MULTI-LEVEL PROBABILISTIC MODEL FOR POPULATION SYNTHESIS AND VEHICLE OWNERSHIP MODELING BASED ON SAMPLES WITH MISSING VALUES

Zhenyuan Ma

Department of Civil Engineering

McGill University, Montreal

December 2020

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science.

Contents

Acknowledgements **5**

Contribution of Authors **5**

Abstract **7**

Résumé **9**

1 Introduction **11**

1.1 Research Background 11

1.2 Population Synthesis 13

1.3 Research Scope and Objectives 17

1.4 Thesis Organization 18

2 Literature Review **20**

2.1 Population Synthesis 20

2.2 Multi-Level Modeling on Population 22

2.3 Missing Value Imputation 23

3 Probabilistic Modeling and Missing Value Imputation **26**

3.1 Basic Single-Level Model 26

3.2 Multi-Level Model 28

3.3 Model Inference 29

3.4 Model Selection 32
3.5 Missing Value Imputation 34
3.6 PUMS Data Description 37
3.7 Imputation Experiments 37

4 Case Study: Vehicle Ownership Modeling **42**
4.1 Background 42
4.2 NHTS Data Description 44
4.3 Prediction on Vehicle Ownership 45
4.4 Prediction on the Joint Attributes that Contain Vehicle Ownership 51

5 Conclusion and Discussion **56**

References **58**

List of Figures

1.1 Schematic diagram for multi-level PUMS census data. 14

3.1 Model selection results. (a) and (b) are for person-level latent class, while (c) and (d) are for household-level. 34

3.2 Schematic diagram of individuals with missing values. (a) gives a complete individual and (b) shows two complete attributes. 39

4.1 Schematic diagram of vehicle ownership modeling. (a) input X is fully observed and (b) input X is partially missing. 47

4.2 Prediction accuracy on HHVEHCNT with fully observed input. 49

4.3 Prediction accuracy on HHVEHCNT with 20% missing in input. 50

4.4 Prediction accuracy on HHVEHCNT with 50% missing in input. Data is imputed by mean imputation method before prediction. 51

4.5 Correlation between HHVEHCNT and other attributes. 52

4.6 Expected and actual prediction results for four comparison methods on different attribute pairs with fully observed input. 54

4.7 Expected and actual prediction results for four comparison methods on different attribute pairs with 20% missing in input. 55

List of Tables

1.1 PUMS data multi-level samples. 15

1.2 PUMS data multi-level samples with missing values. 15

3.1 PUMS data samples with household-level ignored. 27

3.2 Person-level and household-level attributes selected in PUMS data. 38

3.3 Accuracy of our model and KNN methods on missing value imputation. 41

3.4 Accuracy of our model on missing value imputation when using different number of attributes. 41

4.1 Attributes selected related to vehicle ownership in NHTS data. 46

Acknowledgements

Foremost, I would like to pay my special regards to my supervisor, Prof. Lijun Sun, for providing guidance and recommendations during this 2-year period of my master's research life. His patience and encouragement help me going on and trying my best to finish this thesis work. I cannot imagine a better supervisor than Prof. Lijun Sun, and without him, I may not make it through the master degree. Moreover, I wish to show my gratitude to all my fellow labmates in the smart transport group. They provide me inspiration and practical advice on my thesis. The physical and technical help of all staff in the Civil Engineering department is truly appreciated. They are so friendly for always reminding me to seek their help when I have difficulties. Last but not least, I would like to acknowledge the love of my family, who always gives me support, attention, and understanding in these years.

Contribution of Authors

In this thesis, both my supervisor Prof. Lijun Sun and I contribute. Prof. Lijun Sun and I designed the research and developed the methodologies. I collected and preprocessed the data, carried out the experiments, analyzed the results, and wrote the thesis. Prof. Lijun Sun also provided guidance and proofreading to this thesis.

Abstract

Agent-based modeling has become increasingly important in urban transportation planning. Census and survey data are useful in providing data sources for agent-based simulation and modeling the characteristics and distributions of population and attributes. However, public datasets such as the Public Use Microdata Sample (PUMS) and the National Household Travel Survey (NHTS) contain only a small fraction of the total population. The objective of population synthesis is to generate fake agents with similar characteristics and distributions that served as a replacement for the census or survey data. Besides, the common missing data problem in the census and survey data may cause the available samples to be less representative and may lead to only a few useful samples for analysis. To address the confidentiality and missingness in the census and survey data, we model the population in household-based survey data using a multi-level probabilistic model. Moreover, as an unsupervised method, it can also model and predict vehicle ownership in a supervised way. By analyzing vehicle ownership based on census and survey data, our model can predict the number of cars in a household given related attribute information, and provide data sources on solving how vehicle ownership affect land use and travel behavior. In this thesis, we mainly focus on the missing value imputation and vehicle ownership prediction problem in the census and survey data, particularly when the missing rate of the input data is large. Experimental results show that our model has good performances on both tasks. We achieve superior performance compared with existing methods such as K-Nearest Neighbors methods on imputing missing values at random, where few models can work. The advantage of applying our model on vehicle ownership prediction appears when the input data contains missing values. Moreover, when we pay atten-

tion on vehicle ownership and other attributes altogether, our model can predict directly without making additional efforts.

Résumé

La modélisation basée sur les agents est devenue de plus en plus importante dans la planification du transport urbain. Les données de recensement et d'enquête sont utiles pour fournir des sources de données pour la simulation basée sur des agents et modéliser les caractéristiques et les distributions de la population et des attributs. Toutefois, les ensembles de données publiques tels que le Public Use Microdata Sample (PUMS) et le National Household Travel Survey (NHTS), ne contiennent qu'une petite fraction de la population totale. L'objectif de la synthèse de la population est de générer des faux agents ayant des caractéristiques et des distributions similaires qui ont servi de remplacement aux données du recensement ou de l'enquête. En outre, le problème commun des données manquantes dans le recensement et les données d'enquête peut rendre les échantillons disponibles moins représentatifs et peut conduire à quelques échantillons utiles pour l'analyse. Pour répondre à la confidentialité et à la non-confidentialité dans les données du recensement et de l'enquête, nous modélisons la population en données d'enquête sur les ménages à l'aide d'un modèle probabiliste à plusieurs niveaux. En outre, en tant que méthode non supervisée, elle peut également modéliser et prédire la propriété du véhicule de manière supervisée. En analysant la propriété des véhicules sur la base de données de recensement et d'enquête, notre modèle peut prédire le nombre de voitures dans un ménage donné des informations d'attributs connexes, et fournir des sources de données sur la résolution de la façon dont la propriété du véhicule affecte l'utilisation des terres et le comportement de déplacement. Dans cette thèse, nous nous concentrons principalement sur l'imputation de la valeur manquante et le problème de prédiction de la propriété des véhicules dans le recensement et les données d'enquête, en particulier lorsque le taux manquant

des données d'entrée est important. Les résultats expérimentaux montrent que notre modèle a de bonnes performances sur les deux tâches. Nous obtenons des performances supérieures par rapport aux méthodes existantes telles que les méthodes K-Nearest Neighbors pour imputer des valeurs manquantes au hasard, où peu de modèles peuvent fonctionner. L'avantage d'appliquer notre modèle à la prévision de la propriété du véhicule apparaît lorsque les données d'entrée contiennent des valeurs manquantes. De plus, lorsque nous accordons une attention particulière à la propriété du véhicule et à d'autres attributs, notre modèle peut prédire directement sans faire d'efforts supplémentaires.

Chapter 1

Introduction

1.1. Research Background

Recently, agent-based modeling (ABM) has become an increasingly important tool in urban transportation planning. ABM simulates individuals' decisions and activities over time, provides detailed and accurate information, and reflects the relationships among behaviors in the real world. The first step to build an agent-based model is to prepare agent data sources with attributes that will affect individuals' travel behaviors and decisions. The most ideal data should contain the complete information of the whole population, but these kinds of data are not available in the real world. Another ideal agent data sources are the census and survey data, which characterize and represent the whole population, while other research methodologies may not achieve this goal. However, because of the privacy issue, the census and survey data are confidential and highly sensitive. Directly using the data will affect the simulation process and result in replicated samples in the population.

One of the most popular census and survey data is the Public Use Microdata Sample (PUMS) provided by the American Community Survey. The files are a set of untabulated records about both individual people and housing units (American Community Survey, 2018). Because of the confidentiality issue, PUMS only contains about 1% of the US population from the actual responses for each year. As a result, only PUMS may not be sufficient for better analysis. In this case, a model that could make full use of the available data should be created so that we can have more accurate analysis to make essential conclusions. A critical point to mention is that we can model the data

distributions and generate more samples for future use, since originally we do not have enough data. The model can help in describing complex data and producing simulation results that are systematic, interpretable, and easy to understand. What is needed to model the census and survey data is population synthesis.

Population synthesis is defined as using the available census and survey data and generating samples to represent the whole population. The samples generated have the same attribute structure as the target population, and they are called the synthetic population. Currently, population synthesis is less of a modeling exercise (Rich et al., 2019). It focuses more on aggregating geographic and sociodemographic attributes from different data sources and providing detailed analysis for further planning. For transportation planning applications, we would like to understand how sociodemographic attributes will affect travel patterns and behaviors, and then we could implement such knowledge into agent-based simulation models.

When the census and survey data are collected, there will be missing values in the data for reasons. In general, respondents may skip the questions that they are not willing to answer, or they may ignore the requirements and thus wrongly answer the questions. Both cases are common and result in a large number of missing values in the census and survey data. Another reason for missing values appeared in data is that people have mistakes when preprocessing the raw data source, for instance, wrongly defining the data type or deleting good values by accident. The appearance of missing values in the census and survey data can be avoided to some extent. However, it is still a critical problem to be solved since we require a detailed and representative dataset.

In order to build ABM for further applications, the requirements for input data are crucial. Data deficiency motivates researchers to study more on population synthesis and how to deal with missing values in the census and survey data. We would like to fully understand the available data in the

presence of missing values. If we could contextually impute the missing values, the characteristics of the synthetic population will be more accurate compared with the target population.

1.2. Population Synthesis

Population synthesis is the process of generating a synthetic population of individuals and households such that the synthetic population represents a target population with respect to geographic and sociodemographic attributes (Rich et al., 2019). The synthetic population can be used as the input for some traditional agent-based transport simulation and travel behavior modeling problems such as activity prediction. Since modeling on the population will provide the characteristics or distributions of the population and corresponding attributes, population synthesis is commonly used to find the social and geographical relationship of the population (Borysov et al., 2019). The motivation of population synthesis is to generate fake but representative synthetic data in order to have more samples to do the simulation, and accurately characterize the population. From the census data that only contain 1% of the population, we can model the available data for multinomial distributions on attributes and individuals. After that, we can generate 100% fake synthetic individual samples that served as a replacement for the census or survey data by using these distributions, with privacy and confidentiality being protected (Sun et al., 2018).

Naturally, the simulated agents can be households or individuals for agent-based travel demand forecasting, depending on the continuing study direction (Müller and Axhausen, 2011). Census data provide both household and individual information, and we can aggregate them. Since there is a containing relationship among households and individuals, we cannot analyze them separately, and thus we need a multi-level model to consider more information. Multi-level modeling helps in obtaining more accurate characteristics of the population. Recently, more studies concentrate on using inner relationships among different levels to improve their analysis. Public Use Microdata

Sample (PUMS) provides record files for both individual people and housing units. Figure 1.1 shows the schematic diagram for multi-level PUMS census data. The most top level denotes the whole population in the US. The housing units level is in the middle part, while the individual people level is under the housing units level. Data samples can be viewed in Table 1.1. In general, the PUMS dataset has two levels: person level and household level. Each level contains corresponding attributes in which we are interested. By definition, the person-level attributes belong to the household-level, since the household contains individual persons. Table 1.1 shows that one household has at least one individual person, and the persons in one household share the same household-level attribute values.

FIGURE 1.1 : Schematic diagram for multi-level PUMS census data.

