An Overview of Population Size Estimation where Linking Registers Results in Incomplete Covariates, with an Application to Mode of Transport of Serious Road Casualties

Peter van der Heijden

Abstract

We consider the linkage of two or more registers in the situation where the registers do not cover the whole target population, and relevant categorical auxiliary variables (unique to one of the registers; although different variables could be present on each register) are available in addition to the usual matching variable(s). The linked registers therefore do not contain full information on either the observations (often individuals) or the variables. By treating this as a missing data problem it is possible to construct a linked data set, adjusted to estimate the part of the population missed by both registers, and containing completed covariate information for all the registers. This is achieved using an Expectation-Maximization (EM)-algorithm. We elucidate the properties of this approach where the model is appropriate and in situations corresponding with real applications in official statistics, and also where the model conditions are violated. The approach is applied to data on road accidents in the Netherlands, where the cause of the accident is denoted by the police and by the hospital. Here the cause of the accident denoted by the police is considered as missing information for the statistical units only registered by the hospital, and the other way around. The method needs to be widely applied to give a better impression of the range of problems where it can be beneficial.