Simulating a Close-to-Reality Synthetic Population of the Greater Accra Region

Tyler J. Frazier, AICP, PhD
Senior Scientist, Department of Transportation System Planning and Telematics
Fakultät V, Technische Universität Berlin
Sekr. SG12, SG 4.1
Salzufer 17-19
10587 Berlin
+49 (0)30-314 22418 (office)
+49 (0)30-314-26269 (fax)
http://www.vsp.tu-berlin.de/frazier/
http://www.vsp.tu-berlin.de/gauss/
tyler.j.frazier@tu-berlin.de

Andreas Alfons, PhD (forthcoming)
ORSTAT Research Center
Faculty of Business and Economics
Katholieke Universiteit Leuven
Naamsestraat 69, bus 3555
3000 Leuven, Belgium
Phone: +32 16 326817
http://www.econ.kuleuven.be/andreas.alfons/public/
andreas.alfons@econ.kuleuven.be

Abstract

The purpose of this research is to simulate a synthetic population of the Greater Accra Metropolitan Region (GAMA) from the 2005 Ghana Living Standards Survey (GLSS5) for use in the Greater Accra Urban Simulation System (GAUSS). A primary goal in simulating the synthetic population of GAMA is to employ a method which generates *close-to-reality* population data rather than repeatedly drawing samples. In order to generate *close-to-reality* synthetic data, combinations which were not represented in the original household survey but are likely to occur in the true population must occur in the synthetically generated data. The author estimates the conditional distributions with multinomial logistic regression models in order to simulate categorical and continuous variables. The simulation of random zeros as opposed to structural zeros, are also reflected in the synthetically generated Greater Accra population. One of the main purposes for avoiding pure replication of units from the underlying sample is because this generally leads to small variability of units within smaller subgroups, which results in an increase in unrealistic model behavior when population data is used as input for agent-based simulations of urban dynamics.

Introduction

Synthetically generated population data is generally an important first step in running microsimulations or agent based models used to predict urban dynamics and/or transportation activities. Microsimulation models often attempt to reproduce the behavior of individual persons, households or firms over the course of several years in order to quantitatively and qualitatively visualize potential scenarios which could occur as well as their associated costs and benefits. In order to reduce potential prediction error, using population data that most closely reflects the existing population inhabiting the

geographic area of study is desirable. Generating this synthetic population has typically been achieved by either repeatedly drawing samples from sample data or using the iterative proportional fitting method (IPF), a common method employed by transportation models. By simulating the existing population data, a realistic framework for comparing the implementation of different policy cadres (business as usual, weak sustainability, strong sustainability) under different growth scenarios (low, medium or high economic or demographic growth rates) can be projected.

One of the advantages of synthetic data is its cost effectiveness when compared to comprehensive and detailed population data, which is in effect nearly impossible to obtain for every living person inhabiting a significantly sized urban geographical area. Additionally, generating synthetic data serves to meet the need for observing statistical disclosure limitations. Generating synthetic data not only presents the researcher with the base year data needed to simulate different potential urban simulation scenarios it also presents the public statistician with the means for releasing these base year datasets for practical application, while protecting rights to privacy as well as maintaining the likelihood of receiving authentic data from individual survey observations. (Reiter 2007)

In order to generate a synthetic population from a sample such as the GLSS5, several conditions need to be met. First the actual size of regions and strata must be reflected in the survey weights. Secondly, marginal distributions and interaction between variables should be reflected correctly, while heterogeneities between subgroups, especially regional aspects, should be allowed. Finally, pure replication of units from the underlying sample should be avoided, as this generally leads to extremely small variability of units within smaller subgroups. Following these conditions, the synthetic data should include univariate distributions overall and in subpopulations as well as multivariate relations among the variables. In order to meet these conditions, multinomial logistic regressions can be used to predict possible outcomes of a dependent variable from probabilities derived from a given set of independent variables. Also used in synthetically generating the household structure, categorical and continuous variables for Great Accra is the conditional probability distribution which is the probability distribution of variable Y when variable X is known to be a particular value. (Munnich et al 2003, Alfons et al 2010)

