The coverage estimation strategy for small communal establishment of the 2021 Census of England & Wales

Abu Hossain

Office for National Statistics
Newport, South Wales, UK

August 2, 2021

1 Key Messages

1.1 Purpose

The purpose of this paper is to propose a method to estimate the net coverage error for persons within small Communal Establishments (CEs). The proposed method is similar to the method proposed for the 2021 census person coverage estimation.

1.2 Recommendation

We recommend the use of logistic regression (LR) or mixed effect logistic regression (MELR) based methods for estimating persons within the small CEs as 2021 small CE coverage estimation strategy.

2 Executive Summary

This paper proposes the use of logistic regression to model census response measured by CCS in the estimation of net coverage error of persons in communal establishments. In the 2011 census, post stratification was used to estimate the coverage error for small communal establishments. The proposed Dual System Estimator based on logistic regression estimates the size of the closed population (e.g. census) in the presence of heterogeneous capture probabilities using census and the coverage survey as capture-recapture data. We expect to use DSE based on mixed effect logistic regression, however a comparison between LR and MELR will be done once we have the data. We will also maintain the post-stratification method as an option.

3 Background and Introduction

3.1 Overview

In the 2011 census communal establishments (CEs) were defined as establishments with 10 or more bed spaces, which provide managed residential accommodation. However, in the 2021 census, hotels, guest houses, B& Bs, inns and pubs with residential establishments with 7 or more bed space are also defined as communal establishments. Communal establishment residents represented 1.7 % (937,000) of all usual residents in England & Wales (56.1million) in 2011. Communal establishments fall into 2 categories:

3.2 Small CEs (7-49 bed spaces)

In 2011 communal establishments with less than 100 bed spaces were defined as small communal establishments. In 2021 small CE's will be defined as communal establishments with less than 50 bed spaces. These are enumerated using a special questionnaire in both Census and CCS. The focus of this paper is to propose an estimation strategy for the small CEs.

3.3 Large CEs with 50 or more bed spaces

Large CEs are not included in the CCS. Field and administrative data will be used to assess coverage within these large establishments. As these establishment groups are excluded from the CCS the proposed estimation strategy will not be applied.

Figure 1 shows the response rate for small CEs in 2011 census by establishment nature. The lowest response rate (i.e. 88 %) was recorded for sheltered accommodation while 100 % response rate was achieved for caravan, campsites and marinas, embassies and consulates, homes for terminally ill, mission and night shelters.

4 2011 small CE coverage estimation Strategy

In 2011 small CEs were included in the CCS, and therefore matched data were available for those CEs that were in the sample areas. The CCS sample design did not take CEs into account, and therefore there was no control over the size of the small CE sample. The sample did include a range of different types of small communal establishments.

A dual system and ratio estimation method was used to estimate coverage within the CEs. The method was estimated by higher geographic region, type and collapsed age-sex group to have sufficient sample to support such estimates. The CCS sample contained 520 small CEs and was sufficient to compute a set of regional level DSEs by collapsed age-sex groups (but not by type of CE) to measure coverage within the CEs. The DSE was then used to derive the coverage ratio. This ratio was then applied to the census counts for the people within small CEs, to obtain estimates by local authority, type and age-sex group, by using the adjustment factors and assuming that they apply at the lower level (a synthetic assumption). The estimate of people in small CEs in each

