On the uncertainty and techniques of categorical data fusion

Li-Chun Zhang University of Southampton, UK & Statistics Norway, Norway E-mail: L.Zhang@soton.ac.uk

Abstract

Statistical matching (or data fusion) has long been used to merge separate data files in order to generate a joint fusion data set. Since the target joint data are not observable, it is recognized that, in addition to sampling variations that exist in the separate data files, there is an identification uncertainty associated with the assumptions that underpin the fusion procedure. In this paper, we focus on categorical data fusion, and develop an uncertainty analysis approach that is able to account for both types of uncertainty. Moreover, we discuss various techniques for actually generating the fusion distribution and the fusion data. Particular attention will be given to the so-called proxy variables, which are similar in concept to the target variables and have the same support. Real-life sample survey data will be used to illustrate that the availability of proxy variables can greatly reduce the identification uncertainty, and at the same time widens the scope of data fusion methods.

Key words: identification problem, probability bounds, structure preserving estimation, distribution calibration, longitudinal data.