Transportation Research Part A 85 (2016) 247–264

Contents lists available at ScienceDirect

Transportation Research Part A

journal homepage: www.elsevier.com/locate/tra

An empirical assessment of factors affecting the accuracy of target-year synthetic populations

Lu Ma [a,*], Sivaramakrishnan Srinivasan [b,1]

[a] MOE Key Laboratory for Urban Transportation Complex Systems Theory and Technology, School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, PR China

[b] Dept. of Civil and Coastal Engineering, University of Florida, 513-A Weil Hall, PO Box 116580, Gainesville, FL 32611, United States

ARTICLE INFO

Article history:
Received 16 August 2014
Received in revised form 26 January 2016
Accepted 28 January 2016
Available online 13 February 2016

Keywords:
Synthesized population
Data-fusion method
Target-year
Fitness-based synthesis

ABSTRACT

This study contributes by presenting an empirical assessment of the accuracy of the target-year populations synthesized with different base-year populations, data-fusion methods, and control tables. Forty-five synthetic populations were generated for 12 census tracts in Florida for this purpose. The empirical results indicate the value of synthesizing base-year populations more accurately by accommodating multi-level controls. Although fewer controls are typically available for target years, the use of multi-level controls in the target year with appropriate synthesis methods does benefit the accuracy of the synthetic population. This study also establishes that 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. The improvements in accuracy are statistically significant and hold after controlling for differences in population sizes and growth rates for the different census tracts. Overall, efforts to accurately synthesize base-year populations and to good forecasts of target-year controls can help synthesize accurate target-year populations.

1. Introduction

The need for detailed socio-economic population characteristics as an important pre-requisite to the application of disaggregate, microsimulation-based travel-demand models such as the activity-based models has increased significantly in the recent years. The true characteristics of the population of an area is usually very time consuming, expensive and sometimes impossible (e.g., for future population) to obtain (Barthelemy and Toint, 2013). Therefore, a synthetic population, described in terms of several household and/or person attributes (socio-economic characteristics), is developed to serve as a proxy for the true population.

Once a synthetic population for a target year (generally a future year for which planning is being undertaken) is available, disaggregate – travel-demand – models can be applied to the artificial households and persons in this population to determine the travel patterns. The travel patterns of the individual persons and households can be suitably aggregated to determine the overall demand and system-performance. This approach improves forecast by addressing the issue of aggregation bias that current aggregate-models suffer from (i.e., the behavior of an average person in the population is not truly representative of the overall behavior of the population; see for example, Koppelman, 1974; Landau, 1978). This is because dis-

* Corresponding author. Tel.: +86 1051684599; fax: +86 1051840080.
E-mail addresses: lma@bjtu.edu.cn (L. Ma), siva@ce.ufl.edu (S. Srinivasan).
[1] Tel.: +1 352 392 9537x1456.

http://dx.doi.org/10.1016/j.tra.2016.01.016
0965-8564/

aggregate models can accurately represent the underlying activity-travel generation process by making the individual decision maker, their choices, and their decision-making processes the center of the modeling paradigm. Further, the approach allows for extensive scenario testing; i.e., assessments of travel patterns under a variety of socio-economic futures. Such scenario-based planning methods are becoming increasing important and relevant (Bartholomew, 2007; Swartz and Zegras, 2013). Finally, the overall disaggregate approach also enables the assessments of benefits and costs of any transportation project/policy on several specific population sub-groups (such as the elderly, the low-income, and the transit-captive). There is increasing emphasis (Alsnih and Hensher, 2003; Chakraborty, 2006) being placed on such detailed environmental justice analyses. Existing aggregate modeling methods are limited in the number of market segments that can be studied; in contrast, disaggregate models with synthetic populations can deal with any number of market segments as each individual/household is explicitly represented in the model.

While the needs and theoretical benefits of the disaggregate approach is established, it is also evident that the benefits realized are subject to accuracy of the synthetic populations and the effectiveness of the demand models. The focus of this study is on the accuracy of the target-year synthetic populations. Toward this end, this study contributes to the literature by presenting an empirical assessment of the accuracy of the target-year populations synthesized with different seed data, data-fusion methods, and control tables.

The rest of this paper is organized as follows. Section 2 presents a synthesis of literature and positions the paper in the context of past literature. Section 3 presents the analysis framework. The data used in the study are discussed in Section 4 and the results are summarized and discussed in Section 5. The paper ends (Section 6) by presenting an overall summary of the work, the major conclusions, and the directions for future work.

2. Literature synthesis

A conceptual overview of the procedure for synthesizing target-year population characteristics is presented in Fig. 1. The population for a "base year" is generated first. The base year is usually the most recent census year in the past. The synthesis of the base-year population is performed by "fusing" aggregate data in the form of control totals of select attributes with detailed (disaggregate) population characteristics available for a sample of households in the area (called the seed data and are typically available from the census). The data fusion was first accomplished using the Iterative Proportional Fitting (IPF) methodology (Beckman et al., 1996) and using only household-level attributes as controls. Other than the commonly used IPF approach, earlier studies also developed combinational optimization approaches for the generation of base-year populations (Williamson et al., 1998; Voas and Williamson, 2000), and more recently, Farooq et al. (2013) developed a method using the Markov Chain Monte Carlo principle to generate base-year populations. Subsequently, other methods for data fusion have also been developed which simultaneously accommodate multi-level controls (such as household and person level) thereby relaxing the IPF procedures' requirement that all control tables be at the same "universe" (see for instance, Srinivasan et al., 2008). In order to combine information from different "universe" (for example, household and person information), approaches based on modified or multi-stage IPF procedures (Arentze et al., 2007; Guo and Bhat, 2007; Ye et al., 2009; Auld and Mohammadian, 2010; Müller and Axhausen, 2011; Pritchard and Miller, 2012; Zhu and Joseph, 2014); entropy optimization methods (Bar-Gera et al., 2009; Lee and Fu, 2011); heuristic search techniques (Ryan et al., 2010; Abraham et al., 2012; Ma and Srinivasan, 2015) or a bipartite graph approach (Anderson et al., 2014) have been proposed during the last several years. Methods to generate populations from only aggregate data also exist (see, for example, Gargiulo et al., 2010; Barthelemy and Toint, 2013). These approaches are particularly useful when the prototypical households (seed data) are not available.

Given the base-year synthetic population, there are two approaches for generating the target-year population (Fig. 1). In one approach (called the evolution approach), each base-year household is "grown" over time to determine its characteristics at the target year. This involves modeling phenomenon 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; Pendyala et al., 2012), DEMOS (Sundararajan and Goulias, 2003), and the HA module of the Oregon2 model system (Hunt et al., 2004). Other studies have focused on evolving specific attributes of the population. For example, Paleti et al. (2011) formulate the automobile holding patterns of households by simulating the activities of disposing, replacing and adding for household vehicles. Zhu et al. (2013) evolve the ageing process to predict the marginal age distribution for a target year. Such methods are appealing as they try to mimic the real processes that households and persons go though and model behavioral decisions made at different stages of the life cycle. However, 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 (Eluru et al., 2008).

An alternate approach for generating the target-year population employs the data fusion technique which is similar to the one used in base-year population synthesis. The base-year synthetic population will serve as seed data in target-year population synthesis along with projected aggregate control totals of select attributes in the target year. Thus, unlike the evolution approach, the data fusion approach does not require evolution models and, therefore, is practical for target-year population synthesis. In this paper, we will focus on data fusion methods for target-year population synthesis.

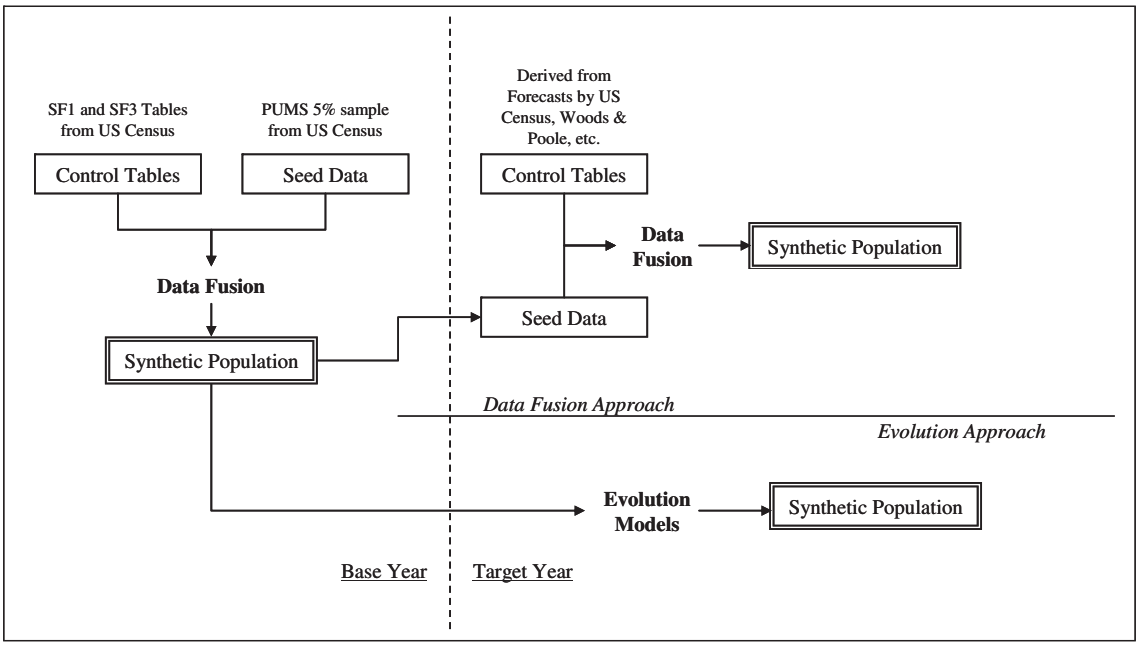

Fig. 1. A conceptual framework for target-year population synthesis.