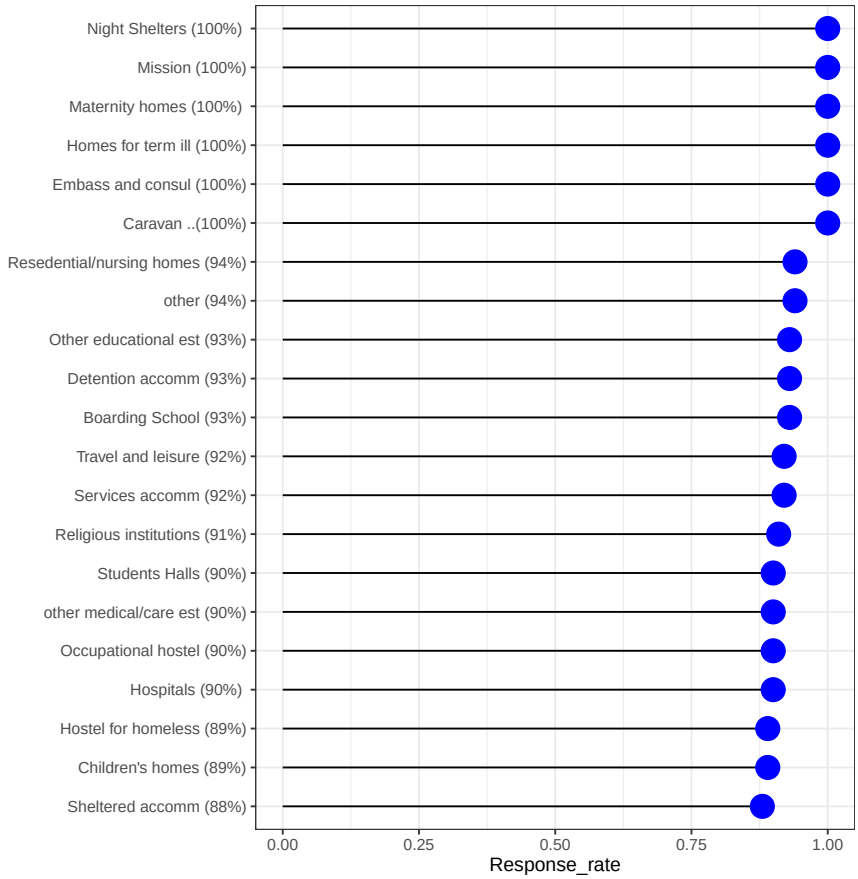

Figure 1: 2011 small CE response by establishment nature (ONS (2012))

local authority by age-sex $\hat{T}_{lea}$(where l is the local authority, e is the establishment type and a is the age-sex-group) is obtained by:

$$\hat{T}_{lea} = \frac{\sum_{sample} \hat{y}_{rgc}}{\sum_{sample} \hat{X}_{rgc}} X_{lea} \tag{1}$$

Where $\hat{X}_{rgc}$ is the census count for the region r grouped establishment type g and collapsed age-sex-group c, X_{lea} is the census total for the local authority by establishment type and age-sex and $\hat{y}_{rgc}$ is the DSE estimated by region r grouped establishment type g and collapsed age-sex-group c (Abbott, 2011). This resulted in the following collapsed age-sex groups:

Outer London

- M and F aged 0 to 29
- M and F aged 30 to 59
- M aged 60 +
- F aged 60 to 84
- F aged 85 +

The South East and Wales

- M and F aged 0 to 59
- M aged 60+
- F aged 60 to 84
- F aged 85+

The North West and South West

- M and F aged 0 to 59
- M aged 60+
- F aged 60+

The remaining regions

- M and F aged 0 to 59
- M and F aged 60+

As can be seen, because of the relatively small number of people within CEs, these strata have been subject to quite a lot of collapsing. Most regions only differentiate between people under 60 and those 60+, with no reference to other age groups or male and female. This reduces the ability to differentiate between the response rates of the different groups.

5 Small CE Coverage estimation for 2021 Census

Like 2011, the 2021 CCS sample design did not take CEs into account. To estimate persons within small CEs, a DSE with logistic regression is proposed - similar to the approach proposed in 2021 Census household coverage estimation ((Račinskij, 2018), US Census Bureau, 2008; US Census Bureau, 2012). Earlier work of Huggins (1989); Alho (1990); Huggins (1991); Alho *et al.* (1993) suggested logistic regression to estimate the size of the closed population based on capture and single recapture. The capture probability p_{ij} is the conditional probability of capturing the i_{th} person in the j_{th} sample given the capture history of samples $1, 2$ and can be written as

$$\texttt{logit}(p_{ij}) = \log\left[\frac{p_{ij}}{1-p_{ij}}\right] = \mu_i + \sigma_j + \gamma V_{ij} \tag{2}$$

Where the parameters $\mu_1, \mu_2, \ldots\ldots, \mu_N$, $\sigma_1, \sigma_2, \ldots\ldots, \sigma_t$, and γ denotes respectively the heterogeneity, time effects and behavioural response effect and V_{ij} is the time dependent variable used to denote the prior capture history of individual i for sample j.

