EasyChair Preprint

№ 7836

Estimates of Total Household Trips for Areas with Zero Sample: a Maryland Statewide Analysis of Household Trip Production

Mohammad B. Al-Khasawneh, Cinzia Cirillo and Partha Lahiri

April 27, 2022

12th International Conference on Transport Survey Methods

Estimates of Total Household Trips for Areas with Zero Sample: A Maryland Statewide Analysis of Household Trip Production

Mohammad B. Al-Khasawneh[a], Cinzia Cirillo[a], Partha Lahiri[b]*

[a] A James Clark School of Engineering, University of Maryland, College Park, MD 20742, USA
[b] Joint Program in Survey Methodology, University of Maryland, College Park, MD 20742, USA

Abstract

This study applies Small Area Estimation (SAE) methods to a Regional Travel Survey (RTS) in order to derive statewide household person trips at the PUMA level when no sample is available. Several methods have been tested; those include: area and unit synthetic model with both Ordinary Linear Square and Poisson regression, and the Fay-Herriot model. Empirical results have been obtained for the State of Maryland, using the 2018 MWCOG Regional Travel Survey and the American Community Survey (ACS). RTS provides both the direct estimates of household person trips and the auxiliary variables for the synthetic model estimation, but only for PUMAs included in the survey. ACS provides the auxiliary variables for the entire state, including PUMAs with no RTS sample. Based on out-of-sample tests, it can be concluded that the area level linear model with RTS auxiliary variables performs better when compared to the other specifications proposed. This model was then applied to estimate household person trips for the area with no sample. We finally applied the Fay-Herriot method to the PUMAs in RTS and found that the combination of direct and synthetic estimation reduces the Coefficient of Variation. This application demonstrates that SAE methods can produce reliable transportation statistics by linking information from several datasets and could potentially reduce survey data collection costs.

Keywords: small area estimation, PUMA, household trips, zero sample

1. Introduction

Household Transportation Surveys gather data about people’s mobility and are mainly used to derive direct statistics of mobility indicators or estimate models of travel behaviour. The National Household Travel Survey (NHTS) conducted in the USA supports analysis at the national and Census region levels; the country is divided into six regions. The add-on program administrated by the Federal Highway Administration (FHWA) provides the opportunity to purchase supplemental samples to support state and metropolitan planning organizations (MPOs) analyses at smaller geographical levels, such as cities or counties. Regional Travel Surveys (RTS) are conducted by local authorities to have more detailed data about daily travel patterns and to estimate and validate large scale model systems (i.e., Four Step or Activity Based models). RTS sample sizes are usually larger than the NHTS sample but are relative to much more limited geographical areas.

This paper proposes the use of Small Area Estimation (SAE) methods to produce transportation-related statistics at a small geographical level when no sample is available. Small Area Estimation (SAE) is a term that embraces different appro

* Corresponding author. Tel.:
E-mail address:

aches and techniques to produce reliable statistics when a very small or even no sample is available. SAE methods are applied to small geographical areas such as a county, a municipality, a census tract, or small domains such as specific groups of people (i.e., low income-race-employment) within a large population (Ghosh & Rao, 1994). Nation or state-wide level surveys do not contain enough data to generate direct estimates for small areas; in such cases, an additional dataset with sufficient information for the small areas of interest may be used to obtain reliable estimates. In this study we estimate the number of household trips at the PUMA level and for the entire State of Maryland. We rely on the American Community Survey (ACS) and the RTS survey collected by the Metropolitan Washington Council of Governments (MWCOG) in 2017-2018 that covers 25 PUMAs only, out of the 44 PUMAs in Maryland. We demonstrate that by linking RTS data to the American Community Survey (ACS) data, it is possible to obtain PUMA level estimates where no sample is available. The method proposed is general and can be adopted to generate different types of small area or domain specific statistics. The remainder of this paper is organized as follows. In Section 2, we review SAE applications to transportation statistics. Section 3 describes the data that support the analysis. Section 4 illustrates the methodology developed, while Section 5 reports on the results obtained. Section 6 presents the main conclusions from the study and proposes new research directions.

2. Literature Review

There is a growing interest in reliable small area statistics; these are routinely used for a variety of purposes, including assessing the economic well-being of a nation, making public policies, and allocating funds at the federal, state and local levels. A comprehensive review of small area methods and their applications can be found in (Jiang & Lahiri, 2006; Rao & Molina, 2015).

