Using Small Area Estimation to Produce Reliable Transportation Statistics: The Case of Household Trips Estimation at The Census Tract Level

Mohammad B. Al-Khasawneh
University of Maryland, College Park
Cinzia Cirillo (ccirillo@umd.edu)
University of Maryland, College Park

Research Article

Keywords: Small Area Estimation, Linked datasets, Transportation statistics, Household person trips, Census tract

Posted Date: February 14th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3920649/v1

Additional Declarations: No competing interests reported.

Using Small Area Estimation to Produce Reliable Transportation Statistics: The Case of Household Trips Estimation at The Census Tract Level

Mohammad B. Al-Khasawneh
University of Maryland
Department of Civil and Environmental Engineering
A. James Clark School of Engineering
3246 Kim Engineering Building
College Park, MD 20742
United States of America
Tel: 740-331-9916; Email: mbmk2020@umd.edu

Cinzia Cirillo*
University of Maryland
Department of Civil and Environmental Engineering
A. James Clark School of Engineering
3250 Kim Engineering Building
College Park, MD 20742
United States of America
Tel: 301-405-6864; Email: ccirillo@umd.edu

*Corresponding author

ABSTRACT

This paper proposes Small Area Estimation (SAE) methods on linked datasets to generate reliable transportation statistics in cases where data on travel behavior is limited or missing. Specifically, household person trips are estimated at the census tract by linking data from the Regional Travel Survey (RTS), the American Community Survey (ACS), and US Census 2020 data. The proposed SAE modeling framework integrates direct and synthetic estimations to produce accurate statistics. Several small area estimation techniques have been employed, including regression-based models and population synthesis for areas with zero samples, as well as the Fay-Herriot model for areas with small samples. For the regression-based models, we assessed several models, including linear, Poisson, negative binomial, and random forest models, using cross-validation analysis. The Fay-Herriot method is also applied to improve estimation precision by combining direct and synthetic estimation approaches. Results showed the proposed methodology's effectiveness in generating reliable estimates in both cases of missing or limited samples. The research highlights the potential of SAE methods in enhancing transportation analysis by integrating diverse datasets and reducing the survey data collection burden. These findings have practical implications for researchers, policymakers, and transportation planners seeking reliable estimates for smaller domains and subgroups using existing data sources.

Keywords: Small Area Estimation; Linked datasets; Transportation statistics; Household person trips; Census tract

INTRODUCTION

Traditional surveys for official statistics are primarily designed to generate reliable estimates at the macro level, such as the national, regional, and state levels. However, they often lack the necessary data for producing reliable estimates in smaller areas, such as census tracts and subpopulations. The reliability of estimated variables is positively correlated with sample size (Wilson et al., 2018). As the sample size increases, our uncertainty regarding the collected information decreases, allowing for more reliable estimates. However, increasing the sample size for smaller domains requires additional effort, time, and financial resources. Moreover, privacy regulations can further complicate access to necessary information in some cases. Researchers and policymakers often seek to answer specific questions by obtaining reliable estimates for smaller domains and subgroups using available data, without the need for additional data collection. The National Travel Survey (NHTS) and the American Community Survey (ACS) serve as notable examples of available data sources. To provide cost-effective solutions, the Center for Statistical Research & Methodology at the United States Census Bureau (USCB) has been exploring new methods to meet user demands for statistical estimates in smaller domains and various subpopulations. One such solution is the utilization of Small Area Estimation (SAE) methods, which integrate different data sources to augment existing information. SAE techniques refer to a range of methods that produce reliable estimates when sample sizes are insufficient or unavailable.

SAE techniques are particularly useful for producing reliable estimates in small geographical areas, such as counties, municipalities, census tracts, or specific subpopulations of interest, such as age-sex-ethnic groups, within larger areas (Ghosh & Rao, 1994). Surveys conducted at the national or state level often lack sufficient data to generate direct estimates for these smaller areas. Therefore, in such cases, additional datasets containing relevant information for the specific areas of interest can be utilized to obtain reliable estimates (Heady et al., 2003). In this research study, we propose a methodological framework for estimating the total number of daily household trips for different purposes (subdomains) in Maryland at the census tract level. This is achieved by linking three datasets: the Regional Travel Survey (RTS), the American Community Survey (ACS), and the Census 2020 data. The remaining of this paper is organized as follows. Section 2 provides a review of SAE applications, with a focus on transportation-related studies. Section 3 describes the different data sources utilized in this study. In Section 4, we present the proposed methodology, while Section 5 summarizes the obtained results. Finally, in Sections 6 and 7, we discuss the implications of our findings and present the conclusion.

2. LITERATURE REVIEW

Small area estimation (SAE) is a statistical technique used to estimate parameters for small subpopulations when direct estimates are unreliable due to small sample sizes (Rao & Molina, 2015; Jiang & Rao, 2020). The term "small area" can refer to geographically small areas, such as a neighborhood or a county, or to any subpopulation group with a small sample size, such as a specific age group or ethnic group. SAE methods can be broadly categorized into two types: direct estimation and indirect estimation (Zhongming et al., 2020). Direct estimation is the simplest and most straightforward approach, utilizing only the primary data source and survey weights to generate estimates within the domain of interest. However, when sample sizes are small, direct estimates can be highly imprecise.

Indirect SAE methods, on the other hand, incorporate additional information, known as auxiliary data, to improve the precision of estimates. Auxiliary data can include information from administrative records, censuses, or other surveys. Indirect SAE methods can be further categorized into two types: (1) SAE methods using auxiliary information and (2) SAE methods using regression-based models. One example of an SAE method using auxiliary information is Survey Weight Reallocation, employed by Martinez (2014) and demonstrated that a survey reweighting method offers a simpler and more direct approach for estimating poverty statistics at the sub-domain level in developing countries. Other SAE techniques using auxiliary information, namely Broad Area Ratio Estimator (BARE) and the Calibration method, are commonly used. Additional details on these techniques can be found in (Zhongming et al., 2020)

Small area estimation (SAE) using regression-based models involves statistical techniques that utilize various regression models, including linear, non-linear, Poisson, or logistic regression. Compared to other SAE techniques, regression-based SAE is a complex technique that requires both primary and secondary datasets, as well as expertise in model selection (Pfeffermann, 2013). It is used to estimate the values of a variable of interest for small areas when direct estimates are unreliable due to the small sample size. The technique involves fitting a regression model to the primary dataset and then using the model to predict estimates for the small areas using the auxiliary dataset. The goal is to obtain estimates with a lower coefficient of variance, making them more reliable (Goerndt et al., 2011; Moretti & Whitworth, 2020). The effectiveness of regression-based SAE depends on the relationship between the variable of interest and the auxiliary variables in the auxiliary dataset. These regression-based models can be applied at either the area level, linking area means to area-specific auxiliary variables, or the unit level, linking unit values to unit-specific auxiliary variables, depending on data availability (Chatrchi, 2019). In the regression process, the primary dataset is always the only source for the dependent variable, regardless of whether the model is at the unit or area level. For independent variables, however, the situation differs between unit-level and area-level models. For unit-level models, the only source for independent variables is the primary dataset. In contrast, for area-level models, independent variables can be drawn from either the primary dataset or a secondary dataset. The output estimates of the regression-based techniques are known as synthetic estimates. When no sample data exists for an area of interest, synthetic estimates, generated by regression-based techniques, are the common way to obtain estimates. However, if direct estimates are available, another method called Empirical Best Linear Unbiased Prediction (EBLUP) offers a more effective approach. The EBLUP estimator is a weighted combination of the direct estimator and the regression-synthetic estimator. It balances the bias of synthetic estimates and the high variance of the direct estimates. It assigns more weight to the direct estimates when their variance is small. Conversely, when the direct estimates are unreliable due to high variance, more weight is given to the synthetic estimator. The EBLUP estimator is also known as the Fay-Herriot model, which was initially implemented by Fay III & Herriot (1979) for use in area-level SAE to estimate the per capita income of the United States population.

Another comprehensive unit-level method is synthetic population generation. Unlike other SAE methods, which focus on a single variable of interest, synthetic population considers multiple variables simultaneously. This method aims to create a detailed population for every unit in the target population, generating populations that share the same structure and characteristics as the real populations (Gargiulo et al., 2010). Several approaches for synthetic population generation exist, with the most common method being Iterative Proportional Fitting (IPF). IPF utilizes survey samples and census data to create synthetic populations. IPF works by iteratively adjusting the weights assigned to each unit within a non-spatial individual-level survey dataset until the desired fit to the target aggregate constraints is achieved (Lovelace et al., 2015). However, IPF has limitations, including integer conversion issues, zero cells, and synthesizing at only one level (Choupani & Mamdoohi, 2016). Alternative approaches for synthesizing populations using statistical methods are outlined in (Saadi et al., 2016; Sun & Erath, 2015), while discussions on the application of machine learning models for this purpose can be found in (Borysov et al., 2019). Lin (2023) implemented a method developed by (Lin & Xiao, 2022) to generate synthetic individual data across the USA with characteristics like age, sex, race, and Hispanic/Latino origin, based solely on aggregated 2010 Census data at block group/tract levels.

