Final Report

E3: EVALUATING EQUITY IN EVACUATION: A PRACTICAL TOOL AND A CASE STUDY

Cinzia Cirillo
Department of Civil and Environmental Engineering,
University of Maryland
3250 Kim Bldg., College Park, MD 20742
Tel: +1 301-405-6864 Fax: +1 301-405-2585; Email: ccirillo@umd.edu

Mohammad Nejad
Department of Civil and Environmental Engineering and National Center for Smart Growth,
University of Maryland
Email: mmnejad@umd.edu

Sevgi Erdogan
National Center for Smart Growth Research and Education,
University of Maryland
112J Preinkert Field House, College Park, MD 20742
Tel: +1 301- 405-9877, Email: serdogan@umd.edu

Celeste Chavis
Transportation and Urban Infrastructure Studies
Morgan State University
1700 East Cold Spring Lane, Baltimore, MD 21251
Tel: +1 443-1-885-5061 Fax: +1 123-456-7890; Email: celeste.chavis@morgan.edu

Date
February 2020

Prepared for the Urban Mobility and Equity Center, Morgan State University, CBEIS 327, 1700 E. Cold Spring Lane, Baltimore, MD 21251

ACKNOWLEDGMENT

The research presented in this report was funded by the Urban Mobility and Equity Center, a Tier 1 University Transportation Center, led by Morgan State University. Its support is greatly acknowledged.

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle E3: Evaluating Equity In Evacuation: A Practical Tool And A Case Study		**5. Report Date** February 2020
		6. Performing Organization Code
7. Author(s) *Include ORCID #* Cinzia Cirillo 0000-0002-5167-0413		**8. Performing Organization Report No.**
9. Performing Organization Name and Address University of Maryland 1179 Glenn M. Hall College Park, MD 20742		**10. Work Unit No.**
		11. Contract or Grant No. 69A43551747123
12. Sponsoring Agency Name and Address US Department of Transportation Office of the Secretary-Research UTC Program, RDT-30 1200 New Jersey Ave., SE Washington, DC 20590		**13. Type of Report and Period Covered** Final
		14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

Natural or man-made hazards that require evacuation put already vulnerable populations in a more precarious situation. When plans and decisions about evacuation are made, access to a private car is typically assumed, and differences in income levels across a community are rarely taken into account. The result is that carless members of a community can find themselves stranded. Low-income carless residents need alternative transportation means to reach shelters in case of an emergency. Thus, evacuation plans, decisions, and models need necessary information that identifies and locates these populations. In this study, data from the American Community Survey, U.S. Census, Internal Revenue Service, and the National Household Travel Survey are used to generate a synthetic population for Anne Arundel County, Maryland, using the copula concept. Geographic locations of low-income residents are identified within each subarea of the county (census tract) and their car ownership is estimated with a binomial logit model. The developed population synthesis method allows officials to have a more accurate account of populations for emergency planning and identify locations of shelters and triage points as well as planning carless transportation services.

17. Key Words: Synthetic population, Archimedean copulas, Accessibility, Car-ownership models, Evacuation planning, low income, carless	18. Distribution Statement		
19. Security Classif. (of this report) : Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 26	**22. Price**

TABLE OF CONTENTS

1 EXECUTIVE SUMMARY 5
2 INTRODUCTION 7
3 LITERATURE REVIEW 9
4 DATA, MODEL SPECIFICATION, AND METHODOLOGY 10
4.1 DATA DESCRIPTION 10
4.1.1 Anne Arundel County Application 10
4.1.2 Data Acquisition 11
4.2 METHODOLOGY 14
4.3 MODEL ESTIMATION 18
4.3.1 Identification of Low-income Carless People 18
5 CONCLUSIONS 22
6 REFERENCES 24

LIST OF TABLES

Table 1: Archimedean three most popular Archimedean copulas 15
Table 2: Estimated parameters for Clayton copula for census tracts within Anne Arundel County 18
Table 3: Results of binomial model used to estimate car ownership 19

LIST OF FIGURES

Figure 1: Median income for different census tracts within Anne Arundel County 11
Figure 2: Anne Arundel County; PUMAs and census tracts 12
Figure 3: A sample of data used from IRS data to generate CDF functions for Income 14
Figure 4: A sample of crosswalks between ZIP code and census tracts 14
Figure 5: Flowchart of the population-synthesizing algorithm 17
Figure 6: Car ownership and income level for census tracts within Anne Arundel County 21

1 EXECUTIVE SUMMARY

Officials at all levels of government understand the importance of adequate planning for and responding to the challenges caused by natural hazards. They have learned that evacuation planning needs to consider the carless, particularly the low-income population. There are many reasons why people may be carless; in many communities, particularly large cities, a significant portion of residents are carless due to unemployment or poverty. Access to transportation is of the utmost importance in the event of an evacuation, especially for the carless population. Having the capability to accurately assess both the population of the low-income individuals as well as their potential need is critical in the event of an evacuation. This information allows transportation planners and emergency managers to deliver the necessary services to those in need. Without accurate estimates, a deflated special needs population estimate can strain service quality and risk lives, whereas overestimating can allocate unnecessary resources to communities that can do without. Inventorying a jurisdiction's carless population can be a daunting task. Using the copula concept and open-sourced data, the authors generated a synthetic population to identify the low-income population of Anne Arundel County, Maryland, and identify the geographical locations of those low-income people.

Anne Arundel County is south of the City of Baltimore and west of the Chesapeake Bay. Its geo-spatial location combined with changing climatic conditions have made this county vulnerable to natural disasters like hurricanes, storm surges and flooding. There are four Public Use Micro Areas (PUMAs) in Anne Arundel County and each PUMA consists of several census tracts. In this project, the data from PUMA and census tract levels are combined with IRS data to generate the synthetic population. In short, the three sources of data used in this study are: (1) American Community Survey (ACS), (2) Decennial Census Data, and (3) IRS Income Data from tax returns. In this study, we fitted the data from Public Micro Area Samples (PUMS) by using the Archimedean family of copulas. The individual- and household-level data of PUMS have been merged to generate an individual-level dataset that includes all the attributes of the interest in fitting purpose. The authors picked nine distinct variables that are mostly used in transportation modeling from the PUMS data attributes and used them to estimate Clayton copula for four PUMAs within the county. Having fitted the copulas, a set of synthetic pseudo-observations can be drawn from the copula for any sub-region of the PUMA of interest within Anne Arundel County. The sub-regions have been set to census tracts. The pseudo-observations have been transformed to real observations using the Inverse Cumulative Distribution functions for the attributes.

