Small area estimation in the survey of Lithuanian census

Andrius Čiginas[1,2] and Ieva Burakauskaitė[2]

[1]Vilnius University
[2]Statistics Lithuania

IASS Webinar | 31 August, 2022

Main objects of the survey

- $\mathcal{U} = \{1, \ldots, N\}$ is the census population, $N = 2810761$.
- There are M domains (areas) $\mathcal{U}_1, \ldots, \mathcal{U}_M$ of known sizes $N_1, \ldots, N_M$ such that $\mathcal{U}_1 \cup \cdots \cup \mathcal{U}_M = \mathcal{U}$ and $\mathcal{U}_i \cap \mathcal{U}_j = \emptyset$ as $i \neq j$. For example, the domains are municipalities, $M = 60$.
- Categorical variables of the survey for small area estimation:
  1. *religion* (16 categories);
  2. *mother tongue* (12 categories);
  3. *knowledge of other languages* (16 languages).

  It is sufficient to consider binary variables. Let y be one of these with the fixed values $y_1, \ldots, y_N$ in $\mathcal{U}$.
- We aim to estimate the domain proportions

$$\theta_i = \frac{1}{N_i} \sum_{k \in \mathcal{U}_i} y_k, \qquad i = 1, \ldots, M,$$

  or totals $N_i \theta_i$.

Sample design and primary sampling weights

- The sample $s \subset \mathcal{U}$ of size $n < N$, $n = 436404$, was drawn according to the sampling design $p(\cdot)$ with inclusion into the sample probabilities $\pi_k = \mathrm{P_p}\{k \in s\} > 0$, $k \in \mathcal{U}$.
- We got the sample $s = s^{(1)} \cup s^{(2)} \cup s^{(3)}$, where
  1. the part $s^{(1)}$ contains individuals from the voluntary sample;
  2. $s^{(2)}$ consists of other units which cannot be included in the sampling frame (this is the part for imputation);
  3. the part $s^{(3)}$ is the probability sample drawn from the sampling frame $\mathcal{U}^{(3)} = \mathcal{U} \backslash \{s^{(1)} \cup s^{(2)}\}$.
- The primary sampling weights are $d_k = 1/\pi_k$, where $\pi_k = 1$ as $k \in s^{(1)} \cup s^{(2)}$, and, in the hth stratum of $\mathcal{U}^{(3)}$,

$$\pi_k \approx m_k n'_h / N'_h, \qquad k \in s^{(3)},$$

where N'_h is the stratum size, n'_h is the number of selected addresses, and m_k is the number of persons in the address.

Regression (calibrated) estimators in domains

- The domain samples $s_i = s \cap \mathcal{U}_i$ are of sizes $n_i \leqslant N_i$.
- Let $\mathbf{x}_k = (1, x_{2k}, \ldots, x_{Pk})'$ be a P-dimensional vector containing the values of auxiliary variables $x_2, \ldots, x_P$ for $k \in \mathcal{U}$, and $\boldsymbol{\theta}_{xi} = \sum_{k \in \mathcal{U}_i} \mathbf{x}_k / N_i$ is the vector of means for each domain $i = 1, \ldots, M$.

The generalized regression estimators (Rao and Molina, 2015)

$$\hat{\theta}_i^{\mathrm{GR}} = \boldsymbol{\theta}'_{xi} \widehat{\mathbf{B}}_i \quad \text{with} \quad \widehat{\mathbf{B}}_i = \left(\sum_{k \in s_i} \frac{\mathbf{x}_k \mathbf{x}'_k}{\pi_k} \right)^{-1} \sum_{k \in s_i} \frac{\mathbf{x}_k y_k}{\pi_k}$$

of θ_i , $i = 1, \ldots, M$, are approximately design unbiased if n_i are not small.

The set of variables $x_2, \ldots, x_P$ includes binary variables on age groups, gender, and religions in 2011 intersected with counties.

The problem

- The direct estimator $\hat{\theta}_i^{\mathrm{GR}}$ of the proportion θ_i is based only on the sample of the ith domain. The domain sample sizes n_i are small for some domains for some survey variables and there the design variances $\psi_i = \mathrm{var_p}(\hat{\theta}_i^{\mathrm{GR}})$ are large.
- The direct estimators (Rao and Molina, 2015)

$$\hat{\psi}_i^{\mathrm{GR}} = \frac{1}{N_i^2} \sum_{k \in s_i} \sum_{l \in s_i} (1 - \pi_k \pi_l / \pi_{kl}) \frac{(y_k - \mathbf{x}_k' \widehat{\mathbf{B}}_i)(y_l - \mathbf{x}_l' \widehat{\mathbf{B}}_i)}{\pi_k \pi_l}$$

  of ψ_i, where $\pi_{kl} = \mathrm{P_p}\{k, l \in s\} > 0$, have high variances themselves for small samples s_i.
- The true proportions are often very small in the estimation domains. For example, the five-number summary for 16 religions in 60 municipalities in 2011 complete data is

$$(0.000000, 0.000095, 0.000681, 0.007380, 0.922597).$$

Smoothing the direct estimators of variances

The estimators $\hat{\psi}_i^{\mathrm{GR}}$ of $\psi_i = \mathrm{var}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{GR}})$ are smoothed by applying the generalized variance function approach (Wolter, 2007).

A simple method of that type is to assume that $\psi_i \approx K N_i^{\gamma}$ and estimate the parameters $K > 0$ and $\gamma \in \mathbb{R}$ using the regression model (Dick, 1995)

$$\log(\hat{\psi}_i^{\mathrm{GR}}) = \log(K) + \gamma \log(N_i) + \eta_i, \qquad i = 1, \ldots, M,$$

where errors η_i are i.i.d. Then the smoothed variances are

$$\hat{\psi}_i^{\mathrm{S}} = \widehat{K} N_i^{\hat{\gamma}}, \qquad i = 1, \ldots, M,$$

which are next multiplied by a bias correction as suggested in Hidiroglou et al. (2019).

Note: that smoothing works well only for the two largest religions.

Alternative smoothing

Let us take $\mathbf{z}_i = (1, z_i)'$ as auxiliary data for the ith domain, where z_i is the proportion of the corresponding variable from the previous complete Census 2011.

Using the approximation (Kish, 1995)

$$\psi_i \approx \psi(\theta_i) := D_i \theta_i (1 - \theta_i)/n_i$$

and assuming that the design effects $D_i = c$ for all $i = 1, \ldots, M$,

$$\hat{\psi}_i^{\mathrm{s}} = \hat{c} z_i (1 - z_i)/n_i, \quad \text{where} \quad \hat{c} = \frac{N^2 \hat{\psi}^{\mathrm{s}}}{\sum_{i=1}^{M} \widetilde{N}_i^2 z_i (1 - z_i)/n_i},$$

are smoothed versions of the variances $\hat{\psi}_i^{\mathrm{GR}}$, $i = 1, \ldots, M$. Here the quantity $\hat{\psi}^{\mathrm{s}}$ smooths the direct estimator $\hat{\psi}^{\mathrm{GR}}$ of the variance of the calibrated estimator for the whole population proportion, and $\widetilde{N}_i$ is the size of the ith domain in Census 2011.

Regression-synthetic estimation

Let $\hat{\psi}_i$ be any estimator of the variance ψ_i. To estimate domain proportion θ_i, one can apply the regression-synthetic estimator

$$\hat{\theta}_i^{\mathrm{S}} = \hat{\theta}_i^{\mathrm{S}}(\hat{\psi}_i) = \mathbf{z}_i'\hat{\boldsymbol{\beta}} \quad \text{with} \quad \hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{M} \frac{\mathbf{z}_i \mathbf{z}_i'}{\hat{\psi}_i} \right)^{-1} \sum_{i=1}^{M} \frac{\mathbf{z}_i \hat{\theta}_i^{\mathrm{GR}}}{\hat{\psi}_i},$$

which is obtained from the basic domain-level model for EBLUP ignoring random area effects (Rao and Molina, 2015). That is, the estimator (predictor) $\hat{\theta}_i^{\mathrm{S}}$ is derived from the model

$$\hat{\theta}_i^{\mathrm{GR}} = \mathbf{z}_i'\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, M,$$

where $\boldsymbol{\beta}$ is the vector of fixed effects and the sampling errors ε_i are assumed independent with $\mathrm{E_p}(\varepsilon_i) = 0$ and $\mathrm{var_p}(\varepsilon_i) = \psi_i$.

Choosing $\hat{\psi}_i$ in $\hat{\theta}_i^{\mathrm{S}} = \hat{\theta}_i^{\mathrm{S}}(\hat{\psi}_i)$ for small proportions

Consider the approximation (Kish, 1995)

$$\hat{\psi}_i^{\mathrm{GR}} \approx \widehat{D}_i \hat{\theta}_i^{\mathrm{GR}}(1 - \hat{\theta}_i^{\mathrm{GR}})/n_i,$$

where $\widehat{D}_i$ is an estimator of the design effect D_i.

Suppose all proportions θ_i are small, say $\theta_i < 0.2$. Let us choose between $\hat{\psi}_i^{\mathrm{GR}}$ and $\hat{\psi}_i^{\mathrm{s}}$ to use in $\hat{\theta}_i^{\mathrm{S}} = \hat{\theta}_i^{\mathrm{S}}(\hat{\psi}_i)$ (Čiginas, 2022).

1. If $\hat{\theta}_i^{\mathrm{GR}} \ll \theta_i$ for the given sample s, then $\hat{\psi}_i^{\mathrm{GR}}$ underestimates ψ_i as well. Therefore, the inequality $\hat{\psi}_i^{\mathrm{s}} > \hat{\psi}_i^{\mathrm{GR}}$ should often hold, that is, the smoothed variance $\hat{\psi}_i^{\mathrm{s}}$ could be a better choice than $\hat{\psi}_i^{\mathrm{GR}}$.
2. If $\hat{\theta}_i^{\mathrm{GR}} \gg \theta_i$, then $\hat{\psi}_i^{\mathrm{GR}}$ overestimates ψ_i. Then the inequality $\hat{\psi}_i^{\mathrm{s}} < \hat{\psi}_i^{\mathrm{GR}}$ should hold for the outlier $\hat{\theta}_i^{\mathrm{d}}$. That larger estimate $\hat{\psi}_i^{\mathrm{GR}}$ down-weights the outlying observation used in $\hat{\theta}_i^{\mathrm{S}}$.

We get the estimators $\hat{\psi}_i^{\mathrm{c}} = \max\{\hat{\psi}_i^{\mathrm{s}}, \hat{\psi}_i^{\mathrm{GR}}\}$ of ψ_i to use in $\hat{\theta}_i^{\mathrm{S}}(\hat{\psi}_i)$.

Classification of religions

- We call the religion *very small* if for the set

$$A = \{i = 1, \ldots, M : \hat{\theta}_i^{\mathrm{GR}} > 0\} \quad \text{holds} \quad \#A < 2M/3;$$

or, for the data triplets $(\hat{\theta}_i^{\mathrm{GR}}, \mathbf{z}_i = (1, z_i)', \hat{\psi}_i^{\mathrm{c}})$, $i \in A$, the slope or the whole $\boldsymbol{\beta}$ (jointly) does not differ statistically significantly from zero in the regression model

$$\hat{\theta}_i^{\mathrm{GR}} = \mathbf{z}_i' \boldsymbol{\beta} + \varepsilon_i, \qquad i \in A, \tag{RS}$$

used to build the regression-synthetic estimator with weights $\hat{\psi}_i^{\mathrm{c}}$; or the correlation between $\hat{\theta}_i^{\mathrm{GR}}$ and z_i, $i \in A$, is lower than 0.4.

- All other can be treated as *small* except *Roman Catholics* (consisting of about 77% of the Census 2011 population) and *Not indicated* (10%), and maybe *None* (6%) and *Orthodox* (4%). But let us call the latter 4 religions *small*, too.

Auxiliary simulations using Census 2011 data

To decide on the model choices for small area estimation and to test hypotheses, we imitate Census 2021 survey using completely known data for Census 2011 and 2001.

1. We transfer the non-probability sample of 2021 to the Census 2011 population (only about 7% of the sample disappears) to get the sample part $s^{(1)}$.
2. We construct the part $s^{(2)}$ (of persons not in the sampling frame) similarly as in Census 2021.
3. We draw $R = 200$ probability samples $s^{(3)}$.

Estimations and other procedures of interest are repeated for R samples $s = s^{(1)} \cup s^{(2)} \cup s^{(3)}$.

Alternative synthetic estimation

We cannot build proper regression-synthetic estimators for *very small* religions using regression model (RS).

Some statistical facts from the data for these religions:

- the simulation using Census 2011 data shows that the joint hypothesis $H_0 : \boldsymbol{\beta} = (0, 1)'$ cannot be rejected for a large part of repeated samples;
- and, for the single sample of Census 2021, the same hypothesis cannot be rejected for some of the religions.

Therefore, for each *very small* religion, we apply the synthetic estimators

$$\hat{\theta}_i^{\mathrm{S}} = z_i, \qquad i = 1, \ldots, M,$$

which are the constants.

Design-based composite estimation

We introduce the composite (shrinkage) estimators (Čiginas, 2022)

$$\hat{\theta}_i^{\mathrm{C}} = \hat{\lambda}_i \hat{\theta}_i^{\mathrm{GR}} + (1 - \hat{\lambda}_i)\hat{\theta}_i^{\mathrm{S}}(\hat{\psi}_i^{\mathrm{c}}) \quad \text{with} \quad \hat{\lambda}_i = \frac{\min\{\hat{\psi}_i^{\mathrm{s}}, \hat{\psi}_i^{\mathrm{GR}}\}}{\hat{\psi}_i^{\mathrm{c}}}$$

of θ_i. If the direct estimator $\hat{\theta}_i^{\mathrm{GR}}$ is an outlier by its too small or too large value, then relatively more weight is attached to the synthetic part $\hat{\theta}_i^{\mathrm{S}}$ due to the monotonicity of the function $\psi(\theta_i)$.

Remark: these design-based compositions are applicable also if the proportions θ_i are not small, but then the inequalities

$$\max\{\theta_i, \hat{\theta}_i^{\mathrm{GR}}\} < 1/2 \quad \text{or} \quad \min\{\theta_i, \hat{\theta}_i^{\mathrm{GR}}\} > 1/2$$

should be satisfied. If these inequalities are not valid for some domains, that composite estimation is less efficient.

Estimation of mean square errors

General method. Treating the composition $\hat{\theta}_i^{\mathrm{C}}$ as a synthetic estimator, one can use the estimator (Gonzalez and Waksberg, 1973)

$$\mathrm{mse}_{\mathrm{u}}(\hat{\theta}_i^{\mathrm{C}}) = (\hat{\theta}_i^{\mathrm{C}} - \hat{\theta}_i^{\mathrm{GR}})^2 - \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{C}} - \hat{\theta}_i^{\mathrm{GR}}) + \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{C}})$$

of $\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{C}})$, where $\hat{\sigma}^2(\cdot)$ stands for any estimator of $\mathrm{var}_{\mathrm{p}}(\cdot)$. It is approximately unbiased but has a large design variance and can take negative values (Rao and Molina, 2015).

Alternative method. Assuming that $\hat{\theta}_i^{\mathrm{C}}$ approximates the optimal linear combination of $\hat{\theta}_i^{\mathrm{GR}}$ and $\hat{\theta}_i^{\mathrm{S}}$ well, we apply the estimator (Čiginas, 2021)

$$\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{C}}) = \hat{\lambda}_i(1 - \hat{\lambda}_i)\hat{\psi}_i^{\mathrm{s}} + \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{C}})$$

of $\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{C}})$. The estimator takes only non-negative values but can be biased. We use Rao et al. (1992) bootstrap to get $\hat{\sigma}^2(\hat{\theta}_i^{\mathrm{C}})$.

EBLUP based on the Fay–Herriot model

The basic Fay–Herriot (FH) model (Fay and Herriot, 1979)

$$\hat{\theta}_i^{\mathrm{GR}} = \mathbf{z}_i'\boldsymbol{\beta} + v_i + \varepsilon_i, \qquad i = 1, \ldots, M,$$

extends the regression model for $\hat{\theta}_i^{\mathrm{S}}$, where random area effects v_i are i.i.d. with $\mathrm{E}(v_i) = 0$ and $\mathrm{var}(v_i) = \sigma_v^2$ and independent of ε_i.

EBLUP of θ_i is expressed as the linear combination

$$\hat{\theta}_i^{\mathrm{EBLUP}} = \hat{\gamma}_i \hat{\theta}_i^{\mathrm{GR}} + (1 - \hat{\gamma}_i)\mathbf{z}_i'\hat{\boldsymbol{\beta}} \quad \text{with} \quad \hat{\gamma}_i = \hat{\sigma}_v^2/(\hat{\psi}_i^{\mathrm{s}} + \hat{\sigma}_v^2)$$

and

$$\hat{\boldsymbol{\beta}} = \Big[\sum_{i=1}^{M} \mathbf{z}_i\mathbf{z}_i'/(\hat{\psi}_i^{\mathrm{s}} + \hat{\sigma}_v^2)\Big]^{-1} \sum_{i=1}^{M} \mathbf{z}_i\hat{\theta}_i^{\mathrm{GR}}/(\hat{\psi}_i^{\mathrm{s}} + \hat{\sigma}_v^2),$$

where $\hat{\sigma}_v^2$ is an estimator of the variance σ_v^2.

For *very small* religions, we simplify the FH model and EBLUP by assuming that $\boldsymbol{\beta} = (0, 1)'$.

Mean square error estimation for EBLUP

We use the estimator $\hat{\sigma}_v^2$ of σ_v^2 based on the method of moments according to Fay and Herriot (1979). An approximately unbiased estimator of MSE of $\hat{\theta}_i^{\mathrm{EBLUP}}$ was derived in Datta et al. (2005):

$$\begin{aligned}\mathrm{mse}(\hat{\theta}_i^{\mathrm{EBLUP}}) =& \hat{\gamma}_i \hat{\psi}_i^{\mathrm{s}} + (1-\hat{\gamma}_i)^2 \Bigg[\mathbf{z}_i' \left(\sum_{j=1}^{M} \frac{\mathbf{z}_j \mathbf{z}_j'}{\hat{\psi}_j^{\mathrm{s}} + \hat{\sigma}_v^2} \right)^{-1} \mathbf{z}_i \\ &+ \frac{4M}{\hat{\psi}_i^{\mathrm{s}} + \hat{\sigma}_v^2} \left(\sum_{j=1}^{M} \frac{1}{\hat{\psi}_j^{\mathrm{s}} + \hat{\sigma}_v^2} \right)^{-2} \\ &- 2\hat{\sigma}_v^2 \left(\sum_{j=1}^{M} \hat{\gamma}_j \right)^{-3} \left\{ M \sum_{j=1}^{M} \hat{\gamma}_j^2 - \left(\sum_{j=1}^{M} \hat{\gamma}_j \right)^2 \right\} \Bigg].\end{aligned}$$

For comparison, if to ignore the covariance term, we have

$$\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{C}}) \approx \hat{\lambda}_i \hat{\psi}_i^{\mathrm{s}} + (1-\hat{\lambda}_i)^2 \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{S}})$$

for the design-based composite estimators.