SAE technique has been successfully applied across diverse fields, including poverty, employment, health, and transportation. In the context of poverty estimation, the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program stands as a prime example, providing valuable insights into the prevalence of poverty among school-age children at the school district, county, and state levels. This program aims to deliver updated income and poverty data, guiding federal program administration and fair funding allocation across local areas. The World Bank (WB) is a notable example, has been using a method developed by Elbers et al. (2003) that integrates household surveys and census data to produce poverty and income inequality estimates for different countries worldwide.

Moving to the domain of employment, SAE has proven instrumental in estimating labor force dynamics at granular levels. A major example of small-area estimation initiatives in the United States is the Local Area Unemployment Statistics (LAUS) program conducted by the Bureau of Labor Statistics. This program generates monthly and annual employment and unemployment estimates for states, metropolitan areas, counties, and specific sub-county regions. For detailed information, you can visit the LAUS website at www.bls.gov/lau/. Further examples outside the United States can be found in López-Vizcaíno et al., (2015) and Bertarelli et al., (2018) who demonstrate SAE's versatility in different contexts. López-Vizcaíno et al. (2015) used SAE to estimate the employed, unemployed, and inactive populations in Galicia, Spain, leveraging data from the Spanish Labor Force Survey (EPA). Their findings showcase SAE's power in providing detailed labor market insights. Bertarelli et al. (2018) pushed the boundaries further, employing SAE to obtain reliable labor force statistics for local Labor Market Areas (LAMs) in Italy. Their work underscores SAE's adaptability to diverse geographic units.

Within the healthcare sector, SAE has played a crucial role in analyzing health indicators at small geographic scales. Take substance abuse, for example. The Substance Abuse and Mental Health Services Administration (SAMHSA) uses SAE to estimate rates for states and metropolitan areas, helping paint a clearer picture of this complex issue. For more detailed information, you can visit http://www.samhsa.gov/. Another application is presented by Li et al., (2009), who employed SAE to study the temporal and geographic variations in community-level obesity prevalence in 398 communities in Massachusetts. Their study effectively combined individual-level data from the Behavioral Risk Factors Surveillance System (BRFSS) survey with community-level data from the 2000 Census Data.

In the transportation domain, SAE has proven to be a valuable tool for estimating transportation-related statistics at smaller geographic scales. Long et al., (2009) employed SAE to estimate the total number of workers per household in Des Moines at the individual and census tract levels using data from the Census Transportation Planning Package (CTPP) and the 2001 NHTS and concluded that SAE is a promising alternative approach for predicting travel information. Their study compared three SAE models: generalized regression estimator method (GREG), empirical best linear unbiased predictor method (EBLUP), and EBLUP without random area effects. (Vaish et al., (2010) utilized the survey-weighted hierarchical Bayes (SWHB) methodology to estimate the percentages of making a high mileage trip per day among different age groups in all U.S. States. Goulias et al., (2014) developed a synthetic population approach to address missing information at smaller geographical scales. This approach was applied using a sample dataset from the California Household Travel Survey (CHTS) and Census Data. Where, Salon (2016) employed the Broad Area Ratio Estimator to estimate the total daily miles traveled by walking and bicycling for all census tracts in California around 2010. Two household-based travel diary surveys were utilized: the 2009 National Household Travel Survey (NHTS) and the 2010-2012 California Household Travel Survey (CHTS). Our study extends the application of Small Area Estimation (SAE) in the field of transportation, demonstrating the effectiveness of SAE through a more comprehensive methodological framework. This framework incorporates three distinct SAE techniques: regression-based models, a composite estimator represented by the FH model, and synthetic population generation. Additionally, the study integrates data from three sources: the American Community Survey (ACS), US census data, and the Regional Travel Survey (RTS). The goal of this integration is a trip generation model to produce accurate and reliable estimates for the total number of daily trips, categorized into six purposes that correspond to those defined in the Maryland Statewide Transportation Model (MSTM), covering all census tracts in Maryland.

3. DATA SOURCES

The empirical analysis proposed is based on several data sources that are linked to overcome the deficiencies of a single dataset and leverage their collective strengths. These are the US Decennial Census data, the American Community Survey, and two Regional Travel Surveys conducted in the State of Maryland. The study refers to two geographical levels: PUMA (Public Use Microdata Area), which is a geographic unit defined by the U.S. Census with at least 100,000 people, and census tracts which are smaller units within a PUMA, typically with a population size of 1200 to 8000 people. The sub-sections below provide more detailed information about the three datasets.

U.S. Decennial Census Data

The United States Census Data is conducted every ten years (i.e., in 2000, 2010, and 2020) and provides total counts of a limited number of socio-demographic variables at various geographic scales. However, the U.S. census data available to the public does not offer household or person microdata. It only provides summary data. For this research study, the 2020 data at census tract level (2020 US Census) was utilized to generate cumulative distribution functions (CDFs) for the variables of interest, which were then used to synthesize the relevant information at the census tract level.

American Community Survey (ACS)

The ACS was selected as a secondary dataset for this analysis due to its relevance in capturing essential household factors aligned with the RTS and its comprehensive coverage of the entire State of Maryland. ACS is an extensive nationwide household survey conducted annually by the United States Census Bureau, collecting demographic, economic, and social information for millions of individuals across the country at the PUMA level. The ACS offers two versions of their sample datasets: the 1-year and the 5-year samples. The 5-year sample is chosen for this analysis because it provides a more comprehensive, accurate, and representative depiction of the population. In order to align with the time frame of the RTS data, which was collected across 2017 and 2018, we utilized the five-year dataset (ACS 2014-2018) for our analysis. However, it's important to note that the ACS does not include information on household travel behavior, except for commuting trips, or offer data at the census tract level. To address this limitation and obtain information at the census tract level, we incorporated the U.S. Census data into our study. This additional dataset fulfills the need for detailed information at the census tract level, complementing the micro-level data provided by the ACS for households.

Regional Travel Survey (RTS)

The study used two Regional Transportation Surveys (RTSs) conducted by the Washington Metropolitan Council of Government (MWCOG) and the Baltimore Metropolitan Council (BMC) respectively between 2017 and 2018. Both surveys had similar designs and data collection methods, covering some parts of Maryland (as shown in Figure 1 and Table 2). When combined, they cover all of Maryland. However, they don't have enough samples at the census tract level to produce reliable estimates. It's important to note that during the initial stage of the study, the authors only had access to RTS 2018 related to the MWCOG region. The RTS 2018 related to the BMC region became available at a later stage. It was decided to use RTS MWCOG for estimation and RTS BMC for validation. RTS MWCOG is the primary data source for this study, providing valuable information on household characteristics, daily travel patterns, modes of transportation, and trip purposes. The data was collected from approximately 16,000 random households in the District of Columbia, suburban Maryland, and northern Virginia, at various geographic levels such as PUMA (Public Use Microdata Area) and census tract. We merged household and trip data to create records for each household's characteristics and daily trips at both PUMA and census tract levels. In Maryland, there are 44 PUMAs and over 1400 census tracts; however, the RTS MWCOG survey covers only 25 PUMAs and about 750 census tracts in the suburban part of Maryland. During the advanced stages of our research, RTS BMC data, which covers the remaining parts of Maryland, became available and was used for validation only.

We carefully selected a set of dependent variables for this study, considering their availability in both datasets (RTS 2018 and ACS 2018) and their relevance to the independent variable (i.e., total household trips). The variables we chose include the number of individuals in the household (NP), the number of workers in the family (WIF), the number of cars owned by the household (VEH), and the household income (HINCP). Table 1 presents the descriptive statistics of these selected variables. To reinforce the importance of these variables in trip generation, we can refer to relevant studies that have examined the relationship between these variables and travel behavior. Many scholarly works have explored the impact of household characteristics, such as the number of individuals, workers, cars, and income on travel patterns, emphasizing their role in trip generation modeling (Badoe & Steuart, 1997; Hu, 2010; Mukherjee & Kadali, 2022; Rashidi et al., 2010).