A binomial car-ownership model has been estimated for the State of Maryland and has been examined on synthetic populations obtained for different census tracts. Through the generated synthetic population and the simple car ownership model, a census tract level map has been generated. The resulting binomial model, which was based on the copula-generated synthetic population, successfully captured the expected dependency between car ownership status and income level of the individuals within the census tracts. From the result, it is observed that most of the people with no car are located within the northern part of the county. The percentage of people with no car varies between 2 percent and 11 percent, while the low-income percentage varies between 12 to 44.

Future research directions were also identified. If accurate data becomes available, the team intends to measure the accessibility and the network connectivity, to understand travel and behavioral pattern for individuals during emergency situations. This would help planners and policy makers better examine alternative scenarios, improve the infrastructure system, address the needs of underserved communities, and measure the accessibility of different population segments for effective and equitable evacuation planning.

2 INTRODUCTION

Transportation engineers, planners, and policy makers are rigorously working to identify successful evacuation strategies, taking on one of the most challenging problems that today's societies face. With the unprecedented increase in frequency and intensity of climatic disasters experienced in the U.S. and around the world, which cost thousands of lives each year, developing methods and tools that would help identify effective evacuation plans remains a crucial research need. The conditions under which evacuations happen and the communities that are impacted may vary significantly depending on many factors including, for example, the nature and intensity of the hazard, geography, socio-demographics, institutional structure, and the preparedness level. However, it is an unfortunate fact that in many cases low-income, typically carless residents are the ones that take the hardest hit as observed in hurricanes Katrina and Maria. Therefore, it is crucial to develop specific strategies and plans that consider the accessibility of vulnerable populations to shelters and whether they have the means to travel.

Under an evacuation, people with personal cars typically choose to drive to destinations other than emergency shelters while low-income carless residents tend to rely on transit services, if provided, or personal connections. While not all low-income people lack access to personal vehicles, they form the majority of the public transportation users under normal conditions, and their transportation to shelters or safe zones needs to be provided by transit agencies under emergency conditions as well.

While man-made hazards can be considered random, natural hazards are not random events. In order to prepare for a disaster, the potential of a location for a disaster needs to be determined, which involves a risk assessment procedure. The risk is typically defined as probability of a disaster times the consequence to the human environment. The estimation of consequences to the human environment requires information on geographic and demographic characteristics of the impacted areas. In this study, we focus on identifying the location of low-income residents in conjunction with car-ownership status in a target area for evacuation. The main objective is to identify low-income carless residents, who impose particular challenges in preparing for evacuation, as they solely rely on public transportation or other transportation services. A robust method is needed to identify low-income people with no access to a car in a geographical area with a medium to high risk of a natural hazard so that effective evacuation transportation services can be planned.

The behavioral patterns and characteristics of people affect their car-ownership status and indirectly influence their decision to own a car. These effects can be captured by deriving the statistical correlation between characteristics of individuals and their car-ownership status by utilizing various data sources. It is

possible to identify carless people based on socio-demographic characteristics if strong correlation is obtained between their characteristics and car ownership status. The biggest challenge in this regard is that the characteristics of the whole population are usually either unavailable or inaccessible due to privacy and security reasons. Only the characteristics of a random sample are usually available for public use. In addition, these samples typically belong to a population that is larger than the geographic area that experiences the hazard, which usually is a small nest within the sampled area. An example is the Public Use Microdata Area (PUMA)-level information that is available from American Community Survey (ACS) data [1]. The PUMA-level information is gathered and updated annually for samples within each PUMA in the United States. Each PUMA consists of several areas with different families and income levels. Income values for each PUMA vary within a large range. An appropriate sublevel of PUMA that may reduce this variation is the census tract. Census tracts are geographically small enough to show low variability in income levels of their residents.

Fortunately, census tracts are perfectly nested within PUMAs. However, census tracts are very small, and therefore very few of the samples from ACS observations might fall into each census tract. For example, some census tracts have only two observations, which is not adequate to represent the whole population. This shortcoming motivated studies to find a practical solution to generate a synthetic population that is highly representative of the characteristics of the people within a small area. In this study, we develop a statistical method to synthetically represent the population characteristics of a small geography so that the necessary information for evacuation planning, such as car ownership and income level, can be obtained. Specifically, we use the data from large-scale samples (e.g., ACS and IRS data) and generate a synthetic population for smaller subareas (census tract). The generated synthetic population is then used for estimating car ownership of low-income residents in the subareas. The mapping of the results for the car-ownership model (i.e., the percentage of the people who do not own a car) can then be used in other transportation and optimization models to optimize evacuation logistics.

3 LITERATURE REVIEW

In the context of sociodemographic data analysis, either the individual-level data are not available or only a sample from the total population is available. Therefore, a synthetic population needs to be generated for an area of interest to be used for various purposes such as microsimulation models, optimization models, and, as is the case for this study, evacuation planning. Synthetic populations are used commonly as input for activity-based models in which the travel patterns of individuals and their transportation choices are modeled at a granular level [1-8]. In addition, synthetic populations have been used in establishing measures of accessibility [9-11]. There are a few population-synthesizing approaches, which generate population while preserving the general characteristics of the population and mask the identity of the people for privacy purposes. One of the widely used methods to generate synthetic population is the Iterative Proportional Fitting (IPF) technique, which was first used by Beckman et al. (1996) in transportation studies. They utilized IPF for U.S census data structure [12], and implemented the method into TRANSIMS, a microsimulation modelling software (Choupani and Mamdoohi 2016) [13]. Although Beckman et al. (1996) were the pioneers who proposed IPF in transportation-related studies, the original method dates back to the 1940s [14]. IPF updates the weights of the counts from a sample until the sum of the weights for all the attributes matches with the marginal values available for the area of the interest. This process includes fitting and allocation steps that consist of cross-tabulation, integer conversion, and selection. For detailed properties of the IPF, readers are referred to Ireland and Kullback (1968), Deming and Stephan (1940), Fienberg (1970), and Mosteller (1968) [14-17].

