GENERATING DISAGGREGATE POPULATION CHARACTERISTICS FOR INPUT TO TRAVEL-DEMAND MODELS

By

LU MA

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2011

To my parents

ACKNOWLEDGMENTS

First of all, I would like to thank my supervisor Dr. Siva Srinivasan who always inspires me for my study. From him, I have received not only knowledge but also the way to thinking and serious attitude on research, which are beneficial for my future career and life.

I also want to thank Dr. Trevor Park, Dr. Lily Elefteriadou, Dr. Yafeng Yin, Dr. Scott Washburn and Dr. Ruth Steiner for their valuable suggestions and comments on this study.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

GENERATING DISAGGREGATE POPULATION CHARACTERISTICS FOR INPUT TO TRAVEL-DEMAND MODELS

By

Lu Ma

December 2011

Chair: Sivaramakrishnan Srinivasan
Major: Civil Engineering

The past several years have seen tremendous developments in disaggregate travel-demand models. The application of such models for predictions and policy evaluations requires as inputs detailed information on the socio-economic-mobility characteristics of the population. Synthesis methods are used to first generate the population for a base year (current year/census year) and this, in turn, is used as an input to generate the target year (forecast year) population.

The state-of-the-practice approach to population synthesis involves the use of the Iterative Proportional Fitting (IPF) method. While there have been several applications of this approach, several issues still remain. First, the number of controls used in the synthesis of the base-year population has been limited. In particular, most practical applications control only for household-level attributes (e.g., household size and dwelling-unit type) and not for person-level attributes such as age and gender. Thus, the synthesized base-year population may not truly match the observed person-level distributions. This would affect the accuracy of the target-year population as the synthesized base-year population is used as an input to generate the target-year population. Second, documentation of the validation of the synthesis procedure,

especially in the context of a “target” year population, is limited. The broad focus of this dissertation is to contribute towards synthetic population generation by addressing these issues.

To generate a synthetic population as inputs on disaggregate travel-demand models, this dissertation proposes a new framework for synthetic population generation through a fitness based synthesis (FBS) method in which multi-level (household level and personal level etc.) attributes can be controlled simultaneously. During simulation, several socioeconomic variables (such that household size, income, gender, age and etc) under census tract level are chosen as control information and the 5% sample from corresponding PUMA (Public Used Microdata Area) forms the seed data. Empirical results indicate that controlled attributes of synthetic population can match the true population almost perfectly. Furthermore, this dissertation also proposes a validation idea in which a set of household is transferred into the distribution of household type and several criteria are also introduced for measuring the difference between true population and synthetic populations. As expected, the synthetic population under household and person level controls will have much similarity with the true population comparing to the synthetic population using only household level controls. On the other hand, this dissertation also compares the FBS method with other population synthesizers through proposed validation criteria. Then these synthetic populations are compared based on the difference with the true population.

Even though the generation of synthetic target-year population is similar to the one of base-year methodologically, there are two factors which affect the quality of synthetic populations for target year exclusively. More specifically, the seed data, namely the

synthetic population of the base year could be one of the factors. Another important factor is the controlled attributes used in target-year population synthesis. Unlike base-year population synthesis, the controls used in target year population synthesis are come from population projection models in which these controls are inaccurate. Toward these aspects, the target-year synthetic populations are applied through the proposed framework and IPF separately. By a back-casting analysis, target-year (1990) populations for twelve census tracts in Florida are generated. For each census tract, three different populations of base year are synthesized according to different controls, and for each base-year population two different methods are applied for target-year populations. The results from target year population analysis indicate that using more accurate base-year populations as seed data is more likely to end up with more accurate target-year populations. Then similar analysis is repeated for the inaccurate controls. By comparing the population with the true controls, it is indicated that the additional errors of synthetic populations introduced by the inaccurate controls are linearly related to the amount of errors of controls.

With the proposed method, this dissertation also assesses travel demand models applied to synthetic populations using NHTS (National Household Travel Survey) dataset. First, two trip generation models are estimated using different specification of explanatory variables. One of them is in a more disaggregate fashion since a lot of household characteristics are involved while another model only adopts several characteristics and hence it can be considered as an aggregated model. Second, two different synthetic populations are generated by different controls. Then, the two models are applied to the two populations separately.

In sum, this dissertation develops a fitness based synthesis methodology that can be applied to synthesize populations by controlling several attributes at both household and person levels simultaneously. In order to assess the similarity between the true population and synthetic populations, this dissertation proposed a validation idea as well as several validation measurements.

The procedure was applied to synthesize both base-year and target-year populations for twelve census tracts in Florida. The analysis indicates that the proposed approach results in synthetic populations that match rather closely with the true distributions. Further, the results also highlight the improvements that can be achieved by controlling for both household and person level attributes.

CHAPTER 1
INTRODUCTION

The past several years have seen tremendous developments in disaggregate travel-demand models and activity-based transportation planning models (e.g. Bhat and Koppelman, 1993). This interest is motivated by several factors such as (1) reduction of aggregation errors, (2) ensuring sensitivity to demographic shifts like the ageing of the population, (3) capturing differential response of travelers to policy actions, and (4) addressing special travel-needs of certain population groups.

The application of such models (e.g. Bhat et al., 2004) for predictions and policy evaluations requires as inputs detailed information on the socio-economic-mobility characteristics of the population (some examples of these characteristics include household-level attributes such as size and composition, income, dwelling-unit type, auto ownership, and person-level attributes such as age, gender, and employment status). However systematic procedures to forecast these attributes required by such disaggregate travel-demand models are under development only recently and the experiences with the empirical application of such models are limited. Therefore, developments in the area of population synthesis methods are important for furthering disaggregate-modeling efforts and for their adoption as state-of-practice.

The methodology for producing such a disaggregate population is generally referred to as Synthetic Population Generation (SPG). The classical SPG framework (Beckman et al., 1996) comprises two major inputs: (1) the marginal distributions (control tables) of certain attributes (e.g. household size and income) for certain target area (census tract), and (2) a sample of households (“seed data”) with detailed attribute values often from a larger area (such as the Public Use Microdata Area or PUMA,

discussed in more detail later on). The Iterative Proportional Fitting (IPF) technique (see Deming and Stephan, 1940; Ireland and Kullback, 1968) is employed to generate the multivariate distribution of all attributes of interest. Using the cell values of this joint distribution as selection probabilities, households are drawn from the seed data thereby generating the synthetic population of interest.

An important short-coming of the classical SPG procedure is that the population is generated to match only one-level controls (the household-level in almost all practical applications). Information about the person-level attributes (such as age and gender) is ignored, even though such marginal controls are available. Therefore, there is potential for improving the accuracy of the synthesized population by employing methods that can handle multi-level controls. Within the last couple of years, there is growing interest in population synthesis using multi-level controls and a primary objective of this dissertation is to contribute towards this end. Specifically, a fitness based approach is developed to synthesize populations that fit controls at multiple-level simultaneously.

A second important issue is that of validation. The populations have to be synthesized only because the true values are unknown. However, the lack of the true population also makes it difficult to validate the methods developed to synthesize the population. In this dissertation, rigorous validation procedures are described and applied in the contexts of base-year population synthesis. The proposed validation metrics are also used to compare our fitness based methodology to other approaches that have been recently developed.

Furthermore, unlike the case of base-year population synthesis, the documentation of results on target-year population synthesis is limited (Bowman2004

and Bowman and Rousseau 2008). Although, conceptually, the application of the approach for target-year synthesis is similar to its application for base year synthesis, there are three important issues of concern. First, the target-year synthesis uses the base-year synthesized population as seed data. Thus, the methodology and controls used in the base-year synthesis impact the accuracy of the base-year population, and in turn, the target-year population. Second, one can expect significantly fewer control tables available for the target year synthesis as opposed to the base year synthesis. In this situation, there might be benefits to using approaches that control for both person- and household- level information as opposed to methods that control for only household-level information so as to take advantage of all the minimal data available. Third, the target year control tables are projections in contrast to base year control tables which are derived from the census counts. It has been well documented (Smith and Shahidullah 1995) that there are significant errors in these projected aggregate distributions of population characteristics. Therefore, examining the effects of errors in control tables is of interest.

Finally, it is important to remember that these procedures ultimately provide data that are in turn fed into disaggregate travel-demand models. Disaggregate models capture travel behavior of the fundamental decision-making units and include several explanatory variables (including socio-economic and mobility characteristics).

Consequently, one may expect such models to provide more accurate predictions of the travel characteristics than aggregate models which include fewer explanatory variables. However, this depends on the accuracy of the socio-economic-mobility characteristics of the synthesized population. Specifically, if the synthesized population

is an inaccurate representation of the true population, gains because of a disaggregate model could be offset by the errors in the synthesized population. In light of the above discussion, predictions from aggregate trip-generation models will be compared with those from disaggregate trip-generation models (with synthetic populations as inputs) to assess the true value of synthetic populations and disaggregate travel-demand models.

In summary, the goals of this dissertation are to: (1) develop a procedure for population synthesis that can allow for multi-level control and demonstrate its benefits over conventional, IPF-based methods, (2) develop a systematic framework for validating synthetic population generators and apply it in the context of the new procedure developed and two other major methods, (3) Synthesize target year populations (using back-casting methods) and assess the value of the new synthesis procedure and quantify the impacts of erroneous controls, and (4) Apply the synthetic populations to trip-generation models and assess the overall accuracy of the trip-rate predictions.

The rest of the dissertation will be arranged as follows. Chapter 2 presents the conceptual framework of population synthesis. Chapter 3 gives a thorough review of population synthesis as well as validation methods. Chapter 4 introduces the framework of fitness based population synthesis. Chapter 5 applies the proposed fitness based approach on the base-year population synthesis and several validation methods are also introduced. Chapter 6 synthesizes target-year populations and an analysis of accuracy of target-year populations is conducted. Then, Chapter 7 applies the synthetic population to trip generation models. And, Chapter 8 closes this proposal with summary and conclusions.

CHAPTER 2
CONCEPTUAL FRAMEWORK

This chapter presents an overview of the methods currently available for population synthesis. The first section in this chapter presents a conceptual overview of the overall synthesis procedure. Although the intent is generally to generate a population for a target year, the synthesis procedure begins with generating a population for a base year. The second section discusses methods for synthesizing base-year population whereas the third section describes the methods for target-year population synthesis. At the end of this chapter, the synthetic population is discussed in terms of inputs for travel demand models.

Conceptual Overview of the Generation of Synthetic Population

A conceptual overview of the SPG procedure is presented in Figure 2-1. In the beginning of a SPG procedure, the base year population is generated first. The base year is defined as the survey year and usually the most recent census year in the past (currently, this would be year 2000). After generating the population for a base year, it will serve as an input in the SPG of a target year. A target year is defined as any year beyond the base year and may or may not be a year for which the decennial census has been planned. That is, if the base year is 2000, years 2003, 2010, and 2025 would all be qualified as target years.

The synthesis of the base year population is performed by data fusion techniques. Broadly, aggregate control-tables (often at the census-tract level) are fused with disaggregate data on population characteristics (seed data) available for a sample of households in the area (often at the PUMA level) to which the census tract belongs.

The result is a synthetic population for base year comprising households drawn (with replacement) from the seed data of corresponding PUMA in the way that the aggregate characteristics of the synthetic population is matched with these aggregated control-tables.

Given the base-year synthetic population, there are two approaches for generating the target-year population. The first approach uses the data fusion technique which is similar to the one used in base-year population synthesis. The synthetic base-year population will be served as seed data in target-year population synthesis. Another approach is called evolution approach and it involves “growing” each base-year household over time to determine its characteristics at the target year. This involves modeling complex phenomenon such as household formation, dissolution and migration. Once target-year populations are generated, they can be used as inputs to travel demand models.

Synthesizing the Base-Year Population

The state-of-the-practice approach to base-year population synthesis involves fusing aggregate control tables with disaggregate seed data. Control tables are one-way or multi-way marginal distributions. Each of these tables corresponds to the joint distribution of a subset of the required population attributes. Typically, these distribution tables are available from the census SF1 and SF3 files and at the spatial resolution of census block groups or census tracts. The population is synthesized at the spatial resolution of the control tables (this is referred to as the “synthesis area” in the rest of this document). The seed dataset comprises a sample of population records with each household/person characterized by all the attributes of interest. The location of these households is typically known only at a more aggregate spatial scale (in contrast to the

finer spatial resolution of the control tables). Typically, such household-level information is obtained from the US census Public Use Microdata Samples (PUMS) and the location is defined in terms of the Public Use Microdata Areas (PUMAs).

The state-of-the-practice data fusion procedure involves two major steps. First, a joint multi-way distribution of all attributes of interest is generated using the Iterative Proportional Fitting (IPF) procedure (conceptually, the procedure is analogous to the Fratar balancing technique; detailed algorithm of the IPF procedure is available from Beckman et al., 1996). The IPF procedure ensures that, when the multi-way distribution is appropriately aggregated, the results match the marginal distributions provided by the control tables (the extent of “matching” depends on the tolerance used). The result of this iterative procedure is a multi-way distribution table that provides the number of households of each type in the synthesis area. In the second step, individual household records are drawn from the seed dataset using Monte-Carlo simulation so as to satisfy the joint multi-way distributions.

This methodology has been applied to support travel-demand modeling in several areas such as Portland Metro, San Francisco, New York, Columbus, Atlanta, Sacramento, Bay Area, and Denver. Bradley and Bowman (2006) and Bowman (2004) provide a general overview of these applications. The Sacramento application is available in Bowman and Bradley (2006) and the Atlanta application and validation results are presented in Bowman and Rousseau (2008).

All the applications discussed thus far control for only household-level attributes. There are several researches (e.g. Guo and Bhat, 2007; Arentze et al., 2007; Ye et al., 2009; Auld and Mohammadian, 2010) provide extensions to incorporate both

household- and person-level controls in the IPF-based population-synthesis procedure. Detailed descriptions of these procedures are presented in Chapter 3.

Synthesizing the Target-Year Population

The data-fusion approach for the synthesis of the target-year population is conceptually similar to the one used for generating the base-year population. Once again, aggregate control tables and disaggregate seed data are the inputs.

The control-tables represent the aggregate socio-economic-mobility characteristics of the synthesis area in the target year. There are two key differences between the control tables used in the base-year synthesis and those used in the target-year synthesis. First, for the target year, the number of controls available is limited (and often multi-dimensional controls may not be available). In contrast, the base year would have several (and multi-dimensional) controls from the Census data. Second, the control tables for the target year may not even be available at the synthesis-area level and may have to be derived from more aggregate spatial units (such as the county).

The structure of the seed data for the target-year population synthesis is the same as the one for the base year. This is because the synthesized base-year population is taken as the seed data. The reader will note that the seed data for the base year are at the PUMA level, but from the same year which is in contrast to the seed data for the target year which is from the same census tract but is from the base year.

The methodology used for the target-year population synthesis is predominantly the same as the one used in the base year. However, some of the attributes of interest may not be directly synthesized due to lack of control data. For these cross-section models can be used. A classic example of an attribute which is forecasted in such a manner is automobile ownership [see for example, the Oregon2 Model (Hunt et al.,

2004) or the SACOG model (Bowman and Bradley, 2007)]. Typically, US census does not provide projections of aggregate auto-ownership levels for any future year for use in a data-fusion approach. However, it is possible to develop cross sectional models of auto ownership (as a function of household characteristics, land use patterns, transportation system characteristics, etc.) using data from local household travel surveys or the PUMS. Thus, once the appropriate socio-economic characteristics for a forecast year have been determined using data-fusion techniques, the cross-sectional model can be applied to each household to generate the auto-ownership levels.

For evolution method, each household in the base-year synthetic population database is evolved or aged though time to determine its characteristics for any future year. This involves the development of a system of models that describe the common demographic/economic transitions that take place over the life-cycle of a household. These transitions include processes such as ageing, births, deaths, formation (marriage) and dissolution (divorce) of households, employment and education choices, children moving out of the household, automobile ownership decisions, and emigration from or immigration to the study region. Some of the currently available model systems that adopt such an approach include MIDAS (Goulias and Kitamura, 1996), MASTER (Mackett, 1990), CEMSELTS (Eluru et al., 2008), DEMOS (Sundararajan and Goulias, 2003), and the HA module of the Oregon2 model system (Hunt et al., 2003). Such methods are appealing as they try to simulate the real processes households go though and model behavioral decisions made at different stages of the life cycle. However, as identified by Eluru et al., (2008), limited theoretical knowledge on the complex socio-economic evolution processes and the minimal availability of relevant data at the

household level limit our ability to specify and estimate good models of household evolution.

Application on Travel Demand Models

As mentioned before, synthetic populations are required as input for disaggregate travel-demand models and activity-based transportation planning models (e.g. Bhat and Koppelman, 1993). And this is the ultimate goal for generate synthetic populations. Undoubtedly, synthetic populations have more disaggregate information than aggregated data even through such information is inaccurate in some degree. So, the accuracy of travel demand forecasting depends on the accuracy of the socio-economic-mobility characteristics of the synthesized population. Specifically, if the synthesized population is an inaccurate representation of the true population, gains because of a disaggregate model could be offset by the errors in the synthesized population.

The details of the proposed fitness based synthesis framework for the “data fusion” are presented in Chapter 4 of this dissertation. The procedure is applied and validated in the context of the base-year synthesis and these are discussed in Chapter 5. The analysis of the population generation procedure in the context of the target-year is described in Chapter 6. In Chapter 7, synthetic populations are applied and assessed under trip generation models.

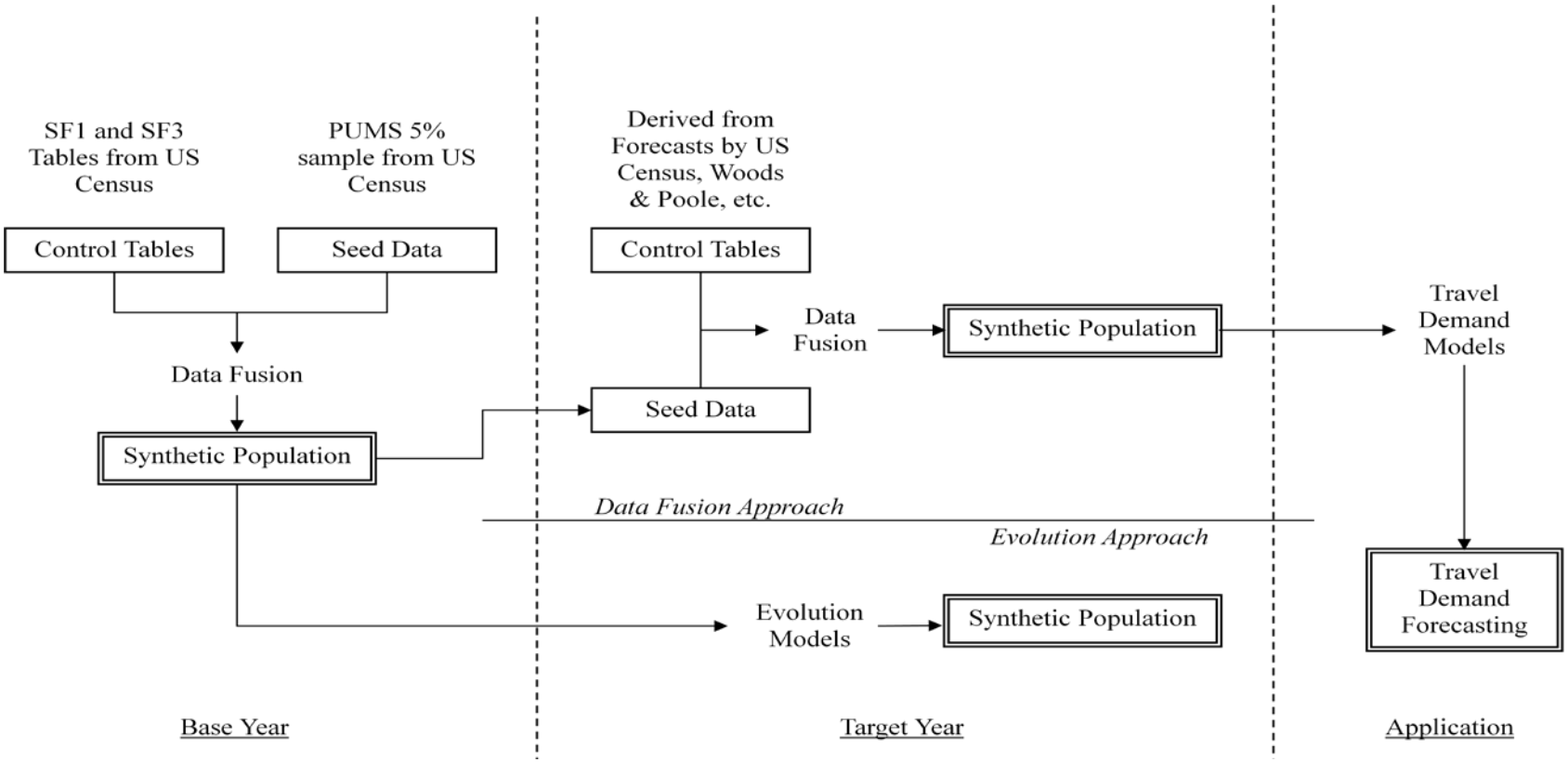

Figure 2-1. Conceptual framework of the population synthesis procedure

CHAPTER 3
LITERATURE REVIEW

This chapter presents a literature review of the procedure of population synthesis based on the data fusion methods. The first section of this chapter summarizes the past researches and applications of population synthesis and in this section, the methods for single level controls are reviewed. Then the ideas and current methods for controlling multilevel characteristics are described in more details. At the end of this chapter, validation methods for population synthesis are reviewed.