Pollock*et al.* (1984) were the first to incorporate covariates in the logistic model for analysis. If the covariates can be used to account for the heterogeneity, then the logistic model 2 becomes

$$\texttt{logit}(p_{ij}) = \log\left[\frac{p_{ij}}{1-p_{ij}}\right] = \mu_i + \sigma_j + \gamma V_{ij} + \beta^{'} x_i \tag{3}$$

Where $x_i^k = (x_{i1}, x_{i2}, \ldots.., x_{ik})$ individual covariates for i_{th} person and $\beta^{'} = (\beta_1, \beta_2, \ldots.., \beta_k)$ denotes the effect of the covariates. The above model 3 can be extended in more general

form (Huggins, 1991). Since the explanatory variables for the uncaptured individuals are not available, the parameter estimates are obtained conditional on the captured individuals. The size of the population is then estimated by the Horvitz-Thomson estimator.

We are aiming to estimate the total population size of a domain of interest. This domain comprises certain individual characteristics within CE as well as some geography attribute: say, an age-sex group a in an area L (usually a local authority). We use the vector of covariates x (which includes a, some other variables that do not constitute the domain, interactions of a with other variables) and optionally an effect of L in estimation.

Let us say our domain of interest is τ_L, where τ is a covariate or combination of covariates (say, age-sex, establishment nature etc.); L is local authority. The domain of interest τ_L can be estimated using the following mixed effects logistic regression based estimator similar to the approach proposed for the 2021 census coverage ((Račinskij, 2018), (Račinskij , 2019),Racinskij 2021):

$$p_{ij} = \left[\frac{1}{1 + \exp\left(- \left[\mathbf{y}_i^T \hat{\beta} + \mathbf{z}_i^T \hat{\kappa} + \hat{\epsilon}_L \right] \right)} \right] \tag{4}$$

Where: the probability p_{ij} that an individual i is captured in census j; $\mathbf{y_i^T}$ – vector of main effects and interactions based on census / coverage survey covariates (such as age-sex, establishment nature, ethnicity etc.); $\mathbf{z_i^T}$ – vector of main effects and interactions based on design variables and field management information (hard-to-count-index, observed census return rate at the local super output area; $\hat{\epsilon}_L$ – random local authority effect.

Estimated census non-response weights $\hat{p}_{ij}^{-1}$ (reciprocals of estimated census response probabilities) for an individual in a communal establishment can then be applied to each census CE person observation with the corresponding characteristics [US Census Bureau, 2012]. Summing up all weighted census observations with the characteristic of interest will produce an estimated population size of units with the characteristic is shown in the following equation:

$$\hat{T}_{\tau_L} = \sum_{r \in \tau_L} \frac{1}{\hat{p}_{ij}} \tag{5}$$

There are some advantages expected from the proposed modelling approach. The flexibility of accommodating multiple covariates and their interactions may improve the estimation, and thus the precision of all parameter estimates is increased. Modelling covariates also provides a clear description of the source of the heterogeneity. The proposed mixed effect model for coverage estimation not only accommodates the heterogeneity due to individual characteristics (e.g. age-sex, establishment nature, ethnicity) but also uncertainty that measurable individual characteristics cannot explain Akand (2014). Another advantage is that the proposed estimation method does not rely on synthetic assumptions when estimating LA level population size. Račinskij (2018)

6 Simulation

A simulation study has been conducted to explore the performance of the proposed mixed effect logistic regression model alongside logistic regression and DSE. These simulations build on Brown *et al.* (2019) and further work of Račinskij (2018) , Račinskij (2019). A series of multilevel logistic regression models were fitted for England and Wales, based on the linked 2011 Census and CCS data. Two logistic models were fitted using matched data one for coverage of individuals within small CEs in the census measured by CCS response and one for coverage of individuals within small CEs in the CCS measured by the census response. The models were used to predict an individual coverage probability for person in each responding small communal establishment. Four hundred simulation were undertaken across England and Wales.

