Population Synthesis for Long-Distance Travel Demand Simulations using Mobile Phone Data

Maxim Janzen

Kirill Müller

Kay W. Axhausen

Institute for Transport Planning and Systems **September 2017**

Institut für Verkehrsplanung und Transportsysteme
Institute for Transport Planning and Systems

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Institute for Transport Planning and Systems

Population Synthesis for Long-Distance Travel Demand Simulations using Mobile Phone Data

Maxim Janzen
IVT
ETH Zurich
CH-8093 Zurich
phone: +41-44-633 33 40
email:
maxim.janzen@ivt.baug.ethz.ch

Kirill Müller
IVT
ETH Zurich
CH-8093 Zurich
email: kirill.mueller@ivt.baug.ethz.ch

Kay W. Axhausen
IVT
ETH Zurich
CH-8093 Zurich
phone: +41-44-633 39 43
email:
axhausen@ivt.baug.ethz.ch

September 2017

Abstract

Analysis of long-distance travel demand has recently become more relevant. The reason is the growing share of traffic due to journeys related to remote activities, which are not part of daily life. In today's mobile world, these journeys are responsible for almost 50 percent of traffic overall. Consequently, there is a need for reliable long-distance travel forecasting tools, such as agent-based simulation with a suitable synthetic population that also covers irregular long-distance travel demand. In addition to socio-demographic attributes, each agent requires information on the number of long-distance tours, and the purpose and duration of each tour. Accuracy of the synthetic population is crucial in order to obtain valid results from the simulations. We will show how existing data sources can be utilized to synthesize a population with these characteristics.

Usually, two data sources are used to synthesize a population for agent-based simulations. First, travel surveys are performed to obtain detailed information on the persons and their travel behavior for a sample of persons. Second, official statistics (register data) provide information on the marginal totals. A fitting algorithm is then applied to create a population that matches the travel behavior reported in the survey data as well as the marginal totals of the register data. The most popular approach is the iterative proportional fitting algorithm, but also other approaches have been employed, e.g. Bayesian Networks. In case of long-distance travel behavior, these two data sources are not sufficient, because it is known that travel surveys under-report long-distance travel heavily. Therefore an additional data source is needed. We propose to add passively collected mobile phone data, which is based either on GPS or on GSM. Mobile phone data is helpful since the obtained information on long-distance travel behavior is more reliable than results from survey data. On the other hand, the available samples of phone data lack socio-demographic information. Thus, it can not fully replace the travel diary data. Potentially, phone data can be substituted by any reliable source on long-distance travel behavior.

In order to overcome the drawbacks described, we propose to utilize all three data sources in our framework. Firstly, all respondents from the survey are treated as possible agents of the synthetic population. Socio-demographic variables that are of interest (e.g. sex, age class, education), the zone of the residence and the number of long-distance tours per purpose are taken from the survey. Previous research has shown that the number of tours reported in surveys is too low. Assuming uniform underreporting, we use the passively collected phone data to adjust this underreporting of long-distance tours in the travel survey data. The adjustment will preserve the dependency structure between socio-demographic attributes and the relative frequency of long-distance tours, using a technique akin to histogram matching in image processing. In the next step, generalized raking, a generalization of IPF, is used to find a weight for each person. Generalized raking can be used to compute weights for agents of a population and account for the marginal totals of the controlled variables. In this case, controlled variables are the number of tours per purpose (taken from phone data) and socio-demographics (taken from register data). The number of agents required for the population can be drawn with respect to the calculated weights. Finally, a duration needs to be assigned to each tour of the drawn agents. The duration can be imputed using a model based on the duration distributions given in the survey data. This model has to take into account zone types, purposes and some of the socio-demographic attributes. The result is a population of arbitrary size including information on socio-demographics and zone of residence as well as detailed long-distance travel demand. The framework takes advantage of the knowledge of marginal totals from register data, reliable travel information from phone data and socio-demographic influences from survey data. Thus, it is more reliable than traditional approaches using just survey and register data.

The framework presented above will be applied with three available data sources covering France in 2007. A national travel survey including a long-distance travel questionnaire was carried out at this time. In addition, register data is collected on a yearly basis by the National Institute of Statistics and Economic Studies (INSEE). Furthermore, the authors had access to aggregated results of a large sample of mobile phone billing data from a French mobile network operator. These three data sources are combined following the framework described above in order to synthesize a population. In case of the phone data, purpose imputation has to be performed, and the data set has to be rescaled because it covers just domestic travel within five months. We assume that the distribution of the number of long-distance tours is similar for all (central) European countries. Therefore, we will use the phone data from France in combination with survey data and register data from Switzerland in order to generate a synthetic population representing Swiss residents.

Keywords

long-distance travel demand; agent-based simulation; population synthesis; mobile phone billing data