It is also conceivable that the target-year populations can be directly synthesized using target-year controls and seed-data from the base year. While this eliminates the need for synthesizing the base year population, this procedure also ignores data available from the base-year control tables at a finer spatial geography. In situations in which the expected future growth patterns are in line with past trends, having this additional information from the base-year controls can be expected to increase the accuracy of the target year population. However, in situations in which the target year characteristics are expected to be significantly different, making the population more like the base year (by using base-year controls) can lead to increased errors rather than increase accuracy.

The literature on population synthesis using data-fusion methods is indeed quite extensive. A summary list is presented in Table 1. The reader is referred to Ma (2011) and Ma and Srinivasan (2015) for further discussions. It is quickly evident that the substantive focus of research thus far has been on synthesizing and validating the base-year populations. This is not surprising given that the primary goal of these studies has been to improve on the data-fusion methods in order to address issues with the basic IPF procedure such as the zero-cell problem (Beckman et al., 1996; Auld et al., 2009), the need for accommodating multi-level controls (Arentze et al., 2007; Guo and Bhat, 2007), and the computational issues (Rizi et al., 2013; Ma and Srinivasan, 2015) associated with dealing with large number of controls. Indeed, the spectrum of improvements proposed do empirically demonstrate that base-year populations can be synthesized more accurately by controlling for multi-level attributes

However, the analysis of the accuracy of results on target-year population synthesis is still limited. 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 (the base-year synthesis used PUMS (Public Use Microdata Sample) data as seed and these are directly obtained from the census). Thus, the methodology and controls used in the base-year synthesis impact the accuracy of the base-year population, and in turn, the accuracy of the target-year population. Second, one can expect significantly fewer socio-economic controls to be available for the target year synthesis as opposed to the base year synthesis (the base-year synthesis can take advantage of a large number of control tables available from census data bases). Some of these controls could be at the household level (such as the household size distribution) and others at the person level (such as age and gender distributions). 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 (see for instance, McCray et al., 2012; 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.

Few studies have empirically assessed the accuracy of target-year populations. Bowman and Rousseau (2008) conducted back-casting analysis (the year 2000 was used as the base year and the year 1990 was used as the target year). This analysis concluded that the accuracy of synthetic population heavily depends on the controlled tables, and for either base-year

Table 1
Literature on population synthesis using data fusion.

Studies	IPF involved	Multi-level controls	Base year synthesis	Base year validation	Target year synthesis	Target year validation
Beckman et al. (1996)	✓		✓	✓		
Williamson et al. (1998)			✓	✓		
Voas and Williamson (2000)			✓	✓		
Frick and Axhausen (2004)	✓		✓	✓		
Simpson and Tranmer (2005)	✓		✓	✓		
Bowman and Bradley (2006)			✓			
Arentze et al. (2007)	✓	✓	✓			
Guo and Bhat (2007)	✓	✓	✓	✓		
Bowman and Rousseau (2008)	✓		✓	✓	✓	✓
Srinivasan et al. (2008)		✓	✓	✓	✓	✓
Auld et al. (2009)	✓		✓	✓		
Bar-Gera et al. (2009)	✓	✓	✓	✓		
Ye et al. (2009)	✓	✓	✓	✓		
Auld et al. (2010)	✓				✓	✓
Auld and Mohammadian (2010)	✓	✓	✓	✓		
Gargiulo et al. (2010)		✓	✓	✓		
Ryan et al. (2010)		✓	✓	✓		
Lee and Fu (2011)		✓	✓	✓		
Müller and Axhausen (2011)		✓	✓	✓		
Abraham et al. (2012)		✓	✓	✓		
Kao et al. (2012)			✓	✓		
Otani et al. (2012)	✓		✓	✓		
Pritchard and Miller (2012)	✓	✓	✓	✓		
Rich and Mulalic (2012)	✓				✓	✓
Rizi et al. (2013)	✓		✓	✓		
Barthelemy and Toint (2013)	✓	✓	✓	✓		
Farooq et al. (2013)			✓	✓		
Anderson et al. (2014)		✓	✓	✓		
Zhu and Joseph (2014)	✓	✓	✓	✓		
Ma and Srinivasan (2015)		✓	✓	✓		

populations or target-year populations, the uncontrolled variables will be less accurate. Another study (Srinivasan et al., 2008) performed a similar back-cast validation for several census tracts in Florida. This study suggests that a more accurate base-year population (which is usually synthesized by controlling both household-level and person-level characteristics), is likely to lead to a more accurate target-year population. Auld et al. (2010) introduced methods for forecasting the control tables of target-year. These methods are implemented in a routine with flexible features to define different scenarios of target-year demographic changes. The method was applied to forecast the marginal controls of household size and number of workers, which are believed the important variables for target-year input of population synthesis.

However, the studies discussed above assume that the target-year controls are known accurately. The target-year control information is usually from demographic forecasts and it is well documented that these projected numbers might be significantly inaccurate (Tayman, 1996; McCray et al., 2012; Smith and Shahidullah, 1995; Rayer and Smith, 2014). The accuracy of population projections is affected through factors such as population size, growth rate, the base period length, projection horizon (Rayer, 2008) and projection methodologies vary from the simple extrapolation models which extent historical population trends into target years, to more complicated cohort-component models in which births, deaths, and migration are projected separately for different age–sex cohort (Smith and Rayer, 2011). However, the pattern of performance of different methodologies is not clear and complicated models may not necessarily provide consistently better results (Smith and Tayman, 2003). On the other hand, the pattern of some factors is clear. For example accuracy tends to decline with increasing time length between the base and forecast years (Duthie et al., 2007) and between the launch and target years (Rayer, 2008, 2010). Recently, Rich and Mulalic (2012) examined the influence of the length of forecast period on accuracy of synthesis population and the results shows that, as expected, longer forecast period will result in greater inaccuracy. Further, if such erroneous population data are fed as input to further models, there is a chance for the population forecast errors to propagate; see for example applications in transportation and land-use models (Duthie et al., 2010), health tracking and analytic epidemiology (Baker et al., 2013) and the target-year population synthesis (Srinivasan et al., 2008). Therefore, there is a need to understand the effect of inaccurate (aggregate) target-year controls on the accuracy of the synthesized (disaggregate) populations.

Overall, many factors could affect the accuracy of target-year synthetic population, including the quality of base-year population, the presence of multilevel controls of target-year population and more notably the accuracy of target-year control tables. However, past studies have not provided a comprehensive assessment of target-year populations with the consideration of all of these factors. Further, the past studies have not established that the differences in accuracy from alternate approaches are statistically significant.

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?

3. Analysis framework

A total of forty-five synthetic populations (five ways of generating the seed data × three sets of control-attributes/data-fusion methods for the target year × three levels of accuracy for target controls) 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 generally begins with the synthesis of base-year populations as these provide the seed data for target-year synthesis. Four different base-year populations were generated for each census tract with varying number of controls and differing in data-fusion methods to serve as the seed data for target-year population synthesis. 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 second base year population was synthesized using the Fitness-Based Synthesis (FBS; see Ma, 2011; Ma and Srinivasan, 2015) approach with the same controls as the first base year population, and this population is referred to as B-FBS0. The other two base-year populations were synthesized using the FBS approach but with both household- and person-level controls. These populations 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-FBS0 ∼ B-IPF. In addition to the four base-year synthetic populations, the base-year data from the PUMS were also directly as a source for the seed data for the target year synthesis. These populations are referred by B-PUMS. Overall, five ways of generating the seed-data for target-year synthesis are used in this analysis.

