Using external datasets for small area estimation - a simulation example using Swiss forest management inventory data

Liviu T. Ene & Leo Bont

Swiss Federal Institute for Forest, Snow and Landscape Research WSL

NB-Nord, Ås, 19-20 June 2018

Motivation

- The work was initialy focused on integrating the National Forest Inventories (NFIs) and Forest Management Inventories (FMIs)
  - Task 2.3.3 of the EU project ‘’DIABOLO’’

- The information needs for operational and tactical forest management and planning cannot be fully supported by the NFIs (Kangas et al 2018)
  - Information that can improve the end-product value not produced by the NFIs
  - The NFIs estimates for the initial state of the forest management units (stands) or at higher management (domains) level not very accurate

Kangas et al (2018). Remote sensing and forest inventories in Nordic countries – roadmap for the future. Scandinavian Journal of Forest Research, 33:4, 397-412, DOI: 10.1080/02827581.2017.1416666

Background

- Although the stand-level predictions methods supported by remotely sensed data is a model-dependent inferential approach, FMIs still rely heavily on probability sampling for acquiring the field information
  - The use of probability sampling guarantees the (approximate) design-unbiasedness at enterprise-level, but the prediction error at management unit is unknown

- In the current parlance, a forest stand can be consider as a ‘’small-area’’ receiving a very small (most often zero) sample size

- With the raise of the “big data” applications in forestry, it is expected to witness an abundant flow of data that may or may not provide an appropriate basis for statistical inference for small-areas.

Background

- From a statistical point of view, the “big data” can be seen as a nonprobability sample
- For instance, ‘’big data’’ can be a database containing field and auxiliary information provided by different sources
  - Data streams are also possible to consider
- The observations in a such database could be related to:
  - individual tree information acquired by harvesters, from which “pseudo-plots” can be created by aggregating the tree measurements based on relative or absolute positions
  - the observations could also consist of field observations (plots) provided by other inventories (even by the NFI)
- It is also assumed that some kind of auxiliary information related to the individual trees or to the entire field plot is also available.

Inference for nonprobability sampling

- Two general approaches (Baker at al 2013, Elliott & Valliant 2017):
  - The quasi-randomization approach
  - The superpopulation model approach

- The *quasi-randomization* assigns so-called "pseudo-inclusion probabilities" to the nonprobability sample data,
  - The classical estimators for unequal probability sample (like the Horvitz-Thompson estimator) can be used
  - Despite the cosmetic resemblance to the inference for probability sampling, the quasi-randomization approach does not guarantee the design-unbiasedness, being in fact a model-dependent inferential approach.
  - **Does not allow for unit-level (spatial) predictions**

- The *superpopulation model* approach assumes the existence of a common data generating process common to the population of interest and to the observations (or at least for some) in the non-probability sample.
  - **Unit-level (spatial) predictions possible**

Baker R, Brick MJ, Battaglia M, Couper MP, Dever JA, Gile KJ et al (2013). Report of the AAPOR Task Force on Non-probability Sampling. Technical report, American Association for Public Opinion Research, Deerfield, IL.

Elliott MR & Valliant R (2017). Inference for nonprobability sampling. Statistical Science, 32(2), 249-264

Inference for nonprobability sampling

- These two main approaches can be further combined in a double-robust (DR) estimation procedure (Kang & Schafer 2007 and references therein)

- DR estimation assumes the existence of two models:
  - A probability model (π-model)
  - A response model (y-model)
  - Several ways of sing the synergies between the π- and y-models are described by Kang & Schafer (2007)

- The minimum requirements for DR:
  - the existence of a common, consistent set of auxiliaries in the non-probability sample and population of interest

Kang DY & Schafer JL (2007). Demystifying double-robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science, 22(4), 523-539.

The π-models

- The π-models are probability models
  - Predict the probability that a sample observation *i* could belong to the population of interest (the small-area in the current parlance)
    - The sample data can be probabilistic or not

- The specification of the π-model requires:
  - A binary (0-1) response vector
    - t_i=1 for the small-area observations
    - t_i= 0 for the sample observations
  - A set of auxiliaries X common for a particular small-area and for the sample

The π-models

- Define the response probability for a unit *i* as:

$$P(t_i = 1|X_i) = \pi_i(X_i) = \pi_i$$

  - π_i - propensity score (Rosenbaum & Rubin 1983)
    - propensity score - the probability that a unit with certain characteristics will be assigned to the treatment group, i.e., the probability of an observation in group label as 0 to belong to group labeled as 1
  - The functional form for predicting π_i is called as the π-model
    - examples: binomial GLM models

- The predicted propensity scores π_i should be close to 1 for the sample observations that share common traits (given the auxiliaries) with the small-area units

Rosenbaum PR & Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70, 41–55.

The y-models

- Links the field attributes to auxiliaries
- The functional form of the y-model could be, f. instance:

$$E(y_i|X_i) = \xi(X_i) + \varepsilon_i, \text{ with } E(\varepsilon_i) = 0$$

- In the simplest case, the coefficients of $\xi(X_i)$ can be estimated by ordinary least squares
- Nonprametric formulation for $\xi(X_i)$ possible, such as nearest neighbor imputations

DR estimation for small areas

- Double-robust (DR) procedures integrate both the π-models and the y-models
  - The double-robustness is based on the assumption that at least one of the model is appropriate
- If both the π-models and the y-models are wrong, then there is no gain in using DR
- Either way, we will never know the model biases
  - The curse of the model-dependent inference

DR estimation for small areas

- Under the common FMIs, most of the forest stands will not receive sampling points
  - This makes the case for the model-dependent inference via a synthetic estimator trained on data which is external to the small-area
- Relative to the forest stand, the sampling design for acquiring the sample data is not relevant anymore
  - if the sampling is non-informative
- The DR approach allows for the y-models to be tailored to each particular small-area via the propensity scores predicted by the π-models

Analyses

- The analysis workflow can be summarized as follows:
  - generate large training (TRAIN) and validation (VAL) datasets using the empirical observations from field plot inventories and corresponding plot-level ALS auxiliaries;
  - create grouping structures within the validation datasets, as proxies for real forest stands;
  - estimate the group means using probability and nonprobability samples

Material

- The artificial datasets were created from emirical data from 2016, consisting of:
  - A field sample of 137 field plots from a local Forest Management Inventory in Fribourg canton, Switzerland
  - ALS data (GPD ~ 5 points m^2)
    - The usual hight percentiles and density features were extracted from the point cloud data
  - Very good temporal match between ALS and terrestrial data

Field data

Table 1. Summary of Fribourg field sample data:

Development stage	No. field plots	Standing volume (mc ha^{-1})				
		Min	Max	Mean	SD[(1)]	$CV_{\%}^{(2)}$
1	25	2.00	99.40	42.48	31.47	74.08
2	12	11.00	233.00	71.48	62.48	87.41
3	5	46.67	137.80	103.13	37.61	36.47
4	22	43.90	184.00	121.18	42.67	35.21
5	32	30.18	290.41	142.89	55.35	38.73
6	24	15.60	308.40	131.43	65.26	49.65
Other	17	0.00	196.80	27.45	64.32	42.68
Overall	137	0.00	308.40	97.04	68.76	70.85

[(1)] Standard deviation; [(2)] Coefficient of variation (%)

Artificial datasets

- The terrestrial data and the corresponding ALS features were used to create large three artificial datasets of 100,000 observations each
  - A training dataset (TRAIN)
  - Two validation datasets (VAL.1 and VAL.2)

- The artificial datasets were generated by copula functions, using the R packages “CDVine” (Brechmann et al 2013) and “VineCopulas” (Schepsmeier et al 2017).

- The TRAIN dataset plays the role of an external database of plot-level field inventory data and ALS auxiliaries, and it can be seen as a very large non-probability sample relative to VAL.1 and VAL.2

Brechmann E C & Schepsmeier U (2013). Modeling Dependence with C- and D-Vine Copulas: The R Package CDVine. Journal of Statistical Software, 52 (3), 1-27. http://www.jstatsoft.org/v52/i03/

Schepsmeier U, Stoeber J, Brechmann EC, Graeler B, Nagler T & Erhardt T (2017). VineCopula: Statistical Inference of Vine Copulas. R package version 2.1.3. (URL https://CRAN.R-project.org/package=VineCopula)

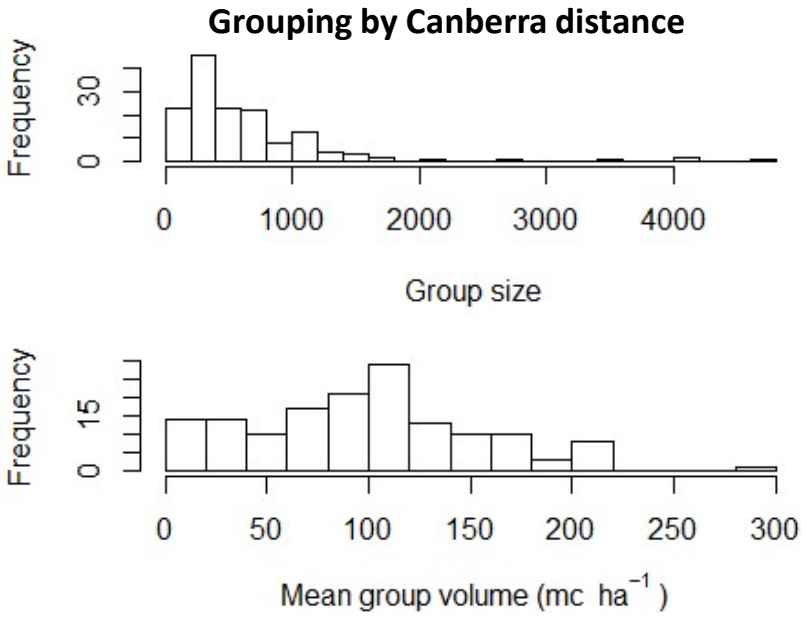
Create the grouping structures for the validation datasets

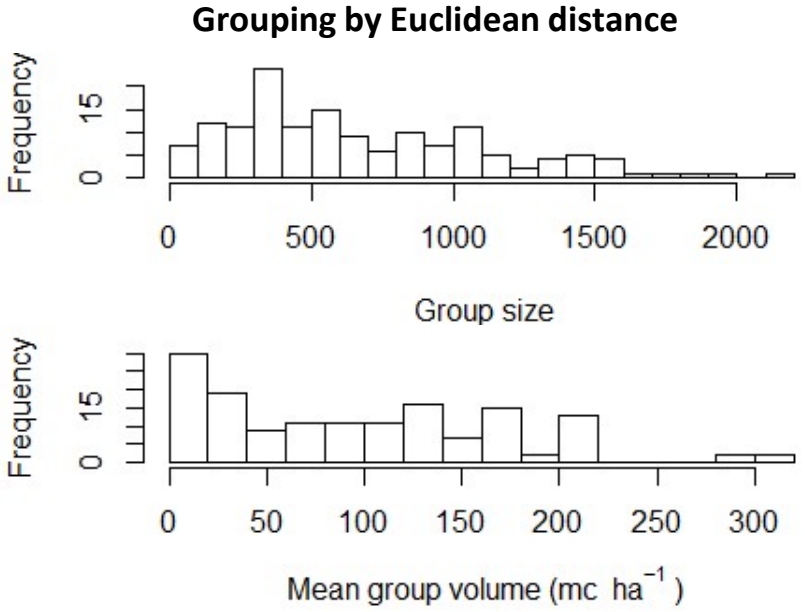
- For each validation dataset, several grouping structures were generated using 25% of the observations:
  - hierarchical clustering using Ward's method on the Euclidean and Canberra distances
  - cut-off to 150 groups

- The rest of 75% of the observations in the validation datasets were assigned to one of these groups by nearest-neighbor imputations

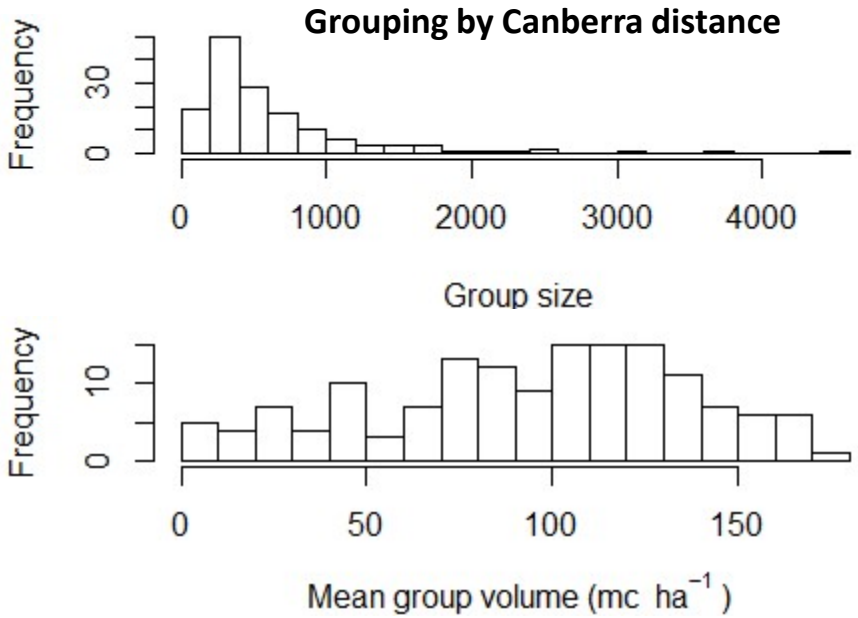
- The groups with less than 25 observations were merged
  - for field plots of 400 m^2, a group of 25 observations would correspond to forest stand of approximately 1.0 ha.