In transportation the number of applications of SAE methods are rather limited. Reuscher et al., (2002) estimated vehicle and person trips, and miles of travel with data from the 1996 Nationwide Personal Transportation Survey (NPTS), an earlier version of NHTS, using clustering techniques aimed at grouping census tracts based on similarity in travel behavior indicators. A similar approach was followed by Hu et al., (2007) to derive cluster-specific transportation statistics using NHTS 2001 and Census data; the method was validated based on add on samples. Vaish et al., (2010) used the 2001 NHTS to derive SAE estimates of percentages of individuals among different age groups with high daily mileage travelled for each state in the USA. They used a survey weighted hierarchical Bayes SAE method (Folsom et al., 1999) and reported significant gains in the Prediction Intervals (PIs) especially for small states with fewer observations. Long et al., (2009) have evaluated three different models of SAE: the generalized regression estimators (GREG), the empirical best linear unbiased predictor (EBLUP), and the EBLUP without area effects to obtain estimates of the total number of workers per household at tract-level and individual-level using the 2001 NHTS dataset. They validated the results obtained using the three models with actual values obtained from the 2000 U.S. census (CTPP). They concluded that SAE methods can be used to obtain unbiased travel statistics for local areas.

A number of studies have focused on the transferability of transportation statistics. According to Koppelman & Wilmot, (1982) model transferability can be implicit when a model estimated on historical data is used to predict the future or explicitly when the model estimated in one area is used to make predictions in another area. Koppelman & Pas, (1986) has studied transferability of joint and sequential choice models for vehicle ownership and mode to work based on goodness of fit measures. A study by Wilmot, (1995) has carried out a comprehensive investigation of the transferability for 19 different linear models of trip-generation within several cities and among areas in one region. He found that models transfer better within areas that have similar characteristics, such as average income. Also, models that have high R2 values showed better transferability than other models with lower R2. (P. Stopher et al., 2005; P. R. Stopher et al., 2003) proposed a method to synthesize household travel survey data from Census data and a national transport survey. The procedure creates distributions of the variables relevant for travel-demand analysis. A sample of local residents is drawn from disaggregate census data, providing detailed information on the socioeconomic characteristics of the sample. Using these socioeconomic characteristics, travel data are simulated from the transport data distributions using Monte Carlo simulation. This procedure was also applied to Adelaide, South Australia, in (P. R. Stopher et al., 2003). Bayesian based updating techniques have been used to improve the transferability of household travel surveys to small and midsized urban areas in (Mohammadian & Zhang, 2007; Zhang & Mohammadian, 2008).

Recently the Bureau of Labor Statistics has released the LATCH model (Local Area Transportation Characteristics for Household) (LATCH, 2021), that provides estimates of average weekday household person trips, vehicle trips, person miles traveled, and vehicle miles traveled per day at the Census tract level in the United States. The methodology adopted is quite simple and based on classical SAE methods. The country is divided into six Census regions and three areas of urban/suburban/rural; using NHTS, models are estimated for each of the variables listed above, selecting dependent variables that are also available in ACS. Model estimates are transferred to Census tracts using the ACS data. This application attests the importance of producing small area transportation statistics for federal, state and local agencies. A similar effort is ongoing for the Freight Analysis Framework (FAF, 2021) (maintained by FHWA and BTS; both agencies are working to produce more granular statistics and implement concepts of SAE in the survey design in order to improve estimates' quality and their transferability.

The literature review reveals that SAE methods are becoming an essential statistical tool for governmental agencies, and applications in transportation are not numerous and limited to basic SAE methods. Also, there is no evidence of SAE applications to areas where no sample is available.

3. Data Sources

The analysis carried out in this paper is based on two datasets. The primary dataset is the Regional Travel Survey collected by the Metropolitan Washington Council of Governments (MWCOG) in 2017-2018, called here RTS 2018. The secondary dataset is the American Community Survey data relative to years 2014 - 2018, called here ACS 2018. ACS is administrated by the U.S. Census Bureau and collects monthly samples that are used to update annual estimates at small areas (census tracts and block groups); five years of samples are necessary to produce these small-area data. This is the reason why we used 5 years ACS data prior to 2018, which is the year when RTS data collection was finalized.

A total of 8,839 randomly selected households are available in RTS 2018 and 135,590 in ACS 2018 for Maryland. In Table 1 we summarize the main characteristics of both datasets; in particular, we identify the finest spatial unit for which the data is available, the area coverage, and the main limitations in the context of transportation statistic estimation.

