2012

Generating a Close-to-Reality Synthetic Population of Ghana

Tyler Frazier

Andreas Alfons

Recommended Citation

Frazier, Tyler and Alfons, Andreas, Generating a Close-to-Reality Synthetic Population of Ghana (2012). *SSRN*.
https://www.doi.org/10.2139/ssrn.2086345

Generating a Close-to-Reality Synthetic Population of Ghana

Tyler Frazier*†and Andreas Alfons‡

Abstract

The purpose of this research is to generate a close-to-reality synthetic human population for use in a geosimulation of urban dynamics. Two commonly accepted approaches to generating synthetic human populations are Iterative Proportional Fitting (IPF) and Resampling with Replacement. While these methods are effective at reproducing one instance of the probability model describing the survey, it is an instance with extremely small variability amongst subgroups and is very unlikely to be the real population. IPF and Resampling with Replacement also rely on pure replication of units from the underlying sample which can increase unrealistic model behavior. In this work we present a sequential logic for estimating variables using multinomial logistic regressions and the conditional probabilities amongst each variable in order to generate combinations which were not represented in the original survey but are likely to occur in the real population. We also present a model based approach to imputing missing observation responses and apply the methodology to the Ghana Living Standard Survey 5 (GLSS5) in order to generate a comprehensive synthetic population for the Republic of Ghana, including such household and person variables as household size, tribal affiliation, educational attainment and annual income, amongst others. The R language and environment for statistical computing was used as well as the packages `VIM` and `simPopulation` in developing and executing the code. Contingency coefficients, cumulative distributions, mosaic plots, and box plots are presented for evaluation in order to demonstrate the effectiveness of the new method in its application to Ghana.

1 Introduction

One of the major challenges of our increasingly complex and urbanizing modern world is the conflict arising from expressed individual rights such as pursuing economic liberties or democratic freedoms with the need to regulate and control that environment in a manner which maintains social welfare and preserves the common good. Multi-agent systems, agent-based models, micro and/or geosimulations are becoming increas-

*Department of Transportation Planning and Telematics, TU Berlin
†correspondances to tyler.j.frazier@tu-berlin.de
‡ORSTAT Research Center, Faculty of Business and Economics, KU Leuven

ingly accepted approaches to predicting urban dynamics and their intrinsic emergent phenomenon, such as land use and development or transportation activities as they enable stakeholders to visualize large cities and regions as computer animated, scalable, graphical representations comprised of their highly disaggregate parts. Improving the predictive power of a geosimulation likewise improves the ability of stakeholders to 1) understand their urban phenomenon and dynamics and 2) manage and improve governance systems, social services, public infrastructures and other mechanisms which enable and promote a sustainable urban and regional environment. The use of geosimulations within the development context where institutional development has historically been lacking, is of the utmost importance for making lasting and meaningful contributions towards the transformative process of a sustainable urban future (Frazier, 2011b).

A geosimulation is defined here as a collection or series of different demographic, economic, statistical and engineering models which are used to predict urban dynamics and their associated emerging phenomenon in terms of the individual behavior of agents (decisions of persons, households, businesses and institutions) within a defined environment, usually a large city and its region. In order to better understand and predict activities and decisions through different time scales, research serving to integrate short term agent behavior (such as decisions about temporary locations which is generally exhibited as transportation activities in days and weeks) with long term agent behavior (such as decisions about more permanent locations which is generally exhibited as decisions regarding land use over months and years) is becoming increasingly important. Integrating different time scales requires accurate and detailed data, which describes all of the individual agents located throughout a city and region at the beginning point in time of the simulation. This dataset is commonly referred to as the base year data, and Figure 1 presents a schematic overview of some of these datasets, tables and variables used in an urban simulation system. This work presents a method for generating close-to-reality synthetic population data from a survey sample, which is one of the initial steps in the process of constructing and developing a geosimulation. The base year data generated by the presented methodology is used in the Greater Accra Urban Simulation System (GAUSS; Frazier, 2011a), which was contructed within the Open Platform for Urban Simulation (OPUS; Waddell, 2002, 2011).

Due to the size and cost, it is generally not feasible to conduct a census which includes comprehensive and detailed attribute data for a large population such as a mega-city, region or country. A more reasonable approach to creating base year data is to generate a synthetic population from a sample. Generating synthetic data is not only cost effective but it also serves to protect individual and institutional rights to privacy as well as improving the likelihood of receiving authentic data from individual survey observations in the future. Using advanced statistical modeling techniques, it is possible to generate a complete synthetic population with statistical parameters very similar to those of the real population, and thus accurately reflecting the demographics of the real population at the point in time the survey was conducted.