It is useful to emphasize that there are several data-fusion methods available now (see Table 1) many of which are fairly recent contributions to the literature. As a starting point, and to focus the effort, we limit our analysis to only two methods; the IPF and the FBS. In the IPF procedure, the cell values of a multi-way contingency table are estimated such that the marginal totals are fixed to target values and the odds ratios among the attributes are retained from the seed data (Beckman et al., 1996; Deming and Stephan, 1940; Ireland and Kullback, 1968). In the FBS procedure (Ma, 2011; Ma and Srinivasan, 2015), households are iteratively selected 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. Thus, in every iteration a household is either added or removed to improve the overall fitness of the synthesized population to the control targets and the algorithm terminates when there are no more households to add or remove (i.e., fitness does not improve any further). The FBS approach is also conceptually similar to the stochastic hill climbing (Abraham et al., 2012) and simulated annealing (Harland et al., 2012) methods that have also been recently proposed. The reader is referred to Ma (2011) and Ma and Srinivasan (2015) for an extensive discussion of the FBS approach, it computational performance, and validations (against known, "true" populations) in the context of base-year synthesis.

The second research question relates to the attributes in the target-year controls and the corresponding data-fusion method employed. To address this, the target-year populations were synthesized using three different methods and controls. The first target year population was generated using only household-level controls and the IPF method (referred to as T-IPF). The second and third target year populations were synthesized using the FBS method, where one population adopted the same household-level controls as the population synthesized using IPF (referred to as T-FBS0), and the other population controlled both household- and person-level controls (referred to as T-FBS1). Each of the five seed data (four base year synthetic populations and the base-year PUMS) were used with each of the three target-year controls/data fusion methods giving a total of fifteen target year populations. These are referred to as T-IPF-B-FBS2 (i.e., target year IPF and base year FBS2), T-IPF-B-FBS1, T-IPF-B-FBS0, T-IPF-B-IPF, T-IPF-B-PUMS, T-FBS0-B-FBS2, T-FBS0-B-FBS1, T-FBS0-B-FBS0, T-FBS0-B-IPF, T-FBS0-B-PUMS, T-FBS1-B-FBS2, T-FBS1-B-FBS1, T-FBS1-B-FBS0, T-FBS1-B-IPF, and T-FBS1-B-PUMS reflecting the population used as seed data and the target-year synthesis methodology.

The fifteen populations described above were first 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. Therefore, fifteen additional populations were synthesized using these approximate control tables and there are referred to as T*-IPF-B-FBS2, T*-IPF-B-FBS1, T*-IPF-B-FBS0, T*-IPF-B-IPF, T*-IPF-B-PUMS, T*-FBS0-B-FBS2, T*-FBS0-B-FBS1, T*-FBS0-B-FBS0, T*-FBS0-B-IPF, T*-FBS0-B-PUMS, T*-FBS1-B-FBS2, T*-FBS1-B-FBS1, T*-FBS1-B-FBS0, T*-FBS1-B-IPF, and T*-FBS1-B-PUMS. In these populations, *all* target year controls were assumed to be erroneous. In addition, another set of populations assumed that only *one* control table, namely household size for the target year was erroneous and the rest were taken to be the true values. These fifteen populations are referred to as T#-IPF-B-FBS2, T#-IPF-B-FBS1, T#-IPF-B-FBS0, T#-IPF-B-IPF, T#-IPF-B-PUMS, T#-FBS0-B-FBS2, T#-FBS0-B-FBS1, T#-FBS0-B-FBS0, T#-FBS0-B-IPF, T#-FBS0-B-PUMS, T#-FBS1-B-FBS2, T#-FBS1-B-FBS1, T#-FBS1-B-FBS0, T#-FBS1-B-IPF, and T#-FBS1-B-PUMS.

Note that, in the labels used to describe the synthetic populations, the "*" is used to indicate that all target-year control tables were approximate and the "#" indicates that only one target-year control table (household size) was approximate. If the label has neither a "*" nor "#", this means that the population was synthesized using accurate values for all target-year controls.

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 marginal tables also include attributes that were uncontrolled in the synthesis procedure. This study uses absolute percentage errors which is the most commonly used measure for judging the accuracy of population projections (Smith and Tayman, 2003; Rayer, 2007). The error specific to marginal table j, D_j is calculated as follows:

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

Here, 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.

4. Data

Data for 12 census tracts and their corresponding PUMAs (Public Use Microdata Areas) and counties from Florida were used for this analysis (Table 2). These census tracts are from some of the major urban regions in Florida where advanced travel-demand models are likely to be needed or developed. Data were collected for years 1990 and 2000. The reader will note that there are wide variations in the populations and the changes in population between the years. 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 in Bowman and Rousseau, 2008 and Srinivasan et al., 2008) 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 more readily for US census 2000.

Table 3 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. Each of these tables corresponds to the joint distribution of a subset of the population attributes typically required as inputs for travel models. All these 11 tables were controlled for in the synthesis of the B-FBS2 (base-year) population. For the synthesis of the other three base-year populations, only a subset of these controls were used with only household-level controls being used for the B-IPF and B-FBS0 population. Table 3 also identifies 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 for the year 2000.

For the target-year synthesis, the marginal distributions of household size and dwelling-unit type were used as controls in the T-IPF and T-FBS0 procedures. These are two major attributes used as inputs in four-step travel demand forecasting models. Person level controls for age and gender were used in addition to the two household level controls in the T-FBS1 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 actually used in the synthesis of the population for the target year, several other marginal tables are also available for the target year from the SF1 and SF3 census files of 1990 (Fig. 2). These were used to assess the accuracy of the synthesized target-year populations. Fig. 2 also indicates whether any or all of the attributes of the

Table 2
Population characteristics of the 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 3
Control tables for base-year population synthesis.

No.	Control tables	Controlled in				Universe	Dimension 1		Dimension 2	
		B-FBS2	B-FBS1	B-FBS0	B-IPF		Attribute	Categories	Attribute	Categories
1	H15	Y	Y	Y	Y	Households	TENURE	Own, Rent	HHSIZE	1, 2, 3, 4, 5, 6, 7+
2	H32	Y				Households	TENURE	Own, Rent	DUTYPE	Single Family, Multi-Family
3	H44	Y	Y	Y	Y	Households	TENURE	Own, Rent	NUMAUTO	0, 1, 2, 3, 4, 5+
4	P26	Y				Households	HHSTRUCT	Family, Non-Family	HHSIZE	1, 2, 3, 4, 5, 6, 7+
5	P34	Y				Families	HHSTRUCT	Married couple, Other family	CHAGE[a]	None, Only <6 years, Only ⩾6 years, Both <6 years and ⩾6 years
6	P48	Y				Families	HHSTRUCT	Married couple, Other family	NUMWORK[b]	0, 1, 2, 3+
7	P52	Y				Households	INCOME	<30 K, 30–50 K, 50–75 K, 75–125 K, more than 125 K	NA	
8	P7	Y	Y			Total Population	ETHNICITY	White, Black, Other, and Multiple Race	NA	
9	P12	Y	Y			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
10	P21	Y				Total Population	CITIZEN	Native, Naturalized, Non Citizen	NA	
11	P47	Y				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 h per week in 1999).
[c] Hours of work per week in 1999.

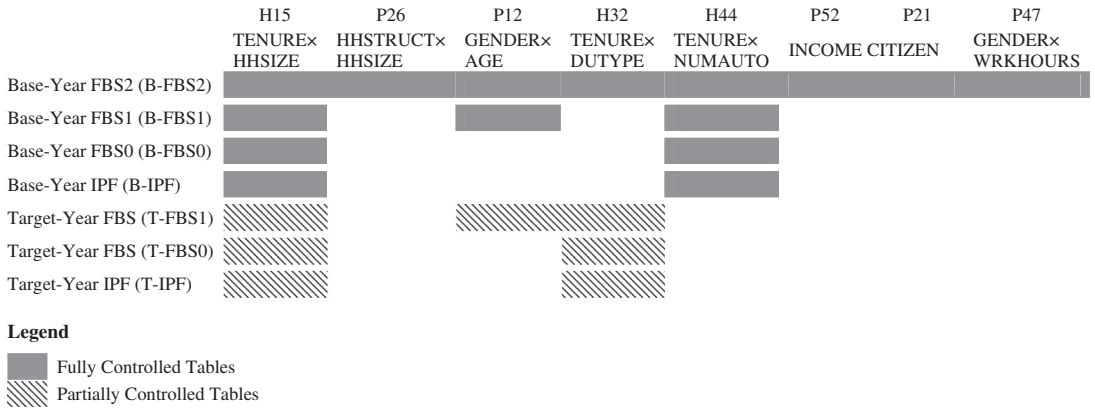

Fig. 2. Marginal tables for assessing the target-year populations.

different marginal tables used for validation were controlled for in synthesizing either the base-year or the target-year populations.