Table 1: Data characteristics and limitations

	American Community Survey (ACS 2018)	Regional Travel Survey (RTS 2018)
Provided Information	It provides detailed population and housing information.	It provides demographics about households and individuals and travel information, including vehicles.
Spatial domain	County, PUMA.	County, PUMA level, zip code
Area covered	All states in the USA. Particularly, it covers all the 44 PUMAs in Maryland.	District of Columbia and parts of Maryland and Virginia. Particularly, it covers only 25 PUMAs out of the 44 PUMAs in Maryland.
Limitations	It does not provide the output variable of interest (i.e., the total number of household person trips).	It does not cover all the small areas in Maryland selected for this study.

The study area covers the State of Maryland, which contains 44 PUMAs (Public Use Microdata Area). PUMAs are the geographic units defined by the U.S. Census and contain at least 100,000 people; PUMAs do not overlap, and are contained within a single state. Among the 44 PUMAs (see Figure 1) in Maryland, only 25 PUMAs are covered in the primary dataset (RTS 2018). Out of these 25 PUMAs, 20 were used as the training set to estimate the proposed models and the remaining five were selected as the testing set (out-of-sample). The remaining 19 PUMAs represent the small areas with zero sample in the primary dataset for which we want to transfer model estimates based on RTS 2018.

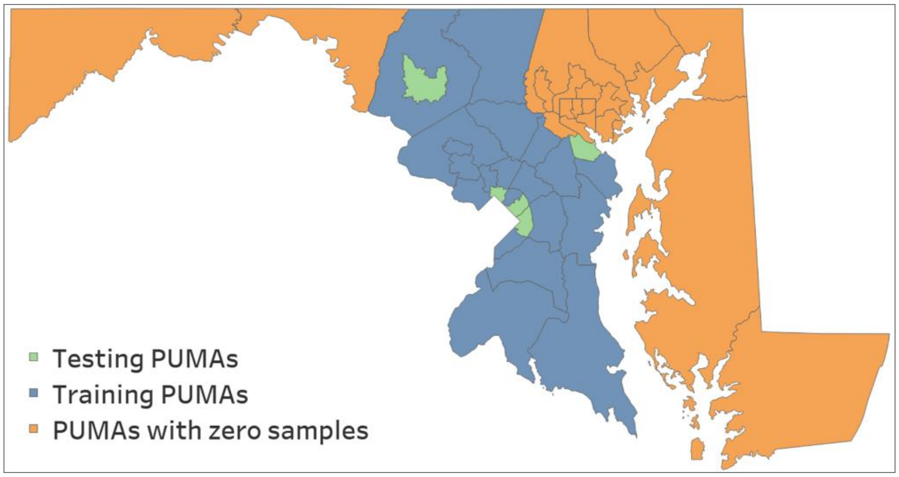


Figure 1: Training, testing, and zero sample PUMAs

4. Methodology: A Three Step SAE Based Procedure

In this section we first present the SAE method adopted in this study, including the Fay-Heriot model, the framework developed for model transfer to areas with no sample, and the tests used to compare models' performance.

4.1 SAE method with no sample

Different formulations of Linear models and Poisson models were estimated using the data for the PUMAs with available samples in the RTS 2018 dataset (the primary dataset).

Linear regression model

The simple linear regression model can be expressed as the following:

$$Y = B_O + X_i^T B + \varepsilon \qquad (1)$$

Where Y is the known dependent variable, X_i^T is a vector of known independent variables, B is a vector of the regression coefficients of the model, B_O is the intercept of the model, ε is the error term.

Poisson regression model

The Poisson model was used in order to address the count data of the household total trips. The model assumes that the dependent variable Y has a Poisson distribution with probability mass function:

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad , x = 0, 1, 2, \dots , \qquad (2)$$

Where λ is Mean of occurrences in the interval. The Poisson model is a natural log of the dependent variable as a linear function of the independent variables. The log of the predicted counts of the output variable of interest in the ith small area unit can be expressed as the following:

For i = 1,2, …n

$$\log\left(\mathrm{Y}_{\mathrm{i}}^{\mathrm{p}}\right) = X_i^T B + \mu_i \qquad (3)$$

Where X_i^T is a vector of known auxiliary count data, B is a vector of the regression parameters of the Poisson model, and μ_i is a vector of independent random errors.

Then the best-performed model was transferred to be applied to the ACS 2018 dataset (the secondary dataset) to generate synthetic estimates for PUMAs with zero samples using equation 4. The two used datasets have the same auxiliary variables. Therefore, no changes of any kind to the estimation coefficients from the primary dataset had to be made before applying the estimated model to the secondary dataset.

$$\hat{Y}_{i_{NS}}^{S} = f(X_i) \qquad (4)$$

Where $\hat{\mathrm{Y}}_{\mathrm{i}_{\mathrm{NS}}}^{\mathrm{S}}$ is the synthetic estimates for areas with no samples in the primary dataset, $\boldsymbol{f}$ is the best performed estimated function that can explain the relationship between the predictor and the auxiliary variables for other areas in the primary dataset, and X_{i} is a vector of auxiliary variables from the secondary dataset for the specific areas.