Shrinkage estimator vs EBLUP for Census 2011 data

Using the described simulation setup, we compare the root mean square errors (RMSEs) and the absolute biases (ABs) for:

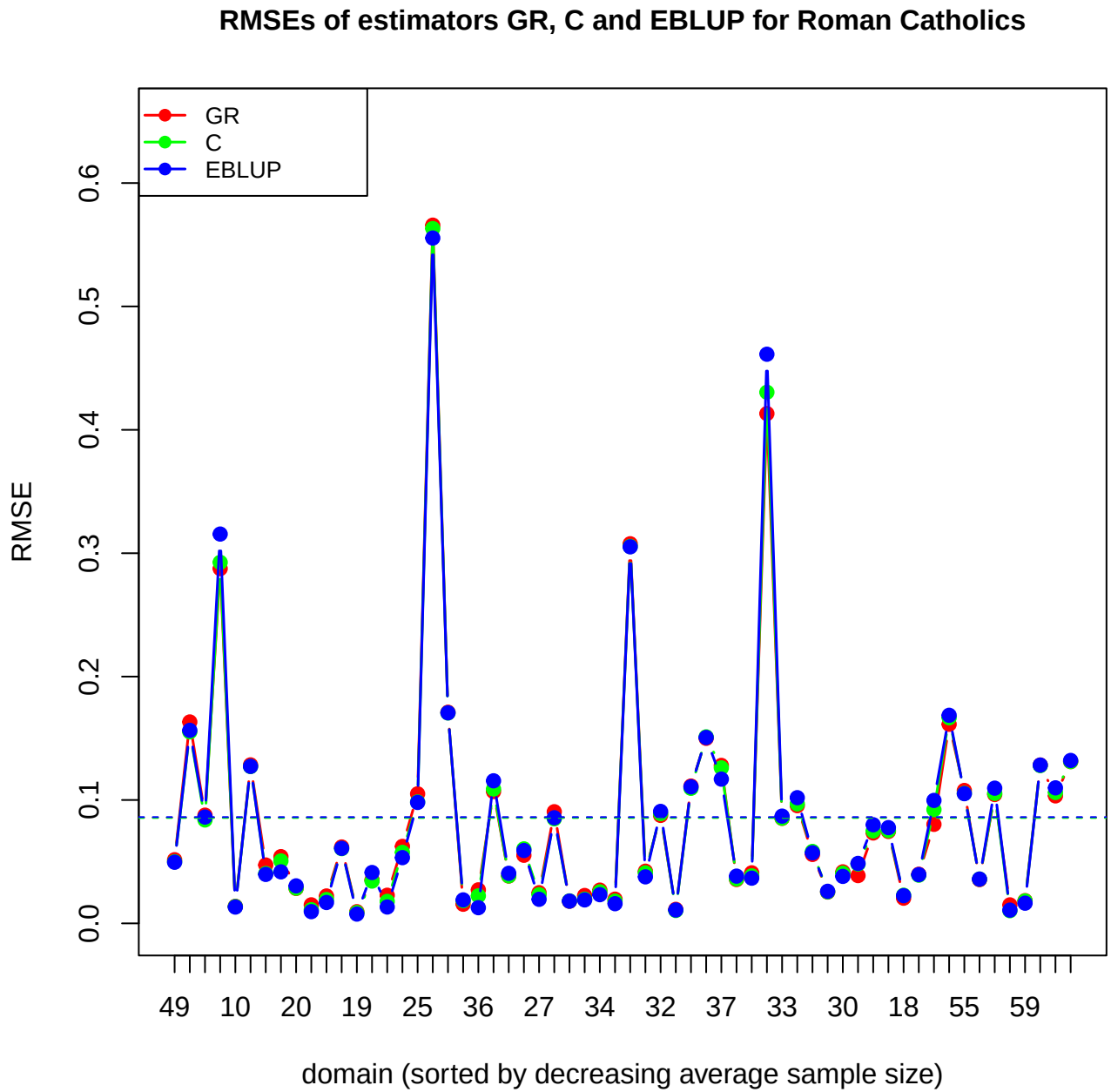
- the estimators $\hat{\theta}_i^{\mathrm{GR}}$, $\hat{\theta}_i^{\mathrm{C}}$, and $\hat{\theta}_i^{\mathrm{EBLUP}}$ of θ_i;
- the estimators $\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{C}})$ and $\mathrm{mse}(\hat{\theta}_i^{\mathrm{EBLUP}})$ of $\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{C}})$ and $\mathrm{MSE}(\hat{\theta}_i^{\mathrm{EBLUP}})$, respectively.

We evaluate all estimators for each of the R samples and calculate approximations to their RMSEs and ABs. That is, we use the accuracy measures

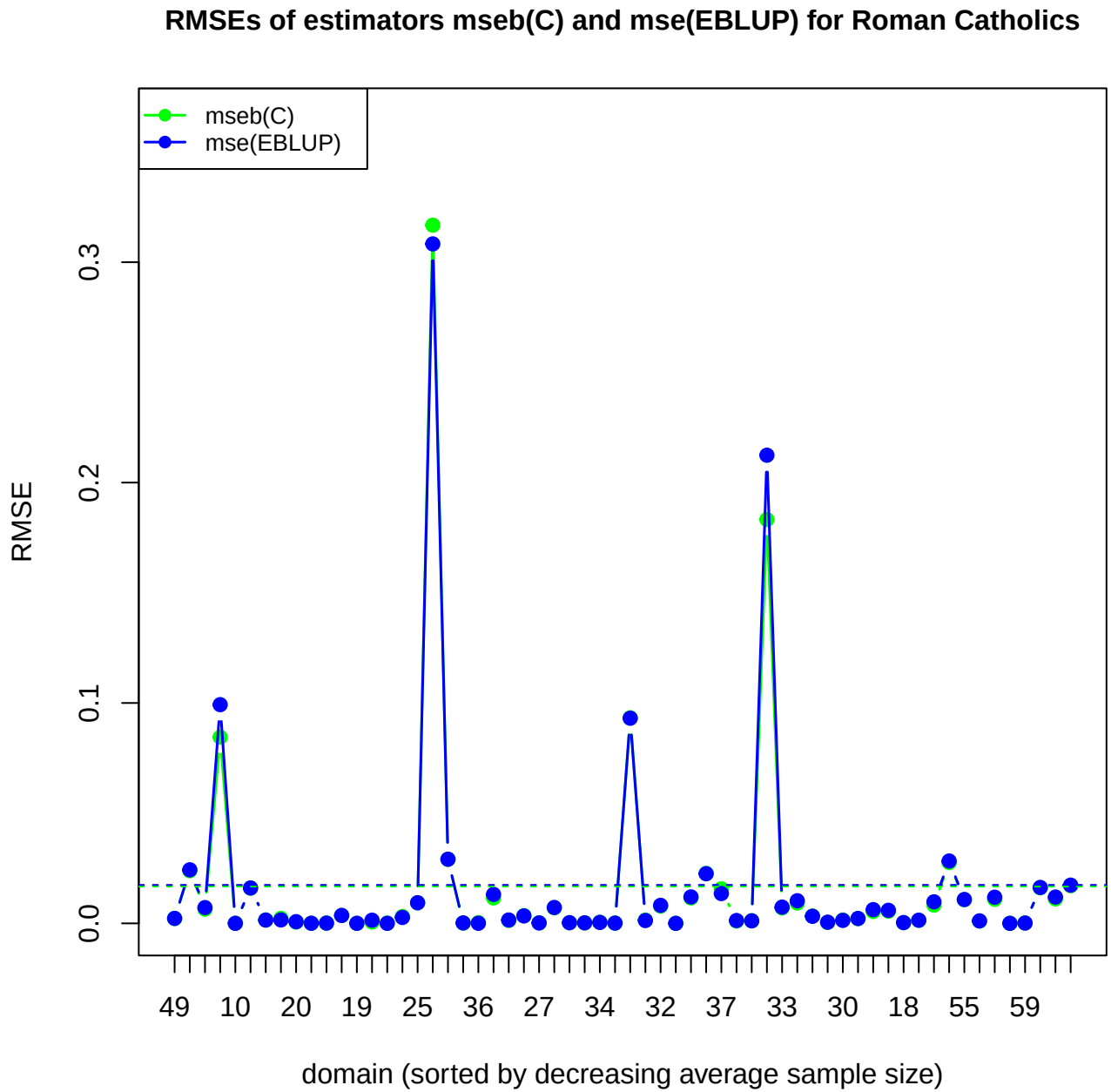
$$\mathrm{RMSE}(\hat{\mu}_i) = \left(\frac{1}{R} \sum_{r=1}^{R} (\hat{\mu}_i^{(r)} - \mu_i)^2 \right)^{1/2}, \mathrm{AB}(\hat{\mu}_i) = \left| \frac{1}{R} \sum_{r=1}^{R} \hat{\mu}_i^{(r)} - \mu_i \right|,$$

where $\hat{\mu}_i^{(r)}$ is a realization of the specific estimator $\hat{\mu}_i$ of the parameter μ_i, based on the rth sample.

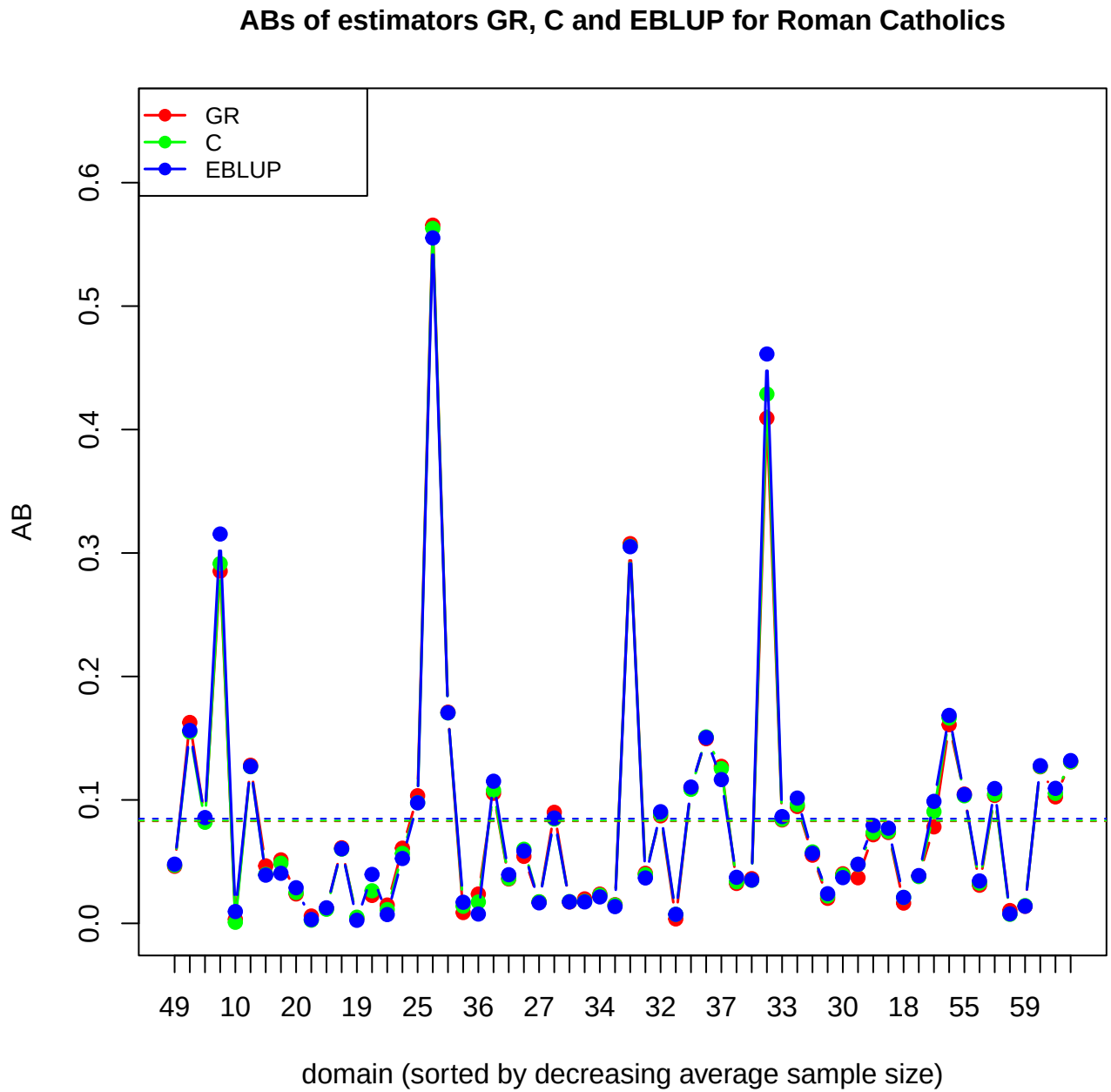
Small religion (i). RMSEs of estimators



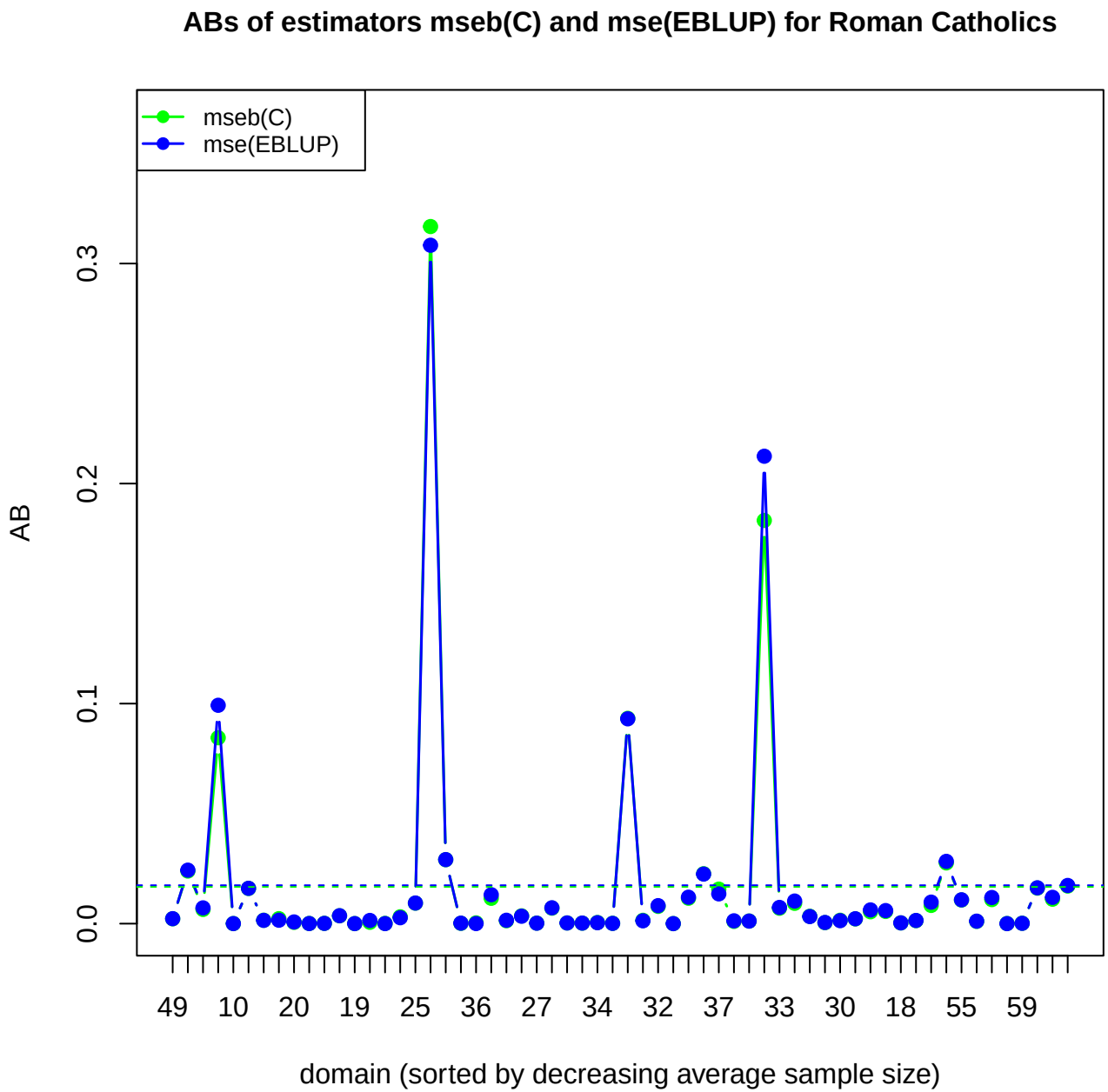
Small religion (i). RMSEs of MSE estimators