However, one critical problem in the census and survey data, which is overlooked in previous studies, is the missing data problem. Table 1.2 shows the situation of actual PUMS data, which contains lots of missing values compared with Table 1.1. Missing values may appear in any attributes except `PID` and `HID` (i.e. person ID and housing ID). In PUMS data, `Age` in the person-level and the number of persons in household `NP` in the household-level are also fully observed. Most of the individual samples have at least one missing attribute, and we name them 'incomplete samples'.

TABLE 1.1 : PUMS data multi-level samples.

PID	Age	Sex	Marital	Edu	Wages	HID	NP	Tenure	Veh	Income
1	80-89	F	Widowed	College	$25-50k	1	2	Occupied	1	$100-150k
2	39-39	F	Single	College	$75-100k					
4	50-59	F	Married	Bachelor	$100-125k	2	3	Rented	1	$100-150k
5	40-49	M	Married	Bachelor	$25-50k					
6	10-19	M	Single	Grade 8	$0-25k					
7	60-69	F	Divorced	Associate	$0-25k	3	2	Owned	2	$0-50k
8	30-39	M	Single	College	$25-50k					
i	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	$x_{i,5}$	h_i	$x_{i,6}$	$x_{i,7}$	$x_{i,8}$	$x_{i,9}$
...	...	...	...	...	...					

The individual samples with all attributes observed are rare, and we define them as 'complete samples'. The complete sample example in Table 1.2 is the person with `PID` equals to 4, while the other seven samples are incomplete.

TABLE 1.2 : PUMS data multi-level samples with missing values.

PID	Age	Sex	Marital	Edu	Wages	HID	NP	Tenure	Veh	Income
1	80-89	F	Widowed	-	$25-50k	1	2	-	1	$100-150k
2	39-39	-	Single	College	$75-100k					
4	50-59	F	Married	Bachelor	$100-125k	2	3	Rented	1	$100-150k
5	40-49	M	-	Bachelor	-					
6	10-19	-	Single	Grade 8	-					
7	60-69	-	Divorced	Associate	$0-25k	3	2	Owned	-	$0-50k
8	30-39	M	-	-	$25-50k					
i	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	$x_{i,5}$	h_i	$x_{i,6}$	$x_{i,7}$	$x_{i,8}$	$x_{i,9}$
...	...	...	...	...	...					

Missing data is a common problem in the census and survey data. Missing values in data will cause a few useful samples, and the available samples will also be less representative. Several factors are leading to the missing data problem. For example, the design of the questionnaire by default contains options such as "not to answer". Some choices in the question such as "do not

know" and "not willing to respond" will be declared as missing (de leeuw, 2001). Also, wrongly respondings such as answering words in numerical question, or male answers the question only for females, will be marked as missing when statistic software deals with the data. When performing preprocessing of data, it will happen that some values are invalid or out of range.

Therefore, it is critical to develop solutions that could handle missing values while avoiding the distortion of population information. The understanding of missing data will also improve the way of data collection in the future (de leeuw, 2001). Besides fully observed data, samples with missing values also contain information in both missing attributes and observed attributes. All the data will help in learning underlying distributions of the population, and better characterizing the population.

Population synthesis has been well-studied in existing literature, including statistical learning and deep learning methods for modeling population. Some studies using the Iterative proportional fitting algorithm (Frick, 2004) and combinatorial optimization methods (Williamson et al., 1998) suffer from high-dimensionality and scalability problems, while some deep generative models (Borysov et al., 2019) need high computational resources and are not interpretable. Recent models consider more on the correlation within multi-level (Sun et al., 2018) or on the scalability and efficiency of models (Borysov et al., 2019). However, few of the published papers consider the missing value problem in the census and survey data. They usually remove or discard data samples with missing values for convenience, and therefore resulting in losing the information of data attributes. As for the modeling, previous works generally use household information and at most the personal information of the household head. In reality, members of the household also contribute to the number of vehicles in the household. Moreover, the information of different years and regions could be borrowed to some extend (Adjemian and Williams, 2009).

1.3. Research Scope and Objectives

In this thesis, we propose a probabilistic model for multi-level population synthesis and two popular applications. This model is a latent class model that could capture the patterns for both individuals and households, and their attributes. Multivariate multinomial distributions represent the patterns that can be used to generate synthetic population samples served as input in agent-based simulation models and urban transportation planning. Also, these distributions are useful resources to analyze the clustering of households and individuals, as well as how attributes, such as age and income, vary among different classes of households and individual, or different years and regions.

Instead of generating synthetic households, our model uses synthetic individuals to represent the whole population. In this case, all the household members contribute, rather than using only the single householder. By using all the persons in one household, we fully utilize the relationship between individuals and the multi-level idea. Nevertheless, individuals could borrow information from the other samples in the same class, and they are also affected by the persons in the same household.

The expectation-maximization (EM) algorithm is implemented for better model inference. We also select the best combination of hyperparameters through different model selection criteria. We need to define the range of hyperparameters and test on different combinations to find the best setting. Some criteria are used for parameter selection.

The first problem that our model could solve is missing value imputation. Because of the incomplete census and survey data, we need to make full use of the available data to analyze or even impute the missing values. As a probabilistic model, our model gives the joint probability of attributes. We can impute according to the latent class of one individual and its household, as well as its other attributes. It could solve the randomly missing case and achieves high performance

even with a larger missing rate. Therefore, there is no doubt that our model could solve more straightforward cases, such as with a low missing rate or when only specific attributes are partially missing. Our model is flexible and reusable when it comes to the selected attribute information. All the attributes of interest can be taken into account. When a new attribute comes in the future with only a few available data, we can still infer the distributions and characteristics.

Our second application is vehicle ownership modeling, which is an extension of the combination of modeling on population and missing value imputation. Actually, it is a supervised learning problem with the presence of missing values in input data, but our unsupervised learning model also works. Instead of using the analyzed distribution, we can model directly on the original data, and predict missing vehicle ownership values. Comparing with the traditional supervised learning model, we achieve better performance when the input data contain missing values. The result shows that our model is valuable since the missing value problem is common for survey and census data. Moreover, if we want to understand more than only vehicle ownership, or equivalently there are missing values in two or more attributes, traditional supervised learning models cannot handle it. Usually, supervised learning models ignore the correlation and model on attributes separately, or combine attributes into a new attribute. However, we can directly apply our model on the input data without any further process, and get results with no additional computation resources.

1.4. Thesis Organization

The remainder of this thesis is structured as follows. Chapter 2 provides some previous works. It includes statistical learning and deep learning methods to solve population synthesis problems, and how models make full use of the household data files and the individual data files to perform multi-level modeling. Literature about the two main applications, which are missing value imputation and vehicle ownership prediction, is also provided. Chapter 3 gives our probabilistic model, from

simple single-level to complex multi-level, and shoes how we perform the EM algorithm to find the optimal solution in the model inference part. How we do missing value imputation and the experimental results are also available in this chapter. In chapter 4, we use the same model presented in chapter 3.2, and apply our unsupervised learning model on a supervised learning problem, which is vehicle ownership modeling. Our model can predict not only one single vehicle ownership, but also two or more attributes containing vehicle ownership. Comparing to other traditional supervised learning models, we achieve superior performance, especially on the dataset with missing values. Finally, chapter 5 concludes this thesis and gives some future directions to be researched.

Chapter 2

Literature Review

In this chapter, we will review related literature on population synthesis, and also focus on one essential problem - missing values imputation.

2.1. Population Synthesis

Modeling on population mainly focuses on fitting a few available data into a desirable model and generating individual and household samples from learned distributions. There have been many existing methods for fitting a model, including statistical learning and deep generative learning.

Generally, statistical methodologies estimate the weights or some latent factors for attributes. Bar-Gera et al. (2009) applied Entropy Maximization methodology to find household factor weights to match the given distribution. The method could estimate the weights under reasonable limitations of the population, but it is not scalable for high-dimensional data. Deming and Stephan (1940) applied Iterative Proportion Fitting (IPF) method on 1940 census data. They used the data samples as the initialization and then estimated the characteristics for the whole population to find the cell frequencies for certain individuals and certain attributes. From the samples, Deming and Stephan (1940) made one estimate for each dimension and used a matrix fitting way to combine all the estimates, and enforced conditions on the marginal totals. The limitation of IPF is that it only samples from current data rather than creating synthetic samples. In this case, samples are only replicated from the original data, keeping the weights and distributions. However, what we want in agent-based modeling are newly observed samples different from the original data sources, and

are randomly drawn from the underlying distributions about population.

A common problem for traditional statistical methods is that they replicate samples instead of estimating the underlying distributions. However, the goal of population synthesis is to characterize the joint probabilistic distributions. Probabilistic models, in this case, have advantages in describing the population. Recent statistical learning methods use simulation to solve the generalization problem, flexibly generating synthetic samples and allowing effective computation for high-dimensional data. Sun and Erath (2015) proposed a graphical model to find the joint distributions of attributes and generate synthetic data from the underlying probabilities. Hidden Markov Model (HMM) (Saadi et al., 2016) is used to model the correlation into a sequence or a chain structure, with all the attributes have their transition matrices. There is no specific ordering of attributes to be sampled when using HMM, but the probability of the latter attribute should be calculated conditional on the former ones. However, the statistical methods mentioned above suffer from scalability issues, and some cannot work when there are missing values in data. Also, hyperparameters such as the number of latent factors and the number of states need to be decided. Bhattacharya and Dunson (2012) mentioned a way for the latent factor model to decide the number of latent factors automatically. Nonparametric Bayes helps define a prior of attribute distributions, and it is more flexible for different kinds of latent structures. Nonparametric Bayes provides advantages in the large dataset and among different choosing criteria (Dunson and Xing, 2009), while for the Bayesian network (BN) proposed by Barash et al. (2003), there was a tendency to be overfitting in large model space.

The combination of deep learning and generative models could deal with a high-dimensional and large dataset. Variational Autoencoders (VAEs) (Kingma and Welling, 2013) and Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) are popular methods in deep learning

area. Applying deep neural network structures will provide detailed modeling and better results than general statistical models with the same dataset. Borysov et al. (2019) tried to embed the VAE framework into population synthesis. They outperformed in high-dimensional cases, when using as many as attributes to describe the population, and tried to generate complex agent information. However, these deep generative models need higher computational costs and resources. Interpretability is another challenge when deep learning models are compared to traditional statistical models. Deep learning models also suffer from the overfitting problem. When we have limited resources, it is necessary to consider the trade-off between model accuracy and generalization.

2.2. Multi-Level Modeling on Population

Previous studies that consider the population by only using household distributions ignore the relationship between households and individuals within households. Since different levels provide different information and attribute correlations, a multi-level model should be considered to make full use of the multi-level data information. Hu et al. (2014) estimated the joint distributions of the population by a nested Dirichlet model. They assumed two separate latent structures for the group (i.e., household) and the unit (i.e., individual in a household), where the unit is nested within the group. This model tends to cluster data samples and finds the dependence among attributes and levels for a better analysis. However, Hu et al. (2014) did not use the prior information of the population, and they mentioned that the model was not suitable for complex attribute structures.

Sun et al. (2018) also proposed a hierarchical model that consists of the household level and individual level. Each level has its own latent structure, and the attributes are independent among levels. The population is characterized by the households and all the individuals within the households, and the assumption is that individuals are conditional on households. By producing the marginal distributions for all the attributes and latent classes on each level, the model could calcu-

late the joint probability and generate the synthetic population. Sun et al. (2018) used a universal individual latent class, assuming that different households share the same individual latent classes. Therefore the model is simplified for their case study. However, the model suffers from the scalability issue and the process of variable selection.

2.3. Missing Value Imputation

Census and survey data contain missing values due to many factors, and the appearance of missing values will highly impact the characterizations of the population. Some existing publications have already paid attention to the missing value problem in the census and survey data. Barnett et al. (2017) used a simple statistical method and they modeled patterns to determine question types and examine the group of people who may miss similar questions. By using the rules to cluster samples, they identified and understood the respondents, and learned from the results to design future surveys. The work of Barnett et al. (2017) was a trial on using missing values rather than simply discarding them.