Application to Ghana Living Standard Survey 5 (GLSS)

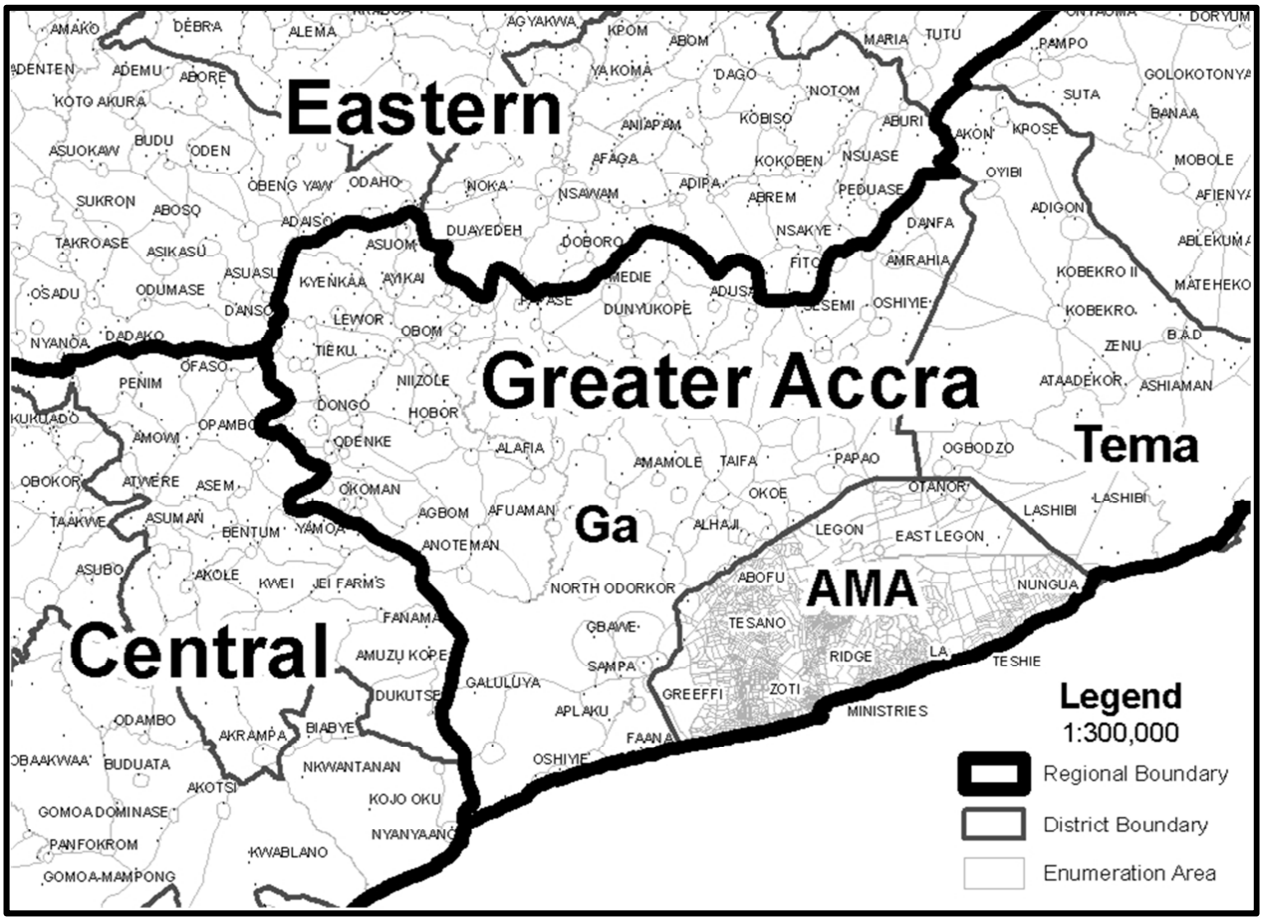
The Ghana Living Standards Survey-Round Five (GLSS 5), like earlier ones, focuses on the household as a key socio-economic unit and provides valuable insights into living conditions in Ghana. The fifth round of the GLSS was conducted by the Ghana Statistical Service (GSS) from 4th September 2005 to 3rd September 2006. A nationally representative sample of 8,687 households in 580 enumeration areas, containing 37,128 households members were covered in GLSS5. Detailed information was collected on demographic characteristics of respondents and all aspects of living conditions including health, education, housing, household income, consumption and expenditure, credit, assets and savings, prices and employment. For the purposes of this work, sections on Demography, Education and Employment were used. While the initial intent was to simulate the entire population of Ghana from the GLSS5 for the chosen variables and focus on the urban population of the Ghanaian capital, the total GLSS5 sample size of 37,128 household members overextended hardware capabilities (particularly RAM), and thus the scope was limited to the Greater Accra Region. The Greater Accra Region portion of the GLSS5 is comprised of 4254 persons as members of 1257 households. (GSS 2008)