Generation of Synthetic Populations

Broadly, the procedure for synthesizing populations involves selecting a set of households from the seed data in such a way that the controls are satisfied.

A classic framework adopts IPF as the kernel part of data-fusion technique. Include IPF, two major steps are involved in the procedure of population synthesis. First, the sample frequencies of each cell of the household contingency table are projected to match the marginal counts. And IPF (Deming and Stephan, 1940 and Wong, D.W.S., 1992) is used to project the seed data by mating the control data. Second, the number of households to be generated of each type (according to a set of combination of household characteristics) will be determined by the projected cell counts. In fact, almost all current approaches involve IPF in their first step and this dissertation will briefly introduce the IPF method through a two way contingency table.

Iterative Proportional Fitting (IPF)

In a two way contingency table, let the cell counts in i-th row and j-th column for seed data be n_{ij} and the marginal control for i-th row and j-th column be m_{i+} and m_{+j}

respectively. Then the estimated cell counts $\hat{\mu}_{ij}$ can be calculated by an iteratively fitting procedure.

During each iteration, the estimated cell counts will be updated by a row-fitting and a column-fitting. Let I be the number of rows in the contingency table. With the initial value of each cell counts $\mu_{ij}^{(0)} = n_{ij}$, a row-fitting is implemented as,

$$\mu_{ij}^{(2k-1)} = \mu_{ij}^{(2k-2)} \frac{m_{i+}}{\mu_{i+}^{(2k-2)}}, \qquad i = 1, \cdots, I \qquad (3-1)$$

where $\mu_{ij}^{(2k-1)}$ is the cell counts in $2k-1$ iteration and $\mu_{i+}^{(2k-2)}$ is the row summation of cell counts in $2k-2$ iteration. Then, the column-fitting can be implemented in the same manner,

$$\mu_{ij}^{(2k)} = \mu_{ij}^{(2k-1)} \frac{m_{+j}}{\mu_{+j}^{(2k-1)}}, \qquad j = 1, \cdots, J \qquad (3-2)$$

where $\mu_{+j}^{(2k-1)}$ is the column summation of cell counts in $2k-1$ iteration and J is the number of columns in the contingency table.

Implementing such procedure until convergence, the estimated cell counts can match the marginal row distribution and marginal column distribution simultaneously. Moreover, the underline nature of the IPF procedure actually maximizes the entropy or minimizes the discrimination information of the joint distribution of the contingency table (see Ireland and Kullback, 1968. and Ruschendorf, 1995) and also maximizes the likelihood function under a log-linear model (see Agresti, 2002).

Above procedure can be readily extended for multi-way contingency tables. In case of more than two attributes, cell counts will be projected to fit the marginal distribution for each dimension of the contingency table.

As mentioned before, IPF has very nice property e.g. maximize the entropy and keep the odds ratio from seed data and also IPF is the foundation of almost all current researches about population synthesis. However, when the number of controls is large, it is no surprise that some of the cells are zero, especially when the size of the seed data is small, some types of household are missing due to the variation of sampling. Also, some combinations of household characteristics are rarely distributed and even do not exist in reality. For example, the chance for a single household with 6 or more vehicles or a single household with an only 5 years old child is very small and the household size of a family household cannot be one. The later situation is also refer to as fixed zeros (Fienberg, 1970b) while the former is called sampling zeros.

More specifically, the values of μ_{ij} are not all positive and if a huge number of $\mu_{ij} = 0$ in the contingency table, such situation is called zero cell problems (Auld et al. 2008 and Beckman et al. 1996). In the research of Ireland and Kullback (1968), they shows that the iterative procedure is convergent, however the situation in which the values of μ_{ij} are not all positive is avoided (Fienberg, 1970a). Such contingency table with some zero entries is referred to incomplete contingency tables (see Yvonne et al., 1969). In fact the incomplete contingency tables always converge under regularity conditions but it is slower than the related complete contingency tables (Fienberg, 1972).

Other than the situation that $\mu_{ij} = 0$, it is also possible that $\mu_{i+} = 0$ and hence the program will have no feasible solution because the required types of households do not exist in seed data. Furthermore, in such situation the IPF cannot proceed because the denominator is zero and all the related objective function, e.g. discrimination information

and entropy become infinity. Another zero-value situation during IPF is that the marginal total $m_{i+} = 0$. This situation will also encounter numerical issues under the original IPF procedure, but these issues can be dodged by making some minor changes to the procedure. One of the suggested changes is to define zero divided by zero as zero (Fienberg, 1970b). More detailed discussion about zero-cell situations can be found in Fienberg (1970b).

Other than IPF based approaches, another branch for generating synthetic population is the combinational optimization method (Williamson et al. 1998. and Voas and Williamson, 2000). Contrast to IPF, combinational optimization methods does not require an initial distribution but weights are associated to each household sample. These weights are readjusted to yield the best fit for matching controlled tables. It has been debated between the two methods for a long time. For example, Ryan et al. (2009) states that combinational optimization methods are superior to the IPF based methods whereas Pritchard and Miller (2009) point several weakness of combinational optimization method.

Intuitively, the IPF can copy the relationship among controlled characteristics and hence produce a more accurate population if the seed data is highly correlated to the true population. For example, if the area from which the seed data sample come is geographically related to the area of true population, it is reasonable to assume the high correlation between populations of the two areas. In this dissertation, the household sample is from the area (e.g. PUMA) which contains the analysis area (e.g. census tract) and we assume the information contained in seed data correlates to the synthetic populations. Even though the new proposed framework also does not rely on IPF, the

fundamental mechanism will require a high correlation among characteristics between seed data and true population.

To be noted, the projected cell counts may not be integers and an empirical way is to adopt the rounded cell counts. Consequently, some types of households are slightly overestimated and some types of households are slightly underestimated.

Moreover, instead of using the rounded cell counts, a household can be selected from the seed data base on the probability of each cell (Beckman, et al., 1996). And the selection probability may be updated (Guo and Bhat, 2007) after a new household is selected.

Methodology for Single-Level Controls

The tradition population synthesizers only focus on the controls under single level, in which only household level characteristics, such as household size, household income and vehicle ownership are controlled. Since the work of Beckman et al. (1996), several researches have contributed in this area. Auld et al. (2008) develops a routine to automatically aggregate categories with zero cells to adjacent categories; Zhang and Mohammadian (2008) try to synthesize the population of New York Metropolitan Statistical Area by a two stage population synthesis procedure; Simpson and Tranmer (2005) illustrate that using a seed data from larger areas containing the analysis area can gain accuracy in some extent and they also developed a SPSS based routine for IPF implementation and Moeckel et al. (2003) using IPF to generate a disaggregate population on Netanya, Israel and Dortmund, Germany.

Beckman et al. (1996) provide a comprehensive framework for generating synthetic population. In this research, an IPF procedure is performed to predict the joint distribution among several demographic characteristics (e.g. age of householder, family

income). The marginal control data comes from census Summary Files. After IPF, the synthetic population of households is constructed from census PUMS data. Beckman et al. (1996) assign a selection probability to each household in PUMS data. The probability is computed based on the "distance" between the selected households and the households characterized by a cell in the multi-way contingency table. Basically, if a household sample in PUMS is similar to the most desirable household in the contingency table, a higher probability will be assigned to this household. Beckman et al. (1996) also mentioned the zero cell issue of IPF.

Frick and Axhausen (2004) generate the population for Switzerland using the method similar to Beckman et al. (1996). In this research, two major steps are conducted and each of them includes several IPFs. The first step estimates a multi-way contingency table for a low spatial area and in the second step it will combine the marginal distribution of high spatial area or small area in which it contained to generate the population for the small area. The purpose of two steps of IPF is to correct the correlation between high spatial area and low spatial area.

Moeckel et al. (2003) using IPF to generate a disaggregate population on Netanya, Israel and Dortmund, Germany. During the step of the Monte Carlo sampling, the location/address of each household is selected based on the densities of different residential locations. Such consideration enriches the generated disaggregate population information in terms of spatial location which is required by activity-based transportation microsimulation models.

As mentioned before, when the number of attributes is large, the number of cells of the household contingency table will grow exponentially and hence the computation

resource may be occupied substantially. Pritchard and Miller (2009) suggest a list based data structure to accommodate this issue, because for a large multi-dimensional contingency table, a lot of cells of households does not exist or rarely distributed, and hence such sparse matrix can be stored as a list of physically exist households. With such specification of contingency tables, IPF only adjust the existed households during the procedure. However, the zero-cell issue still exists if some types of households are missing.

Several methods have been proposed for the zero-cell issue. Beckman et al. (1996) simply “tweak” the zero entry by an arbitrary small value (e.g. 0.01). And this idea was showed to be no benefit of accuracy by Beckman et al. (1996) and it may also introduce bias during prediction. Another major method aggregates the zero cell categories with adjacent categories (e.g. Auld, et al., 2008). In fact, most of the current SPGs adopt one of the two methods or a combination of the two methods for dealing with the zero-cell problem (Auld, et al., 2008).

Other than the zero-cell issue related to IPF, a downside of this state-of-practice procedure is that it requires all the control tables to belong to the same “universe”. Therefore, it is not possible to apply this procedure directly to synthesize populations by simultaneously controlling for household- and person- level attributes. However, for achieving more accurate populations, it would be desirable to control for a wide range of attributes. In the last few years, there have been several efforts to modify the traditional approach to deal with multi-level controls.

Methodology for Multi-Level Controls

In fact, person level attributes are correlated with household level attributes. For example, a two-person family household is more like to have one male and one female

as its members than other possible household structures. Therefore, introducing person level attributes can benefit the accuracy of household level attributes of synthetic population.

Several research studies are conducted towards this point and Müller and Axhausen (2011) gives a detailed review on population synthesizers which can control multilevel attributes. Among these researches, Arentze et al. (2007) adopt a two-step procedure each involving IPFs. Another three methods (Guo and Bhat, 2007; Ye et al., 2007; and Auld and Mohammadian, 2010) to be discussed all begin by generating multi-dimensional distributions for household- and person-level attributes independently using IPF techniques.

Guo and Bhat (2007) presented a methodology in which household level joint distribution and person level distribution are generated by IPF separately. Based on the joint distribution of household level controlled variables, a household is randomly selected from the seed data under a specified probability. Then the desirability of the selected household is checked by two requirements. First, adding the selected household will not make the number of such households exceeds a pre-specified threshold value (e.g. 120% of this kind of households in the multidimensional contingency table from IPF). Secondly, for each of the affiliated persons of the selected household, after adding this household the number of persons of this type is lower than a pre-specified threshold value. If a selected household meets both requirements, a copy of this household with all affiliated persons is added into the synthetic population. Otherwise the selected household is ignored.

Therefore, one of the issues concerning this method is to decide the designed threshold values for each type of households and persons. Furthermore, this method does not control household level characteristics and person level characteristics through the whole selection procedure. Because during the beginning of selection, all feasible types of persons do not exceed the pre-specified threshold values, the second desired requirement cannot be violated by any households.

Arentze et al. (2007) uses a different method to deal with the two level controls. In this research, an algorithm of two-step IPF is proposed. The first IPF converts person level attributes (e.g., age and work status) into household level. Through the interaction between age, gender and household size, a household age composition variable is defined based on the combination of the three variables. For example, one male household with his age under some category or two adult household with the age specification of the two members can be consider as different households. In the research of Arentze et al. (2007), age and work status are combined with household demographic variables separately. They called the two new variables (based on age and work status of each person in household) household age composition and household work composition. During the second IPF, the new generated variables together with other household level variables are served as controls in a regular population synthesis. This method takes some person level characteristics into account, but for each person level characteristics, a special design for the structure between person level attributes and household level demographic attributes is need. Furthermore, since not all marginal counts are available for the first IPF, several additional constraints are assumed toward this point.

Ye et al. (2009) proposed an algorithm which can make the synthetic population closely match the true population in terms of household level and person level marginal controls under certain conditions. In this research, household level joint distribution and personal level joint distribution are formulated individually by IPF, and then household types and person types are defined according to cells in household level and person level multidimensional tables. At the beginning of data fitting, each household in seed data is assigned an initial weight value (usually one). The weight values are literately updated until the weighted seed data matches the two multidimensional table generated by IPF. The final weight can be considered as the number of such household which is needed for constructing the synthetic population. Unfortunately, these weight values are usually not integers and hence cannot truly represent the number of household of each type. A simple way for solving this issue is to round off these weight values to the most closed integers. However, for some types of households which have very small weight (e.g. some households have weight less than one), the round off method will introduce some bias. Another method uses Monte-Carlo simulation to select households from seed data. Similar to the research conducted by Beckman et al. (1996), the selection probability for each household is calculated based on the weight values.

Auld and Mohammadian (2010) consider the joint multi-level controls in the household selection stage. In the first stage, the multidimensional tables under household level and person level are generated by IPF separately. Then during the household selection stage, the selection probability is computed based on household level and person level contingency tables together. In the classic selection probability (e.g. Beckman et al. 1996; Guo and Bhat 2007) associated with each household, only

household level information is considered. That is, if a particular household is the type which has large counts in joint tables a higher probability will be assigned. Auld and Mohammadian (2010) treat person level information in the similar manner and the desirability for each person within a particular household are take into account through a new version of selection probability.

Another research study conducted by Pritchard and Miller (2009) conceptually illustrate a mechanism in which, a household could be formulated by combine the pre-generated household level population and person level population. For example, a married couple household must be constructed by one male person and one female person and these two people will also need to meet some age criteria, e.g. the age difference is not exceed some value. It is a complex mechanism at lest because we need to consider a lot of rules such as age difference and gender consistency and some of the rules are subjectively formulated. Moreover, due to the lack of relationship information of non-family household, errors are expected to be large for such types of household. Unfortunately, this literature does not provide detailed numerical example for the method.

Other than the IPF based researches, Ryan et al. (2010) design a protocol to link the generated individual record to household record by specifying the relationship between household members' characteristics. Instead of IPF, they use a combinational optimization method (Williamson, et al., 1998 and Voas and Williamson, 2000) to generate the populations before linking persons to households.

Lee and Fu (2011) proposed a cross entropy optimization model for population synthesis. The cross entropy are defined similarly to the discrimination information

(Ireland and Kullback, 1968) which is also the entropy of the joint household level distribution. Lee and Fu (2011) actually try to formulate entropy for household distributions defined based on multilevel characteristics. More specifically, “cells” in the traditional household level contingency table are further divided by person level characteristics. Because IPF actually is a numerical approach for solving the entropy problem, the cross entropy method will face the same numerical issues as IPF. Namely, if a type (in terms of control table) of households in seed data is missing, the entropy goes to infinity.

In sum, several researches have contributed toward the point of multi-level controls. As reviewed, the methodology for simultaneously controlling household level and person level controls is varied enough. Starting from different perspectives and ideas, these methods utilizes person level information in order to gain more accurate synthetic populations. However, several issues still exist. For example, the stopping criteria is more subjective if threshold values are assigned and there no guarantee that the synthesis population can match the true marginal controls even the weighted household sample (Ye et al. 2009) can match the true marginal controls perfectly. Furthermore, the considered household/person characteristics are limited in past population synthesizes. As a result, if a large set of control variables (e.g. 20 household level controls and 10 person level controls) involves, IPF may not easily to be implemented since the high-dimensional contingency table faces the zero-cell problem as well as convergence problem of IPF procedure.

Towards to this point, this dissertation proposes a new framework to solve or dodge these issues. As the reader will see shortly, the proposed method can

simultaneously control multi-level characteristics (e.g. household level, person level and family level) and the synthesis population can match these controls almost perfectly. In the new developed framework, IPF procedure is optional and hence the zero-cell and convergence problem can be avoided. Furthermore, the new framework also allows partial IPF during synthesizing populations. More specifically, the joint distribution of some attributes can be generated through IPF while other attributes remain their original data format. The more detailed discussion will be presented in Chapter 4.

Applications of Population Synthesis

Since the work of Beckman et al. (1996), this methodology has been applied to support travel-demand modeling in several areas such as Portland Metro, San Francisco, New York, Columbus, Atlanta, Sacramento, Bay Area, and Denver. Bowman and Bradley (2006) and Bowman (2004) provide a general overview of these applications. The Sacramento application is available in Bowman and Bradley (2006) and the Atlanta application and validation results are presented in Bowman and Rousseau (2008).

As the requirement of future consideration of transportation planning, the final objective of SPG is to generate the synthetic population for a future year or sometimes called a target year. Bowman and Rousseau (2008) illustrate the population synthesizer for Atlanta, Georgia, regional commission (ARC) as well as its validation. The base year SPG of ARC is common to many population synthesizers. Specifically, they define the marginal controls per Transportation Analysis Zone (TAZ) by the data source of census summary file and CTPP with PUMS data performs as seed data. Once the base year synthetic populations have been generated, it will be used as seed data for a forecast year. The marginal controls for the forecast year in ARC come from the land use

forecast. Thus the SPG of ARC connects between land use information and travel demand model.

The base-year population synthesis of MORPC (Mid-Ohio Regional Planning Commission) is designed to have identical data structure as the available data for target year (Bowman, 2004).

However, in most of situation, the control information of target year is limited and usually in person level such as the gender and age distribution for an area. Thus the data structure for target year and base year is so different that additional adjusting procedure may be required. For example, in SFCTA (San Francisco County Transit Authority) the available social economic forecasts cannot be used directly in generation of target-year population. More specifically, SFCTA has the social economic characteristics in person level whereas the base year population synthesis is based on household attributes. An additional procedure is used in SFCTA to adjust the corresponding base year control variable.

The population synthesis of ARC also faces the similar issue due to the inconsistent categorization of control variables between base-year population synthesis and target-year population synthesis. To solve this problem, ARC aggregates the control variables in both base-year and target-year population synthesis to closely match the control variables between the two synthetic populations.

Validation Methods

In order to evaluate the performance of population synthesis, we need to measure the accuracy of synthetic population, e.g. the difference between synthetic population and true population. Since SPG is motivated by the situation that the true population is not available, there is no true population for validation. Beckman et al. (1996) use the

sample in a Public Use Microdata Area (PUMA) as several artificial census tracts and a 5% sample from these artificial census tracts is served as seed data. Then the synthetic population is ready to compare with the true population.

Other than the comparison between true population and synthetic population, Auld et al. (2008) verify their SPG with a baseline population from another method. Instead of measure the distance between true population and synthetic population, this research uses weighted average absolute percent difference between synthetic population and the population from baseline SPG as the validation criterion.

Another commonly used validation method dose not requires the true population or some baseline SPG. Under the purpose of validation, some attributes form control data are considered as validation variables and the SPG is implemented without controlling of these validation variables. And then, the distance between the validation variables from synthetic population and the corresponding variables from true population is measurable and hence can be used as a criterion for validations and evaluations (e.g. Auld et al., 2008).

Bowman and Rousseau (2008) validate the performance of SPG with back-casting to 1990. The seed data for back-casting comes from the base year (2000) SPG and the control data is the land use forecast data. With the SPG of 1990, the validation is implemented by comparing the synthetic population of 1990 with the population from census 1990. They assume the land use information is accurate enough and hence the validation is only for SPG itself. However if the assumption false, the effect of SPG and land use information may confounded with each other.