Although the IPF method is well established in transportation studies, there are multiple problems that can obscure the validity of the synthetic population. One of the important issues is the generation of zero cells, zero marginals, and table sparsity which is caused by zero or small values. Another issue is that IPF is capable of synthesizing at one level [17-19]. To overcome the zero-marginal issue, the Iterative Proportional Updating (IPU) method was developed by Ye et al. (2009) [6]. A few statistically driven population-synthesizing techniques have been proposed, but none of them were as successful as their IPF competitor was. A Bayesian inference system method proposed by Schafer (1997) [20], combinatorial optimization method proposed by Openshaw and L. Rao (1995) [21] and Williamson et al. (1998) [22], simulated annealing proposed by Voas and Williamson (2000) [23], and Monte Carlo Markov Chain (MCMC) proposed by Farooq et al. (2013) [24], are among these methods.

Another method that can be used in population synthesis is copulas. Copulas are robust methods that are useful in capturing the dependency between the variables of interest. While this unique feature of the copulas gained popularity among engineers, they have rarely been used in the transportation engineering

field. Kao et al. (2012) proposed a dependence preserving approach to synthesizing household characteristics by fitting multivariate normal distributions and modifying the covariance matrix [25]. Jeong et al. (2016) used copulas to validate the dependencies that have been captured by IPF methods [26]. Yang et al. (2019) used copula theory to predict the short-term passenger flow for high-speed railway [27].

In this study, we adopt the population synthesis method based on copula theory and first developed by Kaushik et al. (2019) [28]. The copula-based method allows the generation of synthetic agents at a very fine resolution (i.e., census tract level), which is particularly convenient for the problem of emergency evacuation. In what follows, the concept of copula will be described. Then, the methodology to obtain the synthetic population is described using data for Anne Arundel County, Maryland, at the census tract level. The synthetic population is then utilized to obtain the percentage of the people with low income. The people with no access to personal vehicles are also identified through a binomial logit model.

4 DATA, MODEL SPECIFICATION, AND METHODOLOGY

4.1 DATA DESCRIPTION

4.1.1 Anne Arundel County Application

Anne Arundel County in Maryland is located south of the City of Baltimore and west of the Chesapeake Bay. The county's east border is entirely water, including the Chesapeake Bay and its numerous tributaries, as well as the various rivers, creeks, streams, and inlets, covering 29 percent of the county's area. Because of its geo-spatial location combined with changing climatic conditions, the county is vulnerable to natural disasters like hurricanes, storm surges, and flooding. The county has experienced significant damage from tropical storm Isabel (September 2003), and hurricanes Ernesto (September 2006), Irene (August 2011), and Sandy (October 2012). The tragic examples of hurricanes Katrina (2005) and Maria (2017) that cost thousands of lives have taught that the evacuation of carless, typically low-income populations is crucial. Therefore, the need for a tool that could help planners, first responders, and others to ensure the safety of these vulnerable populations is essential. The population synthesis method we developed in this study is applied to Anne Arundel County as a foundation for such a tool.

Figure 1 shows the median household income for census tracts within Anne Arundel County in the State of Maryland. Anne Arundel County includes four PUMAs that are almost equal in terms of area. Within each PUMA there are a few census tracts that are occupied by people with low median household income. The

number of observations for some of the census tracts is too low, and therefore the implementation or validation of any accessibility model would be obscured using these few observations. Copulas provide practical tools to generate a synthetic population provided that a sample of attributes of the interest and their marginal distributions for the study region are available.

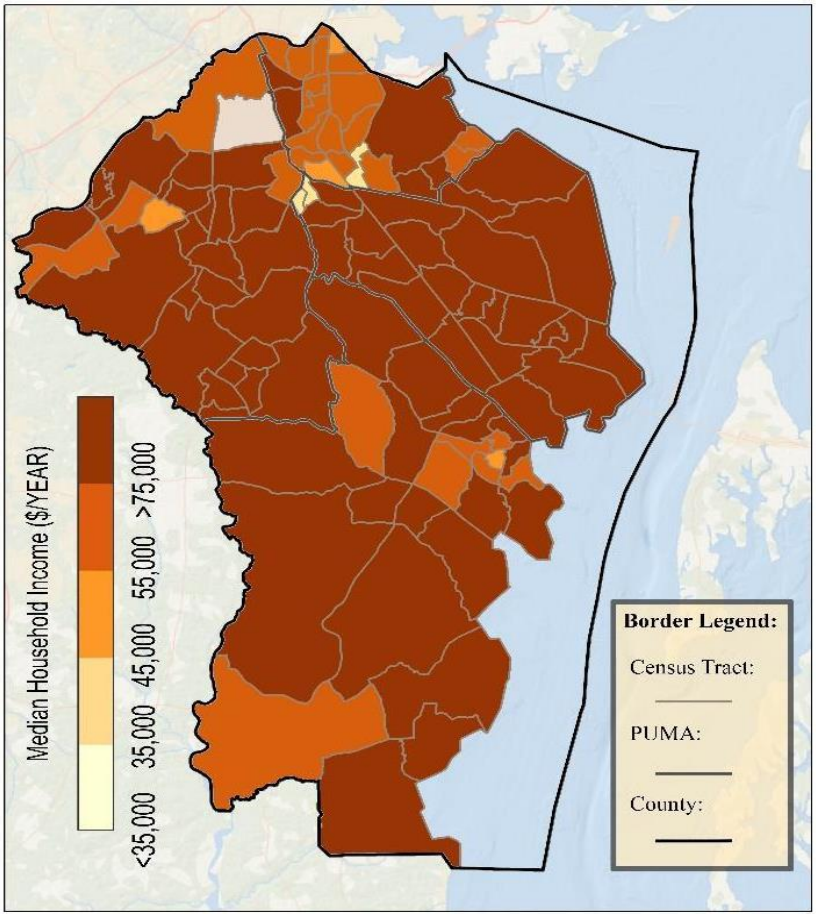

Figure 1: Median income for different census tracts within Anne Arundel County

4.1.2 Data Acquisition

We began by acquiring the spatial data for the study area, consisting of three layers, each finer layer nested within the larger layer. The geographic structure and the resolution of the data are depicted in Figure 2.