The synthetic population generation cannot be applied to the GLSS5 directly if the data includes missing attributes from observations. While in the univariate case the observations with missing information could simply be deleted, this can result in a severe loss of information in the multivariate case. Multivariate observations usually form the rows of a data matrix, and deleting an entire row implies that cells carrying available information are lost for the analysis. Instead of deleting observations with

Table 1: Variables of the GLSS5 used in **simPopulation** application

Variable (missing counts)	Name	Type
Region (0)	`region`	Categorical
Enumeration Area (0)	`cluster`	Strata
Gender (0)	`sex`	Categorical
Age (0)	`age`	Categorical
Nationality (0)	`nation`	Categorical
Ethnicity/Tribal Affiliation (175)	`ethnic`	Categorical
Religion (2)	`religion`	Categorical
Sample weights (0)	`weight`	Continuous
Educational attainment (259)	`highest_degree`	Categorical
Occupation (618)	`occupation`	Categorical
Household Income (737)	`annual_income`	Continuous
Total persons 4254 / total households 1257		

Map 1: Enumeration Areas throughout Greater Accra

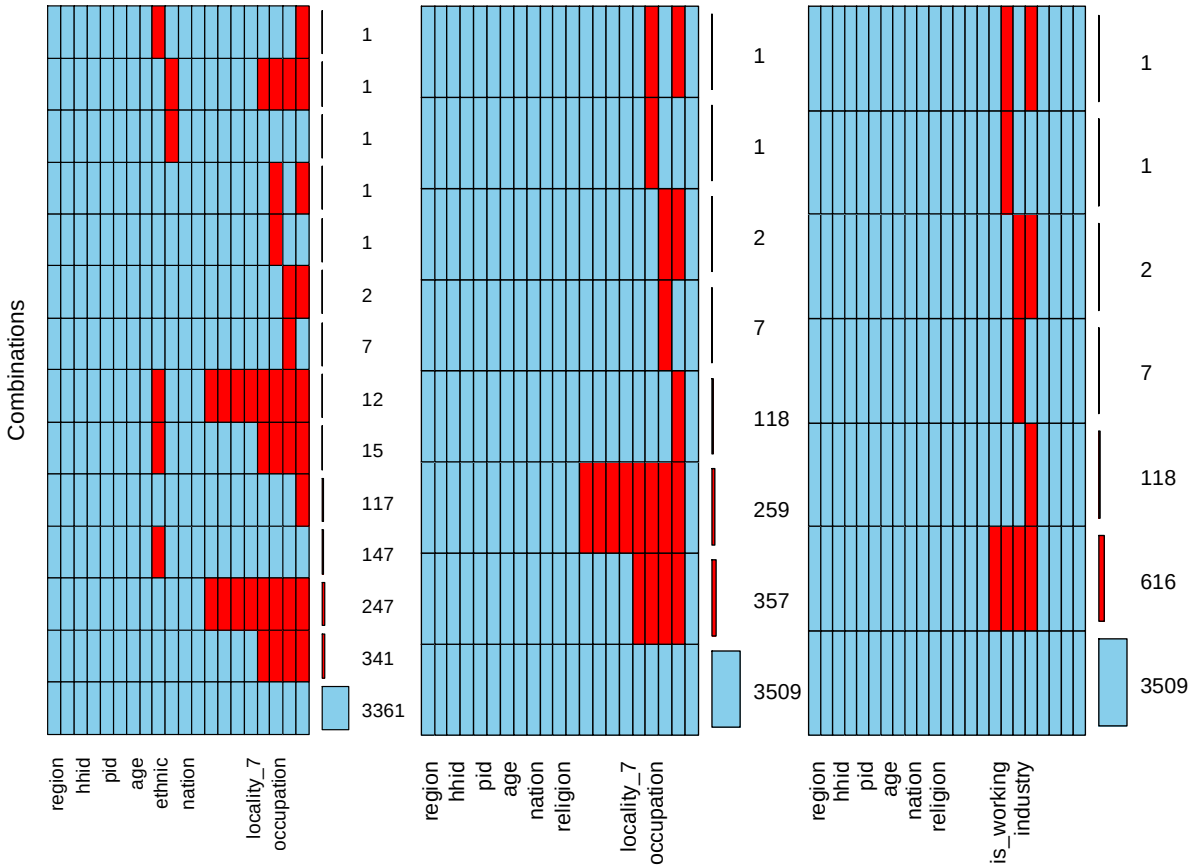


missing values, it is better to fill in the missing cells with appropriate values, which is possible with multivariate data sets. (Hron 2008)