Perhaps the most common and widespread approach to generating a synthetic human

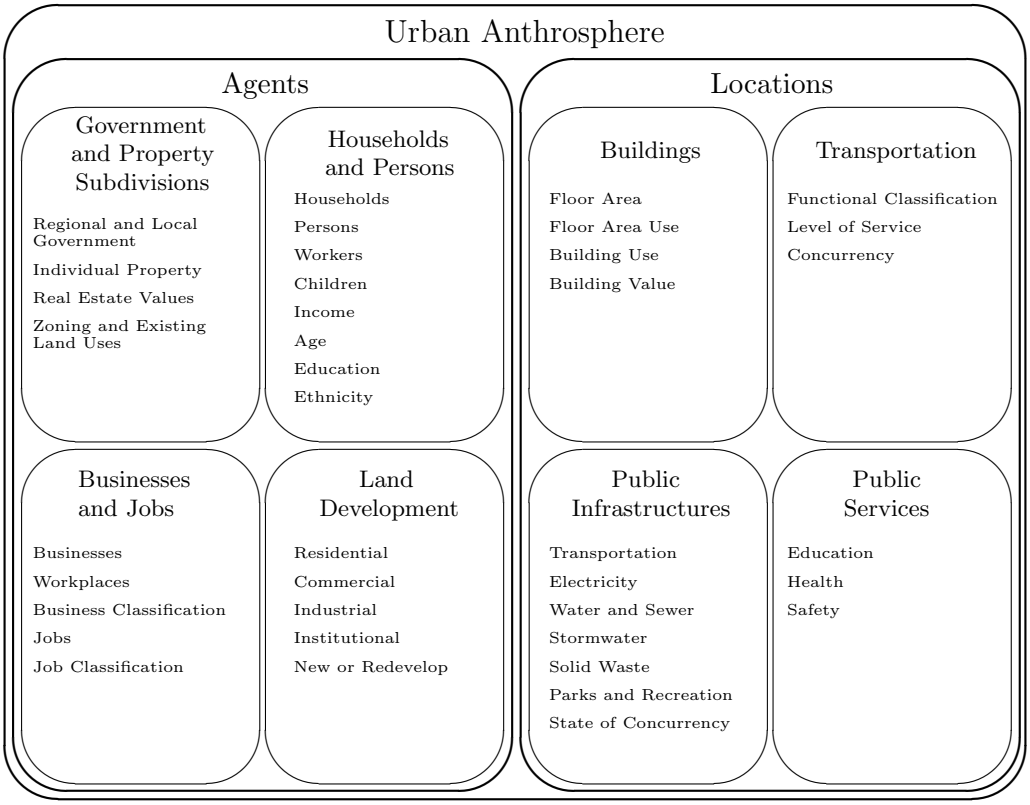

Figure 1: Some of the datasets, tables and variables used as base year data in an urban simulation system.

population for simulating urban dynamics is a method first introduced by Beckman et al. (1996) which combined Iterative Proportional Fitting with a Monte Carlo simulation by integrating detailed but aggregated control variables with generalized data but for more disaggregated target areas. The IPF method yields the constrained maximum entropy for a set of variables by proportionately estimating and controlling the joint and marginal distributions while preserving the survey's correlation structure. Then the procedure samples with replacement complete survey observations to produce a synthetic population with a joint distribution consistent with the constrained maximum entropy (Guo and Bhat, 2007). The IPF method has been traditionally popular in the United States and Europe, where it is common practice to remove sample weights and remove or aggregate specific location identifiers to meet statistical disclosure limitations, since IPF enables generating a plausible synthetic population from primary sampling units

without such information.[1]

It has not been demonstrated that the IPF method is capable of generating a synthetic population which is consistent with the sample design used to create the survey. This becomes more evident when attempting to match hierarchical distributions, for example from households and the persons who are members of each household. Several efforts have been made to develop algorithms for matching these distributions such that the observation weights for households, as well as for all persons within those households, are equal (e.g., Ye et al., 2009). Additionally, the IPF method results in a large majority of incorrect zero cell values in the contingency table largely due to the sample size's likelihood for not representing all of the possible combinations of attributes occurring in the real population (the zero cell problem; see Guo and Bhat, 2007). Pritchard and Miller (2009) also recognized the introduction of bias by the "integerization" procedure for rounding counts used in the Monte Carlo simulation when drawing observations from the survey data.

Most IPF research fails to address missing responses within individual survey observations except to delete entire observations with missing values. Conducting a demographic and economic survey for a large population is an incredible task and investment, which due to the logistics will inherently have some missing observations. One reason is because national or regional surveys are conducted in stages, and at each successive stage the number of observations typically reduce in number, leading to missing data when joined to the master file. Additionally, interviewers, data input personnel, or the responder may contribute to the missing data. If we were only considering responses from a single variable then missing data could be dropped, but since we are considering responses from many variables across thousands of respondents, deleting all of the observations that carry missing responses results in a severe loss of information. Deleting an entire observation also implies that cells carrying available information are noncontributory and/or insignificant. Instead of deleting entire observations with missing values, a better strategy for multivariate data sets is to impute missing responses.

