An efficient hierarchical model for multi-source information fusion

Ismaïl Saadi[1*], Bilal Farooq[2], Ahmed Mustafa[1], Jacques Teller[1], Mario Cools[134]

[1]Local Environment & Management Analysis (LEMA), Urban and Environmental Engineering (UEE), University of Li`ege, All´ee de la D´ecouverte, Quartier Polytech, 4000 Li`ege, Belgium

[2]LITrans, Ryerson University, Department of Civil Engineering, 350 Victoria St, Toronto, Ontario M5B 2K3, Canada

[3]KULeuven Campus Brussels, Department of Informatics, Simulation and Modeling, Warmoesberg 26, 1000 Brussels, Belgium

[4]Hasselt University, Faculty of Business Economics, Agoralaan Gebouw D, 3590 Diepenbeek, Belgium

Email addresses: ismail.saadi@uliege.be (I. Saadi), bilal.farooq@ryerson.ca (B. Farooq), a.mustafa@uliege.be (A. Mustafa), jacques.teller@uliege.be (J. Teller), mario.cools@{uliege.be,kuleuven.be,uhasselt.be} (M. Cools)

[*]Corresponding author: Tel.: +32 4 366 96 44, Fax: +32 4 366 29 09

Forecasting activity-travel patterns is relevant to many applications and research domains, e.g. urban/transportation research, and social sciences. The behavioral realism associated to the simulation of complex urban and transportation systems requires highly disaggregated and reliable datasets. A major problem is that such disaggregated data are not always available. Moreover, sampling rates are generally low, i.e. in the best case at most 10% of the total population, as data collection for travel surveys/micro-samples is costly, and large-scale surveys, i.e. censuses, are systematically subjected to privacy and confidentiality issues. Therefore, in urban and transportation research, efficient and flexible methods are required to fuse information stemming from multiple micro-samples and aggregate statistics, e.g. socio-demographic marginal distributions.

Furthermore, in urban and transportation research, important information is often scattered over a wide variety of independent datasets which vary in terms of described variables and sampling rates. As activity-travel behavior of people depends particularly on socio-demographics and transport/urban-related variables, there is an increasing need for advanced methods to merge information provided by multiple urban/transport household surveys. In this study, we propose a hierarchical algorithm based on a Hidden Markov Model (HMM) and an Iterative Proportional Fitting (IPF) procedure to obtain quasi-perfect marginal distributions and accurate multi-variate joint distributions. The model allows for the combination of an unlimited number of datasets.

In the literature, three families of population/data synthesis techniques have been identified: (a) fitting approaches based on Iterative Proportional Fitting (IPF) or Iterative Proportional Updating (IPU) (a derived version of IPF), (b) Combinatorial Optimization (CO) and (c) Markov Process-based methods. IPF-based approaches are commonly used for modeling populations for transport and urban systems. (a) IPF procedures consist of fitting a multi-dimensional contingency table given a set of target marginal distributions and a single micro-sample derived, for instance, from a travel survey. Besides, (b) CO can be defined as a micro-data reconstruction approach which performs a random selection of households from micro-samples in order to reproduce the characteristics of a specific geographical unit. Different statistical metrics have been proposed to assess the goodness-of-fit of the model. With respect to the Markov Process-based methods (c), Farooq et al. (2013) used, for instance, an MCMC method for population synthesis. Both the full and partial conditional distributions used by MCMC method can be calibrated on multiple micro-samples. Saadi et al. (2016) used an HMM-based approach for synthesizing the population of Belgium. The method is highly flexible for fusing multiple micro-samples and shows competitive prediction capabilities. Nonetheless, the full dependency on micro-samples often leads to less accurate simulated marginal distributions despite accurate simulated joint distributions. In this paper, we propose an extension of the HMM by integrating IPF, allowing an efficient multi-source information fusion.

The structure of the hierarchical model (HM) enables multi-source information fusion. HM includes two important components, i.e. HMM and IPF. The *N* micro-samples and the *M* marginal distributions can be used simultaneously as inputs within the HM framework. The scaled-up and fused micro-sample enables the connection between HMM and IPF. As the multi-source fusion process already takes place within the HMM component, the scaled-up and fused micro-sample will systematically include all the variables of interest. IPF enables a direct fitting of the marginal distributions based on the observed targets, i.e. second set of inputs. Of course, using all the marginal distributions is not mandatory. HM is designed to

enable flexibility towards unavailable marginal distributions. It is indeed possible to fit data against a number of marginal distributions which is lower than the total number of variables of interest, i.e. *M*. Finally, HM results in a fused and more accurate dataset that can be used in multiple applications, e.g. agent-based modeling of complex urban and transportation systems.

The hierarchical model is tested based on a synthetic dataset of 1,000,000 observations and 8 random variables with 128, 16, 8, 8, 4, 4, 3 and 2 categories respectively. Data are deliberately heterogeneous and designed in the image of real world situations. In urban and transportation research, variables contain multiple categories for representing socio-demographics/transport-related variables. And the number of categories is even more important if spatial information is included. Therefore, we also chose a complex categorical variable with 128 levels. In order to underline the influence of the sampling rate on model outputs, five bootstrap samples are derived from the original dataset in the following order 10%, 5%, 1%, 0.1% and 0.06%. There is no point in considering sampling rates higher than 10%, since such data are typically not available.

In this paper, we present the practical procedure for model estimation using a single micro-sample and all the marginal distributions. The results are compared on the basis of the joint and marginal distributions to highlight the performances of HM. Furthermore, we illustrate how to fuse multi-source information based on another case study considering multiple micro-samples and all the marginal distributions.

Based on the results of the current study, the main concluding remarks can be formulated as follows:

1) HM provides the best trade-off in terms of RMSE minimization, when marginal distributions and joint distributions are simultaneously compared. This can be explained by the fact that the principal key features of IPF and HMM are combined within a single unified framework
2) Multiple micro-samples and aggregate marginal distributions can be integrated within HM for allowing multi-source information fusion. Also HM shows a lot of flexibility in terms of data availability. We mentioned that a partial set of marginal distributions can be used if there is absolutely no data.
3) HM is extremely competitive and relatively robust with respect to sampling rate variability. This means that with a sampling rate of only 1%, it is possible to achieve results which are almost comparable to a HM calibrated with a micro-sample of 10%. Several applications within the field of urban and transportation research assume sampling rates which are around 1% using standard methods, i.e. IPF. But the results show that with IPF, a still commonly used method, the RMSE is equal to 13.65. In this context, HM emerges as a far better alternative for mitigating the error in micro-simulation.