Section 1 addresses the primary demographic household characteristics for all 4254 persons, with the exception of the variables ethnic and religion. The GLSS5 only defined tribal affiliation for members who were born in Ghana, thus the 175 "missing" observations were recoded to "unknown." This was done due to the fact that valid inferences can only be made if the missing data are missing completely at random (MCAR), which was not true in this instance. (Little and Rubin 1987) Additionally, there were 2 missing attributes for the variable religion, which were imputed from the existing 4254 members in order to obtain a complete set of observations for all Section 1 variables for Greater Accra.

Section 2 of the GLSS5 addressed characteristics related to educational achievement while Section 4 was related to employment and household income. Four variables from Section 2 were used; with each one have 259 missing observations from the original survey data. This data was imputed using 9 variables from Section 1 using the k-nearest neighbor (KNN) imputation method; in the same manner the 2 missing attributes from the variable religion were imputed. Once Sections 1 and 2 were both complete, the 616 to 737 missing attribute data from Section 4 were imputed, again using the kNN method and the variables from Sections 1 and 2, thus resulting in a complete data set for all 4254 household members

Table 2. Combinations of Missing data during kNN Imputation: original data (left), after imputing Section 1 (middle), after imputing section 2 (right)



residing in Greater Accra, and enabling application of the synthetic population generation method.