Summary

Synthesis methods are used to first generate the population for a base year (current year/census year) and it, in turn, is used as an input to generate the target year (forecast year) population. The state-of-the-practice approach to population synthesis involves the use of the Iterative Proportional Fitting (IPF) method. While there have been several applications of this approach, the following issues still remain. First, the number of controls used in the synthesis of the base-year population has been limited. In particular, most practical applications control only for household-level attributes (e.g., household size and dwelling-unit type) and not for person-level attributes such as age and gender. Thus, the synthesized base-year population may not truly match the observed person-level distributions. This would affect the accuracy of the target-year population as the synthesized base-year population is used as an input to generate the target-year population. Second, documentation of the validation of the synthesis procedure, especially in the context of a “target” year population, is limited. The broad focus of this research is to contribute towards synthetic population generation (SPG) literature by addressing these issues.

CHAPTER 4
FITNESS BASED SYNTHESIS METHODOLOGY

This chapter presents the idea of fitness based synthesis (FBS) methodology for generating a population with multilevel characteristics being controlled. In the first section of this chapter, the framework of the proposed method is presented and then several aspects of this method are discussed in the following sections.

Framework

Broadly, the procedure involves selecting a set of households (with replacement) from the seed data in such a way that the controls are satisfied. This study is also an extension to our previous work (Srinivasan et al., 2008 and Srinivasan and Ma, 2009).

The fitness-based-synthesis methodology provides a mechanism to synthesize population under multilevel control tables. The proposed method works in the way that households are iteratively selected (with replacement) from seed data until the control tables (at multiple levels) are matched. Further, during the iterative procedure, some households already selected are allowed to be removed if losing such household can contribute to reducing the matching error of control tables.

Figure 4-1 gives a flowchart of the fitness-based synthesis method. Generally, three major components define a FBS framework. The first component is the initial household sets. The initial population can be any reasonable populations. For example, the empty set or the entire seed data could be served as a start point for the whole procedure. There are several considerations upon the choice of initial household sets and the detailed discussion will present shortly. Specifically, in the analysis of Chapter 5, the entire data set is served as the start point and in contrast Chapter 6 and Chapter 7 will use the empty set as the start point.

The second component is the metric used for calculate the fitness for each household. Broadly, the fitness measures the marginal contribution of adding or removing the household for matching with controls. In the procedure, two different fitness values which are referred as “Type I Fitness” and “Type II fitness” are defined for each household. More specifically the Type I Fitness computes the contribution if the household is added to current selection where as the Type II Fitness computes the contribution if the household is removed from current selection.

After the two types of fitness values are calculated for all households. All the households with positive fitness values are considered as candidate. Rigorously speaking, a household is eligible for candidate if it has positive type I fitness or it has positive type II fitness and at least one such household is already in current selected households. There are a lot of choice of fitness functions and any function for measuring the difference between two distributions or vectors could be served as fitness functions. This research uses the square-error based fitness measurement and under such fitness definition, Type I and Type II fitness values cannot be positive simultaneously. It is a very nice property. There are also other properties for square-error based fitness function and detailed explanation will be present shortly.

The third component is the mechanism for assigning probability to each candidate households. In this research, an equal probability is assigned to each candidate households during iteration. Then a household will be selected randomly and after adding or removing this household, the overall match against control tables will be enhanced for sure. Detailed discussion about the three components is present as follows.

Initial Household Sets

This dissertation suggests three different initial household sets. They are (1) the empty sets, (2) the entire data sets of seed data and (3) a random selected sample of seed data. Other than these three types of sets, readers can design their own initial household sets based some particular reason. For example, you can even use the population from other population synthesis. For the three proposed types of initial household sets, the empirical results show that they perform similar in the context of validation. Another consideration of initial household sets is the running time of the procedure. The empty sets are the simplest specification of the initial household sets but the running time probability is the longest. Because in the beginning of the procedure, if the empty sets are severed as initial household sets, almost all households own positive Type I fitness and hence the procedure is identical to randomly select household from seed data without check the fitness. Therefore, using a randomly selected sample of seed data is computationally efficient.

Fitness Functions

Two fitness values of each household of seed data can be computed by equation (4-1) and equation (4-2). The two fitness functions are called as Type I fitness and Type II fitness respectively.

$$F_I^{in} = \sum_{j=1}^{J} \sum_{k=1}^{K_j} \left[\left(R_{jk}^{n-1} \right)^2 - \left(R_{jk}^{n-1} - HT_{jk}^{i} \right)^2 \right] \qquad (4-1)$$

$$F_{II}^{in} = \sum_{j=1}^{J} \sum_{k=1}^{K_j} \left[\left(R_{jk}^{n-1} \right)^2 - \left(R_{jk}^{n-1} + HT_{jk}^{i} \right)^2 \right] \qquad (4-2)$$

where $R_{jk}^{n-1} = T_{jk} - CT_{jk}^{n-1}$

Conceptually, F_I^{in} calculates the reduced sum of squared error of controlled tables if the ith household being selected for adding at the nth iteration while F_{II}^{in} is the corresponding error the ith household being selected for removing at the nth iteration. In the above formula, j is an index representing the control (and the corresponding count) tables and J is the total number of control or count tables. Here count tables have the same structure as control tables and they are used to aggregate current selected households in the structure of corresponding control tables. For example, $j = 1$ could represent the joint distribution of household size against tenure; $j = 2$ could represent the joint distribution of age against gender; and so on.

For each control (count) table j, k is an index representing the different cells in that table. For example, in table $j = 1$ (e.g. household size against tenure), k could have values from 1 through 14 representing the 14 different cells (7 categories for household size multiplied by the 2 categories for tenure). Therefore, for this table, $k = 1$ represents the first cell (1 person / own household), $k = 2$ represents the second cell (2 person / own household), and so on. T_{jk} represents the value of cell k in control table j, and it is also the target number of households of a particular type to be synthesized. CT_{jk}^{n-1} represents the value of cell k in count table j after iteration $n-1$. At initialization $n = 1$, all values of the count tables are set to be the value according to the start point of population. After each operation (adding or removing), the values of the cells in the count tables are updated based on the type of the household drawn.

Based on the above definitions $R_{jk}^{n-1} = T_{jk} - CT_{jk}^{n-1}$is the number of households/persons required to satisfy the target for cell k in control table j after iteration $n-1$. This is calculated as the difference between the corresponding cell

values of the control and the count tables. HT_{jk}^{i} is the contribution of the ith household in the seed data to the kth cell in control table j.

$R_{jk}^{n-1} - HT_{jk}^{i}$ is the required number households to achieve the target in cell k of control table j if household i is added and $R_{jk}^{n-1} + HT_{jk}^{i}$ is the required number households to achieve the target in cell k of control table j if household i is removed. With the terms R_{jk}^{n-1}, $R_{jk}^{n-1} - HT_{jk}^{i}$ and $R_{jk}^{n-1} + HT_{jk}^{i}$, functions can be constructed to calculate the overall fitness of the household. The one adopted in this study is presented in equation (4-1) and (4-2). A comparison of the performance of the algorithm for under different functional forms for the fitness calculations is an area of future study. In addition, it is also useful to note here that the present algorithm assumes that all control tables are equally important. If this is not the case, (for example, if matching the household-size distribution is more important than matching the ethnicity distribution) weights can be added to reflect the relative importance of the different tables.

Selection Mechanism

We believe the seed data is the best available sample to represent the true population. In according to assumption, all households in the candidate set should have equal probabilities. In addition, some household survey provides a sampling weight associate with each household and instead of the equal probability, the selection probabilities can be assigned based on the sampling weight.

Some Properties

It is can be proved that for fixed i and n, F_{I}^{in} and F_{II}^{in} cannot be positive simultaneously. Combine the equations of (4-1) and (4-2), it is easy to show that

$$F_I^{in} + F_{II}^{in} = -2\sum_{j=1}^{J}\sum_{k=1}^{K_j}\left(HT_{jk}^{i}\right)^2 \qquad (4-3)$$

We know that $\sum_{j=1}^{J}\sum_{k=1}^{K_j}\left(HT_{jk}^{i}\right)^2 \geq J$, because any household will contribute to each household level control table by one and to each person level control table by at least one.

So, $F_I^{in} + F_{II}^{in} \leq -2J$ for fixed i and n. Therefore for one certain type of households, only one possible operation between adding and removing can reduce the overall matching error of control tables. More specifically, Table 4-1 lists the feasible values of F_I^{in}, F_{II}^{in} and corresponding operations.

Then all households in seed data with positive Type I fitness or positive Type II fitness could be considered as the candidate household for selection. In addition, households with positive type II fitness will be ignored if there is no such household in the current selection. The fitness-based mechanism actually forces the matching error smaller during each iteration and close to zero eventually if there are sufficiently various types of households in seed data.

A natural termination criterion for the algorithm is when no household can meet the requirement of being a candidate of selection. It is necessary to proof the existence of such criterion, namely the number of iterations is bounded above.

Let's assume the sequence of selected households are indexed by $s_1, s_2, \cdots$, and $T_{jk} = R_{jk}^{0}$. If we denote the operation of household s_l be u_l where $u_l = 1$ if the household was added and $u_l = -1$ if the household was removed. Then we have,

$$\sum_{j=1}^{J}\sum_{k=1}^{K_j}\left(R_{jk}^{n}\right)^2=\sum_{j=1}^{J}\sum_{k=1}^{K_j}\left(T_{jk}\right)^2-\sum_{l=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K_j}\left[\left(R_{jk}^{n-1}\right)^2-\left(R_{jk}^{n-1}-u_l HT_{jk}^{s_l}\right)^2\right]$$

$$=\sum_{\mathrm{j=1}}^{\mathrm{J}}\sum_{\mathrm{k=1}}^{\mathrm{K_j}}\left(\mathrm{T_{jk}}\right)^2-\sum_{\mathrm{l=1}}^{\mathrm{n}}\mathrm{F^{s_l l}},\qquad(4-4)$$

where $F^{s_l l}=max\left(F_I^{s_l l},F_{II}^{s_l l}\right)$, namely the positive fitness of household s_l at the lth iteration. We already know $F^{s_l l}>0$ for all $l=1,\cdots,n$. Because all the calculations are based on basic mathematical operation (addition, subtraction, multiplication and no division) of integers, $F^{s_l l}\geq 1$ for all $l=1,\cdots,n$. Therefore,

$$n\leq\sum_{l=1}^{n}F^{s_l l}=\sum_{j=1}^{J}\sum_{k=1}^{K_j}\left(T_{jk}\right)^2-\sum_{j=1}^{J}\sum_{k=1}^{K_j}\left(R_{jk}^{n}\right)^2\leq\sum_{j=1}^{J}\sum_{k=1}^{K_j}\left(T_{jk}\right)^2\qquad(4-5)$$

So, we just find an upper bound for number of iterations $n\leq\sum_{j=1}^{J}\sum_{k=1}^{K_j}\left(T_{jk}\right)^2$, which justifies the proposed natural termination criterion. Actually such upper bound is very conservative since in the beginning of the algorithm, the fitness value of each household is very larger.

Empirical results show that the synthesis population can match the control tables almost perfectly. Namely, the number of required households for each category of each control tables is around zero. But it is difficult to calculate the fitting errors before the procedure. Now, a geometrical explanation about the fitting errors is present.

Because at the end of the procedure as no household can meet the requirement of being a candidate of selection, most of households will have negative Type I and negative Type II fitness values. The following illustration ignores the households with positive Type II fitness values but do not exist in current selection.

Assume the fitting errors for each category is E_{jk} and the two types of fitness at the end of procedure are F_I^i and F_{II}^i for ith household. Then we have,

$$F_I^i = \sum_{j=1}^{J}\sum_{k=1}^{K_j}\left[\left(E_{jk}\right)^2 - \left(E_{jk} - HT_{jk}^i\right)^2\right] = \sum_{j=1}^{J}\sum_{k=1}^{K_j}\left[2E_{jk}HT_{jk}^i - \left(HT_{jk}^i\right)^2\right] \quad (4-6)$$

$$F_{II}^i = \sum_{j=1}^{J}\sum_{k=1}^{K_j}\left[\left(E_{jk}\right)^2 - \left(E_{jk} + HT_{jk}^i\right)^2\right] = \sum_{j=1}^{J}\sum_{k=1}^{K_j}\left[-2E_{jk}HT_{jk}^i - \left(HT_{jk}^i\right)^2\right] \quad (4-7)$$

By the stop criterion, $F_I^i \le 0$ and $F_{II}^i \le 0$ for all i. Therefore for all i

$$\begin{cases} \sum_{j=1}^{J}\sum_{k=1}^{K_j}\left[2E_{jk}HT_{jk}^i - \left(HT_{jk}^i\right)^2\right] \le 0 \\ \sum_{j=1}^{J}\sum_{k=1}^{K_j}\left[-2E_{jk}HT_{jk}^i - \left(HT_{jk}^i\right)^2\right] \le 0 \end{cases} \quad (4-8)$$

$$\Rightarrow \begin{cases} \sum_{j=1}^{J}\sum_{k=1}^{K_j}E_{jk}HT_{jk}^i \le \frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{K_j}\left(HT_{jk}^i\right)^2 \\ \sum_{j=1}^{J}\sum_{k=1}^{K_j}E_{jk}HT_{jk}^i \ge -\frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{K_j}\left(HT_{jk}^i\right)^2 \end{cases} \quad (4-9)$$

or,

$$\left|\sum_{j=1}^{J}\sum_{k=1}^{K_j}E_{jk}HT_{jk}^i\right| \le \frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{K_j}\left(HT_{jk}^i\right)^2 \quad (4-10)$$

Strictly speaking, the number of effective inequalities in (4-10) is the number of distinct households in seed data, since $HT_{\boldsymbol{jk}}^{\boldsymbol{i}}$ is the contribution of ith household in table j at category k.

For fixed household $\boldsymbol{i}$, $\sum_{j=1}^{J}\sum_{k=1}^{K_j} E_{jk} HT_{jk}^{i}$ is a linear combination of E_{jk} and HT_{jk}^{i} is the correspond coefficients. So the vector of E_{jk} will locate between the two hyperplanes of $\sum_{j=1}^{J}\sum_{k=1}^{K_j} E_{jk} HT_{jk}^{i} = \pm\frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{K_j}\left(HT_{jk}^{i}\right)^2$ in the Euclid space of dimension $\sum_{j=1}^{J} K_j$. Therefore the system of inequality (4-10) represents an interaction of all space of all distinct households in seed data. When the number of distinct of households are large, it is expect that the vector of E_{jk} are bounded in a small space around origin. In the other words, a diverse seed data is the foundation to have small fitting errors.

The proposed method can also be considered as a heuristic solution search for the following integer program (IP) problem.

Min $\quad Distributional\ discrepancy\ between$

$\quad\quad true\ population\ and\ synthetic\ population$

Subject to $\quad control\ tables \pm tolerance$

Above IP problem try to minimize the distribution discrepancy between true population and synthetic population and it can be specified by some discrepancy measurement e.g., entropy and Hellinger distance. The constraints are specified by the matching errors with control tables. The integer decision variables actually are the number of household selected for each household in seed data. By FBS method, in each iteration the random-selection mechanism tries to transfer the distributional information from seed data and hence reduce the distribution discrepancy. The fitness functions make the controlled tables are satisfied at the end of the procedure. For practical purposes, when the number of decision variables is huge, it is difficult to solve

the IP problem. However, the proposed FBS method actually provides a heuristic search method for solving the IP problem and it can easily handle the problem with huge number of decision variables.

Conceptual Comparison with IPF Based Methods

In fact such random-selection mechanism is analogous to IPF. As mentioned before, IPF actually keeps the odds ratio between different attributes and also minimized the discrimination information under the restriction of controls. Now the random-selection mechanism also inherits the relationship among different characteristics even in person level and the interaction between household level and person level. Because of the presence of constrains (control tables), the candidate set for selection is part of seed data otherwise the discrimination can achieve zero for a sufficiently large population. So, the FBS procedure naturally tries to find an integer solution to “minimize” the discrimination information under the controlled table.

There is no zero-cell problem and hence the issue of infinite entropy and infinite discrimination information can be avoided. Based on the new mechanism, controlled attributes from different universe e.g. household level, person level, family level and so on can be readily matched by incorporate these control table to the fitness functions. For the same reason, huge number of attributes can also be controlled simultaneously through the new approach while the IPF or entropy based method will face zero-cell problem if there are too many attributes and categories.

The new approach has a highly integrated framework, it can directly select households during iterations and hence all the resulting numbers of each type of households are integers. Empirical results show that synthetic population can match the controlled tables almost perfectly. In contrast, the conventional methods need additional

steps converting the results from IPF to appropriate integers and hence the controlled information may not be perfectly satisfied.

Summary

The Fitness Based Synthesis (FBS) approach is introduced in this chapter. This method is designed to select households in a way that marginal control tables are matched. The proposed framework contains three major components, namely initial household sets, fitness functions and selection mechanism. The initial household set is prepared before the procedure and proper selection of initial household sets can reduce the running time of procedure. Fitness function measures the marginal contribution of each household for reducing matching errors. In this study, we choose fitness function based on the criterion of sum of squared errors and it can be replaced by other criteria. During the selection process, this study assigns equal probability to the candidate households since we believe that household in the seed data are equally important. However, if there is some information indicates that some types of households are more important to the area of interest, the selection probabilities can be adjusted accordingly.

The proposed method is not a member of IPF-based population synthesizers and hence there is no related zero-cell issue and also no convergence issue under square error based fitness function (adopted in this dissertation) because the number of iterations is proved to be bounded. Furthermore, under square error based fitness functions, the synthetic population can match the marginal control tables almost perfectly if there are sufficient variable types of households.

Table 4-1. Feasible value of fitness functions and corresponding operations

F_I^{in}	F_{II}^{in}	Operation
> 0	< 0	Adding
< 0	> 0	Removing
$= 0$	< 0	Nothing
< 0	$= 0$	Nothing
< 0	< 0	Nothing

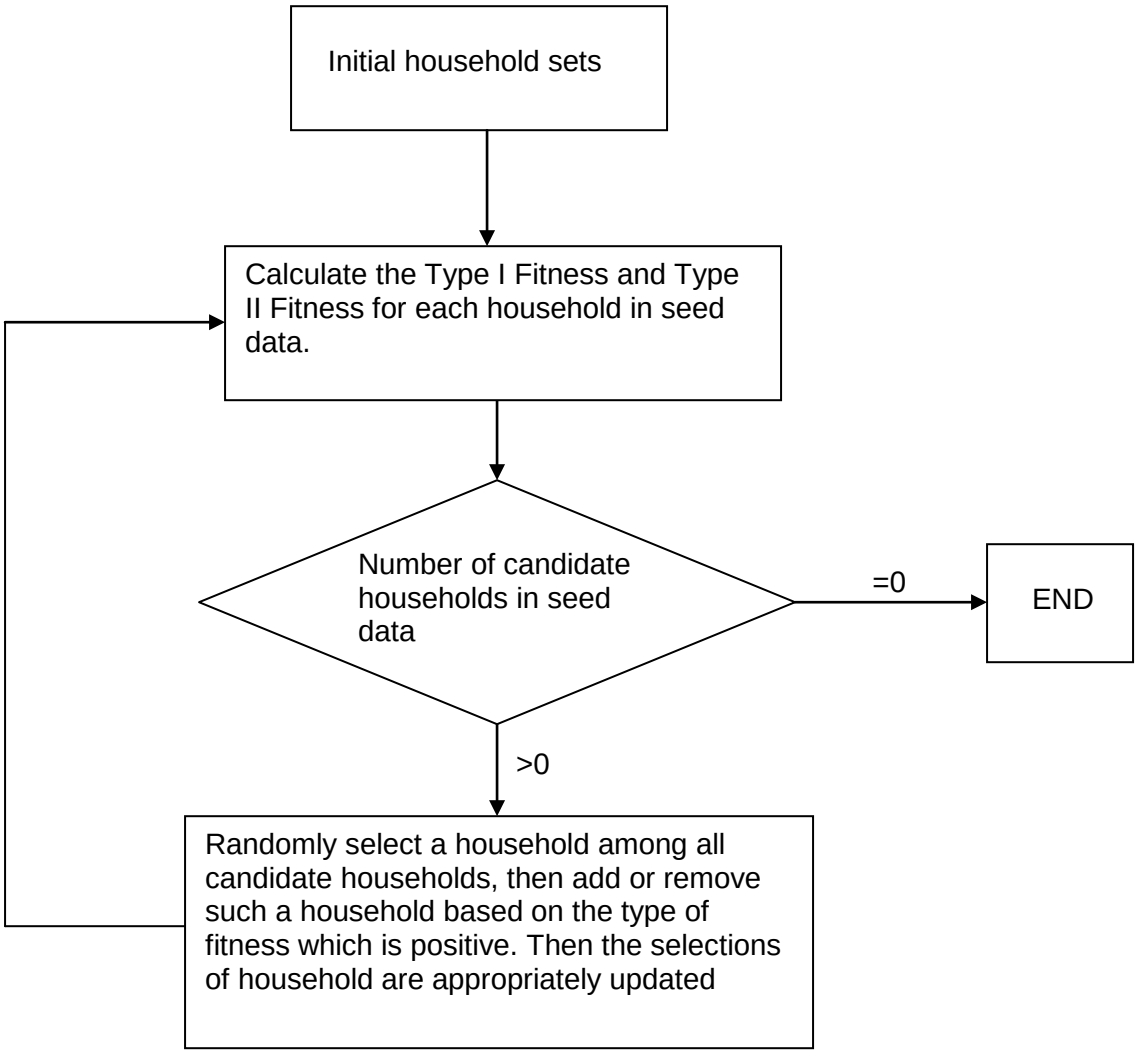

Figure 4-1. Flowchart of the Fitness Based Synthesis method

CHAPTER 5
BASE-YEAR POPULATION SYNTHESIS: COMPARISON AND VALIDATION

In order to evaluate the quality of synthetic populations, it is necessary to measure the accuracy of synthetic population, e.g. the difference between synthetic population and true population. However, the reason for generation synthetic populations is the lack of true population. So, it is difficult to compare the synthetic population with the true population in practice, but sometimes a designed or artificial true population can help to do such analysis.