5. Results

For each of the twelve census tracts analyzed, forty-five populations were synthesized (five ways of generating the seed data × three sets of control-attributes/data-fusion methods for the target year × three levels of accuracy for target controls). For each tract and synthetic population, error measures were calculated for each of the marginal tables available for validation using the formula described in Section 3. The errors are then averaged across the 12 census tracts for each of the forty-five synthetic populations and each of the marginal tables available for validation. Thus, we obtain the average error of each synthetic population in replicating each of several marginal tables for the target year.

Section 5.1 examines the marginal impact of the accuracy of the base-year population on the accuracy of the target-year population. Section 5.2 examines the impact of target-year controls and methods. Section 5.3 presents the impact of the accuracy of the target year control tables. Section 5.4 presents a simple regression model to determine the relative impacts of all three factors (target-year seed data, target-year controls and methods, and accuracy of target-year controls) on target-year population errors.

5.1. Impact of accuracy of the base year population

Prior to assessing the impact of the base-year synthetic population on the accuracy of the target-year population, it is useful to examine the accuracy of the base-year synthetic population. Table 4 compares the four base-year synthetic populations using the error measures as defined before. The B-FBS2 population (synthesized with the maximum number of controls and using the FBS methodology) is generally the most accurate in replicating most tables (for each control table, the least error is highlighted in bold font). Further, the numbers also indicate that increasing accuracy (or decreasing error) with the addition of more controls to the base-year synthesis).

Fig. 3 includes three sets of graphs which compare the accuracy of the target-year populations synthesized with five different seed-data (four synthetic populations for the base year and the base-year PUMS) but with the same (accurate) target year controls and data fusion methodology. The top graph is for the target-year synthesis with FBS methodology and both household and person controls (T-FBS1), the middle and the bottom graph is for the cases when the target year synthesis was undertaken with only household controls using the FBS and IPF procedures respectively (T-FBS0 and T-IPF). The entries along the "X" axis represent the target-year marginal tables used for validation. All these are for the case when the accurate tract-level controls were used (similar trends were observed for the case of approximate controls and hence these are not presented graphically here). In general, we observe that the errors for most marginal-tables are least for populations synthesized using B-FBS2 as the seed data and are maximum for the populations synthesized using B-IPF or B-PUMS as the seed data for most cases. The differences are particularly striking for marginal tables such as P52 (17% errors for T-FBS/IPF-B-FBS2 versus 32–38% errors for T-FBS/IPF-B-PUMS) which has attributes that are not controlled for in the target year synthesis. 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. Of course, this result holds over a ten-year projection horizon and for the range of changes in the population observed across the census tracts analyzed.

Table 4
Accuracy of synthesized base-year populations.

Control tables Attributes	H15 TENURE × HHSIZE	P26 HHSTRUCT × HHSIZE	P34 HHSTRUCT × CHAGE	P7 ETHNICITY	P12 GENDER × AGE	H32 TENURE × DUTYPE	H44 TENURE × NUMAUTO	P48 HHSTRUCT × NUMWORK	P52 INCOME	P21 CITIZEN	P47 GENDER × WRKHOURS
B-FBS2	0.03	**0.02**	**0.03**	**0**	0.01	**0.01**	0.02	**0.04**	**0.01**	**0**	**0**
B-FBS1	0.03	0.05	0.15	0	**0**	0.22	**0.01**	0.22	0.24	0.07	0.13
B-FBS0	**0**	0.05	0.2	0.23	0.17	0.22	**0.01**	0.23	0.28	0.08	0.16
B-IPF	0.01	0.05	0.2	0.26	0.17	0.21	**0.01**	0.25	0.28	0.09	0.18

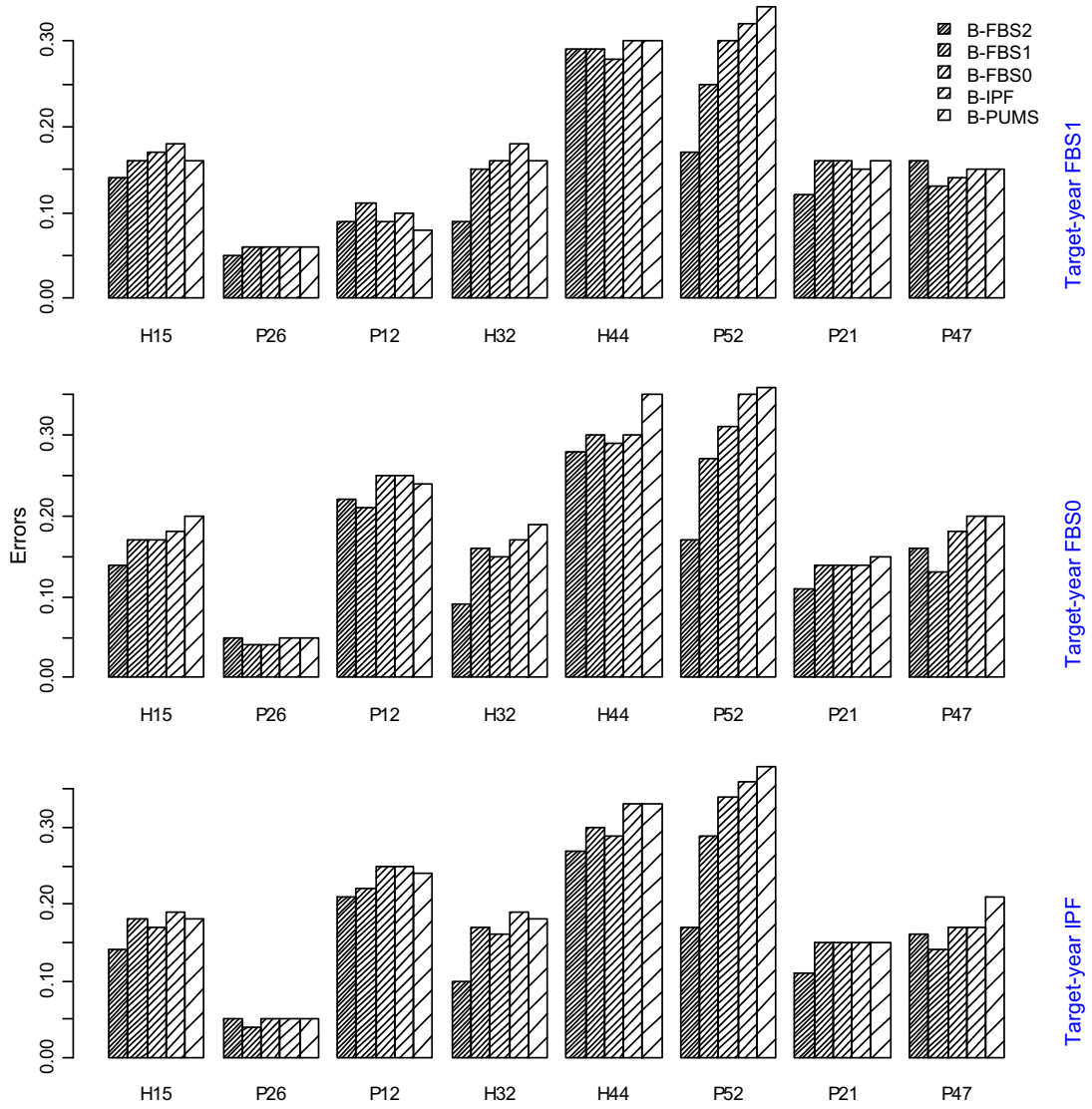

Fig. 3. Impact of base-year populations on the accuracy of target-year population.

5.2. Impact of target-year control tables and methods

Fig. 4 includes five 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 methodologies. Each graph compares the (target) population synthesized with both household and person controls and using the FBS methodology against the population synthesized with only household controls using the FBS and IPF methodologies. The topmost graph is for the cases when the seed data were the B-FBS2 population (base year population synthesized with most controls). This is followed by the results when the seed data were B-FBS1. The bottom-most graph represents the seed data of B-PUMS. 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).

For each fixed seed-data, the three target year populations perform similar, especially for the T-FBS0 and T-IPF populations in the context of accuracy with the T-FBS1 populations 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 is the control for age in the T-FBS1 populations. 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.