TABLE 1 Descriptive Statistics of Selected Variables

RTS (MWCOG)							
Independent Variables	**Min**	**Pctl.25**	**Median**	**Pctl.75**	**Max**	**Mean**	**Std. Dev.**
HINCP	1	4	5	7	8	5.113	2.056
NP	1	1	2	3	7	2.395	1.331
WIF	0	0	1	2	3	1.198	0.901
VEH	0	1	2	2	4	1.834	1.041
Dependent Variable							
Total Trips	1	4	7	11	74	8.123	5.89
RTS (BMC)							
Independent Variables	**Min**	**Pctl.25**	**Median**	**Pctl.75**	**Max**	**Mean**	**Std. Dev.**
HINCP	1	5	7	8	8	6.191	1.776
NP	1	2	2	4	7	2.753	1.436
WIF	0	1	1	2	3	1.392	0.871
VEH	0	1	2	3	4	1.989	0.951
Dependent Variable							
Total Trips	1	4	7	12	44	8.994	6.445
ACS							
Independent Variables	**Min**	**Pctl.25**	**Median**	**Pctl.75**	**Max**	**Mean**	**Std. Dev.**
HINCP	1	5	6	8	8	5.887	1.989
NP	1	2	3	4	7	3.471	1.562
WIF	0	1	2	2	3	1.68	0.849
VEH	0	1	2	3	4	2.099	1.039

*HINCP: 01=Less than $15,000; 02=$15,000 to $24,999; 03=$25,000 to $34,999; 04=$35,000 to $49,999; 05=$50,000-$74,999; 06=$75,000-$99,999; 07=$100,000-$149,999; 08=$150,000 or more.

For a better understanding of the study area and the overlapping coverage of the selected datasets, please refer to Figure 1 and Table 2. Figure 1 displays all of the census tracts in Maryland grouped into either Group 1 or Group 2. These groups are then cross-referenced in Table 2, which provides information on the coverage area of each dataset, the finest spatial unit available, and the available information.

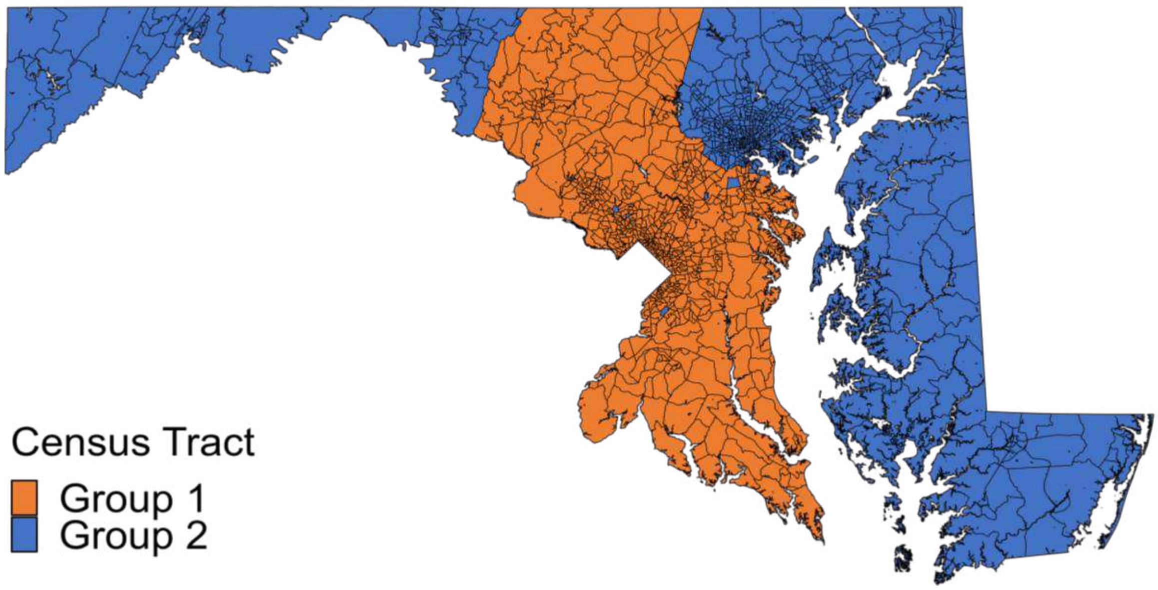


Figure 1: Geospatial Overview of Maryland Census Tracts: Group 1 and Group 2

Table 2: Summary of Selected Datasets: Spatial Characteristics and Coverage Areas

Dataset	**RTS MWCOG 2018 (Primary)**	**ACS 2018 (Auxiliary)**	**US Census 2020 (Auxiliary)**	**RTS BMC 2018 (Auxiliary)**
Coverage (Figure 1)				
Sample vs. Complete Count	Sample	Sample	Complete Count	Sample
Provided Information	Demographics+ travel info	Demographics only	Demographics only	Demographics + travel info
Reliable Estimates	PUMA level	PUMA level	Census tract level	PUMA level
Used in	Model Estimation &Validation	Model Estimation & Synthetic Population		Validation

4. METHDOLOGY

This study estimates the total number of households trips per different trip purposes: Home-Based Work (HBW), Home-Based School (HBSCH), Home-Based Shopping (HBS), Home-Based Other (HBO), Non-Home-Based Work (NHBW), and Non-Home-Based Other (NHBO) aggregated at the census tract level. The trip purposes have been defined consistently with the Maryland Statewide Transportation Model (MSTM), a trip based four step model currently available for the State of Maryland and maintained by the Maryland State Highway Administration. The estimates are obtained by combining information from three sources of data described in Section 3.

SAE methods are adopted to address two fundamental challenges. First, the RTS MWCOG 2018 provides information on the variable interest at the census tract level, but it does not cover all areas in the State of Maryland. Also, the sample size for the subdomain of interest (the total number of trips per different trip purposes) at the census tract level is too small to make reliable direct estimates. Second, the ACS 2018 covers the entire state of Maryland with various demographics and socioeconomic characteristics of households, but it does not have information about the total number of households at census tract level. Also, the public version of ACS 2018 provides location related information at PUMA level and not at the census tract level.

We start addressing the identified challenges associated with ACS 2018 by combining the ACS 2018 sample with U.S. census data to synthesize the population at the census tract level. By doing so, we were able to obtain disaggregated data for every household in Maryland and their census tract location. However, we still lack information about their daily trips. To address this gap, we propose using the RTS MWCOG 2018 data to develop a model that predicts the total number of household trips. By applying the estimated model to the created synthetic population, we can generate synthetic estimates for the variable of interest in regions with no samples.

The SAE methodological framework employed in this study can be outlined in the following steps and are illustrated in (**Figure 2**):

Step 1 (Synthetic Population): Generate a synthetic population at the census tract level by synthesizing Maryland household information using a sample dataset from ACS 2018 and 2020 U.S. census data. This step is crucial as the synthetic population needs to be statistically representative of the actual population to proceed to the next step.

Step 2(Model Selection): Test and evaluate various models to estimate the total number of household trips for different trip purposes at the census tract level, using the information available in RTS MWCOG 2018. This step involves examining multiple models with diverse scenarios and cases to determine the most suitable ones. Both unit-level and area-level models are considered, leveraging RTS MWCOG 2018 data and the synthetic population as sources for dependent variables. By thoroughly testing different models, we ensure a comprehensive analysis and accurate estimation of household trips.

Step 3 (Synthetic Trip Estimation): Utilize the selected best-performing models within the context of the created synthetic population to generate synthetic estimates of total household trips in census tracts with zero sample. Additionally, employ a composite estimator to enhance trip estimates for areas with small sample size.

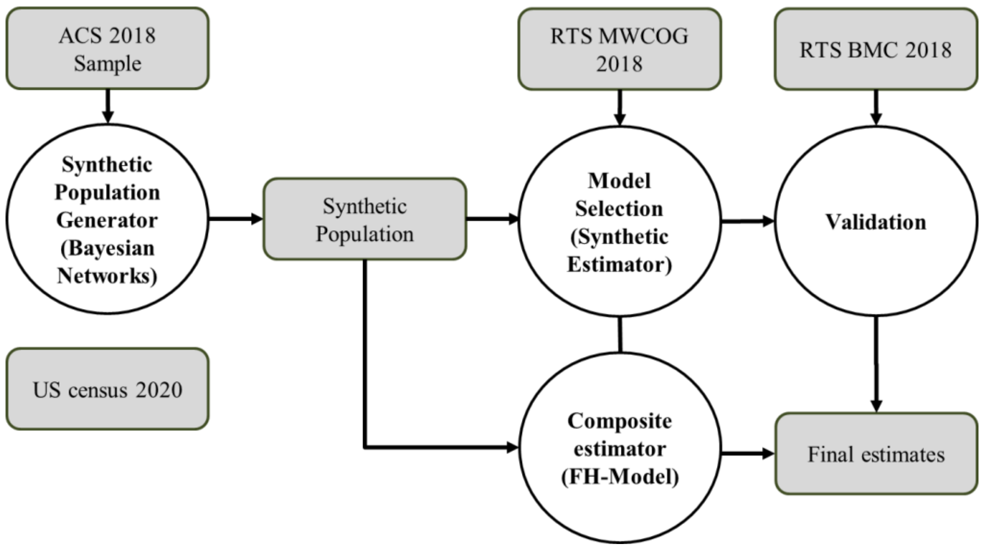


Figure 2: Methodological Framework

Synthetic Population

The existing PUMS sample lacks data at the census tract level, which is crucial for our analysis. While the US census itself provides tract-level estimates, these are insufficient for our needs. Therefore, we require disaggregated data for each individual unit at the census tract level within the population to build our unit-level models, which are currently unavailable in the PUMS data. In addition, these tract-level estimates are based on the IPF technique, which has limitations, so we aim to produce more reliable and updated ones. Finally, our aim is to develop a comprehensive and more general methodological framework that can be replicated by others in similar situations.