Beckman et al. (1996) use several PUMS samples as artificial census tracts and a 5% sample from these artificial census tracts is served as seed data. Then it is feasible to compare synthetic populations with the designed true population.

As mentioned before, in synthetic population procedure, the true population is unknown. In order to capture the true population for validation purpose, this study is going to design several artificial census tracts which are similar to the idea of artificial census tracts provided by Beckman et al. (1996). In this chapter, PUMS data from 22 PUMAs in Florida are served as the artificial census tracts. These are sets as the “true” populations of these tracts. The control tables are generated by aggregating attributes of these artificial true populations. The seed data for synthesis are drawn randomly from these PUMAs. The designed control tables and the designed seed data are then fused to the Fitness Based Synthesis framework described in Chapter 4 for generating the population of these artificial census tracts. With the artificial true populations and synthetic populations, all information for input to validation is available at this point.

Before validation, list-based households set are transferred to frequency-based household sets. More specifically, PUMS data lists households row by row and each

household is defined by household characteristics as well as all members' characteristics and households in different row may have same specification of attributes, especially when the sample size is large. Now, if a household is defined by a particular combination of household characteristics and its members' characteristics, a set of households can also be defined by a list of frequencies of household types. Based on the new definition, we can compare two sets of households by comparing the two lists of frequencies. Conceptually, validation is the same as measuring the distance of two vectors or similarity between two distributions.

According to the above validation idea, the following analysis will compare several synthetic populations from different methods which include the conventional IPF method (Beckman et. al. 1996) and the recently developed IPU method (Ye et. al. 2009) with true population. This is accomplished by comparing the actual number of households of each "type" in the true population with the corresponding number in the synthesized populations.

Dataset

Before validation, synthetic populations of 22 artificial census tracts are generated first. As illustrated in Figure 2-1, base year population generation requires two types of data of input, seed data and control tables. The seed data comes from the 5% random samples of the true population of 22 artificial census tracts. And five control tables (an example in Figure 5-1) adopted for the validation analysis.

Pre-Treatment of Seed Data

The seed data for the population synthesis come from the PUMS. As the PUMS data represent only a 5% sample of the overall population, it is possible that there are certain types of households (especially "rare" households) which are represented in the

tract-level control tables but are not present in the PUMS data from the corresponding PUMA. For example, a control table may indicate few “4-persons and 1-car” households to be present in a census tract. However, the 5% PUMS data from the PUMA to which this tract belongs may not have any such households. The pre-treatment procedure simply augments the PUMS data for each PUMA by adding such missing household types from other PUMAs. Our current procedure ensures that each PUMA has at least one household that satisfies each cell (independently) in each of the control tables. One household of the missing type is borrowed (arbitrarily, in the current implementation) from some other PUMA to satisfy this requirement. Overall, the pre-treatment procedure broadly ensures consistency between the seed data and the control tables and, therefore, it would always be possible to find a household in the seed data towards satisfy each cell of the control tables.

To be noted, it is not necessarily to pre-treat the seed data for the new method and the result synthetic population will leave those missing cell unsatisfied, which is still the “best” population under the original seed data. However, if some categories of control tables are missing samples from seed data, the IPF cannot have a feasible solution since some constrains are not satisfied.

The following analysis will also incorporate two IPF related population synthesizers. So it is necessarily to pre-treat the seed data to serve the three population synthesizers.

Control Tables

Five control tables are considered here and Figure 5-1 provides an example of these tables which is also control tables for artificial census tract with ID A6. The five control tables consists two tables (household size × tenure and number of

vehicles×tenure) from household level, two tables (age × gender and ethnicity) from person level and one table (age × gender) from group quarter level.

Validation Method

Essentially, validating a synthetic population is to measure the difference or similarity between synthetic population and true population. Therefore, two questions are raised, namely how to define a population more comprehensively? And how two measure the difference between two populations based on the definition?

Defining Household Types

As mentioned before, the validation is based on frequency-based household sets. Namely, a list-based household sets are transferred into the distribution of household types. This study defines a type of household based on its household level attributes and each member's person level attributes. It is not recommended to define household types only through a particular controlled attribute since population synthesis can force the generated population to match controlled population and as a result the difference can be close to zero which makes the population looks accurate but the intrinsic errors are veiled.

Conventionally, a household are characterized by only household level attributes and hence the maximum possible number of types is the product of number of categories of attributes even though some of these are not practically feasible (such as a one person seven car household). However, if the household types are defined based on a combination of household and person level attributes, it is difficult to predefine all types of households. That is due to that we do not know which types of households exist in reality and if we enumerate all types of household the number is huge and more likely that most of them are rarely distributed. This study finally counts all types of

households in the true population of 22 artificial census tracts as the set of possible household type.

The household and person attributes considered for defining household types are listed in Table 5-1 and corresponding 8529 types of households are classified. Such household type definition considers the interaction between household level attributes and person level attributes and we can consider that it is in the household composite level. For instance, a household type under this classification scheme could be a two person family household with two cars with one person being a white male in age category 35-44 years and the other person is a white female in the age category 25-34 years. More precisely, if two households are counted into one type, these two households will have identical household level attributes and persons in any household are one to one corresponding to persons in another household with identical person level attributes.

Measures of Dissimilarity between True- and Synthesized- Populations

Once the household types have been defined and the frequency of each type are determined for the true- and synthesized- populations, we need measures to determine the extent to which these are dissimilar. Let t_i and s_i be the number of households for $i\text{th}$ type of true population and synthetic population respectively and I be the total number of household types (e.g. $I = 8529$ for this study). There are several measures can be defined.

$$\text{MAE} = \frac{1}{I}\sum_{i=1}^{I}|t_i - s_i| \qquad (5-1)$$

Equation (5-1) defines the mean absolute error (MAE) as the average absolute value of the difference between the actual number of households and the synthesized number of households of different types.

$$\mathrm{MSE} = \frac{1}{I}\sum_{i=1}^{I}(t_i - s_i)^2 \qquad (5-2)$$

In equation (5-2) the mean squared error (MSE) is determined as the average of square differences across all household types.

$$\mathrm{H}^2(\mathrm{T},\mathrm{S}) = \frac{1}{2}\sum_{i=1}^{I}\left(\sqrt{t_i/t} - \sqrt{s_i/s}\right)^2 \qquad (5-3)$$

$$\text{where } t = \sum_{i=1}^{I} t_i \text{ and } s = \sum_{i=1}^{I} s_i$$

An alternative distance measure is Hellinger distance (see Simpson, 1987 and Karlis and Xekalaki, 1998 for more information) presented in Equation (5-3). The Hellinger distance is used to measure the distance between two probability distributions or non negative vectors. If the distributions of the two populations match perfectly, the Hellinger distance is zero and if the distributions are perfectly disjoint, the distance is 1. Therefore, smaller values of the Hellinger distance indicate a closer match of the synthetic population with the true population. Unlike MSE and MAE, Hellinger distance is more intend to capture the distributional difference between two distributions without sensitivity to the size of population and it is also not sensitive to outliers.

Another measurement is the proportion of types of household such that the absolute difference $|t_i - s_i|$ exceeds a predefined threshold values. Following analysis uses 5% of type size and they are denoted as “Threshold 5%” in the following analysis.

Comparison with Other Methods

This study compares three different synthetic populations over 22 artificial census tracts. Besides the proposed FBS approach, the conventional IPF approach (refer to Beckman et al. 1996) and IPU approach (refer to Ye et al. 2009) are also implemented on these artificial census tracts. All the three methods use the same PUMS data and the FBS approach and IPU approach use all the 5 control tables (an example in Figure 5-1). However, IPF method only adopts household level controls.

Table 5-2 summarizes the numbers of households, male individuals and female individuals for these generated populations. Comparing these numbers to true populations, the three methods generate populations that match to the total number of household of true population. However, IPF approach cannot match the gender distribution since the person level controls are not included. Since the control tables are true information of census tracts, matching with controlled information (number of household, number of male and number of female are controlled information here) is an important aspect of population synthesizers. Nevertheless more comprehensive comparisons should be applied through uncontrolled information, namely the validation.

In Table 5-3, validation of populations of the three methods is presented. For each of the 22 artificial census tracts, four different measurements are applied to three different synthetic populations. By compare the errors across the three method, the FBS approach and IPU approach generate populations closer to true population than the one from conventional IPF method because this two methods have smaller errors for most of census tracts. Since the former two methods control person level characteristics, they are expected to be more accurate. In term of validation criteria there are no superior between FBS and IPU. More specifically, FBS is better in some cases while IPU is

better in other cases. However, IPU cannot provide solution for some cases due to the convergence issues.

Generally, there are three important aspects which affect the accuracy of synthetic populations. The most fundamental factor is the seed data. More specifically, if the seed data is more similar to true population, it is expected to be more accurate for synthetic population. The second factor is the population synthesizer itself. A good synthesizer should preserve the information from seed data while match the control information. The last factor is the relationship between the controlled information and the information of interest (usually is not controlled but used for validation). In other words, if the controlled attributes are highly correlated with the uncontrolled attributes, the synthetic population is more likely to have less error if validated by the uncontrolled attributes.

In addition, Table 5-4 gives the number of iterations of FBS procedure for the 22 artificial census tracts. In this example, the number of adding iterations is constantly increasing as the increase of tract-level population size while the number of remove iterations does not has such pattern. To be remembering, the initial household sets are the seed data (1542 households in this analysis). Therefore, for smaller census tracts (e.g., A1 and A2) most of the iterations remove households. Irrespective of tract size, it is interesting that the number of iterations is correlated to Hellinger distance, e.g., for case A22, even it is a large tract, numbers of adding and remove iterations are relatively smaller than the one of A20 and A21. Intuitively, smaller Hellinger distance reflects more similarity between synthetic population and true population and hence the selection procedure can easily fit to the controlled tables. In other words, if the seed

data is highly related to true population, there is less iterations with other factors fixed. Moreover, the total number of iterations is compatible to the tract size in this example.

Figure 5-2 demonstrates more detailed difference between synthetic population and true population for the census tract ID A6. Each point represents a household type, X axis is the size of this type of true population and Y axis is the size of this type of synthetic population. Therefore a perfect synthetic population (zero validation error) will locate on the 45 angle straight line and we can roughly observe the validation error through the concentration of the point cloud on the 45 angle straight line.

Summary

This study provides a more comprehensive validation conception in which a set of households first are transferred to a representation of household-type distribution and then the synthetic population is validated by measuring the difference between two household-type distributions. Such difference can be evaluated by the difference of two vectors or similarity of two distributions. Several criteria are introduced toward this point.

Based on the proposed FBS method, synthetic populations from 22 artificial census tracts are generated. As an important objective, this dissertation compares the proposed method with other synthesis method. So, the synthetic populations of the same 22 census tracts are also generated by the conventional IPF method and a recently developed IPU method. According to validation, the FBS approach and IPU approach generate populations closer to true populations than the one from conventional IPF method.

There are several avenues for the future research. First, it is useful to test different fitness functions under the new mechanism. Second, the quality of synthetic population

is affected by several aspects, e.g. seed data. So the analysis of impact of seed data or other aspects can also be conducted.

Table 5-1. Characteristics for defining household types

Attributes	Categories
household level	
Household Size	1,2,3,4,5,6,7+
Tenure	Rent, Own
Number of Vehicles	0,1,2,3,4,5+
person level	
Ethnicity	White, Black, Other, Multiple Race
Gender	Male, Female
Age	0-5,6-15,16-17,18-24,25-34,35-44,45-54,55-64,65-74,Over 75

Table 5-2. Aggregate comparisons of the true- and synthesized- populations for 22 artificial census tracts

Case ID	Households[a]				Male Population				Female Population			
	True Values	Synthesized			True Values	Synthesized			True Values	Synthesized		
		FBS	IPF	IPU		FBS	IPF	IPU		FBS	IPF	IPU
A1	225	231	225	213	251	248	218	228	249	252	291	260
A2	460	455	462	439	469	469	506	435	532	533	503	513
A3	616	617	612	585	715	714	733	668	785	785	756	790
A4	801	806	801	761	1006	1008	974	958	994	991	1030	970
A5	1014	1015	1017	1015	1197	1198	1215	1165	1305	1304	1308	1314
A6	1400	1391	1404	1389	1440	1442	1463	1399	1560	1562	1546	1548
A7	1794	1783	1794	1790	1881	1882	1868	1880	2032	2034	2038	2045
A8	1673	1665	1672	1661	2131	2137	2046	2100	2062	2057	2151	2067
A9	1489	1514	1487	NA[b]	2237	2236	2250	NA	2485	2482	2480	NA
A10	1965	1967	1968	1926	2479	2480	2398	2430	2493	2492	2599	2466
A11	2002	2010	1997	1978	2403	2402	2411	2367	2617	2615	2616	2612
A12	2214	2217	2216	2195	2439	2438	2446	2411	2707	2708	2685	2700
A13	2588	2576	2589	2585	2718	2723	2756	2701	3094	3093	3065	3117
A14	2363	2365	2363	2352	2961	2960	2869	2928	3013	3014	3100	3021
A15	2474	2479	2479	2454	3137	3138	3084	3105	3056	3054	3137	3056
A16	3110	3133	3109	NA	3192	3187	3142	NA	3464	3465	3515	NA
A17	3335	3361	3334	NA	3443	3438	3373	NA	3713	3714	3752	NA
A18	2979	2986	2973	2966	3676	3676	3608	3647	3798	3798	3852	3821
A19	3342	3340	3343	NA	3866	3869	3837	NA	3795	3792	3843	NA
A20	2982	3022	2983	NA	3598	3597	3703	NA	4102	4096	4004	NA
A21	3654	3644	3654	3656	3790	3792	3829	3775	4124	4125	4074	4153
A22	3763	3763	3765	3730	4413	4414	4419	4377	4454	4453	4466	4434

a Group quart population are counted as household

b “NA” means that Ye’s method does not converge and hence no feasible solutions

Table 5-3. Validation results of population from three population synthesizers

Case ID	MSE			MAE			Hellinger			Threshold 5%		
	FBS	IPF	IPU	FBS	IPF	IPU	FBS	IPF	IPU	FBS	IPF	IPU
A1	0.059	0.049	0.059	0.040	0.037	0.040	0.683	0.626	0.718	0.033	0.032	0.034
A2	0.124	0.142	0.127	0.072	0.076	0.072	0.581	0.622	0.601	0.055	0.058	0.058
A3	0.157	0.167	0.130	0.090	0.091	0.087	0.544	0.532	0.537	0.071	0.069	0.074
A4	0.252	0.308	0.227	0.120	0.128	0.118	0.540	0.575	0.562	0.092	0.090	0.095
A5	0.264	0.303	0.223	0.137	0.141	0.134	0.487	0.491	0.476	0.103	0.098	0.108
A6	0.576	0.669	0.725	0.190	0.199	0.190	0.461	0.481	0.467	0.113	0.120	0.115
A7	0.938	0.935	0.714	0.252	0.252	0.232	0.480	0.483	0.443	0.137	0.146	0.146
A8	0.672	0.787	0.549	0.232	0.240	0.221	0.482	0.487	0.464	0.148	0.147	0.154
A9	0.630	0.632	NA[a]	0.256	0.254	NA	0.641	0.646	NA	0.165	0.169	NA
A10	0.788	0.918	0.713	0.261	0.271	0.228	0.460	0.472	0.442	0.161	0.159	0.166
A11	0.668	0.750	0.761	0.247	0.256	0.252	0.418	0.439	0.441	0.154	0.157	0.155
A12	0.658	0.891	0.680	0.256	0.273	0.252	0.395	0.414	0.387	0.163	0.158	0.165
A13	1.284	1.625	0.987	0.266	0.302	0.247	0.324	0.367	0.311	0.139	0.149	0.144
A14	0.978	1.150	0.840	0.309	0.318	0.297	0.448	0.461	0.431	0.180	0.177	0.183
A15	1.452	1.462	1.069	0.342	0.343	0.319	0.459	0.459	0.438	0.188	0.190	0.197
A16	1.925	2.006	NA	0.401	0.392	NA	0.429	0.408	NA	0.201	0.206	NA
A17	2.168	2.109	NA	0.429	0.407	NA	0.428	0.402	NA	0.211	0.215	NA
A18	1.508	1.610	1.214	0.381	0.382	0.366	0.436	0.431	0.416	0.201	0.202	0.208
A19	2.941	3.112	NA	0.447	0.443	NA	0.422	0.425	NA	0.211	0.217	NA
A20	2.700	2.100	NA	0.461	0.448	NA	0.513	0.531	NA	0.222	0.242	NA
A21	3.616	3.163	3.480	0.515	0.487	0.467	0.447	0.426	0.408	0.207	0.220	0.214
A22	2.169	2.074	1.681	0.470	0.465	0.444	0.404	0.406	0.388	0.230	0.232	0.238

a “NA” means that IPU method does not converge in that situation

Table 5-4. Number of iterations for generating population of 22 artificial census tracts

Case ID	Number of generated household	Number of adding iterations	Number of remove iterations	Number of total iterations
A1	231	127	1438	1565
A2	455	260	1347	1607
A3	617	203	1128	1331
A4	806	257	993	1250
A5	1015	284	811	1095
A6	1391	927	1078	2005
A7	1783	1360	1119	2479
A8	1665	733	610	1343
A9	1514	1345	1373	2718
A10	1967	1155	730	1885
A11	2010	1127	659	1786
A12	2217	1094	419	1513
A13	2576	2067	1033	3100
A14	2365	1631	808	2439
A15	2479	1830	893	2723
A16	3133	2863	1272	4135
A17	3361	3216	1397	4613
A18	2986	2570	1126	3696
A19	3340	3133	1335	4468
A20	3022	4274	2794	7068
A21	3644	4285	2183	6468
A22	3763	3327	1106	4433

Household Size × Tenure (Household level)							
	1	2	3	4	5	6	7+
Own	165	253	98	82	36	12	3
Rent	278	192	94	53	12	6	8

Number of Vehicles × Tenure (Household level)						
	0	1	2	3	4	5+
Own	25	209	301	90	19	5
Rent	85	338	166	38	14	2

Age× Gender (Person Level)										
	0-5	6-15	16-17	18-24	25-34	35-44	45-54	55-64	65-74	Over 75
Male	95	164	31	351	218	171	191	102	65	52
Female	106	163	32	367	223	184	198	120	77	90

Ethnicity (Person Level)			
White	Black	Other	Multiple Race
1910	952	88	50

Age× Gender (Group Quarter Level)			
	0-17	18-64	Over 65
Male	0	54	5
Female	0	41	8

Figure 5-1. Examples of control tables

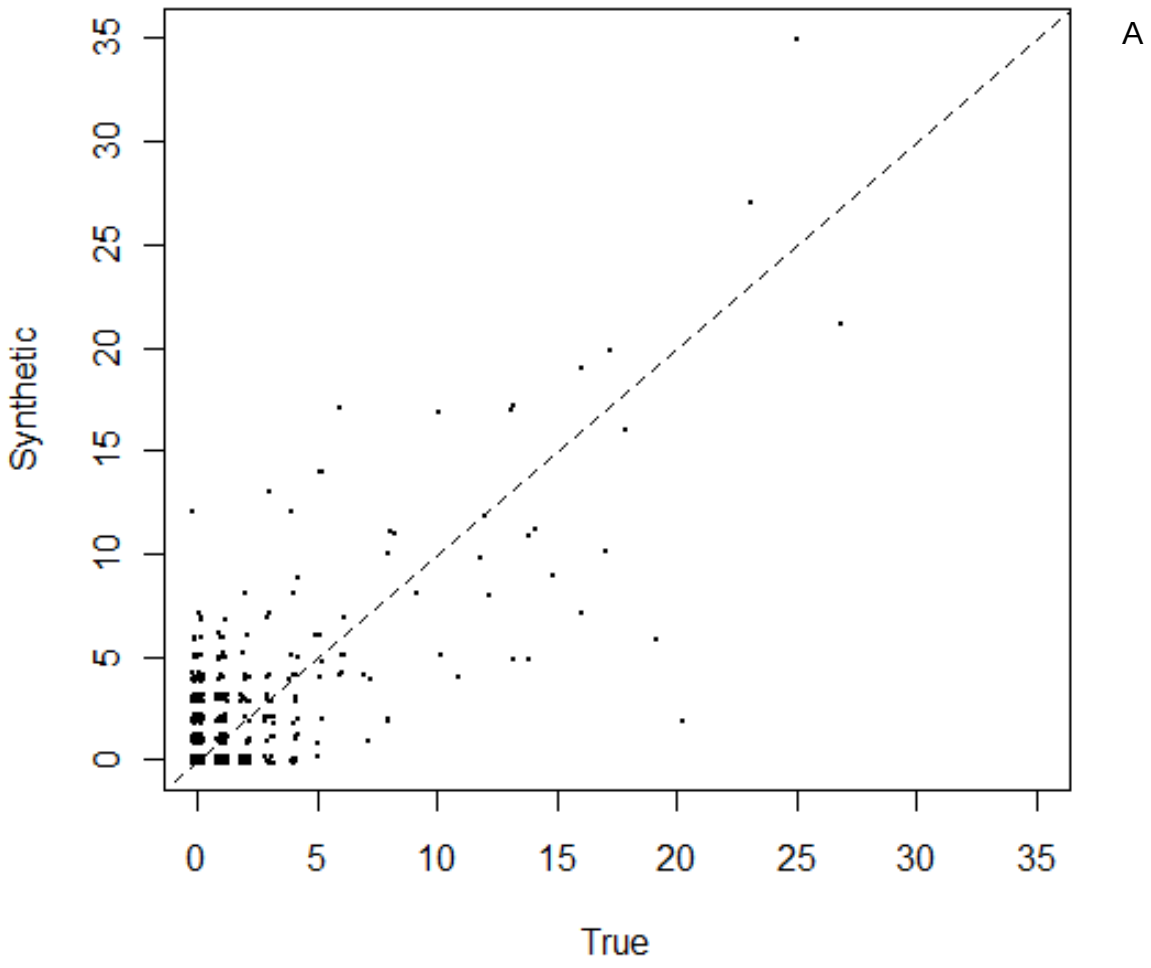

Figure 5-2. Scatter plots of three different synthetic populations against true population on artificial census A6. A) FBS Method. B) IPF Method. C) IPU Method.

