy*, 87(3), 925–933.

Lee, D., Derrible, S., and 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: Journal of the Transportation Research Board, 036119811879697.

Lee, D., Mulrow, J., Haboucha, C. J., Derrible, S., and Shiftan, Y. (2019). “Attitudes on Autonomous Vehicle Adoption using Interpretable Gradient Boosting Machine.” *Transportation Research Record: Journal of the Transportation Research Board*.

Lee, S.-J., Chang, H., and Gober, P. (2015). “Space and time dynamics of urban water demand in Portland, Oregon and Phoenix, Arizona.” *Stochastic environmental research and risk assessment*, 29(4), 1135–1147.

Lee, S.-J., Wentz, E. A., and Gober, P. (2010). “Space–time forecasting using soft geostatistics: a case study in forecasting municipal water demand for Phoenix, Arizona.” *Stochastic Environmental Research and Risk Assessment*, 24(2), 283–295.

Lindenbaum, M., Markovitch, S., and Rusakov, D. (2004). “Selective sampling for nearest neighbor classifiers.” *Machine learning*, 54(2), 125–152.

Lozano, S., and Gutiérrez, E. (2008). “Non-parametric frontier approach to modelling the relationships among population, GDP, energy consumption and CO2 emissions.” *Ecological Economics*, 66(4), 687–699.

Lundberg, S., and Lee, S.-I. (2017). “A Unified Approach to Interpreting Model Predictions.”

Marschak, J. (1950). “Rational behavior, uncertain prospects, and measurable utility.” *Econometrica: Journal of the Econometric Society*, 111–141.

Masters, T. (1995). *Advanced algorithms for neural networks: a C++ sourcebook*. John Wiley & Sons, Inc.

Mayer, P. W., DeOreo, W. B., Opitz, E. M., Kiefer, J. C., Davis, W. Y., Dziegielewski, B., and Nelson, J. O. (1999). “Residential end uses of water.”

Mazzanti, M., and Montini, A. (2006). "The determinants of residential water demand: empirical evidence for a panel of Italian municipalities." *Applied Economics Letters*, 13(2), 107–111.

McFadden, D., and Train, K. (2000). "Mixed MNL models for discrete response." *Journal of applied Econometrics*, 15(5), 447–470.

Miller, T. (2017). "Explanation in artificial intelligence: Insights from the social sciences." *arXiv preprint arXiv:1706.07269*.

Mohammadian, A., and Miller, E. (2002). "Nested logit models and artificial neural networks for predicting household automobile choices: comparison of performance." *Transportation Research Record: Journal of the Transportation Research Board*, (1807), 92–100.

Molnar, C. (2019). "Interpretable machine learning. a guide for making black box models explainable."

Monfort, M., Lake, B. M., Ziebart, B., Lucey, P., and Tenenbaum, J. (2015). "Softstar: heuristic-guided probabilistic inference." 2764–2772.

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

Natekin, A., and Knoll, A. (2013). "Gradient boosting machines, a tutorial." *Frontiers in Neurorobotics*, 7.

Ng, A. Y., and Jordan, M. I. (2002). "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes." 841–848.

Olson, R. S., La Cava, W., Mustahsan, Z., Varik, A., and Moore, J. H. (2017). "Data-driven advice for applying machine learning to bioinformatics problems." *arXiv preprint arXiv:1708.05070*.

Park, H.-S., and Jun, C.-H. (2009). "A simple and fast algorithm for K-medoids clustering." *Expert systems with applications*, 36(2), 3336–3341.

Parzen, E. (1962). "On estimation of a probability density function and mode." *The annals of mathematical statistics*, 33(3), 1065–1076.

"Pattern Recognition and Machine Learning." (2007). *Journal of Electronic Imaging*, 16(4), 049901.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). "Scikit-learn: Machine learning in Python." *Journal of Machine Learning Research*, 12(Oct), 2825–2830.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Model-agnostic interpretability of machine learning." *arXiv preprint arXiv:1606.05386*.

Robert, C. (2014a). "Machine learning, a probabilistic perspective."

Robinson, C., Dilkina, B., Hubbs, J., Zhang, W., Guhathakurta, S., Brown, M. A., and Pendyala, R. M. (2017). "Machine learning approaches for estimating commercial building energy consumption." *Applied Energy*, 208(Supplement C), 889–904.

Rohekar, R. Y. Y., Nisimov, S., Koren, G., Gurwicz, Y., and Novik, G. (2018). "Constructing Deep Neural Networks by Bayesian Network Structure Learning." *arXiv:1806.09141 [cs, stat]*.