A second commonly accepted practice to generating synthetic population is Resampling with Replacement from the survey itself based on the weights and locations. This approach has the advantage of permitting a synthetic population to be generated without making assumptions about the form of the population or its distribution. Like IPF, Resampling with Replacement also suffers from the zero cell problem. Frazier (2011a) provides an example, using the Ghana Living Standard Survey 5 (GLSS5; Ghana Statistical Service, 2008), a two-stage stratified random sample of 8,687 households containing 37,128 persons from 580 enumeration areas, to generate the GAUSS base year data, a synthetic population of 186,593 persons residing in 43,763 households as well as 79,413

[1] An alternative and possibly more popular approach in the UK is called hill climbing or combinatorial optimization (e.g., Huang and Williamson, 2001; Ryan et al., 2007; Harland et al., 2012), which randomly populates disaggregated spatial subdivisions with household observations and randomly swaps households across the landscape to match the real and simulated marginals, in accordance with several different proposed algorithms.

jobs throughout Accra's Korle Bu district.

In all of the aforementioned examples, the resulting synthetic data is never comprised of anything more than a combination of the complete observations from the original sample. The real population is far more heterogeneous, and includes many more combinations of possible outcomes than found in the survey. Both IPF and Resampling with Replacement, will generate one instance of the probability model, but it is an instance with extremely small variability of units within subgroups and is very unlikely to be the real population. In order to simulate a population which is close-to-reality, a statistical method is needed which is capable of generating combinations which were not represented in the original survey but are likely to occur in the real population. The framework of generating a close-to-reality synthetic population by drawing from conditional distributions was first presented by Münnich and Schürle (2003); Münnich et al. (2003), and was later significantly modified and extended to estimate the conditional distributions via multinomial models by Alfons et al. (2011), motivated by an application to EU-SILC data. In this work we extend and apply this sequential logic for estimating variables using multinomial logistic regression models to the Ghana Living Standard Survey 5 (GLSS5) in order to simulate a close-to-reality population of the Republic of Ghana.

Finally we conclude our introduction by distinguishing between original survey data and generating synthetic survey data. Rather than devising an algorithm which attempts to unmask the intentionally encrypted sample design or invest significant amounts of time into building institutional capacity for gaining access to original, raw survey data, a preferable method was first discussed by Rubin (1993), to generate fully synthetic microdata sets using multiple imputation which not only provide researchers and modelers with weights and locations, but also preserves the identity of the respondent. Raghunathan et al. (2003) and Reiter (2009) develop and extend the methodology for generating synthetic sample data from a survey, which is likely the preferable methodology going forward since it meets Statistical Disclosure Limitations as well as provides the capability to generate a close-to-reality synthetic population.

2 Methodology

The GLSS5 has been provided by the Ghana Statistical Service as Stata files (`.dta`) and the raw data have been imported to R (R Development Core Team, 2011). Then several modifications to the original data have been made with regard to new variables needed, simplifying variables by aggregation, and in some instances editing raw data responses based on conclusions drawn from the survey's logic. The variables household size and annual income were added by calculations from the existing dataset; in the instance of annual income, the amount was normalized from different periods of payment and converted from old to new Ghana Cedis. The variables for ethnicity and occupation were aggregated in order to simplify the numerous response categories in accordance with the

Table 1: Variables from the Ghana Living Standard Survey 5.

Variable	No. Observations	Type	No. Outcomes
Region	37,128	Macro Strata	10
Cluster	37,128	Micro Strata	580
Sex	37,128	Categorical	2
Age	37,128	Categorical	21
Household Relationship	37,128	Categorical	10
Household Size	37,128	Categorical	29
Nationality	37,128	Categorical	11
Tribal Ethnicity	37,128	Categorical	10
Religion	37,118	Categorical	11
Highest Degree	36,656	Categorical	16
Occupation	29,334	Categorical	11
Annual Income	23,498	Semi-Continuous	–

data provided by the ISSER / Yale University code book (Economic Growth Center, Yale University and Institute of Statistical, Social and Economic Research, University of Ghana, Legon, 2009). Upon completion of the modifications for each of the three sections, the data was merged into a single R object.

Table 1 lists each variable used in this work as well as the number of observations for each as found in the original survey data. Popular approaches for impuation in multivariate data are distance-based methods such as k-nearest neighbor (kNN), or model-based methods using regression techniques. The kNN approach is popular but has some limitations when data has outliers, is skewed or has multimodal distributions, all of which are frequent situations with practical data sets. For this reason, model-based imputation, which applies regression methods to the EM-algorithm's iterative procedure of estimating, adapting and re-estimating, have been increasing in popularity. However, model-based imputation typically requires much more computation time than distance-based methods, in particular for large data sets. The R package `VIM` provides functions for kNN imputation based on a variation of the Gower distance (Gower, 1971), which is suitable for complex survey data, as well as the iterative robust model-based imputation (IRMI) algorithm by Templ et al. (2011). In addition, it also includes several other practical tools for visualizing missing survey responses (see Templ et al., 2012). Due to the large number of observations in GLSS5, the results in this paper are based on kNN imputation using the implementation in `VIM`. Once the GLSS5 has been edited and missing values imputed, a data set of 37,128 observations, each with a complete set of responses, has then been prepared in order to apply the methodology of Alfons et al. (2011) for generating a synthetic population of Ghana.