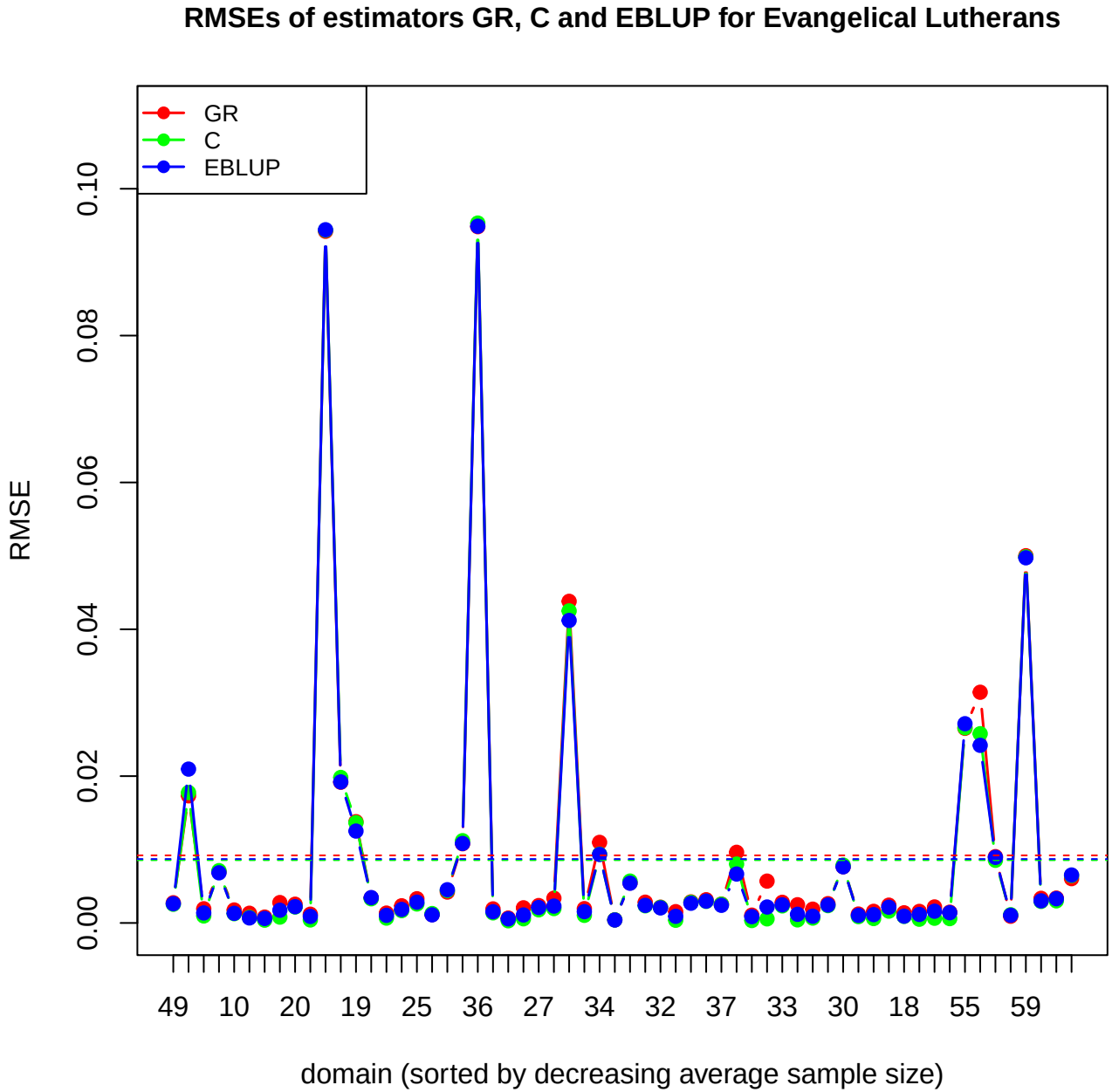
Small religion (i). ABs of estimators



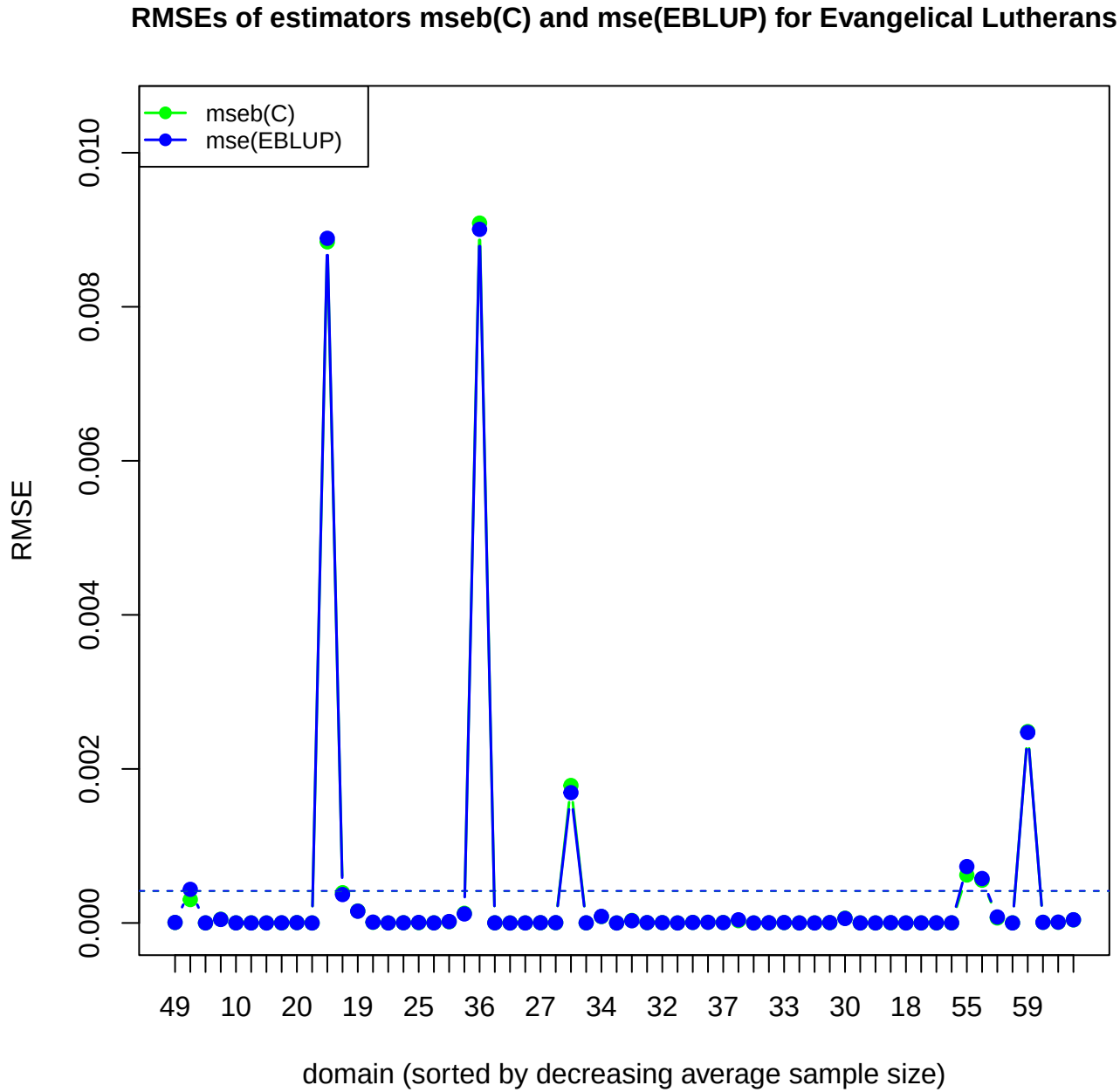
Small religion (i). ABs of MSE estimators



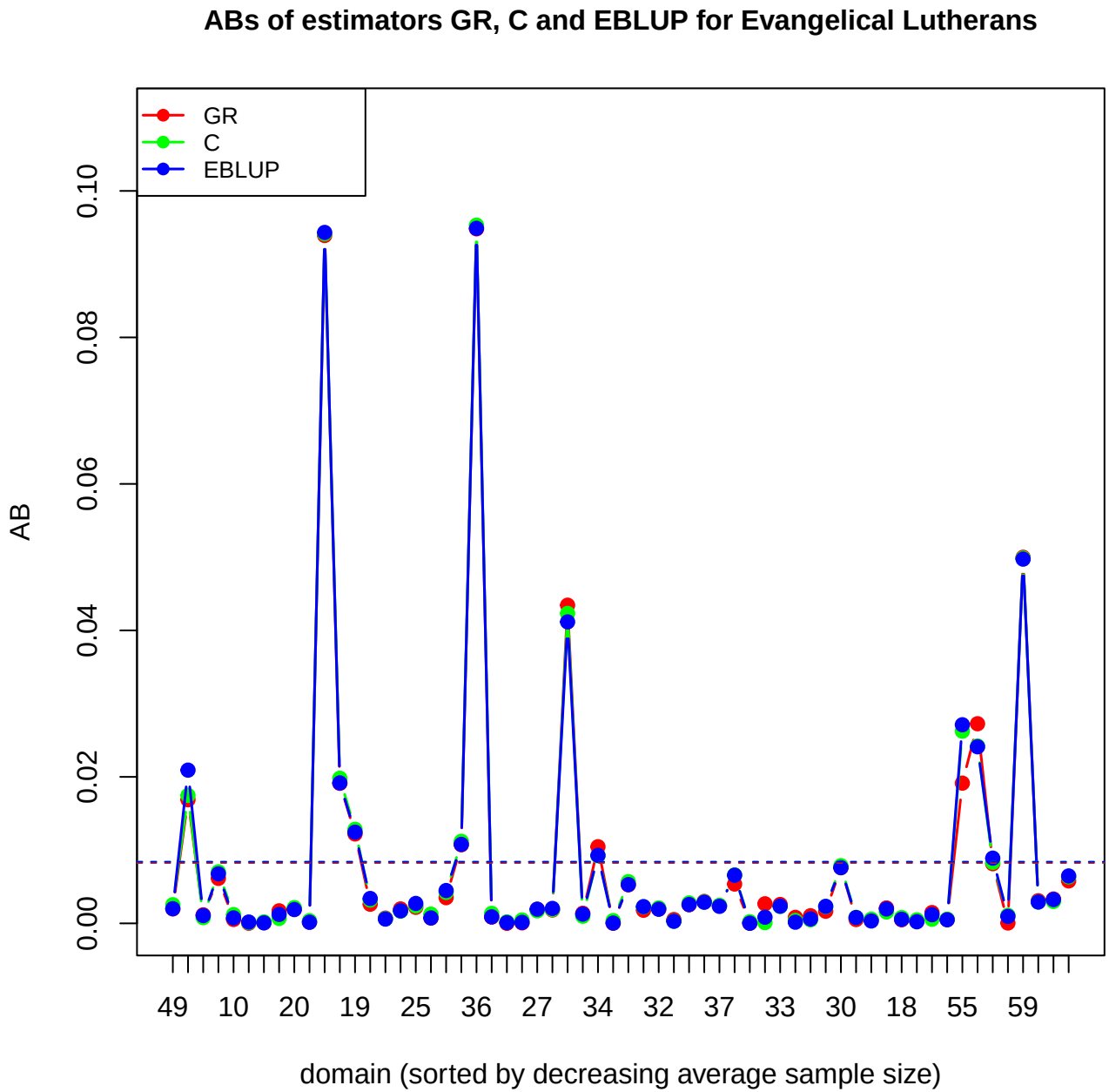
Small religion (ii). RMSEs of estimators



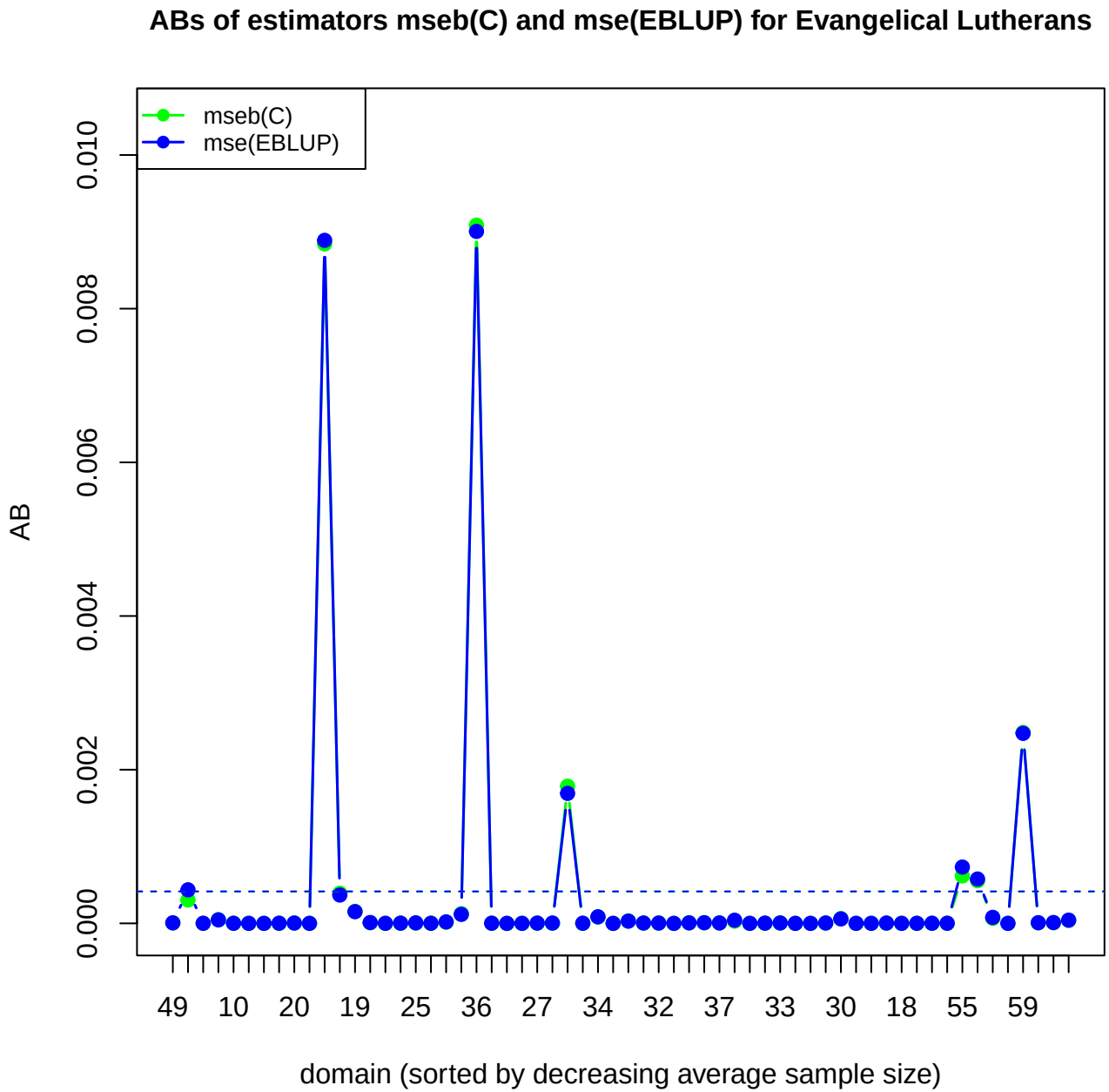
Small religion (ii). RMSEs of MSE estimators