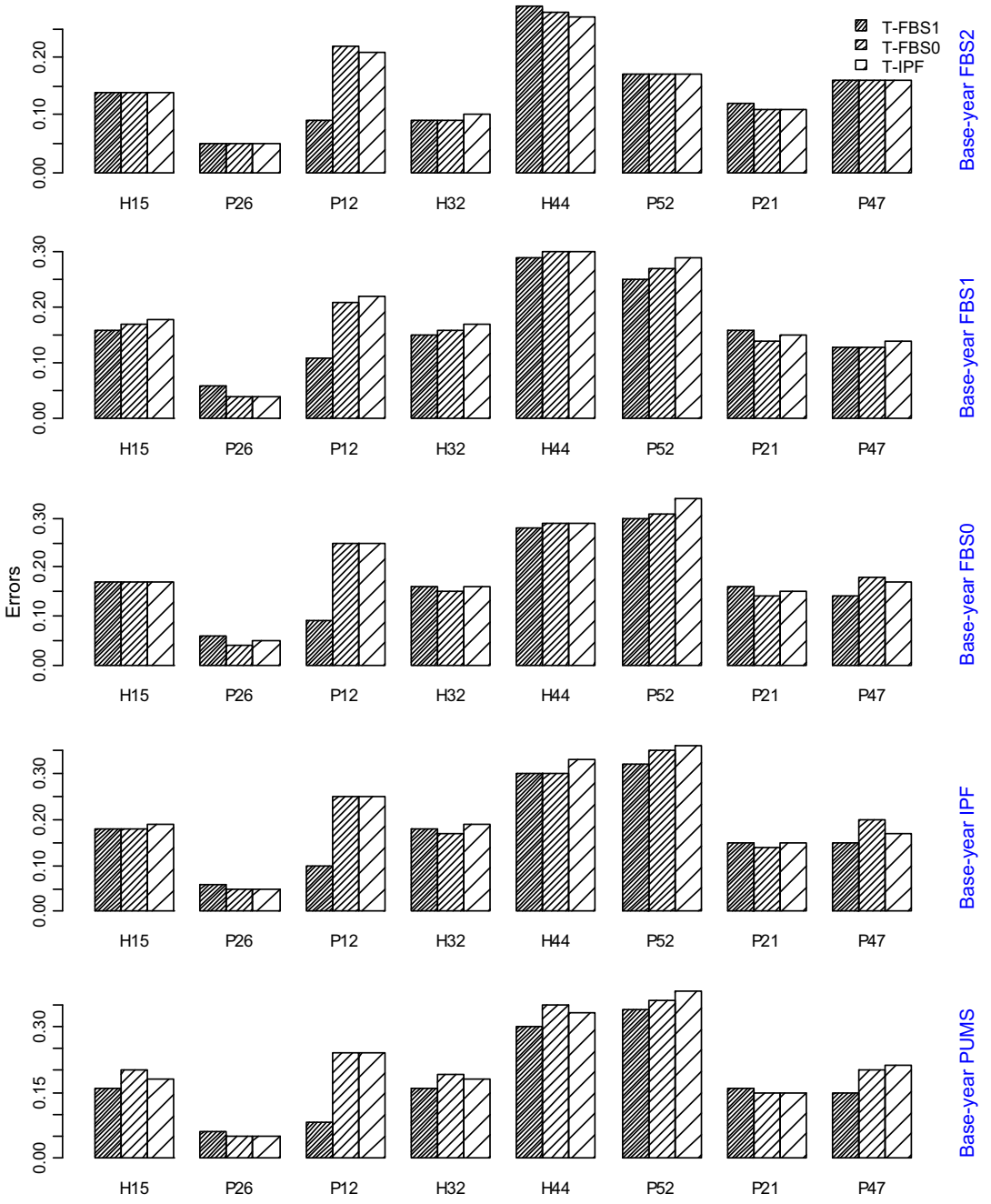

Fig. 4. Impact of target-year controls and data-fusion methodology on the accuracy of target-year population.

5.3. 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 5, 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. Note that Table 5 presents the results of all 45 synthetic populations organized in groups of three. The first row in any group of three represents the case when true target-year control totals are used in the synthesis (T). The second row represents the case in which approximate control totals are used for all attributes (T*). The third row represents the case in which the approximate controls were used only for one table (household size distribution) and accurate controls were used for the rest (T#). Therefore, for any column, the value in the first row of any group would be lowest (less error or

Table 5
Accuracy of target-year synthetic populations.

Control tables Attributes	H15 TENURE × HHSIZE	P26 HHSTRUCT × HHSIZE	P12 GENDER × AGE	H32 TENURE × DUTYPE	H44 TENURE × NUMAUTO	P52 INCOME	P21 CITIZEN	P47 GENDER × WRKHOURS
T-FBS-B-FBS2	0.14	0.05	0.09	0.09	0.29	0.17	0.12	0.16
T*-FBS-B-FBS2	0.34	0.22	0.28	0.41	0.41	0.20	0.13	0.25
T#-FBS-B-FBS2	0.25	0.22	0.10	0.1	0.27	0.18	0.12	0.16
T-FBS-B-FBS1	0.16	0.06	0.11	0.15	0.29	0.25	0.16	0.13
T*-FBS-B-FBS1	0.27	0.23	0.26	0.41	0.32	0.32	0.14	0.20
T#-FBS-B-FBS1	0.26	0.22	0.11	0.15	0.28	0.28	0.15	0.14
T-FBS-B-FBS0	0.17	0.06	0.09	0.16	0.28	0.30	0.16	0.14
T*-FBS-B-FBS0	0.27	0.23	0.26	0.41	0.33	0.36	0.15	0.22
T#-FBS-B-FBS0	0.27	0.22	0.10	0.16	0.29	0.32	0.16	0.15
T-FBS-B-IPF	0.18	0.06	0.10	0.18	0.30	0.32	0.15	0.15
T*-FBS-B-IPF	0.26	0.23	0.26	0.41	0.31	0.40	0.16	0.22
T#-FBS-B-IPF	0.28	0.22	0.10	0.16	0.27	0.35	0.15	0.15
T-FBS-B-PUMS	0.16	0.06	0.08	0.16	0.30	0.34	0.16	0.15
T*-FBS-B-PUMS	0.39	0.24	0.25	0.46	0.49	0.41	0.16	0.24
T#-FBS-B-PUMS	0.26	0.22	0.09	0.15	0.35	0.38	0.17	0.15
T-FBS0-B-FBS2	0.14	0.05	0.22	0.09	0.28	0.17	0.11	0.16
T*-FBS0-B-FBS2	0.34	0.24	0.29	0.40	0.36	0.21	0.18	0.20
T#-FBS0-B-FBS2	0.27	0.23	0.26	0.09	0.30	0.21	0.15	0.17
T-FBS0-B-FBS1	0.17	0.04	0.21	0.16	0.30	0.27	0.14	0.13
T*-FBS0-B-FBS1	0.27	0.23	0.25	0.41	0.30	0.30	0.17	0.15
T#-FBS0-B-FBS1	0.30	0.23	0.26	0.16	0.30	0.29	0.16	0.15
T-FBS0-B-FBS0	0.17	0.04	0.25	0.15	0.29	0.31	0.14	0.18
T*-FBS0-B-FBS0	0.27	0.23	0.29	0.41	0.32	0.37	0.18	0.18
T#-FBS0-B-FBS0	0.29	0.23	0.29	0.16	0.32	0.35	0.15	0.17
T-FBS0-B-IPF	0.18	0.05	0.25	0.17	0.30	0.35	0.14	0.20
T*-FBS0-B-IPF	0.28	0.24	0.29	0.42	0.33	0.40	0.17	0.20
T#-FBS0-B-IPF	0.31	0.23	0.29	0.17	0.32	0.38	0.15	0.19
T-FBS0-B-PUMS	0.20	0.05	0.24	0.19	0.35	0.36	0.15	0.20
T*-FBS0-B-PUMS	0.40	0.25	0.29	0.47	0.49	0.43	0.19	0.20
T#-FBS0-B-PUMS	0.29	0.24	0.28	0.19	0.38	0.40	0.16	0.21
T-IPF-B-FBS2	0.14	0.05	0.21	0.10	0.27	0.17	0.11	0.16
T*-IPF-B-FBS2	0.33	0.24	0.28	0.40	0.37	0.22	0.17	0.21
T#-IPF-B-FBS2	0.28	0.23	0.25	0.09	0.28	0.19	0.15	0.18
T-IPF-B-FBS1	0.18	0.04	0.22	0.17	0.30	0.29	0.15	0.14
T*-IPF-B-FBS1	0.26	0.24	0.26	0.42	0.31	0.33	0.18	0.16
T#-IPF-B-FBS1	0.29	0.23	0.27	0.17	0.32	0.32	0.17	0.16
T-IPF-B-FBS0	0.17	0.05	0.25	0.16	0.29	0.34	0.15	0.17
T*-IPF-B-FBS0	0.28	0.24	0.28	0.42	0.31	0.39	0.16	0.19
T#-IPF-B-FBS0	0.30	0.23	0.30	0.16	0.31	0.35	0.16	0.17
T-IPF-B-IPF	0.19	0.05	0.25	0.19	0.33	0.36	0.15	0.17
T*-IPF-B-IPF	0.28	0.24	0.29	0.43	0.34	0.40	0.16	0.20
T#-IPF-B-IPF	0.30	0.24	0.29	0.18	0.33	0.39	0.15	0.19
T-IPF-B-PUMS	0.18	0.05	0.24	0.18	0.33	0.38	0.15	0.21
T*-IPF-B-PUMS	0.41	0.25	0.29	0.47	0.49	0.45	0.18	0.22
T#-IPF-B-PUMS	0.31	0.24	0.29	0.18	0.37	0.42	0.16	0.22

Table 6
Difference between true controlled tables and erroneous tables.