Fay-Herriot Model

In regression-synthetic estimation, the Fay-Herriot model estimator is a weighted combination of the direct estimator and the synthetic estimator as the following:

For i= 1,2,…n

$$\hat{Y}_i^S = \omega_i \hat{Y}_i^D + (1 - \omega_i) X_i^T \beta \qquad (5)$$

Where $\hat{Y}_i^S$ is the Fay-Herriot estimator, $\hat{Y}_i^D$ is the direct estimator, $X_i^T \beta$ is the synthetic estimator, and ω_i is the weight of the direct estimator as a ratio ranging from 0 to 1 and can be computed using the variances of the model $\hat{\sigma}_u^2$ and the sampling errors $\hat{\sigma}_e^2$:

$$\omega_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_{e_i}^2} \qquad (6)$$

The Fay-Herriot model minimizes the mean square error in the final estimator to produce more reliable estimates. If the sampling error is small, the direct estimate will have more weight on the Fay-Herriot estimator. However, if the direct estimate is not reliable, the synthetic estimate will have more weight. In the case of zero sample observations, the direct estimate is impossible to be calculated. Therefore, the Fay-Herriot estimator reduces to the synthetic estimate. The unbiased estimator of the mean square error of the Fay-Herriot estimator can be approximated by the formula given by Prasad and Rao in 1990:

$$\mathrm{MSE}\ (\hat{Y}_i^S) = \omega_i \hat{\sigma}_{e_i}^2 + (1 - \omega_i) X_i^T \left[\sum \frac{X_i X_i^T}{\hat{\sigma}_u^2 + \hat{\sigma}_{e_i}^2}\right]^{-1} X_i + \frac{\hat{\sigma}_{e_i}^2}{\widehat{(\sigma}_u^2 + \hat{\sigma}_{e_i}^2)^3} \frac{4 \sum \left(\hat{\sigma}_u^2 + \hat{\sigma}_{e_i}^2\right)^2}{n^2} \qquad (7)$$

4.2 The modeling framework

The methodology developed and applied in this study is schematically represented in Figure 2, where it can be seen that we estimated both area level models based on aggregate data for each PUMA, and unit level models on disaggregate households' observations. A three-step procedure for model estimation, validation and transfer was implemented as follows:

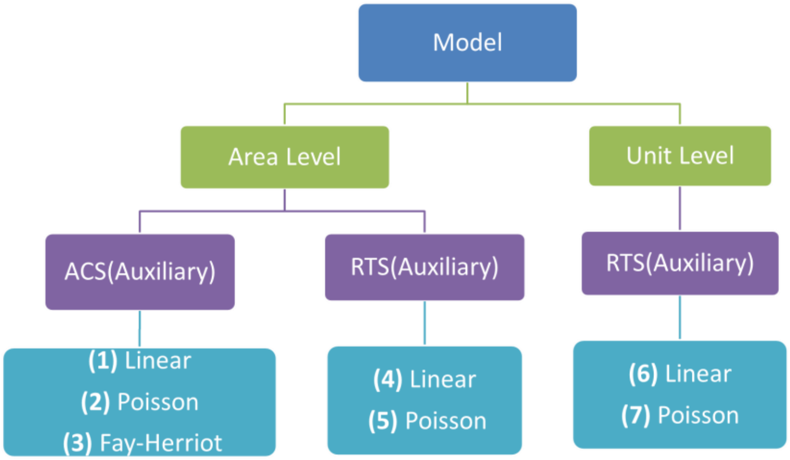


Figure 2: Methodological Framework

Step1: Generate direct total estimates of the output variable (i.e., the total number of household person trips) from the primary dataset (RTS 2018), and generate the direct estimates of the auxiliary variables from both the primary and secondary datasets (RTS 2018 and ACS 2018).

Step 2: Produce synthetic estimates of the output variable (i.e., the total number of household person trips) using different formulations for the unit level models and area level models based on the direct estimates generated in Step 1.

Step 3: Select the best-performing model by comparing the error between synthetic estimates and the actual direct estimates for the covered areas in the primary dataset. Then using the selected model, produce synthetic estimates for the areas with zero sample in the primary dataset to cover all the PUMAs in Maryland.