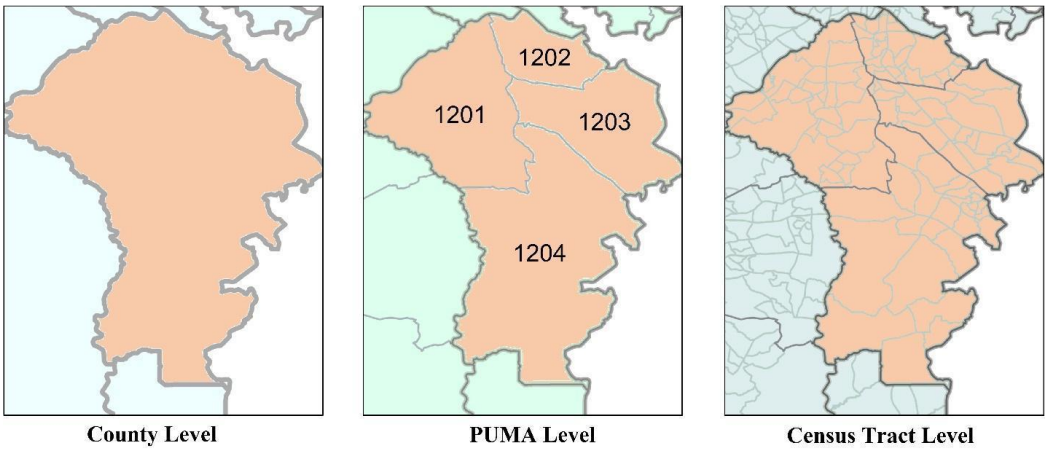

Figure 2: Anne Arundel County; PUMAs and census tracts

Anne Arundel County is divided into four PUMAs and each PUMA consists of several census tracts. Perfectly nested within each other, the structure of the data ensures that no data is mistakenly duplicated. In this study, the data from PUMA and census tract levels is combined with IRS data and used to generate the synthetic population, which to the best of authors' knowledge have not been consolidated before for population synthesis purposes. The three sources of data consolidated in this study are: (1) American Community Survey (ACS), (2) Decennial Census Data, and (3) IRS Income Data from tax returns.

The American Community Survey provides samples of information for Public Use Micro Areas (PUMAs), called Public Use Micro Samples (PUMS). PUMS consists of two different levels of the data: household level and individual level. These two levels of data share a common attribute (i.e., serial number) that is suitable for data integration purposes. The serial number indicates the attachment of each individual to a certain family. Each of the two data sets benefits from a weight column that determines the weight of the observation within the sample. The weights can be used to replicate the data, before the copula fitting process. The ACS suggests that for individual-level studies, the household-level data can be integrated with individual-level data, and then the integrated data can use the individual-level weights to replicate the observations. The attributes that are used in this study from the household level (HH) and individual level (IL) are listed below.

1. NP: Number of person records following the housing record (HH),
2. HHT: Household or family type (HH),
3. HINCP: Household income (past 12 months) (HH),
4. HUPAC: Household presence and age of children (HH),

5. WIF: Workers in family during the past 12 months (HH),
6. AGEP: Age of the person (IL),
7. SEX: Sex of the person (IL),
8. ESR: Employment status of the person record (IL),
9. RAC1P: Recorded detailed race code (IL).

These variables can be fitted to a copula as a multivariate problem. The copula can capture the dependency between these nine variables and generate pseudo-observations from the estimated copula.

As mentioned before, preserving the dependency between the variables is not enough for synthetic population generation. Another key piece of the puzzle is the Cumulative Distribution Function (CDF) of each variable. Inverse CDF of the variables transforms the pseudo-observations generated by the copula to real synthetic observations. In this study, two data sources have been used to construct CDF functions for the variables of interest. The first data is Decennial Census Data 2010, which includes tables required to generate CDF functions for six of the nine attributes mentioned above. These attributes are NP, HHT, HUPAC, AGEP, SEX, and RAC1P. Unfortunately, CDF functions for WIF and ESR are not currently available, and, therefore, the CDF for these two variables has been obtained from the PUMA samples. For income variable, which is the key attribute in identifying the low-income people, IRS data has been used to generate the CDF function. Figure 3 shows an example of data provided by the IRS that has been used to generate CDF functions for Income.

ZIP	INCOME	CLASS	N_return	N_single	N_joint	N_head	N_dpnd	N_TOT
20601	Total	0	13160	6120	3950	2520	8390	24930
20601	$1 under $25,000	1	3720	2790	230	620	1390	5260
20601	$25,000 under $50,000	2	2820	1450	450	750	1940	5040
20601	$50,000 under $75,000	3	2100	890	510	560	1560	4030
20601	$75,000 under $100,000	4	1640	540	660	340	1120	3320
20601	$100,000 under $200,000	5	2440	450	1690	250	1940	6020
20601	$200,000 or more	6	440	0	410	0	440	1260
20602	Total	0	13280	6390	3180	3190	9210	25150
20602	$1 under $25,000	1	4100	2880	240	890	1930	6180
20602	$25,000 under $50,000	2	3080	1560	390	980	2300	5620
20602	$50,000 under $75,000	3	2280	1000	490	650	1610	4240
20602	$75,000 under $100,000	4	1620	590	530	410	1270	3330
20602	$100,000 under $200,000	5	1990	360	1340	260	1880	5180
20602	$200,000 or more	6	210	0	190	0	220	600
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 3: A sample of data used from IRS data to generate CDF functions for Income

The IRS reports the number of returns for different types of income return forms. The report includes number of returns, number of single returns, number of joint returns, number of head of family returns, and number of dependents recorded. In order to obtain the CDF function for income, we doubled the record for the number of joint returns and summed with single, head of family, and dependents records. The last column in Figure 3 represents the calculated total returns for individual-level data. It should be noted that since the individual-level data has been fitted to the copulas, CDF function of the variables must be in individual level.

Another issue associated with IRS data is due to geographical resolution: The data is reported for zip codes; therefore, it cannot directly be used for generating the CDFs for census tract levels. Fortunately, the Office of Policy Development and Research provides a crosswalk between zip code-level and census tract-level data that indicates the percentage of each zip code that belongs to a census tract. The data consist of the ratio for residential and business addresses. In the IRS data case, since individuals tend to file their tax return forms based on their residential addresses, we used the residential ratio as a base to generate income CDFs for different census tracts. Figure 4 shows an example of the crosswalk for two census tracts.

