SAE 2013

Freie Universität Berlin

Full bias-correction of spatial robust small area estimators

Session: SAE Using Time Series or Spatial Models

SAE 2013, Bangkok

Timo Schmid

Monday, September 2, 2013

SAE 2013

Freie Universität Berlin

Contents

Motivation

- Classical small area models are based on strong distributional assumptions, which are often not fulfilled in the case of business data.
- Especially in business surveys, outliers and skewed distributions are very common in the sample data.
- Skewed distributions and outliers are violating the strong assumptions of small area models.
- These phenomena have great impact on the estimators and lead to a substantial bias especially within small sample sizes.
- Beyond that, spatial dependencies occur very often in business data (e.g. similar industry segments).
- Thus, there is a need to investigate spatial outlier robust small area models.

Robust "Plug-In" methods

Assumption: All non-sampled values are not outliers

- The sample includes all outliers of the population.
- Chambers et. al. (2013) called this approach *projective* because they *project* the working model onto the whole non-sampled part of the population.
- Examples:
  - Robust EBLUP (Sinha and Rao, 2009)
  - M-Quantile methods (Chambers and Tzavidis, 2006)
  - Spatial robust EBLUP (Schmid and Münnich, 2013)
- These "Plug-In" methods may suffer from a bias in situations with representative outliers or non-symmetric contamination.

References: Chambers (1986) and Chambers et al. (2013)

Bias-corrected robust methods

Assumption: Some non-sampled units are outliers

- The sample includes only some of the outliers in the population.
- Chambers et. al. (2013) called this approach *predictive* because they use the sample outlier information to *predict* contamination on the target variable.
- Define a robust bias correction to the robust "Plug-In" estimators. Two concepts: Locally vs. fully bias corrections.
- Examples for the REBLUP:
  - Locally: CCST (Chambers et al., 2013)
  - Fully: CHAM (Dongmo-Jiongo et al., 2013)
  - Fully: CB (Dongmo-Jiongo et al., 2013)

References: Chambers (1986) and Chambers et al. (2013)

Bias-corrected robust methods

Assumption: Some non-sampled units are outliers

- The sample includes only some of the outliers in the population.
- Chambers et. al. (2013) called this approach *predictive* because they use the sample outlier information to *predict* contamination on the target variable.
- Define a robust bias correction to the robust "Plug-In" estimators. Two concepts: Locally vs. fully bias corrections.
- Concepts for the SREBLUP:
  - Locally: SCCST
  - Fully: SCHAM
  - Fully: SCB

References: Chambers (1986) and Chambers et al. (2013)

SAE 2013

Freie Universität Berlin

Contents

Basic model and notations

General linear mixed model:

$$y = X\beta + Zv + e$$

- Individual level covariates X and area level covariates Z
- SAR model: Area random effect $v \sim N(0, G)$ with $G = \sigma_v^2 \left((I - \rho W)(I - \rho W^T)\right)^{-1}$
- W describes the neighbourhood structure of the areas i
- $\rho \in [-1, 1]$ defines the strength of the spatial relationship among the areas
- Error term $e \sim N(0, R) = N(0, diag(\sigma_e^2))$
- Variable of interest is $y \sim N(X\beta, V)$ with $V_\theta = R + ZGZ^T$
- Target variable is $\overline{y}_i$

The benchmark: EBLUP

Empirical best linear unbiased predictor (EBLUP) for $\overline{y}_i$ is:

$$\hat{\overline{y}}_i = N_i^{-1}\Big\{\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} \hat{y}_{ij}\Big\} = N_i^{-1}\Big\{\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} (x_{ij}^T\hat{\beta} + z_{ij}^T\hat{v}_i)\Big\}$$

where

$$\begin{aligned}\hat{\beta} &= (X^T V_{\theta}^{-1} X)^{-1}(X^T V_{\theta}^{-1} y)\\ \hat{v} &= GZ^T V_{\theta}^{-1}(y - X\hat{\beta})\end{aligned}$$

$\hat{\theta}$ is e.g. the REML- or ML-Estimator of the variance component.

Reference: Rao (2003).

Robust EBLUP

Basic idea: Substitute $\hat{\beta}$ and $\hat{v}$ with robust estimators $\hat{\beta}^{\psi}$ and $\hat{v}^{\psi}$, leading to a robust estimator $\hat{y}_{ij}^{\psi}$.

Robustified ML-Equations:

$$\begin{aligned} \alpha(\beta) &= X^T V^{-1} U^{\frac{1}{2}} \psi(r) = 0 \\ \Phi(\theta_l) &= \psi^T(r) U^{\frac{1}{2}} V^{-1} \frac{\partial V}{\partial \theta_l} V^{-1} U^{\frac{1}{2}} \psi(r) - \mathrm{tr}(V^{-1} \frac{\partial V}{\partial \theta_l} K) = 0 \end{aligned}$$

- ψ is an influence function
- $r = U^{-\frac{1}{2}}(y - X\beta)$ and $U = diag(V)$

REBLUP for $\overline{y}_i$ is:

$$\hat{\overline{y}}_i = N_i^{-1}\Big\{\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} \hat{y}_{ij}^{\psi}\Big\} = N_i^{-1}\Big\{\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} (x_{ij}^T \hat{\beta}^{\psi} + z_{ij}^T \hat{v}_i^{\psi})\Big\}$$

Reference: Fellner (1986), Sinha and Rao (2009).

Spatial REBLUP

Basic idea: Substitute $\hat{\beta}$ and $\hat{v}$ with spatial robust estimators $\hat{\beta}^{\psi,sp}$ and $\hat{v}^{\psi,sp}$, leading to a spatial robust estimator $\hat{y}_{ij}^{\psi,sp}$.

