Fractional Counting for administrative based population statistics

*Daniel Ward[1]; Jonathan Rees[1]; Iva Spakulova[1]; Greg Payne[1], Matthew Plummer[1]; Michael Hawkes[1]; Alison Whitworth[1]

[1] Office for National Statistics

Abstract:

The fractional counting framework takes a hypercube, constructed from administrative data, and uses regression and machine learning techniques to fractionally distribute individuals across their possible addresses. These modelled hypercube weights can then be used to produce granular multivariate population and related characteristic statistics, with reduced bias, on a rolling basis. Alternative administrative data-based approaches, which place individuals at addresses in an integer fashion, can be subject to bias when producing estimates. The use of fractional counting aims in part to reduce this, whilst also allowing for the production of smaller area multivariate estimates.

The integrated extended population hypercube is constructed by linking together anonymised record-level data between several administrative data sources (patient register (PR), personal demographic service (PDS), higher education statistics agency (HESA), and the English and Welsh School Census' (ESC/WSC)). The hypercube is then used in combination with the 2011 Census to produce a 'truth' state of information (target population inclusion and true address) for each individual. This linked dataset currently contains attribute information such as age, sex, and location of residence, from which multivariate outputs can then be derived.

We model the likelihood of inclusion in the target population, and subsequently the likelihood or residency at any conflicting addresses, in a dual-stage pipeline. Using regression and machine learning techniques (logistic regression (LGRG), support vector machines (SVM), random forests (RF) and gradient boosted trees (XGBOOST)) we (1) determine target population inclusion likelihoods and (2) to produce residency weights for each record in the hypercube (ranging from 0-1). From these individual weights it is then possible to produce multivariate estimates at a range of resolutions, from local authority (LA) to national level.

Initial results look promising, with each model producing population estimates that are all qualitatively accurate and in line with known totals taken from the 2011 Census, with RF performing best and SVM performing least best. A challenge for the fractional counting project is to continuously adjust the weights as the hypercube 'rolls forward' each year following the reference 'truth' state (currently 2011 Census). Going forward, we will consider the use of survey data to adjust the weights as additional administrative data becomes available or is updated, to produce continuously updated characteristic estimates.

Keywords:
Census; Multivariate; Characteristics; Hypercube; Rolling;

1. Introduction:

<Introduction>

2. Methodology:

<Methodology>

3. Result:

<Result>

4. Discussion and Conclusion:

<Discussion and Conclusion>

References:

1.
2.
3.

NOTE: THE MAXIMUM NUMBER OF PAGES FOR THE PAPER IS SIX PAGES