ZIP	TRACT	STATE	RES_RATIO	BUS_RATIO	OTH_RATIO	TOT_RATIO
21054	24003702204	24	0.013561211	0.03500761	0.029126214	0.016115904
21114	24003702204	24	0.146610815	0.043153527	0.013071895	0.132833958
21114	24003702205	24	0.292650419	0.546058091	0.575163399	0.325260092
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 4: A sample of crosswalks between ZIP code and census tracts

4.2 METHODOLOGY

The mathematician Abe Sklar first introduced the copula theory and its unique feature into probability theory in 1959. Based on his definition, copulas are any kind of mathematical functions that perform a mapping between marginal distributions of a d- dimensional multivariate and their joint cumulative distribution:

$$C: \underbrace{[0,1]^d}_{marginal\ distributions\ of\ d\ dimentional\ multivarite\ data} \rightarrow [0,1]^1 \quad (1)$$

Where C is a copula function. In other words, the inputs of any copula function are the marginal distributions of the multivariate data, and the output is a unique value between zero and one, which is the joint cumulative distribution of the random variables. Assuming that H is the joint cumulative distribution function (CDF) of any *d*-dimensional continuous random variable, Sklar Theorem states that:

$$H(x_1, x_2, \dots, x_d) = C\{F_1(x_1), F_2(x_2), \dots, F_d(x_d)\} \qquad x_1, x_2, \dots, x_d \in \mathbb{R} \quad (2)$$

Where F_d is the marginal distribution of the d^{th} variable. By definition, an empirical copula can be constructed using the cumulative marginal distributions. For example, for a bivariate random variable, the empirical copula function can be written as follows:

$$C_n\big(f_1(x_1), f_2(x_2)\big) = \frac{1}{n}\sum_{i=1}^{n} 1(F_1(x_1) < f_1,\ F_2(x_2) < f_2) \quad (3)$$

Sklar (1959) showed that, for any multivariate set of continuous variables, there exists at least one functional (parametric) form of the copula that can perform the mapping mentioned in Equation 1. Several families of copulas have been described in the literature; each family consists of multiple copula functions. Copulas that belong to one family usually share some common property. For example, Archimedean copulas admit an explicit formula, and allow modeling dependency of multivariate random variables with only one parameter (θ), which is also the reason for its popularity. Table 3 shows the three most popular Archimedean copulas that are also used in this study.

Table 1: Archimedean three most popular Archimedean copulas

Copula	**Multivariate Copula C_θ**	**Range of the θ**
Clayton	$\left[\max\{f_1^{-\theta} + f_2^{-\theta} + \cdots f_d^{-\theta} - d + 1; 0\}\right]^{-1/\theta}$	$\theta \in [-1, \infty]\backslash\{0\}$
Frank	$-\frac{1}{\theta}\log\left[1 + \frac{\prod_{i=1}^{d}(\exp(-\theta f_1) - 1)}{\exp(-\theta) - 1}\right]$	$\theta \in \mathbb{R}\backslash\{0\}$
Gumble	$\exp\left[-\left(\sum_{i=1}^{d}(-\log f_1)^{\theta}\right)^{1/\theta}\right]$	$\theta \in [1, \infty)$

There are deterministic approaches estimating the copula parameters for bivariate random samples that rely on the dependency measurements. For multivariate samples, maximum log-likelihood estimation is a popular method. For big sample data, however, maximum log-likelihood estimation usually converges to a solution, although this method is computationally exhausting. The estimated parameter for a copula preserves the dependency of the variables that exists within a multivariate dataset. In order to reduce the longitudinal dimension of the problem (i.e., the number of observations), one can write the log-likelihood in its weighted form. In this form, the repeated observations within the data appear only once in the log-likelihood with their associated weights. Assuming that n^* is the modified sample number after deleting the identical observations and ω_i is the weight for each observation, one can write the log-likelihood function as follows:

$$\mathcal{L}(\theta) = \frac{1}{n^*}\sum_{i=1}^{n^*} \omega_i \log\left[C_\theta\left(\widehat{U}_i\right)\right] \tag{4}$$

$\widehat{U}$ in the above equation is a vector that preserves the normalized rank of the variables within the sample, which is called pseudo-observation. The concept of the rank is utilized to introduce the marginal distribution of the data to the copula. The rank of the variables preserves the dependency of the variables for a multivariate problem. This is because the rank of the data remains unchanged under any strictly decreasing or increasing transformation function. The rank of the variables divided by the number of observations augmented by one results in the normalized rank for each realization within the sample. Note that the copula maps the normalized ranks of the variables to the joint distribution of the variables.

Once the data is fitted to a copula, the fitted copula must be checked for goodness of fit. The Cramér-Von-Mises statistic can be used to perform a bootstrap algorithm and calculate the P-values for estimated parameters. The initial step of the bootstrap method is to fit the available data to a targeted copula. The Cramér-Von-Mises statistic is then constructed and calculated as the sum of squares for differences between empirical and fitted copulas (*Sn*). This procedure is repeated for *k* times (a statistically sufficient times needs to be selected), each time with the data obtained by sampling with replacement from the initial dataset. The bootstrap P-value can be calculated from the following equation (for a 95% confidence interval):

$$P = \frac{1}{K+1}\left[\sum_{i=1}^{K} 1(S_n^* \geq S_\mathrm{n}) + 0.5\right] \tag{5}$$

Once a copula has passed the statistical test for the goodness of fit, there are statistical methods to generate synthetic data from the fitted copula. It should be noted that estimated copulas produce pseudo-observations, which are realizations from the cumulative distribution function of the population.

The last piece of the problem that completes the procedure is the utilization of the inverse cumulative distribution function of each variable. Each pseudo-observation drawn from the copula is transformed into a real synthetic observation by using the inverse cumulative distribution function of that variable. Within the mentioned procedure, the copula preserves the dependency of the variables, while the CDF of the variables evokes the distribution and realization of the observations, independent of the copula. Figure 6 shows the flowchart of the entire procedure for generating the synthetic population. The process starts with a sample of observations as input. Then, ranks of variables in the data set are determined followed by computation of pseudo-observations. The pseudo-observations are then fitted to a copula and the goodness of fit test is performed using a bootstrapping algorithm. After the test, n observations are drawn from the pseudo-observations to represent the population and they are transformed into a synthetic population using CDFs.

