Thomas Zimmermann (Federal Statistical Office of Germany)

Evaluation of small area estimates for hypercubes in the German census

Topic 6 – Statistics closer to the ground

Keywords: small area estimation, synthetic estimation, log-linear models, register-based census

Introduction

In 2011, the German census was conducted as a register-based census, where the register information was supplemented by a probability sample of approximately 10% of the population.

This sample was designed such that precise design-based direct estimates were obtained for the over- and under-counts in the population registers to determine the number of inhabitants in large communities.

In addition to that, the census sample was used to provide information about variables which were not part of the register. Examples include the employment status according to the ILO definition as well as the migrant status and the highest level of school education.

Estimates on multi-dimensional hypercubes constructed as cross-classifications of these additional variables with further demographic variables are of special interest.

Methods / Problem statement

Currently, we use the census data from 2011 to evaluate different approaches to provide reliable estimates for the hypercube cells at detailed regional breakdowns (e.g. counties) in the next German census.

Owing to small sample cell counts, the relative standard errors of the direct estimates do not meet our internal quality requirements in most cases. SPREE (Purcell and Kish, 1980) and GSPREE (Zhang and Chambers, 2004) approaches are not directly applicable either, as the initial cell proportions cannot be obtained from the register. Instead, we consider a two-step procedure similar to Dostal et al (2016), where synthetic estimates of the cell proportions obtained from aggregate sample data are adjusted to agree with known marginal totals at the regionally disaggregated level.

Results / Proposed solution

First, we considered the adjustment by minimising the chi-square distance function as proposed by Dostal et al (2016). In our application, this produced negative estimates for some of the hypercube cells. Hence, we decided to enfore coherence of the cell estimates with known marginal totals by applying log-linear models as discussed by Noble et al (2002).

This approach comprises iterative proportional fitting as a special case but is also applicable when some of the hypercube cells are structural zeros. We employ a design-based resampling approach that captures the additional uncertainty due to the estimation of initial cell proportions at the aggregate level to estimate the MSE of the cell estimates.

Conclusions

Our results show a higher precision of the cell estimates using our two-step procedure when compared to a direct design-based estimator.