Using the sample weights the Ghana Statistical Service has provided for households (and persons), we begin generating our synthetic population with computing the expansion factors

Table 2: Categorized variables created for use as predictors during the simulation.

Variable	Categories
Age Category	[0,6], (6,8], (8,10], (10,12], (12,14], (14,16], (16,18], (18,20], (30,35], (35,40], (40,45], (45,50], (50,55], (55,60], (60,65], (65,70], (70,75], (75,80], (80,99]
Annual Income Category	[0], (0, 50], (50, 80], (80, 156], (156, 312], (312, 600], (600, 1 040], (1 040, 1 800], (1 800, 2 600], (2 600, 4 160], (4 160, 7 680], (7 680, 13 000], (13 000, 38 400], >38 400

$$\hat{M}_{kl} := \sum_{h \in H_{kl}^S} w_h \tag{1}$$

where $\hat{M}_{kl}$ is the number of households in region k with household size l as estimated by the summation of all weights w_h for each household h found in the corresponding index set of households H_{kl}^S from the GLSS5.

Since no information is yet available on the population level, a basic household structure needs to be resampled with replacement from the survey data. With a very small number of variables that have a limited number of possible responses, it is still likely that all unique response combinations occurring in the real population are represented in the sample.[2] Additionally, resampling basic information from survey households serves to prevent unrealistic structures in the synthetic population. For each population household h in region k with household size l, our next step is therefore to select a survey household $h' \in H_{kl}^S$ with probability $w_{h'}/\hat{M}_{kl}$ and set the population household structure (given by the variables age, sex and household relationship) to

$$x_{hij}^U := x_{h'ij}^S, \qquad i = 1, \ldots, l, \tag{2}$$

where x_{hij}^U denotes the value of person i from household h in variable j for the population data, and $x_{h'ij}^S$ is defined accordingly for the sample data. Note that for the remaining steps, the variable age is then categorized as indicated in Table 2 to avoid having too many possible combinations of outcomes among the categorical variables.

With the addition of each new categorical variable, the number of potential combinations of responses increases exponentially, and the likelihood that resampling complete observations will be capable of realistically capturing conditional probabilities decreases. In order to accurately generate combinations which are likely to occur in the real population but do not appear in the survey dataset, the conditional distributions amongst variable responses are estimated via multinomial logistic regression models. In the following, the basic procedure to generate the variables nationality, ethnicity, religion,

[2] assuming a relatively large sample size

highest degree and occupation is described. We use age category, sex and household size as predictors in the multinomial models, with each new variable being iteratively added to the predictors for the generation of the remaining variables. Furthermore, separate models are fitted for each region to better account for regional heterogeneities. With x_{ij}^U denoting the value of population individual i in the response variable and $x_{i1}^U, \ldots, x_{i,j-1}^U$ denoting the corresponding values in the predictors, the conditional probabilities $p_{ir}^U := P(x_{ij}^U = r | x_{i1}^U, \ldots, x_{i,j-1}^U)$ are estimated by

$$\begin{aligned} \hat{p}_{i1}^U &:= \frac{1}{1 + \sum_{r=2}^{R} \exp(\hat{\beta}_{0r} + \hat{\beta}_{1r} x_{i1}^U + \ldots + \hat{\beta}_{j-1,r} x_{i,j-1}^U)}, \\ \hat{p}_{ir}^U &:= \frac{\exp(\hat{\beta}_{0r} + \hat{\beta}_{1r} x_{i1}^U + \ldots + \hat{\beta}_{j-1,r} x_{i,j-1}^U)}{1 + \sum_{r=2}^{R} \exp(\hat{\beta}_{0r} + \hat{\beta}_{1r} x_{i1}^U + \ldots + \hat{\beta}_{j-1,r} x_{i,j-1}^U)}, \quad r = 2, \ldots, R, \end{aligned} \tag{3}$$

where $\hat{\beta}_{0r}, \ldots, \hat{\beta}_{j-1,r}$ are the estimated coefficients from the multinomial logistic regression in the corresponding region, and R is the number of possible outcomes of the response variable. The values x_{ij}^U are then drawn from these estimated conditional distributions for each individual i in the population.