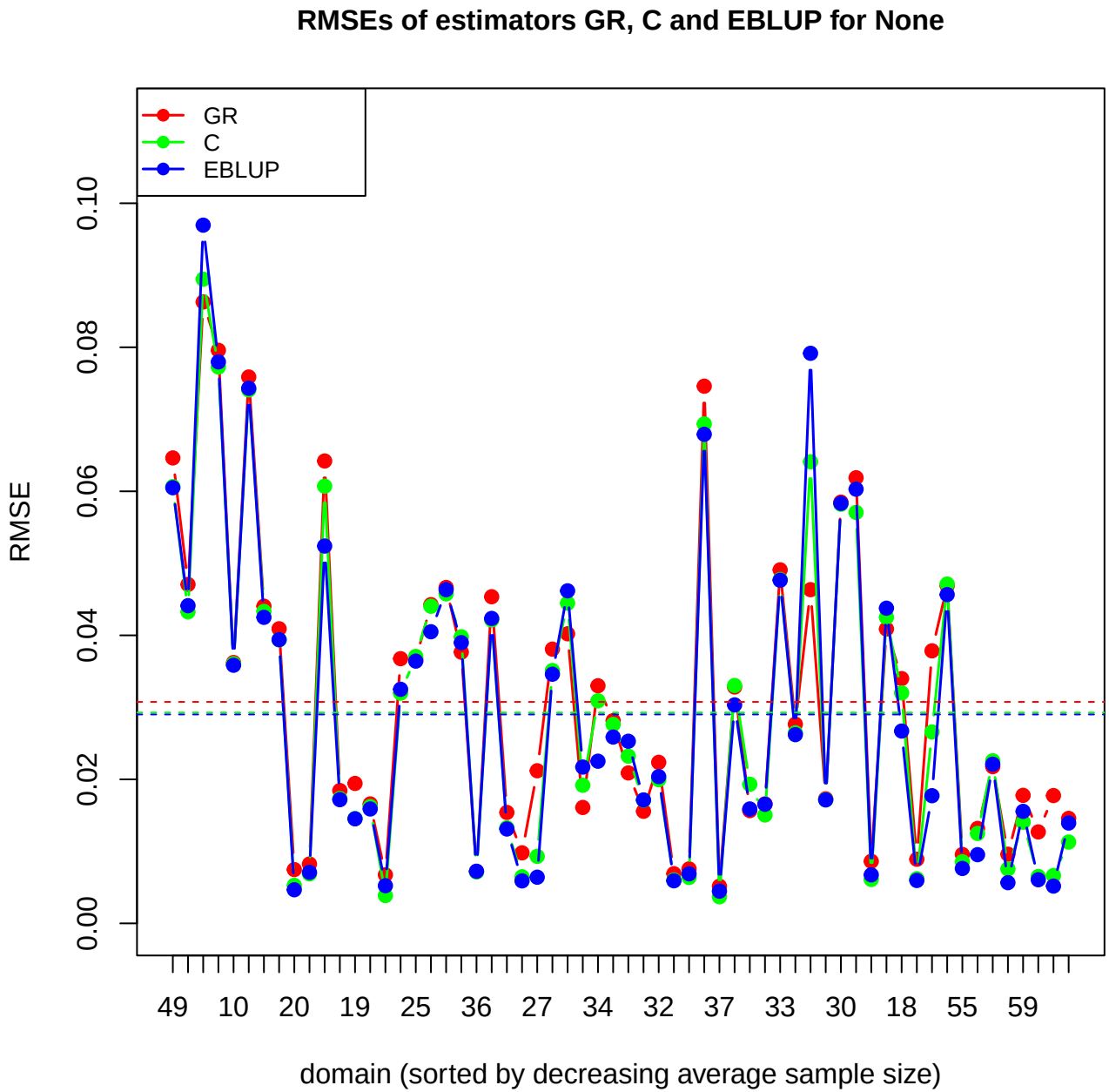
Small religion (ii). ABs of estimators



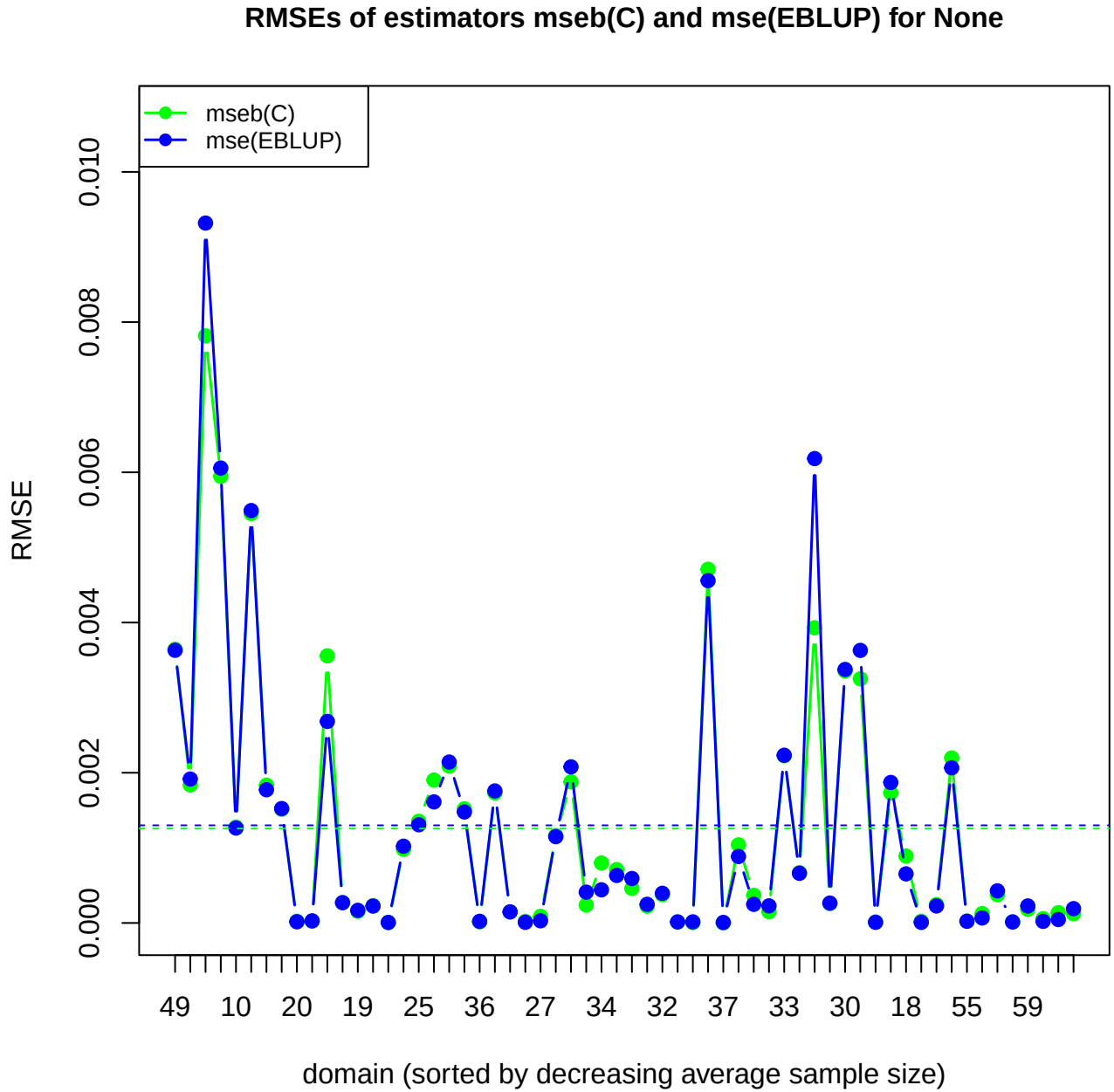
Small religion (ii). ABs of MSE estimators



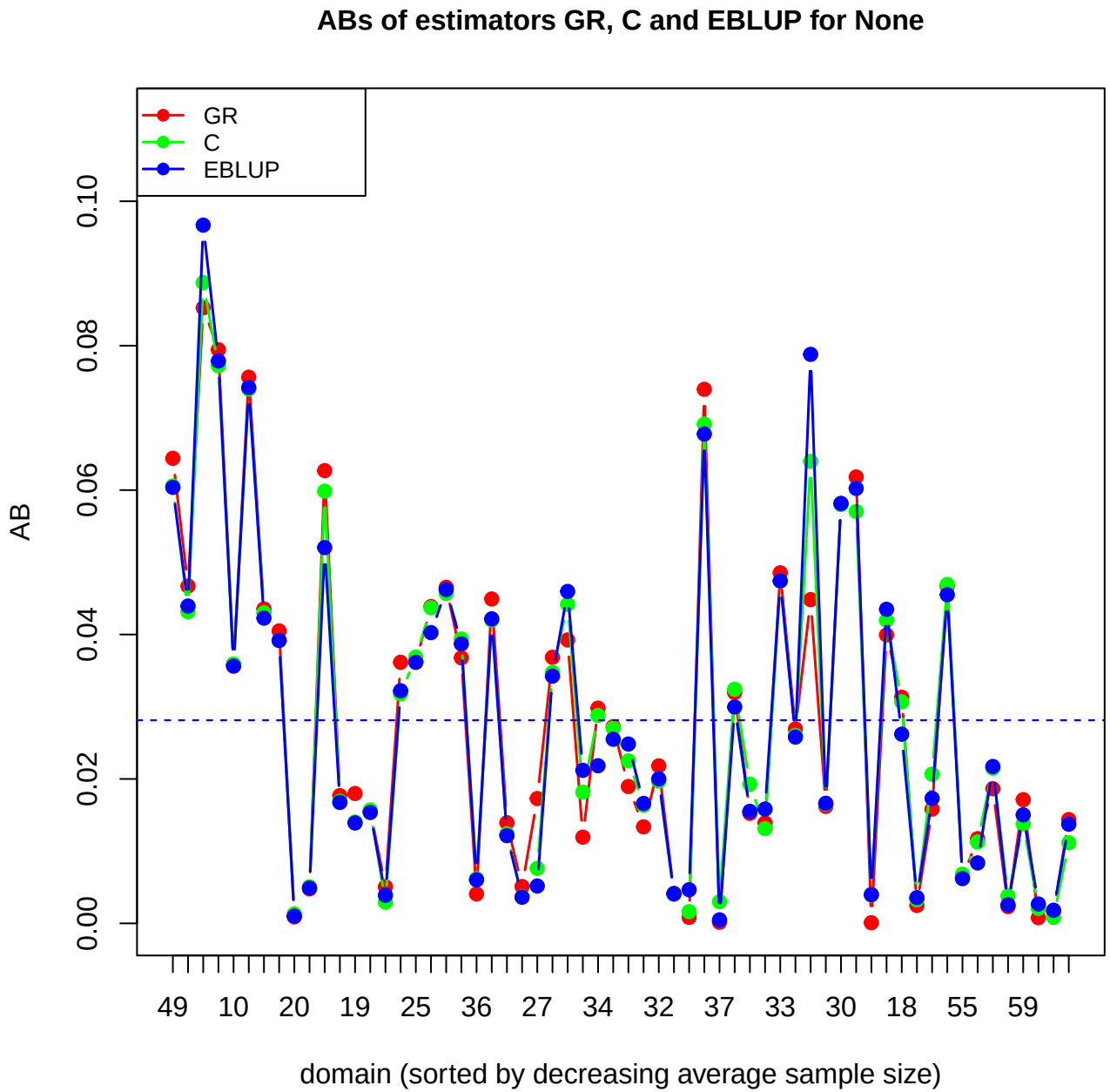
Small religion (iii). RMSEs of estimators



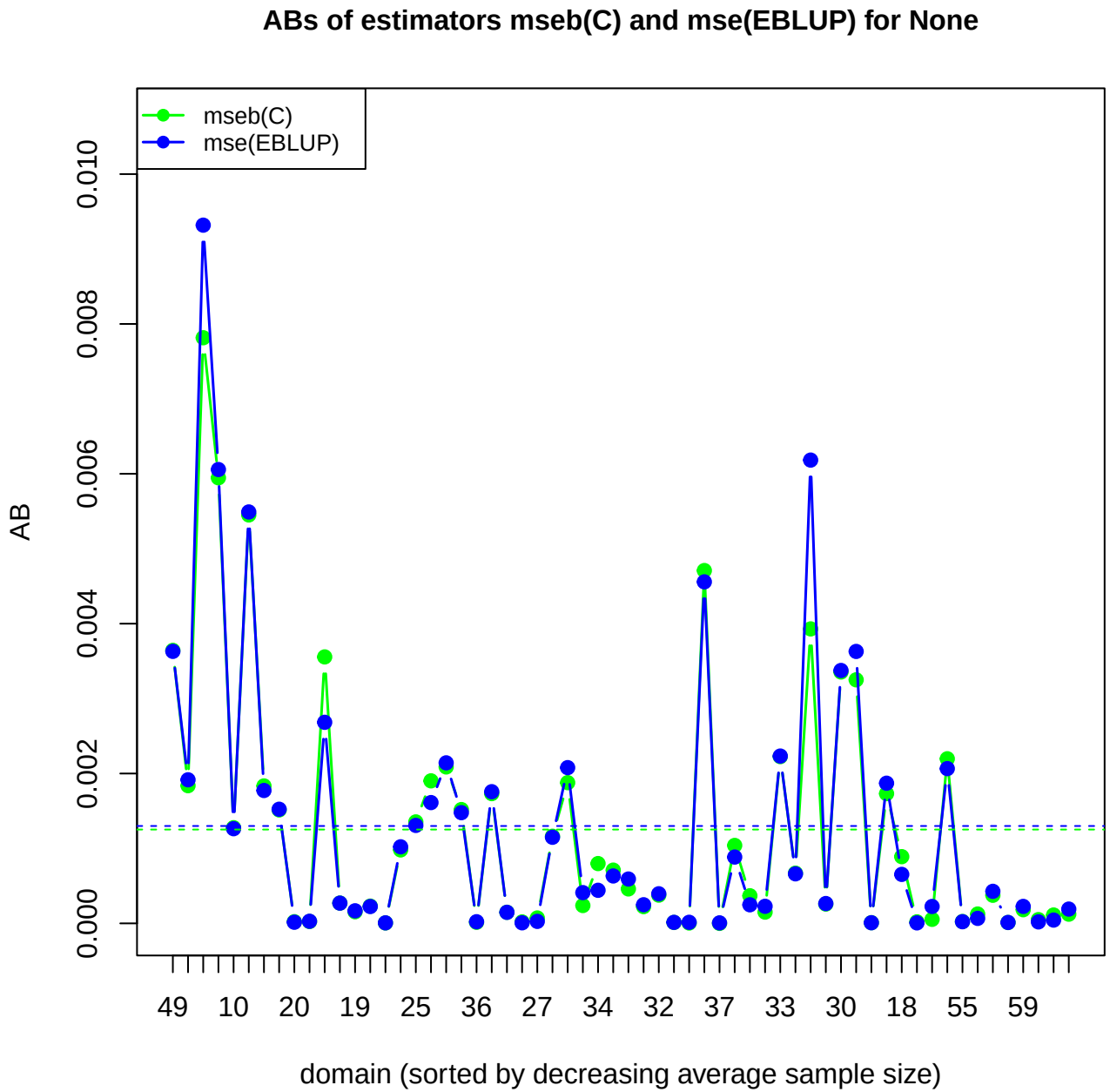
Small religion (iii). RMSEs of MSE estimators



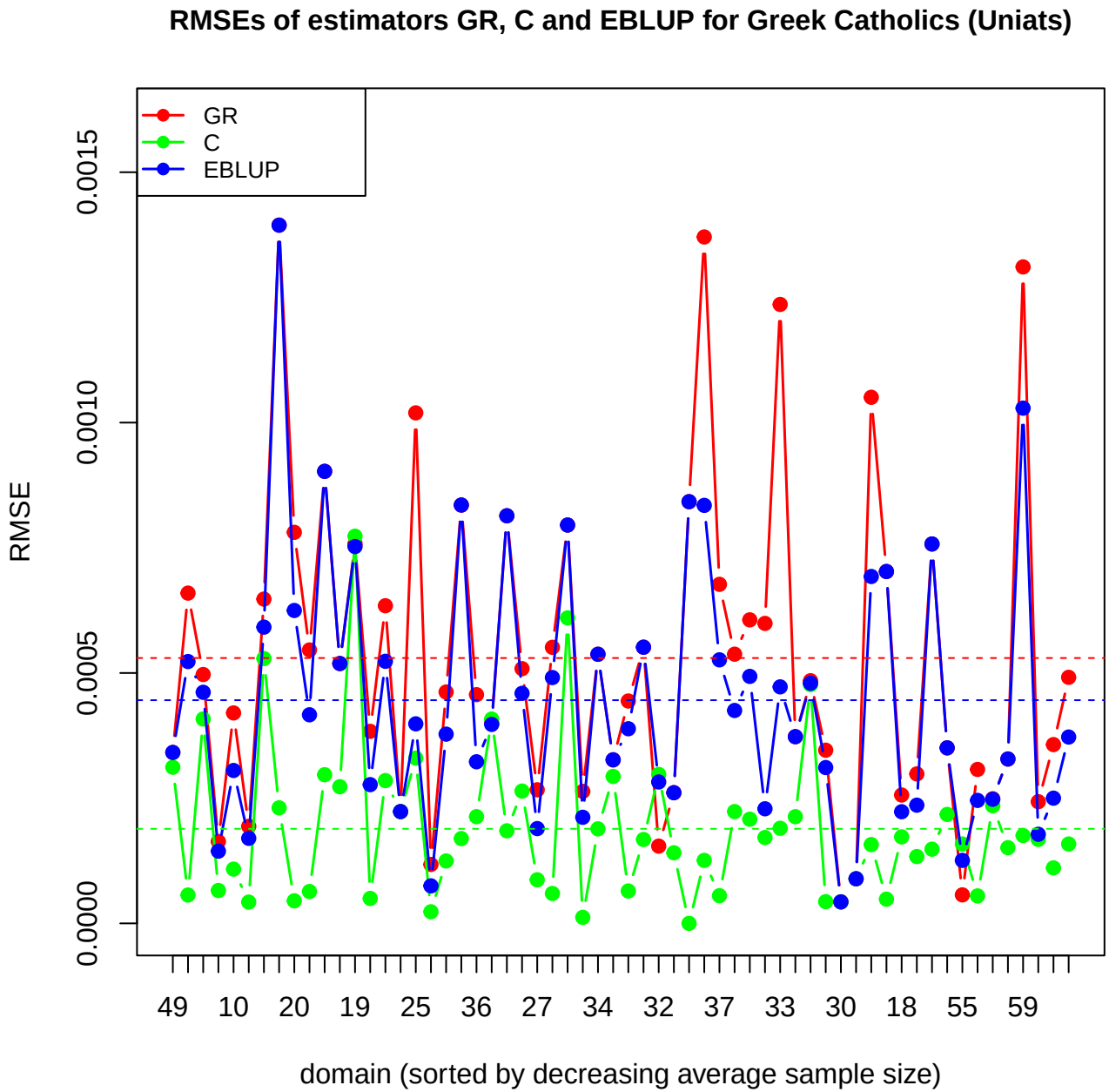
Small religion (iii). ABs of estimators



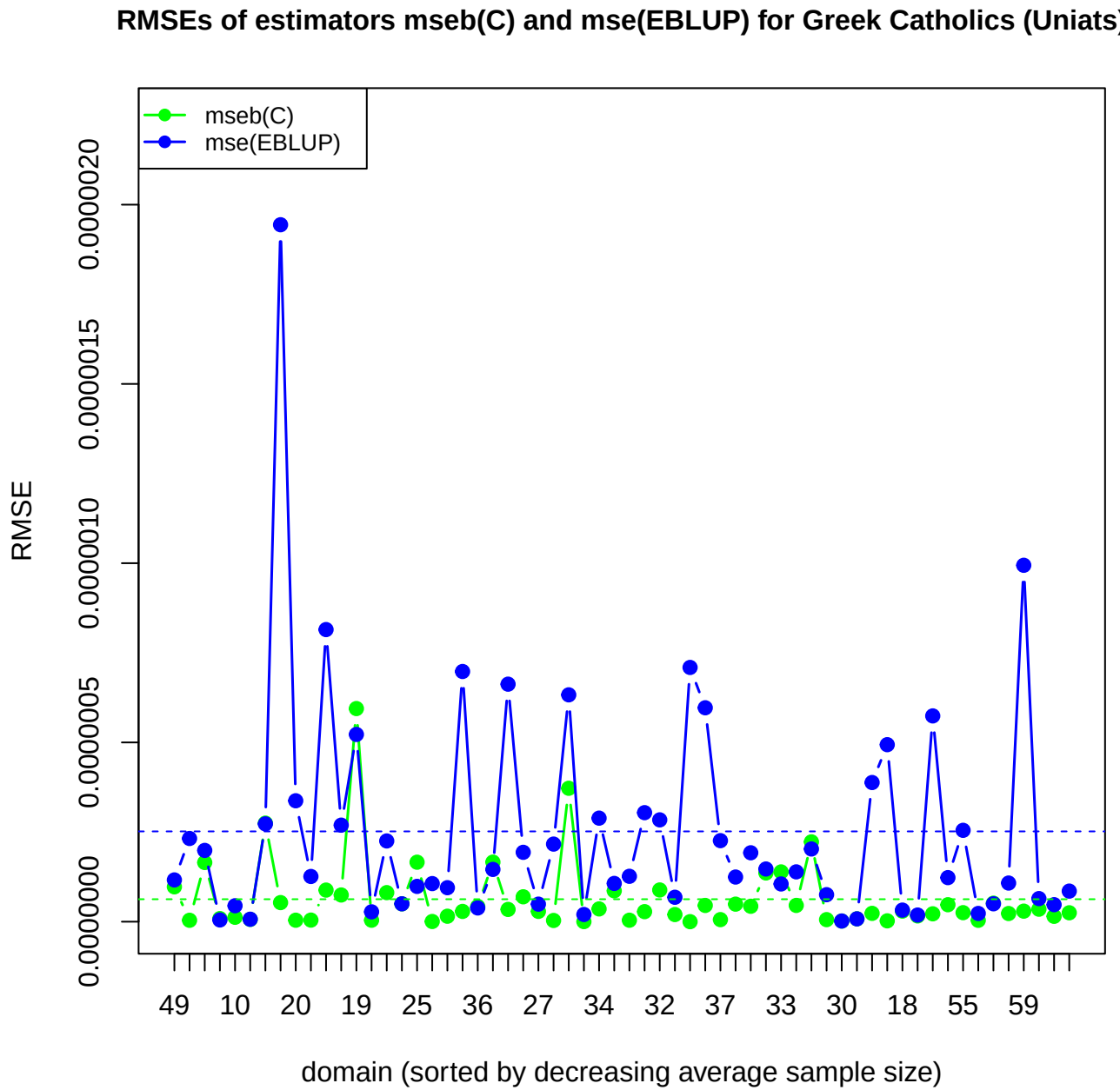
Small religion (iii). ABs of MSE estimators