Karanja et al. (2013) summarized how researchers in Management Information Systems handle missing data for survey data. According to their study, three generations of techniques have been used to estimate, predict, or recover missing values. The 1st generation techniques such as Listwise Deletion (Gilley and Leone, 1991) and Pairwise Deletion (Marsh, 1998) remove variables with missing values. The 2nd generation discards the samples with missing values and utilizes fully observed samples. Meanwhile, the 3rd generation of methods, such as expectation-maximization (Dempster et al., 1977) and Full Information Maximum Likelihood (Enders and Bandalos, 2001), try to preserve as much data as possible. Liao et al. (2014) introduced a concept named as imputability, which defined whether a missing value is imputable by borrowing information from other values. Their results showed that the imputed values with low imputability would affect the

imputation accuracy and analysis of the input data.

Young et al. (2011) provided a survey about popular machine learning methods. Methods such as nearest neighbors (Chen and Shao, 2000) can handle multiple missing values with few complete data samples. K-Nearest Neighbors (Batista and Monard, 2003) is a generalization of the nearest neighbor, where the missing values are imputed according to the majority of k neighbors. Expectation-Maximization algorithm (Dempster et al., 1977) and Artificial Neural Network (Gupta and Lam, 1996) are conventional machine learning methodologies, trying to add as little as information about data and capture the trends to improve the imputation accuracy.

Moreover, Vermunt et al. (2008) proposed a latent class model for multiple categorical data imputation. The model is based on log-linear modeling (Schafer, 1997), especially for cross-classified categorical data, but it is improved to high-dimensional data with lots of variables. The latent class model is a flexible model both for data type and interpretability. It allows a mixture of any distributions for variables and cluster data using different latent classes. However, rather than the prediction accuracy of missing values, Schafer (1997) cared more about the coefficient estimate results on the logit model of one dependent variable. If we would like to apply the latent class model on complex datasets with multilevel or multiple-group, Vermunt (2003) provided a graphical latent class model to analyze the associations among different groups.

Similarly, there is a multiple imputation solution for large assessment survey data (Si and Reiter, 2013). When there is a large number of variables to be used, methods such as log-linear modeling (Schafer, 1997), cannot handle the interdependency among variables. In this case, Schafer (1997) used a fully Bayesian modeling for multiple imputations, which was related to the work of Dunson and Xing (2009) and Vermunt et al. (2008), the one using the Dirichlet process and the one using a mixture of distributions, respectively.

Another way to impute missing values is matrix factorization. The most popular application is on the recommendation system (Koren et al., 2009) with sparse user-movie matrix data. Matrix factorization also finds patterns and clusters the samples by borrowing information from samples in the same class. It can be used in large-scale dataset (Mitra et al., 2010) and survey data such as voting (Agathokleous and Tsapatsoulis, 2013). Moreover, matrix factorization is extended to deal with different kinds of data such as binary, categorical, and continuous data, by embedding probabilistic modeling (Hernández-Lobato et al., 2014) or Bayesian methods (Yang and Dunson, 2016).

Chapter 3

Probabilistic Modeling and Missing Value Imputation

The probabilistic model presented in this chapter is based on the model of Sun et al. (2018) and works on both the person-level and household-level. Instead of generating synthetic household samples, our model focuses on individuals and uses individual information conditional on household information. Moreover, our model can impute all the missing values that appeared in the input data to make full use of the available samples, and provide a more accurate characterization of the population based on the census and survey data in the presence of missing values. In this chapter, we will start from the single-level model, and then show the multi-level model and how it performs missing value imputation. Model selection on hyperparameters will also be provided, and the superior results on missing value imputation compared to the K-Nearest Neighbors model.

3.1. Basic Single-Level Model

In the single-level model, we care about all the individual and household information rather than the subordinate relationship between individuals and households. By simply assigning household-level attributes to the belonging individuals, we ignore the multi-level information and different latent structures of each level. The sample data are shown in Table 3.1.

The input of the model consists of the attribute values of both individuals and households. We use $i = 1, ..., N$ to index the individuals and $j = 1, ..., J$ to index all the attributes. The value of individual i belongs to attribute j is denoted by $x_{i,j}$. The vector $\mathbf{x}_i = (x_{i,1}, x_{i,2}, ..., x_{i,j}, ..., x_{i,J})$ denotes the full information of individual i, containing both person-level and its household-level

TABLE 3.1 : PUMS data samples with household-level ignored.

PID	Age	Sex	Marital	Edu	Wages	NP	Tenure	Veh	Income
1	80-89	F	Widowed	College	\$25-50k	2	Occupied	1	\$100-150k
2	30-39	F	Single	College	\$75-100k	2	Occupied	1	\$100-150k
4	50-59	F	Married	Bachelor	\$100-125k	3	Rented	1	\$100-150k
5	40-49	M	Married	Bachelor	\$25-50k	3	Rented	1	\$100-150k
6	10-19	M	Single	Grade 8	\$0-25k	3	Rented	1	\$100-150k
7	60-69	F	Divorced	Associate	\$0-25k	2	Owned	2	\$0-50k
8	30-39	M	Single	College	\$25-50k	2	Owned	2	\$0-50k
i	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	$x_{i,5}$	$x_{i,6}$	$x_{i,7}$	$x_{i,8}$	$x_{i,9}$
...	...	...	...	...	...	...	...	...	...

attributes. In the survey and census data, most of the questions are provided with categorical choices, or answers are be easily categorized. As a result, we treat all the values of attributes as categorical, either nominal or ordinal, and define $x_{i,j} \in \{1,...,d_j\}$ where d_j is the number of categories in attribute j and the set of possible categorical values is a discrete set starting from one. For each individual, we assume that it belongs to a latent class $z_i \in \{1,...,M\}$, and let π_m be the probability that the individual i belongs to latent class m for any m (i.e. $\pi_m = \Pr(z_i = m)$). Moreover, we use $\theta_{c_{i,j},m} = \Pr(x_{i,j} = c_{i,j}|z_i = m)$ to represent the probability of the truly observed category value of individual i in attribute j is $c_{i,j}$, conditional on that individual i belongs to latent class m.

In the probabilistic model, we would like to find the best parameter setting that could maximize the probability of a individual which is $\Pr(\mathbf{x}_i)$ for individual i where $\mathbf{x}_i = (x_{i,1},...,x_{i,J})$. We assume there is independence among each attribute j conditional on a selected latent class. We can model the joint distribution as:

$$\Pr(\mathbf{x}_i) = \sum_{m=1}^{M} \pi_m \cdot \left[\prod_{j=1}^{J} \theta_{c_{i,j},m}\right] = \sum_{m=1}^{M} \Pr(z_i = m) \cdot \left[\prod_{j=1}^{J} \Pr(x_{i,j} = c_{i,j}|z_i = m)\right] \tag{3.1}$$

The joint probability can be viewed as a latent class model, in which each attribute has a multinomial distribution for its categorical values. In this case, the model works similarly to a factorization model where low-dimensional factors are used to model high-dimensional data efficiently. The decrease of latent classes will reduce the number of parameters in modeling high-dimensional data and easily model an individual using simpler latent class distribution and attribute distributions. We use person-level attributes and household-level attributes together and only consider one latent class structure in the single-level. However, there are differences between person-level and household-level in general. Also, individuals in the same household will definitely affect other household members. In this case, we need a more complex model to capture the patterns of different levels of attributes to understand and utilize available data.

3.2. Multi-Level Model

In the multi-level model, we consider the subordinate relationship between households and individuals. The data samples are shown in Table 1.1. The individuals in the same household share the same household-level attribute values and have their own person-level attributes. Using a multi-level model, we can capture the latent class distribution of each level and the relationship between the two levels.

Similar to the single-level model, we use $i = 1, ..., N$ to index the individuals. In total we have J attributes for each individual i. We define $j = 1, ..., k$ for the person-level attributes (e.g., `AGEP`: age, `SCHL`: educational attainment, ...) and $j = k+1, ..., J$ for the household-level attributes (e.g., `VEH`: number of vehicles in household, `HHT`: household / family type, ...). We use $x_{i,j} \in \{1, ..., d_j\}$ to show the possible discrete set of category values for individual i and attribute j, where d_j is the number of categories in attribute j. Normally, person-level attributes and household-level attributes should have different latent distributions, thus we consider a multi-level model to show detailed

statistical information. We use $p = 1, ..., P$ and $h = 1, ..., H$ to denote the latent class at person-level and at household-level respectively. We then use $\Pr(z_{p,i} = p)$ and $\Pr(z_{h,i} = h)$ to denote the probabilities that individual i belongs to person-level latent class p and the household of individual i belongs to household-level class h for any p and h. In this case, the joint probability of individual i becomes:

$$\Pr(\mathbf{x}_i) = \left\{ \sum_{h=1}^{H} \Pr(z_{h,i} = h) \cdot \left[\prod_{j=k+1}^{J} \Pr(x_{i,j} = c_{i,j} | z_{h,i} = h) \right. \right. \\ \left. \left. \cdot \left(\sum_{p=1}^{P} \Pr(z_{p,i} = p | z_{h,i} = h) \cdot \prod_{j=1}^{k} \Pr(x_{i,j} = c_{i,j} | z_{h,i} = h, z_{p,i} = p) \right) \right] \right\} \quad (3.2)$$

The idea is that one individual i firstly belongs to a household where the household has its household-level latent class $z_{h,i}$, following a multinomial distribution. Within a household, the individual i has its own person-level latent class $z_{p,i}$. For simplicity, we assume independence among person-level attributes and household-level attributes conditional on their latent classes.

3.3. Model Inference

We use one of the commonly used methods, which is the expectation-maximization (EM) algorithm for parameter estimation. EM algorithm (Dempster et al., 1977) is used to find the maximum a posterior estimate of parameters through an iterative way when equations of the model cannot be directly solved. It generally takes the derivatives of the expectation of log-likelihood function with respect to the unknown latent parameters, and solves the derivatives for the maximum log-likelihood function. In the Expectation step, we define the expectation of the log-likelihood function using the current parameters and compute the expected value by using the current estimates of parameters. In the Maximization step, we update the parameters by maximizing the

expectation of the log-likelihood function defined in the Expectation step. The parameters calculated in the Maximization step are used for the expectation in the next iteration. We start from random initialization and perform the EM procedure until convergence.

We assume the latent classes of individuals and corresponding attributes all follow independent multinomial distributions:

$$\begin{aligned} z_{h,i}|\pi &\sim \text{Multi}(\pi_1,...,\pi_H) \text{ for all } i \\ z_{p,i}|z_{h,i},\omega &\sim \text{Multi}(\omega_{h,1},...,\omega_{h,P}) \text{ for all } i \\ x_{i,j}|z_{h,i},\theta &\sim \text{Multi}(\theta_{1,h},...,\theta_{d_j,h}) \text{ for all } i, j = k+1,...,J \\ x_{i,j}|z_{h,i},z_{p,i},\phi &\sim \text{Multi}(\phi_{1,h,p},...,\phi_{d_j,h,p}) \text{ for all } i, j = 1,...,k \end{aligned} \tag{3.3}$$

For simplicity, we use the notations provided in Equation 3.3 to represent the joint probability. Let $\pi_h = \Pr(z_{h,i} = h)$ denote the probability of the household of individual i belongs to household-level latent class h, and $\omega_{h,p} = \Pr(z_{p,i} = p|z_{h,i} = h)$ denote the probability of individual i belongs to person-level latent class p if its household belongs to household-level latent class h. For notations related to the attributes, $\theta_{c_{i,j},h} = \Pr(x_{i,j} = c_{i,j}|z_{h,i} = h)$ represents the probability of truly observed household-level attributes is equal to $c_{i,j}$, for any $j = k+1,...,J$ given the household-level latent class. Similarly, $\phi_{c_{i,j},p,h} = \Pr(x_{i,j} = c_{i,j}|z_{h,i} = h, z_{p,i} = p)$ is the probability of truly observed person-level attributes for any $j = 1,...,k$ conditionally on both household-level and person-level latent classes. Then we could simplify the joint distribution as Equation 3.4:

$$\Pr(\mathbf{x}_i) = \left\{ \sum_{h=1}^{H} \pi_h \cdot \left[\prod_{j=k+1}^{J} \theta_{c_{i,j},h} \cdot \left(\sum_{p=1}^{P} \omega_p \cdot \prod_{j=1}^{k} \phi_{c_{i,j},h,p} \right) \right] \right\} \tag{3.4}$$