Nevertheless, there are strong dependencies for the values of nationality, ethnicity and religion among household members. Based on the information in the variable household relationship, the procedure of Alfons et al. (2011) is extended to a two-step procedure for the generation of these three variables. First, the values of the household heads are generated with a multinomial model as described above. Second, the values of the other individuals are generated with another multinomial model, using the values of the respective household heads as an additional predictor.

In order to simulate the variable for annual income, it is necessary to use a method that is capable of handling a semi-continuous variable, i.e., a variable that contains a large number of zeros, which is particularly true with the GLSS5. Hence we use a method based on the generation of categorical variables that is recommended by Alfons et al. (2011). The idea is to divide the data into relatively small subsets such that the empirical distribution is well reflected in the simulated population variable. We therefore discretize the semi-continuous variable annual income as indicated in Table 2, with zero being defined as category of its own. Then the income category of each population individual is generated via fitting a separate multinomial logistic regression model for each region, using all previously generated variables as predictors. Finally, the values of annual income are simulated by random draws from the respective income category. To be more precise, this is done by random draws of a uniform distribution within the category, except for the two highest income categories. There the values are drawn from a (truncated) generalized Pareto distribution (see, e.g., Kleiber and Kotz, 2003), which is a distribution frequently used in economics to model high incomes.

The R commands used to generate the synthetic population of Ghana are included in package `simPopulation` as a demo, which can be run by typing the following command

into the R console:

```
R> demo("ghana", package = "simPopulation")
```

However, the original GLSS5 data are of course confidential, therefore synthetic GLSS5 data are included in the package and used for the demo.

3 Results and Evaluation

In this section, the structure of the simulated categorical variables is evaluated with mosaic plots and contingency coefficients, while the semi-continuous variable annual income is evaluated by investigating the cumulative distribution function as well as producing conditional box plots.

First we create mosaic plots, which are a graphical representation of the cells in a contingency table. The frequencies are thereby represented by the area of rectangles. Figure 2 illustrates the expected and realized frequencies of region, sex and household size *(left)* as well as sex, ethnicity and occupation *(right)*. The left plot presents a visualization for the resampled basic household structure, while the right plot also includes selected variables generated with multinomial logistic regression models. Both show very similar structures in the sample and population data, demonstrating the accurate reflection of interactions amongst the categorical variables in the synthetic population. While the left plots are nearly indistinguishable, the plots on the right reveal some minor differences between the sample and population data upon close inspection. These differences

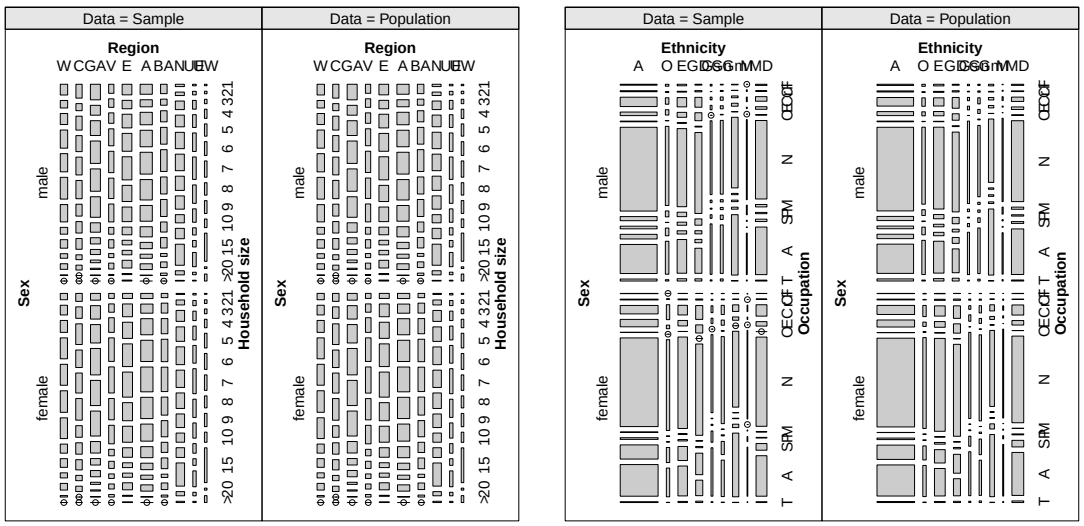

Figure 2: Mosaic Plots of sex, region and household size *(left)*, as well as sex, ethnicity and occupation *(right)*

Table 3: Pearson's pairwise contingency coefficients from the GLSS5 *(top)*, from the synthetic population *(middle)*, as well as the relative differences in *(bottom)*.