The direct estimates of the total number of household person trips (HHTRIPS) and of the selected auxiliary variables were estimated for the 25 PUMAs in the testing and training sets using the survey package in R software (Lumley, 2020). The auxiliary variables were chosen based on their availability in both ACS and RTS datasets and their relation to the output variable of interest in this study. They are: household size (HHSIZE), household income (HHINCOME), and the number of vehicles in the household (NUMVEHICLE). Table 2 lists the selected variables and their names, availability in surveys, definitions and associated values.

Table 2: List of variables used in this study

Auxiliary Variables	**Values**	**Survey**
Number of household vehicles (NUMVEHICLE)	0,1,2,…6+	ACS & RTS
Number of household people (HHSIZE)	0 to 20	ACS & RTS
Household income (HHINCOME)	1 to 9999999	ACS & RTS
Dependent Variable (Outcome)	**Values**	**Survey**
Total Household Trips (HHTRIPS)	1 to 9999999	RTS only

We considered different model specifications: Linear regression, Poisson model, and empirical best linear unbiased prediction (EBLUP) using Fay-Herriot model (Fay III & Herriot, 1979). The linear and the Poisson models were estimated using the survival package in R software (Therneau & Lumley, 2015), and the Fay-Herriot model was estimated using the "SAE" package (Molina & Marhuenda, 2015).

A total of seven different regression alternatives were considered for the synthetic estimation of the household person trips in each PUMA. We estimated five area level models:

(1) Ordinary linear model using the auxiliary variables in the RTS 2018.
(2) Poisson linear model using the auxiliary variables in the ACS 2018.
(3) Fay-Herriot model using the auxiliary variables in the ACS 2018.
(4) Ordinary linear model using the auxiliary variables in the ACS 2018.
(5) Poisson linear model using the auxiliary variables in the RTS 2018.

A two unit-level models:

(6) Ordinary linear model using the auxiliary variables in the RTS 2018;
(7) Poisson linear model using the auxiliary variables in the RTS 2018.

4.3 Statistical test for model comparison

All the regression models were tested by comparing the synthetic estimates with the direct estimates computed in Step 1 for the five PUMAs in the training set defined in Table 1. Two indicators were chosen to measure the errors: The Mean Absolute Percentage Error (MAPE) and the Root-Mean-Square deviation (RMSE). They are defined as follows:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right| \quad (8)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_t - F_t)^2} \quad (9)$$

where, A_t = actual value; F_t = forecast value; n = number of observations.
Moreover, the coefficient of variance (CV) was used to test the performance of the Fay-Herriot model:

$$CV = \frac{StandardError(SE)}{Estimates} \quad (10)$$

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

$$SE = \sqrt{MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y_\iota}\right)^2} \quad (11)$$

Where, Yi = actual values; $\widehat{Y_\iota}$ = predicted values

5. Results: SAE Estimates for Areas with No Sample

According to the framework in Section 4.3, we first calculate the direct estimates of the independent variables from both RTS and ACS data, and direct estimates of the dependent variable using RTS only. The outcome is reported in Table 3, where direct estimates are reported for a sample of the 25 PUMAs covered by RTS and ACS and a sample of the 19 remaining PUMAs covered by the ACS only. It can be noted that the direct estimates of each of the auxiliary variables have similar patterns and very similar values across the two surveys and for both the training and the testing areas. This finding supports our approach that intends to use ACS 2018 as a secondary dataset to infer the number of household person trips for PUMAs with no sample.

Table 3: Direct estimates

PUMA	Direct Total Estimates from RTS 2018				Direct Total Estimates from ACS 2018		
	HHSIZE	NUMVEHICLE	HHINCOME	HHTRIPS	HHSIZE	NUMVEHICLE	HHINCOME
302	125553	99032	286385	403920	117083	86752	251557
1007	112971	66334	290209	347806	110363	63023	244633
1103	104390	56464	201719	302215	98398	56761	181934
1104	115476	65651	225118	292323	105946	64051	208134
1202	106185	85368	218103	318722	110809	82559	222869
301	131380	108244	283800	405041	123872	105977	266245
400	166553	144967	360651	523232	162259	137256	342142
901	140156	106930	330252	477981	132544	104503	316809
902	169694	128525	406835	589222	169664	122211	384401
1001	139708	109942	336110	463088	133039	103750	298192
1002	140362	87924	287493	412366	127266	82012	256084
1003	181379	121336	445926	579498	178231	117812	400924
1004	190480	126504	528252	669010	178930	125307	486955
1005	146141	87244	327495	428798	134098	87758	272628
1006	132846	84508	272068	466919	116814	77973	238629
1101	99796	54866	188847	285350	96185	49941	157277
100	No samples				91996	72001	161242
200					138861	106986	259444
501					115094	97510	263606
502					125928	85776	262160
503					112288	79891	251567
504					103695	73967	233730
505					104373	76112	218450
506					106322	66145	184429
507					105388	71167	212109
601					131809	110268	284081
602					106296	79220	222164
700					93032	74685	189184
801					116344	52400	194005
802					92896	48633	193329