The EM procedure for our probabilistic model is shown as follows:

E step. We define $\lambda_{i,h} = p(z_{h,i} = h|\mathbf{x}_i)$ as the expectation of the household of individual i belongs to household-level latent class h given the full information of this individual i. Similarly, $\lambda_{i,p} = p(z_{p,i} = p|\mathbf{x}_i, z_{h,i} = h)$ works as the expectation of the individual i belongs to person-level latent class p, given the full information of individual i, and conditional on the household of this individual belongs to household-level latent class h. We can calculate the expectation as follows:

$$\lambda_{i,h} = \frac{\pi_h \cdot \left[\prod_{j=k+1}^{J} \theta_{c_{i,j},h} \cdot \left(\sum_{p=1}^{P} \omega_p \cdot \prod_{j=1}^{k} \phi_{c_{i,j},h,p}\right)\right]}{\sum_{h=1}^{H} \pi_h \cdot \left[\prod_{j=k+1}^{J} \theta_{c_{i,j},h} \cdot \left(\sum_{p=1}^{P} \omega_p \cdot \prod_{j=1}^{k} \phi_{c_{i,j},h,p}\right)\right]} \tag{3.5}$$

$$\lambda_{i,p} = \frac{\omega_p \cdot \prod_{j=1}^{k} \phi_{c_{i,j},h,p}}{\sum_{p=1}^{P} \omega_p \cdot \prod_{j=1}^{k} \phi_{c_{i,j},h,p}} \tag{3.6}$$

M step. We use the expectation calculated in the Expectation step to update our distribution parameter π_h, ω_p, $\theta_{c_{i,j},h}$ and $\phi_{c_{i,j},h,p}$ using following equations:

$$\begin{aligned}
\pi_h &= \frac{\sum_{i=1}^{N} \lambda_{i,h}}{\sum_{i=1}^{N}\sum_{h=1}^{H} \lambda_{i,h}} \\
\omega_p &= \frac{\sum_{i=1}^{N} \lambda_{i,h}\lambda_{i,p}}{\sum_{i=1}^{N}\sum_{p=1}^{P} \lambda_{i,h}\lambda_{i,p}} \\
\theta_{c_{i,j},h} &= \frac{\sum_{i=1}^{N} \lambda_{i,h} \times \mathbb{I}(x_{i,j} = c_{i,j}|z_{h,i} = h)}{\sum_{i=1}^{N}\sum_{h=1}^{H} \lambda_{i,h}}) \\
\phi_{c_{i,j},h,p} &= \frac{\sum_{i=1}^{N} \lambda_{i,h}\lambda_{i,p} \times \mathbb{I}(x_{i,j} = c_{i,j}|z_{h,i} = h, z_{p,i} = p)}{\sum_{i=1}^{N}\sum_{p=1}^{P} \lambda_{i,h}\lambda_{i,p}}
\end{aligned} \tag{3.7}$$

We randomly initiate the values for π_h, ω_p, $\theta_{c_{i,j},h}$ and $\phi_{c_{i,j},h,p}$, and use the EM algorithm to update

these four parameters. The algorithm will stop when convergence, and then it will provide the optimal values of the parameters.

3.4. Model Selection

There are two hyperparameters in this probabilistic model: the number of person-level latent classes z_p and the number of household-level latent classes z_h. By using the criteria such as Bayesian information criterion (BIC), perplexity, and Standard Root Mean Square Errors (SRMSE), we may find the most appropriate values for z_p and z_h after testing on different values.

Bayesian information criterion is formally defined by Schwarz (1978):

$$BIC = \ln(n)k - 2\ln(L) \tag{3.8}$$

where L is the likelihood of the model, n is the sample size, and k is the number of parameters estimated by the model. When the number of latent class increases, the penalty term $\ln(n)k$ will increase while term $-2\ln(L)$ will decrease. By using the penalty term, the overfitting problem will be avoided. More parameters may result in better modeling, but it will increase the complexity of a model. However, there are limitations when using BIC. BIC is suitable when the number of samples N is much larger than the number of parameters k. Also, it cannot deal with complex collections of models in high-dimensional data (Giraud, 2014).

Perplexity is a measurement for a probabilistic model showing how well the model predicts a sample. An interpretation for perplexity is to what extent the model can use the learned patterns to predict future behavior (Sun et al., 2019). The lower the perplexity is, the prediction will be better. Here, we use a formula that is similar to using cross-entropy (Goodfellow et al., 2016), and it is

calculated by:

$$perplexity = b^{H(p)} = b^{-\sum_{x_i} \frac{1}{N} \log_b p(x_i)} = b^{-\frac{1}{N} \sum_{x_i} \log_b p(x_i)} \tag{3.9}$$

where b is 2 in most cases. In the cross-entropy $H(p)$, the empirical distribution of the samples $p(x_i)$ is $\frac{1}{N}$, that is, a uniform distribution. Lower perplexity represents a better job as it requires less information for future samples.

The other popular method is a distance-based Standard Root Mean Square Errors (SRMSE) defined by Müller and Axhausen (2011):

$$SRMSE = \frac{\sqrt{\prod_j c_j \times \sum_j (F_j - N_j)^2}}{\sum_j N_j} \tag{3.10}$$

where c_j is the number of categories for attribute j. F_j, and N_j are the number of synthetic individuals and true population respectively.

Figure 3.1 presents the model selection results. We run experiments for different combinations of z_p and z_h, and separately show the results. Apparently, SRMSE is not a suitable evaluation metric in our case. It will show an increasing trend for either z_p or z_h, so that fewer latent classes will be better. However, too few numbers of latent classes will not model the distributions of attributes well. From Figure 3.1 (a) and 3.1 (b), we could also find out the increase trends for the number of person-level latent classes. Since we use a hierarchical probabilistic model and the person-level is under the household-level, a linear increase of z_p will result in a polynomial increase in perplexity and BIC rather than a similar linear increase.

The convex appeared in Figure 3.1 (c) gives an appropriate range of z_h, which is $(6, 10)$, considering both perplexity and BIC. From Figure 3.1 (a) and 3.1 (b), we could not find the appropriate

values for z_p. While a smaller number of the person-level latent class will provide a better performance, it will result in a rough characterization of the distributions. We would like to choose a median number of person-level latent classes to balance the trade-off between the performance of the model and the characterization of person-level attributes.

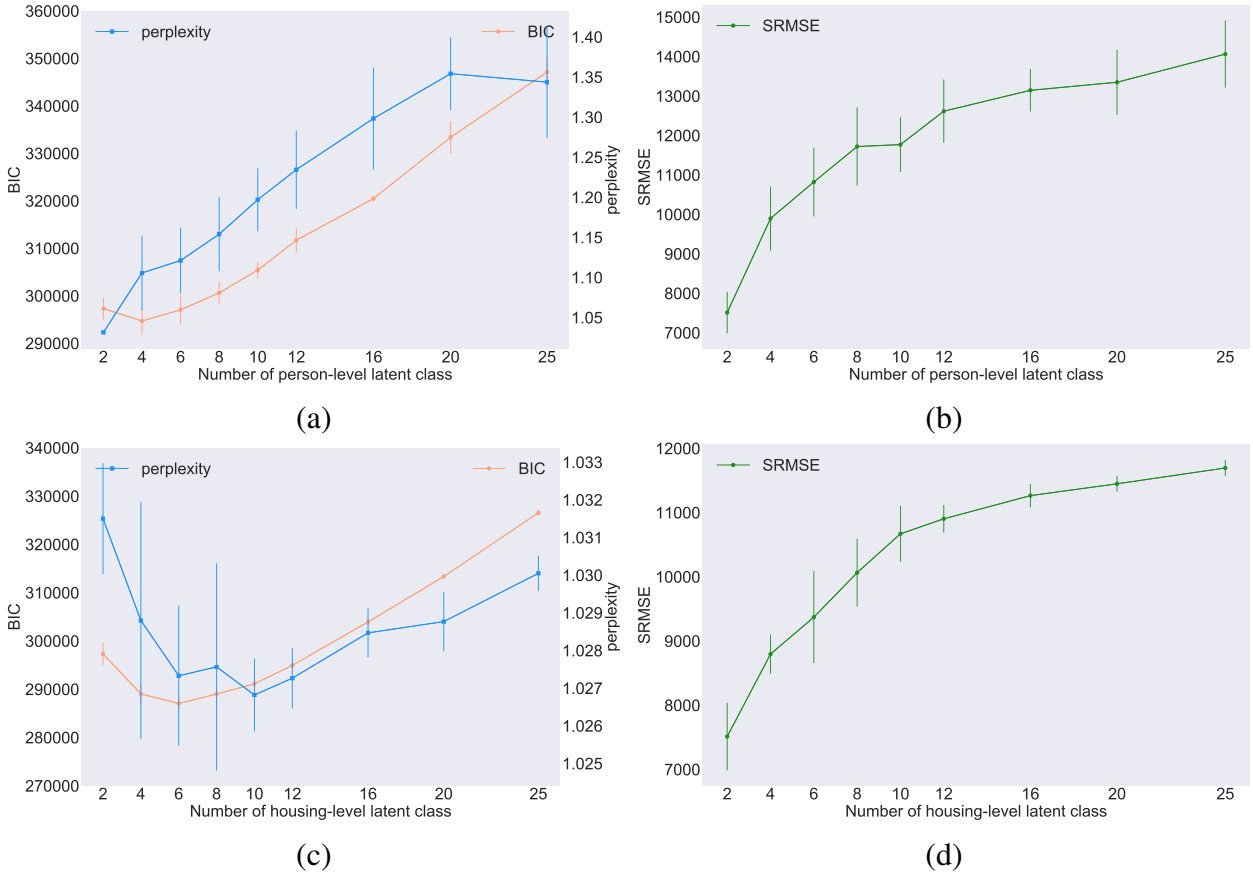


FIGURE 3.1 : Model selection results. (a) and (b) are for person-level latent class, while (c) and (d) are for household-level.

3.5. Missing Value Imputation

Using the probabilistic model, we could deal with the census and survey data that contain missing values, and we could also impute these missing values. We assume that the set of missing values

follows the same distributions as the set of the observed values. Thus, we can use the calculated distributions of all the attributes, individuals, and households to impute the probabilities for all the missing values. Imputing missing values by using the calculated distributions will not change the structure and information of the census and survey data. It is similar to the mean imputation method that the mean value of the dataset is unchanged. However, by using distributions and latent class information, we will have more accurate imputation results with a large variance to represent the whole population than the mean imputation method. We can get the distributions of each attribute value for one individual by using the following equation:

$$\Pr(x_{i,j}) = \begin{cases} \sum_{h=1}^{H} \pi_h \cdot \theta_{c_{i,j},h} & \text{,for } j \in [k+1, J] \\ \sum_{h=1}^{H} \pi_h \left[\sum_{p=1}^{P} \omega_p \cdot \theta_{c_{i,j},h,p} \right] & \text{, for } j \in [1, k] \end{cases} \tag{3.11}$$

Note that the person-level attributes and household-level attributes are separately calculated using different equations, since we use a multi-level probabilistic model and different latent structure between two levels. After getting the array of probability for each value, we have several criteria to select one value as the missing one. To select the one with the highest probability is an interpretable way.

Our model works in an unsupervised learning way to find patterns of individuals, households, and attributes. It imputes missing values without explicitness of how the attributes contribute. A couple of methods can be used to impute missing values, such as mean or median imputation, K-Nearest Neighbors, and matrix decomposition. However, there are additional requirements for imputing in the census and survey data. Firstly, most of the attribute values in the census and survey data can be treated as nominal categorical values. Also, the imputation results should make sense when compared to the context of the original dataset.

Mean imputation and median imputation are simple to implement and easy to understand. These two approaches provide good performances for some specific datasets because they do not change the mean or median value of the dataset. However, they cause a bias problem and decrease the variance of data samples. Moreover, in reality, these data being imputed should have different values since the census and survey collect data from various respondents. Merely using the mean or median value to substitute missing data also ignores the relationship among different attributes. For example, the number of vehicles in a household is related to household income and whether there are drivers' licenses for household members.