Case ID	Household size	Dwelling type	Age	Gender	Average 1[a]	Average 2[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] Average 1 is the average of all four attributes listed in the table.
[b] Average 2 is the average of "household size" and "dwelling type".

more accurate) and the value in the second row of any group would be the highest (more error or less accurate). 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. As the errors are introduced by replacing the true tract-level marginal tables with the corresponding one from the county, the magnitudes of the errors are not equal across the tracts. Table 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, Average 1 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 and approximate-household and person controls for target year synthesis (T-FBS1). Average 2 is calculated across the two household-level control tables and, hence, it may be interpreted as the input error (or discrepancy) introduced in populations employing only approximate household controls only for target year synthesis (T-FBS0 and T-IPF).

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 (all control tables are approximate) is calculated (for each base year population and target year synthesis approach). This difference is averaged across all marginal tables and is defined as the loss of accuracy for the census tract. Fig. 5 plots the input error (discrepancy) against the loss of accuracy for each census tracts (the numbers within the charts identify the census tracts) and for each of the fifteen types of synthetic populations (five ways of generating the seed data × three sets of methods and controls for the target year). Fig. 6 plots discrepancy against loss of accuracy for the cases in which only one control table (household size) was assumed to be approximate (Fig. 5 corresponds to the cases in which *all* control tables are approximate). By comparing Figs. 5 and 6, it is interesting that the slope of the estimated linear function between the discrepancy and the loss of accuracy in Fig. 5 is larger than the one in Fig. 6, which reflects that more inaccurate tables leads to additional loss of accuracy.

Note that census tract 11 is most similar to the county in which it is located (based on Table 6). Therefore, when the tract-level controls are replaced by the county level data, the input errors introduced are small and this translates into a small loss of accuracy in the synthetic populations. However, tract 2 is most dissimilar to the county in which it is located. Therefore, when the tract-level controls are replaced by the county level data, the input errors introduced are large and this translates into a larger loss of accuracy in the synthetic populations. In Figs. 5 and 6, tract 11 shows up in the bottom left corner of the charts and tract 2 is located in the top-right corner. In general, the loss of accuracy is greater with greater input errors and this relationship appears to be linear.

It is also useful to acknowledge that this analysis assumes that all marginal tables are equally important and therefore a simple average of the errors across these tables may be used as an indicator of the overall population accuracy. Conceptually, this analysis can also be performed with a weighted average if accuracy in certain attributes is more important than others.

5.4. Overall accuracy assessment

The previous sections presented the marginal effects of each of three factors affecting the accuracy of target-year populations independently. In this section, we present a statistical comparison of the accuracies of all 540 synthetic target-year populations (45 target-year populations for each twelve census tracts). This is accomplished by running a regression model (Table 7) on the tract-level errors for each synthetic population (averaged across all marginal tables available for validation, see Fig. 2) against the key characteristics of the method used to synthesize the corresponding target-year population. These key characteristics include seed data (5 types), target-year synthesis method (3 types), and the discrepancy introduced in control table (zero if true controls were used).

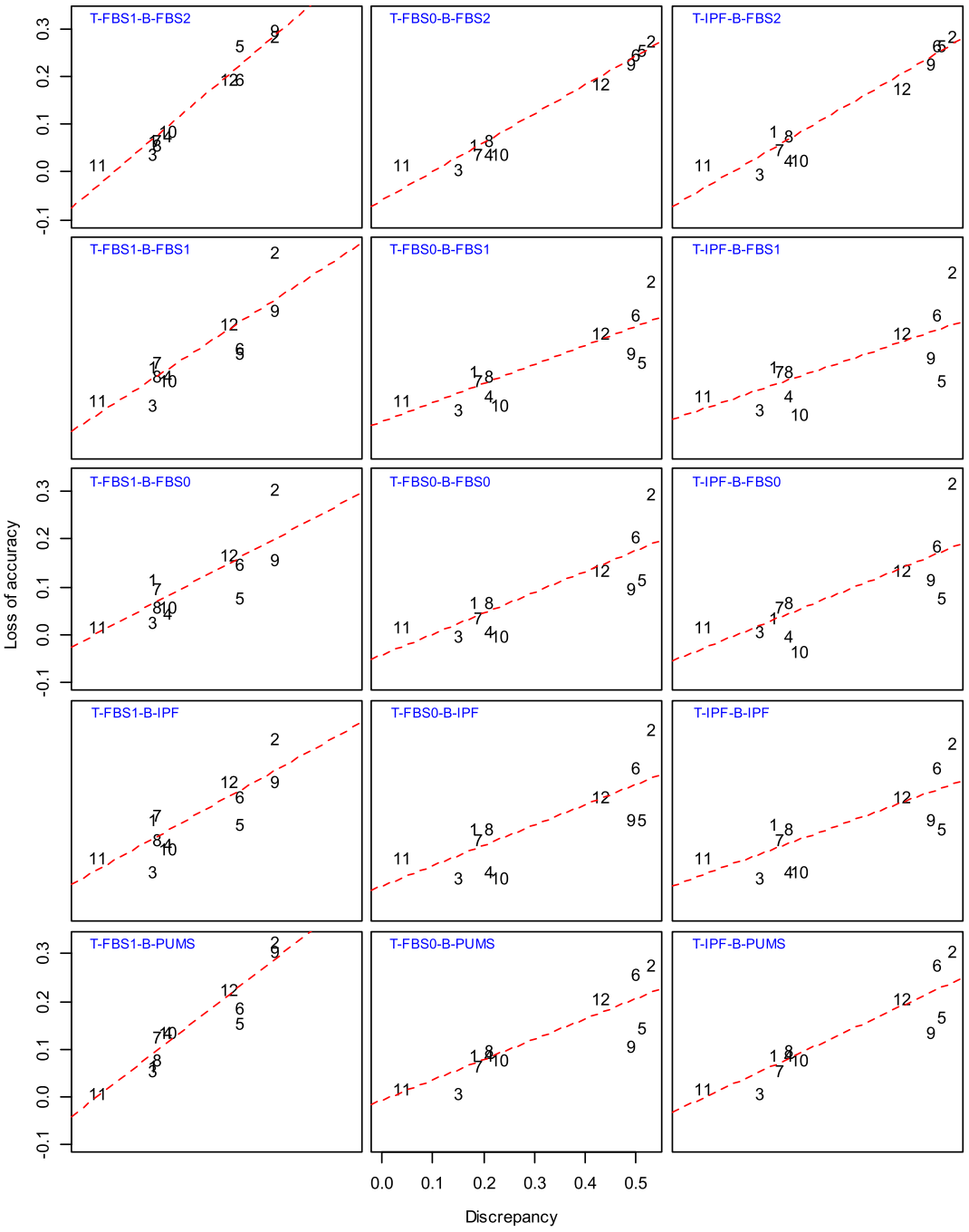

Fig. 5. Impact of inaccurate control tables on the change of accuracy of target-year populations.

To examine the impact of seed data, B-PUMS was used as the reference (coefficient is zero). The coefficients on each of the other four base-year synthetic populations are negative indicating that using a synthesized based year population leads to more accurate (less error) target-year populations instead of simply using the base-year PUMS data as the seed. We conclude that over a ten-year forecast horizon and for the range of changes seen in the census tracts analyzed, the use of base-year synthesis data do indeed outperform the use of PUMS. The reader will recall that we had also discussed situations in which the use of PUMS could outperform the use of base-year synthetic data (see Section 2). Further, the coefficients are decreasing (more negative) from B-IPF to B-FBS2. This indicates that controlling for more attributes for the base-year does statistically improve the accuracy of the target-year populations.