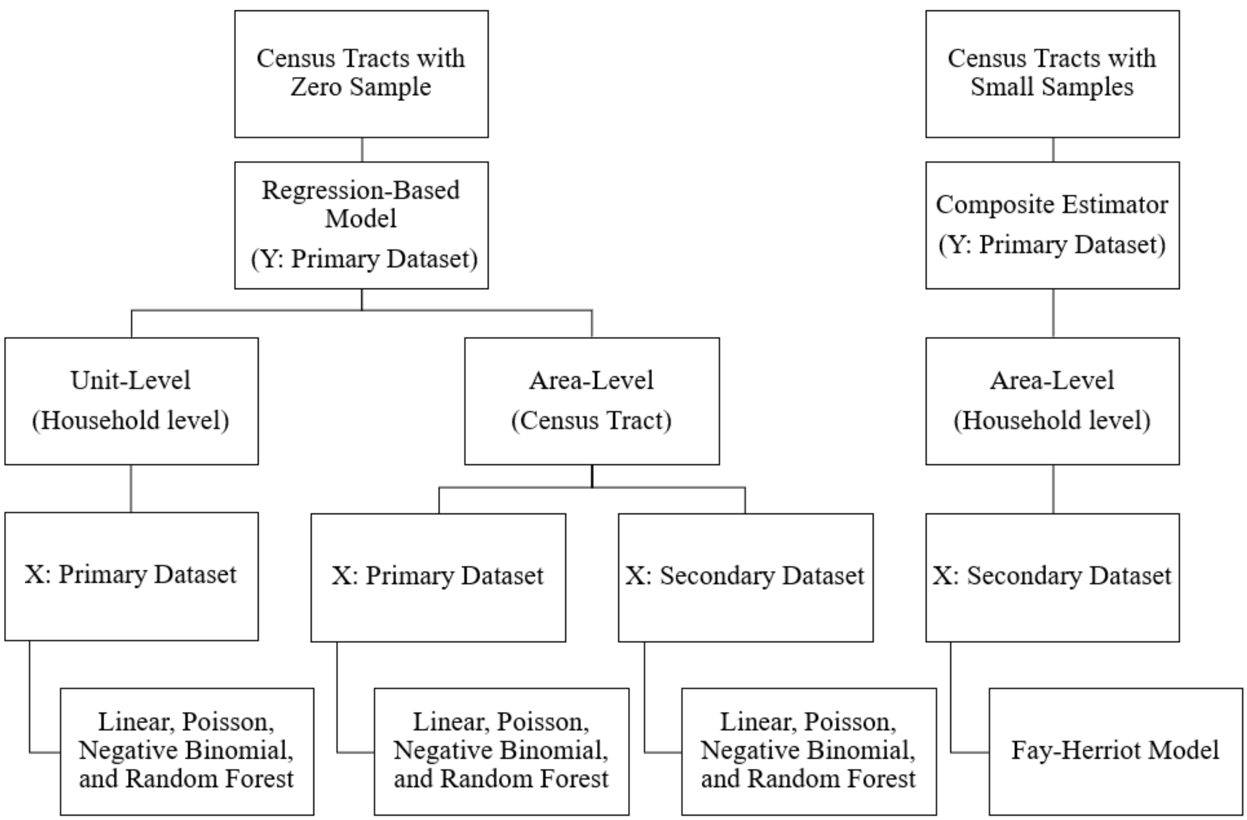
We combined the ACS 2018 sample with 2020 U.S. census data to generate a disaggregated synthetic population of households in Maryland at the census tract level. While various approaches exist for synthetic population generation, we chose to use a Bayesian Network (BN) method in this study. A Bayesian Network is a probabilistic model that defines variables and their conditional relationships through directed acyclic graphs (DAG) (Frick, 2004). Nodes and directional links in the DAG represent variables and their dependencies, with each node carrying the probability under a given condition.

Sun & Erath (2015) previously used a BN to create a synthetic population for Singapore, demonstrating superior performance compared to other synthesizers like Iterative Proportional Fitting (IPF) and the Markov chain Monte Carlo algorithm (MCMC), particularly in addressing overfitting, heterogeneity, and missing data. Moreover, the BN approach has proven effective in generating disaggregated data from traditional survey data or open data for transportation models (Hörl & Balac, 2021; Zhang et al., 2019).

Our synthetic population simulation for Maryland's households at the census tract level considered household income, household size, number of workers, and number of vehicles as dependent variables. To achieve this, we transformed the sample population into pseudo-observations for a given Public Use Microdata Area (PUMA) by considering the empirical cumulative distribution function (ECDF) for each variable. Subsequently, we utilized pseudo-observations to learn the dependency structure among the selected variables using Bayesian Network learning techniques. Using the constructed network, we generated synthetic pseudo-observations for each census tract within the selected PUMA. In the final step, we transformed synthetic pseudo-observations into a synthetic population by applying the inverse CDF sampling method, using the ECDF obtained from U.S. census data. For more details on the BN method applied in this paper, refer to (Jutras-Dubé et al., 2023).

Model-based Estimates (SAE Method)

After generating a reliable synthetic population that incorporates all the necessary independent variables at the census tract level, the subsequent step involves integrating it with the available information from RTS 2018 and identifying the most effective models for describing the total number of trips based on various trip purposes. **Figure 3** summarizes the model selection process, where Y refers to the source of dependent variable, and X refers to the source of



independent variables.

Figure 3: Model Selection Process

A) Census Tracts with Zero Sample:

We used four different models: Linear, Poisson, Negative Binomial, and Random Forest, to estimate the total number of household trips and trip rates for the purposes of interest (HBW, HBSHOP, HBS, HBO, NHBW, and NHBO). We evaluated them at two levels: the unit level with one source of auxiliary variables (RTS MWCOG 2018) and the area level with two sources of auxiliary variables (RTS MWCOG 2018 and Synthetic Population). This resulted in a total of 12 models per trip purpose, totaling 72 models overall. To determine the best-performing model, we conducted a comprehensive evaluation using a cross-validation method.

For model estimation, in the unit-level models, we use RTS MWCOG 2018 data only for the independent variables to predict the total number of trips for each household unit then we aggregate unit predictions to the census tract level. In area-level models, we test alternative models with two sources for the independent variables: RTS MWCOG 2018 and the synthetic population. We aggregate total estimates for the independent variables at the census tract level. Then, we use them to predict the total number of household trips for each variable of interest aggregated at the census tract level. Finally, for model application, we apply the estimated model to the synthetic population to make predictions in unit-level and area-level models for the areas with zero sample.

Then, the best-performing model $\boldsymbol{f}$ is applied to the synthetic population $\boldsymbol{X_i}$ to produce synthetic estimates $\widehat{\boldsymbol{Y}}_{\boldsymbol{i_{NS}}}^{\boldsymbol{S}}$ for census tracts with missing information.

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{i_{NS}}}^{\boldsymbol{S}} = \boldsymbol{f}(\boldsymbol{X_i}) \tag{1}$$

A) Census Tracts with Small Samples:

For the census tracts covered by the RTS MWCOG 2018, we can obtain direct estimates for the variables of interest. However, the associated coefficients of variance tend to be high in some areas and even higher over the specified subdomains of interest. Therefore, we considered the empirical best linear unbiased prediction (EBLUP) using the Fay-Herriot model (*34*) to improve direct survey estimates of the total number of trips.

The Fay-Herriot model is a composite estimator $\widehat{Y}_i^S$ of the direct estimator $\widehat{Y}_i^D$ and the synthetic estimator $X_i^T\beta$ as the following:

$$\widehat{Y}_i^S = \omega_i\widehat{Y}_i^D + (1-\omega_i)X_i^T\beta \tag{2}$$

Where ω_i is the weight of the direct estimator, ranging from 0 to 1. The direct estimator weight is calculated using the variances of the model $\widehat{\sigma}_u^2$ and the sampling errors $\widehat{\sigma}_e^2$:

$$\omega_i = \frac{\widehat{\sigma}_u^2}{\widehat{\sigma}_u^2+\widehat{\sigma}_{e_i}^2} \tag{3}$$

The Fay-Herriot model aims to minimize the mean square error in the final estimator of the parameter by giving more weight to the estimator with less variance to produce more reliable estimates. In 1990, Prasad and Rao provided the following formula, which can be used as an approximation for the unbiased estimator of the mean square error of the Fay-Herriot estimator.

$$\text{MSE}\ (\hat{Y}_i^S) = \omega_i \hat{\sigma}_{e_i}^2 + (1-\omega_i) X_i^T \left[\sum \frac{x_i x_i^T}{\hat{\sigma}_u^2 + \hat{\sigma}_{e_i}^2}\right]^{-1} X_i + \frac{\hat{\sigma}_{e_i}^2}{(\hat{\sigma}_u^2 + \hat{\sigma}_{e_i}^2)^3} \frac{4\sum\left(\hat{\sigma}_u^2 + \hat{\sigma}_{e_i}^2\right)^2}{n^2} \quad (4)$$

Evaluation Methods

We evaluated the regression models using a cross-validation method and assessed their performance based on the Mean Absolute Percentage Error (MAPE) and accuracy (ACC).