Rosca, M., Lakshminarayanan, B., and Mohamed, S. (2018). "Distribution Matching in Variational Inference." *arXiv:1802.06847 [cs, stat]*.

Samek, W., Wiegand, T., and Müller, K.-R. (2017). "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models." *arXiv preprint arXiv:1708.08296*.

Sayed, T., and Razavi, A. (2000). "Comparison of Neural and Conventional Approaches to Mode Choice Analysis." *http://dx.doi.org/10.1061/(ASCE)0887-3801(2000)14:1(23)*, American Society of Civil Engineers.

Schleich, J., and Hillenbrand, T. (2009). "Determinants of residential water demand in Germany." *Ecological economics*, 68(6), 1756–1769.

Semanjski, I., and Gautama, S. (2015). "Smart city mobility application—gradient boosting trees for mobility prediction and analysis based on crowdsourced data." *Sensors*, 15(7), 15974–15987.

Shen, J. (2009). "Latent class model or mixed logit model? A comparison by transport mode choice data." *Applied Economics*, 41(22), 2915–2924.

Shevchuk, Y. (2015). "Neupy: Neural Networks in Python."

Specht, D. F. (1988). "Probabilistic neural networks for classification, mapping, or associative memory." 525–532.

Specht, D. F. (1990). "Probabilistic neural networks." *Neural networks*, 3(1), 109–118.

Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G., and Richardson, T. (2000). *Causation, prediction, and search*. MIT press.

Sucar, L. E. (2015). "Probabilistic Graphical Models." *Advances in Computer Vision and Pattern Recognition. London: Springer London. doi*, 10, 978–1.

Teh, Y. W., and Jordan, M. I. (2010). "Hierarchical Bayesian nonparametric models with applications." *Bayesian Nonparametrics*, N. L. Hjort, C. Holmes, P. Muller, and S. G. Walker, eds., Cambridge University Press, Cambridge, 158–207.

Train, K. (2016). "Mixed logit with a flexible mixing distribution." *Journal of Choice Modelling*, 19, 40–53.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

Tso, G. K. F., and Yau, K. K. W. (2007). "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks." *Energy*, 32(9), 1761–1768.

Tversky, A., and Kahneman, D. (1974). "Judgment under uncertainty: Heuristics and biases." *science*, 185(4157), 1124–1131.

Vij, A., and Krueger, R. (2017). "Random taste heterogeneity in discrete choice models: Flexible nonparametric finite mixture distributions." *Transportation Research Part B: Methodological*, 106, 76–101.

Vitter, J. S., and Webber, M. E. (2018). "A non-intrusive approach for classifying residential water events using coincident electricity data." *Environmental Modelling & Software*, 100, 302–313.

Wainwright, M. J., and Jordan, M. I. (2008). "Graphical models, exponential families, and variational inference." *Foundations and Trends® in Machine Learning*, 1(1–2), 1–305.

Wang, F., and 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: Journal of the Transportation Research Board*, 036119811877355.

Willis, R. M., Stewart, R. A., Panuwatwanich, K., Williams, P. R., and Hollingsworth, A. L. (2011). "Quantifying the influence of environmental and water conservation attitudes on household end use water consumption." *Journal of Environmental Management*, 92(8), 1996–2009.

Wisetjindawat, Derrible, and Kermanshah. (2018). "Modeling the Effectiveness of Infrastructure and Travel Demand Management Measures to Improve Traffic Congestion During Typhoons." *Transportation Research Record: Journal of the Transportation Research Board*, in press.

Wolman, A. (1965). "The metabolism of cities." *Scientific American*, 213(3), 178–193.

Wong, M., Farooq, B., and Bilodeau, G.-A. (2017). "Latent behaviour modelling using discriminative restricted Boltzmann machines." *5th International Choice Modeling Conference*.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Xie, C., Lu, J., and 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), 50–61.

Yi, J.-H., Wang, J., and Wang, G.-G. (2016). "Improved probabilistic neural networks with self-adaptive strategies for transformer fault diagnosis problem." *Advances in Mechanical Engineering*, 8(1), 168781401562483.

Yuan, Y., You, W., and Boyle, K. J. (2015). "A guide to heterogeneity features captured by parametric and nonparametric mixing distributions for the mixed logit model."

Yurdusev, M. A., Firat, M., and Turan, M. E. (2010). “Generalized regression neural networks for municipal water consumption prediction.” *Journal of Statistical Computation and Simulation*, 80(4), 477–478.

Zhang, K., Schölkopf, B., Spirtes, P., and Glymour, C. (2017). “Learning causality and causality-related learning: some recent progress.” *National Science Review*, 5(1), 26–29.