Grouping structure for VAL.1 population





Grouping structure for VAL.2 population



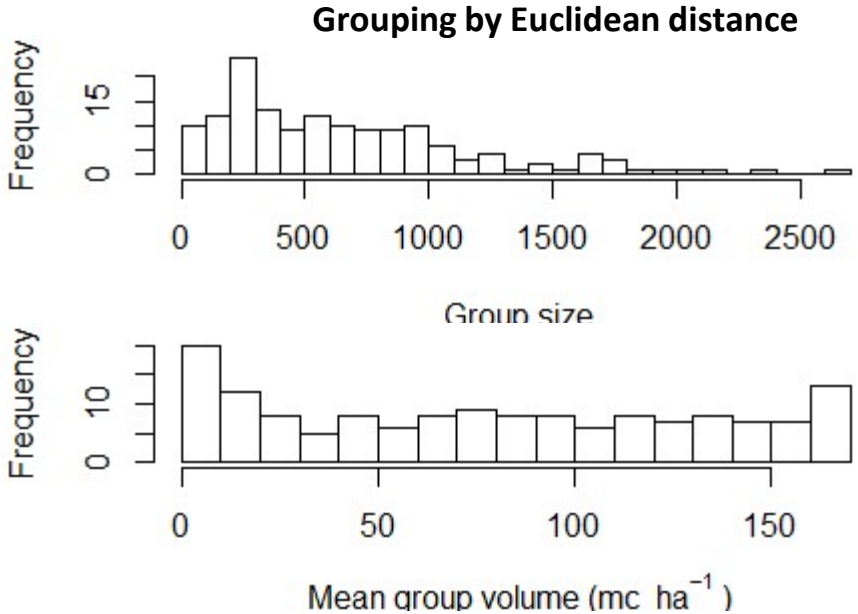


Table 2. Realized number of groups and median group size for artificial populations:

Population	Grouped by	No. Groups	Median group size
VAL.1	Euclidean distance	148	547
	Canberra distance	150	473
VAL.2	Euclidean distance	148	568
	Canberra distance	148	438

Estimation strategies

- Probabilistic models (π-models)
  - The propensity scores were predicted by two types of estimators:
    - a parametric model formulated as generalized linear model for binary responses (GLM)
      - Feature selection and model fitting using the "speedglm" R package (Enea 2017)
    - Support Vector Machines (SVM) implemented in the "kernlab" R package (Karatzoglou et al 2004), using a Gaussian kernel with automatic tuning of the kernel coefficients

Enea M (2017). speedglm: Fitting Linear and Generalized Linear Models to Large Data Sets. R package version 0.3-2. (URL https://CRAN.R-project.org/package=speedglm)

Karatzoglou A, Smola A, Hornik K, Zeileis A (2004). kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software 11(9), 1-20. URL http://www.jstatsoft.org/v11/i09/

Estimation strategies

- Response models (y-models):
  - Linear regression models estimated by OSL and WLS;
  - Linear regression models estimated by the Least Absolute Shrinkage and Selection Operator (LASSO, Tibshirani 1996) using the "glmnet" R package (Friednam et al 2010);
  - Nearest-neighbor imputations (k=1) performed using the "FNN" R package (Beygelzimer et al 2013).

Beygelzimer A, Kakadet S, Langford J, Arya S, Mount D & Li S (2013). FNN: Fast Nearest Neighbor Search Algorithms and Applications. R package version 1.1.(URL https://CRAN.R-project.org/package=FNN)

Friedman J, Hastie T & Tibshirani R (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/.

Tibshirani R (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, 58(1), 267-288.

Estimators

- Under the *superpopulation model* approach
  - Synthetic linear regression estimators
    - OLS regression (REG)
    - Lasso regression (LAS)
  - Double robust (DR) estimation:
    - Linear regression with residual bias correction, using the propensity weights as pseudo-inclusion probabilities to adjust the predictions of from OLS regression (REG.c) and LASSO (LAS.c)
    - Propensity weighted regression estimation (REG.w), using the propensity weights as precision weights for weighted least square (WLS) fitting

Estimation strategies

- Under *quasi-randomization* approach
  - propensity weighting of the responses in the sample data (PW)

- In addition:
  - Nearest neighbour (NN) imputations using the Euclidean distance
  - Weighted nearest neighbour (NN.w) imputations using the Euclidean distance weighted by the propensity scores of a π-model

Estimation strategies

- The inference to VAL.1 and VAL.2 was addressed considering the following case studies:
  - using TRAIN as an external database (large non-probability sample)
    - In order to obtain a balanced dataset for the π-model, SRSwoR samples of size equal the median size of the small-areas were selected from TRAIN
    - Note that more data do not compensate for model bias
    - The samples should be large enough for NN imputations
  - using probability samples of size n=50 observations selected by SRSwoR from VAL.1 and VAL.2

Case studies

- The robustness of the estimators was assessed by running simulation trials
- The factors considered
  - Grouping structure: two similarity measures for hierarchical clustering
    - Euclidean and Canberra distances
  - Two formulations for the π-model
    - Binomial logistic regression (GLM)
    - Support Vector Machines (SVM)

