Simulation studies for small area estimation

New Challenges for Statistical Software - The Use of R in Official Statistics in Bucharest, Romania

Jan Pablo Burgard, Ralf Münnich, Jan Seger and Thomas Zimmermann

Bucharest, 23rd April 2015

Simulation studies for small area estimation

Monte-Carlo simulation studies: pros and cons

A practical example

Summary and Outlook

Why do we need simulations studies?

- According to Münnich (2014)
  - You can learn from simulations – changing simulation settings may lead to interesting behavior or to the discovery of peculiarities which hardly can be found with mathematical proofs
  - Sometimes these observations may constitute the point of origin of a proof
  - Using careful set-ups, one can learn about the applicability of the different kind of estimators in various situations
  - In practice, we have only one sample. How do we know that our approach is adequate?
- The setup including the corresponding evaluation of a MC study is utmost important
- One has to differ between examples and simulation studies

Classification of Monte-Carlo studies

Münnich and Burgard (2014)

Design-based Inference based on the sampling design

- Universe: (real) finite population
- Sampling is included by construction
- True parameters: fixed (by calculation)

Model-based Inference based on realizations of the superpopulation model

- No universe but superpopulation
- Sampling is outcome of random variables; designs are difficult to include
- True parameters have to be derived from the superpopulation model

Classification of Monte-Carlo studies

Quasi design-based Inference based on the sampling design

- Universe: finite population not (fully) available; has to be expanded or further generated
- Sampling is generically included
- True parameters: fixed (by calculation)

Quasi model-based Inference based on random variables in the model (independent variable, in general, fixed)

- Universe is repeatedly drawn from a model
- Sampling designs can be integrated in each run

Design-based under model data One outcome of the model is used as universe – rest as design-based simulation

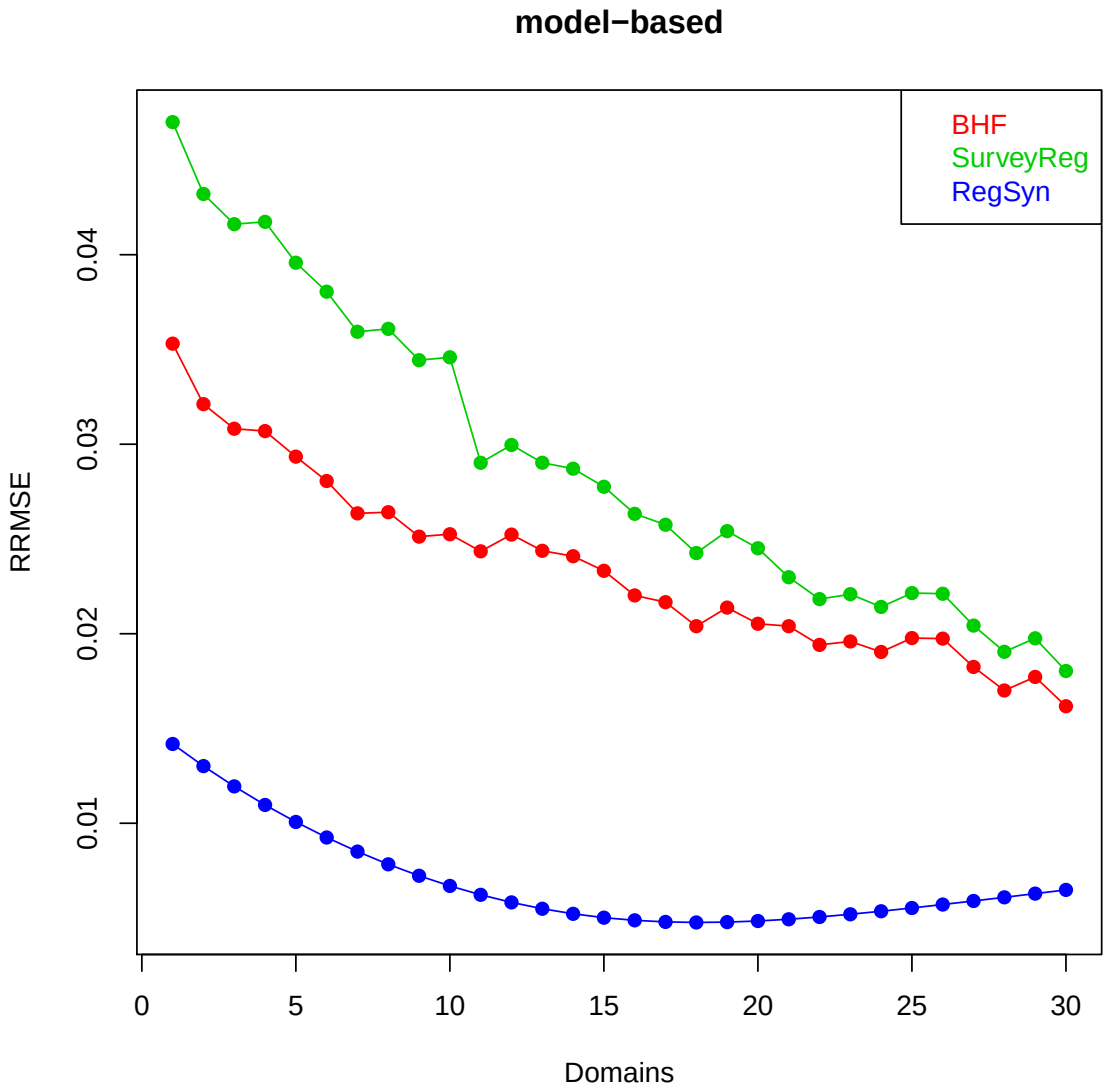
Simulation setup

- Comparison of the performances of three estimators in a model-based and a quasi model-based simulation study
- Consider a variable of interest **y** linked to a matrix of auxiliary variables **x** via a one-fold nested unit-level model:

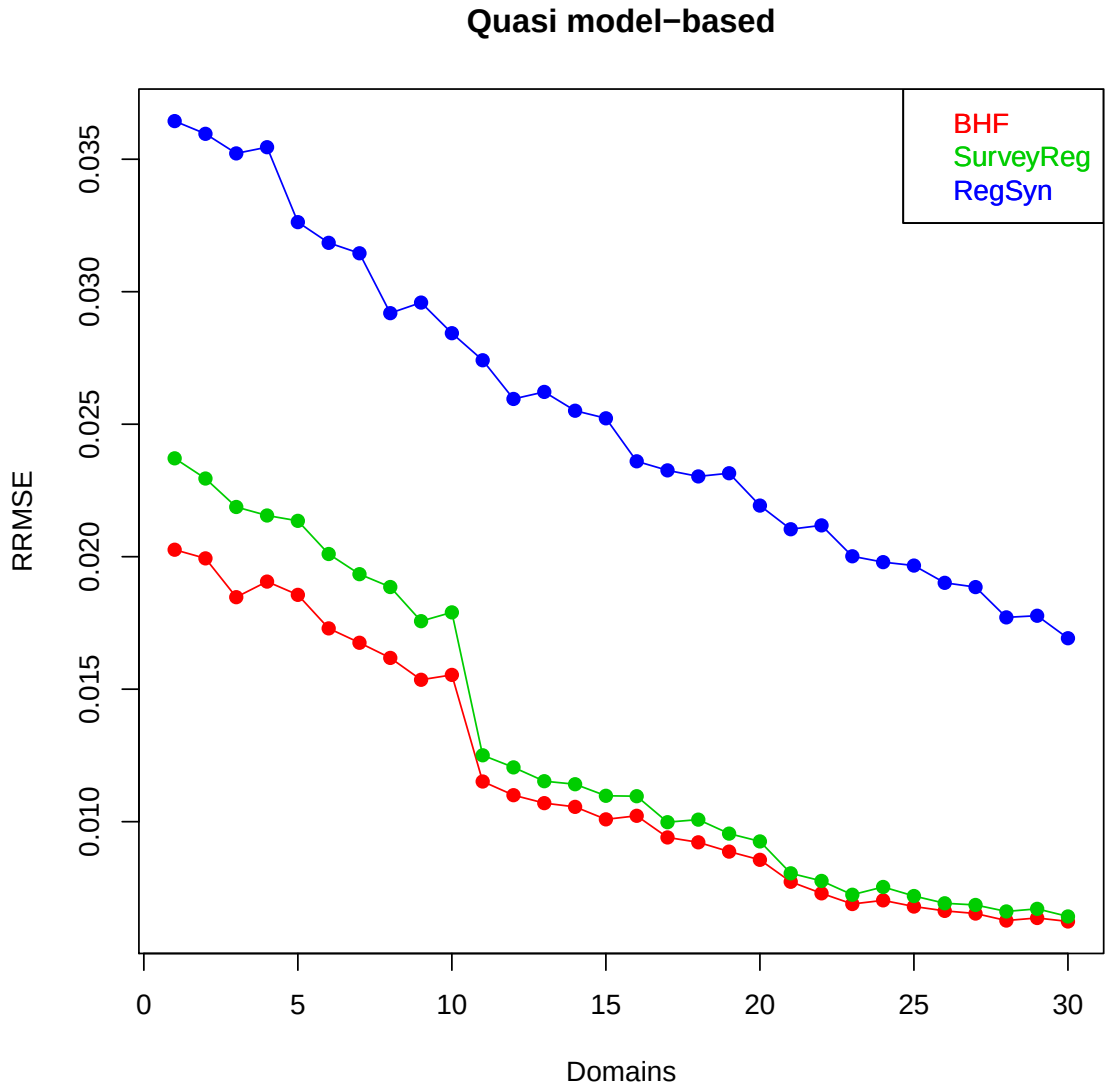
$$y_{dj} = \mathbf{x}'_{dj}\boldsymbol{\beta} + u_d + e_{dj} \; j = 1, \ldots, N_d, \; d = 1, \ldots, D \qquad (1)$$

  with $u_d \overset{\text{iid}}{\sim} N(0, \sigma_u^2)$, $e_{dj} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$ and independence of the error term e and the area effects
- Estimators: Survey regression estimator, synthetic regression estimator, EBLUP according to model (1) (BHF)
- 30 domains with domain specific sample sizes of 10, 20, and 30 for 10 areas each

RRMSE model-based simulation



RRMSE quasi model-based simulation



Summary and Outlook

- In general, simulation studies are useful to test and understand the behavior of estimators
- The setup of a simulation study may have a large influence on the results
- Differences between setup specific results may yield additional important information
- Safe datasets and samples for comparative research:
  - AMELIA based on European SILC data plus a bunch of sampling designs
    `http://ameli.surveystatistics.net`