Because of the categorical attribute values, the matrix decomposition method may not work. Popular matrix decomposition applications on categorical datasets are either 0-1 binary or ordinal data (i.e., ordered categorical data). However, for census and survey data, most of the attribute values are nominal, which means unordered categorical data. In this case, the imputation results of matrix factorization may ignore the contextual meaning of attribute values.

Finally, we choose K-Nearest Neighbors (KNN) compared with our probabilistic model, which works fine in missing value imputation for the nominal categorical dataset. KNN has been proved to be simple and effective for numeric input, and it surpasses commonly used and simple methods such as mean imputation and median imputation (Kuhn et al., 2013). KNN does not create an explicit model (Batista and Monard, 2003), and it is flexible when adding or changing data samples or attributes. Also, it can deal with missing values randomly appeared in multiple attributes. However, since KNN will find the nearest samples, the training process costs much more time. We need at least one complete sample row in the data so that all other rows contain missing values will be imputed according to the complete row. When the missing rate becomes higher, there is a chance that there is no complete row in the data. In this case, we will consider a tricky process

by adding one sample using the mean or median values and treating it as the complete sample. The added sample originally should not exist, and the reason that it is added is to make the KNN method work. Although this sample does not change the mean value of the dataset, it will affect the performance and make the accuracy higher than real.

3.6. PUMS Data Description

PUMS dataset contains samples from the actual responses to the American Community Survey (ACS). Person-level files and household-level files are provided to better analyze people within their families and with other household members. Dataset for each year, for example, `data2017` contains about 1% of the US population, from nearly every town and county.

We select several attributes from the raw dataset, including those containing individual and household information and those related to the transportation area. The selection process refers to Hu et al. (2014), Potoglou and Kanaroglou (2008) and Bhat and Pulugurta (1998). Details of the attributes and the possible values of each attribute are shown in Table 3.2. The first half of attributes are person-level attributes, while the rest half are household-level attributes. Some of the attributes are categorical. For instance, the ability to speak English represented by `ENG` has four categories: `1. very well`, `2. well`, `3. not well`, and `4. not at all`. While some attributes such as `AGE` and `WAGP` can be treat as continuous. We restrict the number of categories for each attribute to a maximum of ten when running experiments for simplicity. The discretizing procedure uses the same size of intervals between every two categories.

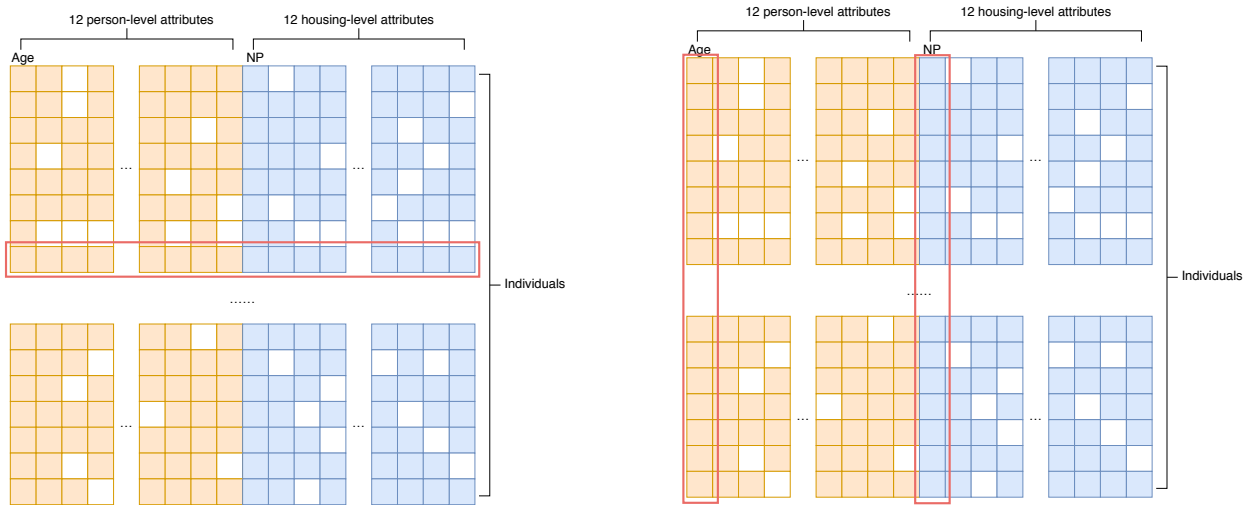
3.7. Imputation Experiments

In the imputation experiments, we consider missing values randomly appeared in data. We generate a full observation dataset by simply deleting all individuals that contain at least one missing value

TABLE 3.2 : Person-level and household-level attributes selected in PUMS data.

name	description	possible value
AGEP	age	00 - 99
ENG	ability to speak English	1 - 4
JWRIP	vehicle occupancy	1 - 10
MAR	marital status	1 - 5
RELP	relationship	00 - 17
SCHL	educational attainment	01 - 24
SEX	sex	1 - 2
WAGP	wages or salary income past 12 months	000000 - 999999
HICOV	health insurance coverage record	1 - 2
HISP	recorded detailed Hispanic origin	01 - 24
PINCP	total person's income	-019998 - 4209995
RAC1P	recorded detailed race code	1 - 9
NP	number of persons in this household	00 - 20
TYPE	type of unit	1 - 3
ACR	lot size	1 - 3
TEN	Tenure	1 - 4
VEH	vehicles available	0 - 6
FES	family type and employment status	1 - 8
FINCP	family income (past 12 months)	-059999 - 9999999
FPARC	family presence and age related children	1 - 4
HHT	household / family type	1 - 7
HINCP	household income (past 12 months)	-059999 - 9999999
WIF	workers in family during the past 12 months	0 - 3
WORKSTAT	work status of householder or spouse in family households	01 - 15

from the original PUMS dataset. After combining all regional data in the year 2017, we have one sample dataset `data2017`. Originally, the missing rate for data in 2017 is 17.76%. However, the proportion of complete data samples is only 2.59%, which means in 40 individuals, we could only find one fully observed individual sample. Most individuals have at least one attribute value missing. The schematic diagram of the data is shown in Figure 3.2. There is only one complete individual among the given individuals, which is framed by red in Figure 3.2 (a), while the missing rate is 17.1% (i.e., similar to PUMS dataset). Additionally, attribute `AGE` in person-level and `NP` in household-level are fully observed, which is clearly shown in 3.2 (b).



(a) (b)

FIGURE 3.2 : Schematic diagram of individuals with missing values. (a) gives a complete individual and (b) shows two complete attributes.

For consistency of the categories in each attribute, the preprocessing procedures on data are performed after combining. Missing values will be created randomly in all possible person-level attributes and household-level attributes, with different missing rates. Then we could apply our model to the data and output the multinomial distribution of the attributes and the latent classes.

We can impute the missing values by using the methods we have introduced previously.

The first experiment is on the different maximum number of categories of each attribute. It is rare for categorical survey data that the number of categories for each attribute is too large. Even for continuous data, we can preprocess and categorize them. For attributes such as SEX, it originally has two categories (i.e., male and female). But for continuous data related to income, such as WAGP, PINCP, FINCP, and HINCP, the number of categories will be more than 20,000. In the experiment, we can treat continuous data as categorical directly or discretize into fewer categories. Since the computation time of using the original number of categories will be more than 50 times than categorizing it into 25 categories, we tend to discretize the data.

The results are shown in Table 3.3. The accuracies of our model are always above or around 60%, regardless of the missing rate and the maximum number of categories. Even when 70% of the data are missing, we could still reach a 59% accuracy when the number of categories is 25. For our model, a large number of categories will reduce the imputation accuracy, but the decrease is not large. However, general KNN does not provide satisfactory results. The accuracy drops quickly after the missing rate is larger than 30%. For a missing rate larger than 50%, there is a chance that data do not contain a complete individual sample so that we could not impute according to nearest neighbors.

We also examine the effect of the number of attributes used when performing missing value imputation. In the original PUMS dataset, we have 286 person-level attributes and 230 household-level attributes. We only choose a small fraction of attributes that are related to the transportation domain. The following three sets of attributes shown in Table 3.4 are randomly chosen. From the results, we could conclude that more attributes used will increase the imputation accuracy. The result of more information provided will counteract the result of more possible attribute values to

TABLE 3.3 : Accuracy of our model and KNN methods on missing value imputation.

# of categories	model	missing rate						
		0.1	0.2	0.3	0.4	0.5	0.6	0.7
10	our model	0.676	0.675	0.673	0.669	0.667	0.665	0.661
	KNN	0.556	0.547	0.522	0.202	-	-	-
15	our model	0.611	0.645	0.643	0.641	0.64	0.639	0.636
	KNN	0.493	0.49	0.448	0.342	0.205	-	-
20	our model	0.626	0.622	0.621	0.619	0.614	0.613	0.609
	KNN	0.461	0.451	0.358	0.188	-	-	-
25	our model	0.61	0.61	0.606	0.603	0.598	0.595	0.592
	KNN	0.471	0.465	0.393	0.192	-	-	-
original	our model	0.502	0.496	0.499	0.494	0.495	0.492	0.486
	KNN	0.377	0.374	0.329	0.194	-	-	-

be imputed. However, there should be an upper bound of accuracy regardless of the number of attributes used. Also, too many uncorrelated attributes used may possibly result in a decrease in imputation accuracy.

TABLE 3.4 : Accuracy of our model on missing value imputation when using different number of attributes.

# of attributes	missing rate						
	0.1	0.2	0.3	0.4	0.5	0.6	0.7
9 person-level	0.495	0.495	0.502	0.502	0.44	0.508	0.506
10 person-level + 5 household-level	0.587	0.58	0.575	0.572	0.567	0.562	0.557
12 person-level + 12 household-level	0.676	0.675	0.673	0.669	0.667	0.665	0.661

Chapter 4

Case Study: Vehicle Ownership Modeling

4.1. Background

With the development of urban transport technology and the increase of household income, vehicle ownership in a family grows rapidly. The development of society and the urban transportation system are highly related to the demand for vehicles (Basu and Ferreira, 2020). The member composition and income of households will affect the number of vehicles in a household. Vehicle ownership reversely indicates how the life of a household and sustainability to some extent. In the person-level, it reflects a person's travel behavior, showing which mode he uses to school, to work, or for other personal issues. The preference of a person, the income, and whether the person has a driver's license will affect using vehicles. In the household-level, more attributes are related to vehicle ownership. Not only the householder but also other members of the household will contribute. The total income, the existence of children and older adults, and the household location will also affect the number of vehicles in one household. Moreover, other social attributes such as the price of both vehicles and gasoline and government policy have an effect on people who are willing to have a car but hesitate when buying it. The analysis of vehicle ownership will become an indicator. Thus, government and industries could make decisions on public transport and gasoline prices and plan the land use and transport arrangement. Furthermore, more opportunities will be provided to employees if the demand for vehicles grows.

Vehicle ownership modeling is a hot topic to be researched and a special case that our model can

solve. The average number of cars owned per household will highly impact land use, vehicle, and energy pricing, and, most importantly, travel behavior (Sinha, 2003). It determines the travel mode of family members and the number of trips. Ryan and Han (1999) suggested that the structure of the household and the attributes should be taken into account when analyzing vehicle ownership. They used the 1990 census data to model vehicle ownership, compared it with other American cities, and found that the results were similar and could be borrowed by other regions. There is a proxy model considering both travel survey data and census files to characterize area-level vehicle ownership (Adjemian and Williams, 2009). By aggregating data information, the model could avoid the problem of insufficient data samples and non-representative samples.

Useful factors for vehicle ownership prediction are provided by Cirillo and Liu (2013), Ryan and Han (1999), and Adjemian and Williams (2009). Commonly used household attributes are household income, household size, household type, number of members who have a driver's license, and number of workers in a household. Since the prediction is on the household-level, only the personal information of the household head contributes. These kinds of attributes include the education level, race, and Hispanic status. Besides, region-level information such as household location, urban size, gasoline price, and population size will count. In some population synthesis papers written by Hu et al. (2014), Potoglou and Kanaroglou (2008), and Bhat and Pulugurta (1998), these attributes also contribute to the characterization of the household. Using similar attributes inspires us to use population synthesis to characterize on vehicle ownership, and then perform prediction and analysis based on other attributes provided.