In the estimation system logistic regression, mixed effect logistic regression and DSE are used to model response rates of persons within CEs. In the logistic regression model covariates include age-sex, hard-to-count index, region, establishment nature and marital status. The mixed effect logistic regression for CE persons includes a random local authority effect and age-sex, hard-to-count index, establishment nature, region and marital status as main effects. This model is not the necessarily the optimal one and so careful model selection once census data is available may increase model performance.

7 Results & discussion

The analysis was divided into two parts. The first part analysed the relative bias (RB) and relative root mean square error (RRMSE) for the total population estimates of each LA, over 400 simulations, across the different estimators (DSE, LR and MELR). The second part compared both the RB and RRMSE obtained from DSE, LR and MELR by LA by age and sex groups and LA by communal establishment nature. To assess estimator performance percentage relative bias, $PRB = 100 \times (E[\hat{N}] - N) \div N$ where $E(\hat{N})$ is estimated by the $AVE(\hat{N})$ and empirical relative root mean square error, $RRMSE = \sqrt{\left[V\hat{A}R(\hat{N}) + bias^2\right]}/N$.

7.1 Results for the total population

This section assesses the performance of the estimator - as measured by the relative bias and the relative root mean square error for the total small CE population in England and Wales. Figure 2 shows the relative bias at LA level for three different models. It can be observed that the proposed MELR and LR have relatively little bias in most of the local authorities, although DSE has the lowest bias in some LAs. The bias in the DSE may come through the synthetic application of regional estimated ratio to LAs.

When looking at the relative root mean square error for the local authority estimates (Figure 3) variability is relatively lower in the case of LR/MELR