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{A_i - F_i}{A_i} \right| \quad (5)$$

where, A_t = actual value; F_t = forecast value; n = number of observations.

The accuracy was defined as the following:

$$Accuracy = \text{Max}\ (0, 1 - \text{MAPE}) \quad (6)$$

Moreover, we used the coefficient of variance (CV) to evaluate the performance of the Fay-Herriot model:

$$CV = \frac{StandardError\,(SE)}{Estimates} \quad (7)$$

The standard SE error is the square root of the mean absolute error as the following equation:

$$SE = \sqrt{MSE = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2} \quad (8)$$

Where, Yi = actual values; $\hat{Y}_i$ = predicted values

5. RESULTS

Synthetic Population Generation and Validation

As mentioned earlier, the initial stage of our methodological framework involves generating a synthetic population for Maryland at the census tract level. This step is crucial to ensure that the constructed population accurately represents the actual population in terms of the selected independent variables. To validate the accuracy of our synthetic population in each respective region, we utilized household samples from RTS 2018, which were available for those regions.

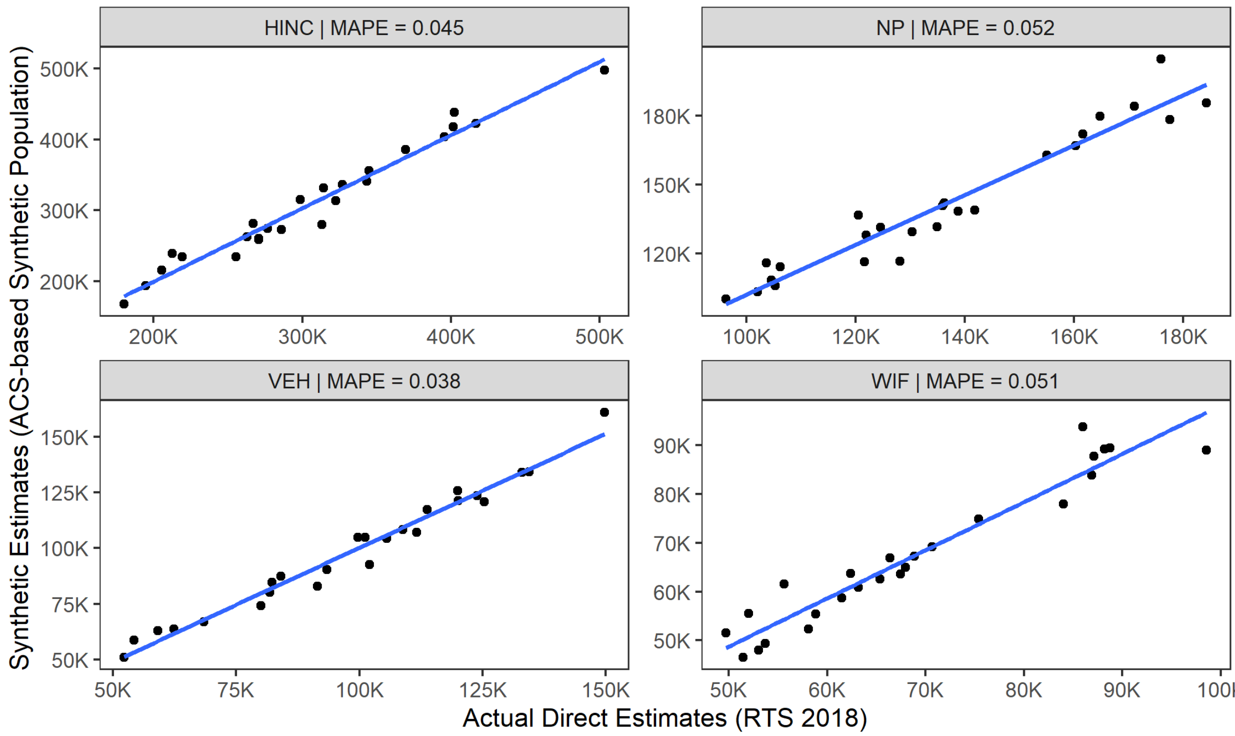


Figure 4: Comparison of Synthetic Population and Actual Population (RTS 2018)

First, we aggregated the results from RTS MWCOG 2018 at the census tract level and found that the coefficient of variance (CV) values, on average, were higher than 40%. These high CVs indicate that the estimates at the census tract level are not statistically reliable enough to be compared with the synthetic estimates (Aronhime et al., 2014). Therefore, we validated the synthetic population at the PUMA levels where the CVs were found to be less than 10%, which can be considered acceptable (Aronhime et al., 2014). Table 2 illustrates the average values of the associated CVs for total estimates from RTS 2018 at the PUMA and census tract levels respectively.

TABLE 3 CV Comparison for Total Estimates in Maryland (PUMA vs. Census Tract)

Geographic Level	CV%_WIF	CV%_VEH	CV%_NP	CV%_HINC
Census tract	49.29	44.85	47.98	42.66
PUMA	9.32	8.51	9.94	7.61

The results from validation illustrated in (**Figure 4**) demonstrate a very good match between the actual and synthetic populations, with an average MAPE of less than 5%. This indicates that a reliable synthetic population was obtained with the BN method.

Regression-Synthetic Estimators

We employed the k-fold cross-validation procedure to test various alternatives of unit-level and area-level models for estimating the six variables of interest at the census tract level. The dataset was divided into 25 folds using the PUMA criteria. The analysis was repeated 25 times, with each run excluding one PUMA for validation while training the model on the remaining 24 PUMAs. To determine the overall accuracy, we calculated the mean accuracy from each run. For unit-level models, each fold of the PUMA contained household observations. In contrast, area-level models utilized aggregated estimates at the census tract level in each fold. The results of accuracy for all the tested models are summarized in (**Table 4**).

TABLE 4 Comparison of Accuracy Results for Tested Unit-level and Area-level Models

Unit Level (RTS-AUX)	HBO	HBW	NHBO	NHBW	HBSHOP	HBS	Overall
Linear	0.829	0.872	0.845	0.844	0.874	0.737	0.847
Poisson	0.819	0.875	0.842	0.846	0.874	0.666	0.835
Negative Binomial	0.801	0.876	0.833	0.849	0.874	0.218	0.767
Random Forest	0.856	0.870	0.854	0.834	0.878	0.769	0.858
Area Level (RTS-AUX)							
Linear	0.841	0.888	0.877	0.846	0.875	0.710	0.853
Poisson	0.838	0.848	0.865	0.832	0.871	0.740	0.842
Negative Binomial	0.838	0.830	0.868	0.820	0.861	0.740	0.833
Random Forest	**0.861**	**0.888**	**0.873**	**0.846**	**0.877**	**0.725**	**0.859**
Area Level (SYN-BN-AUX)							
Linear	0.849	0.877	0.882	0.844	0.867	0.727	0.854
Poisson	0.832	0.878	0.881	0.841	0.870	0.715	0.849
Negative Binomial	0.833	0.868	0.875	0.831	0.854	0.715	0.840
Random Forest	0.859	0.865	0.855	0.844	0.864	0.718	0.847

Overall, the differences in accuracy among the tested models are generally insignificant. However, it is worth noting that the unit-level negative binomial model exhibited relatively low accuracy in predicting HBO trips. On the other hand, the results indicate that the Random Forest area-level model, using RTS MWCOG 2018 as the source of the dependent variables, performed the best among all other tested models, demonstrating a mean accuracy of 0.859. This was the model chosen for model application to areas with no sample.

Composite Estimator (Fay-Herriot Model)

Many census tracts covered by the RTS MWCOG 2018 lack sufficient observations to provide reliable estimates. Additionally, estimating the total number of trips for six different trip purposes has become more challenging due to a decrease in reported cases. The current direct estimates at the census tract level exhibit high coefficient of variation (CV) values, with an average exceeding 30%. To address this issue, we employed the Fay-Herriot model (FH) to enhance the direct estimates in areas with small sample sizes. The FH model yielded composite estimates of total trips that demonstrate lower CVs. (**Figure 5**) illustrates the average CV values for the total number of trips obtained directly from RTS 2018 and synthetically estimated using the FH model.