	age	sex	nation	ethnic	religion	highest_degree	occupation
hsize	0.41	0.10	0.14	0.44	0.33	0.24	0.28
age		0.09	0.20	0.19	0.25	0.65	0.61
sex			0.02	0.02	0.07	0.13	0.18
nation				0.57	0.20	0.09	0.10
ethnic					0.62	0.24	0.19
religion						0.25	0.20
highest_degree							0.62
	age	sex	nation	ethnic	religion	highest_degree	occupation
hsize	0.41	0.10	0.20	0.39	0.31	0.27	0.30
age		0.09	0.12	0.15	0.20	0.63	0.60
sex			0.02	0.02	0.08	0.12	0.18
nation				0.54	0.18	0.10	0.09
ethnic					0.60	0.25	0.19
religion						0.27	0.20
highest_degree							0.60
	age	sex	nation	ethnic	religion	highest_degree	occupation
hsize	0.01	−0.19	47.83	−11.98	−3.94	9.87	6.13
age		0.08	−39.34	−24.50	−19.59	−3.26	−2.23
sex			9.12	17.10	4.60	−2.15	−1.56
nation				−4.91	−8.22	11.07	−6.96
ethnic					−3.90	3.55	−2.07
religion						6.60	−3.52
highest_degree							−1.89

are primarily due to the fact that the multinomial models permit simulating combinations that do not occur in the sample, but are in fact likely to occur in reality. When considered in combination with the fact that the expected frequencies of the different combinations are solely determined by the sum of the corresponding sample weights, such slight differences may be interpreted as corrections to the expected frequencies.

Since presenting mosaic plots for all possible combinations of categorical variables would exponentially increase this length of this work, the generated population is further assessed by contingency coefficients. Pearson's coefficient of contingency measures the association amongst categorical variables and is defined as $P = \sqrt{\chi^2/(n + \chi^2)}$, where χ^2 is the test statistic of the χ^2 test of independence and n is the number of observations (see, e.g., Kendall and Stuart, 1967). Table 3 compares the pairwise contingency coefficients obtained from the sample and the population by their relative differences. Only some coefficients including age or nationality present some degree of error. For age, this can be explained with using age categories rather than age itself in the simulation of the additional variables. Nationality, on the other hand, is dominated by one category (about 98% of the individuals are Ghanaian by birth), leading to a poor χ^2 approximation. Also the relative difference for the contingency coefficients of sex and

ethnicity is rather large, but this is due to the contingency coefficients being close to 0. All in all, the contingency coefficients indicate relatively good association, which we also confirmed by inspecting further mosaic plots that are not shown in this paper.

For simulating annual income, all other variables are used as predictors in models computed for each region. We therefore not only investigate the generated income with respect to the distribution over the entire population, but also with respect to heterogeneities between subgroups. In Figure 3 *(left)*, the cumulative distribution functions (CDF) of annual income in the sample and the synthetic population are compared, and indicate an excellent fit. It is also worth noting that about 65% of the individuals have zero income. In the following, the income distribution within and between subgroups is evaluated with box plots. The box plots are therby generated only from the nonzero observations, while the proportion of individuals with nonzero income is reflected in the box widths. Figure 3 *(right)* shows the resulting box plots for annual income with regard to sex. Furthermore, Figure 4 contains box plots of the conditional distributions of annual income with respect to region *(top left)*, ethnicity *(top right)*, occupation *(bottom left)* and highest degree *(bottom right)*. The proportion of zeros and the distributions of the non-zero observations appear to be in general well reflected in the synthetic population, which again underlines the good fit of the models and illustrates their success in accounting for heterogeneities found in the GLSS5 survey.

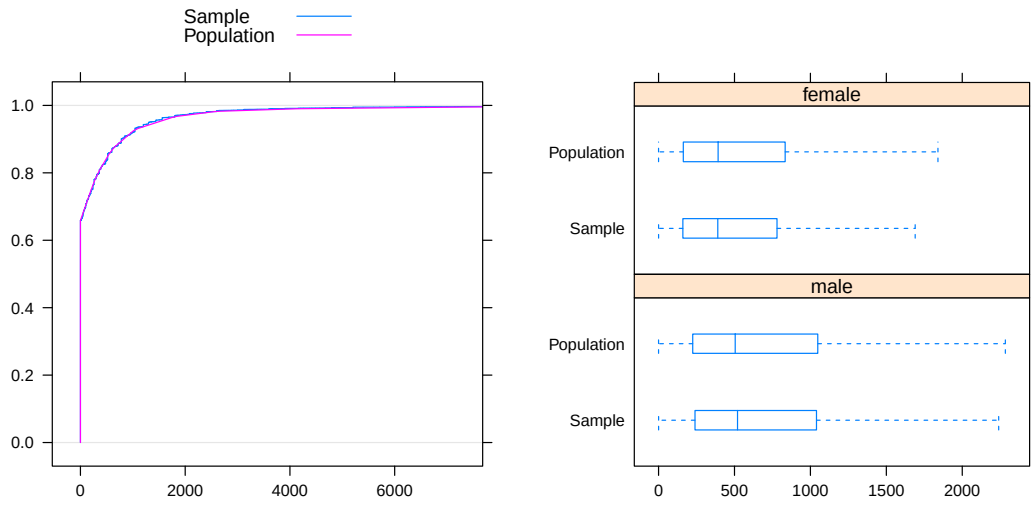

Figure 3: Cumulative distribution functions of income (*left*) and box plots of income by sex (*right*).