Robustified spatial ML-equations:

$$\begin{aligned}
\alpha(\beta) &= X^T V^{-1} U^{1/2} \psi(r) = 0 \\
\Phi(\theta_l) &= -\mathrm{tr}(V^{-1} \frac{\partial V}{\partial \theta_l} K) + U^{1/2} \psi(r) V^{-1} \frac{\partial V}{\partial \theta_l} V^{-1} U^{1/2} \psi(r)^T = 0 \\
\Omega(\rho) &= -\mathrm{tr}(V^{-1} \frac{\partial V}{\partial \rho} K) + U^{1/2} \psi(r) V^{-1} \frac{\partial V}{\partial \rho} V^{-1} U^{1/2} \psi(r)^T = 0
\end{aligned}$$

Spatial REBLUP for $\overline{y}_i$ is:

$$\hat{\overline{y}}_i = \frac{1}{N_i} \Big\{ \sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{y}_{ij}^{\psi,sp} \Big\} = \frac{1}{N_i} \Big\{ \sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} (x_{ij}^T \hat{\beta}^{\psi,sp} + z_{ij}^T \hat{v}_i^{\psi,sp}) \Big\}$$

Reference: Schmid and Münnich (2013).

Bias-corrections for the REBLUP

Locally: Chambers et. al. (2013) used an approach similar to the one of Welsh and Ronchetti (1998) for a robust prediction of the empirical distribution function of y leading to

$$\hat{\bar{y}}_i^{CCST} = \hat{\bar{y}}_i^{REBLUP} + (1 - \frac{n_i}{N_i})\frac{1}{n_i}\sum_{j \in s_i} \omega_{ij}^{\psi}\psi_c\left\{(y_{ij} - \hat{y}_{ij}^{\psi})/\omega_{ij}^{\psi}\right\}$$

Fully: Dongmo-Jiongo et al. (2013) used ideas similar to Chambers (1986) and the fact that the EBLUP can be written as a weighted linear function of the sample leading to

$$\begin{aligned}\hat{\bar{y}}_i^{CHAM} &= \hat{\bar{y}}_i^{REBLUP} + N_i^{-1}\sum_{j \in s_i}\psi_{k_1}\left\{(w_j - 1)(y_{ij} - \hat{y}_{ij}^{\psi})\right\} \\ &+ N_i^{-1}\sum_{\substack{h=1 \\ h \neq i}}^{m}\sum_{j \in s_h}\psi_{k_1}\left\{w_j(y_{hj} - \hat{y}_{hj}^{\psi})\right\} + N_i^{-1}\sum_{h=1}^{m}\psi_{k_2}\left\{\varpi_h \hat{v}_h^{\psi}\right\}\end{aligned}$$

Reference: Chambers et al. (2013) and Dongmo-Jiongo et al. (2013)

Bias-corrections for the SREBLUP

Locally: The approach of Chambers et. al. (2013) can be extended to the case of spatial correlation leading to

$$\hat{\bar{y}}_i^{SCCST} = \hat{\bar{y}}_i^{SREBLUP} + (1 - \frac{n_i}{N_i})\frac{1}{n_i}\sum_{j \in s_i} \omega_{ij}^{\psi,sp}\psi_c\left\{(y_{ij} - \hat{y}_{ij}^{\psi,sp})/\omega_{ij}^{\psi,sp}\right\}$$

Fully: The ideas of Dongmo-Jiongo et al. (2013) can be applied for the SREBLUP leading to

$$\begin{aligned}\hat{\bar{y}}_i^{SCHAM} = \hat{\bar{y}}_i^{SREBLUP} &+ N_i^{-1}\sum_{j \in s_i}\psi_{k_1}\left\{(w_j^{\psi,sp} - 1)(y_{ij} - \hat{y}_{ij}^{\psi,sp})\right\} \\ &+ \frac{1}{N_i}\sum_{\substack{h=1 \\ h \neq i}}^{m}\sum_{j \in s_h}\psi_{k_1}\left\{w_j^{\psi,sp}(y_{hj} - \hat{y}_{hj}^{\psi,sp})\right\} + \frac{1}{N_i}\sum_{h=1}^{m}\psi_{k_2}\left\{\varpi_h^{\psi,sp}\hat{v}_h^{\psi,sp}\right\}\end{aligned}$$

Reference: Schmid et al. (2013)

Ideas behind the SCHAM estimator

Basic idea: The spatial EBLUP of Petrucci et al. (2005) can be written as a weighted linear function of the sample leading to

$$\hat{\bar{y}}_i^{SEBLUP} = N_i^{-1} \sum_{j \in s} w_j^{sp} y_j.$$

Calculation leads to:

$$\begin{aligned}\hat{\bar{y}}_i^{SEBLUP} &= \hat{\bar{y}}_i^{SREBLUP} + N^{-1} \sum_{j \in s_i} \left(w_j^{sp} - 1\right)\left(y_j - x_j^T \hat{\beta}^{\psi,sp} - \hat{v}_i^{\psi,sp}\right) \\ &\quad + N^{-1} \sum_{\substack{h=1 \\ h \neq i}}^{m} \sum_{j \in s_h} w_j^{sp} \left(y_j - x_j^T \hat{\beta}^{\psi,sp} - \hat{v}_h^{\psi,sp}\right) + N^{-1} \sum_{h=1}^{m} \varpi_h \hat{v}_h^{\psi,sp}.\end{aligned}$$

Reference: Schmid et al. (2013)

Ideas behind the SCHAM estimator

Basic idea: The spatial EBLUP of Petrucci et al. (2005) can be written as a weighted linear function of the sample leading to

$$\hat{\bar{y}}_i^{SEBLUP} = N_i^{-1} \sum_{j \in s} w_j^{sp} y_j.$$

Robustification:

$$\begin{aligned}\hat{\bar{y}}_i^{SCHAM} &= \hat{\bar{y}}_i^{SREBLUP} + N_i^{-1} \sum_{j \in s_i} \psi_{k_1}\{(w_j^{\psi,sp} - 1)(y_{ij} - \hat{y}_{ij}^{\psi,sp})\} \\ &+ \frac{1}{N_i} \sum_{\substack{h=1 \\ h \neq i}}^{m} \sum_{j \in s_h} \psi_{k_1}\{w_j^{\psi,sp}(y_{hj} - \hat{y}_{hj}^{\psi,sp})\} + \frac{1}{N_i} \sum_{h=1}^{m} \psi_{k_2}\{\varpi_h^{\psi,sp} \hat{v}_h^{\psi,sp}\}.\end{aligned}$$