Most previous studies use self-defined logit equations and supervised learning methods to do vehicle ownership modeling (Basu and Ferreira, 2020). Cirillo and Liu (2013) generated different utility functions for selected attributes and used the estimated coefficients to determine the per-

formance. Liu et al. (2014) used discrete choice modeling to extract information about vehicle ownership, vehicle type, and vehicle usage decisions. However, this kind of modeling requires humans to define functions, categories, and metrics. In this case, preferences will be added to vehicle ownership analysis.

In principle, vehicle ownership modeling using the census and survey data is a supervised learning problem in the presence of missing values. Our model, which is an unsupervised learning model, can work in this case. Not only can it predict vehicle ownership, but it can also model the distributions and characteristics of vehicle ownership. Moreover, when we care about more than vehicle ownership, our model can model these attributes altogether and give satisfying results without consuming more computational resources. This case study chapter mainly focuses on vehicle ownership, which is an essential attribute in understanding households and individuals. However, our model can generalize to all attributes we may care about.

4.2. NHTS Data Description

For vehicle occupancy, we consider another dataset: the National Household Travel Survey (NHTS) (Federal Highway Administration (FHWA), 2017) conducted by the Federal Highway Administration. Comparing to the PUMS dataset, the NHTS dataset focuses more on travel behaviors. The survey is used to analyze the travel information in the US, and plan for urban infrastructures and selling in facilities. It is conducted with an interval of several years, and we use the eighth survey, which is `2017NHTS` file. The NHTS data contain information about the individual person, household, vehicle, and daily travel trip. Four parts of the data are related to the attributes in the household data file.

To research on the vehicle ownership, we select some related attributes appeared in four data files and list in Table 4.1. `HHVEHCNT` is the number of vehicles in one household, and it is

the attribute we care about. All the values of attributes can be viewed as categorical values, either nominal or ordinal. In addition, category values which are negative numbers are different types of missing values: for example, `'-1'` means `'Appropriate Skip'`, `'-7'` represents `'prefer not to answer'` and `'-8'` for `'do not know'`. We treat these values all as missing when doing experiments.

4.3. Prediction on Vehicle Ownership

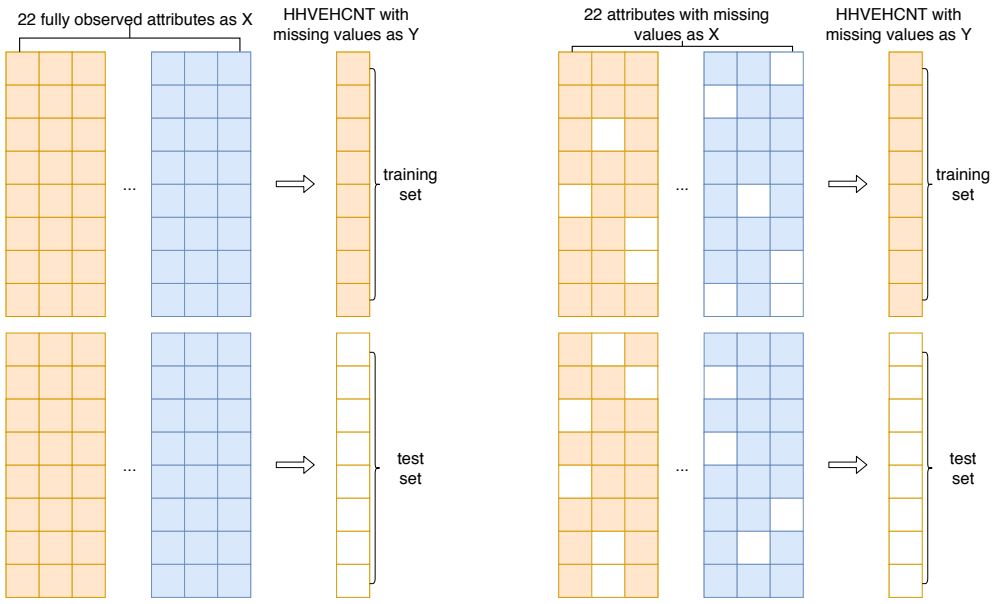
Prediction on vehicle ownership based on other available attributes information is a traditional supervised learning problem. The schematic diagram is shown in Figure 4.1. The other 22 attributes work as input, and the attribute `HHVEHCNT` with unknown missing values works as output. All the complete samples are divided into the training set, while samples with unknown `HHVEHCNT` are in the test set. Ideally, in supervised learning, input X should be fully observed to preserve information and predict what we care about. However, in our census and survey dataset, missing values appear as shown in Figure 4.1 (b). Thus, the problem is to model and predict vehicle ownership based on attributes with missing values.

There exist a massive amount of supervised learning algorithms, and they could achieve satisfying performance in modeling vehicle ownership. Some supervised learning methods are used as a comparison with our model, including Multinomial Logistic Regression (MLR), Random Forest (RF), Ensemble of learners (ENS), and Naive Bayes (NB) methods. We also use a machine learning method K-Nearest Neighbors (KNN) to compare.

Multinomial logistic regression (Nelder and Wedderburn, 1972) is a generalization of logistic regression for multi-class prediction problem. Similarly to our model, it will output the probabilities of possible categories as a multinomial distribution, and we can use the one with the highest probability as the output result. Random forest (Ho, 1995) is an ensemble method using multiple

TABLE 4.1 : Attributes selected related to vehicle ownership in NHTS data.

name	description	Appeared in data file			
		Person	Household	Vehicle	Trip
HHVEHCNT	count of household vehicles	✓	✓	✓	✓
CNTTDHH	count of household trips on travel day	✓			
DRVRCNT	number of drivers in household	✓	✓	✓	✓
HBHTNRNT	percentage of occupied housing in the census block of household	✓	✓	✓	✓
HBHUR	urban / rural of block group	✓	✓	✓	✓
HBRESDN	housing units per square miles in the census block of household	✓	✓	✓	✓
HHFAMINC	household income	✓	✓	✓	✓
HHSIZE	count of household members	✓	✓	✓	✓
HH_HISP	Hispanic status of household respondent	✓	✓	✓	✓
HH_RACE	race of household respondent	✓	✓	✓	✓
HOMEOWN	home ownership	✓	✓	✓	✓
LIF_CYC	life cycle classificationfor the household	✓	✓	✓	✓
MSACAT	Metropolitan Statistical Area (MSA) category	✓	✓	✓	✓
MSASIZE	population size of Metropolitan Statistical Area (MSA)	✓	✓	✓	✓
NUMADLT	count of adult household members	✓	✓	✓	✓
PRICE	price of gasoline affects travel	✓			
RAIL	MSA heavy rail status for household	✓	✓	✓	✓
URBAN	household's urban area classification	✓	✓	✓	✓
URBANSIZE	urban area size where home address is located	✓	✓	✓	✓
URBRUR	household in urban / rural area	✓	✓	✓	✓
WEBUSE	frequency of internet use	✓			
WRKCOUNT	number of workers in household	✓	✓	✓	✓
YOUNGCHILD	count of persons with age between 0 and 4 in household	✓			



(a) (b)

FIGURE 4.1 : Schematic diagram of vehicle ownership modeling. (a) input X is fully observed and (b) input X is partially missing.

decision trees, and output an average prediction result of the decision trees. In the training step, the random forest method applies the bagging technique to multiple tree learners and decreases variance to achieve a better result. For ensemble of learners, we consider adaptive boosting method named AdaBoostM2 (Eibl and Pfeiffer, 2005). It trains learners sequentially and uses the weighted pseudo-loss for multiple classes prediction with weak base classifiers. Naive Bayes classifier assumes independence among input attributes and considers equal contributions of each attribute. It is a conditional probabilistic model, and it only requires a small number of training data to estimate (Murty and Devi, 2011).

For supervised learning methods, the task is to make predictions on the test set after training the model on the training data, as shown in Figure 4.1. On the other hand, for unsupervised learning methods such as our model and KNN, the task is to impute missing values that appeared in data, like shown in Figure 3.2, without splitting data into training/test set. There is no difference between the results of supervised learning and unsupervised learning methods, although they have different names or definitions. For unsupervised learning methods, by rearranging the samples in the dataset, we can treat all the samples with missing in certain attribute as a set, where it is the same as the test set in supervised learning methods. In this case, the term 'missing rate' in unsupervised learning is the same as the term 'test data rate' in supervised learning.

Figure 4.2 shows the prediction accuracy results under the ideal situation, where the input is fully observed. Compared to the five baseline methods, our model does not work very well as an unsupervised learning method, although it is not the worst. Our model is better than the Naive Bayes classifier, which is also a probabilistic model. However, the baseline models are well-defined for many years and suitable for prediction with no missing values in the input.

As for the real case, we test on input with 20% of missing values. This missing rate of 20%

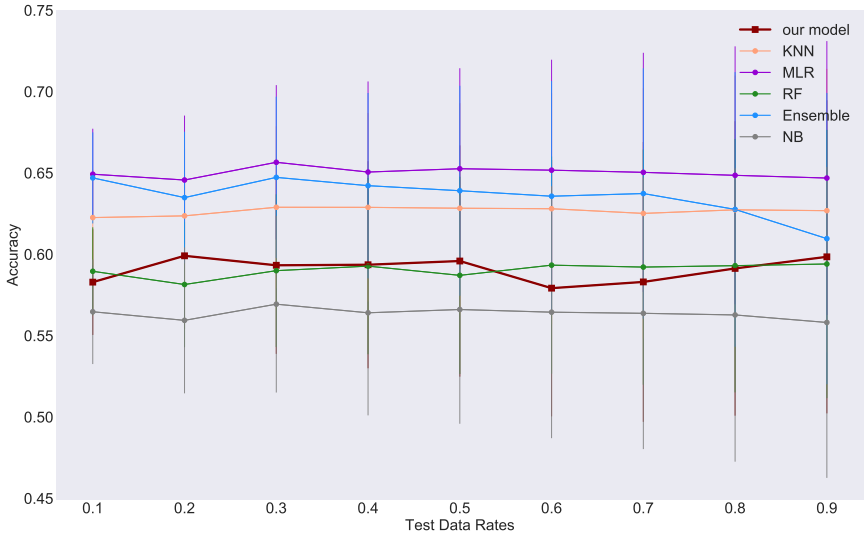


FIGURE 4.2 : Prediction accuracy on `HHVEHCNT` with fully observed input.

is similar to the missing data rate of the PUMS dataset, which means that the 20% missing rate is ordinary for some famous census and survey data. As shown in Figure 4.1 (b), even with 20% of missing values, there are only a few numbers of complete samples. In this case, the simplest way to handle missing values is to discard the incomplete samples, and this is the way how these baselines supervised learning methods deal with `'NaN'`s (i.e., missing values) in Matlab (MathWorks, 2020). Using the input that contains 20% missing values, our model, K-Nearest Neighbors, and random forest have good performances since they all can make full use of the available data samples, although samples contain missing values. However, discarding incomplete samples will lose information and decrease the accuracy, shown in the results of multinomial logistic regression, the ensemble of learners using AdaBoostM2 and Naive Bayes.