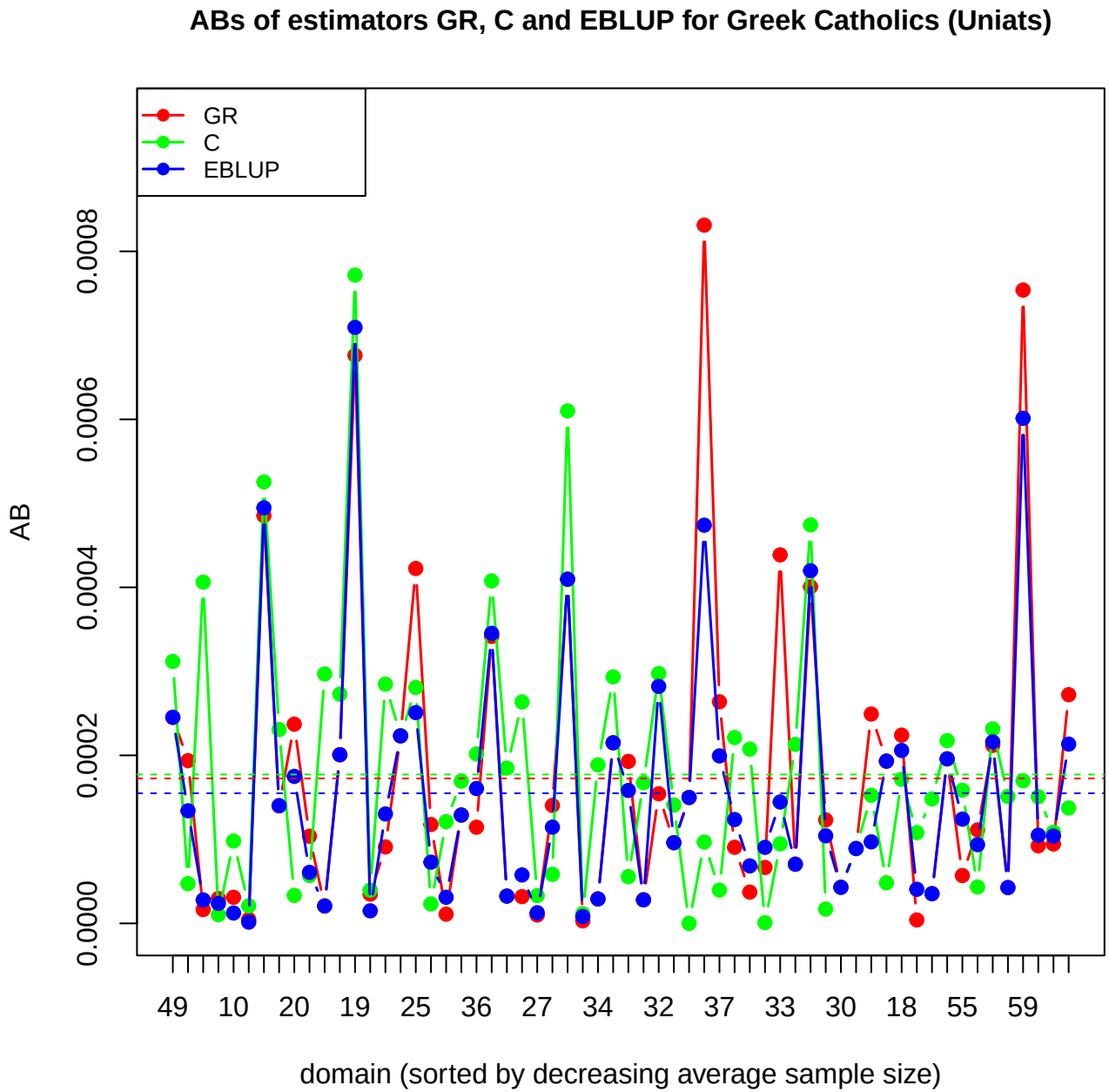
Very small religion (i). RMSEs of estimators



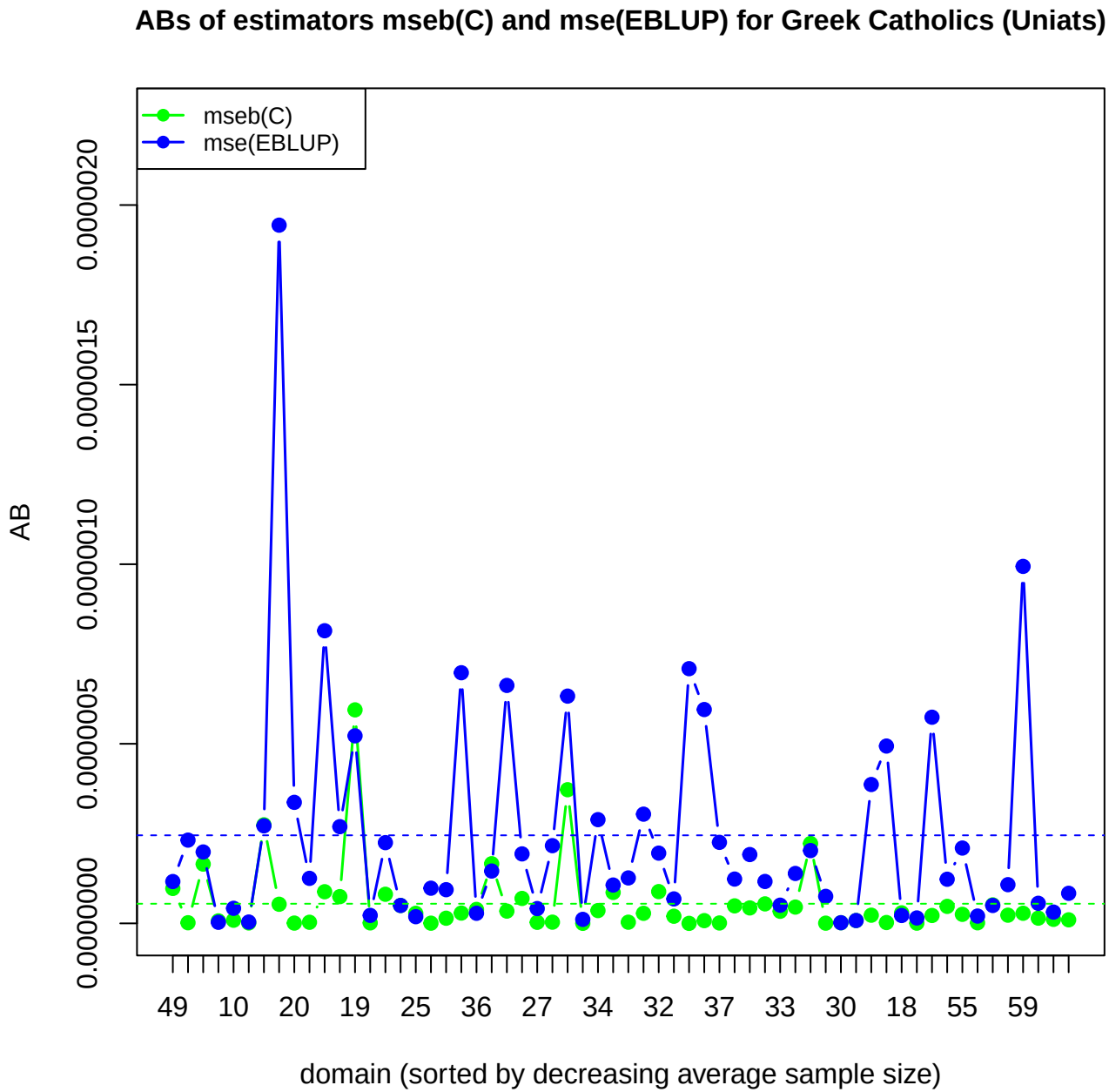
Very small religion (i). RMSEs of MSE estimators



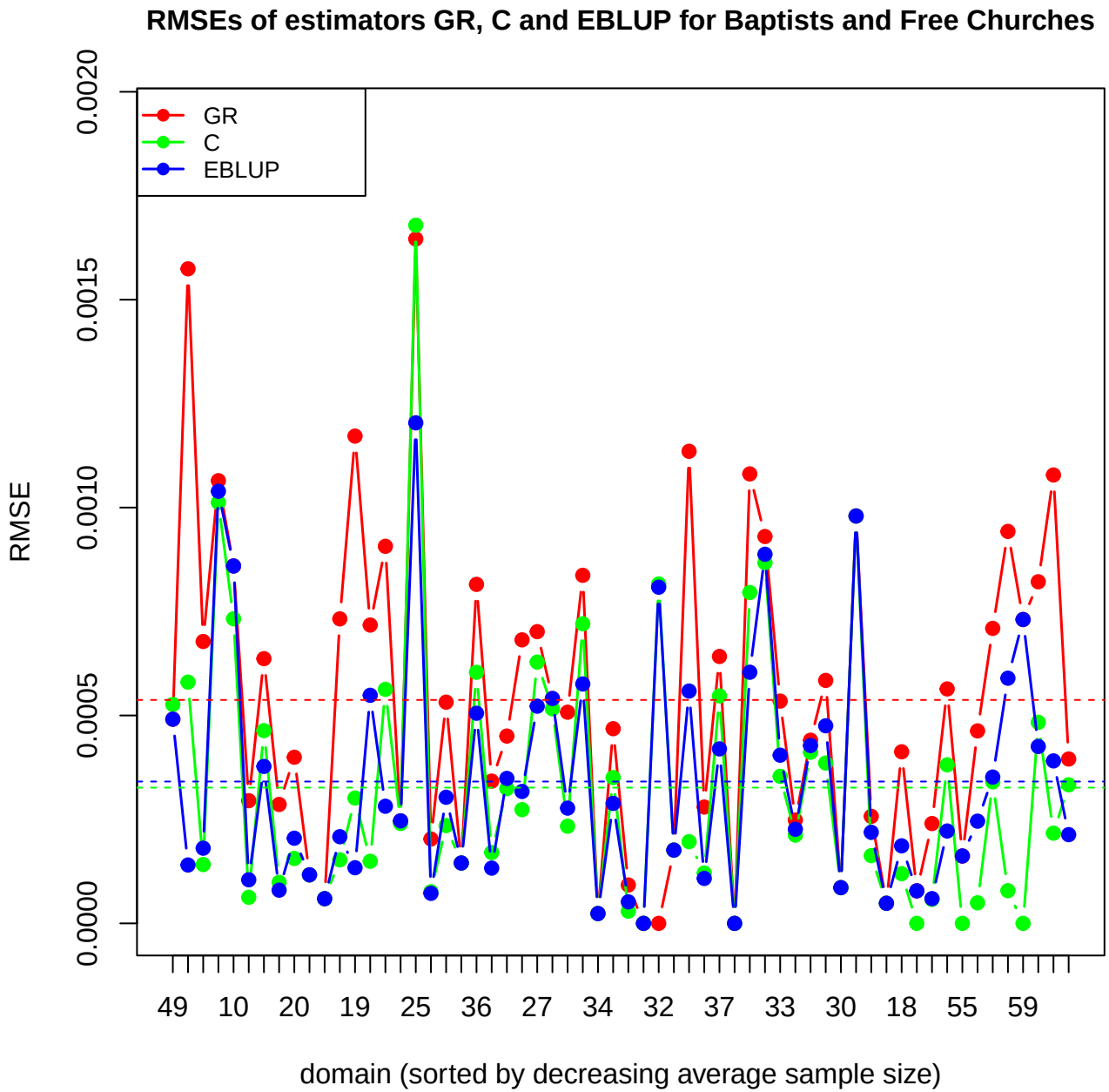
Very small religion (i). ABs of estimators



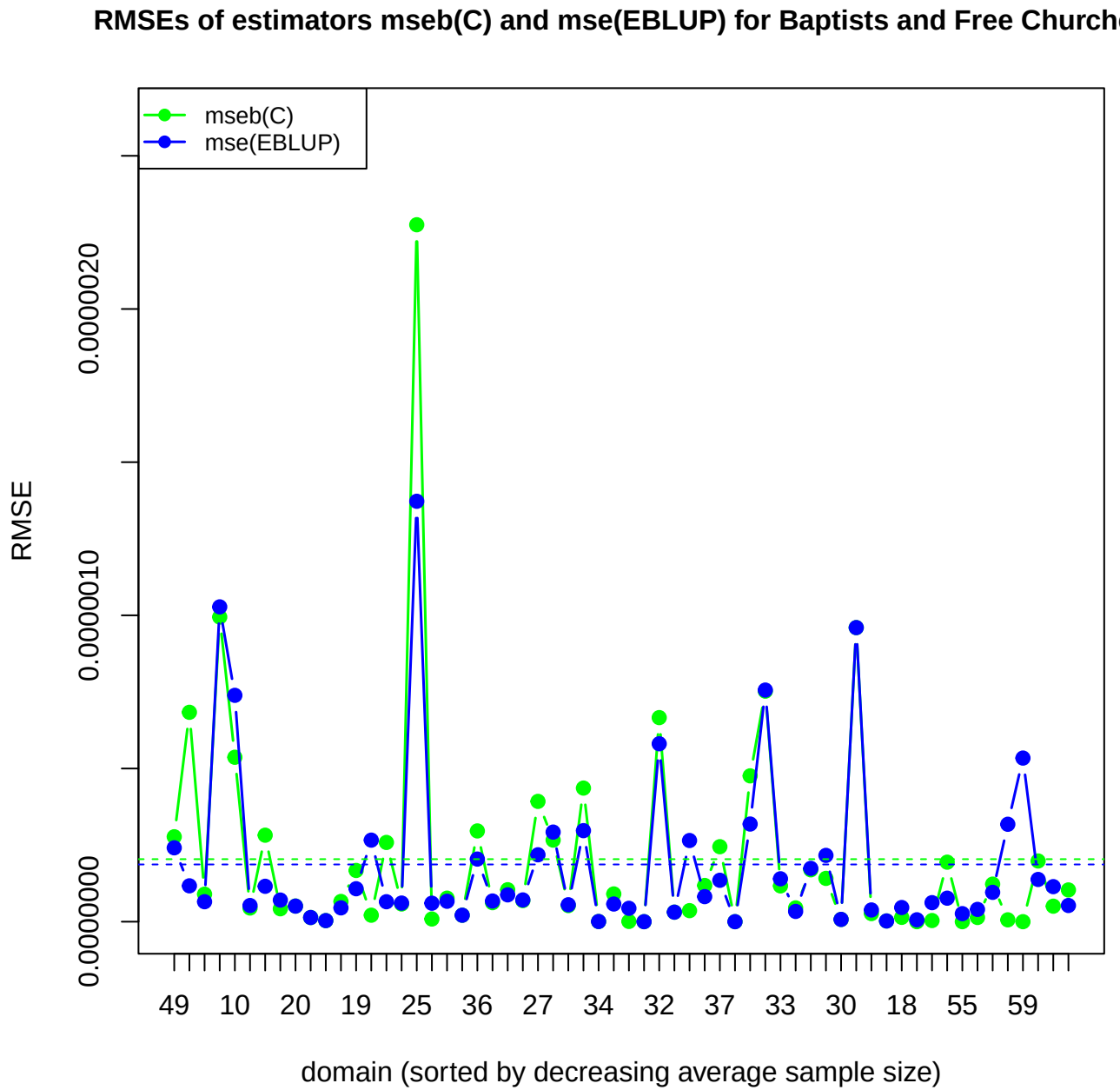
Very small religion (i). ABs of MSE estimators



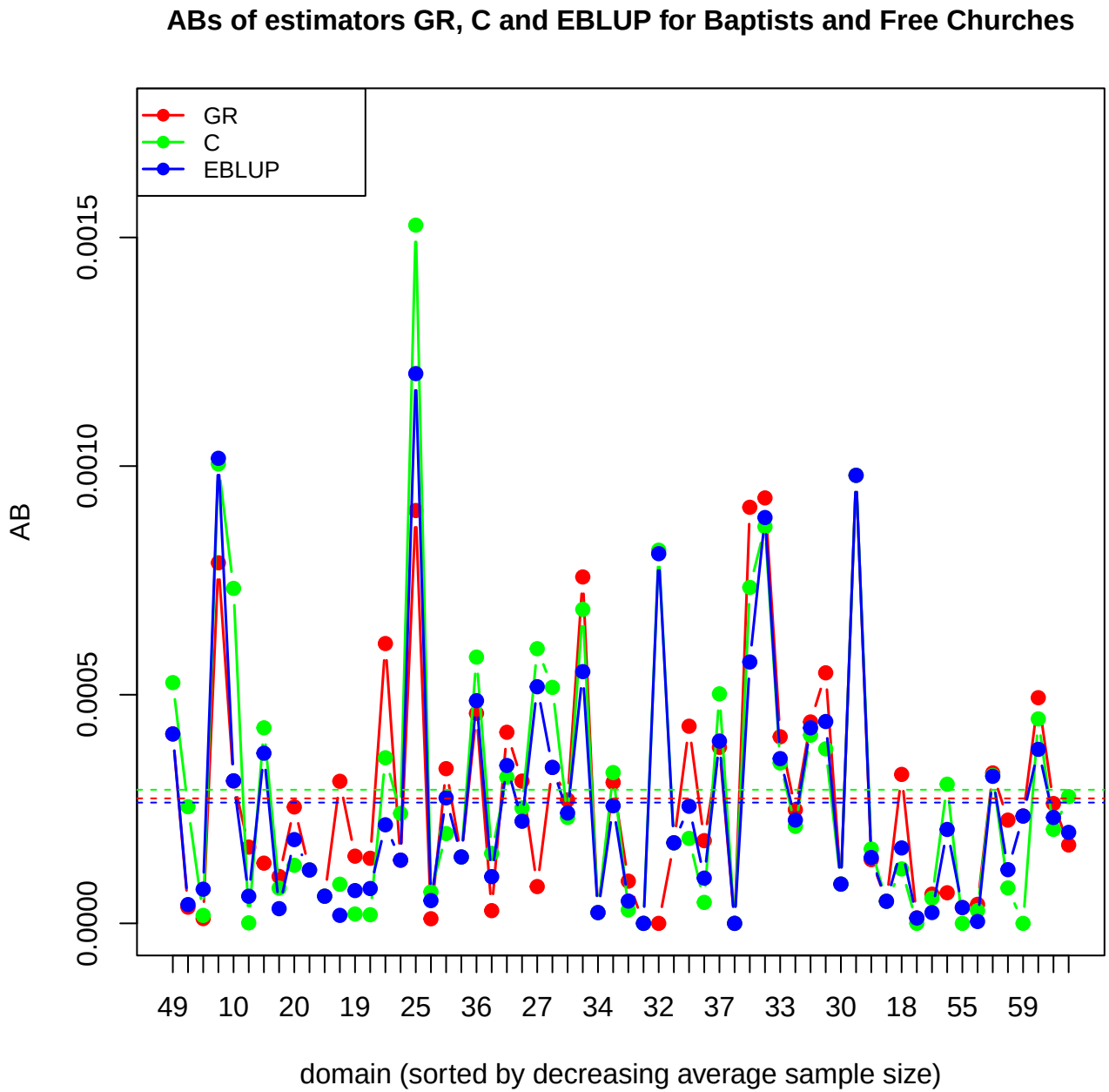
Very small religion (ii). RMSEs of estimators