Reference: Schmid et al. (2013)

SAE 2013

Freie Universität Berlin

Contents

Model-based simulation setup

- Population data is generated for $m = 100$ small areas via

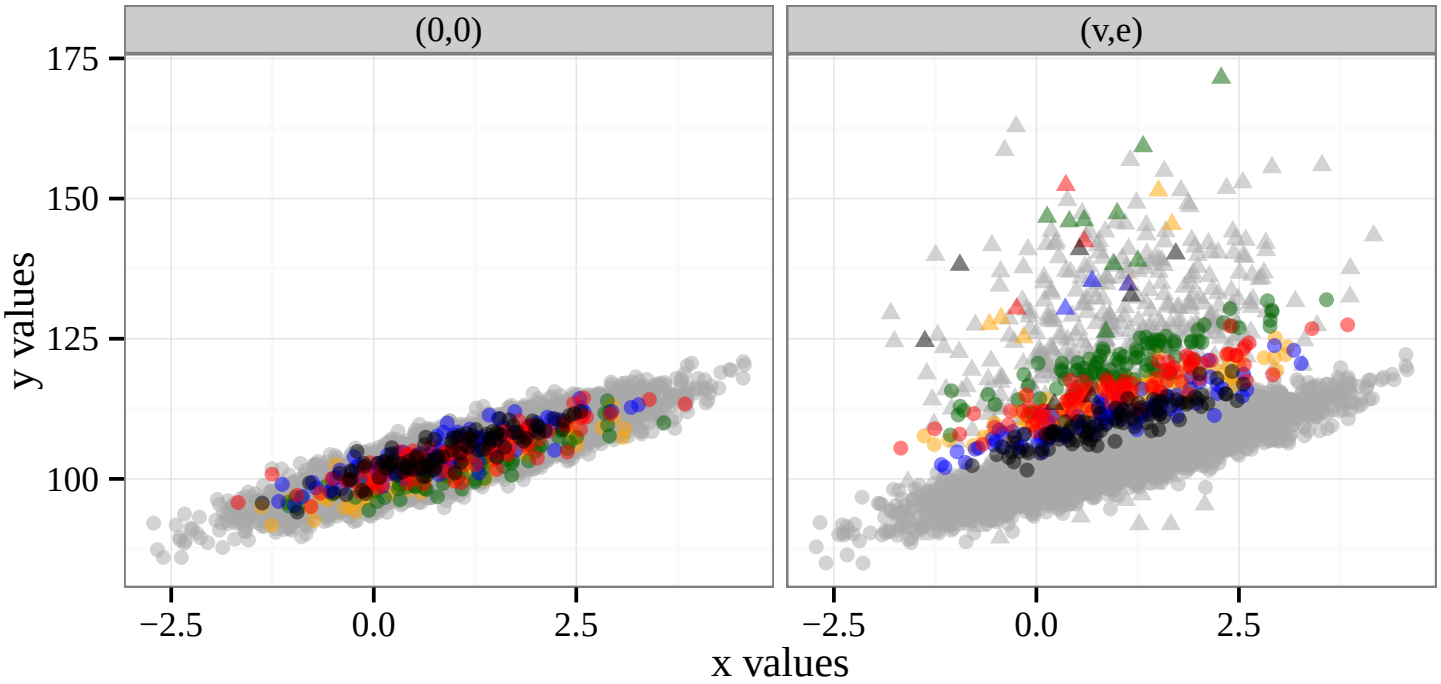
$$y_{ij} = 100 + 4x_{ij} + v_i + e_{ij}$$

- X is generated from a normal distribution with mean 1 and standard deviation 1

$$x_{ij} \sim N(1, 1)$$

- v_i and e_{ij} are generated according to two spatial and two non-spatial scenarios:
  (1) No outliers
  (2) Area and individual outliers in v_i and e_{ij}
- Samples were selected by simple random sampling without replacement within each area, $N_i = 100$ and $n_i = 5$
- Each scenario was simulated 500 times