B

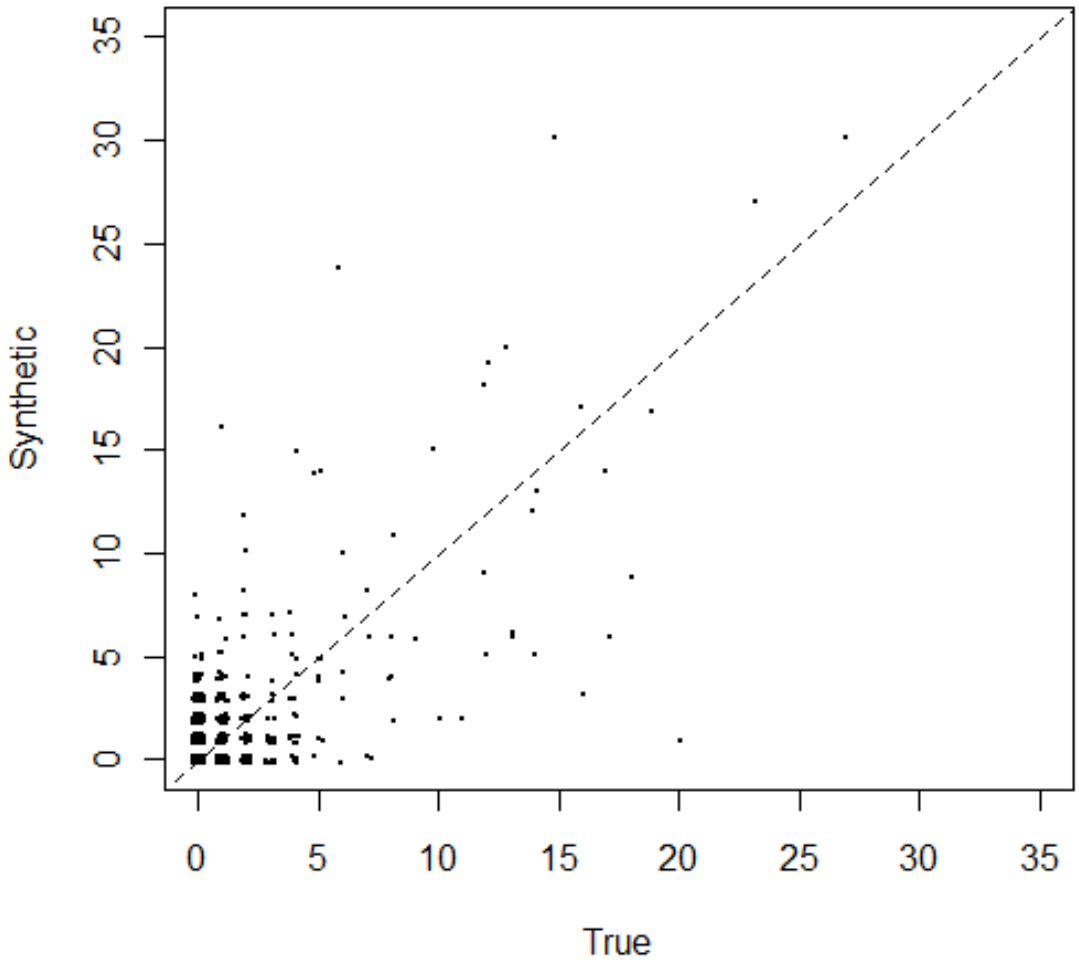

Figure 5-2. Continued

C

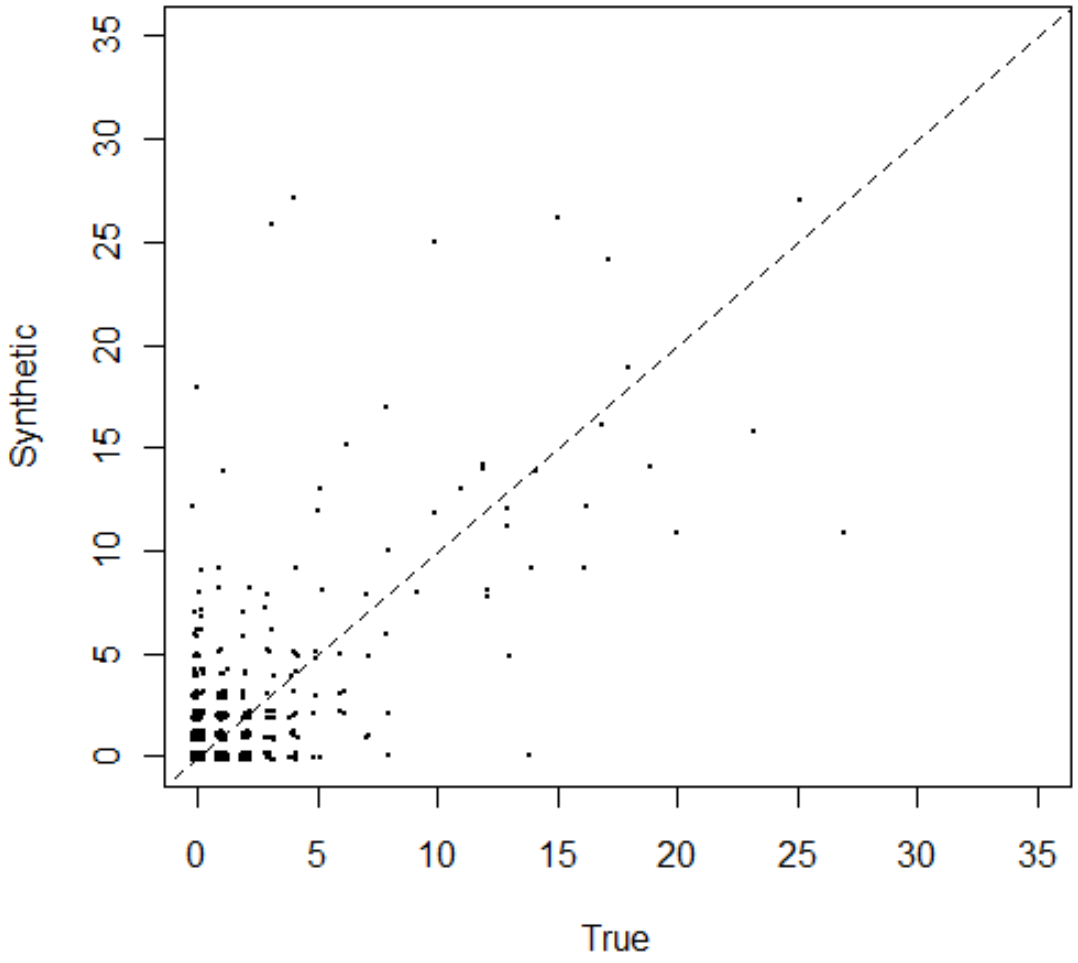

Figure 5-2. Continued

CHAPTER 6
TARGET-YEAR POPULATION SYNTHESIS: APPLICATION AND VALIDATION

This study contributes by presenting an empirical assessment of target year populations synthesized with different base-year populations, data-fusion methods, and control tables. Twelve synthetic populations were synthesized for 12 census tracts in Florida. The empirical results indicate the value of synthesizing more accurate base-year populations by accommodating multi-level controls. The impact of the data fusion methodology applied in the target year context is more modest possibly because there are fewer control tables available in the target year. Finally, errors in the target year control tables significantly reduce the accuracy of the synthesized populations. The magnitude of the overall error in the synthesized population appears to be linearly related to the magnitude of the input errors introduced via the control tables. Overall, efforts to accurately synthesize base-year populations and obtain target-year controls can help synthesize good target-year populations.

Unlike the case of base-year population synthesis, the documentation of results on target-year population synthesis is limited (Bowman 2004 and Bowman and Rousseau 2008). Although, conceptually, the application of the data-fusion approach for target-year synthesis is similar to its application for base year synthesis, there are three important issues of concern. First, the target-year synthesis uses the base-year synthesized population as seed data. Thus, the methodology and controls used in the base-year synthesis impact the accuracy of the base-year population, and in turn, the target-year population. Second, one can expect significantly fewer control tables available for the target year synthesis as opposed to the base year synthesis. In this situation, there might be benefits to using approaches that control for both person- and

household- level information as opposed to methods that control for only household-level information so as to take advantage of all the minimal data available. Third, the target year control tables are projections in contrast to base year control tables which are derived from the census counts. It has been well documented (Smith and Shahidullah, 1995 and Stoto, 1983) that there are significant errors in these projected aggregate distributions of population characteristics. In addition these errors are inherent and cannot be avoided. Therefore, examining the effects of errors in control tables is of interest.

In light of the above discussions, the intent of this paper is to contribute to our understanding of target-year population synthesis by addressing the following questions: (1) what is the effect of the accuracy of the base-year population (which will serve as the seed data for target-year synthesis) on the accuracy of the target-year population?, (2) What is the value of controlling both household- and person- level information in the target year versus only household-level controls? (3) How do errors in the projections of target year controls affect the accuracy of the population synthesized?

Analysis Framework

A total of twelve synthetic populations were generated for a target year for each of several census tracts to address the three fundamental research questions of this study.

As already discussed, the synthesis of target-year population begins with the synthesis of base-year populations. In this case, three different base-year populations were generated for each census tract with varying number of controls and differing in data-fusion methods. The first base year population was generated using only household-level controls and IPF as the data fusion methodology (this population is referred to as B-IPF in the rest of this document). The other two base-year populations

were synthesized using the FBS approach with both household- and person- level controls. These are referred to as B-FBS1 and B-FBS2 with the latter having more controls than the former. Thus, given the differences in the number of controlled attributes, one may expect the following order for the accuracy of the synthesized base-year population (B-FBS2 > B-FBS1 > B-IPF).

The second research question relates to the target-year controls. To address this, the target year populations were synthesized using two different approaches: IPF with only household-level controls and FBS with both household- and person- level controls. Each of the three base-year populations were used with each of the two target-year data fusion methods giving a total of six target year populations. These are referred to as T-IPF-B-IPF (i.e., target year IPF and base year IPF), T-IPF-B-FBS1, T-IPF-B-FBS2, T-FBS-B-IPF, T-FBS-B-FBS1, and T-FBS-B-FBS2 reflecting the base-year population used as seed data and the target-year synthesis methodology.

The six populations described above were synthesized using the true tract-level control tables. In order to assess the impact of erroneous target year controls, an approximate control table was generated for each target-year controlled attribute by replacing the true distribution with the distribution of the same attribute in the county to which the tract belongs. Six additional populations were synthesized using these approximate control tables and there are referred to as T*-IPF-B-IPF, T*-IPF-B-FBS1, T*-IPF-B-FBS2, T*-FBS-B-IPF, T*-FBS-B-FBS1, and T*-FBS-B-FBS2. The “*” is used to indicate that approximate controls were used for the target year.

Once the populations were synthesized, they were compared in terms of their ability to accurately replicate several marginal tables available for the target year. The error specific to marginal table j, $D_j$is calculated as follows:

$$D_j = \frac{\sum_{k=1}^{K_j} |T_{jk} - S_{jk}|}{\sum_{k=1}^{K_j} T_{jk}} \qquad (6-1)$$

T_{jk} is the true value of the kth category in table j and S_{jk} is the corresponding value of synthesized marginal table. The synthesized marginal tables were obtained by simply aggregating the synthesized population along the appropriate dimensions. The error measure can be interpreted as the proportion of the synthesized households/persons misclassified in the cells of the corresponding marginal table.

Dataset

Data from 12 census tracts and their corresponding PUMAs and Counties from Florida were used for this analysis. These census tracts and the PUMAs and counties to which they belong are identified in Table 6-1. Data were collected for years 1990 and 2000. The reader will note that the there are wide variations in the populations and the changes in population between the years. Further, these census tracts were chosen to represent some of the major urban regions in Florida where advanced travel-demand models are likely to be needed or developed. Finally, for all these census tracts, the boundaries did not change between 1990 and 2000. The year 2000 was used as the base year in this analysis and the year 1990 was set as the target year. Thus, we adopt a back-casting approach as opposed to a forecasting approach. The primary reason for this was that the PUMA-level data required for base-year population synthesis were available for 2000 and not 1990.

Table 6-2 identifies eleven base-year (2000) control tables (eight two-dimensional tables and three one-dimensional tables) used in this study. These distribution tables were obtained from the US census SF1 and SF3 files. These controls cover most of the important socio-economic-mobility attributes commonly used in travel modeling. Each of these tables corresponds to the joint distribution of a subset of the required population attributes. All these 11 tables were controlled for in the synthesis of the B-FBS2 (base-year) population. For the synthesis of the other two base-year populations, only a subset of these controls were used with only household-level controls being used for the B-IPF population. Table 6-2 also identified the controls used for the synthesis of each of the base-year populations. The seed data of base-year synthesis come from US census 5% PUMS (Public Use Microdata Sample).

For the target-year synthesis, distributions of household size and dwelling-unit type were used as controls for the IPF procedure. Person level controls for age and gender were used in addition to the two household level controls in the FBS procedure. The true tract-level tables were obtained from the US Census SF1 tables of 1990. The approximate control tables were obtained from the counties of the respective census tracts from the US census data of 1990. In generating the approximate control tables for the target year, we assume that the total population (persons and households) is still accurately known at the tract-level. Only the distribution is borrowed from the county level.

In addition to the controls for the target year, several other marginal tables are also available which are used to assess the accuracy of the synthesized populations.

Results

Table 6-3 gives the number of iterations for base-year population synthesis by the method FBS2. In this analysis, the number of adding iterations are roughly twice of number of remove iterations and they are linearly related to the size of generated population.

For each of the 12 synthetic populations, and for each of the twelve census tracts analyzed, error measures were calculated for each of the marginal tables. These are identified in Figure 6-1. Further this figure also indicates whether any or all of the attributes of the different marginal tables were controlled for in synthesizing either the base-year or the target-year populations.

The errors are then averaged across the 12 census tracts for each of the 12 synthetic populations and each of the marginal tables. These results are summarized in Table 6-4. In the rest of this section, these results are systematically analyzed to address the three key questions of interest.

Impact of Accuracy of the Base-Year Population

Table 6-5 compares the three base-year synthetic populations pair-wise using the error measures as defined before. Broadly, the numbers indicate that the three base-year populations are significantly different in terms of accurately replicating various base-year control tables.

Further, the errors of B-IPF relative to B-FBS2 are greater than the errors of B-FBS relative to B-FBS2 reflecting the effects of increased controls in base-year synthesis. The B-FBS2 population was synthesized controlling for all the tables mentioned in Table 6-5 and hence can be expected to replicate the true distributions of all these tables with great accuracy.

Figure 6-2 includes two sets of graphs which compare the accuracy of the populations synthesized with different base-year populations but with the same target year controls and data fusion methodology. The top graph is for the cases when the target year synthesis was undertaken using the IPF procedure and the bottom graph is for the target-year synthesis with FBS methodology. All these are for the case when true tract level controls were used (similar trends were observed for the case of approximated controls and hence these are not presented graphically here). The reader will note that the graphs derive their values from Table 6-4.

In general, we observe that the errors are least for populations synthesized using B-FBS2 as the seed data (see the circles) and are maximum for the populations synthesized using B-IPF as the seed data (see the triangles). The differences are particularly striking for marginal tables such as P52 which has attributes that are not controlled for in the target year. This indicates that if the base-year populations are synthesized controlling for as many attributes as possible, then the corresponding target year populations are also more accurate irrespective of the target-year controls/data fusion methodology employed.

Impact of Target-Year Control Tables and Methods

Figure 6-3 includes three sets of graphs which compare the accuracy of the populations synthesized with the same base-year populations but with different target year controls and data fusion methodology. Each graph compares the population synthesized with both household and person controls and using the FBS methodology against the population synthesized with only household controls using the IPF methodology. The top graph is for the cases when the base year synthesis was undertaken using the IPF procedure, the middle graph is when the base year population

is B-FBS and the bottom graph represents the base year population of B-FBS2. All these are for the case when true tract level controls were used (similar trends were observed for the case of approximated controls and hence these are not presented graphically here). Again, the graphs derive their values from Table 6-4.

For each fixed base-year population, the two target year populations perform similar in the context of accuracy with the FBS approach providing slightly better accuracy. This relatively low magnitude of improvement is as expected as the FBS essentially controls for only age and gender over and above the IPF target-year controls. Further, with the gender being practically equally distributed, the real difference between the two methods is the control for age in the FBS approach. Consistent with this discussion, the reader will note significant differences in the error for control table P12 which is the two dimensional joint table between gender and age. Since gender and age are controlled in the FBS method but not in IPF during target-year synthesis, the population under method “FBS” performs systematically better than “IPF” for these attributes.

Impact of Inaccurate Control Tables

The final research question examines the effect of the inaccuracies in the control tables on the accuracy of the synthetic populations. As shown in Table 6-4, the errors increase significantly when the approximate, county level distributions are used as controls instead of the true controls. This holds irrespective of the base-year synthesis methodology and the target-year synthesis methods. Therefore, it is important to be cognizant of the errors in the target-year controls despite using multi-level population synthesis methods as well as more-accurate base-year synthetic population as seed data.

It is also of interest to assess how the error in the control tables translates into errors in the synthetic populations. Table 6-6 presents the errors between the true and approximate (i.e., county level) control tables for the twelve census tracts. These errors are calculated using procedures previously described. The table also presents the average of these errors across the different control tables. Specifically, Average1 is calculated across all four control tables and, hence, it may be interpreted as the “input” error (or discrepancy) introduced in populations employing the FBS for target year synthesis. Average2 is calculated across all two household-level control tables and, hence, it may be interpreted as the input error (or discrepancy) introduced in populations employing the IPF for target year synthesis.