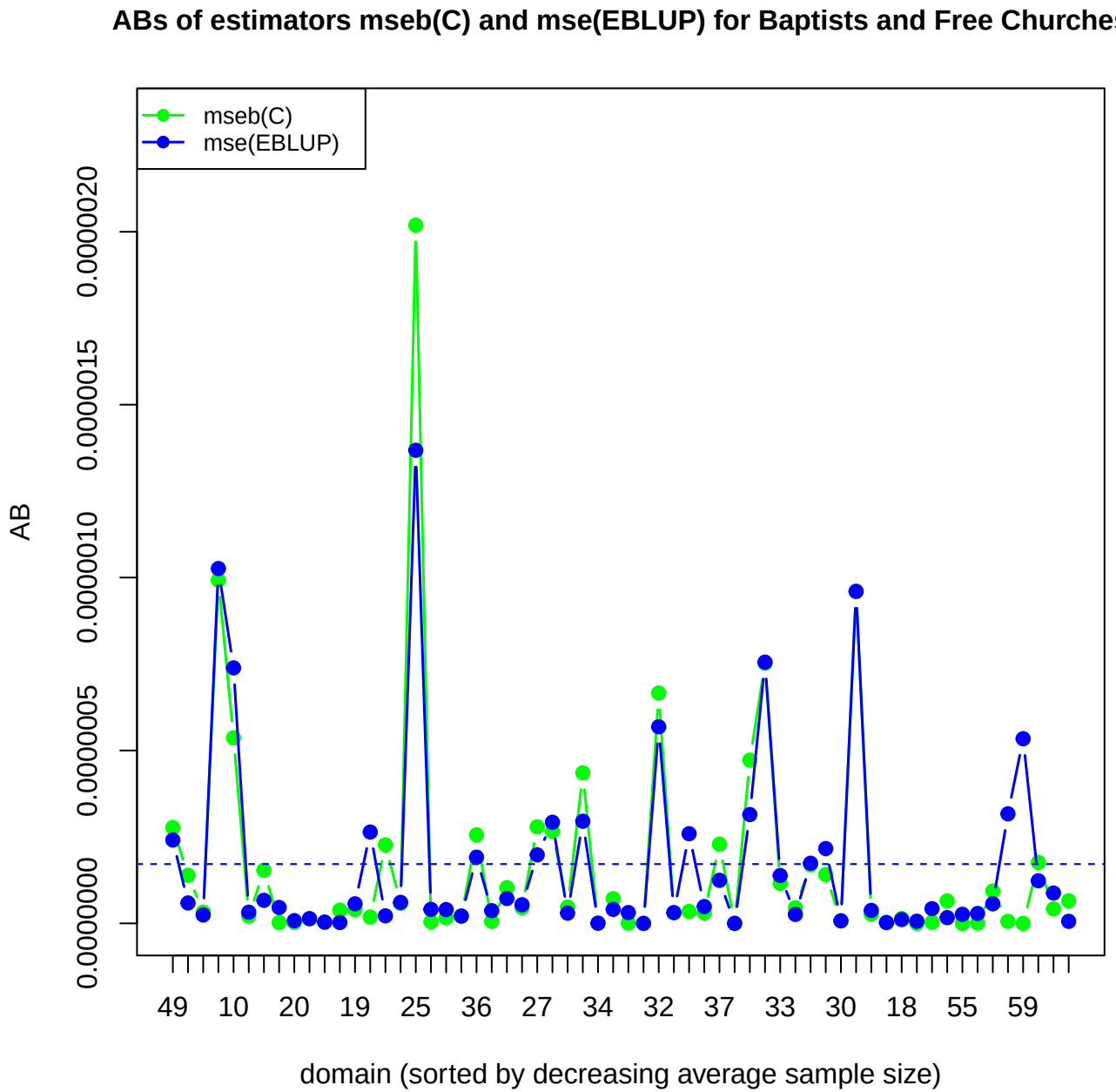
Very small religion (ii). RMSEs of MSE estimators



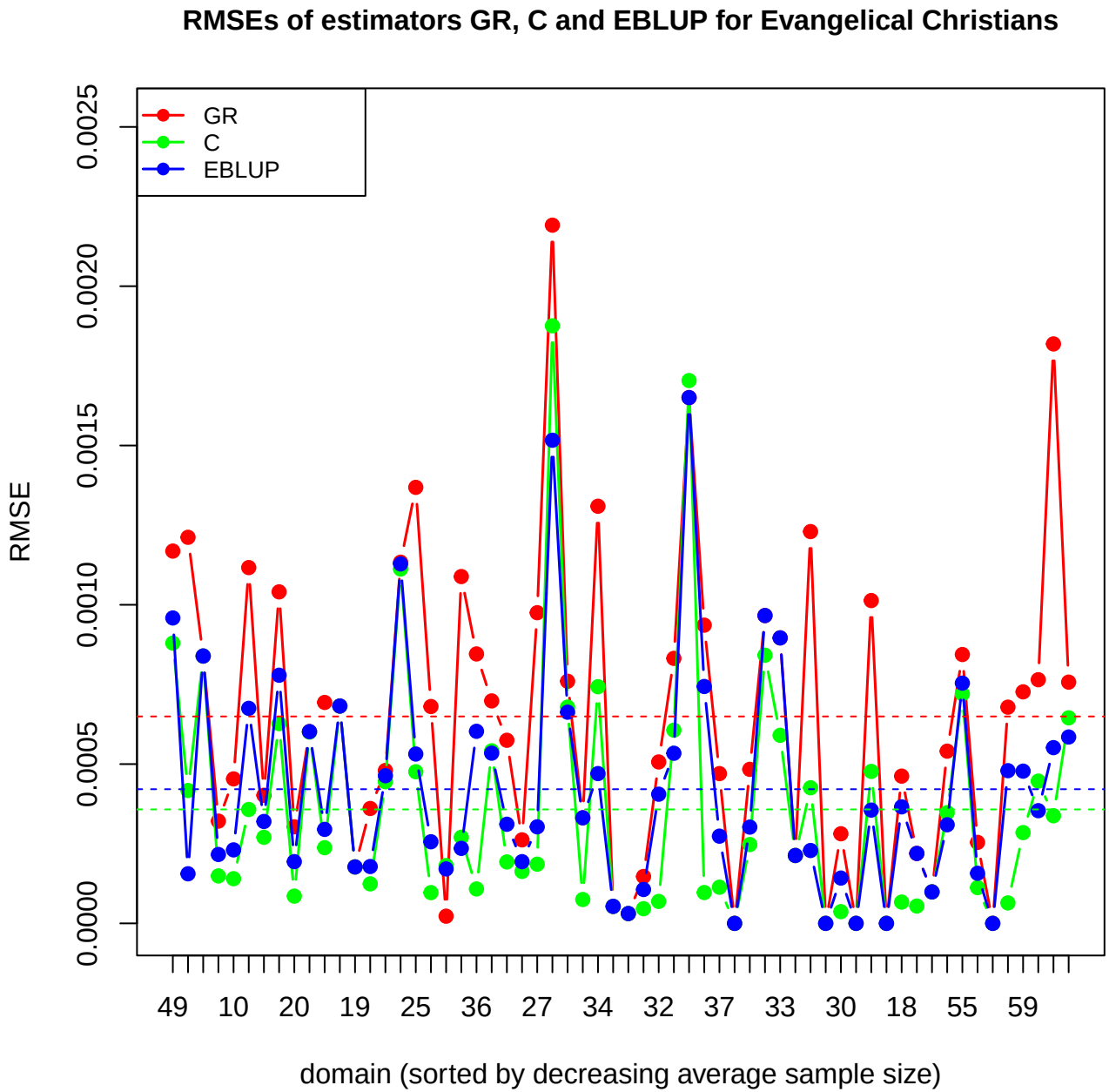
Very small religion (ii). ABs of estimators



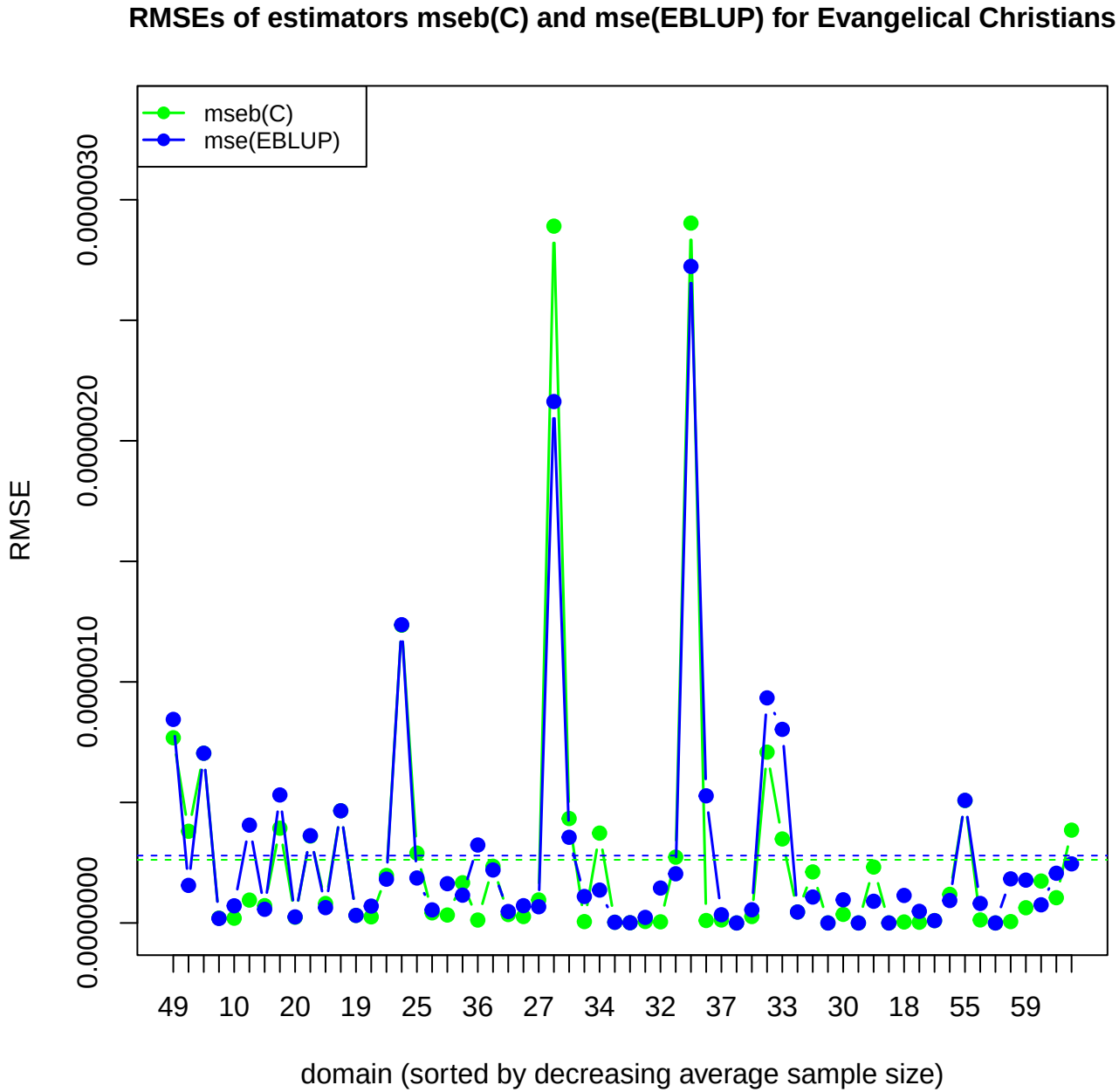
Very small religion (ii). ABs of MSE estimators



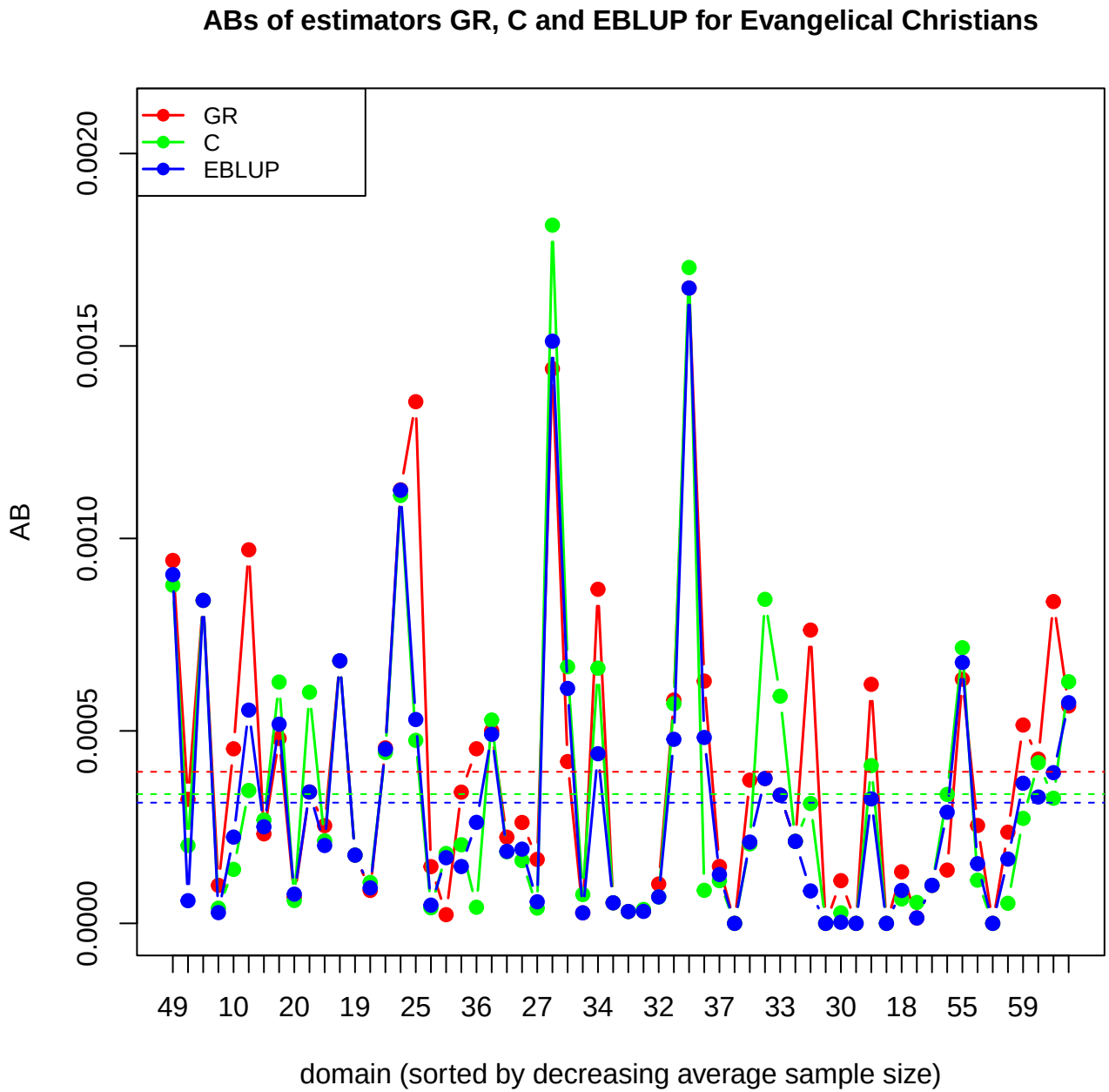
Very small religion (iii). RMSEs of estimators