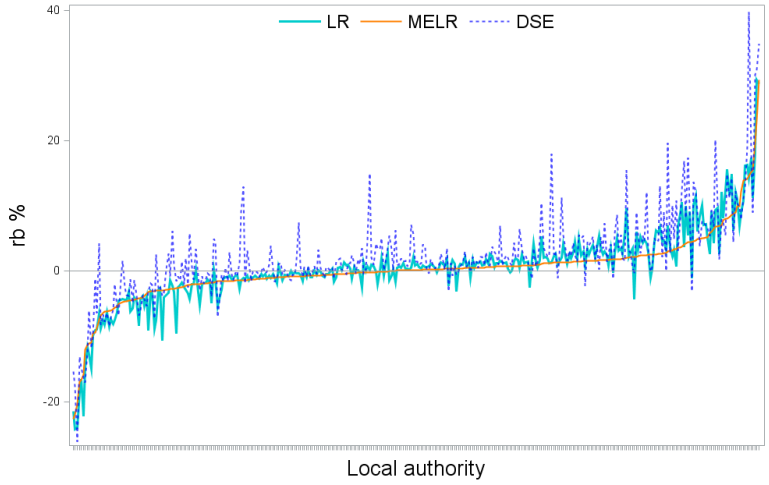

Figure 2: Relative bias for DSE, LR and MELR by local authority

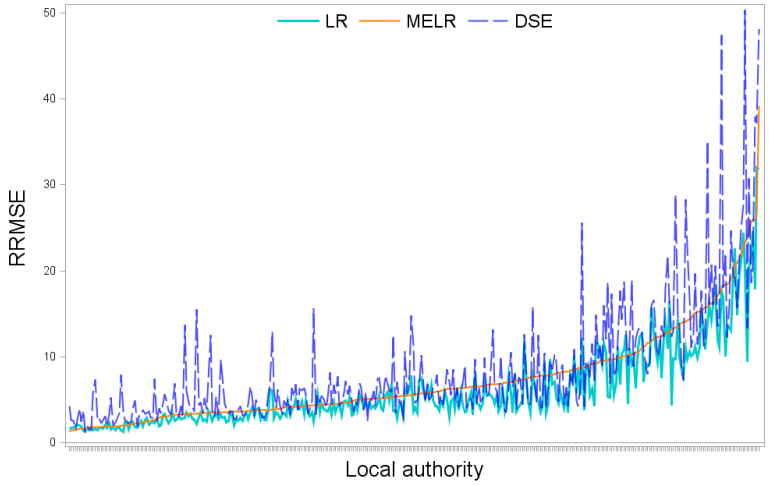

Figure 3: RRMSE for DSE, LR and MELR by local authority

7.2 Results by Age-Sex

This section compares the relative bias and the relative root mean square error of the census estimates obtained from DSE, LR and MELR by LA, by age-sex groups. In terms of relative bias for the age-sex groups in Figure 4 the proposed model performs at least as well as DSE. In age group 1 (Males and Females aged $0-14$) average relative bias for DSE is approximately more than 1%. Figure 5 depicts the relative root mean square error over all the age groups. For RRMSE, the proposed model shows lower error in each local authority and age-sex group domain.

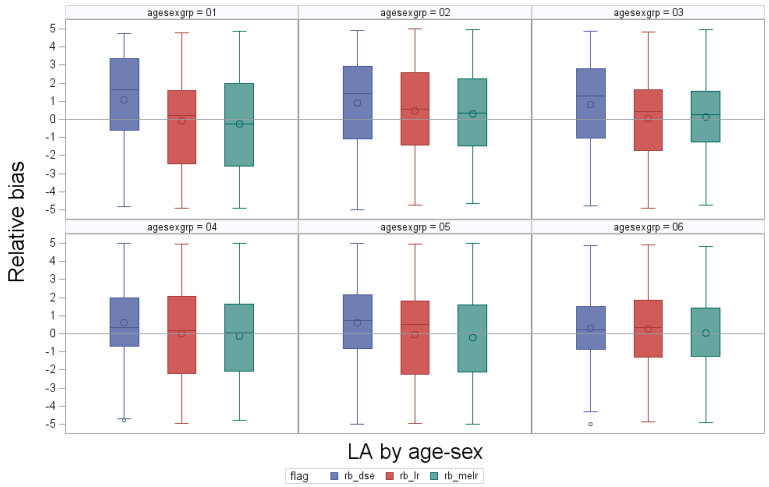

Figure 4: Relative bias by age-sex group (agesexgrp 01 = Males and Females aged $0-14$, agesexgrp 02 = Males and Females aged $15-29$, agesexgrp 03 = Males and Females aged $30-59$, agesexgrp 04 = males 60+, agesexgrp 05 = Females aged $60-79$, agesexgrp 06 = Females aged 80+)

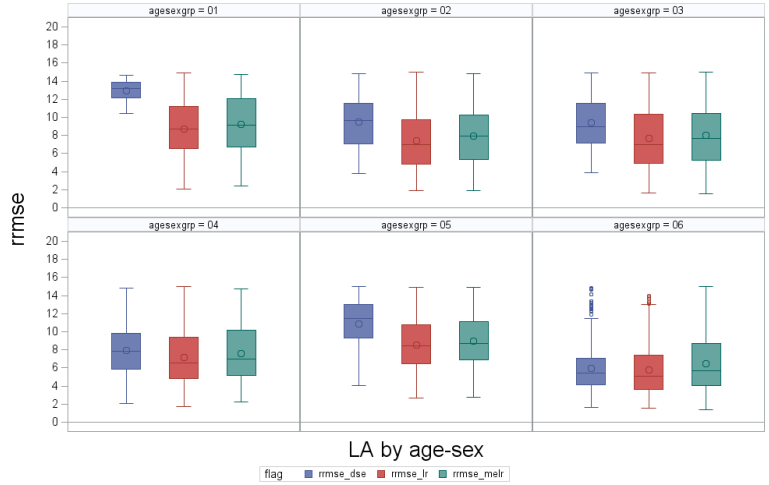

Figure 5: RRMSE by age-sex group (agesexgrp 01 = Males and Females aged $0-14$, agesexgrp 02 = Males and Females aged $15-29$, agesexgrp 03 = Males and Females aged $30-59$, agesexgrp 04 = males 60+, agesexgrp 05 = Females aged $60-79$, agesexgrp 06 = Females aged 80+)