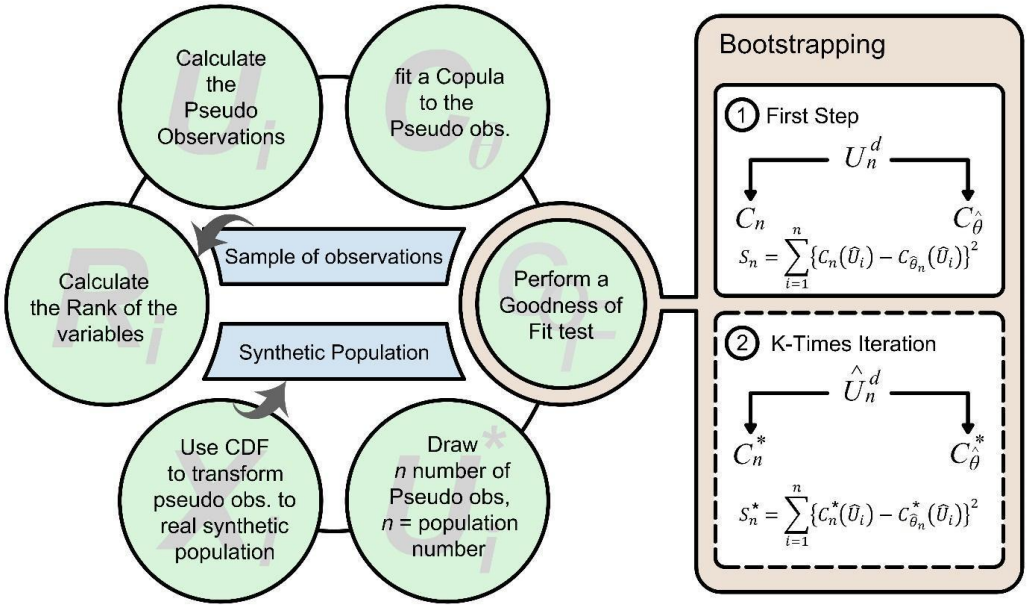

Figure 5: Flowchart of the population-synthesizing algorithm

4.3 MODEL ESTIMATION

4.3.1 Identification of Low-income Carless People

The pseudo-observations calculated for the nine input variables from ACS have been fitted to the four well-known Archimedean copulas (i.e., Clayton, Gumbel, Frank, and Joe copulas). The Clayton copula showed the lowest test statistic, and other than the Frank copula, *p*-values for all other copulas were significant within the 99% confidence interval. Therefore, we used the Clayton copula for population synthesis. The results of the Clayton copula used to generate the synthetic population are shown in Table 1. There are identical observations within the sample which are usually referred to as "ties." Existence of the ties within the sample needs to be accounted for in the calculation of the rank and pseudo-observations. The rank for ties can be set to the minimum, maximum, or average of the index assigned to the identical observations. In this study, the ranks of the variables have been obtained by replacing the index set of the ties with their mean value, and the pseudo-observations have been calculated accordingly. This method keeps the weighted distribution of the pseudo-observations close to the distribution of the real observations.

Table 2: Estimated parameters for Clayton copula for census tracts within Anne Arundel County

PUMA	Copula	Statistics	P-value	Estimated Parameter (θ)
1201	Clayton	156.74	0.9995	0.30082
1202	Clayton	139.65	0.9995	0.33478
1203	Clayton	173.25	0.9995	0.28861
1204	Clayton	158.7	0.9995	0.29214

Having obtained the estimation parameters, we have generated a set of pseudo-observations for each census tract. The size of each set is equal to the population of the corresponding census tract. The CDF functions, produced by using 2010 Decennial Census Data and IRS Income Return Data, have been used to obtain the real synthetic population. The distribution of the synthetic population has been compared with that of the real data from Decennial Census and IRS. The error is found to be statistically zero within the 99% confidence interval.

Once the synthetic population is ready, different models may benefit from the available anonymous data for the entire population of a census tract. These data are suitable for feeding input into activity-based

models, car-ownership, and other models. In this study, we have used a simple binomial logit model using Biogeme software to find the percentage of people who do not own a personal car to use as input for an example of natural hazard evacuation planning. The binomial model has been estimated using the 2017 National Household Travel Survey (NHTS) data for the State of Maryland, and the estimated model has been applied to the synthetic population. The authors evaluated a series of binomial models to find the best attributes that statically describe the car-ownership status of the people. Within any binomial model, there are two choices to make. The choices in our model have been set to no-car and 1+ car. The base choice (i.e., no-car) has been set to zero, which designates zero utility to no-car choice. The results of the initial models showed a high correlation between household size and attributes such as number of children and number of workers in the family. Therefore, we eliminated household size from the final model. In addition, income groups and family race have been aggregated to a fewer clusters because of the collinearity issue observed for some of the clusters. Equation 6 expresses the final utility function defined for the choice of having access to a personal car.

$$U = \beta_0 + \beta_{CH1} + \beta_{inc2} + \beta_{inc3} + \beta_{inc456} + \beta_{R1} + \beta_{R2} + \beta_{WIF} \quad (6)$$

Table 2 shows the results for the estimated binomial logit model, and associated statistical test for a 98% confidence interval.

Table 3: Results of binomial model used to estimate car ownership

Name	Value	Robust std. err	Robust t-test	P-value	significance
ASCN	0.00	Fixed			
ASCC	0.583	0.180	3.24	0.00	
CH	1.01	0.802	1.26	0.21	*
INC2	1.55	0.369	4.20	0.00	
INC3	3.65	0.468	7.81	0.00	
INC4	3.93	0.588	6.68	0.00	
RAC1	-1.35	0.562	-2.40	0.02	
RAC2	-2.06	0.795	-2.59	0.01	

Name	Value	Robust std. err	Robust t-test	P-value	significance
WIF	0.745	0.262	2.85	0.00	

Note: ASCN and ASCC: Alternative specific Constant for no-car and 1+car(s), respectively.

In Table 2, CH stands for the number of children in the family. INC2, INC3, and INC4 are dummy variables indicating the income class of each individual. The intra-thresholds between these three categories are $50,000 and $75,000 (dollars per year of family income). The base cluster is INC1, which indicates people with a family income of less than 25,000 $/year. RAC1 is a dummy variable that indicates if the individual belongs to a black family. RAC2 is a variable of the same type for Asian families. The aggregation of races other than Black and Asian is the base cluster for family race and dropped from the utility function to prevent a multi-collinearity issue. It can be seen from Table 2 that having more children imposes an additional utility to the choice of having a car, which is logical. As income grows, the utility of having car also grows. After the Asian race, being a black family imposes negative utility to the choice of owning a car. As the number of workers grows the utility of having a car grows positively as well.