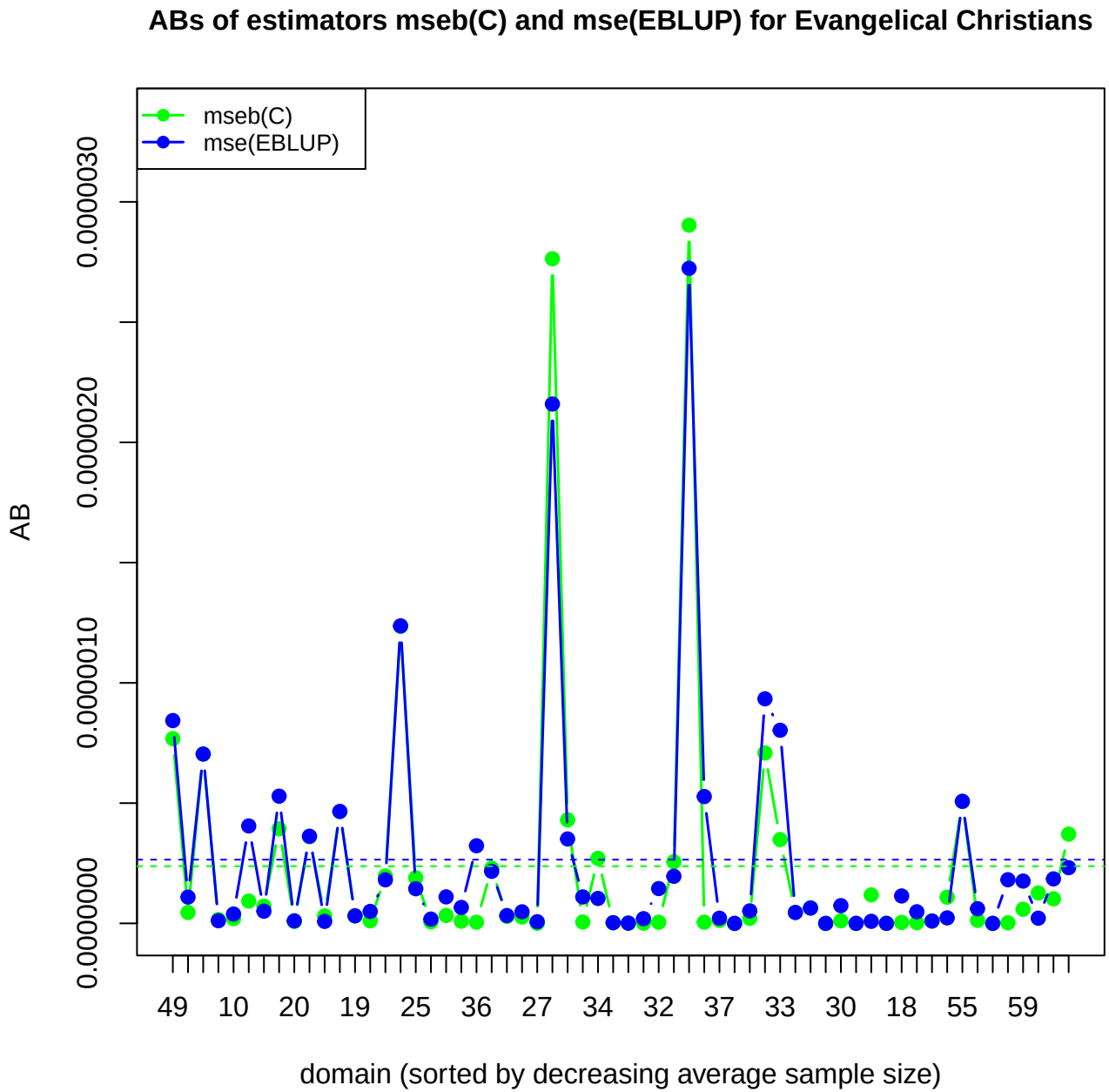
Very small religion (iii). RMSEs of MSE estimators



Very small religion (iii). ABs of estimators



Very small religion (iii). ABs of MSE estimators



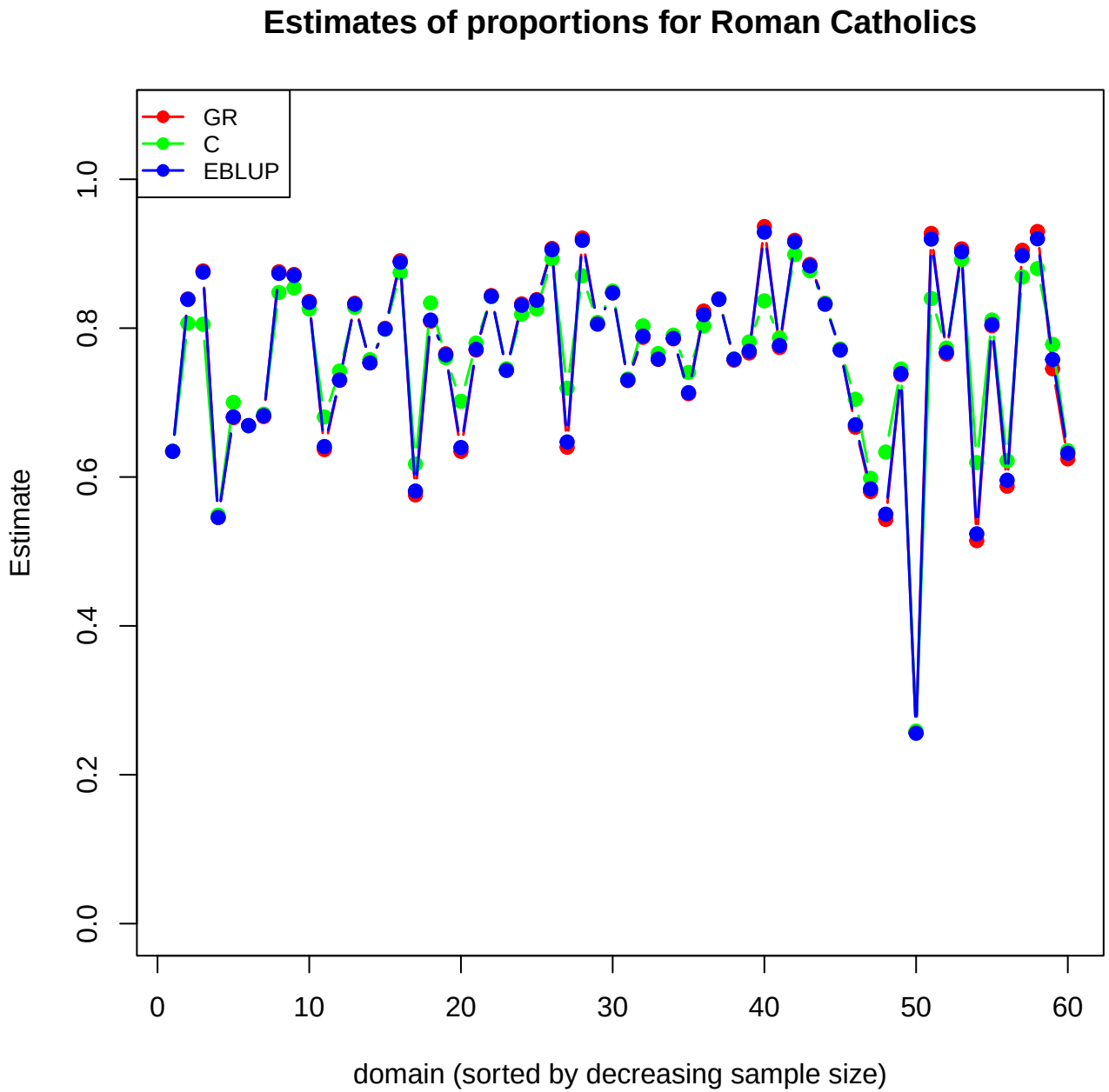
Shrinkage estimator vs EBLUP for Census 2021 data

The decision was based on the simulations using Census 2011 data – the design-based composite shrinkage estimator $\hat{\theta}_i^{\mathrm{C}}$ was applied to estimate all proportions in the Census 2021 survey. The accuracy of $\hat{\theta}_i^{\mathrm{C}}$ was estimated using the estimator $\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{C}})$.

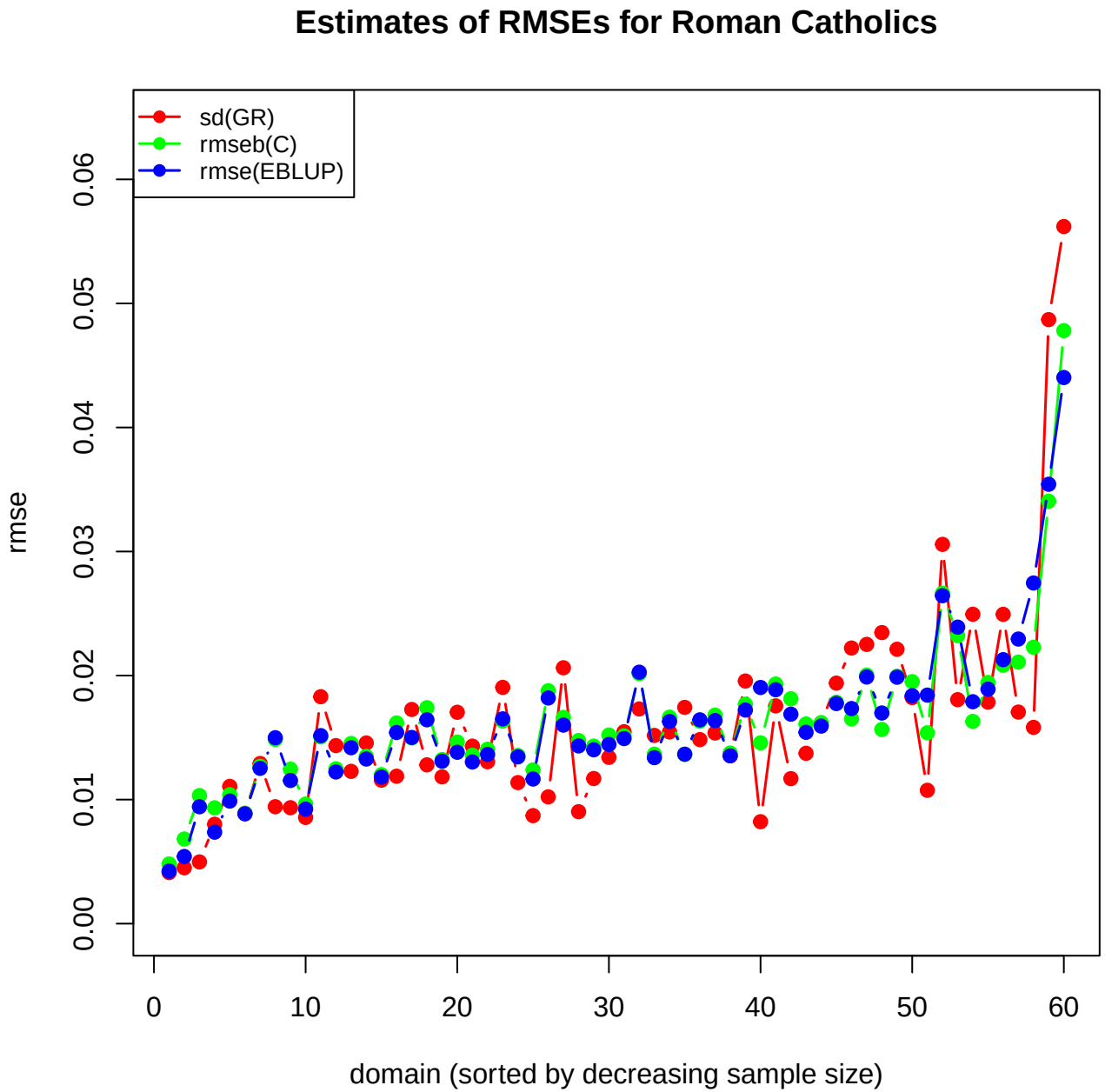
For the single sample of Census 2021 we compare:

- the estimates $\hat{\theta}_i^{\mathrm{GR}}$, $\hat{\theta}_i^{\mathrm{C}}$, and $\hat{\theta}_i^{\mathrm{EBLUP}}$ of θ_i;
- the square roots of the estimates $\hat{\psi}_i^{\mathrm{GR}}$, $\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{C}})$, and $\mathrm{mse}(\hat{\theta}_i^{\mathrm{EBLUP}})$ of $\mathrm{var}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{GR}})$, $\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{C}})$, and $\mathrm{MSE}(\hat{\theta}_i^{\mathrm{EBLUP}})$, respectively.

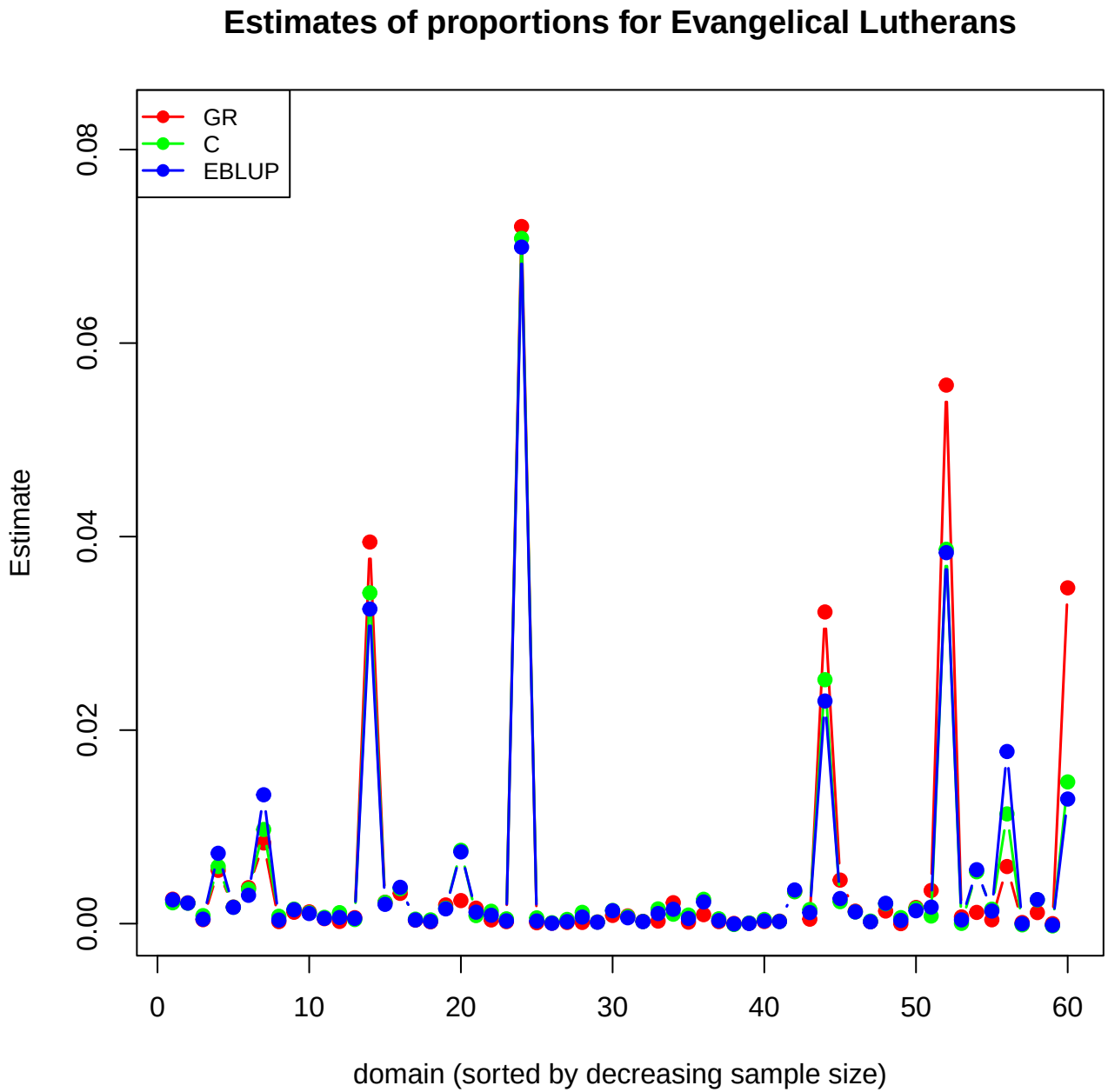
Small religion (i). Estimates of proportions



