Proposal of statistical matching methodology to fuse data from different survey samples

Naroa Burreso Pardo[1], *Edorta Arana Arrieta*[1], *Libe Mimenza Castillo*[1], *Irantzu Barrio*[2,3], *Josu Amezaga Albizu*[1]

[1]naroa.burreso@ehu.eus, Department of Audio-visual Communication and Advertising, University of the Basque Country UPV-EHU

[2]Department of Mathematics, University of the Basque Country UPV-EHU

[3]BCAM, Basque Center for Applied Mathematics

Statistical matching (SM) refers to a series of methods that use different available data sources (usually survey samples), referred to the same target population, with the aim of studying the relationship among variables not jointly observed in a single data source. In the simplest case of statistical matching there are two different databases A and B, each of them with their specific set of variables **X** and **Y** respectively, and they also share a set of common variables **Z**. The objective of the fusion is to investigate the relationship between **X** and **Y**. To study this relationship different methods exist. Most of the SM techniques assume that A and B are random samples of independent and identically distributed observations selected from the same infinite population. However, in most real cases the data come from surveys based on some complex design carried out on the same finite population. One of the characteristics of this type of data is the sampling weights, which indicate the number of units that each sampled observation represents in the finite population.

The goal of this methodological proposal is to use SM in order to create a synthetic database powered by information from two different surveys conducted to the same finite population, where each survey may (or may not) be based on a different sample design. To do so, one of the databases will be considered as the donor (the one with more individuals) and the other one as the receptor (the one with less individuals). The donor will transfer information to the receptor. The proposed methodology consists of five main steps: 1) Identify the common stratification variables and create the donation clases using those variables; 2) find the common variables (matching variables) not used in the sampling design; 3) calculate the distances between the individuals (in the common variables and same donation clases) using Gower´s distance (Gower 1971); 4) establish the donor- receptor relations taking into account the distances among individuals and their sampling weights; and 5) transfer the information from the donor database to the receptors using the relations established before. Once the fusion is completed, two types of analysis will be carried out to check the validity of the synthetic database: the imputed variables conserve the marginal distribution and the imputed variables conserve the joint distribution with the matching variables.

The proposed methodology has been applied to the databases Ikusiker and CIES. Ikusiker is a panel that measures audience consumption in non-traditional media of students between 12 and 21 years old from the four provinces in Southern Basque Country (Araba, Bizkaia, Gipuzkoa and Nafarroa). CIES is a survey that measures audience consumption in traditional media of the population over 14 years old in the same four provinces. The final product of the fusion has been a complete database with the information of the variables available in both original databases.

Keywords: Statistical matching, surveys, design based inference.