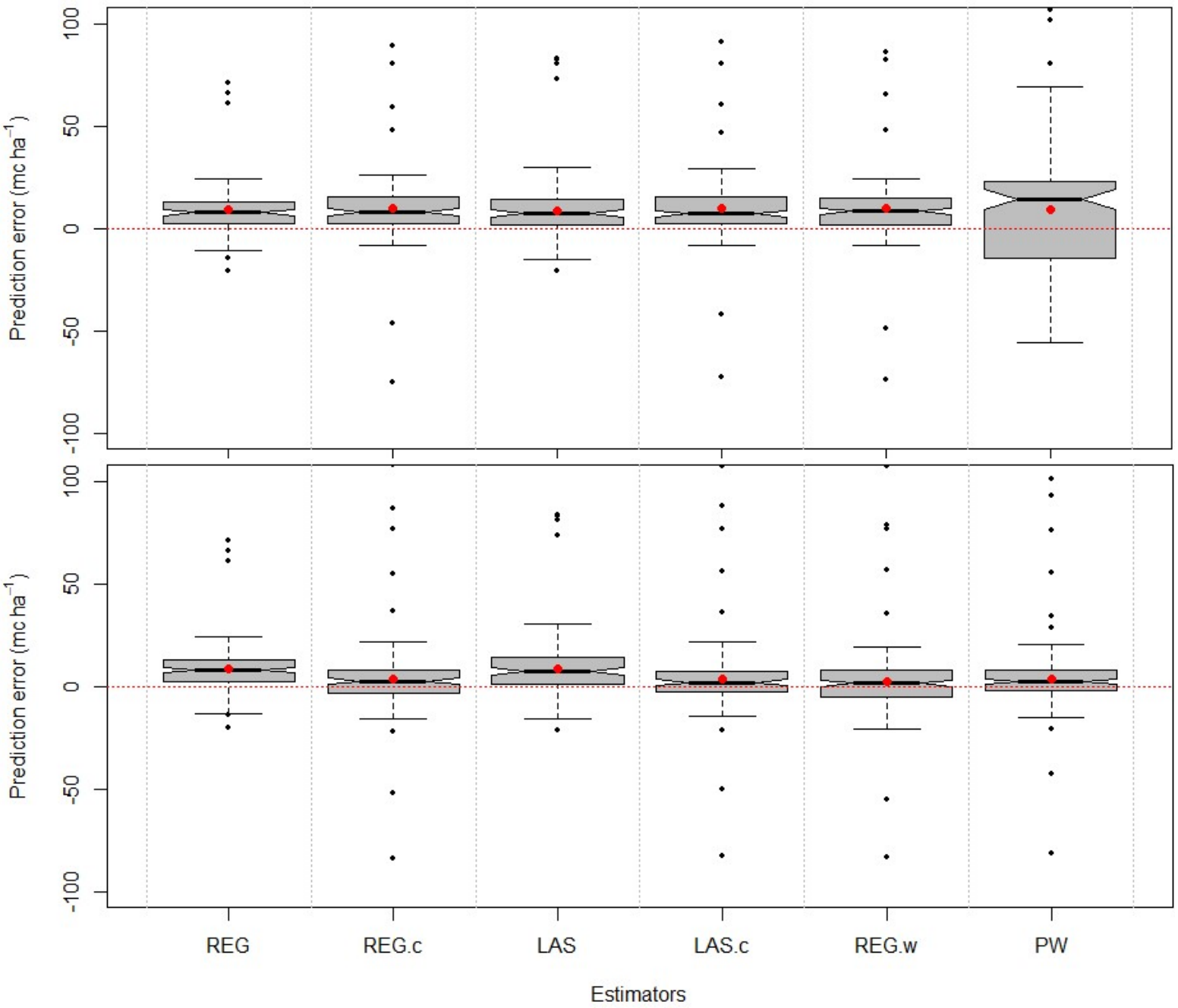
Table 3. Overview of the case studies:

Model-dependent inference for:					
Nonprobability samples			Probability samples		
π-model	Similarity method for clustering		π-model	Similarity method for clustering	
	Euclidean	Canberra		Euclidean	Canberra
SVM	VAL.1,2	VAL.1,2	SVM	VAL.1,2	VAL.1,2
GLM	VAL.1,2	VAL.1,2	GLM	VAL.1,2	VAL.1,2

Results

- Inference for nonprobabilty samples
  - Using the TRAIN dataset to predict on VAL.1 and VAL.2
  - Sample size equal to median group size, by grouping for each population populations (from Table 2)

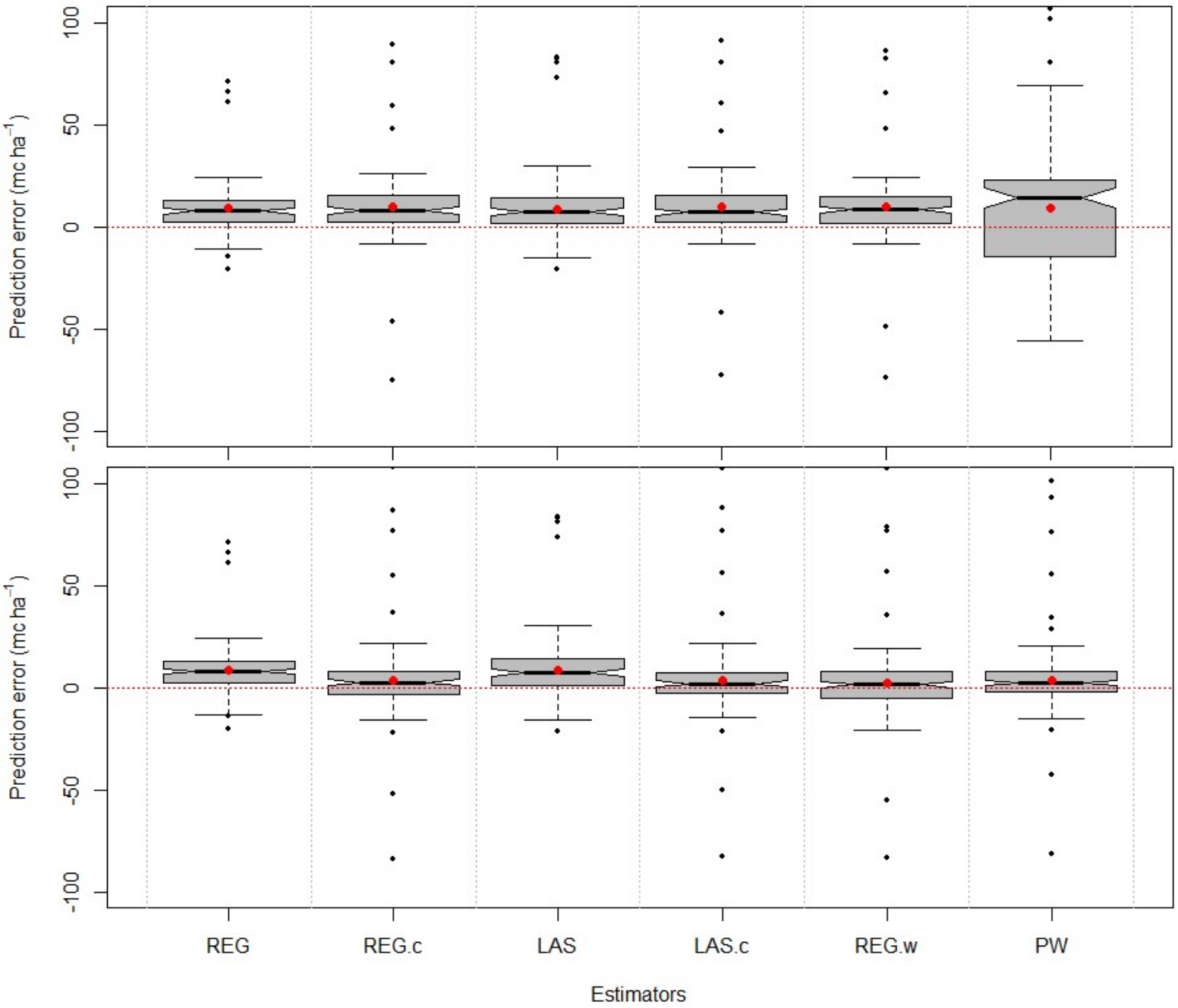
VAL.1 population, external data, grouping by Euclidean distance