Zhou, Z.-H. (2012). *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC.

APPENDIX A: Copyright permission

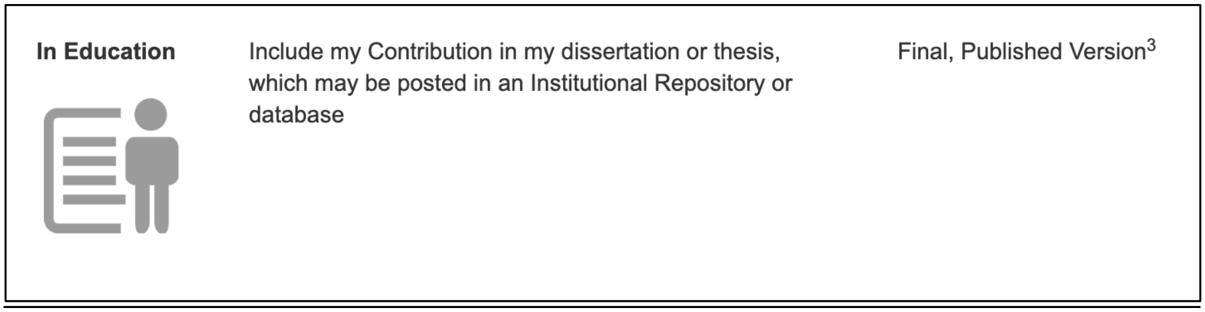
Copyright Statement from SAGE Journal (Transportation Research Record)

For an overview of the below information, please view our printable guide to re-use policies here.

Authors Re-Using Their Own Work

SAGE Journal authors are able to reuse their Contribution in certain circumstances without requiring permission from SAGE. Most SAGE journals allow authors to reuse their Contributions in accordance with SAGE's Green Open Access policy, which is detailed below. If a journal's reuse policy is an exception to the standard policy, the specific terms for that journal will be supplied in the author's Contributor Agreement; for a list of titles with policy exceptions, see below.

Note: For Contributions published as SAGE Choice, or via Gold Open Access journals, see Reusing Open Access and SAGE Choice Content. Authors of papers published under a Creative Commons license (either in an Open Access journal or under the SAGE Choice option) with a non-commercial (NC) or no derivatives (ND) specification may reuse their work under the terms of the Creative Commons license attached to their article and additionally may reuse their work as stated under the Green Open Access policy below. For a list of SAGE's Gold Open Access journals, please see Gold Open Access journals.



In Education	Include my Contribution in my dissertation or thesis, which may be posted in an Institutional Repository or database	Final, Published Version[3]

VITA

NAME: Dongwoo Lee

EDUCATION: B.S., Transportation Engineering, University of Seoul, South Korea, 2011

M.S., Transportation Engineering, University of Seoul, South Korea, 2013

WORK EXPERIENCE: Argonne National Laboratory, U.S., 2015-2016

Urban Science Research Center, Seoul, 2013-2014

ACADEMIC PRESENTATIONS: TRB 98th Annual Meeting, Transportation Research Board, 2019

International Conference on New Horizons in Green Civil Engineering (NHICE-01), 2018

TRB 97th Annual Meeting, Transportation Research Board, 2018

International Society for Industrial Ecology, 2017

Transport Chicago, 2017

HONORS: Korean Science and Engineer Association (KSEA) the best conference paper awards, 2019

David Boyce Scholarship Award, University of Illinois, Chicago, 2018

Korean Science and Engineer Association (KSEA) the best conference paper awards, 2018

Brain Korea 21 (BK21), National Scholarship for Science and Engineering Fields, Ministry of Education, Korea 2011-2013

PROFESSIONAL MEMBERSHIP: Transportation Research Board

Illinois Transportation Engineers

Korean Science and Engineer Association

PUBLICATIONS: Lee, D., Derrible, S., “Predicting Residential Water Demand with Machine-Based Statistical Learning”, ASCE Journal: Journal of Water Resources Planning and Management, In press.

Lee, D., Mulrow, J., Haboucha, C.J., Derrible, S., Shiftan, Y., 2019 “Attitudes on Autonomous Vehicle Adoption using Interpretable Gradient Boosting Machine”, Transportation Research Record: Journal of the Transportation Research Board.

Trelles Trabucco, J., Marai, G.E., Lee, D., Derrible, S., 2019 "Visual Analysis of a Smart City's Energy Consumption", Multimodal Technologies Interact.

Lee, D., Derrible, S., and 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: Journal of the Transportation Research Board.