Many different methods for imputation have been developed over the last few decades. While univariate methods replace the missing values by the coordinate-wise mean or median, the more advisable multivariate methods are based on similarities among the objects and/or variables. A typical distance based method is k-nearest neighbor (KNN) imputation, where the information of the nearest k>=1 complete observations is used to estimate the missing values using the Aitchison distance for measuring compositional datasets. While kNN is numerically stable it has some limitations. First the optimal number of k nearest neighbors needs to be determined, by randomly setting observed cells to missing, estimating these values and measuring the error. Secondly, kNN imputation does not fully account for the multivariate relations between the compositional parts, which are only considered indirectly when searching for the k-nearest neighbors. A next step in this process will be to apply a model-based imputation procedure which relies on a more realistic estimation of the multivariate data structure. (Hron et al 2008)

Application of SimPopulation to GLSS5

The household structure is simulated separately for each combination of stratum k and household size l. First, the number of households is estimated using the Horvitz-Thompson estimator:

$$M_{kl} := \sum_{h \in H_{kl}^S} w_h$$

where H_{kl}^S denotes the index set of households in stratum k of the survey data with household size l, and w_h, $h \in H_{kl}^S$, are the corresponding household weights. (Hortvitz and Thompson 1952) To prevent unrealistic structures in the population households, basic information from the survey households is resampled. (Alfons et al, 2010) Using the R package simPopulation, we start our analysis using the function `simStructure()` and enter the following command from the R command prompt.

```
gamaP <- simStructure(gamaI, hid = "hhid", w = "weight",
strata = "cluster", additional = c("age","sex"))
```

Additional categorical variables are simulated using the simCategorical function which estimates conditional distributions with multinomial logistic regression models for each stratum using survey indices to fit responses and predictors while incorporating survey weights. (Alfons et al, 2010) In order to reduce computation time, age categories are combined into categorical groups, before proceeding with the `simCategorical()` function. The argument `basic` specifies existing generated variables found in the household structure, while the argument `additional` specifies the variables to be simulated in this step.

```
basic <- c("ageCat","sex","hsize")

gamaP_Cat <- simCategorical(gamaI, gamaP, w = "weight",
strata = "cluster", basic = basic, additional = c(“nation”,
”ethnic”, ”religion”, ”highest_degree”, "occupation"))
```

Next the function `simContinuous()` is used to simulate the variable annual income with the `basic` argument modified to include additional predictor variables. This approach is able to handle semi-continuous variables, i.e. variables that contain a large amount of zeros, which is true with regard to the variable annual_income in the GLSS5. Following the approach used for simulating categorical variables, the continuous variable is discretized by breakpoints and zero becoming a category of its own. Multinomial logistic regression models are then fitted for every stratum k separately, as previously described in order to simulate the continuous variable. Finally the values of the variable are generated by random draws from uniform distributions within the corresponding categories.

```
basic <- c("ageCat","sex","hsize", "nation", "ethnic",
"religion", "highest_degree", "occupation")
gamaP_Cont <- simContinuous(gamaI, gamaP, w = "weight",
strata = "cluster", basic = basic, additional =
c("annual_income"))
```

Evaluation of the Simulated Synthetic Population of Greater Accra

In this section the relationship between categorical variables, including variables defining the household structure are evaluated using contingency coefficients. Pearson's coefficient of contingency is a measure of association for categorical data and is defined as

$$P = \sqrt{\frac{X^2}{n + X^2}}$$

where X^2 is the test statistic of the X^2test of independence and n is the number of observations. Tables 3 and 4 present the contingency coefficients obtained from the sample as well as those from the synthetic Greater Accra population. The relative differences are negligible in all instances with the correlation structure of the simulated population being very close to that found in the GLSS5 after application of kNN imputation.

The result is the synthetic generation of 3,111,779 persons being described by the variables for household size, age, sex, religion, educational attainment and occupation, while the variables for nationality, ethnicity and household income will be subsequently included. This synthetic population is ready to be used as the base year data set in an urban simulation system or transportation model. More specifically, the R package **simPopulation()** has been applied to the Ghana Living Standard Survey 5 and presented as a vignette. Additionally, the methodology of using multinomial logistic regression models and the conditional probability distribution was presented and explained. Descriptive mosaic plots, cumulative distribution plots, and box-and-whisker plots will also be added to further demonstrate the effectiveness of the employed method as further emphasis of the validity demonstrated by comparing the pairwise contingency coefficients of the GLSS5 with the Synthetic Population. In conclusion, this work presents a synthetic urban population of Accra, Ghana, which is a step forward towards generating synthetic Close-to-Reality populations with combinations not represented in the original household survey but likely to occur in the true population.