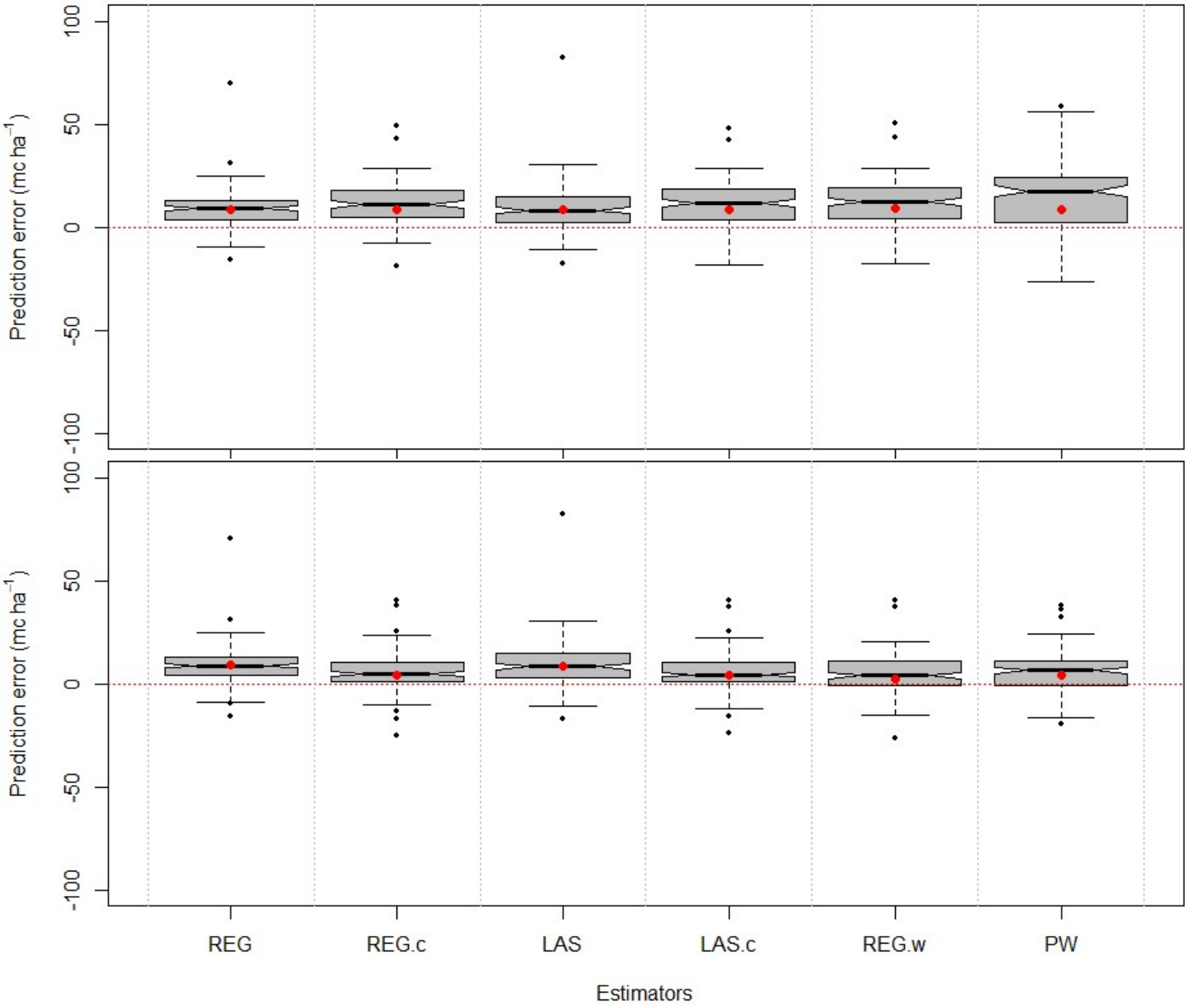
SVM π-model

GLM π-model

VAL.2 population, external data, grouping by Euclidean distance



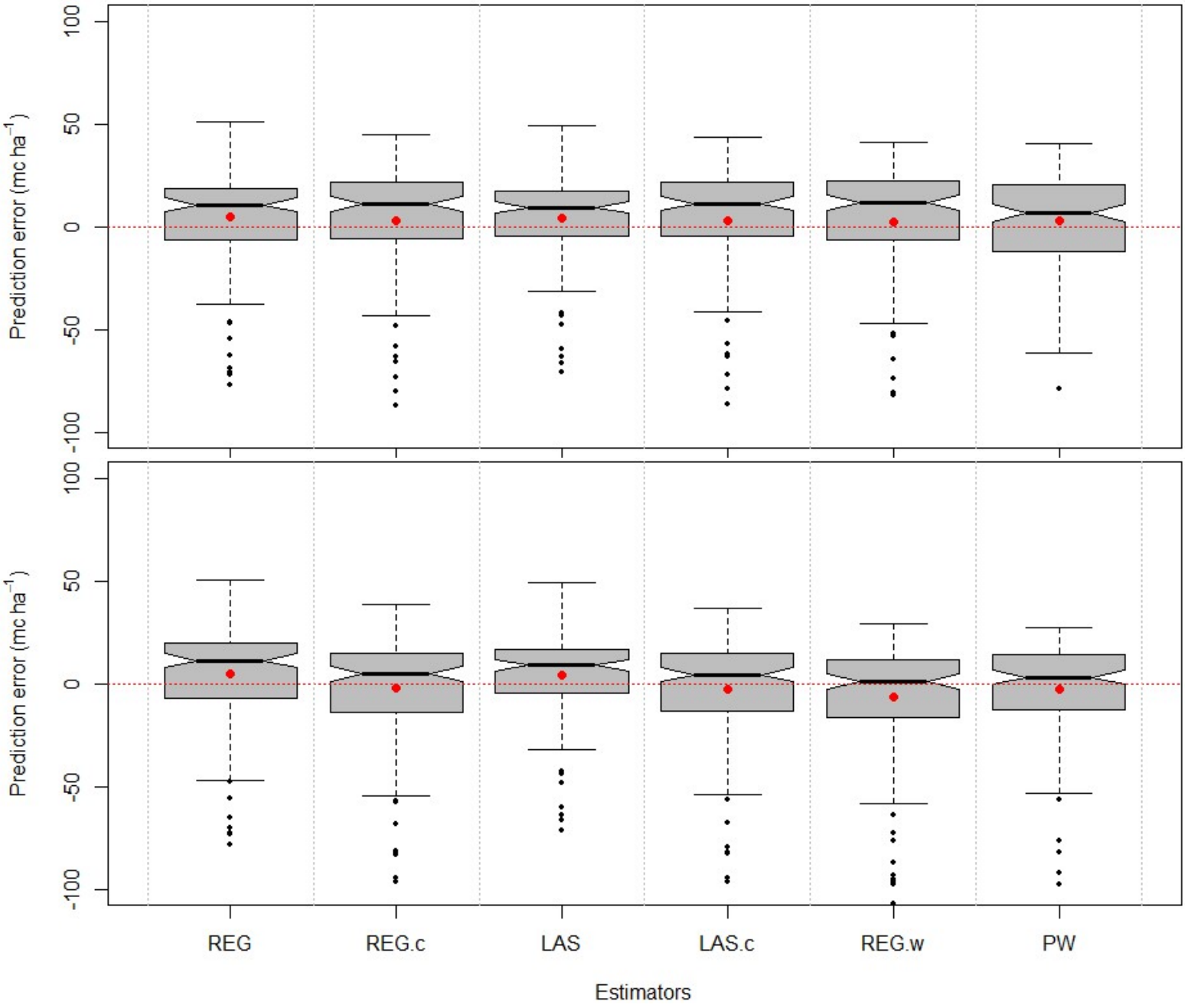
VAL.1 population, external data, grouping by Canberra distance



SVM π-model

GLM π-model

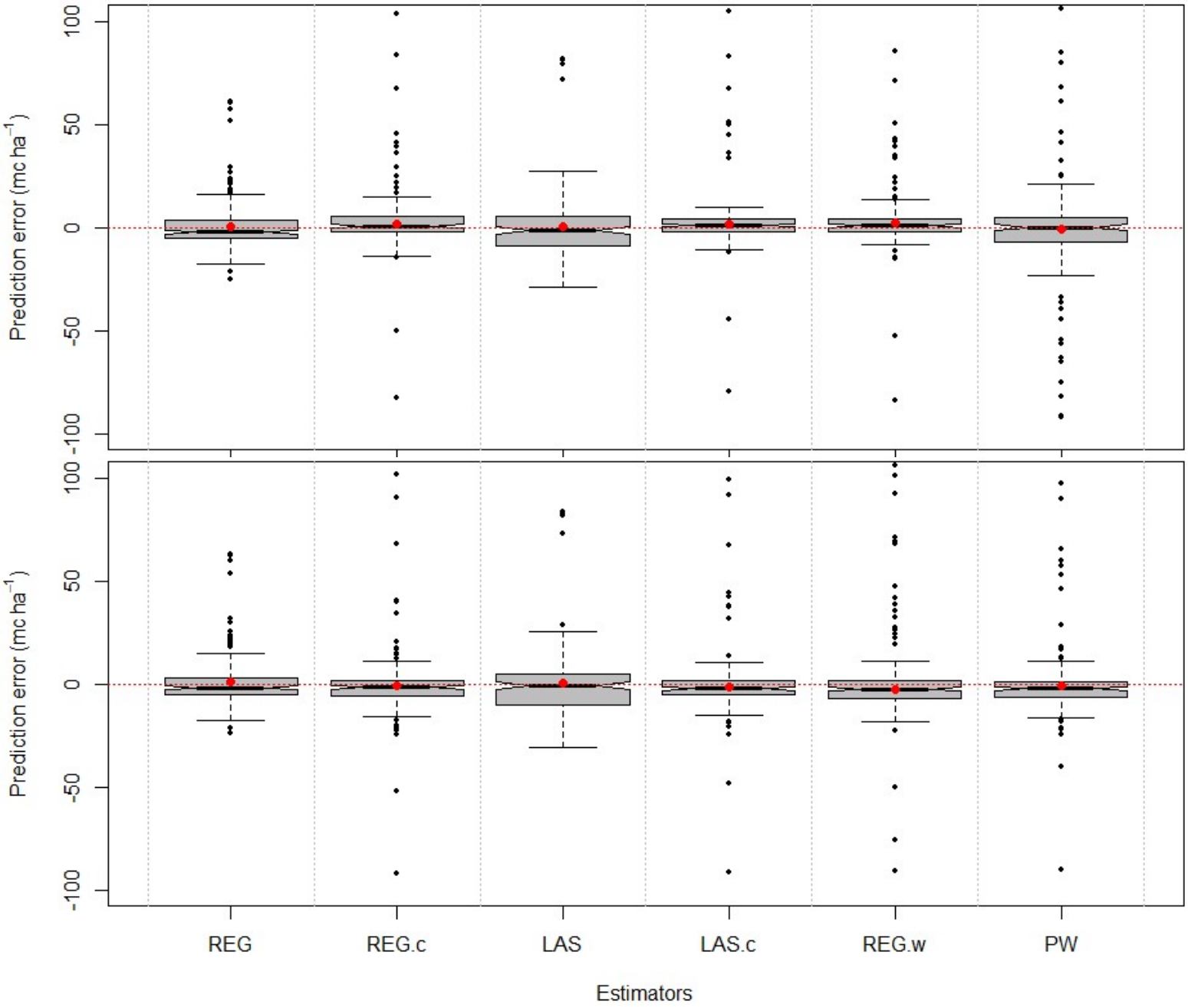
VAL.2 population, external data, grouping by Canberra distance



Results

- Inference for probability samples
  - Estimators trained on current inventory data
    - SRSwoR sampling from the VAL.1 and VAL.2 for model-dependent small-area inference
    - Sample size: 40 observations