Different scenarios



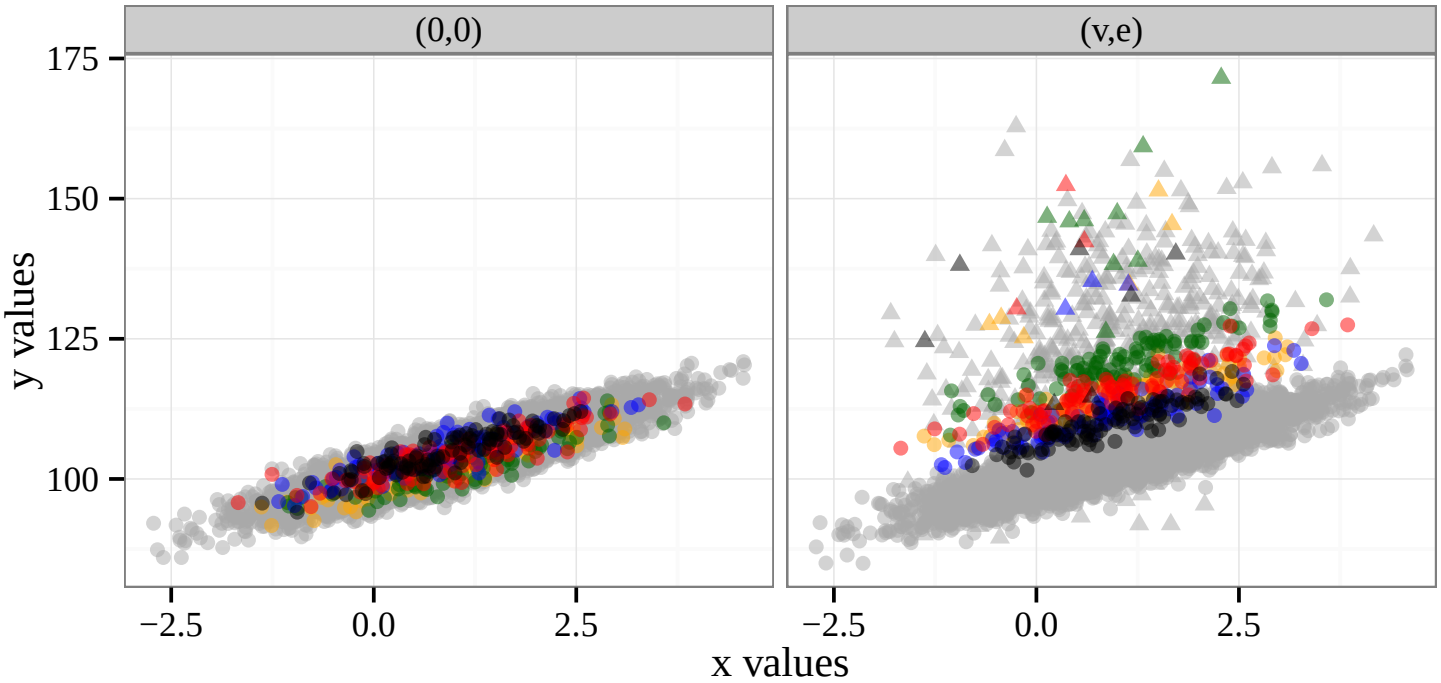
Two non-spatial scenarios:

$(0,0)$ $v \sim N(0,1)$ & $e_{ij} \sim N(0,4)$

(v,e) $v \sim 0.95N(0,1)+0.05N(9,20)$ & $e_{ij} \sim 0.95N(0,1)+0.05N(9,150)$

Appendix

Different scenarios



Two spatial scenarios:

$(0,0)_p$ $v \sim N(0, G)$ & $e_{ij} \sim N(0,4)$

$(v,e)_p$ $v \sim 0.95N(0, G)+0.05N(9,20)$ & $e_{ij} \sim 0.95N(0,1)+0.05N(9,150)$

with $G = \left((I - 0.8W)(I - 0.8W^T)\right)^{-1}$

Appendix

SAE 2013

Freie Universität Berlin

Spatial correlation between the areas

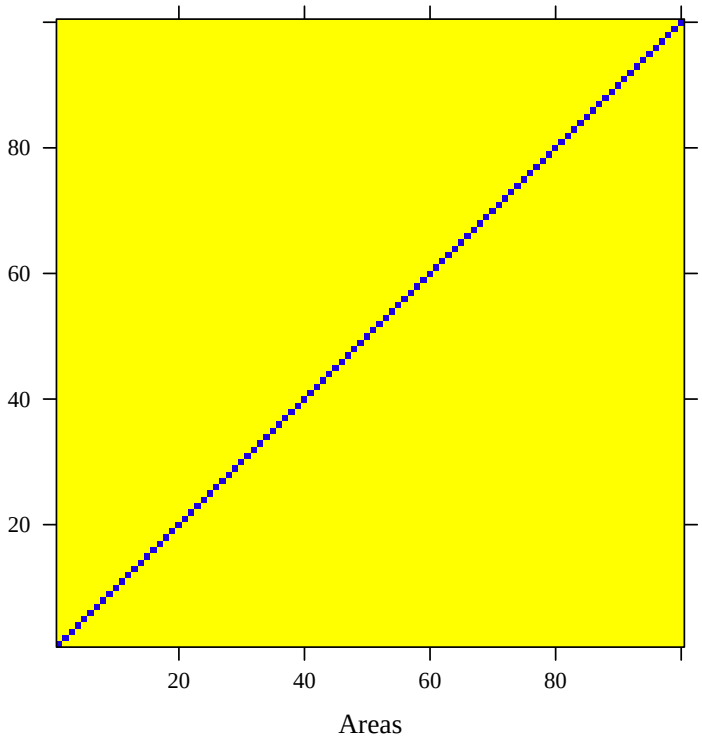


Figure : $p = 0$

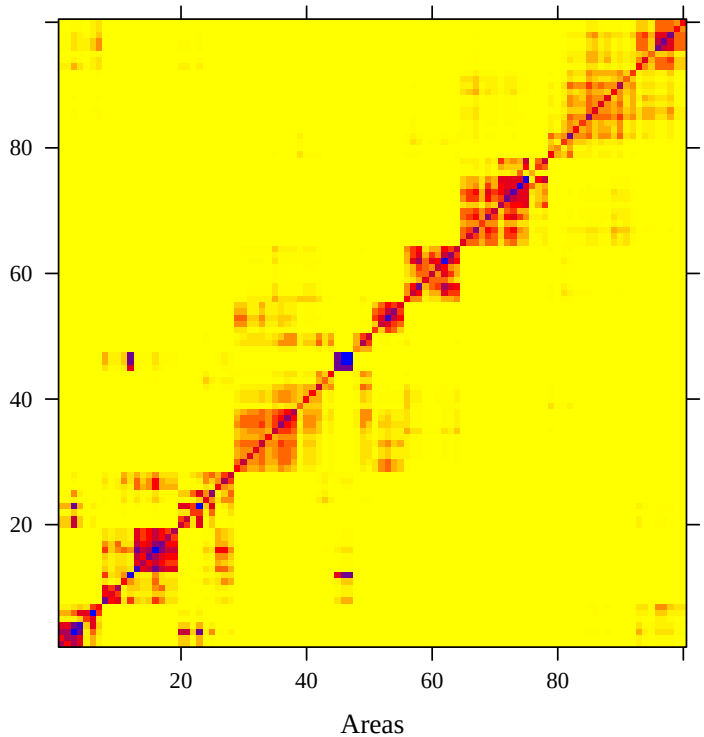


Figure : $p = 0.8$

Quality measures

Relative root mean square error [%]:

$$RRMSE_i^A = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\frac{\hat{\bar{y}}_{i,r}^A - \bar{y}_i}{\bar{y}_i}\right)^2} \cdot 100$$