803	104350	56848	194749

We also calculate the direct estimates of the independent auxiliary variables for the 19 PUMAs with no sample using the ACS 2018 data. Then we proceed to estimate the seven regression models described in Section 4.3 using data from the 20 PUMAs common to RTS 2018 and ACS 2018 and leaving out the five PUMAs that will be used to test model performance. Results from model estimation are reported in Table 4. It can be observed that the area level models have higher R2 compared to the unit level models. The Poisson regression did not improve the accuracy for the proposed models as all Poisson models showed the same or slightly less R2 compared to the linear models for the same category.

Table 4: Model estimation results

Model	**Coefficients:**	**Estimate**	**Std. Error**	**p-value**	**Significance**[†]	**Observations**	**R2**
(Linear) (1)	(Intercept)	64970.000	4.23E+04	0.1425		20 PUMAs aggregated from 7316 households	**0.908**
	ACS_HHSIZE	1.171	7.78E-01	0.1508			
	ACS_HHINCOME	0.777	2.69E-01	0.0101	**		
(Poisson) (2)	(Intercept)	12.200	2.01E-03	<2e-16	***	20 PUMAs aggregated from 7316 households	**0.896**
	ACS_HHSIZE	2.58E-06	3.46E-08	<2e-16	***		
	ACS_HHINCOME	1.56E-06	1.19E-08	<2e-16	***		
(Linear) (4)	(Intercept)	28670.000	5.11E+04	0.5822		20 PUMAs aggregated from 7316 households	**0.896**
	RTS_HHSIZE	1.374	8.65E-01	0.1306			
	RTS_HHINCOME	0.708	2.78E-01	0.0209	*		
(Poisson) (5)	(Intercept)	12.090	2.38E-03	<2e-16	***	20 PUMAs aggregated from 7316 households	**0.874**
	RTS_HHSIZE	3.65E-06	3.82E-08	<2e-16	***		
	RTS_HHINCOME	1.22E-06	1.20E-08	<2e-16	***		
(Linear) (6)	(Intercept)	-1.38789	0.3351	3.49E-05	***	7316 Households based on the training 20 PUMAs	**0.362**
	RTS_HHSIZE	2.826	4.75E-02	< 2e-16	***		
	RTS_NUMVEHICLE	0.305	1.35E-01	0.0243	*		
	RTS_HHINCOME	1.014	1.84E-01	3.48E-08	***		
(Poisson) (7)	(Intercept)	0.745	2.17E-03	<2e-16	***	7316 Households based on the training 20 PUMAs	**0.345**
	RTS_HHSIZE	2.78E-01	2.21E-04	<2e-16	***		
	RTS_NUMVEHICLE	1.43E-01	8.22E-04	<2e-16	***		
	RTS_HHINCOME	1.70E-01	1.11E-03	<2e-16	***		

The performance of the regression-based models was assessed on the testing set composed of five PUMAs using the RMSE and MAPE indicators. Table 5 summarizes the results obtained. The linear area-level model with the auxiliary variables from the primary dataset (Model 4) was found to have the best performance as it showed the lowest RMSE and MAPE values. Therefore, this model was selected to predict the total household trips in the remaining 19 PUMAs with no sample. We finally estimate the Fay-Herriot model as specified in Section 4.2 for the 25 PUMAs in RTS and found that it produced better estimates with respect to the direct estimates since the associated coefficients of variance were lower. Figure 3 illustrates the change in the coefficients of variance between the Fay-Herriot model and the direct estimates. The final estimates are shown in Figure 4 as the following: (1) the final improved synthetic estimates of the total household trips in the 25 PUMAS using the Fay-Herriot model and (2) the synthetic estimates for all the remaining 19 PUMAs with no sample using the linear area-level model with the auxiliary variables from the primary dataset.