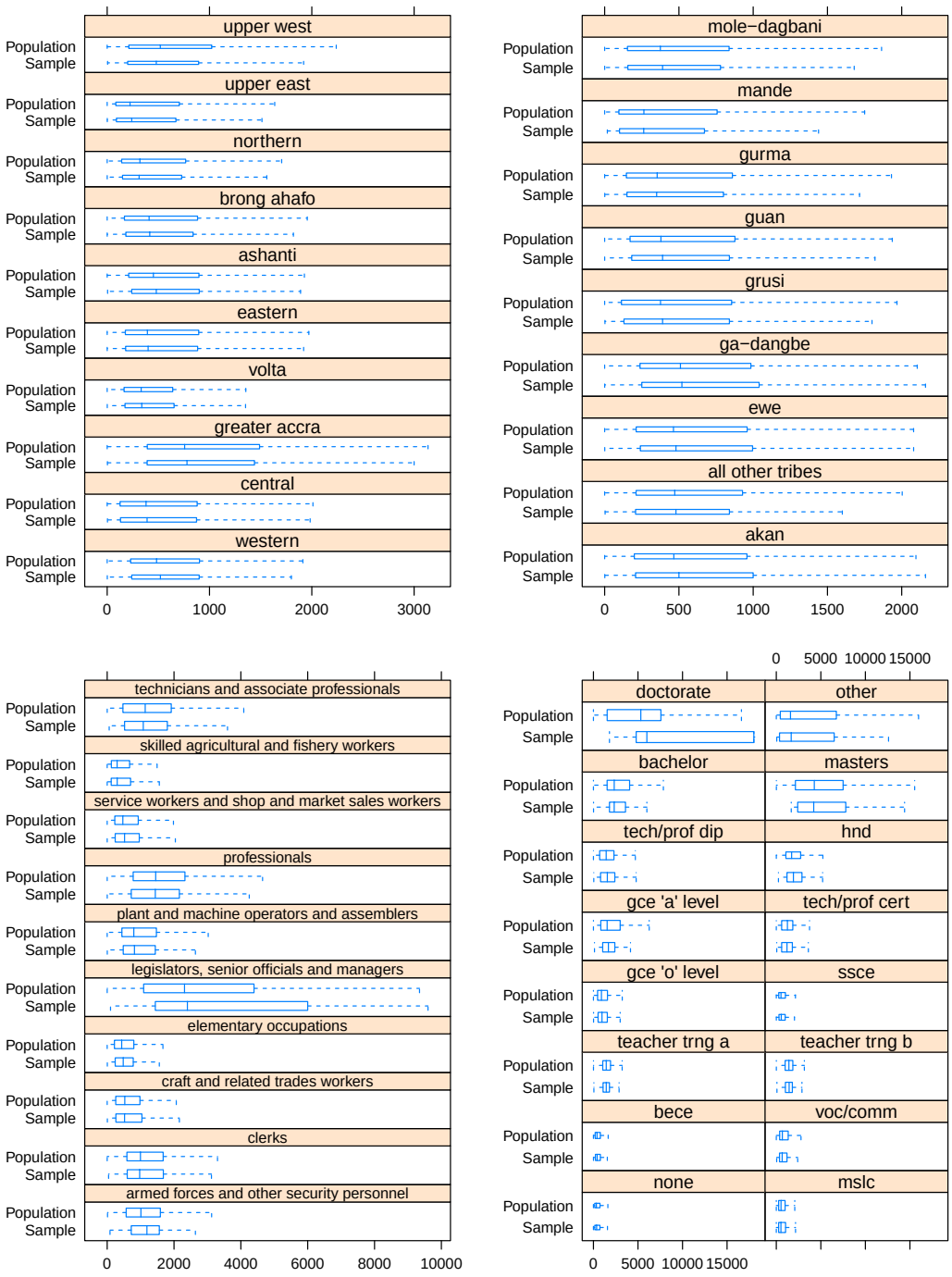

Figure 4: Box plots of income according to region (*top left*), ethnicity (*top right*), occupation (*bottom left*) and highest degree (*bottom right*).