To examine the effect of target-year controls and data-fusion methodology, the use of household-only controls and IPF was taken as the reference (T-IPF). The coefficient on T-FBS1 is negative and significant indicating that controlling for more

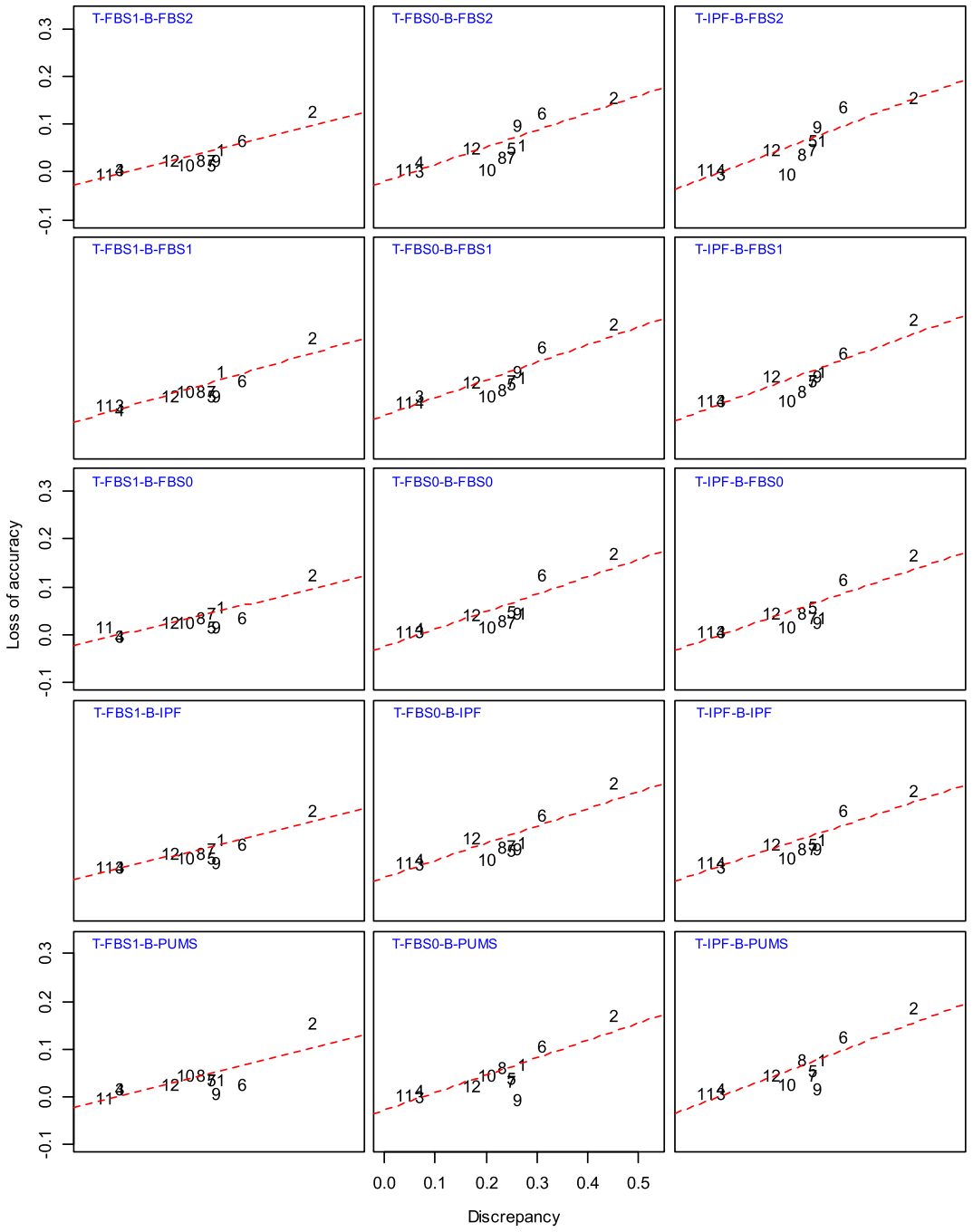

Fig. 6. Impact of inaccurate household-size control on the change of accuracy of target-year populations.

target-year attributes also improves the accuracy of the synthetic population. The coefficient on T-FBS0 was estimated to be statistically insignificant. This indicates that all else being equal, the use of FBS and IPF with the same household-only controls would output target populations with similar accuracy.

The coefficient on discrepancy was estimated to be 0.441 and 0.278 for cases with multiple erroneous controls and cases with only household size control respectively. This implies that increased errors in the projection of the control totals will increase the errors in the synthetic population for the target year. Further, multiple erroneous tables can lead to an even greater increase in the error of the synthesized populations.

It is useful to note that, the regression model also controls for target-year population size and absolute change in population from base year (%) for each tract recognizing that the errors could be impacted by both the size of the population being

Table 7
Regression results for the overall accuracy assessment.

Explanatory variables	Coefficients	*t* value	*P*(>|*t*|)
Intercept	2.24E−01	30.024	<2E−16
Seed data from base year (B-PUMS as the reference)			
B-FBS2	−5.88E−02	−8.695	<2E−16
B-FBS1	−4.45E−02	−6.586	1.09E−10
B-FBS0	−3.13E−02	−4.628	4.64E−06
B-IPF	−2.24E−02	−3.314	9.83E−04
Target year methods (T-IPF as the reference)			
T-FBS1	−1.27E−02	−2.417	1.60E−02
T-FBS0	−3.56E−03	−0.679	0.50
Discrepancy for cases with multiple erroneous controls	4.41E−01	29.325	<2E−16
Discrepancy for cases with erroneous household size control	2.78E−01	14.059	<2E−16
Target year population size	−9.86E−06	−11.181	<2E−16
Absolute change in population from base year (%)	1.13E−03	9.474	<2E−16

synthesized and the growth of the zone relative to the base year. The results indicate that zones with more population have lower error (note that the error represents a fraction as defined in Section 3) and that zones that have experienced a greater increase in population have greater errors.

Overall, the statistical modeling results empirically demonstrate the need for improved (more accurate) base-year population synthesis, multiple target-year controls, and accurate control-table projections for improving the accuracy of the target-year population. These results are obtained after controlling for differences in population sizes and growth patterns across the zones.

6. Summary and conclusions

The application of disaggregate models for predictions and policy evaluations requires as inputs detailed information on the socio-economic characteristics of the target-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 seed data, data-fusion methods, and control tables. Forty-five 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 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 (over a ten-year projection horizon and for the range of population change observed in the census tracts analyzed). The populations synthesized (target year) with multi-level controls do perform better in replicating certain attributes than those synthesized with only household level controls. 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 propagated according to the magnitude of the input errors introduced via the control tables. In sum, accurate base-year population synthesis and accurate projections of target year controls are keys to ensuring the accuracy of target-year populations.

It is anticipated that future studies will examine the accuracy of population estimates obtained using synthesis methods other than IPF and FBS. Examining the accuracy of populations synthesized with controls at different spatial scales is also of interest. In all these analyses it would be of interest to examine alternate measures of error/accuracy of the synthetic population. Finally, it is also important to feed alternate populations into travel demand models to assess the impacts of population errors on travel-demand estimates.

Acknowledgments

This research was supported by the Systems Planning Office of the Florida Department of Transportation, National Natural Science Foundation of China, Nos. 51208032 and 71210001. The authors also acknowledge valuable feedback provided by three reviewers on an earlier draft of this manuscript.

References

Abraham, J.E., Stefan, K.J., Hunt, J.D., 2012. Population synthesis using combinatorial optimization at multiple levels. In: Presented at the 91th Annual Meeting of Transportation Research Board, Washington DC.

Alsnih, R., Hensher, D.A., 2003. The mobility and accessibility expectations of seniors in an aging population. Transport. Res. Part A 37, 903–916.