[†] significance code [p-value]: "***" [0, 0.001], "**" (0.001, 0.01],"*" (0.01, 0.05], "." (0.05, 0.1], " " (0.1, 1]

Table 5: RMSE and MAPE in the testing dataset

		Actual	Synthetic Estimation (Predicted)					
			Area Level				Unit Level	
		Direct Estimates	(Linear) ACS-AUX	(Poisson) ACS-AUX	(Linear) RTS-AUX	(Poisson) RTS-AUX	(Linear) RTS-AUX	(Poisson) RTS-AUX
Testing PUMA		HHTRIPS	HHTRIPS	HHTRIPS	HHTRIPS	HHTRIPS	HHTRIPS	HHTRIPS
1	302	403920	397455	397367	367730	372789	394424	390650
2	1007	347806	384208	386335	351659	355315	367341	358849
3	1103	302215	321500	339616	291221	315391	325314	336003
4	1104	292323	350688	360772	320241	335181	351573	365488
5	1202	318721.7	367827	373839	338910	351824	373523	374149
		Error						
		MAPE	10.8%	13.2%	5.9%	7.9%	6.5%	9.4%
		RMSEE	38875.74	46150.71	22945.20	28746.14	38779.22	44418.40

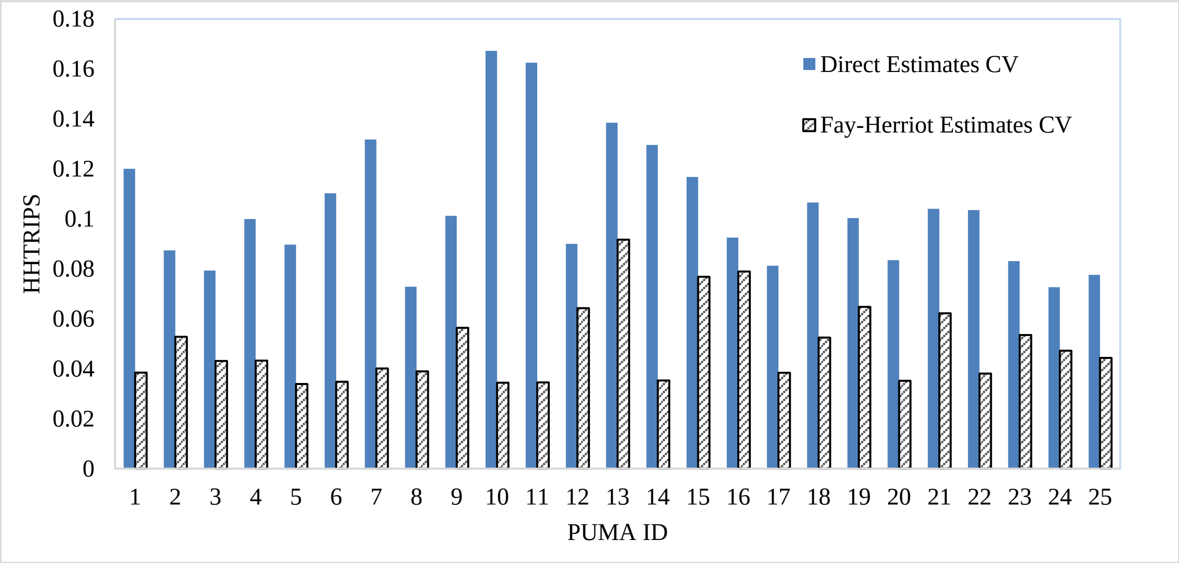
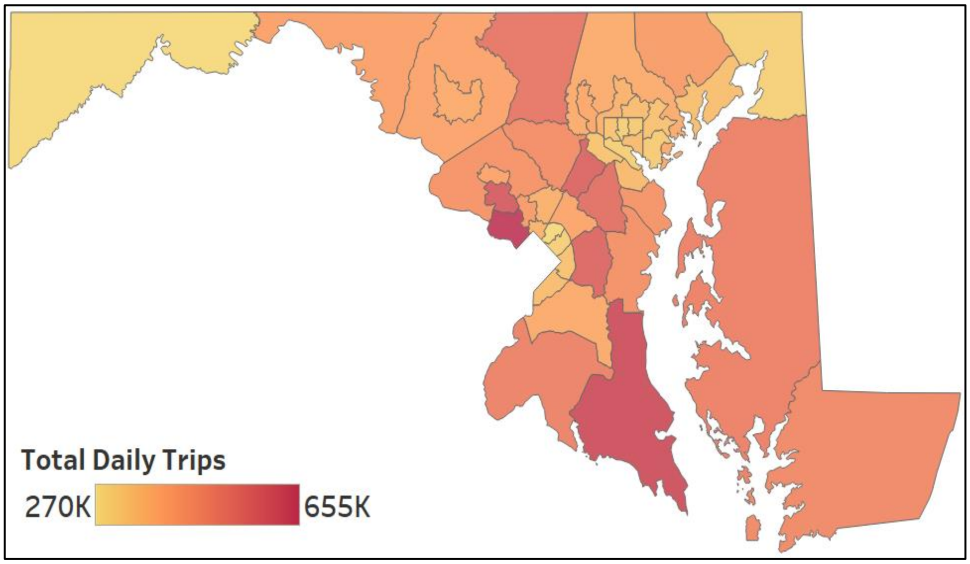