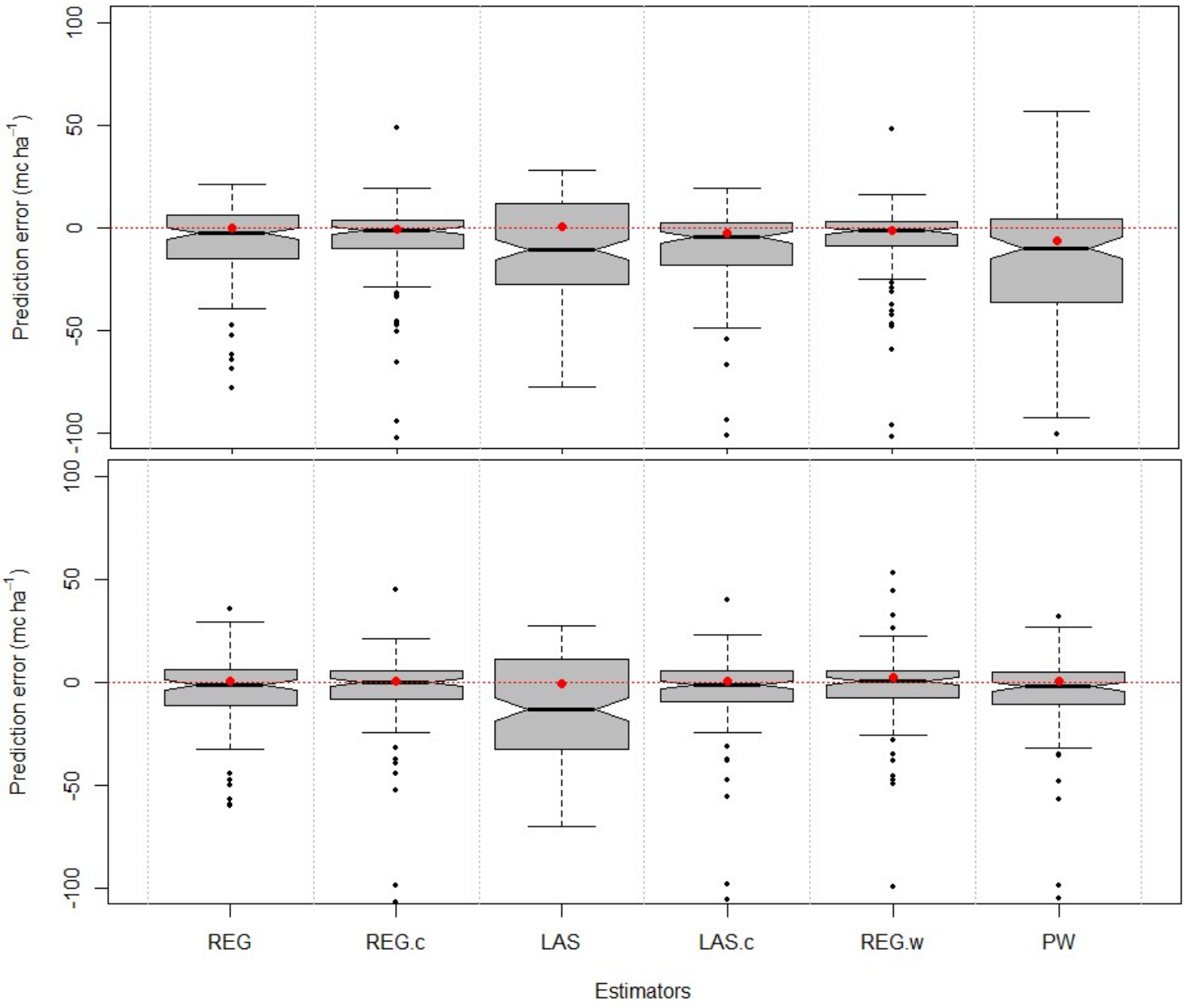
VAL.1 population, external data, grouping by Euclidean distance