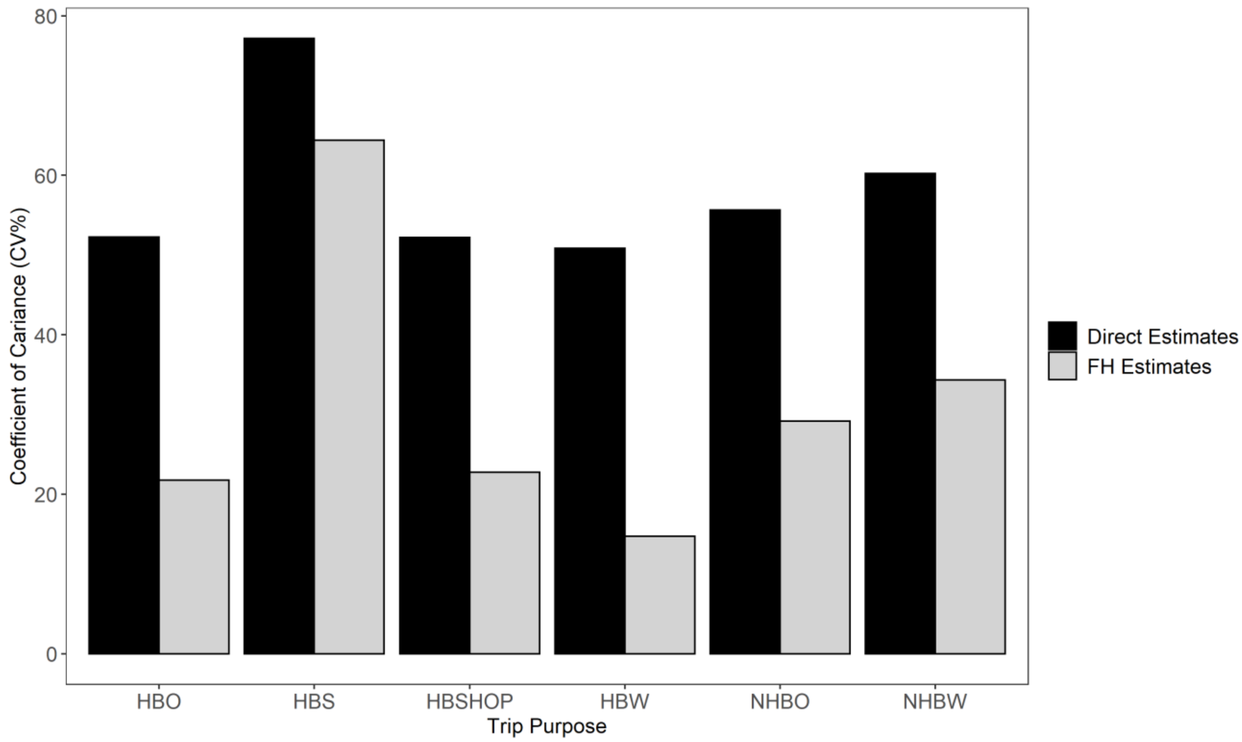


Figure 5: Comparison of Coefficient of Variation Values for Direct and Synthetically Estimated Total Trips Using the Fay-Herriot Model

The Fay-Herriot model has produced estimates with lower coefficients of variation (CV) for all the variables of interest. However, the CV associated with the FH model for HBS trips remains high. This can be attributed to the very low number of reported HBS trips by the sampled households in RTS 2018, which couldn't be significantly improved by the Fay-Herriot model.

Final Estimates of Total Trips

The overall results relative to the total number of household trips in Maryland at the census tract level are reported in (**Figure 6**); they are categorized as HBW, HBS, HBSHOP, HBO, NHBO, and NHBW. In particular we report:

1. Synthetic Estimates for Census Tracts not covered by RTS MWCOG 2018: These estimates were obtained by applying the selected best-performing model, namely the Random Forest area-level model. The model utilized the RTS MWCOG dataset as the source for independent variables. The estimates were generated within the framework of the synthetic population created from ACS 2018 and the 2020 census data.

2. Improved Estimates for Census Tracts covered by RTS MWCOG 2018: These estimates were produced using the Fay-Herriot model, which enhanced the accuracy and reliability of the estimates. The model was specifically applied to the census tracts covered by RTS MWCOG 2018.

Final Estimation Validation

As previously mentioned, the complete set of RTS data for Maryland became available later, which allowed us to validate our final synthetic estimates. However, the direct estimates at the census tract level in RTS BMC 2018 were not reliable and exhibited a high coefficient of variance, making them unsuitable for validation. Therefore, our approach was to validate the final estimates at the PUMA level and compare the estimated total number of trips with the observed in the RTS BMC 2018 data for all purposes. To accomplish this, we plotted the synthetic predicted results against the direct real estimates on a 45-degree plot, which represents perfect accuracy. The outcome of this analysis was highly encouraging, demonstrating an overall accuracy of 92% for the total number of trips. Moreover, the estimates for most trip purposes exhibited good performance, with an average accuracy of over 75%. However, the accuracy for the "Home-Based Social" (HBS) trip purpose was lower at 59%. Validation results are shown in (**Figure 7**), where the correlation between the synthetic predicted results and the direct real estimates for various trip purposes, including the overall total number of trips, are plotted.

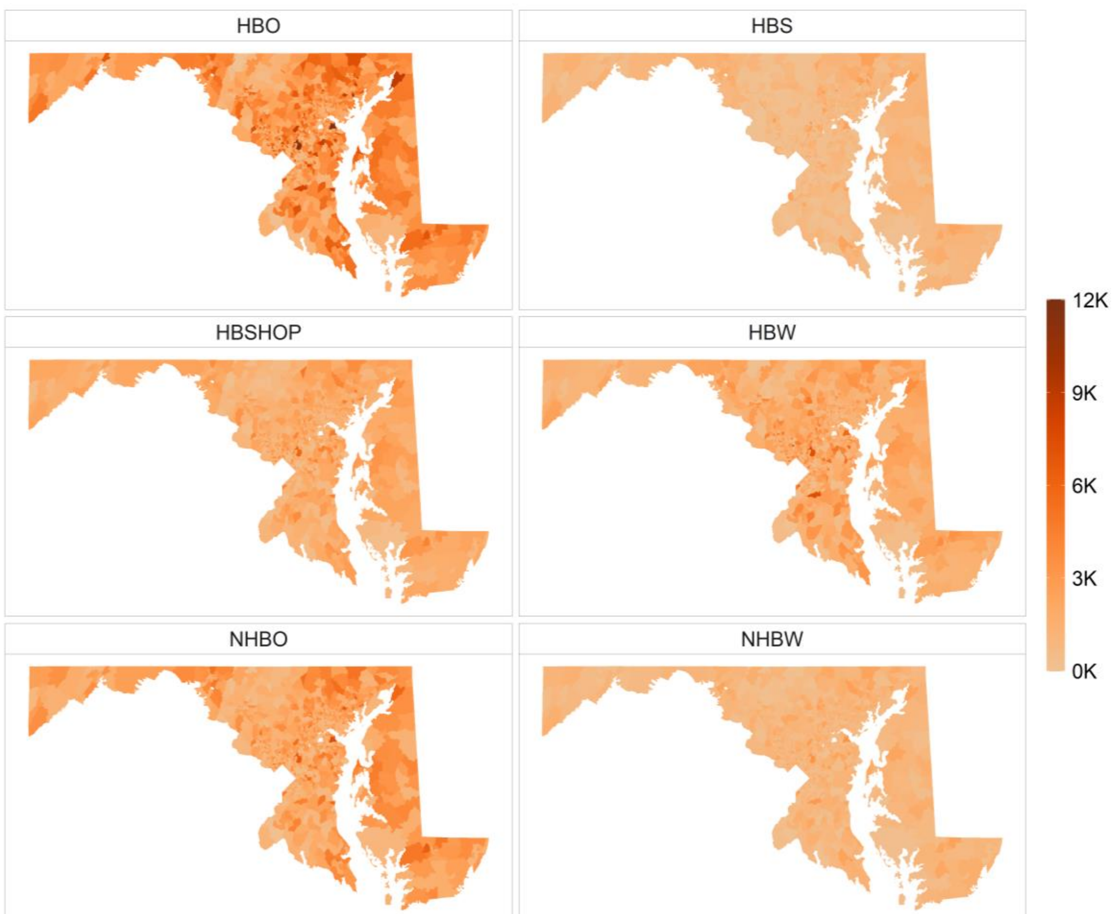


Figure 6: Distribution of Estimated Household Trip Types across Census Tracts in Maryland