Relative bias [%]:

$$RB_i^A = \frac{1}{R}\sum_{r=1}^{R}\frac{\hat{\bar{y}}_{i,r}^A - \bar{y}_{i,r}}{\bar{y}_{i,r}} \cdot 100$$

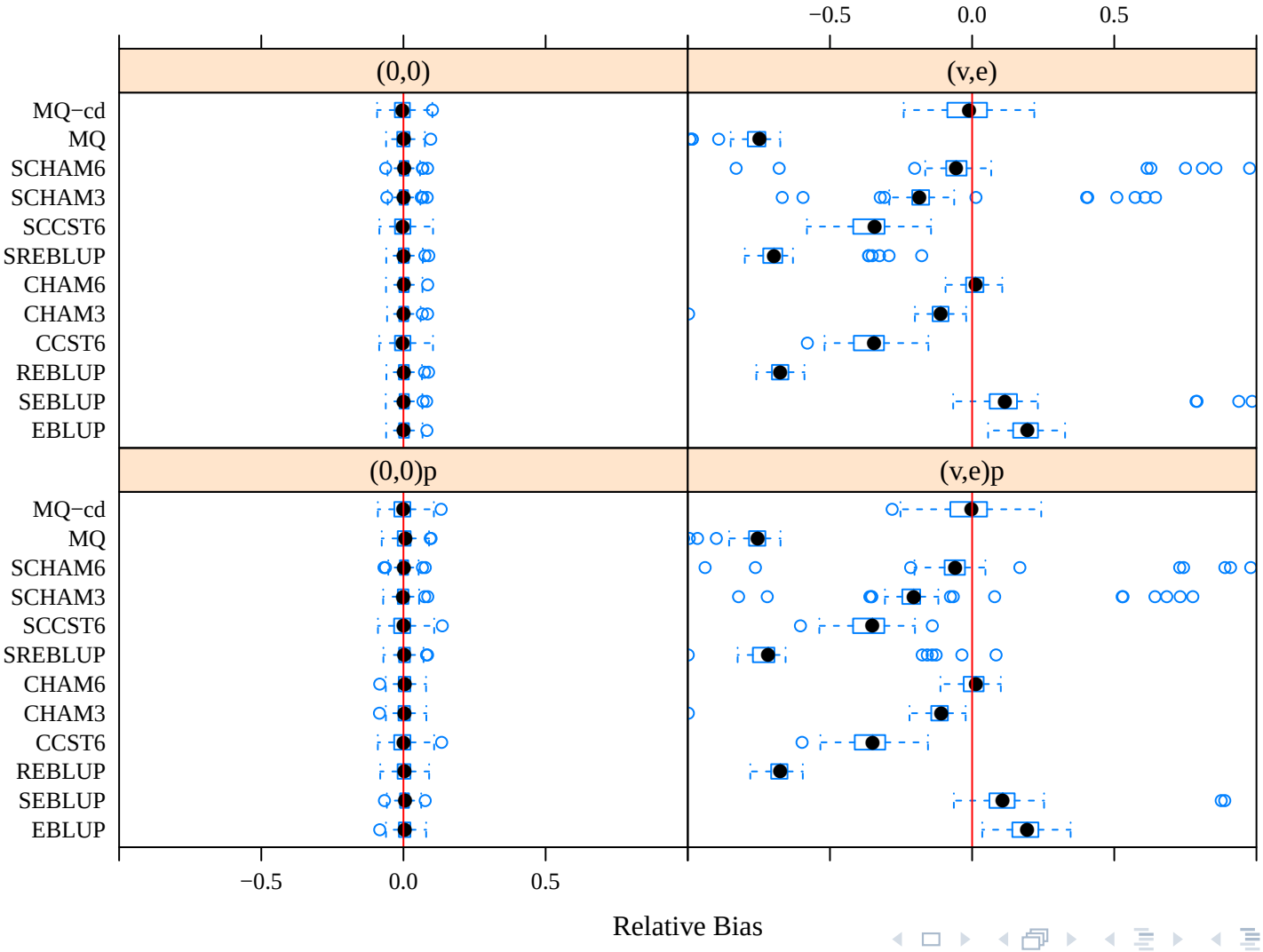
Relative efficiency [%]:

$$RE_i^A = \frac{RRMSE_i^A}{RRMSE_i^{EBLUP}} \cdot 100$$

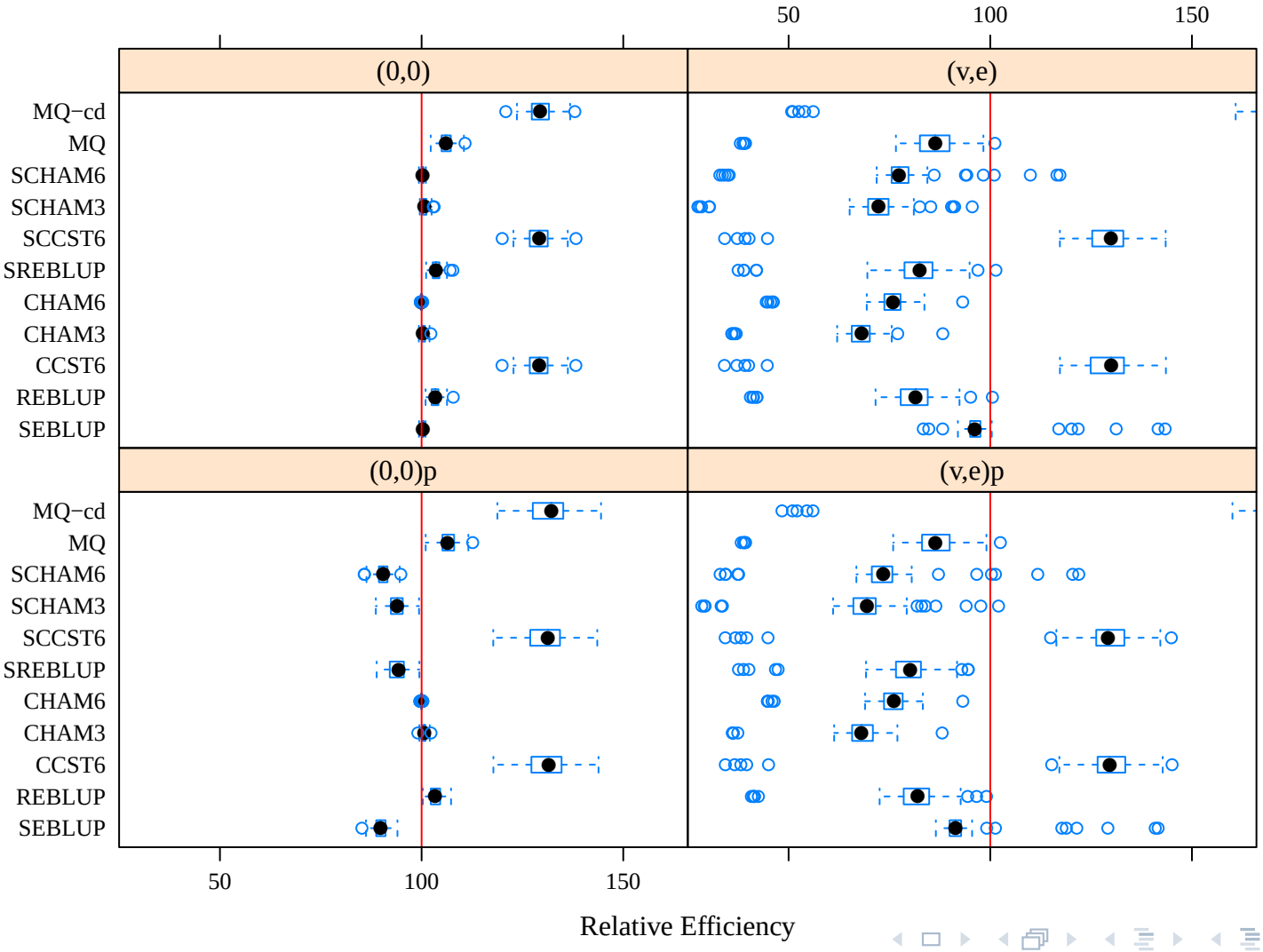
SAE 2013

Freie Universität Berlin

Relative Bias [%]



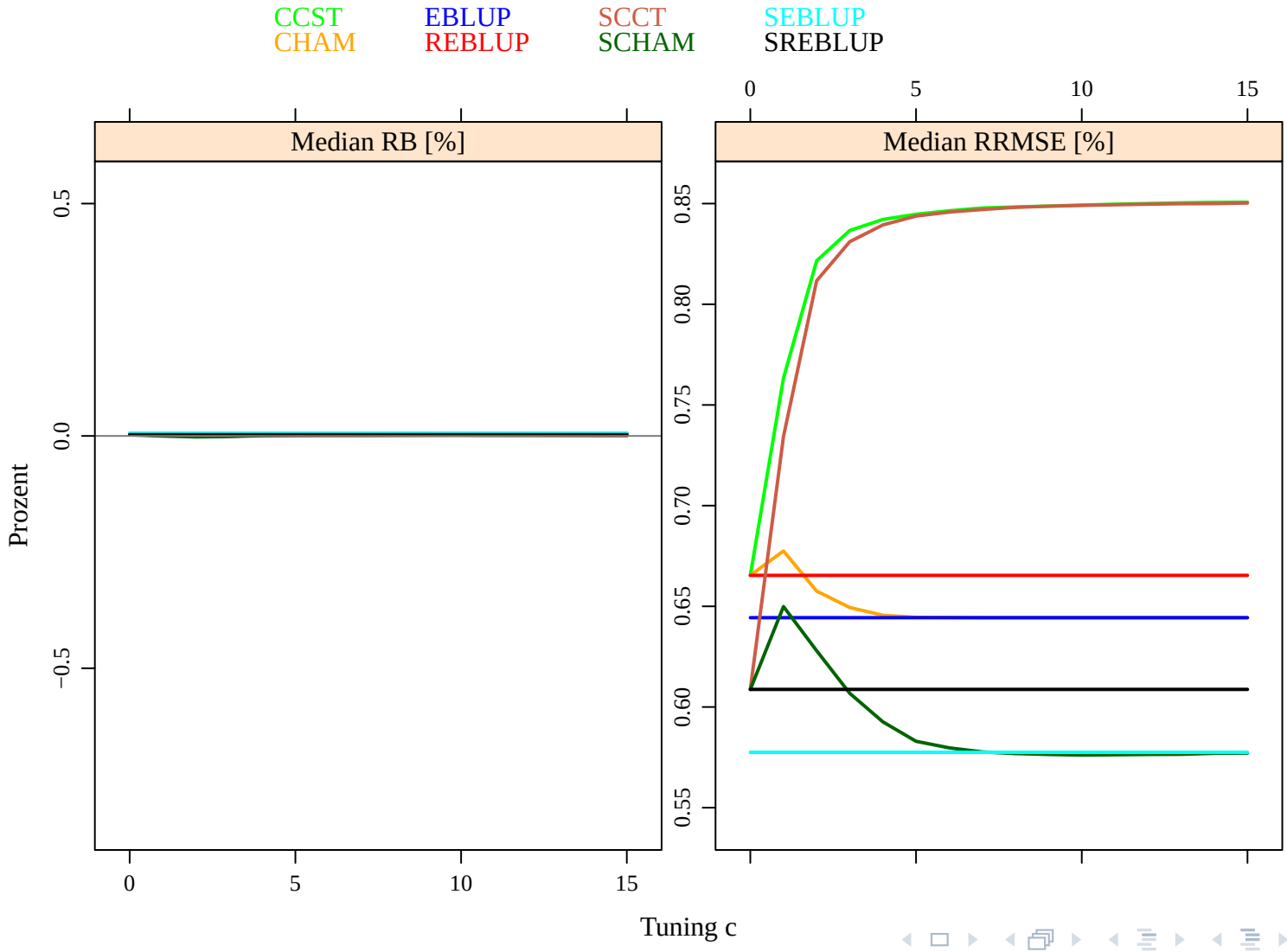
Relative Efficiency [%]