Having predicted the car ownership model, the next step is to predict the probability of having one or more than one car for each individual within the synthetic population by using Equation 7.

$$P(U = 1|x) = \frac{e^{U_1}}{1 + e^{U_1}} \tag{7}$$

This probability has been aggregated over the entire population to find the percentage of the people who do not have access to a personal car. Figure 5 illustrates the car ownership and income level of the people living in different census tracts within Anne Arundel County.

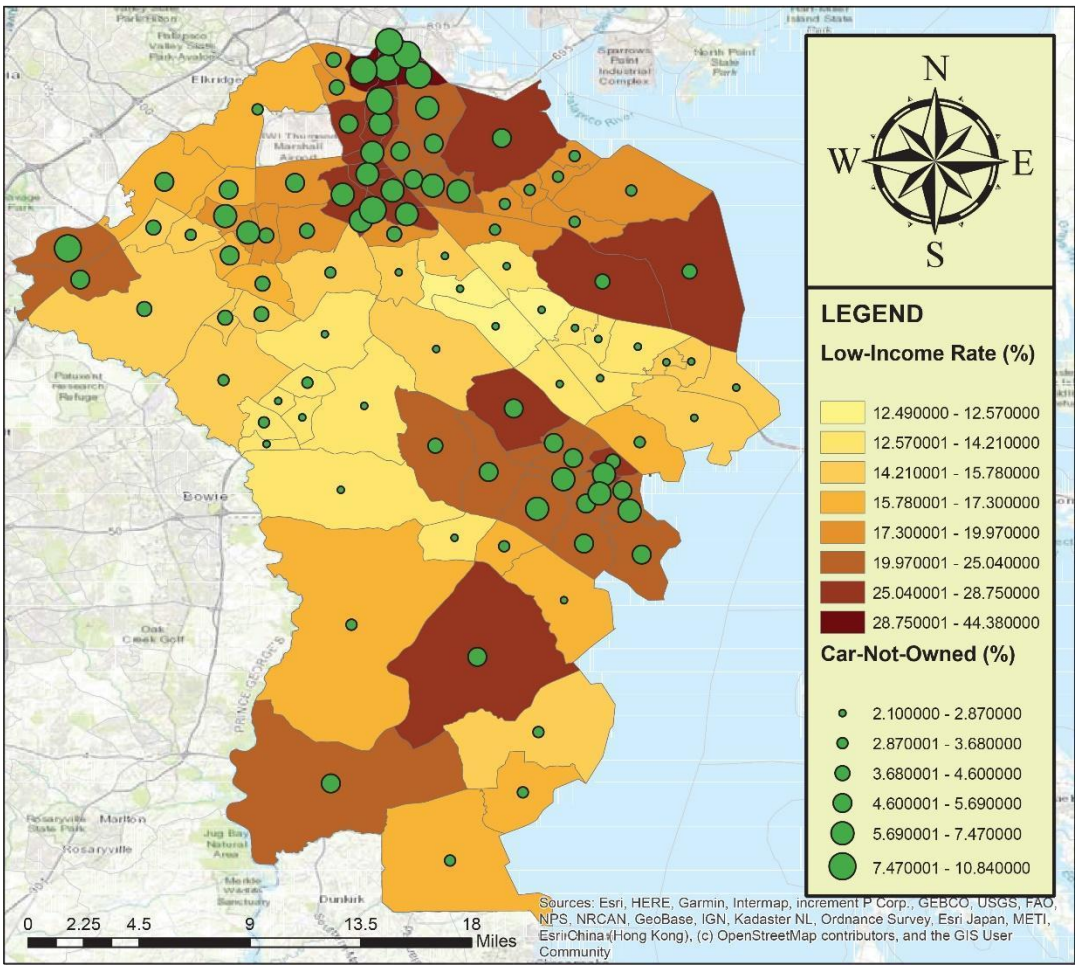

Figure 6: Car ownership and income level for census tracts within Anne Arundel County

It is clear from the figure that as the percentage of low-income people increases within a census tract, the percentage of people with no access to car also increases. We observe that most of the people with no car are located within the northern part of the county, which is adjacent to the southern area of Baltimore City. The percentage of people with no car varies between 2 percent and 11 percent, while the low-income percentage varies between 12 to 44. The percentage of low-income people we obtained is higher than the real percentages. The reason for this discrepancy is our simplifying assumption when defining low-income households: We assumed that people with an annual salary of less than 25,000 dollars belong to the low-income category, but the definition of low income is more complex, involving various thresholds by family types, e.g., number of people in the household and employment statuses. In addition, the available criteria for defining low-income households in the existing standards are used to identify the income status of a family rather than individuals. Since we are interested in the individual-level threshold, we assumed a single value of 25,000 dollars per year as a breakpoint to define a low-income individual.

5 CONCLUSIONS

This study presented a copula-based population synthesizer that interactively utilizes triangulated data to generate populations that maintain the dependency structure and the marginals of the real population. Within the proposed algorithm, the triangulated data is fitted to a copula, which preserves the dependency structure among different attributes of the dataset. This dependency can be defined as the hidden correlation that exists between the different variables of the multivariate problem. By fitting the available data to a copula one can preserve this correlation and produce new synthetic observations from the fitted copula. In this study, we fitted the data from Public Micro Area Samples (PUMS) to the Archimedean family of copulas. The individual- and household-level data have been merged to generate an individual level dataset that includes all the attributes of the interest for fitting purpose. Nine distinct variables that are mostly used in transportation modeling were picked from the PUMS data attributes. These nine variables have been used to estimate the Clayton copula for four Public Use Micro Areas (PUMA) within Anne Arundel County. The results of the estimation were very promising, and the validity of the estimation has been proved with robust statistical evidence.

Having fitted the copulas, a set of synthetic pseudo-observations can be drawn from the copula for any subregion of the PUMA of interest within Anne Arundel County. In this study, we set the subregions to census tracts. We transformed the pseudo-observations to real observations using the Inverse Cumulative Distribution functions for the attributes. This success in generating the synthetic population has been accompanied by a simple transportation application. A binomial car-ownership model has been estimated for the State of Maryland and examined on synthetic populations obtained for different census tracts. Through the generated synthetic population and the simple car ownership model, a census tract level map has been generated. The binomial model results, which were based on the copula-generated synthetic population, successfully captured the expected dependency between car-ownership status and income level of the individuals within the census tracts.