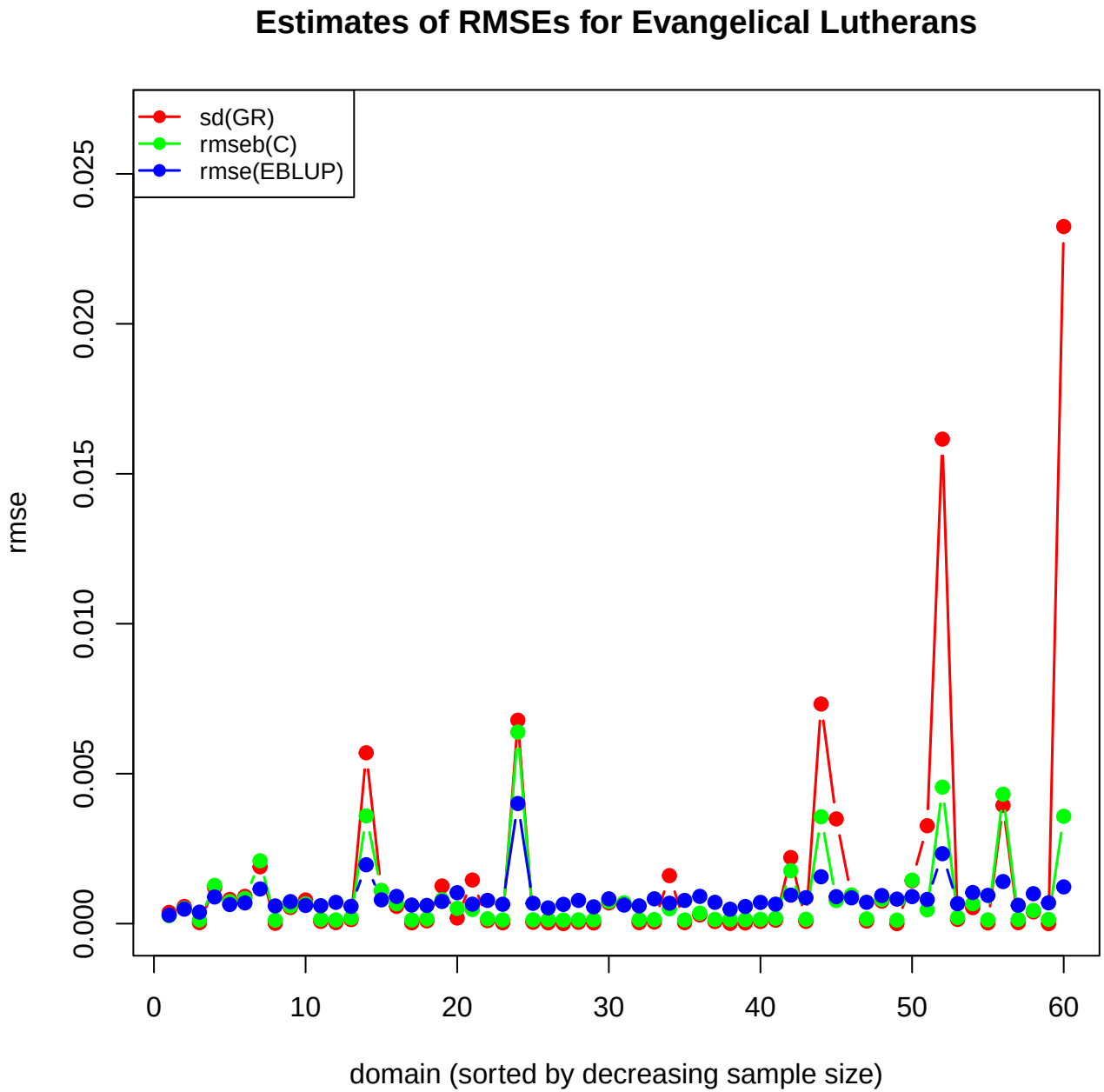
Small religion (i). Estimates of RMSEs



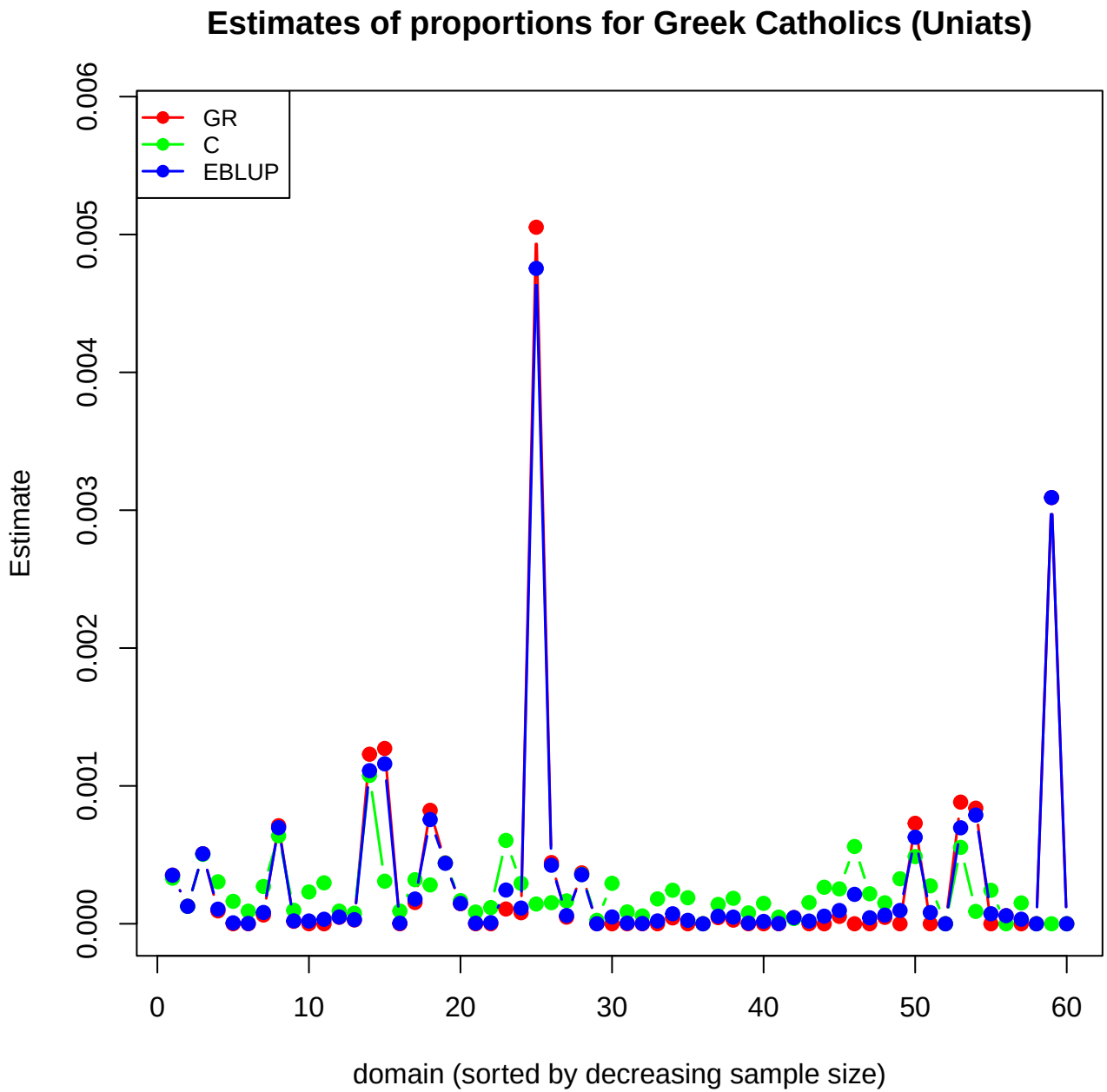
Small religion (ii). Estimates of proportions



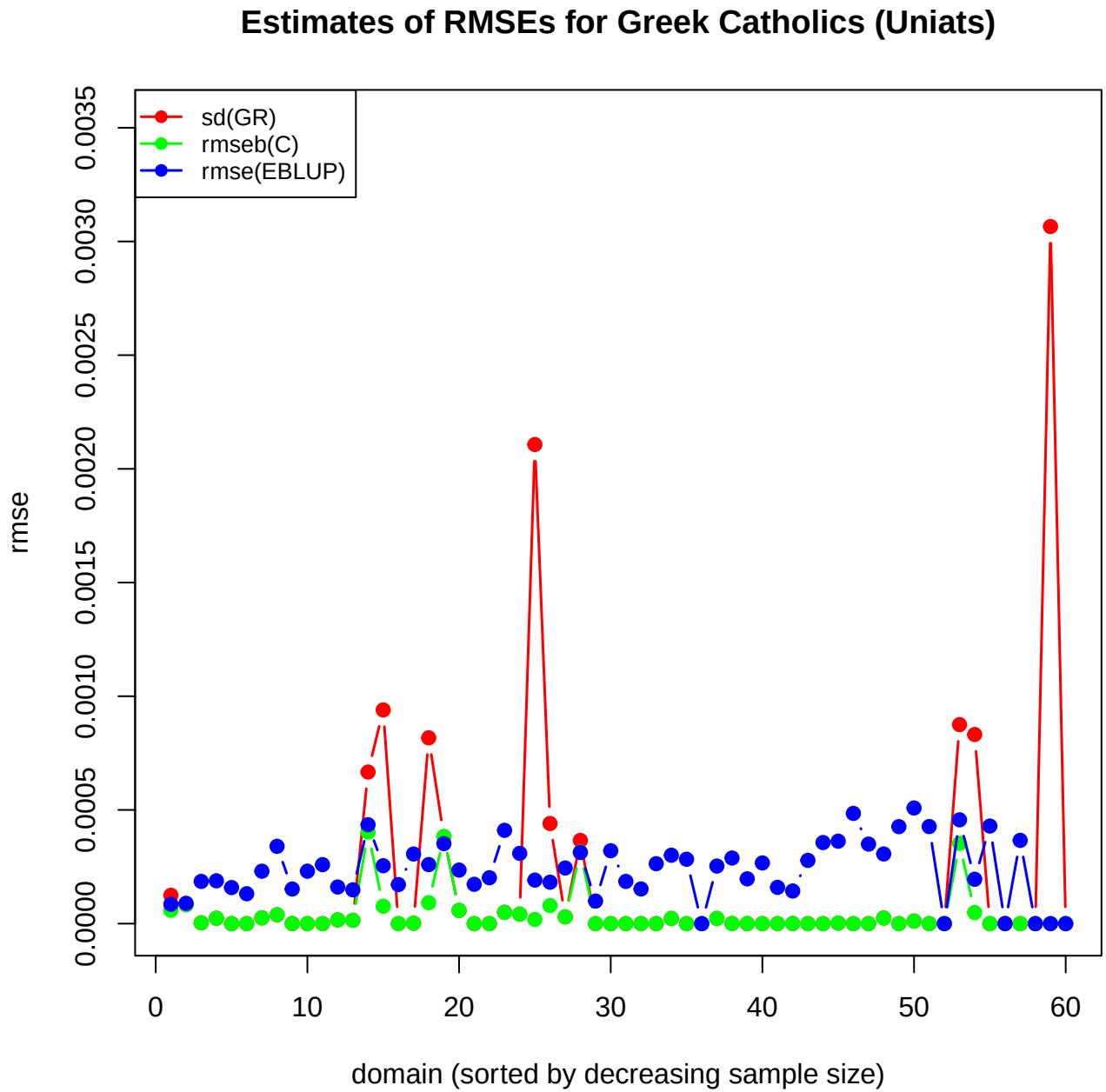
Small religion (ii). Estimates of RMSEs



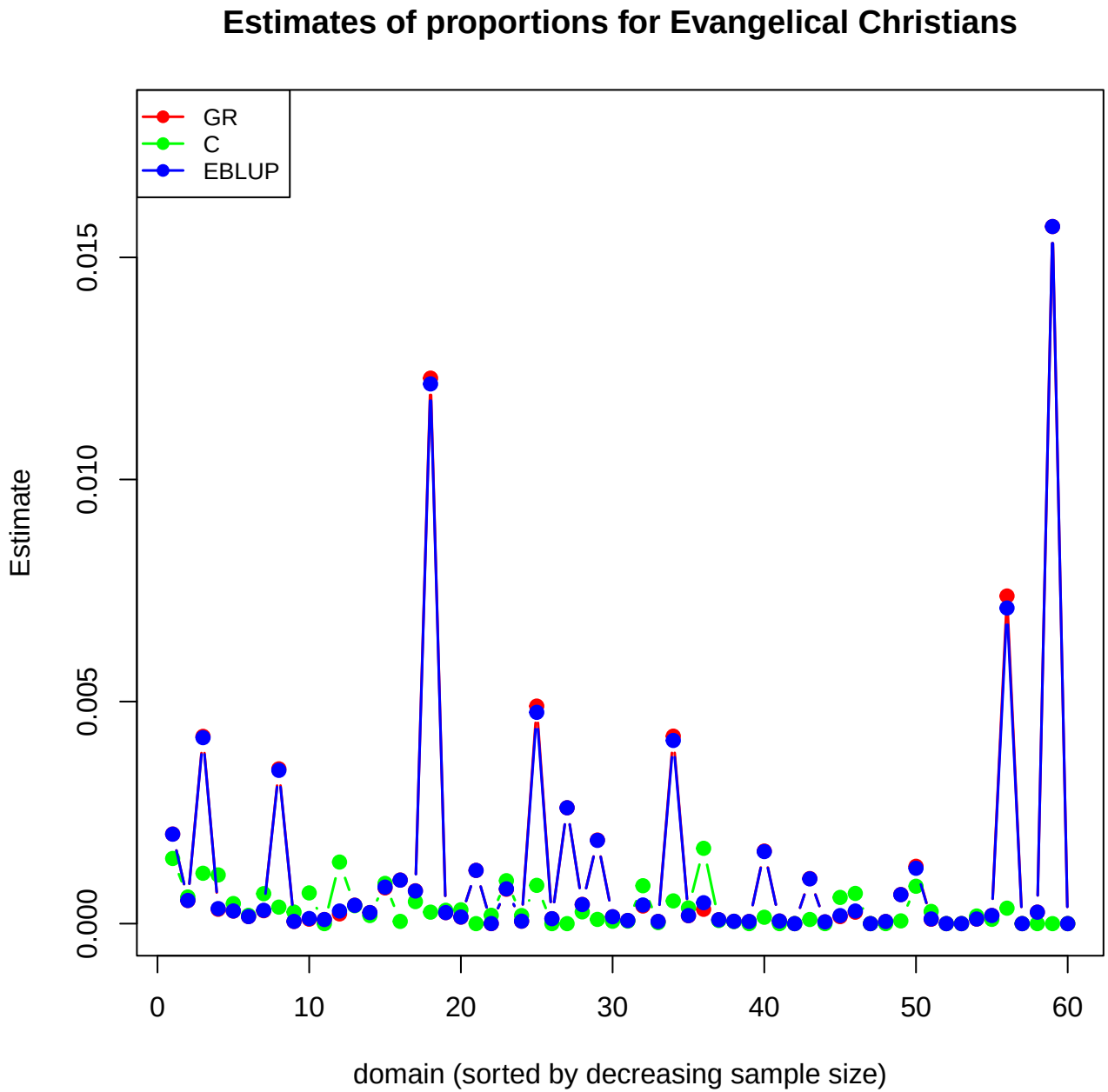
Very small religion (i). Estimates of proportions



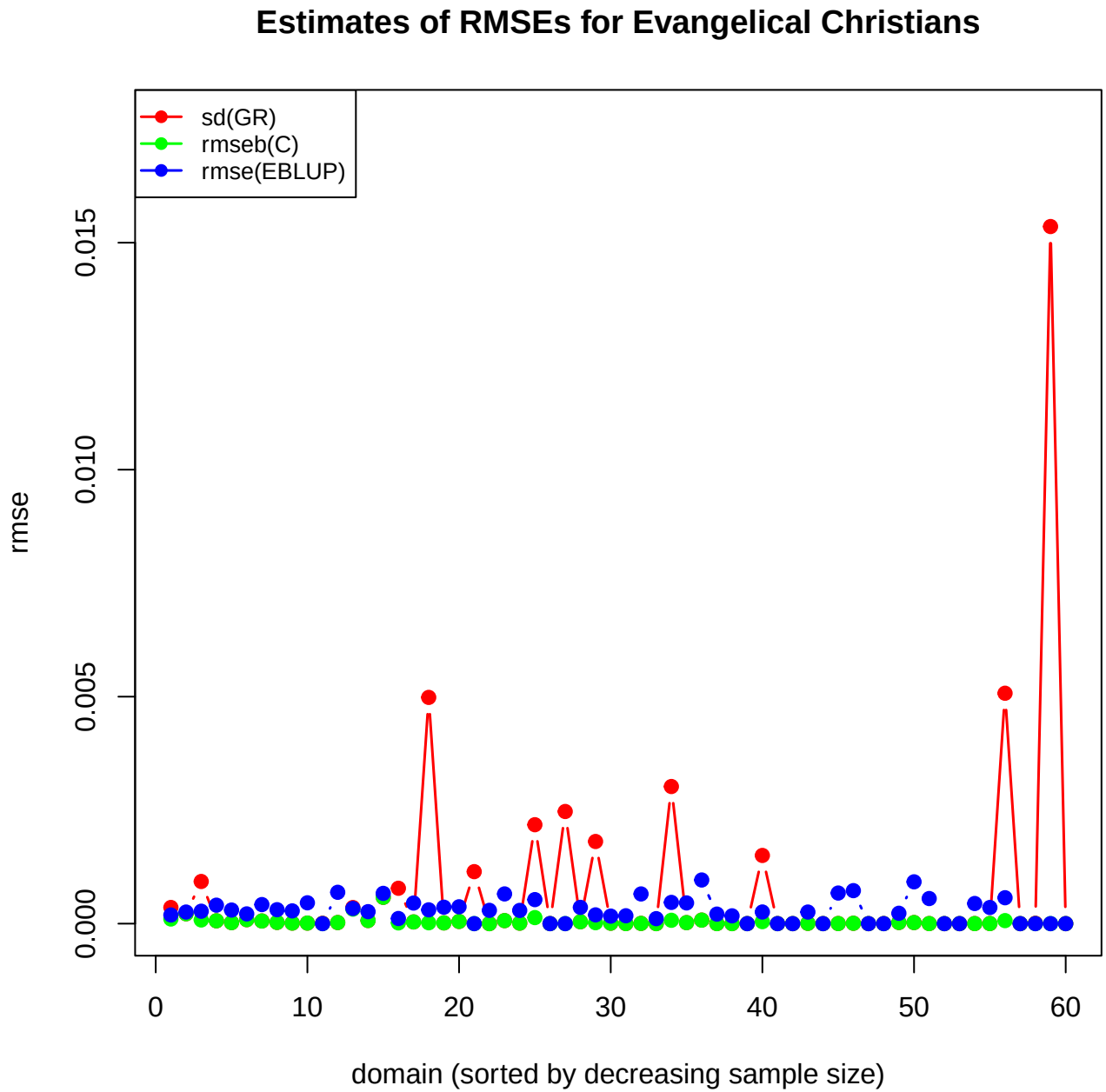
Very small religion (i). Estimates of RMSEs



Very small religion (iii). Estimates of proportions



Very small religion (iii). Estimates of RMSEs



Benchmarking

Let $\hat{\theta}_{ij}^{\mathrm{C}}$, $i = 1, \ldots, M$, be the estimates for $j = 1, \ldots, J$ categories ($J = 16$ for religions). Let $\hat{\theta}_j^{\mathrm{c}}$ be the final estimate of the whole population proportion θ_j for the jth category.

We require that

$$\frac{1}{N} \sum_{i=1}^{M} N_i \hat{\theta}_{ij}^{\mathrm{C}} = \hat{\theta}_j^{\mathrm{c}} \quad \text{for} \quad j = 1, \ldots, J$$

and

$$\sum_{j=1}^{J} \hat{\theta}_{ij}^{\mathrm{C}} = 1 \quad \text{for} \quad i = 1, \ldots, M.$$

The estimates $\hat{\theta}_{ij}^{\mathrm{C}}$, $i = 1, \ldots, M$, $j = 1, \ldots, J$, are benchmarked to satisfy the above conditions using the criterion of weighted least squares with the inverse MSE estimates $\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{C}})$ as the weights (Boonstra et al., 2008).

Summary

- The design-based composite shrinkage estimator and its MSE estimator are efficient compared to the model-based EBLUP and its MSE estimator for small and very small domain proportions, according to the simulation study.
- Applied design-based composite shrinkage estimation based on domain-level models is robust compared to small area estimators supported by unit-level auxiliary data.
- Domain-level data available from the previous full census are crucial for the efficiency of the estimators. That means more challenges in the next census. A possible solution is to collect a much larger non-probability sample.

References

Boonstra, H.J., van den Brakel, J.A., Buelens, B., Krieg, S., Smeets, M. (2008). Towards small area estimation at Statistics Netherlands. *Metron* 66:21–49.

Čiginas, A. (2021). Design-based composite estimation rediscovered. `arXiv:2108.05052 [stat.ME]`.

Čiginas, A. (2022). Design-based composite estimation of small proportions in small domains. `arXiv:2202.13085 [stat.ME]`.

Datta, G.S., Rao, J.N.K., Smith, D.D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika* 92:183–196.

Dick, P. (1995). Modelling net undercoverage in the 1991 Canadian census. *Survey Methodology* 21:45–54.

Fay, R.E., Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association* 74:269–277.

Gonzalez, M.E., Waksberg, J. (1973). Estimation of the error of synthetic estimates. Paper presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria.

Hidiroglou, M.A., Beaumont, J.-F., Yung, W. (2019). Development of a small area estimation system at Statistics Canada. *Survey Methodology* 45:101–126.

Kish, L. (1995). Methods for design effects. *Journal of Official Statistics* 11:55–77.

Rao, J.N.K., Molina, I. (2015). *Small Area Estimation.* 2nd edition, John Wiley & Sons, Inc., Hoboken, New Jersey.

Rao, J.N.K., Wu, C.F.J., Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology* 18:209–217.

Wolter, K.M. (2007). *Introduction to Variance Estimation.* 2nd edition, Springer-Verlag, New York.