4 Conclusion

This paper presents a method for generating a synthetic population of Ghana based on the GLSS5 primarily using model based regression methods in order to realize combinations in the synthetic data which are likely, but do not occur in the sample. A comprehensive discussion of the methodology, its application as well as graphical and quantitative means for evaluating the results have been presented, through incorporating the R package `simPopulation` as well as `VIM`. The result is a synthetically generated population of 22,620,269 Ghanaians, across all 10 regions and 110 districts, with each person being described in terms of their household structure, individual attributes and income. The evaluation section both graphically illustrates and quantitatively demonstrates the successful application of the proposed method to the GLSS5 for use as the base year data set in the Greater Accra Urban Simulation System (GAUSS). Since only a proportionate number of enumeration areas were selected from each region, a "next step" will be to further disaggregate synthetic household locations, either by a spatial regression method which uses the survey enumeration area's location, or by a modification of the method described by Waddell (2011) which identifies each household within their larger geography and then runs a residential location choice model for all households to choose their dwelling unit at the beginning of the simulation.

Acknowledgments The authors would like to thank the Volkswagen Stiftung for their support of this research as well as associated project work related to simulating urban dynamics in the developing world.

References

A. Alfons, S. Kraft, M. Templ, and P. Filzmoser. Simulation of close-to-reality population data for household surveys with application to EU-SILC. *Statistical Methods & Applications*, 20 (3):383–407, 2011.

R.J. Beckman, K. Baggerly, and M.D. McKay. Creating synthetic baseline populations. *Transportation Research A*, 30:415–429, 1996.

Economic Growth Center, Yale University and Institute of Statistical, Social and Economic Research, University of Ghana, Legon. EGC/ISSER Socio-Economic Panel Survey Code Book. Technical report, 2009.

T. Frazier. *Powering Accra: Projecting Electricity Demand for Ghana's Capital City*. Number 85 in Ecology and Development Series. Cuvillier Verlag Göttingen, 2011a.

T. Frazier. Transforming Accra towards a sustainable future: Comprehensive land use planning and the Greater Accra Urban Simulation System (GAUSS). *Viewpoints*, V, 2011b. URL `http://ugec.org/docs/ViewpointsV_final.pdf`.

Ghana Statistical Service. Ghana living standard survey: Report of the fifth round. Technical report, Ghana Statistical Service, 2008.

J.C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4): 857–871, 1971.

J.Y. Guo and C.R. Bhat. Population synthesis for microsimulating travel behavior. *Transportation Research Record: Journal of the Transportation Research Board*, 2014:92–101, 2007.

K. Harland, A. Heppenstall, D. Smith, and M. Birkin. Creating realistic synthetic populations at varying spatial scales: A compariatve critique of population syntheis techniques. *Journal of Artificail Societies and Social Simualtions*, 15, 2012.

Z. Huang and P. Williamson. A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata. *Working Paper, Department of Geography, University of Liverpool*, 2001.

M.G. Kendall and A. Stuart. *The Advanced Theory of Statistics*, volume 2. Charles Griffin & Co. Ltd., London, 2nd edition, 1967.

C. Kleiber and S. Kotz. *Statistical Size Distributions in Economics and Actuarial Sciences*. John Wiley & Sons, Hoboken, New Jersey, 2003. ISBN 0-471-15064-9.

R. Münnich and J. Schürle. On the simulation of complex universes in the case of applying the German Microcensus. DACSEIS research paper series No. 4, University of Tübingen, 2003.

R. Münnich, J. Schürle, W. Bihler, H.-J. Boonstra, P. Knotterus, N. Nieuwenbroek, A. Haslinger, S. Laaksonen, D. Eckmair, A. Quatember, H. Wagner, J.-P. Renfer, U. Oetliker, and R. Wiegert. Monte Carlo simulation study of European surveys. DACSEIS Deliverables D3.1 and D3.2, University of Tübingen, 2003.

D.R. Pritchard and E.J. Miller. Advances in population synthesis: Fitting many attributes per agent and fitting to household and person margins simulataneously. *Transportation*, 2009.

T. Raghunathan, J. Reiter, and D. Rubin. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19:1–16, 2003.

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

J. Reiter. Using multiple imputation to integrate and disseminate confidential microdata. *International Statistics Review*, 77:179–195, 2009.

D. Rubin. Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9:461–468, 1993.

J. Ryan, H. Maoh, and P. Kanaroglou. Population synthesis: Comparing the major techniques using a small, complete population of firms. *Working Paper, Center for Spatial Analysis, McMaster University*, 2007.

M. Templ, A. Kowarik, and P. Filzmoser. Iterative stepwise regression imputation using standard and robust methods. *Computational Statistics & Data Analysis*, 55(10):2793–2806, 2011.

M. Templ, A. Alfons, and P. Filzmoser. Exploring incomplete data using visualization techniques. *Advances in Data Analyis and Classification*, 2012. DOI 10.1007/s11634-011-0102-y, to appear.

P. Waddell. UrbanSim: Modeling urban development for land use, transportation and environmental planning. *Journal of the American Planning Association*, 68:297–314, 2002.

P. Waddell. Integrated land use and transportation planning and modelling: Addressing challenges in research and practice. *Transport Reviews*, 31:209–229, 2011.

X. Ye, K. Konduri, R. Pendyala, B. Sana, and P. Waddell. 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, 2009.