SAE 2013

Freie Universität Berlin

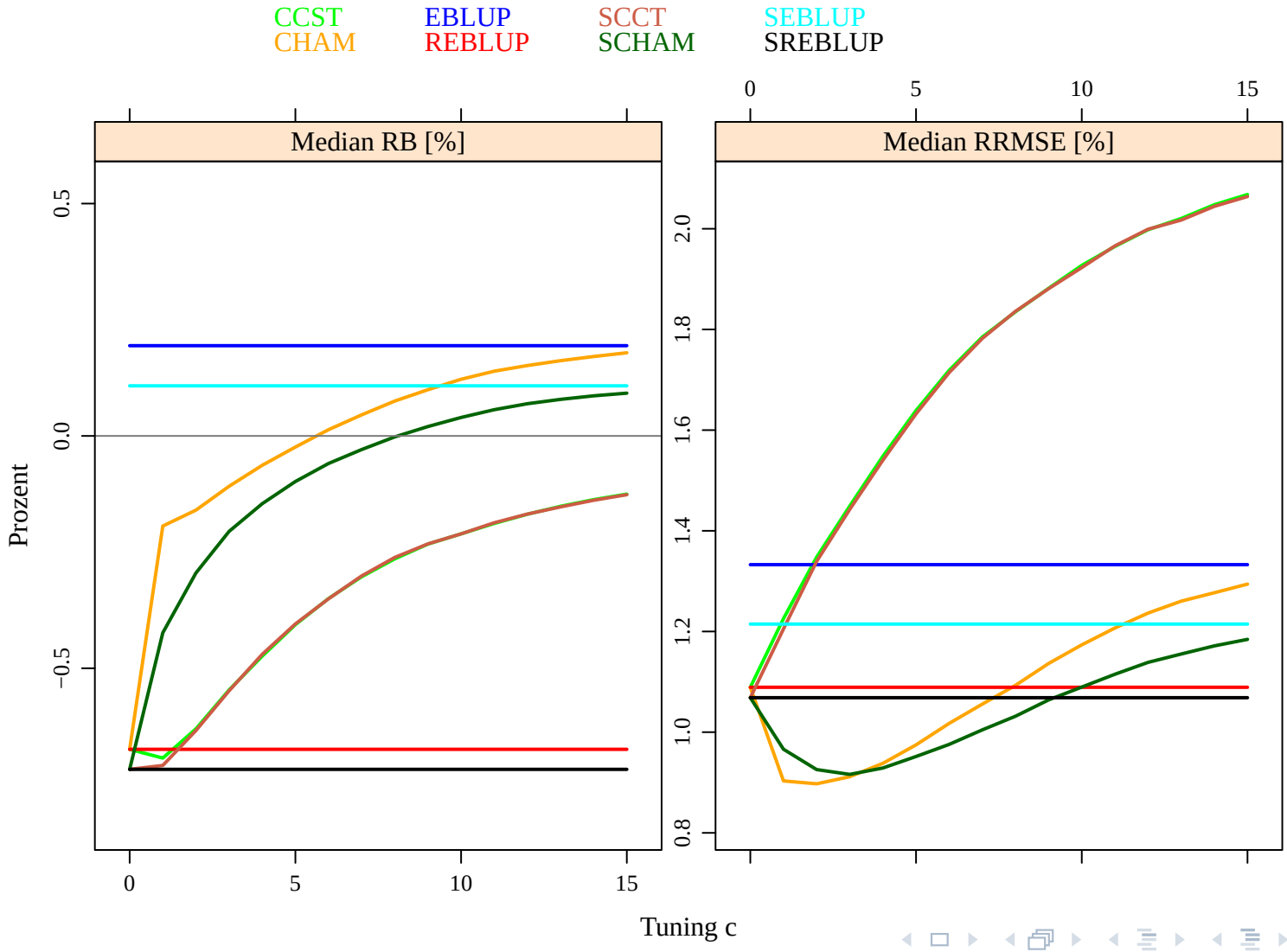
RB and RRMSE [%] - Scenario $(0, 0)_p$



SAE 2013

Freie Universität Berlin

RB and RRMSE [%] - Scenario $(v, e)_p$



SAE 2013

Freie Universität Berlin

Contents

Summary and Outlook

Summary:

- Robust projective methods (MQ, REBLUP and SREBLUP) suffer from a bias in the case of non-symmetric contamination.
- Model-based simulations indicate usefulness of fully bias-corrected spatial robust small area estimators.
- Their implementation remains challenging $\rightarrow$ selection of starting values, convergence issues, handling of large data sets.

