LR5_3 Area-level small area estimation methods for domain statistics

Area-level small area estimation methods for domain statistics	
General setting	
Configuration	Configuration 5
Type of sources	Combination of aggregate-level survey and register data
Statistic of interest	Domain estimates in social or business statistics
Type of errors	Sampling and measurement errors
Quality measure	Estimation of mean squared error for the statistics of interest
Method (*use one block per method*)	
Description	Design-based or direct survey estimates for small subpopulations often have large sampling variances due to small sample sizes. By linking the subpopulation quantities of interest through a model it is usually possible to obtain estimates with smaller mean squared errors. This is the subject of small area estimation (Rao and Molina, 2015). The subpopulations are not restricted to geographical areas, but can be anything as long as they are mutually exclusive, there are not too few of them, and the quantities of interest show some degree of similarity between the areas. A popular model in small area estimation is the basic area-level or Fay-Herriot model (Fay and Herriot, 1979). The input data for this model consist of a set of survey estimates y_i and corresponding variance estimates ψ_i, as well as auxiliary data X_i, where each X_i can be a vector. The index i refers to the subpopulations (areas) of interest in a particular study, such as municipalities, and y_i are direct or initial estimates for the unknown quantities of interest θ_i. The aim is to find estimates for θ_i, which improve on the initial estimates y_i. The model is $$y_i = \theta_i + e_i$$ $$\theta_i = \beta' X_i + v_i \quad (1)$$ with independent errors $$e_i \sim N(0, \psi_i) \quad \text{and} \quad v_i \sim N(0, \sigma_v^2)$$ The first line of (1) is like a measurement equation relating y_i to the latent quantities θ_i, while the second line is the structural part of the model relating the quantities of interest to auxiliary information and to each other via common regression coefficients β. Equations (1) can be combined into $$y_i = \beta' X_i + v_i + e_i$$

	which is a special case of a linear mixed model. The Empirical Best Linear Unbiased Predictor (EBLUP) based on this model is $$\hat{\theta}_i = \hat{\beta}' X_i + \hat{v}_i = \hat{\gamma}_i y_i + (1 - \hat{\gamma}_i)\hat{\beta}' X_i$$ where $$\hat{\beta} = \left(\sum_i \hat{\gamma}_i X_i X_i{}'\right)^{-1} \sum_i \hat{\gamma}_i X_i \, y_i$$ $$\hat{\gamma}_i = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \psi_i}$$ Various methods are available to find $\hat{\sigma}_v^2$, including (restricted) maximum likelihood, see Rao and Molina (2015). Given $\sigma_v^2 = \hat{\sigma}_v^2$, the mean squared error (mse) for the EBLUP is $$mse(\hat{\theta}_i) = \hat{\gamma}_i \psi_i + \hat{\sigma}_v^2 (1 - \hat{\gamma}_i)^2 X_i{}' \left(\sum_j \hat{\gamma}_j X_j X_j{}'\right)^{-1} X_i$$ An additional contribution to the mse comes from uncertainty in the estimate $\hat{\sigma}_v^2$. This contribution is usually relatively small, but analytical expressions can be found in Rao and Molina (2015).
Assumptions	The method is based on the assumption that the model holds. In particular, • the input estimates y_i are unbiased for θ_i • the errors v_i and e_i are normally distributed • the sampling variances ψ_i are assumed known, whereas in practice they must be estimated
Advantages (benefits)	• the method borrows strength over areas, thereby achieving much smaller variances for areas with small sample sizes, at the cost of some bias • estimates can be obtained even for areas without observations • the basic area level model is appealing to many survey statisticians since the initial estimates can account for complex sampling design features and differential response rates. • as the model is specified at an aggregate level, it is typically not very computationally demanding to fit
Disadvantages (costs)	• only auxiliary information at the area level can be used in the model. The model cannot correct for unit-level selection bias. Such correction should already be incorporated in the input estimates. • mean squared errors may be underestimated because the sampling errors are assumed to be known

Case study (per method if needed)	
Agency – country	Statistics Netherlands
Topic	Estimation of municipal unemployment based on Labour Force Survey data. Several estimation methods compared in a simulation study.
Data sets used	LFS data from several years was used to create an artificial population. Auxiliary

	variables from several registrations have been used, including registered unemployment.
Results (e.g. different methods)	Estimates based on the area-level model were found to be more accurate than survey regression estimates, which are similar to direct GREG estimates. Small area methods specified at the unit-level performed even better, for one thing because they better exploit auxiliary information which in this study was available at the unit-level.
Report	See Boonstra et al. (2009).
Final remarks	
Gap analysis	The method has already been extended in most ways one can think of. These extensions include time-series extensions, spatial correlation, modelling of sampling error, non-linear specifications, etc. It would be interesting to combine this line of parametric aggregate-level modelling with macro-integration methods for incorporating equality and inequality restrictions.
Other remarks	
References	Boonstra, H.J., van den Brakel, J., Buelens, B., Krieg, S. and M. Smeets (2008). Towards small area estimation at Statistics Netherlands. *Metron* 66 (1), pp. 21-49. Fay, R. E., and R.A. Herriot (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association* 74 (366a), pp. 269-277. Rao, J.N.K. and I. Molina (2015), *Small Area Estimation*. John Wiley.