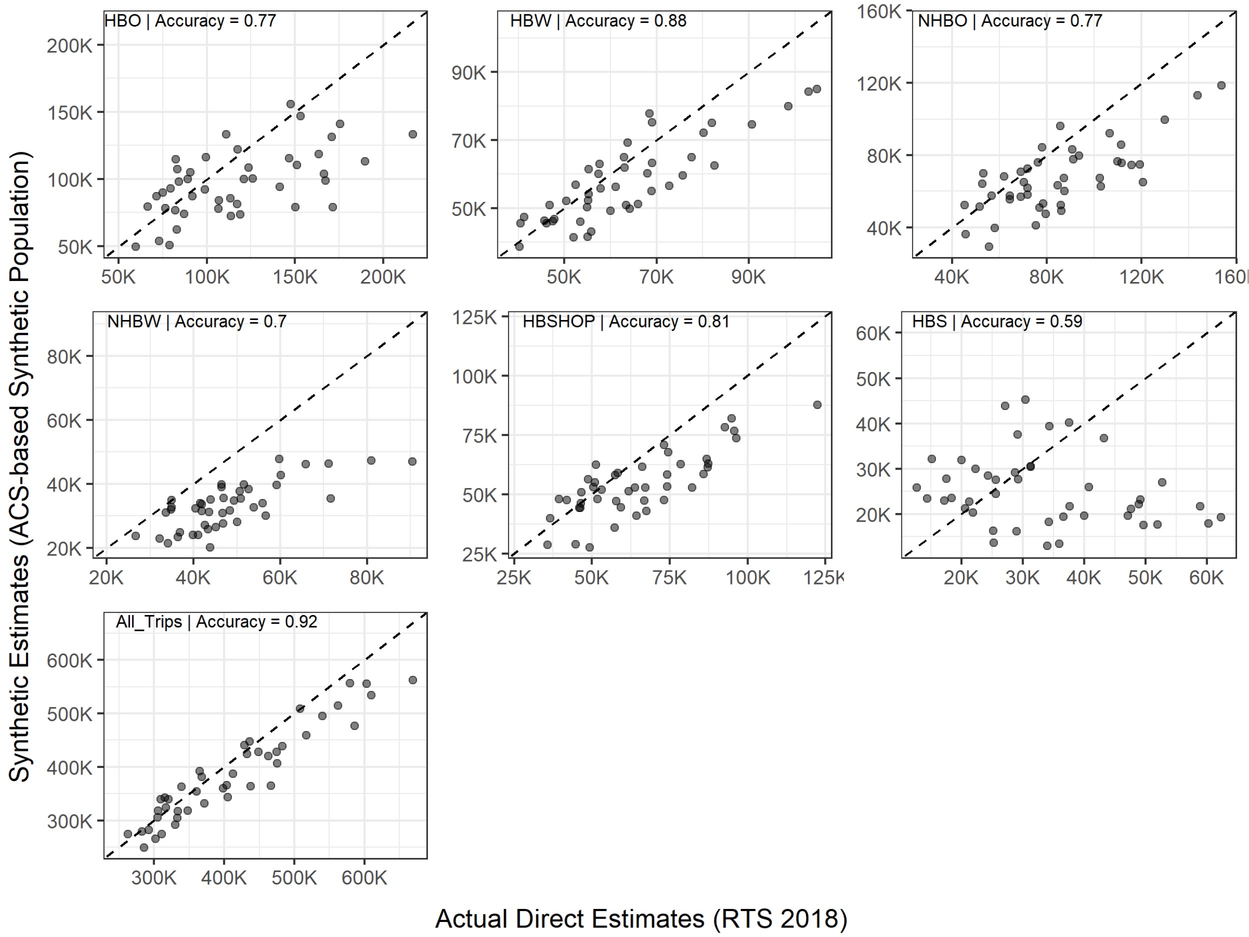


Figure 7: Validation Plot of Synthetic vs. Direct Estimates for Trip Purposes and Total Trips at PUMA Level

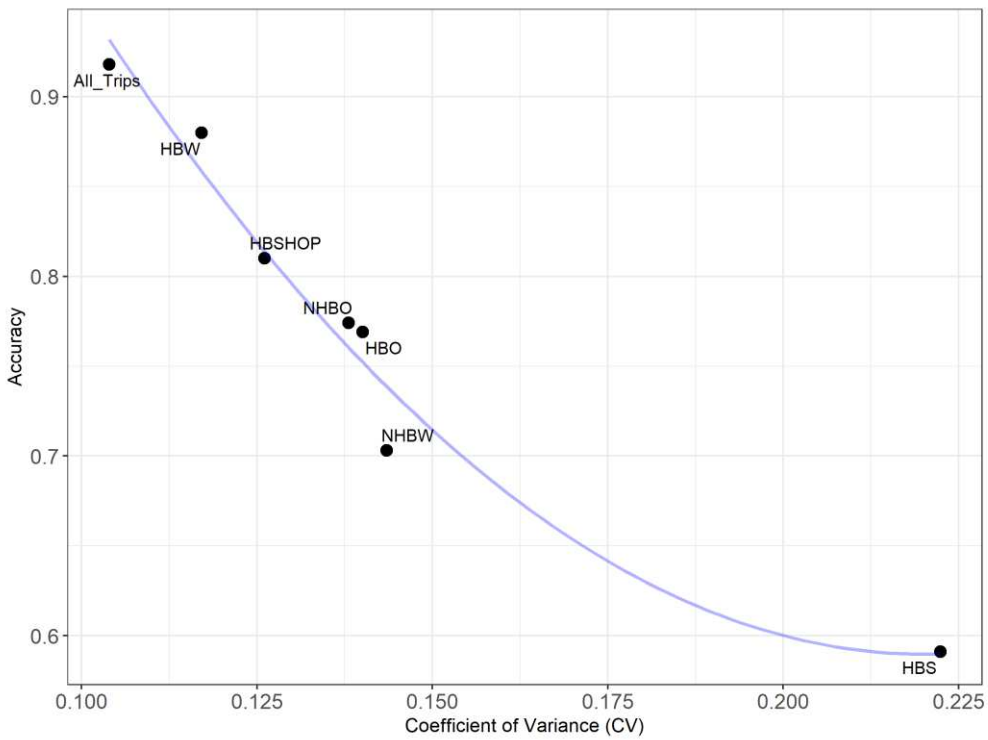


Figure 8: Relationship between Accuracy and Reliability (CV) of Estimated Models

Upon further investigation into the lower accuracy of the HBS estimates, we identified the primary reason could be the reliability of the direct estimates we used for comparison. The number of households reporting HBS trips was relatively small, leading to a high occurrence of zero trips being reported. Consequently, obtaining reliable estimates for the total number of HBS trips became challenging, and this challenge was reflected in the high coefficient of variance associated with these estimates. In essence, the estimated models exhibited excellent performance when compared to reliable estimates, such as those for "All Trips" and "Home-Based Work" (HBW). Conversely, the accuracy of the HBS model was affected by the unreliability of the HBS estimates used for comparison. Thus, it would be incorrect to conclude that the HBS model performed poorly. (**Figure 8**) depicts the relationship between the accuracy and reliability of the estimated models, represented by the coefficient of variance (CV).

In our overall evaluation, the models demonstrated commendable performance, consistently achieving a satisfactory level of accuracy exceeding 70%. However, the accuracy of the HBS model exhibited a noticeable deviation. This variance can be attributed to the utilization of less reliable estimates during the validation process. Hence, it is important to consider the reliability of the comparison data when assessing the performance of these models.

6. DISCUSSION

This study shows the power of combining various SAE techniques to tackle challenges in trip estimation at the census tract level. First, we synthesized disaggregated data at the census tract level across Maryland, integrating data from the American Community Survey (ACS) and the US Census. This synthetic population serves as a secondary dataset alongside the primary datasets of the Regional Travel Survey (RTS) in a regression-based modelling step to estimate trip totals in areas with no RTS samples. Furthermore, we successfully employ the Fay-Herriot model to refine existing estimates in areas with limited samples. However, we acknowledge that the current regression models could benefit from dividing the study area into groups based on shared characteristics. This would enable us to estimate separate models for each group, accounting for potential variations within the overall population. Our current approach assumes uniformity, which might introduce biases in certain areas.

It is important to understand that our goal is to produce dependable estimates by using the available datasets and offer a comprehensive methodological framework that can be adapted to similar research scenarios. We do not aim to recommend a specific model or compare the statistical performance of each model. Instead, we focus on providing a flexible approach. Please note that we do not claim that using an area-level random forest model which utilizes RTS MWCOG 2018 as the source for independent variables will always be the best approach. The same technique applied in different contexts may yield different results.

Moreover, the sample sizes in the ACS and RTS datasets used in this study are limited at the census tract level. The ACS does not provide tract-level samples, while the RTS has limited sample size at that level. The U.S. Census estimates are a form of synthetic population, resulting from a rolling survey approach that involves sampling methodologies. These factors can introduce uncertainties, particularly in tract-level analyses. It is crucial to recognize these limitations, as they can impact the statistical robustness of any model that heavily relies on ACS, RTS or Census data. Despite these challenges, we remain mindful of these limitations and strive to provide interpretations that consider both the strengths and challenges of the data for a nuanced understanding.