Figure 3: Coefficient of variance across direct and Fay-Herriot estimates



6. Conclusion

Transportation agencies rely on quantitative statistics to make information-based decisions about investments and operations. Data about travel behavior is collected through national and regional surveys, but often their representativeness is limited in space and domains. Methods from Small Area Estimation can be applied to produce reliable statistics for areas with sparse or no data by leveraging information contained in different data sets. In this SAE application, the Regional Travel Survey with no sample for about half of Maryland was linked to a secondary dataset (the American Community Survey) that covers the entire state. In order for the linkage to happen the two datasets must contain the same auxiliary variables. Direct and synthetic estimations of the variable of interest (household person trips) were produced at the PUMA level for the area covered by the survey. The best-performing model on the out-of-sample composed of five PUMAs was selected to transfer the estimates and obtain statistics for the area with no sample. The Fay-Herriot model was finally applied to the region covered by MWCOG RTS and was found to produce inferior CVs with respect to the direct estimates.

This paper has provided to the transportation community an SAE modeling framework that is ready to be applied and can be replicated in different areas and domains. A natural extension of this paper is the SAE estimation of person trips at Traffic Analysis Zone (TAZ) level, which is the unit of analysis for large scale transportation model systems. Other applications might include the estimation of non-motorized trip rates at the Census tract level, or the analysis of travel behavior for specific segments of the population (i.e., low income, disabled, and senior citizen). Finally, SAE methods offer a great variety of models for data linkage; these could be applied to link big (traffic) data and/or cell phone data to traditional travel survey data to fully exploit the power of passively collected data

References

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.

Folsom, R. E., Shah, B., & Vaish, A. (1999). Substance abuse in states: A methodological report on model based estimates from the 1994–1996 National Household Surveys on Drug Abuse. Proceedings of the Section on Survey Research Methods, American Statistical Association, 371–375.

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

Hu, P. S., Reuscher, T., Schmoyer, R. L., & Chin, S.-M. (2007). Transferring 2001 National Household Travel Survey. ORNL/TM-2007/013. Washington, DC: US Department of Transportation, Federal Highway Administration. Http://Nhts. Ornl. Gov/Tx/TransferabilityReport. Pdf.

Jiang, J., & Lahiri, P. (2006). Mixed model prediction and small area estimation. Test, 15(1), 1–96.

Koppelman, F. S., & Pas, E. I. (1986). Multidimensional choice model transferability. Transportation Research Part B: Methodological, 20(4), 321–330.

Koppelman, F. S., & Wilmot, C. G. (1982). Transferability analysis of disaggregate choice models. Transportation Research Record, 895, 18–24.

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.

Lumley, T. (2020). Survey: Analysis of complex survey samples. R package version 3.35-1. 2019.

Mohammadian, A., & Zhang, Y. (2007). Investigating transferability of national household travel survey data. Transportation Research Record, 1993(1), 67–79.

Molina, I., & Marhuenda, Y. (2015). sae: An R Package for Small Area Estimation. R J., 7(1), 81.

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

Reuscher, T. R., Schmoyer Jr, R. L., & Hu, P. S. (2002). Transferability of nationwide personal transportation survey data to regional and local scales. Transportation Research Record, 1817(1), 25–32.

Stopher, P., Greaves, S., & Xu, M. (2005). Using national data to simulate metropolitan area household travel data. Journal of Transportation and Statistics, 8(3), 83–95.

Stopher, P. R., Greaves, S., & Bullock, P. (2003). Simulating household travel survey data: Application to two urban areas. 82nd Annual Meeting of the Transportation Research Board, Washington, DC.

Local Area Transportation Characteristics for Household (LATCH Survey). (2021, February 21). Retrieved from https://www.bts.gov/latch

Freight Analysis Framework (FAF). (2021, June 24). Retrieved from https://ops.fhwa.dot.gov/freight/freight_analysis/faf/

Therneau, T. M., & Lumley, T. (2015). Package ‘survival.’ R Top Doc, 128(10), 28–33.

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.

Wilmot, C. G. (1995). Evidence on transferability of trip-generation models. Journal of Transportation Engineering, 121(5), 405–410.

Zhang, Y., & Mohammadian, A. (2008). Bayesian updating of transferred household travel data. Transportation Research Record, 2049(1), 111–118.