Figure 6-4 plots the input error (discrepancy) against the loss of accuracy for each census tracts and for each of the six types of synthetic populations. The loss of accuracy is calculated as follows. First, for each marginal table, the difference in errors between the population synthesized with the true controls and the one synthesized with the approximate controls is calculated (for each base year population and target year synthesis approach). This difference is error averaged across all marginal tables and is defined as the loss of accuracy for the census tract. In general, the loss of accuracy is greater with greater input errors and this relationship appears to be linear.

Summary

The application of disaggregate models for predictions and policy evaluations requires as inputs detailed information on the socio-economic characteristics of the future-year population. Although the IPF-based procedure is most popularly used, this is limited by the need to restrict all controls to the same universe. More recently, new methods have been developed to incorporate multi-level controls in population

synthesis. However, there is limited documentation of the application of IPF and other methods in the context of target-year synthesis. This study contributes by presenting an empirical assessment of target year populations synthesized with different base-year populations, data-fusion methods, and control tables. Twelve synthetic populations were synthesized for 12 census tracts in Florida. The year 2000 was taken as the base year and the 1990 as the target year.

The empirical results indicate the value of synthesizing more accurate base-year populations by accommodating multi-level controls. Target year populations synthesized with more accurate base-year populations as seed data are shown to be more accurate. The impact of the data fusion methodology applied in the target year context is more modest. This is because there are few control tables available in the target year and, hence, there might not be significantly more “information” contained in person-level controls beyond those in household-level controls. Nonetheless, the populations synthesized (target year) with multi-level controls and the FBS methodology do perform better than those synthesized with only household level controls and IPF. Finally, errors in the target year control tables significantly reduce the accuracy of the synthesized populations. The magnitude of the overall error in the synthesized population appears to be linearly related to the magnitude of the input errors introduced via the control tables.

Table 6-1. Characteristics of the twelve census tracts in 1990 and 2000

Case ID	Census Tract ID	PUMA ID	County Name	Households			Population			Group Quarters Population		
				2000	1990	% Change	2000	1990	% Change	2000	1990	% Change
1	0012	701	Leon	474	491	3.59	1030	1094	6.21	0	0	NA
2	0273.09	2601	Pinellas	643	240	-62.67	1606	617	-61.58	55	11	-80.00
3	0215.03	2003	Seminole	593	556	-6.24	1630	1561	-4.23	130	112	-13.85
4	0202	300	Okaloosa	711	612	-13.92	1799	1592	-11.51	0	0	NA
5	0101.24	4016	Miami-Dade	581	429	-26.16	2257	1290	-42.84	87	0	-100.00
6	0142.02	1104	Duval	1992	1797	-9.79	3770	3683	-2.31	30	0	-100.00
7	0016	3502	Palm Beach	1606	1515	-5.67	3875	3423	-11.66	0	34	NA
8	0219.02	2001	Seminole	1862	1857	-0.27	4513	4469	-0.97	14	25	78.57
9	0019.06	3502	Palm Beach	4170	2274	-45.47	7728	4260	-44.88	342	0	-100.00
10	0168.02	1106	Duval	3529	2203	-37.57	8145	5409	-33.59	0	0	NA
11	9801	600	Jefferson	3128	2747	-12.18	8894	7634	-14.17	1034	205	-80.17
12	0054.02	4011	Miami-Dade	3720	3572	-3.98	9426	8855	-6.06	12	0	-100.00

Table 6-2. Control tables for base-year population synthesis

Control Tables	Controlled In	Universe	Dimension 1		Dimension 2	
			Attribute	Categories	Attribute	Categories
H15	B-IPF, B-FBS, B-FBS2	Households	TENURE	Own, Rent	HHSIZE	1,2,3,4,5,6,7+
H32	B-FBS2	Households	TENURE	Own, Rent	DUTYPE	Single Family, Multi-Family
H44	B-IPF, B-FBS, B-FBS2	Households	TENURE	Own, Rent	NUMAUTO	0,1,2,3,4,5+
P26	B-FBS2	Households	HHSTRUCT	Family, Non-Family	HHSIZE	1,2,3,4,5,6,7+
P34	B-FBS2	Families	HHSTRUCT	Married couple, Other family	CHAGE[a]	None, Only <6 years, Only >=6 years, Both <6 years and >= 6 years
P48	B-FBS2	Families	HHSTRUCT	Married couple, Other family	NUMWORK[b]	0,1,2, 3+
P52	B-FBS2	Households	INCOME	< 30K, 30-50K, 50-75K, 75-125K, more than 125K		NA
P7	B-FBS, B-FBS2	Total Population	ETHNICITY	White, Black, Other, and Multiple Race		NA
P12	B-FBS, B-FBS2	Total Population	GENDER	Male, Female	AGE	0-5, 6-15, 16-17, 18-24, 25-34, 35-44, 45-54, 55-64, 65-74, over 75
P21	B-FBS2	Total Population	CITIZEN	Native, Naturalized, Non Citizen		NA
P47	B-FBS2	Population >=16 years	GENDER	Male, Female	WRKHOURS[c]	0,1-14, 15-35, more than 35

a Age distribution of "own children" in the household
b Number of workers (more than 0 hours per week in 1999)
c Hours of work per week in 1999

Table 6-3. Number of iterations for generating population of 12 census tracts by method FBS2 in base year (2000)

Case ID	Number of generated household	Number of adding iterations	Number of remove iterations	Number of total iterations
1	498	1180	682	1862
2	700	1350	650	2000
3	721	1280	559	1839
4	729	1403	674	2077
5	682	1465	783	2248
6	2036	3925	1889	5814
7	1629	2836	1207	4043
8	1909	3496	1587	5083
9	4517	9349	4832	14181
10	2515	5049	2534	7583
11	4185	6363	2178	8541
12	3747	7002	3255	10257

Table 6-4. Accuracy of target-year synthetic populations

	H15	P26	P12	H32	H44	P52	P21	P47
T-FBS-B-FBS2	0.14	0.05	0.09	0.09	0.29	0.17	0.12	0.16
T-FBS-B-FBS	0.16	0.06	0.11	0.15	0.29	0.25	0.16	0.13
T-IPF-B-FBS2	0.14	0.05	0.21	0.10	0.27	0.17	0.11	0.16
T-IPF-B-FBS	0.18	0.04	0.22	0.17	0.30	0.29	0.15	0.14
T-IPF-B-IPF	0.19	0.05	0.25	0.19	0.33	0.36	0.15	0.17
T-FBS-B-IPF	0.18	0.06	0.10	0.18	0.30	0.32	0.15	0.15
T*-FBS-B-FBS2	0.34	0.22	0.28	0.41	0.41	0.20	0.13	0.25
T*-FBS-B-FBS	0.27	0.23	0.26	0.41	0.32	0.32	0.14	0.20
T*-IPF-B-FBS2	0.33	0.24	0.28	0.40	0.37	0.22	0.17	0.21
T*-IPF-B-FBS	0.26	0.24	0.26	0.42	0.31	0.33	0.18	0.16
T*-IPF-B-IPF	0.28	0.24	0.29	0.43	0.34	0.40	0.16	0.20
T*-FBS-B-IPF	0.26	0.23	0.26	0.41	0.31	0.40	0.16	0.22

Table 6-5. Differences among synthesized base-year populations

	H15	P26	P34	P7	P12	H32	H44	P48	P52	P21	P47
Between IPF and FBS	0.03	0.05	0.15	0.26	0.17	0.05	0.02	0.14	0.09	0.04	0.10
Between IPF and FBS2	0.03	0.05	0.19	0.26	0.17	0.20	0.02	0.23	0.28	0.09	0.18
Between FBS and FBS2	0.02	0.05	0.14	0.00	0.01	0.22	0.02	0.20	0.24	0.07	0.13

Table 6-6. Difference between true controlled tables and erroneous tables

Case ID	Household Size	Dwelling type	Age	gender	Average1[a]	Average2[b]
1	0.27	0.09	0.20	0.01	0.14	0.18
2	0.45	0.61	0.41	0.04	0.38	0.53
3	0.07	0.22	0.19	0.06	0.14	0.15
4	0.07	0.35	0.24	0.03	0.17	0.21
5	0.25	0.77	0.18	0.05	0.31	0.51
6	0.31	0.69	0.17	0.05	0.31	0.50
7	0.25	0.13	0.19	0.04	0.15	0.19
8	0.23	0.18	0.14	0.04	0.15	0.21
9	0.26	0.71	0.53	0.03	0.38	0.49
10	0.20	0.25	0.21	0.02	0.17	0.23
11	0.04	0.04	0.04	0.01	0.03	0.04
12	0.17	0.68	0.29	0.03	0.29	0.43

a Average1 is the average of all four attributes listed in the table
b Average2 is the average of “household size” and “dwelling type”

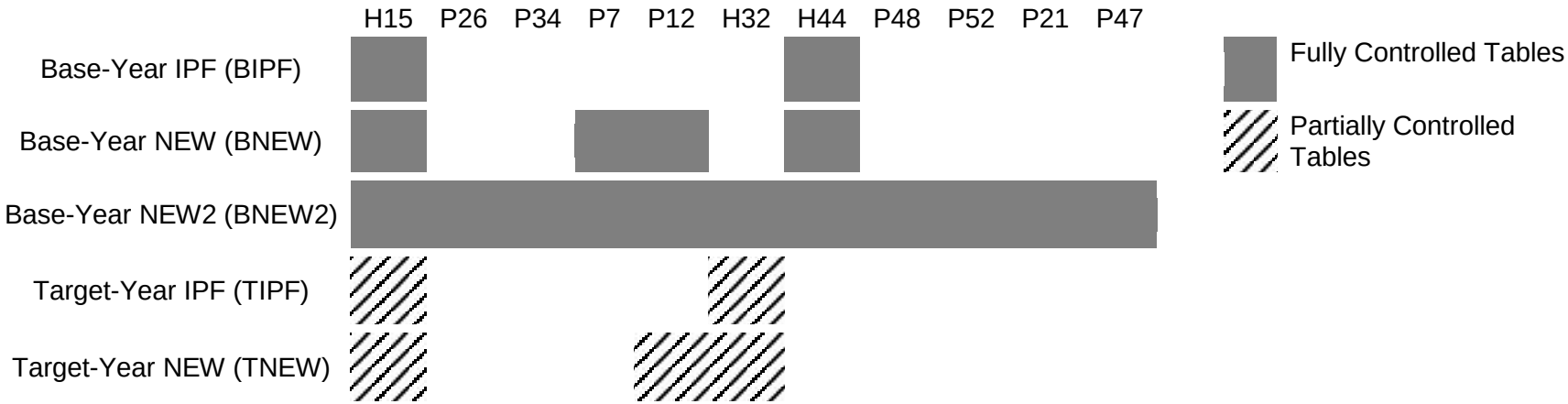

Figure 6-1. Marginal tables for assessing the target-year populations

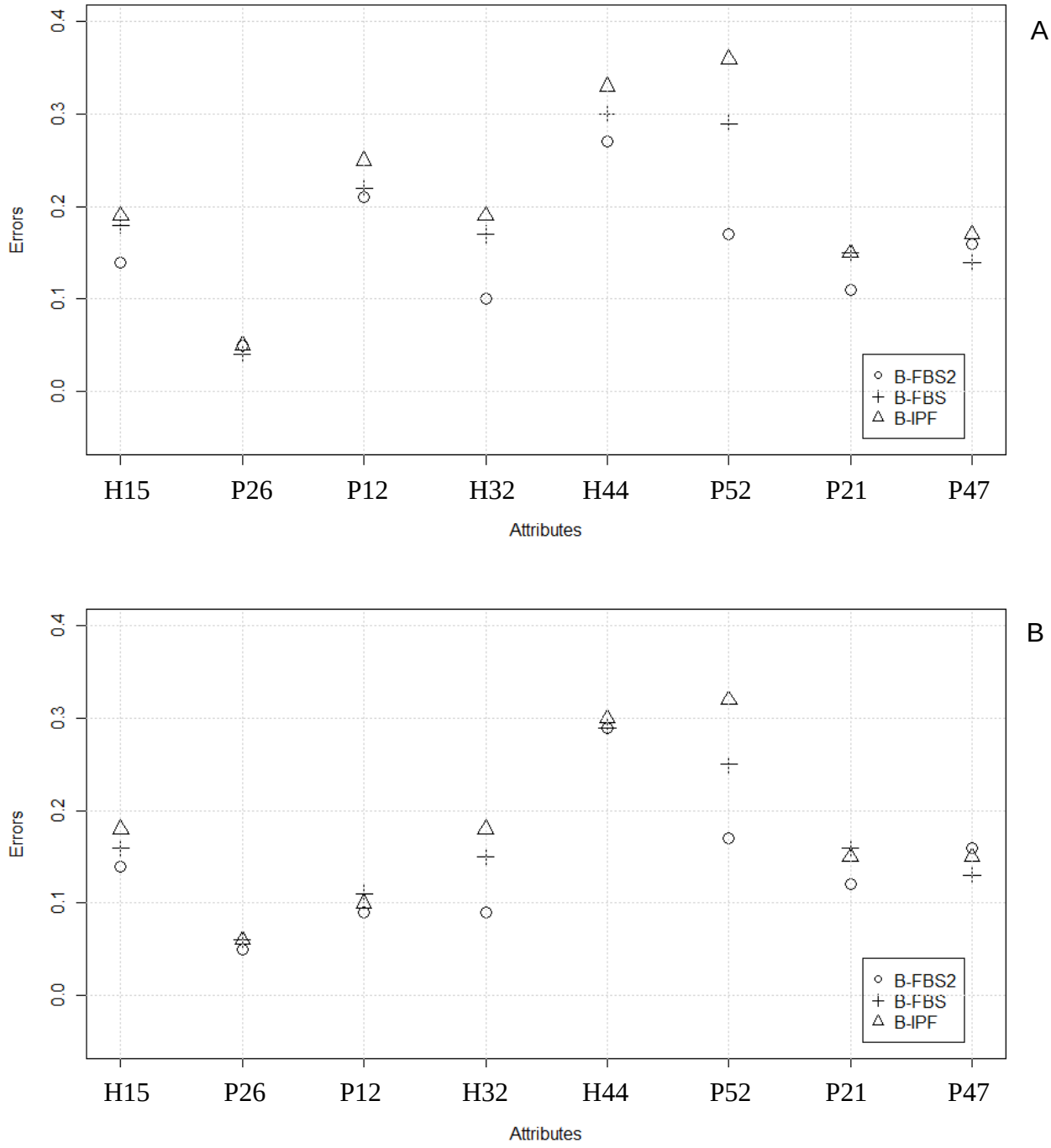

Figure 6-2. Impact of base-year populations on the accuracy of target-year population. A) Target-Year IPF. B)Target-Year FBS.

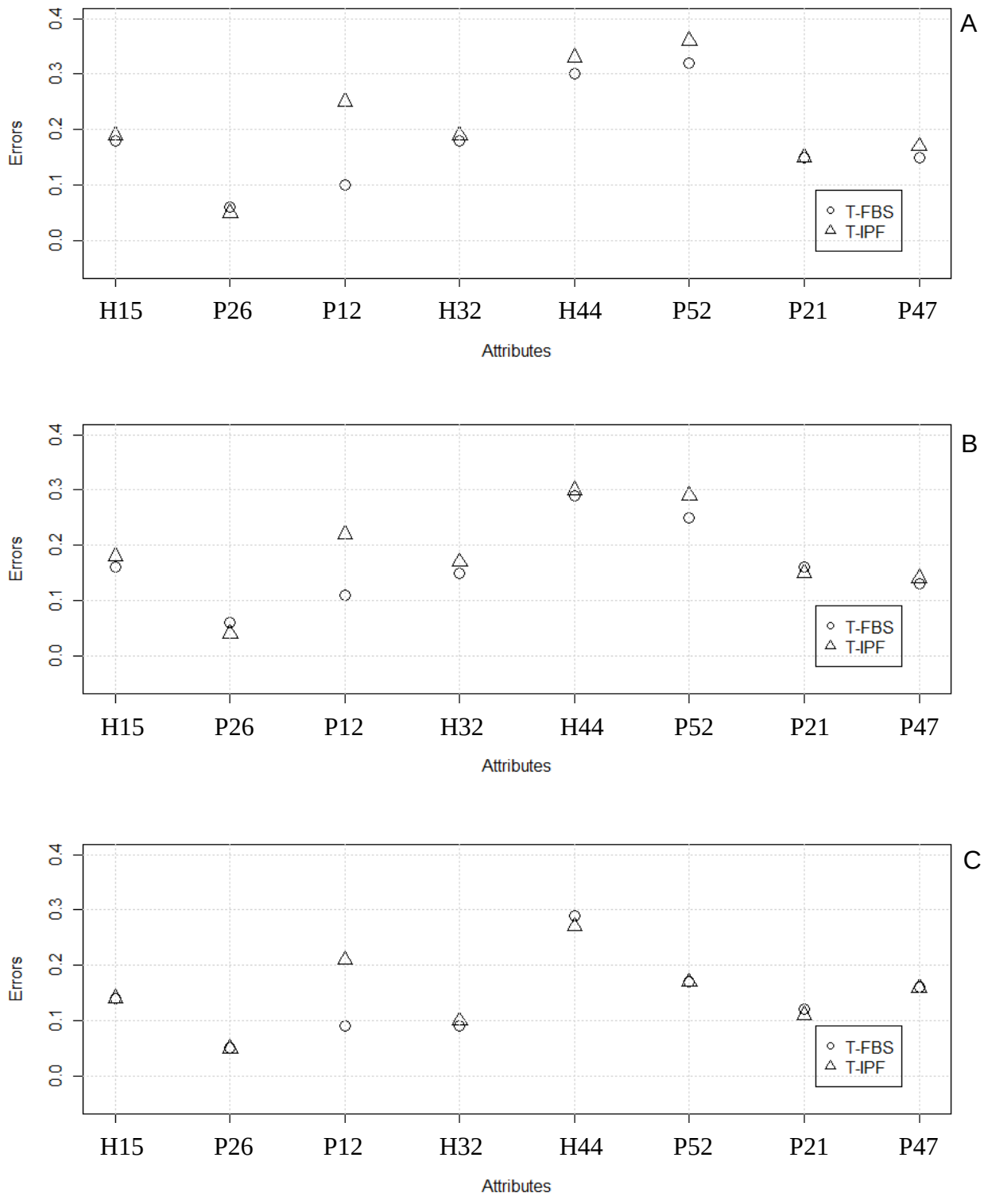

Figure 6-3. Impact of target-year controls and data-fusion methodology on the accuracy of target-year population. A) Base-Year IPF. B) Base-Year FBS. C) Base-Year FBS2.