7. CONCLUSIONS

This research demonstrates that Small Area Estimation (SAE) methods are effective in producing reliable transportation statistics by leveraging multiple datasets. The proposed methodological framework successfully addresses data scarcity challenges and produces accurate estimates of household person trips at the census tract level in Maryland. To develop a comprehensive framework for modeling SAE (Small Area Estimation), this study integrates data from the Regional Travel Survey (RTS), the American Community Survey (ACS), and Census 2020. The framework includes three SAE techniques: regression-based models, population synthesis, and the composite estimator represented by the Fay-Herriot model. The study evaluates various models, including linear, Poisson, negative binomial, and random forest models, using cross-validation techniques, and identifies the area-level Random Forest model, utilizing RTS MWCOG 2018 auxiliary variables, as the best-performing specification for this study. By combining direct and synthetic estimation approaches, the research enhances estimation precision using the Fay-Herriot method.

The findings highlight the potential of SAE methods in transportation analysis by integrating diverse datasets and reducing survey data collection costs. The methodology and results of this study have practical implications for researchers, policymakers, and transportation planners seeking reliable estimates for smaller domains and subgroups using existing data sources. This research contributes to the field by providing a replicable SAE modeling framework for generating accurate transportation statistics. The application of SAE techniques addresses challenges related to limited coverage and small sample sizes, enabling the estimation of variables at the census tract level. The study emphasizes the importance of integrating different datasets to improve the reliability and value of statistical estimates. Furthermore, the research suggests the future exploration and implementation of SAE methods in transportation analysis, integrating diverse datasets such as large-scale traffic data or cell phone data, to further enhance data integration and analysis capabilities.

Declarations

Ethical Approval
Not applicable

Competing interests
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' contributions
Mohammad B. Al-Khasawneh: data collection, linkage and analysis; methodology; interpretation of the results; manuscript preparation; discussion and implications
Cinzia Cirillo: design of the study; methodology definition; advising; funding acquisition; manuscript editing and final draft.

Funding
We acknowledge funding received from the Lieutenant General John W. Morris II Graduate Fellowship, which helped funding the first author while working on this paper

Availability of data and materials
Data and code are available upon request submitted to the authors

REFERENCES

1. Wilson, T., Brokensha, H., Rowe, F., & Simpson, L. (2018). Insights from the evaluation of past local area population forecasts. Population Research and Policy Review, 37, 137-155.

2. Ghosh, M., & Rao, J. (1994). Small area estimation: An appraisal. Statistical Science, 9(1), 55–76.

3. Heady, P., Clarke, P., Brown, G., Ellis, K., Heasman, D., Hennell, S., Longhurst, J., & Mitchell, B. (2003). Model-based small area estimation series no. 2: Small area estimation project report. London, UK, Office for National Statistics.

4. Bertarelli, G., Ranalli, G., Bartolucci, F., d'Alò, M., & Solari, F. (2018). Small area estimation for unemployment using latent Markov models. Surv. Methodol, 44, 167–192.

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

6. Chatrchi, G. (2019). Small Area Estimation: Informative Sampling and Two-fold Models [PhD Thesis, Carleton University]. https://repository.library.carleton.ca/concern/etds/ff3656270

7. Choupani, A.-A., & Mamdoohi, A. R. (2016). Population synthesis using iterative proportional fitting (IPF): A review and future research. Transportation Research Procedia, 17, 223–233.

8. Fay III, R. E., & Herriot, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. Journal of the American Statistical Association, 74(366a), 269–277.

9. Gargiulo, F., Ternes, S., Huet, S., & Deffuant, G. (2010). An iterative approach for generating statistically realistic populations of households. PloS One, 5(1), e8828.

10. Goerndt, M. E., Monleon, V. J., & Temesgen, H. (2011). A comparison of small-area estimation techniques to estimate selected stand attributes using LiDAR-derived auxiliary variables. Canadian Journal of Forest Research, 41(6), 1189–1201. https://doi.org/10.1139/x11-033

11. Goulias, K. G., Ravulaparthy, S. K., Konduri, K., & Pendyala, R. M. (2014). Using Synthetic Population Generation to Replace Sample and 3 Expansion Weights in Household Surveys for Small Area 4 Estimation of Population Parameters 5.

12. Jiang, J., & Rao, J. S. (2020). Robust Small Area Estimation: An Overview. Annual Review of Statistics and Its Application, 7(1), 337–360. https://doi.org/10.1146/annurev-statistics-031219-041212

13. Li, W., Kelsey, J. L., Zhang, Z., Lemon, S. C., Mezgebu, S., Boddie-Willis, C., & Reed, G. W. (2009). Small-area estimation and prioritizing communities for obesity control in Massachusetts. American Journal of Public Health, 99(3), 511–519.

14. Lin, Y. (2023). Synthetic population data for small area estimation in the United States. Environment and Planning B: Urban Analytics and City Science, 23998083231215824. https://doi.org/10.1177/23998083231215825

15. Lin, Y., & Xiao, N. (2022). Developing synthetic individual-level population datasets: The case of contextualizing maps of privacy-preserving census data. ArXiv E-Prints, arXiv-2206.

16. Long, L., Lin, J., & Pu, W. (2009). Model-Based Synthesis of Household Travel Survey Data in Small and Midsize Metropolitan Areas. Transportation Research Record, 2105(1), 64–70.

17. López-Vizcaíno, E., Lombardía, M. J., & Morales, D. (2015). Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. Journal of the Royal Statistical Society: Series A (Statistics in Society), 178(3), 535–565.

18. Lovelace, R., Birkin, M., Ballas, D., & Van Leeuwen, E. (2015). Evaluating the performance of iterative proportional fitting for spatial microsimulation: New tests for an

established technique. Journal of Artificial Societies and Social Simulation, 18(2). https://eprints.whiterose.ac.uk/83990/

19. Martinez, A. (2014). Examining a weight reallocation method for small area estimation of poverty. Electronic Journal of Applied Statistical Analysis, 7(2), 417–431.

20. Moretti, A., & Whitworth, A. (2020). Development and Evaluation of an Optimal Composite Estimator in Spatial Microsimulation Small Area Estimation. Geographical Analysis, 52(3), 351–370. https://doi.org/10.1111/gean.12219

21. Pfeffermann, D. (2013). New important developments in small area estimation. https://projecteuclid.org/journals/statistical-science/volume-28/issue-1/New-Important-Developments-in-Small-Area-Estimation/10.1214/12-STS395.short

22. Rao, J. N., & Molina, I. (2015). Small area estimation. John Wiley & Sons.

23. Saadi, I., Mustafa, A., Teller, J., Farooq, B., & Cools, M. (2016). Hidden Markov Model-based population synthesis. Transportation Research Part B: Methodological, 90, 1–21.

24. Salon, D. (2016). Estimating pedestrian and cyclist activity at the neighborhood scale. Journal of Transport Geography, 55, 11–21.

25. Sun, L., & Erath, A. (2015). A Bayesian network approach for population synthesis. Transportation Research Part C: Emerging Technologies, 61, 49–62.

26. Vaish, A. K., Chen, S., Sathe, N. S., Folsom, R. E., Chandhok, P., & Guo, K. (2010). Small area estimates of daily person-miles of travel: 2001 National Household Transportation Survey. Transportation, 37(6), 825–848.

27. Zhongming, Z., Linong, L., Xiaona, Y., Wangqiang, Z., & Wei, L. (2020). Introduction to Small Area Estimation Techniques: A Practical Guide for National Statistics Offices.

28. Badoe, D. A., & Steuart, G. N. (1997). Urban and travel changes in the greater Toronto area and the transferability of trip-generation models. Transportation Planning and Technology, 20(4), 267–290.

29. Hu, S. (2010). Modelling trip generation/trip accessibility using logit models [PhD Thesis].

30. Mukherjee, J., & Kadali, B. R. (2022). A comprehensive review of trip generation models based on land use characteristics. Transportation Research Part D: Transport and Environment, 109, 103340.

31. Rashidi, T. H., Mohammadian, A., & Zhang, Y. (2010). Effect of variation in household sociodemographics, lifestyles, and built environment on travel behavior. Transportation Research Record, 2156(1), 64–72.

32. Frick, M. A. (2004). Generating synthetic populations using IPF and monte carlo techniques: Some new results. Arbeitsberichte Verkehrs-Und Raumplanung, 225.

33. Hörl, S., & Balac, M. (2021). Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data. Transportation Research Part C: Emerging Technologies, 130, 103291.

34. Zhang, D., Cao, J., Feygin, S., Tang, D., Shen, Z.-J. M., & Pozdnoukhov, A. (2019). Connected population synthesis for transportation simulation. Transportation Research Part C: Emerging Technologies, 103, 1–16.

35. Jutras-Dubé, P., Al-Khasawneh, M. B., Yang, Z., Bas, J., Bastin, F., & Cirillo, C. (2023). Copula-based synthetic population generation. arXiv preprint arXiv:2302.09193.

36. Aronhime, S., Calcagno, C., Jajamovich, G. H., Dyvorne, H. A., Robson, P., Dieterich, D., Isabel Fiel, M., Martel-Laferriere, V., Chatterji, M., & Rusinek, H. (2014). DCE-MRI of the liver: Effect of linear and nonlinear conversions on hepatic perfusion quantification and reproducibility. Journal of Magnetic Resonance Imaging, 40(1), 90–98.