When it comes to input with 50% of missing values, there are nearly no complete samples. The reason to consider a high missing rate is that we may add new attributes that have few available data. These newly added attributes are helpful in describing the population, but they may only be

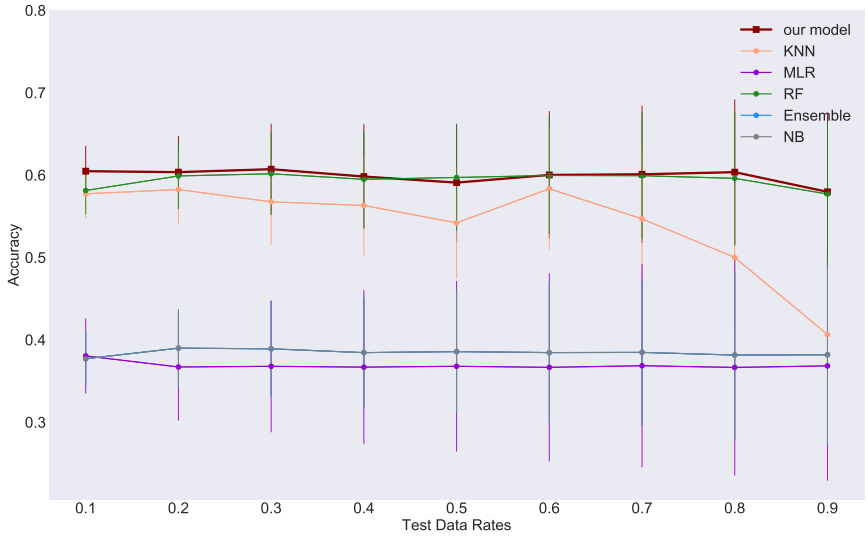


FIGURE 4.3 : Prediction accuracy on `HHVEHCNT` with 20% missing in input.

available for recent years. The extremely high missing rate in new attributes will result in a high missing rate in the whole dataset. In this case, only our model can work. We try to impute all the missing values using mean imputation to solve this problem since we do not want to add additional information to the original data. It uses the mean value of each attribute to replace the missing. Therefore the input data can be viewed as fully observed, with some wrong information. The results in Figure 4.4 show that our model performs better than other baseline methods. Since mean imputation cannot fully represent the samples, K-Nearest Neighbors and random forest, which make predictions by some specific samples, will have worse performances. While multinomial logistic regression and ensemble of learners using AdaBoostM2, which use the distribution of attributes and relationships among attributes, perform better than the case with 20% missing in the input.

To conclude the prediction results on one single attribute: vehicle ownership, our model achieves similar results as traditional supervised learning methods and performs better when the missing

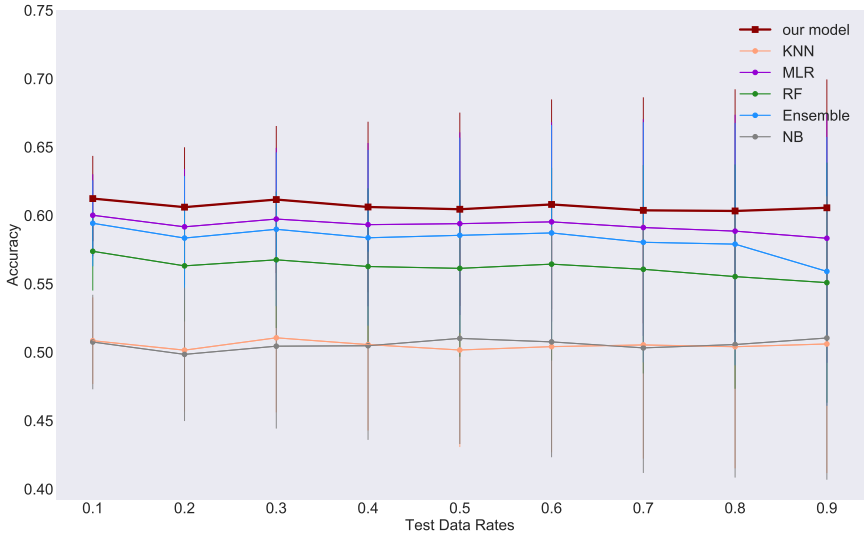


FIGURE 4.4 : Prediction accuracy on `HHVEHCNT` with 50% missing in input. Data is imputed by mean imputation method before prediction.

value rate of input data becomes large. Since our model itself can do missing value imputation, as introduced in chapter 3.5, we can use our model on vehicle ownership modeling when the input data contain lots of missing values.

4.4. Prediction on the Joint Attributes that Contain Vehicle Ownership

Besides vehicle ownership, we may care about other attributes in the meanwhile. Given a small fraction of samples, we would like to know vehicle ownership in one household and other information about households or even for individuals. However, if we want to do so, traditional supervised learning methods such as linear regression and random forest cannot handle two or more output vectors under this circumstance. Generally, we have two ways to solve this problem. One is to separately model each attribute, using the attributes that do not have missing values as input data, and one attribute with missing values as output once a time. Another way is to embed those attributes with missing values altogether into one combined attribute. In this case, the number of categories

will become the multiplication of the number of categories of original attributes. Comparing to the first solution, the embedding one will cost much more computational resources, from linear to polynomial. Also, it does not consider the correlation between two or more attributes for the first solution, which definitely will affect the imputation accuracy. When there exist missing values in the input data, the problem becomes much more complicated.

We consider the simplest joint attributes that contain vehicle ownership and one other attribute, that is, we predict an attribute pair that consists of two attributes simultaneously. We first analyze the correlation between the two attributes since it will affect the imputation results. Figure 4.5 provides the correlation between `HHVEHCNT` and other 23 attributes including itself. Of course, the correlation between `HHVEHCNT` and itself is the highest, which is one. We could also clarify the attribute pairs with positive correlation, negative correlation, or no correlation. The most positively correlated pair is `HHVEHCNT` and `DRVRCNT`, which is intuitively accepted by common knowledge. The most negatively correlated pair is `HHVEHCNT` and `HBHTNRNT`, while there is nearly no correlation between `HHVEHCNT` and `HH_HISP`. These three pairs are used for later experiments.

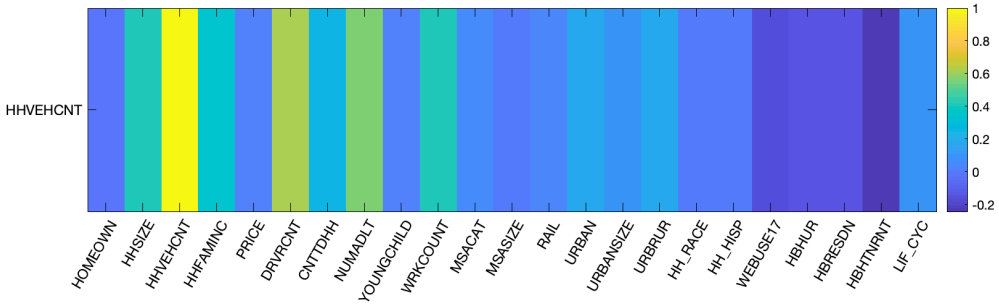


FIGURE 4.5 : Correlation between `HHVEHCNT` and other attributes.

Figure 4.6 provides the prediction accuracy for our model comparing to the other four methods, multinomial logistic regression, random forest, the ensemble of learners with AdaBoostM2, and Naive Bayes classifier, on three different correlated attribute pairs. The bar with the color blue and

green are the expected prediction accuracy for our model and comparison methods. The expected values are calculated by averaging the results of separately training two models on two attributes, which is the first solution we have mentioned previously. The bars with the color orange and red present the results of training one model on the combined joint attribute, and it is the second solution we have mentioned before. When we consider the correlation among attributes, the second solution cannot reach the performance of the first solution, which is shown in the results. We could conclude that the accuracy for the combined attribute for all groups of experiments is below the average of two separate accuracies, especially for the random forest method and negatively correlated pairs. Nevertheless, our model keeps the level of accuracy regardless of which kind of correlation among selected attributes. Moreover, we find that combination will affect less on positively correlated attribute pairs and no correlation attribute pairs for different correlated attribute pairs. In contrast, negatively correlated attribute pairs are highly affected when they are combined. The four supervised learning methods intuitively provide this result. However, for our model, the result is not much affected, which means we can achieve satisfied prediction performance for vehicle ownership and any other attributes simultaneously, without extra efforts to analyze the correlation among attributes.

The data input is similar to predicting single vehicle ownership, which means there are missing values that appeared in the data. We have done experiments on input with 20% of missing values, and the results are shown in Figure 4.7. In this case, every group of the experiment, regardless of methods and attribute pairs, suffers from a decrease in prediction accuracy results. Our model is not much affected by the missing values that appeared in the input, while other methods do not work well, especially for multinomial logistic regression. Another thing that needs to be mentioned is the level of decrease among different correlated pairs. For our model and multinomial logistic

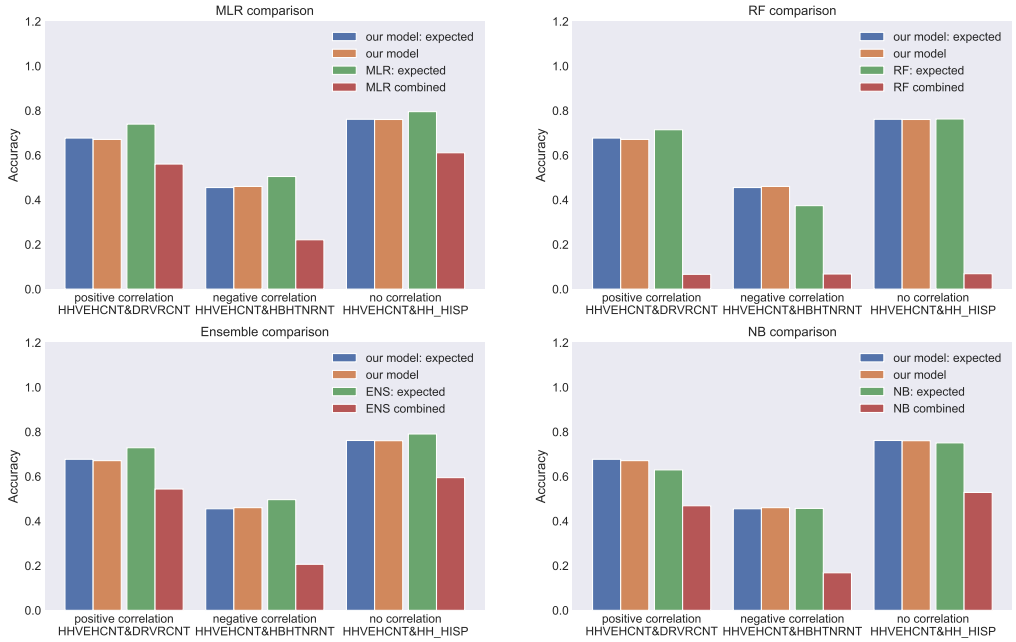


FIGURE 4.6 : Expected and actual prediction results for four comparison methods on different attribute pairs with fully observed input.

regression, negatively correlated attribute pairs are obviously less affected. The reason may be that the loss of information contributes much to correlation when predicting and analyzing two or more attributes.

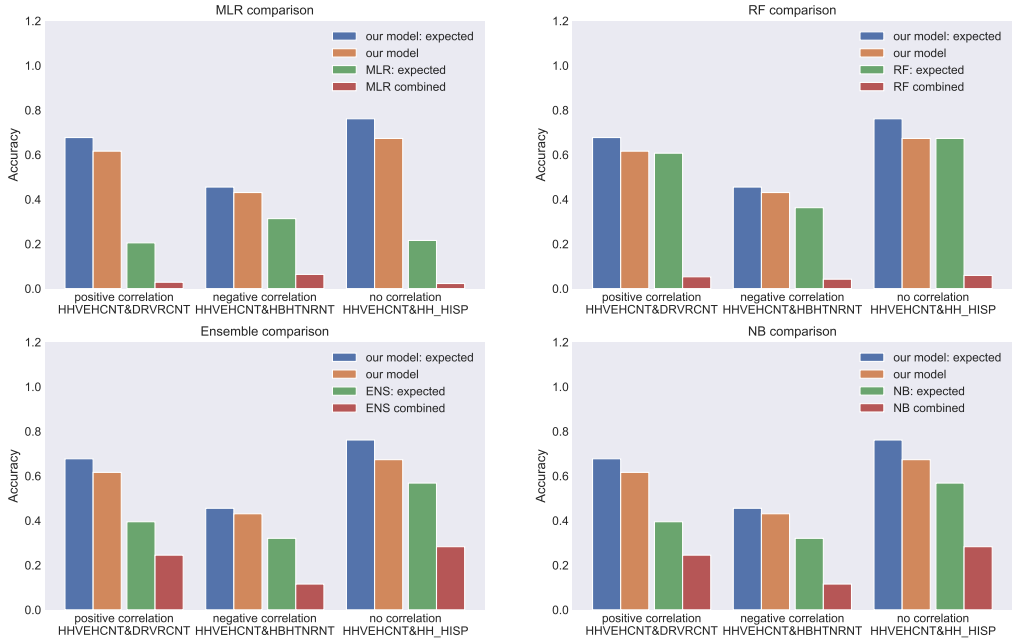


FIGURE 4.7 : Expected and actual prediction results for four comparison methods on different attribute pairs with 20% missing in input.