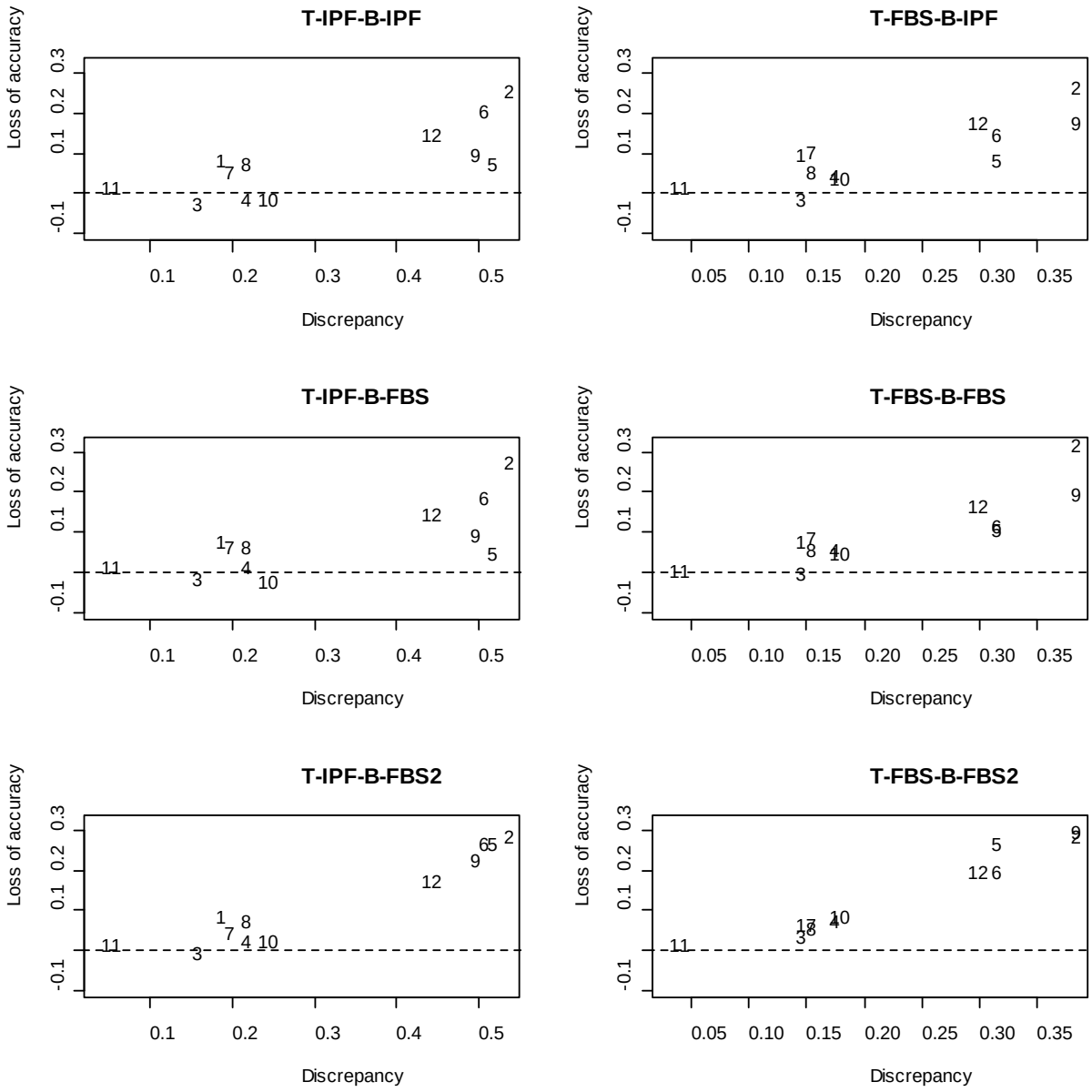

Figure 6-4. Impact of inaccurate control tables on the change of accuracy of target-year populations

CHAPTER 7
ASSESSMENT OF TRAVEL-DEMAND MODELS APPLIED TO SYNTHETIC POLULATIONS

Disaggregate models capture travel behavior of the fundamental decision-making units and include several explanatory variables (including socio-economic and mobility characteristics). Consequently, one may expect such models to provide more accurate predictions of the travel characteristics than aggregate models which include fewer explanatory variables. At the same time, for use in forecasting, disaggregate models require more inputs compared to aggregate models (as the number for explanatory variables are more in the former models). To generate such detailed inputs, population-synthesis procedures such as the one presented in this report are used. Thus, the accuracy of the socio-economic-mobility characteristics of the synthesized population (i.e., the inputs to the disaggregate models) is of particular interest. Specifically, if the synthesized population is an inaccurate representation of the true population, gains because of a disaggregate model could be offset by the errors in the synthesized population. In light of the above discussion, the intent of this chapter is to compare the predictions from aggregate trip-generation models with those from disaggregate trip-generation models. Further, for each of the two types of models, the predictions when the model is applied to the true population are compared to the predictions when the model is applied to a synthesized population. The scope of this comparative analysis is limited to linear-regression-based trip-generation models. The extension of such an analysis for the evaluation of non-linear models such as the multinomial-Logit for mode choice is identified as an important future avenue for study.

Dataset

Data from the NHTS (National Household Travel Survey) 2009 Add-Ons of Florida are used to estimate the aggregate and disaggregate trip-generation models used in this analysis. Further, pre-processing of the data was performed to eliminate those households for which we did not have the travel records for all persons in the household. Additional cleaning was also performed to remove cases with missing values for the explanatory variables of interest.

The final estimation sample comprised 12,577 households. Using the above dataset, linear-regression models are estimated for the trip purpose of Home-Based Non-Work Non-School and Non-Religion (HBNWSR). Frequency distributions of the household HBNWSR trip rates are presented in Table 7-1. This table is interpreted as follows: 2058 out of 12,577 or 16.36% of all households in the estimation sample did not have any HBNWSR trips (see first row in Table 7-1).

Trip-Generation Models

In this study, two trip generation models are estimated. One of the models adopts household size, vehicle ownership, and household tenure type as explanatory variables while another one adopts four additional explanatory variables, e.g. number of workers, urban/rural location, household income and life-cycle characteristics. The number of explanatory variables used in the latter models is greater than those used in the former. Thus, the former model is treated as the “aggregate” model and the latter model is the “disaggregate” model.

The trip generation models relate the total number of HBNWSR trips made by a household to the socio-economic characteristics of the household. For aggregate model, household size, vehicle ownership, and household tenure type were taken as

the explanatory variables as these are the most common variables used in trip generation models. The estimation of aggregate model is presented in Table 7-2. The disaggregate model includes four additional explanatory variables and its estimation is presented in Table 7-3.

As a general note on the models estimated, the linear-regression was chosen considering its simplicity and popularity in practical use. We recognize that this is not the best econometric structure to model integer variables within a finite range (i.e., the number of trips in our context). An examination of the application of methods such as the Poisson-regression and ordered-probit is identified as an avenue for future research.

Based on the estimation results of the two models, the adjusted rho square of disaggregate model (0.2111) is apparently larger than the one of aggregate model (0.1796). Therefore, under this point the disaggregate model is superior and can give a more precise forecasting for HBNWSR trips.

Population Synthesis

One of the objectives of this research is to examine the predictions of a trip-generation model when applied to a true population with the predictions when the same model is applied to synthesized versions of the same population. The Miami-Dade sample (2540 persons from 1191 households) is extracted from the NHTS 2009 Add-Ons of Florida. This sample is treated as the true population as all the socio-economic characteristics are known for each person and household. The same population was also synthesized using the proposed FBS method. The control tables are generated by aggregating the household and population characteristics of the Miami-Dade sample of the NHTS 2009 Add-Ons of Florida (instead of using the US Census SF tables). Then a

10% random sample of the NHTS 2009 Add-ons of Florida is used as the seed data (instead of the PUMS data as described in Chapter 5). It is useful to note that the NHTS does not include institutionalized population and hence the population in group-quarters is not a part of this analysis.

Two populations were synthesized. The first population (called Population I) was synthesized using six control tables (Table 7-4). Four of these tables are two dimensional and the other two are one dimensional. Five of these are household-level tables (i.e., the universe is all households) and another one are person-level tables. Overall, the structures of these tables are largely consistent with those presented in Chapter 5. The second population (called Population II) is synthesized using just four one-dimensional control tables (household size, dwelling type, gender, and age) as opposite to Population I with more controls.

Assessment of Linear Regression Based Trip Generation Models

The aggregate and disaggregate models developed for household HBNWSR trip were presented in Table 7-2 and Table 7-3. These models are applied on three populations. One of these is the “true” population (2540 persons and 1191 households of the Miami-Dade sample and the other two are “synthesized” populations, e.g. Population I which have more control tables and Population II which have few control tables. This chapter presents the results of an analysis of the predictive accuracy of each of the models when applied to each of the populations. The results are examined by the total number of trips by different life-cycle categories and different income categories.

In this study, each of the aggregate and disaggregate models is applied to each of the three populations to predict the number of trips for each household. Then the

predicted trips are aggregated by different categories of life-cycle and household income characteristics.

The six predictions of travel volumes obtained (two models × three populations) are compared with the actual travel volumes as reported in the travel surveys.

Table 7-6 presents the total number of trips for each life-cycle categories of three populations. First, the observed number of trips of true population is 5068 which is closed to corresponding number of trips of the six predications (5014.8, 4980.7, 5149.6, 5145.0, 5131.9, and 5162.0). In fact it is usually difficult to assess synthetic populations by the total number of predicted trips (aggregated trips of all households), as presented different populations and different models can provide similar results. That is because of that the aggregation process will mediate all the errors and make the total trips accurate. So, in this study the total number of trips is further divided into different categories to observe more disaggregate errors.

In this study, the absolute difference of number of trips or average number of trips between predicated trips and true trips is adopted as a criterion of accuracy. Table 7-6 provides number of trips for each life-cycle category for Population I. By such specification, it is observed that evident errors occur in different life cycle categories for both aggregated model and disaggregated model. The overall errors indicate that the disaggregated model is slightly better. In Table 7-6, significant errors are also observed for both aggregated model and disaggregated model under Population II, but the aggregated model is more accurate in terms of overall errors. A superficial explanation is that even the disaggregate model is more accurate but with population with few controls, the compromised data accuracy can make the estimation worse. However,

several factors can lead to such results and detailed discussion will be present shortly. Table 7-6 also shows the application of the two models on true population. The disaggregate model is better than the aggregate as expected. The errors of disaggregate model is uniformly small than aggregate model.

Because that the life-cycle attribute is not controlled in both Population I and Population II, the number of households is different than the one of true population. So, it is necessary to exam the average number of trips under each life-cycle category. Table 7-7 provides the average number of trips of the six predications on the five life-cycle categories. For that comparison, we see that the disaggregate model is superior in most of categories under all three populations.

Now we try to examine the comparison on household income categories. Different than life-cycle attribute, household income is controlled in population I and not controlled in population II where as the life-cycle attribute is not controlled in these two synthetic populations. So, still the comparison is performed in terms of average number of trips. The corresponding result is present in Table 7-8.

Within each income category, evident errors occur for both aggregate model and disaggregate model across the three populations.

Even in the application on true population, disaggregate model is not uniformly better than aggregate model since in the last category of income, disaggregate model has 0.32 off from true average trips while the aggregate model only have 0.06 off from true average trips. To remember, the model is estimated by the whole Florida dataset while the population is from Miami-Dade, some discrepancy is expected between estimation sample and forecasting sample.

Another interesting evident error is in the first income category of Table 7-8 under the Population I. There is a 0.42 off between average true trips and average trips estimated by the disaggregate model under population I while such error (0.08) is very small for the true population application in Table 7-8 under true population. By inspect on the population, it is found that the household size and household location variables have major discrepancy. Table 7-9 presents the distribution of household size and household location of Population II and true population within the first income category. In true population, there are more households with size greater than two and more households live in urban area. In Table 7-3, household size and urban household location have positive impact on number of trips made by households. That is the reason that number of trips made by first income categories of Population I is underestimated.

Summary

Two linear regression based trip generation models are estimated by Florida 2009 NHTS Add Ons. The one with more explanatory variables is referred as disaggregate model while another one with less explanatory variables is referred as aggregate model. These trip generation models are applied on three populations. One of the populations is the observed sample of Miami-Dade County from Florida 2009 NHTS Add Ons and another two populations are synthetic populations of the same region.

With the comparison between the forecasting number of trips and observed number of trips (also referred as true trips), the performance of the two models and three populations are examined by the number of trips and average number of trips within different categories from life cycle and household income attributes. Results indicate that by the categorization of life-cycle attribute, disaggregate model provides

smaller errors on average trips made by most of categories across the three populations. By the categorization of household income, disaggregate model gives relatively large errors on average trips made by the first income category when applied on synthetic populations.

Conceptually, disaggregate models give a more precisely predication on trips made by each household and hence should provide more precise forecasting on average trips by different categories. However, several other factors affect such aggregate predications, especially when applied to synthetic populations. First, after aggregation of trips, predication errors of each household are neutralized and hence the advantage of disaggregate models is veiled. Second, when applied to synthetic populations, the disaggregate model could give less precise predications because of the inaccuracy distribution of some attributes which are the explanatory variables. Also, the adopted trip generation models may have discrepancies between the applied populations. And some errors are expected when the behavior of some group of households is different than such households in the data used to estimation the models.

Table 7-1. Frequency distribution of household HBNWSR trip rates

# Trips	Freq.	%	# Trips	Freq.	%	# Trips	Freq.	%
0	2058	16.36	5	365	2.90	10	424	3.37
1	671	5.34	6	1374	10.92	11	82	0.65
2	2693	21.41	7	228	1.81	12	268	2.13
3	530	4.21	8	915	7.28	13	47	0.37
4	2428	19.31	9	158	1.26	14+	336	2.67
Total	12577	100						

Table 7-2. Aggregate model

Explanatory Variables	Aggregate Model		
	Coefficient	Std. Error	t value
Intercept			
One	1.623	0.153	10.604
Household Size			
One[a]			
Two	1.947	0.086	22.590
Three	3.000	0.124	24.262
Four and more	5.033	0.121	41.462
Tenure			
Owned household[*]			
Rented household	-0.584	0.108	-5.421
Household vehicle ownership			
Zero[*]			
One	0.731	0.157	4.655
Two	0.698	0.168	4.148
Three and more	0.783	0.179	4.382
Number of cases	12577		
Rho square	0.1801		
Adjusted Rho square	0.1796		

a Reference categories

Table 7-3. Disaggregate model

Explanatory Variables	Disaggregate Model		
	Coefficient	Std. Error	t value
Intercept			
One	1.144	0.166	6.900
Household Size			
One[a]			
Two	1.796	0.088	20.435
Three	3.112	0.161	19.347
Four and more	5.414	0.199	27.259
Tenure			
Owned household*			
Rented household	-0.300	0.107	-2.792
Household vehicle ownership			
Zero*			
One	0.721	0.156	4.610
Two	0.809	0.171	4.720
Three and more	0.939	0.185	5.076
Employment Status			
Number of workers	-0.393	0.054	-7.277
Household Location			
Household in urban area*			
Household in rural area	-0.534	0.075	-7.148
Household Income			
<$25,000*			
$25,000-$44,999	0.410	0.090	4.539
$45,000-$64,999	0.599	0.102	5.902
$65,000-$99,999	0.763	0.103	7.373
>$100,000	1.072	0.111	9.680
Life Cycle Characteristics			
No children*			
Youngest child 0-5	-1.129	0.190	-5.935
Youngest child 6-15	1.033	0.173	5.974
Youngest child 16-21	0.420	0.197	2.134
Retired and no children	0.548	0.089	6.165
Number of cases		12577	
Rho square		0.2122	
Adjusted Rho square		0.2111	

a Reference categories

Table 7-4. Control tables for population I

	Universe	Dimension 1		Dimension 2	
		Attribute	Categories	Attribute	Categories
1	Households	TENURE	Own, Rent	HHSIZE	1,2,3,4+
2	Households	TENURE	Own, Rent	DUTYPE	Single Family, Multi-Family
3	Households	TENURE	Own, Rent	NUMAUTO	0,1,2,3+
4	Households	NWORKER	0,1,2,3+		NA
5	Households	HHINCOME	< 25K, 25-45K, 45-65K, 65-100K, more		NA
6	Person	AGE	0-5, 6-15, 16-17, 18-24, 25-34, 35-44, 45-54, 55-64, 65-74, over 75	GENDER	Male, Female

Table 7-5. Control tables for population II

	Universe	Dimension 1		Dimension 2	
		Attribute	Categories	Attribute	Categories
1	Households		NA	HHSIZE	1,2,3,4+
2	Households		NA	DUTYPE	Single Family, Multi-Family
3	Person	AGE	0-5, 6-15, 16-17, 18-24, 25-34, 35-44, 45-54, 55-64, 65-74, over 75		NA
4	Person		NA	GENDER	Male, Female

Table 7-6. Total number of trips for households with different life cycle characteristic

Life Cycle (Under Population I)	Frequency	Number of Trips			Difference between True and	
		True	Aggregate Model	Disaggregate Model	Aggregated Model	Disaggregated Model
No children	361.0	1129.0	1308.9	1146.4	179.9	17.4
Youngest child 0-5	109.0	733.0	717.9	603.7	15.1	129.3
Youngest child 6-15	121.0	982.0	766.4	881.2	215.6	100.8
Youngest child 16-21	51.0	324.0	315.8	321.1	8.2	2.9
Retired and no children	548.0	1900.0	1905.8	2028.3	5.8	128.3
Total	1190.0	5068.0	5014.8	4980.7	424.6	378.6
Life Cycle (Under Population II)	Frequency	Number of Trips			Difference between True and	
		True	Aggregate Model	Disaggregate Model	Aggregated Model	Disaggregated Model
No children	321.0	1129.0	1154.1	995.6	25.1	133.4
Youngest child 0-5	110.0	733.0	724.4	603.2	8.6	129.8
Youngest child 6-15	147.0	982.0	936.7	1069.7	45.3	87.7
Youngest child 16-21	47.0	324.0	297.3	305.2	26.7	18.8
Retired and no children	566.0	1900.0	2037.0	2171.3	137.0	271.3
Total	1191.0	5068.0	5149.6	5145.0	242.8	641.1
Life Cycle (Under true population)	Frequency	Number of Trips			Difference between True and	
		True	Aggregate Model	Disaggregate Model	Aggregated Model	Disaggregated Model
No children	363.0	1129.0	1243.4	1116.0	114.4	13.0
Youngest child 0-5	137.0	733.0	912.3	771.7	179.3	38.7
Youngest child 6-15	134.0	982.0	863.7	996.9	118.3	14.9
Youngest child 16-21	53.0	324.0	305.3	314.0	18.7	10.0
Retired and no children	504.0	1900.0	1807.2	1963.4	92.8	63.4
Total	1191.0	5068.0	5131.9	5162.0	523.6	140.0

Table 7-7. Average number of trips for households with different life cycle characteristic

Life Cycle (Under Population I)	Frequency	Average Number of Trips			Difference between True and	
		True	Aggregate Model	Disaggregate Model	Aggregated Model	Disaggregated Model
No children	361.0	3.11	3.63	3.18	0.52	0.07
Youngest child 0-5	109.0	5.35	6.59	5.54	1.24	0.19
Youngest child 6-15	121.0	7.33	6.33	7.28	0.99	0.05
Youngest child 16-21	51.0	6.11	6.19	6.30	0.08	0.18
Retired and no children	548.0	3.77	3.48	3.70	0.29	0.07
Life Cycle (Under Population II)	Frequency	Average Number of Trips			Difference between True and	
		True	Aggregate Model	Disaggregate Model	Aggregated Model	Disaggregated Model
No children	321.0	3.11	3.60	3.10	0.49	0.01
Youngest child 0-5	110.0	5.35	6.59	5.48	1.24	0.13
Youngest child 6-15	147.0	7.33	6.37	7.28	0.96	0.05
Youngest child 16-21	47.0	6.11	6.32	6.49	0.21	0.38
Retired and no children	566.0	3.77	3.60	3.84	0.17	0.07
Life Cycle (Under true population)	Frequency	Average Number of Trips			Difference between True and	
		True	Aggregate Model	Disaggregate Model	Aggregated Model	Disaggregated Model
No children	363.0	3.11	3.43	3.07	0.32	0.04
Youngest child 0-5	137.0	5.35	6.66	5.63	1.31	0.28
Youngest child 6-15	134.0	7.33	6.45	7.44	0.88	0.11
Youngest child 16-21	53.0	6.11	5.76	5.92	0.35	0.19
Retired and no children	504.0	3.77	3.59	3.90	0.18	0.13

Table 7-8. Average number of trips for households with different income characteristic

Household Income (Under Population I)	Frequency	Number of Trips			Difference between True and	
		True	Aggregate Model	Disaggregate Model	Aggregated Model	Disaggregated Model
<$25,000	349	3.16	3.07	2.74	0.09	0.42
$25,000-$44,999	209	4.35	4.21	4.33	0.14	0.01
$45,000-$64,999	169	4.01	4.50	4.50	0.49	0.48
$65,000-$99,999	221	5.29	4.87	5.00	0.43	0.29
>$100,000	242	4.98	5.07	5.18	0.09	0.20
Household Income (Under Population II)	Frequency	Number of Trips			Difference between True and	
		True	Aggregate Model	Disaggregate Model	Aggregated Model	Disaggregated Model
<$25,000	293	3.16	3.20	2.88	0.05	0.28
$25,000-$44,999	291	4.35	4.23	4.24	0.12	0.11
$45,000-$64,999	196	4.01	4.53	4.63	0.52	0.62
$65,000-$99,999	203	5.29	5.03	5.15	0.27	0.15
>$100,000	208	4.98	5.15	5.37	0.17	0.39
Household Income (Under true population)	Frequency	Number of Trips			Difference between True and	
		True	Aggregate Model	Disaggregate Model	Aggregated Model	Disaggregated Model
<$25,000	349	3.16	3.43	3.24	0.27	0.08
$25,000-$44,999	210	4.35	4.19	4.22	0.15	0.13
$45,000-$64,999	169	4.01	4.42	4.42	0.41	0.41
$65,000-$99,999	221	5.29	4.91	5.05	0.38	0.24
>$100,000	242	4.98	5.04	5.30	0.06	0.32

Table 7-9. Distribution of household size and location for population I and true population

Populations	Household Size				Location	
	1	2	3	4+	Urban	Rural
Population I	199	110	22	18	282	67
True Population	147	144	30	28	341	8

CHAPTER 8
SUMMARY AND CONCLUSIONS

As a growing interest in the development of disaggregate travel-demand models. This dissertation proposed a fitness-based procedure for generating synthetic populations which are the detailed information as inputs to disaggregate travel-demand models and activity-based transportation planning models. Around the proposed method all major steps toward synthetic populations are analyzed comprehensively. These steps include base-year population synthesis, target-year population synthesis and application on travel demand models.