7.3 Results by Establishment Nature

Accross establishment nature categories the relative bias for LR (Logistic regression) and MELR(Mixed effect logistic regression) remain close to zero. However for the DSE

relative bias varies accross establishment type see Figure 6 in the appendix. Figure 7 in the appendix depicts higher variability of DSE model compare to other two models.

Across all of the different domains we have looked at, the LR/MELR model performs as good (as) or better than the DSE approach, which is consistent with the results in the work done on census coverage for households (Račinskij (2018), Račinskij (2019)).

8 Conclusion & Recommendation

In this research we have compared three dual system population size estimators (e.g. DSE using post-stratification, DSE with LR and DSE with MELR). The measured bias and variation in the error are similar for the LR and MELR models. However there was increased bias and variation in the error when using Dual system estimator with poststratification. It may inferred that LR/MELR may improve upon the post-stratification approach using parametric modelling for covariate effects. Actually most of the bias comes through the synthetic application of regional estimated ratio to LAs. The key point is that the LR and MELR are not restricted by small samples in this way and can use a more flexible choice of model to give better LA level estimation. This is not a surprising result given the sample size and collapsing necessary in the 2011 approach. Clearly, having and being able to use data from across England and Wales is an advantage over estimating areas one by one. One of the biggest challenges of modelling CEs is that they are not considered when drawing the CCS sample. Therefore, insufficient sample size may have impact on the quality of estimates. Given the analysis done in this paper, it seems that this risk may be lowered by logistic regression/mixed effect logistic regression methods, and we advocate use of the LR/MELR approach. We also advocate keeping 2011 small CE coverage estimation strategy i.e. dual system/ratio estimator to derive regional level estimates by broad type and broad age-sex as a back up option. In the case of major data collection / processing issues and strong evidence that the modelling approach would produce poorer quality estimates compared to the 2011 approach, the latter could be used (at least in the areas where major issues occurred).

Bibliography

Abbott, O. (2011) Estimation & Adjustment for Communal Establishments in the 2011 Census.

Akanda, M.A.S., Alpizar-Jara, R. (2014) *A generalized estimating equations approach for capture–recapture closed population models Environ Ecol Stat 21*, 667–688.

Agresti, A. (2002) *Categorical Data Analysis* 2nd. edition. Wiley. New York, USA.

Alho, J. (1990) Logistic Regression in Capture-Recapture Models. *Biometrics, 46*, 623-635.

Alho, J., Mulry, M., Wurdeman, K. & Kim, J. (1993) Estimating Heterogeneity in the Probabilities of Enumeration for Dual-System Estimation. *Journal of the American Statistical Association, 88*, 1130- 1136.

Brown, J., Abbott, O. & Smith, P. (2013) Design of the 2001 and 2011 census coverage surveys for England and Wales. *Journal of the Royal Statistical Society Series A, 169*, 883-902.

Brown, J., Sexton, C., Abbott, O. & Smith, P. A. (2019) The framework for estimating coverage in the 2011 Census of England and Wales: combining dual-system estimation with ratio estimation. *Statistical Journal of the IAOS*.

Burke, D. and Račinskij, V. (2020) Census coverage survey 2021 sample allocation strategy. *Report to be presented at the Census External Assurance Panel on 24 March, 2020*.

Huggins, R. M.(1989) On the Statistical Analysis of Capture Experiments. *Biometrika, 76(1)*, 133-140.

Huggins, R. M.(1991) On the Statistical Analysis of Capture Experiments. *Biometrics, 47(2)*, 725-732.

Pollock, K. H., Hines, J. & Nichols, J. (1984) The Use of Auxillary Variables in Capture-Recapture and removal Experiments. *Biometrics, 40(2)*, 329-340.