To conclude, our model can predict two or more attributes for the joint attributes prediction and achieve superior results. To do supervised learning work as an unsupervised learning model, it considers the correlation among attributes, and the prediction results are not affected by different correlations. Moreover, the computation time is acceptable, while other baseline methods on the combined attribute consume much more resources but could not provide satisfying results.

Chapter 5

Conclusion and Discussion

In this thesis, we develop a probabilistic model for multi-level population synthesis, including both person-level and household-level. We focus on using latent classes to find patterns among individuals, households, and attributes. The patterns are presented by multivariate multinomial distributions, which work well for categorical attribute values in census and survey data. The result distributions can be used for agent-based modeling in urban transportation planning and spatiotemporal analysis.

Our model can deal with random missing values in the dataset with few complete individual samples and large missing rates. Imputing missing values can help in mining information about the whole population. We apply the Public Use Microdata Sample (PUMS) dataset and choose attributes from both person-level and household-level. Besides model selection, our experiments mainly focus on missing value imputation since census data are confidential and the data provided are resampled with some missing. From the experiment results of random missing value imputation, our model shows satisfied accuracy on imputation. Regardless of the choice of attributes and the preprocessing procedure, and the different missing rates, our model outperforms the K-Nearest Neighbors model, especially on a larger missing value rate.

Vehicle ownership modeling is a significant problem and reflects urban and land-use planning, energy consumption, industrial decision-making, and travel behavior prediction. When applying census and survey data on modeling, this is a supervised learning problem in the presence of missing values. Traditional supervised learning methods need extra effort to deal with missing values

in the input. In contrast, our model, which is an unsupervised learning model, can do supervised learning work and deal with these missing values. The experimental results show the advantage of our model when there are missing values in the census and survey data. Moreover, when we care about vehicle ownership and other attributes altogether, our model can predict directly without taking more effort. Traditional supervised learning methods have to separately train on each attribute or train on a combined attribute, which will ignore the correlation information or cost many computational resources, respectively.

There are still several directions that could be researched in the future. The first direction is to continue analyzing the multinomial distribution for attributes. We may apply a temporal model on annual data to predict possible population characteristics in the future, or a graphics model on regional data to find out regional correlation and borrow information when analyzing. Secondly, we can process more on the input data to improve the accuracy of missing value imputation. There could be commonly known metrics to determine if individuals contribute to distributions. The third direction is to extend this model to different types of attribute values. Attributes related to age and income could be viewed as continuous instead of categorical values. Gaussian and Beta distribution may help in this case.

References

American Community Survey, *Public Use Microdata Sample*, 1999-2018.

Rich, J., G. Flötteröd, S. Garrido, and F. Pereira, Review of population synthesis methodologies, 2019.

Borysov, S. S., J. Rich, and F. C. Pereira, How to generate micro-agents? A deep generative modeling approach to population synthesis. *Transportation Research Part C: Emerging Technologies*, Vol. 106, 2019, pp. 73 – 97.

Sun, L., A. Erath, and M. Cai, A hierarchical mixture modeling framework for population synthesis. *Transportation Research Part B: Methodological*, Vol. 114, 2018, pp. 199 – 212.

Müller, K. and K. W. Axhausen, Hierarchical IPF: Generating a synthetic population for Switzerland. *Arbeitsberichte Verkehrs-und Raumplanung*, Vol. 718, 2011.

de leeuw, E., Reducing Missing Data in Surveys: An Overview of Methods. *Quality and Quantity*, Vol. 35, 2001, pp. 147–160.

Frick, M., Generating synthetic populations using IPF and monte carlo techniques: Some new results. *Arbeitsberichte Verkehrs-und Raumplanung*, Vol. 225, 2004.

Williamson, P., M. Birkin, and P. H. Rees, The estimation of population microdata by using data from small area statistics and samples of anonymised records. *Environment and Planning A*, Vol. 30, No. 5, 1998, pp. 785–816.

Adjemian, M. and J. Williams, Using census aggregates to proxy for household characteristics: an application to vehicle ownership. *Transportation*, Vol. 36, No. 2, 2009, pp. 223–241.

Bar-Gera, H., K. Konduri, B. Sana, X. Ye, and R. M. Pendyala, Estimating survey weights with multiple constraints using entropy optimization methods. In *88th Annual Meeting of the Transportation Research Board, Washington, DC*, 2009.

Deming, W. E. and F. F. Stephan, On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *Ann. Math. Statist.*, Vol. 11, No. 4, 1940, pp. 427–444.

Sun, L. and A. Erath, A Bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies*, Vol. 61, 2015, pp. 49 – 62.

Saadi, I., A. Mustafa, J. Teller, B. Farooq, and M. Cools, Hidden Markov Model-based population synthesis. *Transportation Research Part B: Methodological*, Vol. 90, 2016, pp. 1–21.

Bhattacharya, A. and D. B. Dunson, Simplex factor models for multivariate unordered categorical data. *Journal of the American Statistical Association*, Vol. 107, No. 497, 2012, pp. 362–377.

Dunson, D. B. and C. Xing, Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, Vol. 104, No. 487, 2009, pp. 1042–1051.

Barash, Y., G. Elidan, N. Friedman, and T. Kaplan, Modeling dependencies in protein-DNA binding sites. In *Proceedings of the seventh annual international conference on Research in computational molecular biology*, 2003, pp. 28–37.

Kingma, D. P. and M. Welling, Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets. In *Advances in neural information processing systems*, 2014, pp. 2672–2680.

Hu, J., J. P. Reiter, and Q. Wang, *Dirichlet Process Mixture Models for Modeling and Generating Synthetic Versions of Nested Categorical Data*, 2014.

Barnett, A. G., P. McElwee, A. Nathan, N. W. Burton, and G. Turrell, Identifying patterns of item missing survey data using latent groups: an observational study. *BMJ Open*, Vol. 7, 2017.

Karanja, E., J. Zaveri, and A. Ahmed, How do MIS researchers handle missing data in survey-based research: A content analysis approach. *International Journal of Information Management*, Vol. 33, 2013, p. 734–751.

Gilley, O. W. and R. P. Leone, A two-stage imputation procedure for item nonresponse in surveys. *Journal of Business Research*, Vol. 22, No. 4, 1991, pp. 281 – 291.

Marsh, H. W., Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 5, No. 1, 1998, pp. 22–36.

Dempster, A. P., N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 39, No. 1, 1977, pp. 1–22.

Enders, C. K. and D. L. Bandalos, The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 8, No. 3, 2001, pp. 430–457.

Liao, S. G., Y. Lin, D. D. Kang, D. Chandra, J. Bon, N. Kaminski, F. C. Sciurba, and G. C. Tseng, Missing value imputation in high-dimensional phenomic data: imputable or not, and how? *BMC bioinformatics*, Vol. 15, No. 1, 2014, p. 346.

Young, W., G. Weckman, and W. Holland, A survey of methodologies for the treatment of missing values within datasets: Limitations and benefits. *Theoretical Issues in Ergonomics Science*, Vol. 12, No. 1, 2011, pp. 15–43.

Chen, J. and J. Shao, Nearest neighbor imputation for survey data. *Journal of official statistics*, Vol. 16, No. 2, 2000, p. 113.

Batista, G. E. and M. C. Monard, An analysis of four missing data treatment methods for supervised learning. *Applied artificial intelligence*, Vol. 17, No. 5-6, 2003, pp. 519–533.

Gupta, A. and M. S. Lam, Estimating missing values using neural networks. *Journal of the Operational Research Society*, Vol. 47, No. 2, 1996, pp. 229–238.

Vermunt, J. K., J. R. Van Ginkel, L. A. Van der Ark, and K. Sijtsma, 9. Multiple Imputation of Incomplete Categorical Data Using Latent Class Analysis. *Sociological Methodology*, Vol. 38, No. 1, 2008, pp. 369–397.

Schafer, J. L., *Analysis of incomplete multivariate data*. CRC press, 1997.

Vermunt, J. K., Multilevel latent class models. *Sociological methodology*, Vol. 33, No. 1, 2003, pp. 213–239.

Si, Y. and J. P. Reiter, Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, Vol. 38, No. 5, 2013, pp. 499–521.

Koren, Y., R. Bell, and C. Volinsky, Matrix factorization techniques for recommender systems. *Computer*, Vol. 42, No. 8, 2009, pp. 30–37.

Mitra, K., S. Sheorey, and R. Chellappa, Large-scale matrix factorization with missing data under additional constraints. In *Advances in Neural Information Processing Systems*, 2010, pp. 1651–1659.

Agathokleous, M. and N. Tsapatsoulis, Voting advice applications: missing value estimation using matrix factorization and collaborative filtering. In *Ifip international conference on artificial intelligence applications and innovations*, Springer, 2013, pp. 20–29.

Hernández-Lobato, J. M., N. Houlsby, and Z. Ghahramani, Probabilistic matrix factorization with non-random missing data. In *International Conference on Machine Learning*, 2014, pp. 1512–1520.

Yang, Y. and D. B. Dunson, Bayesian conditional tensor factorizations for high-dimensional classification. *Journal of the American Statistical Association*, Vol. 111, No. 514, 2016, pp. 656–669.

Schwarz, G., Estimating the Dimension of a Model. *Ann. Statist.*, Vol. 6, No. 2, 1978, pp. 461–464.

Giraud, C., *Introduction to high-dimensional statistics*, Vol. 138. CRC Press, 2014.

Sun, L., X. Chen, Z. He, and L. Miranda-Moreno, Pattern discovery and anomaly detection of individual travel behavior using license plate recognition data, 2019.

Goodfellow, I., Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

Kuhn, M., K. Johnson, et al., *Applied predictive modeling*, Vol. 26. Springer, 2013.

Potoglou, D. and P. S. Kanaroglou, Modelling car ownership in urban areas: a case study of Hamilton, Canada. *Journal of Transport Geography*, Vol. 16, No. 1, 2008, pp. 42 – 54.

Bhat, C. R. and V. Pulugurta, A comparison of two alternative behavioral choice mechanisms for household auto ownership decisions. *Transportation Research Part B: Methodological*, Vol. 32, No. 1, 1998, pp. 61 – 75.

Basu, R. and J. Ferreira, Understanding household vehicle ownership in Singapore through a comparison of econometric and machine learning models. *Transportation Research Procedia*, Vol. 48, 2020, pp. 1674–1693.

Sinha, K. C., Sustainability and urban public transportation. *Journal of Transportation Engineering*, Vol. 129, No. 4, 2003, pp. 331–341.

Ryan, J. M. and G. Han, Vehicle-ownership model using family structure and accessibility application to Honolulu, Hawaii. *Transportation Research Record*, Vol. 1676, No. 1, 1999, pp. 1–10.

Cirillo, C. and Y. Liu, Vehicle Ownership Modeling Framework for the State of Maryland: Analysis and Trends from 2001 and 2009 NHTS Data. *Journal of urban planning and development*, Vol. 139, No. 1, 2013, pp. 1–11.

Liu, Y., J.-M. Tremblay, and C. Cirillo, An integrated model for discrete and continuous decisions with application to vehicle ownership, type and usage choices. *Transportation Research Part A: Policy and Practice*, Vol. 69, 2014, pp. 315–328.

Federal Highway Administration (FHWA), *National Household Travel Survey*, 2017.

Nelder, J. A. and R. W. Wedderburn, Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, Vol. 135, No. 3, 1972, pp. 370–384.

Ho, T. K., Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, IEEE, 1995, Vol. 1, pp. 278–282.

Eibl, G. and K.-P. Pfeiffer, Multiclass boosting for weak classifiers. *Journal of Machine Learning Research*, Vol. 6, No. Feb, 2005, pp. 189–210.

Murty, M. N. and V. S. Devi, *Pattern recognition: An algorithmic approach*. Springer Science & Business Media, 2011.

MathWorks, *Matlab Documentation: mnrfit*, 2020.