How Real Are Synthetic Populations?

Juliet Nakamya; Elke Moons; Geert Wets

Transportation Research Institute, Hasselt University. Wetenschapspark 5, bus 6, B-3590 Diepenbeek,Belgium.

E-mail: juliet.nakamya@uhasselt.be; elke.moons@uhasselt.be; geert.wets@uhasselt.be

Introduction

Within the field of activity-based analysis and transportation research in general, there has been growing interest in the generation and use of synthetic populations over the years. Several disaggregate land use and activity-based travel models, which represent decisions and actions of individual persons and households, incorporate microsimulations. In these models, synthetic populations are initially created and the prediction of the outcomes for each unit of the population is done. The results are then aggregated to guide policy related analyses and decision making. The fundamental goal in the development of a population synthesizer is to synthesize the required population as accurately and precisely as possible, for as many variables as possible that are known to determine travel behavior.

The majority of the currently developed population synthesizers control for variables only at one level,

usually the household level. Beckman et al. (1996) deal with the problem of generating synthetic baseline populations based on sample and census data that are available in the US and controlling for only the household-level variables. The main challenge that is frequently encountered in a broad range of population synthesis studies is that of simultaneously controlling for both household and individual-level distributions in estimating target joint distributions. Recently, Guo and Bhat (2007) proposed a new population synthesis procedure that generally addresses the main problems faced in the application of Beckman et al.(1996)'s approach. The approach provides a solution that deals with simultaneous control for both household and individual-level distributions in estimating target joint distributions as well dealing with the problem related to zero-cell values. Ye et al. (2009) further presented a heuristic approach that was called the Iterative Proportional Updating (IPU) algorithm for generating synthetic populations in a computationally efficient way. In this study, the aim is to create synthetic populations for the application area, the Flemish region of Belgium for the year 2007. An application of Beckman et al. (1996) and Guo and Bhat (2007)'s approaches for generating synthetic populations is presented. The data available here include micro data from the socio-economic population census data of 2001 (SEE'01) conducted in Belgium, marginal data (obtained via the website of the Flemish government) available for the variables of interest that are desired to be controlled for, for the Flemish population in the year 2007 (FL'07), a sample file for 2001 and a sample travel survey (OVG'07) data set for the year 2007. At the household-level, the variables controlled for include: availability

of car(s) in a household, age of the householder and household size. At the individual-level, gender and age are controlled for.

Results and Discussion

To estimate the target joint distributions for Flanders in 2007, the SEE'01 joint distributions were updated using the Iterative Proportional Fitting (IPF) algorithm based on the population marginal of 2007 for Flanders. This was conducted both at the household and the person-level based on the control variables mentioned above. The synthetic population data were then generated following Beckman et al. (1996) and Guo and Bhat (2007)'s procedures. The procedures are referred to herein as Procedure1 and Procedure2 respectively. Overall, Procedure2 provides better results. Using the average absolute relative difference to compare the synthetic with the true joint distribution, a fit of 0.088, which is close to zero, is obtained as compared to 0.57 for Procedure1 at the person-level. At the household-level, Procedure1 maps the target joint distribution perfectly whereas Procedure2 inhibits a deviation of 0.129. The results also reveal that the true marginal distributions from census records of all the control variables at the household and person level are very closely preserved using Procedure2 whereas some slight deviation is observed for Procedure1 at the person-level. Both procedures are able to yield synthetic data that result into a value (2.47 for Procedure1 and 2.43 for Procedure2) that is relatively close to the true average household size of 2.40 that is based on the population register of Flanders for the year 2007. The distributions of the variables that were not

explicitly controlled for were also examined. Impressively, overall, the distributions based on the synthetic data are quite close to the actual distributions for both procedures both at the household and the person level based on most of the variables that were examined. Even Procedure1 which does not involve controlling for person level variables yields reasonable distributions at the person level. Furthermore, for some travel-related variables such as work/school travel distance, the work/school indicator and w hours per week, both procedures perform well as distributions are very close to those from the validation sample data (OVG'07). The estimates of the mean and the standard deviation for some travel-related very good. For example, the actual mean work hours of 39hrs for the working population is preserved and the travel distance for work/school trips which is 16km for the validation data is estimated as 17km with the synthetic data for persons who either work or go to school.

In conclusion, the results obtained from comparing the generated synthetic populations with the real data provided support that both the household and the individual-level distributions of the control and some non-control variables represent the true population rather well and consequently the actual population could be relatively accurately synthesized.

REFERENCES

[1] Beckman, R.J., Baggerly, K.A., & McKay, M.D. (1996). Creating synthetic baseline populations. Transportation Research Part A: Policy and Practice, 30(6), 415-429. doi: 10.1016/0965-8564(96)00004-3.

[2] Guo, J., & Bhat, C. (2007). Population synthesis for microsimulating travel behavior. Transportation Research Record: Journal of the Transportation Research Board, 2014, 92-101. doi: 10.3141/2014-12.

[3] Ye, Xin , X., Konduri, K.C., Pendyala, R.M., Sana, B., & Waddell, P. (2009). Methodology to match distributions of both household and person attributes in generation of synthetic populations. In TRB 88th Annual Meeting Compendium of Papers.