Chapter 3

Synthetic Population Techniques in Activity-Based Research

Sungjin Cho
Hasselt University, Belgium

Tom Bellemans
Hasselt University, Belgium

Lieve Creemers
Hasselt University, Belgium

Luk Knapen
Hasselt University, Belgium

Davy Janssens
Hasselt University, Belgium

Geert Wets
Hasselt University, Belgium

ABSTRACT

Activity-based approach, which aims to estimate an individual induced traffic demand derived from activities, has been applied for traffic demand forecast research. The activity-based approach normally uses two types of input data: daily activity-trip schedule and population data, as well as environment information. In general, it seems hard to use those data because of privacy protection and expense. Therefore, it is indispensable to find an alternative source to population data. A synthetic population technique provides a solution to this problem. Previous research has already developed a few techniques for generating a synthetic population (e.g. IPF [Iterative Proportional Fitting] and CO [Combinatorial Optimization]), and the synthetic population techniques have been applied for the activity-based research in transportation. However, using those techniques is not easy for non-expert researchers not only due to the fact that there are no explicit terminologies and concrete solutions to existing issues, but also every synthetic population technique uses different types of data. In this sense, this chapter provides a potential reader with a guideline for using the synthetic population techniques by introducing terminologies, related research, and giving an account for the working process to create a synthetic population for Flanders in Belgium, problematic issues, and solutions.

DOI: 10.4018/978-1-4666-4920-0.ch003
INTRODUCTION

Since its introduction in transportation, ABM (activity-based model), which purpose is to estimate an individual induced traffic demand derived from activities, have been applied for traffic demand forecasts. The ABM typically uses different types of input data including daily activity-trip survey data and population data. The individual daily activity-trip schedule data describes the different trips, its purpose, locations, transport modes, as well as its temporal dimension. The population data, including socio-demographic features, are used to estimate population characteristics such as gender, household composition, income, home location, etc. In general, it seems to be hard to use those datasets because they are rather expensive and normally protected by a privacy law. Thus, it is indispensable to find a solution to substitute population data in a synthetic manner.

Several synthetic population generators have been used in the literature to generate synthetic population data in transportation. Examples are techniques like Iterative Proportional Fitting and Combinatorial Optimization. Despite these advancements in research, using those techniques is not easy for non-expert researchers not only due to the fact that there are no explicit terminologies and concrete solutions to some existing issues and problems so far, but also every synthetic population technique handles different types or structures of input and output data.

In this sense, the chapter is supporting a potential reader with a guideline for using synthetic population techniques by introducing terminologies and related research, and giving an account of the working process to create a synthetic population, along with problematic issues and solutions. In detail, the following sections provide common terminologies and related research in this field. Then, section 3 introduces related research. The next section describes the whole process of generating a synthetic population, which consists of three steps: data preprocessing, fitting and drawing (sampling). The section of issues and proposed solutions deals with some issues and solutions addressed by previous research. Finally, the chapter ends with a summary and by suggesting future work in this field.

RELATED RESEARCH

Synthetic population techniques can be largely divided into two groups: IPF and CO. Most techniques in these two groups have a similar concept of fitting seed data to a target marginal distribution, but they generate the required synthetic population in totally different way. This section covers the different ways by introducing related research in each group.

IPF

Deming and Stephan (1940) developed a basic algorithm in IPF (Iterative Proportional Fitting), which has been widely applied for synthetic population research in several fields, including transportation. The basic algorithm, which is called ‘a least squares adjustment’, is based on the assumption that the source and target have the same correlation structure. The correlation structure is defined by odds ratios, for example the odds ratio in a 2 x 2 cross-table is calculated as follows:

$$\varnothing = \frac{p_{11}p_{22}}{p_{12}p_{21}}$$

where $p_{ij}$ is a cell proportion of the cell $(i, j)$. Based on that assumption, the IPF adjusts seed data to target marginal distribution to keep the correlation between source and target. We do not explain further details of the IPF algorithm in this chapter, but we are dealing with how it can be applied within the synthetic population process in the next section.