Further research:

- Develop an analytical MSE estimation for the fully bias corrected methods.
- Investigate the proposed methods in design-based simulations.
- Choose the tuning constants by a cross validation criteria where the tuning constant is obtained in the computation and is not fixed.

SAE 2013

Freie Universität Berlin

Essential bibliography

R. Chambers *Outlier robust finite population estimation*, JASA Vol. 81 (2013), 1063-1069.

R. Chambers, H. Chandra, N. Salvati and N. Tzavidis *Outlier robust small area estimation*, Royal Statistical Society: Series B Vol. 75 (2013).

R. Chambers and N. Tzavidis *M-quantile models for small area estimation*, Biometrika Vol. 93 (2006), 255-268.

V. Jiongo, D. Haziza and P. Duchesne *Controlling the bias of robust small area estimators*, Biometrika (2013), forthcoming.

M. Pratesi and N. Salvati *Small Area Estimation in the Presence of Correlated Random Area Effects*, Statistical Methods and Application Vol. 17 (2009), 113-141.

A. Richardson and A. Welsh, *Asymptotic properties of restricted ML estimates for hierarchical mixed linear models*, Australian Journal of Statistics Vol. 36 (1994), 31-43.

S.K. Sinha and J.N.K. Rao, *Robust Small Area Estimation*, The Canadian Journal of Statistics (2009), 381-399.

T. Schmid and R. Münnich *Spatial Robust Small Area Estimation*, Statistical Papers (2013), forthcoming.

T. Schmid *Spatial Robust Small Area Estimation applied to Business Data*, phd thesis (2013), Opus, Trier.

T. Schmid, R. Chambers and R. Münnich *Bias correction of robust small area estimators under spatial correlation*, Working paper (2013).

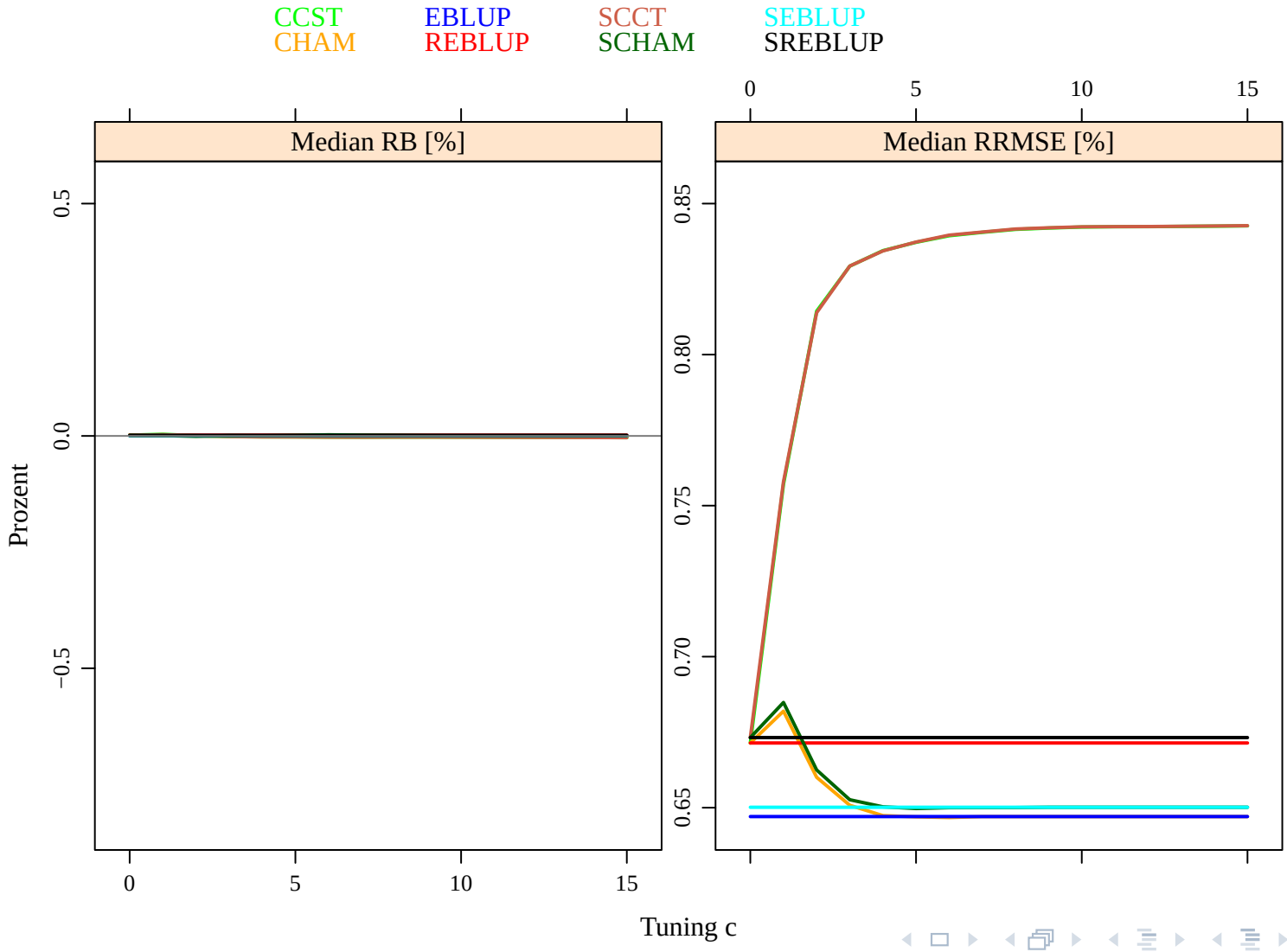
Thank you very much for your attention.

Timo Schmid (timo.schmid@fu-berlin.de)

SAE 2013

Freie Universität Berlin

RB and RRMSE [%] - Scenario $(0, 0)$



SAE 2013

Freie Universität Berlin

RB and RRMSE [%] - Scenario (v, e)