The traditional population synthesis is based on IPF procedure which is most popularly used in current applications. However, IPF is limited by the need to restrict all controls to the same universe and issues like zero-cell problem also impede this method on complex situations.

More recently, new methods have been developed to incorporate multi-level controls in population synthesis. However, there is also limited documentation of the application of IPF and other methods in the context of target-year synthesis. This research is to contribute towards synthetic population generation by addressing these issues.

In Chapter 4, the Fitness Based Synthesis (FBS) approach is proposed. This method is designed to select households in a way that marginal control tables are matched. The proposed framework contains three major components, namely initial household sets, fitness functions and selection mechanism. This study introduces two types of fitness functions to measure the marginal add/remove contribution of each household for reducing matching errors. The initial household set is prepared before the

procedure and proper selection of initial household sets can reduce the running time of procedure. Fitness function measures the marginal contribution of each household for reducing matching errors. In this study, we choose fitness function based on the criterion of sum of squared error and it can be replaced by other criteria.

The proposed method is not a member of IPF-based population synthesizers and hence there is no zero-cell issue related and also no convergence issue under square error based fitness function (adopted in this dissertation) because the number of iterations is proved to be bounded. Furthermore, under square error based fitness functions, the synthetic population can match the marginal control tables almost perfectly if there are sufficient variable types of households.

In Chapter 5, synthetic populations from 22 artificial census tracts are generated based on the propose FBS method. As an important objective, this dissertation compares the proposed method with other method. So, the synthetic populations of the same 22 census tracts are also generated by the conventional IPF method and a recently developed IPU method. According to validation results, the FBS approach and IPU approach generate populations closer to true population than the one from conventional IPF method.

In Chapter 6, this study also contributes by presenting an empirical assessment of target year populations synthesized with different base-year populations, data-fusion methods, and control tables. Twelve synthetic populations were synthesized for twelve census tracts in Florida. The year 2000 was taken as the base year and the 1990 as the target year. The empirical results indicate the value of synthesizing more accurate base-year populations by accommodating multi-level controls. Target year populations

synthesized with more accurate base-year populations as seed data are shown to be more accurate. And, the populations synthesized (target year) with multi-level controls and the FBS methodology do perform better than those synthesized with only household level controls and IPF. Finally, errors in the target year control tables significantly reduce the accuracy of the synthesized populations. The magnitude of the overall error in the synthesized population appears to be linearly related to the magnitude of the input errors introduced via the control tables.

In Chapter 7, two linear regression based trip generation models are estimated by Florida 2009 NHTS Add Ons. The one with more explanatory variables is referred as disaggregate model while another one with less explanatory variables is referred as aggregate model. These trip generation models are applied on three populations. One of the populations is the sample of Miami-Dade County from Florida 2009 NHTS Add Ons and another two populations are synthetic populations of the same region.

Disaggregate models give a more precisely predication on trips made by each household and hence should provide more precisely forecasting on average trips by different categories. However, several other factors affect such aggregate predications, especially when applied to synthetic populations. First, after aggregation of trips, predication errors of each household are neutralized and hence the advantage of disaggregate models is veiled. Second, when applied to synthetic population, disaggregate model could give less precise predications because of the inaccuracy distribution of some attributes which are the explanatory variables. Also, the adopted trip generation models may have discrepancies between the applied populations. And

some errors are expected when the behavior of some group of households is different than such households in the data used to estimation the models.

There are several avenues for the future research. First, it is useful to test different fitness functions under the FBS mechanism. Second, the quality of synthetic population is affected by several aspects, e.g., seed data. So the analysis of impact of seed data or other aspects can also be conducted.

The analysis on the application of trip generation models can be extend to such an analysis for the evaluation of non-linear models such as the multinomial-Logit for mode choice. In addition, an examination of the application of methods such as the Poisson-regression and ordered-probit for household trips is also identified as an avenue for future research.

APPENDIX A
NUMERICAL ILLUSTRATION OF THE FITNESS BASED SYNTHESIS PROCEDURE

Chapter 4 outlines the procedure for population synthesis which involves selecting a set of households from the PUMS data in such a way that the tract-level controls are satisfied. A numerical illustration of this procedure is presented here.

For simplicity, we assume that there are two control tables (Figure A-1) for a hypothetical census tract (synthesis area). The first table (T_{1k}) is a two-dimensional household-level table joint distribution of household-size (household size is limited to being either 1 or 2 persons, again for simplicity) and tenure. The second table (T_{2k}) is a one-dimensional, person-level table representing the distribution of gender. The intent of the population-synthesis procedure is to generate households and persons that satisfy the distributions present in these control tables.

The seed data are presented in Figure A-2. There are two tables. The household-level table presents the tenure and household size of each household. The person-level table presents the gender of each person present in each of the households in the seed data. We see that there are five households and eight persons in this dataset. The reader will note that the seed-data has at least one household of each of the four types (i.e., Own, 1-person; Own, 2-person; Rent, 1-person; and Rent, 2-Person) represented in the control table. Similarly, there are both males and females in the seed data.

The population synthesis is an iterative procedure and one household is selected to be added or removed in each iteration. The selection of households is based on the two types of fitness functions. This fitness function is calculated for each household in the seed data and for each iteration based on the values of the control tables (T), the current selection and the contribution of the corresponding household towards satisfying

the controls. Figure A-3 presents the values in the two HT-tables for the five households in the seed data. As already defined, the HT-tables define the contribution of each household towards satisfying the different controls. Household 1 (HH ID = 1) comprised a single female living in a rental house (Figure A-2). Thus, “selecting” this household would contribute one single-person, rental household (as indicated in the first HT-table in Figure A-3) and one female (as indicated in the second HT-table) to the population. Similarly, selecting the third household from the seed data will contribute one two-person, rental household (as indicated in the first HT-table for the third household) and two males (as indicated in the second HT-table of the third household) to the population.

Once all the tables have been defined, the two fitness values can be calculated for each of the households in the seed data. The fitness of a household i in iteration n, F_I^{in} and F_{II}^{in} are calculated using the following formula.

$$F_I^{in} = \sum_{j=1}^{J} \sum_{k=1}^{K_j} \left[\left(R_{jk}^{n-1} \right)^2 - \left(R_{jk}^{n-1} - HT_{jk}^{i} \right)^2 \right] \qquad (A-1)$$

$$F_{II}^{in} = \sum_{j=1}^{J} \sum_{k=1}^{K_j} \left[\left(R_{jk}^{n-1} \right)^2 - \left(R_{jk}^{n-1} + HT_{jk}^{i} \right)^2 \right] \qquad (A-2)$$

where $R_{jk}^{n-1} = T_{jk} - CT_{jk}^{n-1}$

In the current example, $J = 2$ (there are two control tables) and $K_1 = 4$ (four cells in the first control table) and $K_2 = 2$ (two cells in the second control table).

In this example, the entire seed data is chosen as the initial household sets. The algorithm begins with the calculation of the fitness values for each of the five households. Then a household is selected among the candidate sets.

The fitness values are then recalculated and the iterations continue. A household from the seed data can be selected into the population of the census tract multiple times. The algorithm terminates when there is no household meet the requirement of candidate.

The numerical calculations corresponding to the application of the algorithm to the example problem is presented in table A-1. In the first iteration, all households have positive Type I fitness value and negative Type II fitness value. Hence the candidate sets consists all the five households. By simulation, the household 2 is selected and added into the current selection. After this household is added, the fitness values for all household are then suitably updates. The results for the following iterations are presented in table A-1. After eight iterations, it is found that all fitness values of the five household are negative and hence no household is eligible as a candidate and hence the procedure is finished.

To be noted, in iteration 7, only household 2 has a positive type II fitness value, and it means remove such household from current selection can improve the overall fitting with control tables.

After the procedure, the household counts in iteration 7 construct the synthetic population. After aggregate the synthetic into the structure of control tables it is found that the table by aggregating synthetic population is exactly the same as control tables (refer to Figure A-1 and Figure A-4).

Therefore, it indicates a perfect fit of the synthesized population to the controls. When the number of households is large and there are several control tables, such a perfect fit may not be possible.

Table A-1. Intermediate results during generating the example population

Household ID		1	2	3	4	5
Initial Counts		1	1	1	1	1
	Fitness I	8	8	17	21	21
	Fitness II	-12	-12	-27	-27	-31
Iteration1	Candidate set	X	X	X	X	X
	Selection		X			
	Counts	1	2	1	1	1
	Fitness I	8	4	13	19	17
	Fitness II	-12	-8	-23	-25	-27
Iteration2	Candidate set	X	X	X	X	X
	Selection				X	
	Counts	1	2	1	2	1
	Fitness I	6	2	9	13	11
	Fitness II	-10	-6	-19	-19	-21
Iteration3	Candidate set	X	X	X	X	X
	Selection	X				
	Counts	2	2	1	2	1
	Fitness I	2	2	9	11	11
	Fitness II	-6	-6	-19	-17	-21
Iteration4	Candidate set	X	X	X	X	X
	Selection				X	
	Counts	2	2	1	3	1
	Fitness I	0	0	5	5	5
	Fitness II	-4	-4	-15	-11	-15
Iteration5	Candidate set			X	X	X
	Selection			X		
	Counts	2	2	2	3	1
	Fitness I	0	-4	-5	1	-3
	Fitness II	-4	0	-5	-7	-7
Iteration6	Candidate set				X	
	Selection				X	
	Counts	2	2	2	4	1
	Fitness I	-2	-6	-9	-5	-9
	Fitness II	-2	2	-1	-1	-1
Iteration7	Candidate set		X			
	Selection		X			
	Counts	2	1	2	4	1
Iteration8	Fitness I	-2	-2	-5	-3	-5
	Fitness II	-2	-2	-5	-3	-5

T_{1k}	HHSize = 1	HHSize = 2	Total
Own	1	5	6
Rent	2	2	4
Total	3	7	10

T_{2k}	Total
Male	11
Female	6
Total	17

Figure A-1. Control tables

HH ID (*i*)	Tenure	HH Size
1	Rent	1
2	Own	1
3	Rent	2
4	Own	2
5	Own	2

HH ID (*i*)	Person ID	Gender
1	1	Female
2	1	Male
3	1	Male
3	2	Male
4	1	Male
4	2	Female
5	1	Male
5	2	Male

Figure A-2. Seed data

HH ID (i)= 1

HT^{i}_{1k}	HHSize = 1	HHSize = 2
Own	0	0
Rent	1	0

HT^{i}_{2k}	Total
Male	0
Female	1

HH ID (i)= 2

HT^{i}_{1k}	HHSize = 1	HHSize = 2
Own	1	0
Rent	0	0

HT^{i}_{2k}	Total
Male	1
Female	0

HH ID (i) = 3

HT^{i}_{1k}	HHSize = 1	HHSize = 2
Own	0	0
Rent	0	1

HT^{i}_{2k}	Total
Male	2
Female	0

HH ID (i) = 4

HT^{i}_{1k}	HHSize = 1	HHSize = 2
Own	0	1
Rent	0	0

HT^{i}_{2k}	Total
Male	1
Female	1

HH ID (i)= 5

HT^{i}_{1k}	HHSize = 1	HHSize = 2
Own	0	1
Rent	0	0

HT^{i}_{2k}	Total
Male	2
Female	0

Figure A-3. HT-tables for each of the households in the seed data

A_{1k}	HHSize = 1	HHSize = 2	Total
Own	1	5	6
Rent	2	2	4
Total	3	7	10

A_{2k}	Total
Male	11
Female	6
Total	17

Figure A-4. The representation of synthetic population as the structure of control tables

LIST OF REFERENCES

Agresti, A., 2002. Categorical Data Analysis, 2nd ed. Wiley, New York.

Arentze, T., Timmermans, H.J.P., Hofman, F., 2007. Creating synthetic household populations: problems and approach. Transportation Research Record 2014, 85-91.

Auld, J., Mohammadian, A., Wies, K., 2008. Population synthesis with region-level control variable aggregation. In: Paper presented at the 87th Transportation Research Board Annual Meeting, Washington, DC.

Auld, J., Mohammadian, A., 2010. Efficient methodology for generating synthetic populations with multiple control levels. Transportation Research Record 2175, 138-147.

Beckman, R.J., Baggerly, K.A., Mckay, M.D., 1996. Creating synthetic baseline populations. Transportation Research Part A 30 (6), 415-429.

Bhat, C.R., Koppelman, F.S., 1993. A conceptual framework of individual activity program generation. Transportation Research Part A 27 (6), 433-446.

Bhat, C.R., Guo, J.Y., Srinivasan, S., Sivakumar, A, 2004. Comprehensive econometric microsimulator for daily activity-travel patterns. Transportation Research Record 1894, 57-66.

Bowman, J.L., Bradley, M., 2006. Activity-based travel forecasting model for SACOG: population synthesis. Technical memo number 2, prepared for Sacramento Area Council of Governments, available from http://jbowman.net/ProjectDocuments/SacSim/SACOG%20tech%20memo%202--Pop%20Synth.20060731.pdf

Bowman, J.L., Bradley, M., 2007. Activity-based travel forecasting model for SACOG: household auto availability model. Technical memo number 9, prepared for Sacramento Area Council of Governments, available from http://jbowman.net/ProjectDocuments/SacSim/SACOG%20tech%20memo%209--Auto%20availability.20060914.pdf

Bowman, J.L., 2004. A comparison of population synthesizers used in microsimulation models of activity and travel demand. Draft paper available from http://jbowman.net/papers/

Bowman, J.L., Rousseau, G., 2008. Validation of Atlanta, Georgia, regional commission population synthesizer. Transportation Research Board Conference Proceedings 2 (42), 54-62.

Deming, W.E., Stephan, F.F., 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. The Annals of Mathematical Statistics 11 (4), 427-444.

Eluru, N., Pinjari, A., Guo, J.Y., Sener, I., Srinivasan, S., Copperman., R., Bhat., C.R., 2008. Population updating system structures and models embedded within the comprehensive econometric microsimulator for urban systems. Transportation Research Record 2076, 171-182.

Fienberg, S.E., 1970a. An iterative procedure for estimation in contingency tables. The Annals of Mathematical Statistics 41(3), 907-917.

Fienberg, S.E., 1970b. The analysis of multidimensional contingency tables. Ecology 51(3), 419-433.

Fienberg, S.E., 1972. The analysis of incomplete multi-way contingency tables. Biometrics 28(1), 177-202.

Frick, M., Axhausen, K.W., 2004. Generating synthetic populations using IPF and Monte Carlo techniques: some new results. In Paper presented at the 4th Swiss Transport Research Conference, Ascona.

Guo, J.Y., Bhat, C.R., 2007. Population synthesis for microsimulating travel behavior. Transportation Research Record 2014, 92-101.

Goulias, K. G., Kitamura. R., 1996. A dynamic model system for regional travel demand forecasting. In Panels for Transportation Planning: Methods and Applications, Eds. Golob, T., R. Kitamura, and L. Long, Kluwer Academic Publishers, Boston, Ch. 13, 321-348.

Hunt, J.D., Abraham, J.E., Weidner, T., 2004. The household allocation (HA) module of the Oregon2 model. Transportation Research Record 1898, 98-107.

Ireland, C.T., Kullback, S., 1968. Contingency tables with given marginals. Biometrika 55 (1), 179-188.

Karlis, D., Xekalaki, E., 1998. Minimum hellinger distance estimation for poisson mixtures. Computational Statistics & Data Analysis 29 (1), 81-103.

Lee, D., Fu, Y., 2011. A cross entropy optimization model for population synthesis used in activity-based micro-simulation models. In: Paper presented at the 90th Transportation Research Board Annual Meeting, Washington, DC.

Mackett, R. L., 1990. MASTER Mode. Report SR 237, Transport and Road Research Laboratory, Crowthorne, England.

Moeckel, R., Spiekermann, k., Wegener, M., 2003. Creating a synthetic population. In: Paper presented at the 8th International Conference on Computers in Urban Planning and Urban Management (CUPUM), Sendai, Japan.

Mohammadian, A., Javanmardi, M., Zhang, Y., 2010. Synthetic household travel survey data simulation. Transportation Research Part C 18 (6), 869-878.

Müller, k., Axhausen, K.W., 2011. Population synthesis for microsimulation: state of the art. In: Paper presented at the 90th Transportation Research Board Annual Meeting, Washington, DC.

Pritchard, D.R., Miller, E.J., 2009. Advances in agent population synthesis and application in an integrated land use and transportation model. In: Paper presented at the 88th Transportation Research Board Annual Meeting, Washington, DC.

Ruschendorf, L., 1995. Convergence of the iterative proportional fitting procedure. The Annals of Statistics 23(4), 1160-1174.

Ryan, J., Maoh, H., Kanaroglou, P., 2009. Population synthesis: comparing the major techniques using a small, complete Population of firms. Geographical Analysis 41 (2), 181-203.

Ryan, J., Maoh, H., Kanaroglou, P., 2010. Population synthesis for microsimulating urban residential mobility. In: Paper presented at the 89th Transportation Research Board Annual Meeting, Washington, DC.

Simpson, D.G., 1987. Minimum hellinger distance estimation for the analysis of count data. Journal of the American Statistical Association 82, 802-807.

Simpson, L., Tranmer, M., 2005. Combining sample and census data in small area stimates: iterative proportional fitting with standard software. The Professional Geographer 57(2), 222-234.

Smith, K.S., Shahidullah, M., 1995. An evaluation of population projection errors for census tracts. Journal of the American Statistical Association 90, 64-71.

Srinivasan, S., Ma, L., Yathindra, K., 2008. Procedure for forecasting household characteristics for Input to travel-demand models. Project Report of University of Florida, Gainesville; Florida Department of Transportation, available from http://www.fsutmsonline.net/images/uploads/reports/FDOT_BD545_79_rpt.pdf

Srinivasan, S., Ma, L., 2009. Synthetic population generation: a heuristic data-fitting approach and validations. In: Paper presented at the 12th International Conference on Travel Behaviour Research (IATBR), Jaipur, India.

Stoto, M.A., 1983. The accuracy of population projections. Journal of the American Statistical Association, 78, 13-20.

Sundararajan, A., Goulias, K. G., 2003. Demographic microsimulation with DEMOS 2000: design, validation, and forecasting. In Transportation Systems Planning: Methods and Applications, Eds. K.G. Goulias, CRC Press, Boca Raton, Ch. 14.

Voas, D., Williamson, P., 2000. An evaluation of the combinatorial optimization approach to the creation of synthetic microdata. International Journal of Population Geography 6 (5), 349-366.

Williamson, P., Birkin, M., Rees, P.H., 1998. The estimation of population microdata by using data from small area statistics and samples of anonymised records. Environment and Planning A 30(5), 785–816.

Wong, D.W.S., 1992. The reliability of using the iterative proportional fitting procedure. The Professional Geographer 44(3), 340-348.

Ye, X., Konduri, K., Pendyala, R.M., Sana, B., Waddell, P., 2009. Methodology to match distributions of both household and person attributes in generation of synthetic populations. In: Paper presented at the 88th Transportation Research Board Annual Meeting, Washington, DC.

Yvonne, M., Bishop, M., Fienberg, S.E., 1969. Incomplete two-dimensional contingency tables. Biometrics 25(1), 119-128.

Zhang, Y., Mohammadian, A., 2008. Microsimulation of household travel survey data. In: Paper presented at the 87th Transportation Research Board Annual Meeting, Washington,DC.

BIOGRAPHICAL SKETCH

Lu Ma received his bachelor's degree and master's degree in civil engineering at Tsinghua University, Beijing, China, in 2004 and 2007 respectively. He also earned another master's degree in statistics at the University of Florida in 2010. In August 2007, he began graduate school at the University of Florida in Department of Civil and Coastal Engineering for his Ph.D. study.

Lu Ma has published one journal paper and co-authored four papers for conferences and technique reports. During the Ph.D. study, he was awarded the Short Stay Fellowship from Utrecht University, the Netherlands and visited this university for a study of three month in 2010 and he was also awarded the ITS Florida Annual Anne Brewer Scholarship in 2010.