Anderson, P., Farooq, B., Efthymiou, D., Bierlaire, M., 2014. Associations generation in synthetic population for transportation applications: a graph-theoretic solution. Transport. Res. Rec. 2429, 38–50.
Arentze, T., Timmermans, H.J.P., Hofman, F., 2007. Creating synthetic household populations: problems and approach. Transport. Res. Rec. 2014, 85–91.
Auld, J., Mohammadian, A., Wies, K., 2009. Population synthesis with subregion-level control variable aggregation. J. Transport. Eng. 135 (9), 632–639.
Auld, J., Mohammadian, A., 2010. Efficient methodology for generating synthetic populations with multiple control levels. Transport. Res. Rec. 2175, 138–147.
Auld, J., Rashidi, T.H., Mohammadian, A., 2010. Evaluating transportation impacts of forecast demographic scenarios using population synthesis and data transferability. In: Presented at the 89th Annual Meeting of Transportation Research Board, Washington DC.
Baker, J.D., Alcantara, A., Ruan, X., Vasan, S., Nathan, C., 2013. An evaluation of the accuracy of small-area demographic estimates of population at risk and its effect on prevalence statistics. Popul. Health Metrics 11, 24.
Bar-Gera, H., Konduri, K., Sana, B., Ye, X., Pendyala, R.M., 2009. Presented at the 88th Annual Meeting of Transportation Research Board, Washington DC.
Barthelemy, J., Toint, P.L., 2013. Synthetic population generation without a sample. Transport. Sci. 47 (2), 266–279.
Bartholomew, K., 2007. Land use-transportation scenario planning: promise and reality. Transportation 34, 397–412.
Beckman, R.J., Baggerly, K.A., Mckay, M.D., 1996. Creating synthetic baseline populations. Transport. Res. Part A 30 (6), 415–429.
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. <http://jbowman.net/ProjectDocuments/SacSim/SACOG%20tech%20memo%202–Pop%20Synth.20060731.pdf>.
Bowman, J.L., Rousseau, G., 2008. Validation of Atlanta, Georgia, regional commission population synthesizer. Transport. Res. Board Conf. Proc. 2 (42), 54–62.
Chakraborty, J., 2006. Evaluating the environmental justice impacts of transportation improvement projects in the US. Transport. Res. Part D 11, 315–323.
Deming, W.E., Stephan, F.F., 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Ann. Math. Stat. 11 (4), 427–444.
Duthie, J., Cervenka, K., Waller, S.T., 2007. Environmental justice analysis challenges for metropolitan transportation planning. Transport. Res. Rec. 2013, 98–107.
Duthie, J., Voruganti, A., Kockelman, K., Waller, S., 2010. Highway improvement project rankings due to uncertain model inputs: application of traditional transportation and land use models. J. Urban Plann. Dev. 136 (4), 294–302.
Eluru, N., Pinjari, A.R., Guo, J.Y., Sener, I.N., Srinivasan, S., Copperman, R., Bhat, C.R., 2008. Population updating system structures and models embedded within the comprehensive econometric microsimulator for urban systems. Transport. Res. Rec. 2076, 171–182.
Farooq, B., Bierlaire, M., Hurtubia, R., Flötteröd, G., 2013. Simulation based synthesis of population. Transport. Res. Part B 58, 243–263.
Frick, M., Axhausen, K.W., 2004. Generating synthetic populations using IPF and Monte Carlo techniques: some new results. In: Presented at the 4th Swiss Transport Research Conference.
Gargiulo, F., Ternes, S., Huet, S., Deffuant, G., 2010. An iterative approach for generating statistically realistic populations of households. PLoS ONE 5 (1), e8828. http://dx.doi.org/10.1371/journal.pone.0008828.
Goulias, K.G., Kitamura, R., 1996. A dynamic model system for regional travel demand forecasting. In: Golob, T., Kitamura, R., Long, L. (Eds.), Panels for Transportation Planning: Methods and Applications. Kluwer Academic Publishers, Boston, pp. 321–348 (Chapter 13).
Guo, J.Y., Bhat, C.R., 2007. Population synthesis for microsimulating travel behavior. Transport. Res. Rec. 2014, 92–101.
Harland, K., Heppenstall, A., Smith, D., Birkin, M., 2012. Creating realistic synthetic populations at varying spatial scales: a comparative critique of population synthesis techniques. J. Artif. Soc. Soc. Simul. 15 (1), 1.
Hunt, J.D., Abraham, J.E., Weidner, T., 2004. The household allocation (HA) module of the oregon2 model. Transport. Res. Rec. 1898, 98–107.
Ireland, C.T., Kullback, S., 1968. Contingency tables with given marginals. Biometrika 55 (1), 179–188.
Kao, S., Kim, H., Liu, C., Cui, X., Bhaduri, B.L., 2012. Dependence-preserving approach to synthesizing household characteristics. Transport. Res. Rec. 2302, 192–200.
Koppelman, F.S., 1974. Prediction with disaggregate models: the aggregation issue. Transport. Res. Rec. 527, 73–80.
Landau, U., 1978. Aggregate prediction with disaggregate models: behavior of the aggregation bias. Transport. Res. Rec. 673, 100–105.
Lee, D., Fu, Y., 2011. Cross-entropy optimization model for population synthesis in activity-based microsimulation models. Transport. Res. Rec. 2255, 20–27.
Ma, L., 2011. Generating Disaggregate Population Characteristics for Input to Travel Demand Models. Ph.D. Dissertation, The University of Florida, USA.
Ma, L., Srinivasan, S., 2015. Synthetic population generation with multilevel controls: a fitness-based synthesis approach and validations. Comput.-Aided Civ. Infrastruct. Eng. 30, 135–150.
Mackett, R.L., 1990. MASTER Mode. Report SR 237. Transport and Road Research Laboratory, Crowthorne, England.
McCray, D.R., Miller, J.S., Hoel, L., 2012. Accuracy of zonal socioeconomic forecasts for travel demand modeling: retrospective case study. Transport. Res. Rec. 2032, 148–156.
Müller, K., Axhausen, K.W., 2011. Hierarchical IPF: generating a synthetic population for Switzerland. In: Presented at the 4th European Regional Science Association Conference. <http://www.sustaincity.org/publications/SC_Hierarchical_IPF.pdf>.
Otani, N., Sugiki, N., Vichiensan, V., Miyamoto, K., 2012. Modifiable attribute cell problem and solution method for population synthesis in land use microsimulation. Transport. Res. Rec. J. Transport. Res. Board 2302, 157–163.
Paleti, R., Eluru, N., Bhat, C.R., Pendyala, R.M., Adler, T.J., Goulias, K.G., 2011. Design of comprehensive microsimulator of household vehicle fleet composition, utilization, and evolution. Transport. Res. Rec. 2254, 44–57.
Pendyala, R.M., Bhat, C.R., Goulias, K.G., Paleti, R., Konduri, K.C., Sidharthan, R., Hu, H., Huang, G., Christian, K.P., 2012. Application of socioeconomic model system for activity-based modeling: experience from southern California. Transport. Res. Rec. J. Transport. Res. Board 2303, 71–80.
Pritchard, D.R., Miller, E.J., 2012. Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. Transportation 39 (3), 685–704.
Rayer, S., 2007. Population forecast accuracy: does the choice of summary measure of error matter? Popul. Res. Policy Rev. 26, 163–184.
Rayer, S., 2008. Population forecast errors: a primer for planners. J. Plann. Educ. Res. 27, 417–430.
Rayer, S., 2010. Factors affecting the accuracy of subcounty population forecasts. J. Plann. Educ. Res. 30 (2), 147–161.
Rayer, S., Smith, K.S., 2014. Population projections by age for Florida and its counties: assessing accuracy and the impact of adjustments. Popul. Res. Policy Rev. http://dx.doi.org/10.1007/s11113-014-9325-x.
Rich, J., Mulalic, I., 2012. Generating synthetic baseline populations from register data. Transport. Res. Part A 46 (3), 467–479.
Rizi, S.M.M., Łatek, M.M., Geller, A., 2013. Fusing remote sensing with sparse demographic data for synthetic population generation: an algorithm and application to rural Afghanistan. Int. J. Geogr. Inform. Sci. 27 (5), 986–1004.
Ryan, J., Maoh, H., Kanaroglou, P., 2010. Population synthesis for microsimulating urban residential mobility. In: Presented at the 89th Annual Meeting of Transportation Research Board, Washington DC.
Simpson, L., Tranmer, M., 2005. Combining sample and census data in small area estimates: iterative proportional fitting with standard software. Prof. Geogr. 57 (2), 222–234.
Smith, K.S., Shahidullah, M., 1995. An evaluation of population projection errors for census tracts. J. Am. Stat. Assoc. 90 (429), 64–71.
Smith, K.S., Rayer, S., 2011. An Evaluation of Population Forecast Errors for Florida and its Counties, 1980–2010. Special Population Reports, Bureau of Economic and Business Research, Warrington College of Business Administration, University of Florida. <http://www.bebr.ufl.edu/sites/default/files/population/SPR_9.pdf>.
Smith, K.S., Tayman, J., 2003. An evaluation of population projections by age. Demography 40 (4), 741–757.

Srinivasan, S., Ma, L., Yathindra, K., 2008. Procedure for Forecasting Household Characteristics for Input to Travel-Demand Models. Project Report of University of Florida, Florida Department of Transportation, TRC-FDOT-64011-2008. <http://www.fsutmsonline.net/images/uploads/reports/FDOT_BD545_79_rpt.pdf>.

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

Swartz, P.G., Zegras, P.C., 2013. Strategically robust urban planning? A demonstration of concept. Environ. Plann. B: Plann. Des. 40, 829–845.

Tayman, J., 1996. The accuracy of small-area population forecasts based on a spatial interaction land-use modeling system. J. Am. Plann. Assoc. 62 (1), 85–98.

Voas, D., Williamson, P., 2000. An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. Int. J. Popul. Geogr. 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. Environ. Plann. A 30 (5), 785–816.

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: Presented at the 88th Annual Meeting of Transportation Research Board, Washington DC.

Zhu, X., Mishra, S., Welch, T.F., Pandey, B., Baber, C.M., 2013. A framework for modeling and forecasting population age distribution in metropolitan areas at transportation analysis zone level. In: Presented at the 92th Annual Meeting of Transportation Research Board, Washington DC.

Zhu, Y., Joseph, F.J., 2014. Synthetic population generation at disaggregated spatial scales for land use and transportation microsimulation. Transport. Res. Rec. 2429, 168–177.