Table 3. Pairwise Contingency Coefficients from the GLSS5 after Imputation

	sex	hsize	religion	highest_degree	occupation
age	0.07684275	0.3885837	0.2888835	0.7238775	0.7198963
sex	NA	0.1370375	0.1008813	0.1912632	0.3971899
hsize	NA	NA	0.5395080	0.3009451	0.5537457
religion	NA	NA	NA	0.3287186	0.5677181
highest_degree	NA	NA	NA	NA	0.8378192

Table 4. Pairwise Contingency Coefficients from the Synthetically Generated Greater Accra Population

	sex	hsize	religion	highest_degree	occupation
age	0.07684068	0.3889738	0.2905272	0.7201150	0.7146574
sex	NA	0.1374701	0.1011180	0.1932895	0.3778930
hsize	NA	NA	0.5396357	0.2947910	0.5464265
religion	NA	NA	NA	0.3288038	0.5583367
highest_degree	NA	NA	NA	NA	0.8280430

References

Alfons A. and Kraft S. (2010). Package simPopulation. R package version 0.2.1.

Alfons A., Templ M. and Filzmoser P. (2010a). "Simulation of EU-SILC Population Data: Using the R Package simPopulation." Research Report, Department of Statistics and Probability Theory, Vienna University of Technology. Vienna, Austria.

Alfons A., Templ M. and Filzmoser P. (2010b). "Simulation of synthetic population data for household surveys with application to EU-SILC." Research Report, Department of Statistics and Probability Theory, Vienna University of Technology. Vienna, Austria.

Frick, Martin, Axhausen, Kay W. (2004). Generating Synthetic Populations using IPF and Monte Carlo Techniques: Some New Results. Conference paper STRC 2004, Swiss Transport Research Conference. Zurich, Switzerland.

Ghana Statistical Service (2008). Ghana Living Standard Survey: Report of the Fifth Round (GLSS5). Accra, Ghana.

Hron K., Templ M. and Filzmoser P. (2008). "Imputation of missing values for compositional data using classical and robust methods." Research Report, Department of Statistics and Probability Theory, Vienna University of Technology. Vienna, Austria.

Raghunathan T.E., Reiter J and Rubin D.B (2003). “Multiple Imputation for Statistical Disclosure Limitation.” Research Report, Department of Biostatistics and Institute of Social Research, University of Michigan. Ann Arbor, Michigan, USA.

Reiter Jerome (2007). “Selecting the Number of Imputed Datasets When using Multiple Imputation for Missing Data and Disclosure Limitation.” Research Report, Institute of Statistics and Decision Sciences, Duke University. Durham, North Carolina, USA.

Reiter Jerome (2005). “Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study.” Statistics and Society A, Vol 168, Part 1, pp. 185-205.

Reiter Jerome (2004) “Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation.” Research Report, Institute of Statistics and Decision Sciences, Duke University. Durham, North Carolina, USA.

Reiter J., Raghunathan T. and Kinney S (2006). “The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data.” Research Report, Institute of Statistics and Decision Sciences, Duke University. Durham, North Carolina, USA.

Templ M. and Alfons A (2009). “An application of VIM, the R package for visualization of missing values, to EU-SILC data.” Research Report, Department of Statistics and Probability Theory, Vienna University of Technology. Vienna, Austria.

Ye, Xin, Konduri, Karthik, Pendyala, Ram M., Sana, Bhargava, Waddell, Paul (2009). A Methodology to Match Distributions of Both Household and Person Attributes in the Generation of Synthetic Populations. Transportation Research Board and Department of Civil and Environmental Engineering, Arizona State University. Tempe, Arizona, USA.