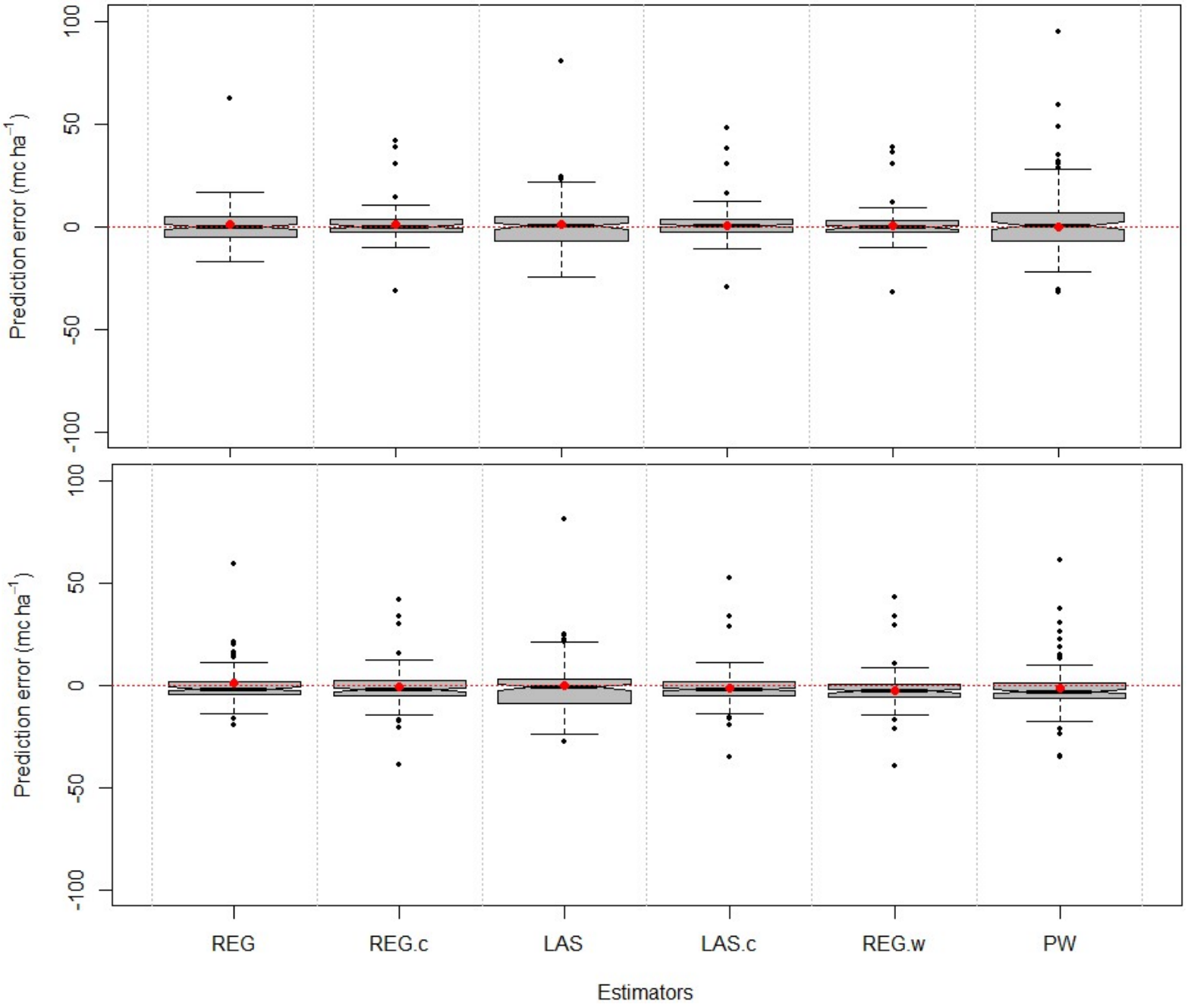
SVM π-model

GLM π-model

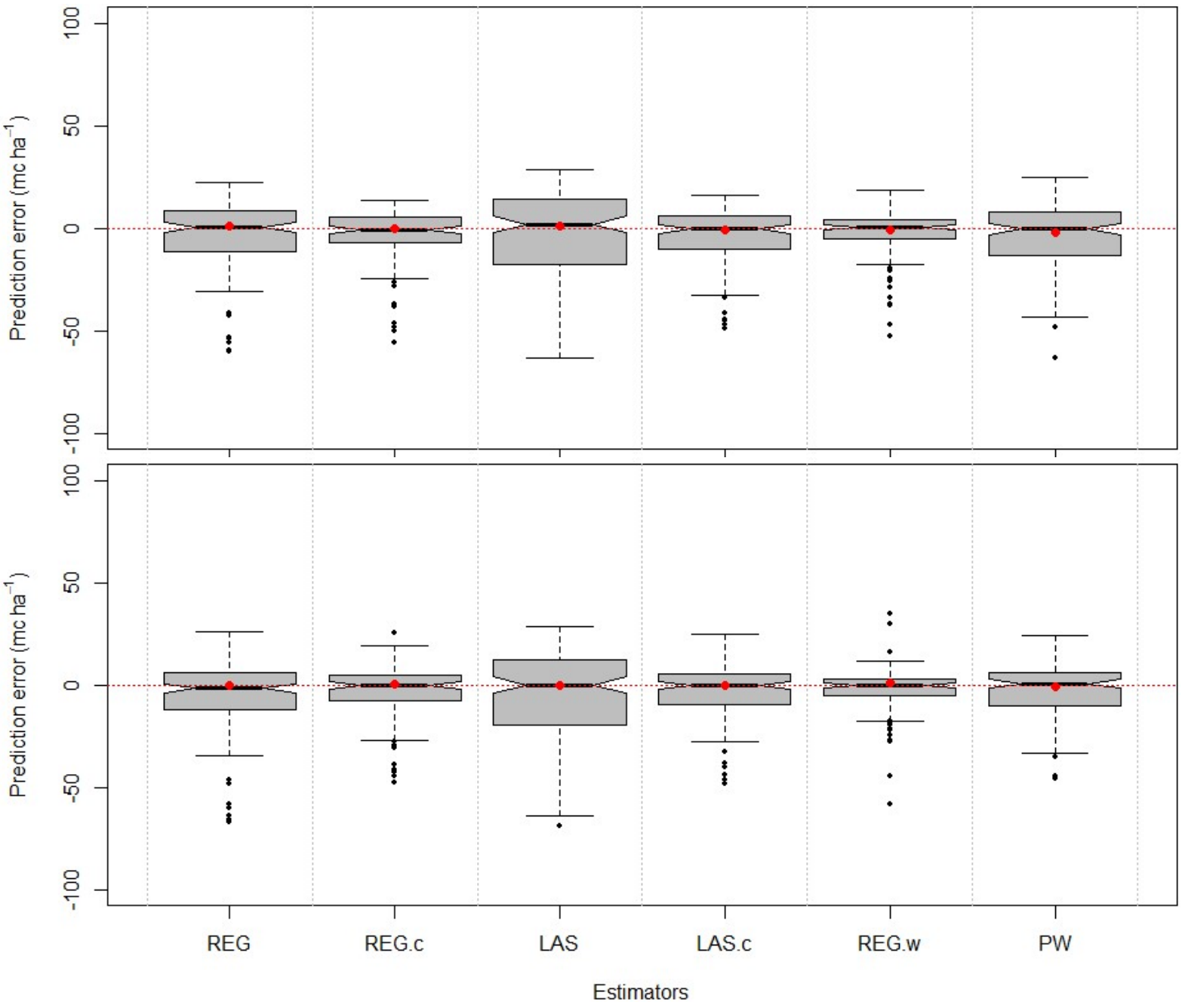
VAL.2 population, external data, grouping by Euclidean distance



VAL.1 population, external data, grouping by Canberra distance



VAL.2 population, external data, grouping by Canberra distance



SVM π-model

GLM π-model

Conclusions

- The double-robust approach reduces the group-level prediction errors for synthetic regression estimator:
  - Adjustments using the fitted residuals
  - Using the propensity scoreas as precision weights for WLS fitting
- In general, the quality of the π-models seems to dominate the results
- NN methods worked very well only for the case of probability samples, and showed little sensitivity to the propensity weighting

Conclusions

- Propensity score weighting works quite well for probability samples
  - Approx. 30-60% reduction in RMSE compared to the default synthetic estimator (i.e., ‘’classical’’ ALS inventory)
  - Binomial GLMs tend to be biased for unbalanced datastes (due to small terrestrial sample)
  - SVM more robust to unbalanced dataste
- For nonprobability samples, the prediction errors are higher (as expected), but the weighting still provide slight improvements
  - The π-models formulated as GLMs perform better, the datasets are well balanced
  - SVM not good for extrapolation

Challenges

- Investigate possible diagnostics for the risk of π-models failure
- Assess combinations of probability and nonprobability samples
  - The workflow for the DR approach does not change
- Use a database of ‘pseudo-plots’ or a set of working models?
  - Accuracy vs. feasibility trade-offs
- How to use incorporate the propensity weighting in the NN imputations
- Since estimating the propensities weights does not reuqire spatial information (like plot coordinates), would we get easier access to the NFI plot information?

Challenges

- How to report the stand-level accuracy
  - Reporting the prediction errors of the synthetic estimator does not capture the lack-of fit with regard to model bias
  - It is exected to have low model errors, but large prediction errors
    - The nominal coverage of the prediction intervals will be too narrow
- What is the effect of using DR estimation for decision-making?