Office for National Statistics (2012) Household bias adjustment (2011 Census Evaluation Report). Office for National Statistics.

Office for National Statistics (2012) 2011 Census (Estimation and Adjustment for Communal Establishments). Office for National Statistics.

Račinskij, V. (2018) Coverage Estimation Strategy for the 2021 Census of England and Wales. *Report presented at the Census External Assurance Panel on 16 October, 2018*.

Račinskij, V. (2019) Estimation of the household population in 2021 Census of England and Wales: initial ideas and results. Internal ONS report. Available on request.

Račinskij, V. & Hammond, C. (2019) Overcoverage Estimation Strategy for the 2021 Census of England and Wales. *Report presented at the Census External Assurance Panel on 17 October, 2019.*

Račinskij, V. (2020) Dealing with informative sampling in 2021 Census of England and Wales. Internal ONS report. Available on request.

Račinskij, V. (2020) Private communication.

Appendix

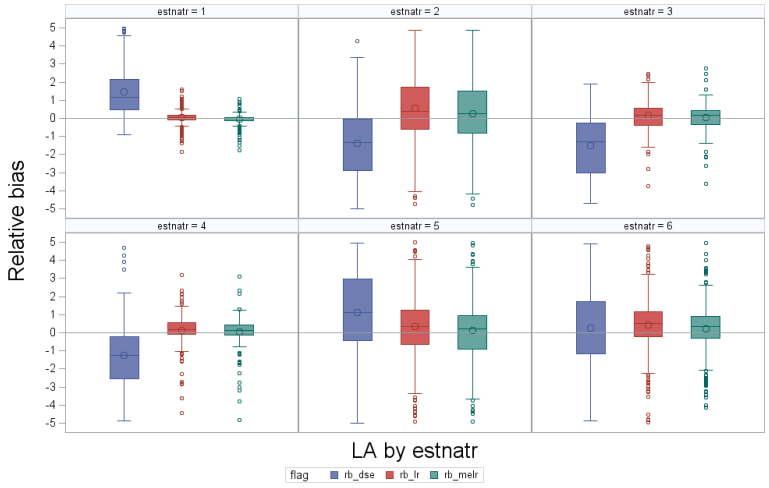

Figure 6: Relative bias by Est-Nature (estnature 01 = Hospitals,Medical and Care establishments, estnature 02 = School, University Halls, estnature 03 = Defence Establishments, estnature 04 = Prisons and Detention Centre, estnature 06 = Religious and Other establishments)

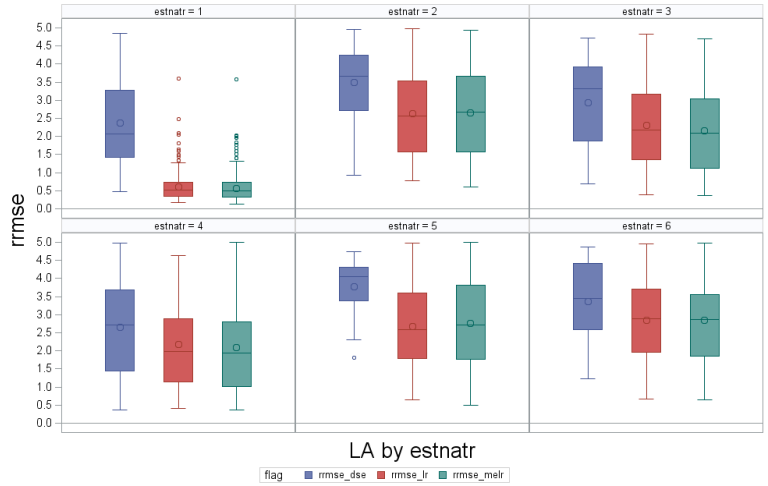

Figure 7: RRMSE by Est-Nature (estnature 01 = Hospitals,Medical and Care establishments, estnature 02 = School, University Halls, estnature 03 = Defence Establishments, estnature 04 = Prisons and Detention Centre, estnature 06 = Religious and Other establishments)