

Title: Estimating hidden population sizes using multiple systems estimation with contaminated mortality data

Authors: Lanxin Li¹, Ruth King¹, Serveh Sharifi Far¹, Vanda Inácio¹, Heather Bradley²

¹ School of Mathematics and Maxwell Institute for Mathematical Sciences, University of Edinburgh, Edinburgh, UK

² Department of Epidemiology, Emory University, Rollins School of Public Health, US

Abstract:

Estimating the size of hidden or hard-to-reach populations is essential in epidemiology and public health for understanding disease burden and the impact of control efforts. Multiple systems estimation (MSE) is often used to estimate the size of such a target population. In MSE, multiple data lists observing individuals in the target population are linked and collated to form an incomplete contingency table and are fitted within a generalised linear model framework. The number of individuals not observed by any list can then be estimated from the fitted model and combined with the number of observed individuals to provide an estimate of the total population size.

In many real-world applications, however, the data collected are often imperfect, making it challenging to model. For example, in our case study relating to estimating the prevalence of opioid use disorder (OUD) in New York in 2020, two challenges were identified regarding the use of the mortality data. First, the inclusion of death lists violates the closed population assumption of MSE, since appearing on the list renders captured individuals ineligible for further recapture. Second, potential false positives were recorded in the death list due to the applied less stringent OUD case definition. This creates contaminated cell counts, where the observed cell value represents only an upper bound on the true cell count. Treating such observed cell counts as true cell counts leads to biased population size estimation. We use a staggered time period in the data preparation process to allow for the inclusion of death record data as a data list, without violating the closed population assumption. We derive an explicit expression for the observed data likelihood in the presence of contaminated cells, extending the standard MSE models. We demonstrate the substantial bias introduced when contamination is ignored in simulated data, before applying the approach to the case study.