Despite these findings, challenges remain regarding synthetic population and associated analysis within the transportation modeling context. For instance, an accurate resemblance to the population of the census tracts requires replacing the CDF functions of the attributes from PUMS (e.g., employment status of the individuals and the number of workers in the household) with the marginals that are accurately obtained from Decennial surveys or the Department of Labor Statistics. Unfortunately, these data either are not available or not in suitable classification order to be used jointly with PUMS data. With updated data in

2020, it is possible to see more attributes recorded within the Decennial Census which leads to increased accuracy of the copula-based synthetic population.

As accurate data becomes available, future research could focus on using the copula-based synthetic population to create models to measure the accessibility and connectivity of areas that are, for example, majority low-income, disabled, or carless people. In addition, the proposed population synthesizing method can provide reliable and granular-level input for activity-based models that are particularly developed to understand the travel and behavioral pattern for individuals within large cities and small communities as well. These studies would help planners and policy makers better examine alternative scenarios to improve the infrastructure, address the needs of underserved communities, and measure the accessibility of different population segments for effective and equitable evacuation planning.

The report is based on the paper (20-05768) "A Statistical Approach to Synthetic Population Generation as a Basis for Carless Evacuation Planning" (20-0576 8) by Mohammad Nejad, Sevgi Erdogan, and Cinzia Cirillo presented at the 99th TRB Annual Meeting, Washington, D.C., January 2020.

6 REFERENCES

1. United States Census Bureau (2016). 2016 American Community Survey. https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/2016/

2. Arentze, T. A., and Timmermans, H. J. P. (Eds.) (2000). *ALBATROSS: a learning based transportation oriented simulation system*. Eindhoven: Technische Universiteit Eindhoven / EIRASS.

3. Salvini, P. and E. J. Miller (2005). "ILUTE: An operational prototype of a comprehensive microsimulation model of urban systems." *Networks and Spatial Economics* 5.2, pp. 217-234.

4. Pinjari, A. R., N. Eluru, R. B. Copperman, I. N. Sener, J. Y. Guo, S. Srinivasan, and C. R. Bhat (2006). *Activity-based travel-demand analysis for metropolitan areas in Texas: CEMDAP Models, Framework, Software Architecture and Application Results*. Tech. rep.

5. Guo, J. and C. Bhat (2007). "Population synthesis for microsimulating travel behavior." *Transportation Research Record: Journal of the Transportation Research Board* 2014, pp. 92–101.

6. Ye, X., K. Konduri, R. M. Pendyala, B. Sana, and P. Waddell (2009). "A methodology to match distributions of both household and person attributes in the generation of synthetic populations." *88th Annual Meeting of the Transportation Research Board, Washington, D.C.*

7. Bradley, M., J. L. Bowman, and B. Griesenbeck (2010). "SACSIM: An applied activity-based model system with fine-level spatial and temporal resolution." *Journal of Choice Modelling* 3.1, pp. 5–31.

8. Javanmardi, M., J. Auld, and K. Mohammadian (2011). "Integration of TRANSIMS with the ADAPTS Activity-Based Model." *4th Conference on Innovations in Travel Modeling*, Tampa, Florida.

9. O'Kelly, M. and Horner, M. Aggregate accessibility to population at the county level: U.S. 1940–2000. *Journal of Geographical Systems* (2003) Volume 5, Issue 1, pp 5–23.

10. Dong, X., Ben-Akiva, M.E. Bowman J.L., and Joan L. Walker J.L. Moving from trip-based to activity-based measures of accessibility. *Transportation Research Part A: Policy and Practice*, Volume 40, Issue 2, 2006, Pages 163-180.

11. Ziemke, D., Joubert, J. and Nagel, K. Accessibility in a Post-Apartheid City: Comparison of Two Approaches for Accessibility Computations. *Networks and Spatial Economics* (2018) vol. 18(2), pages 241-271.

12. Beckman, R. J., K. A. Baggerly, and M. D. McKay (1996). "Creating synthetic baseline populations." *Transportation Research Part A: Policy and Practice* 30.6, pp. 415–429.

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

14. Deming, W. E. and F. F. Stephan (1940). "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known." *The Annals of Mathematical Statistics* 11.4, pp. 427–444.

15. C. T. IRELAND, S. KULLBACK. "Contingency tables with given marginals," *Biometrika,* Volume 55, Issue 1, March 1968, Pages 179–188.

16. Fienberg, S. E., 1970. "An iterative procedure for estimation in contingency tables." *The Annals of Mathematical Statistics*, 41, 907-917

17. Mosteller, F., 1968. Association and estimation in contingency tables. *Journal of the American Statistical Association* 63, 1-28.

18. Pritchard, D. R. and E. J. Miller (2012). "Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously." *Transportation*, 39.3, pp. 685–704.

19. Pukelsheim, F., 2014. Biproportional matrix scaling and the iterative proportional fitting procedure. *Annals of Operations Research*. 215. pp. 269–283.

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

21. Openshaw, S. and L. Rao (1995). "Algorithms for reengineering 1991 Census geography." *Environment and Planning A* 27.3, pp. 425–446.

22. Williamson, P., M. Birkin, and P. H. Rees (1998). “The estimation of population microdata by using data from small area statistics and samples of anonymized records.” *Environment and Planning A* 30.5, pp. 785–816.

23. Voas, D. and P. Williamson (2000). “An evaluation of the combinatorial optimization approach to the creation of synthetic microdata.” *Population, Space and Place* 6.5, pp. 349–366.

24. Farooq, B., M. Bierlaire, R. Hurtubia, and G. Flötteröd (2013). “Simulation based population synthesis.” *Transportation Research Part B: Methodological* 58, pp. 243–263.

25. Kao, S.-C., H. Kim, C. Liu, X. Cui, and B. Bhaduri (2012). “Dependence-Preserving Approach to Synthesizing Household Characteristics.” *Transportation Research Record: Journal of the Transportation Research Board* 2302, pp. 192–200.

26. Jeong, B., W. Lee, D.-S. Kim, and H. Shin (2016). “Copula-Based Approach to Synthetic Population Generation.” *PloS one* 11.8, e0159496.

27. Yuedi Yang, Jun Liu, Minshu Ma, Xudong Jia, Xuchao Chen, Short-Term Passenger Flow Prediction of High-Speed Railway Based on Copula Theory. *Transportation Research Board 2019.*

28. Kartik Kaushik, Cinzia Cirillo, Fabian Bastin, Copula Based Synthetic Population Generation for Microsimulation (under review), 2019.