Random forests and mixed effects random forests for small area estimation of general parameters

Nikos Tzavidis

University of Southampton

Random forests have been shown to excel in terms of predictive performance. The use of few tuning parameters, automated model-selection, and their ability to detect higher order interactions and complex relationships make their use appealing. Some examples of research on tree-based methods for the analysis of complex survey data and survey estimation include Toth & Eltinge (2011), Breidt & Opsomer (2017), Bilton et al. (2017), and Bilton et al. (2020). More recently, Dagdoug et al. (2021) study the theoretical properties of random forests for complex survey data. In this work we study the use of random forests and extensions for estimating general small area parameters. Conventionally, random forests do not include random effects. However, random effects play a central role in small area estimation. We propose an extension of random forests to mixed effects random forests that combines the random forest fitting algorithm with a mixed effects model to exploit clustering in out-of-bag residuals. The proposed fitting algorithm extends previous work by Krennmair & Schmid (2022). It uses non-parametric bootstrap to correct the bias (due to the estimation of the random forest) in the estimated residual variance (Mendez & Lohr, 2011) before proceeding to estimate the variance components and the random effects. Ignoring this bias impacts both on point and mean squared error (MSE) estimation. Estimators of general small area parameters are derived by using a smearing estimator of the area-specific distribution function (Chambers et al., 2014). MSE estimation under machine learning methods remains a largely unexplored research area. In this work we study MSE estimation using non-parametric block bootstrap with appropriate scaling of the residuals. The proposed methods are evaluated in model-based simulations and by using real data from a poverty assessment case study in Mozambique. Comparisons to industry standard methods under a linear mixed model e.g., the Empirical Best Predictor (also with data driven transformations), and to a synthetic estimator under the random forest are presented. The empirical evaluations and the real data application inform us about (a) the impact of including random effects in random forests, (b) the importance of using data transformations with random forests, and (c) the performance of point and MSE estimators. The current approach to including random effects in random forests is more in line with the data modelling culture than the algorithmic modelling culture (Breiman, 2001; Efron, 2020). We critically discuss the proposed approach and outline possible alternatives.

Co-authors: Patrick Krennmair, Timo Schmid, and Nora Wür

References

Bilton, P., Jones, G., Ganesh, S., & Haslett, S. (2017). Classification trees for poverty mapping. *Computational Statistics & Data Analysis*, *115* , 53–66.

Bilton, P., Jones, G., Ganesh, S., & Haslett, S. (2020). Regression trees for poverty mapping. *Australian & New Zealand Journal of Statistics*, *62* (4), 426–443.

Breidt, F. J., & Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, *32* (2), 190–205.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16* (3), 199–231.

Chambers, R., Chandra, H., Salvati, N., & Tzavidis, N. (2014). Outlier robust small area estimation. *Journal of the Royal Statistical Society: Series B*, *76 (1)*, 47-69.

Dagdoug, M., Goga, C., & Haziza, D. (2021). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, 1–18.

Efron, B. (2020). Prediction, estimation, and attribution. *Journal of the American Statistical Association*, *115* (530), 636-655.

Krennmair, P., & Schmid, T. (2022). Flexible domain prediction using mixed effects random forests. *Journal of Royal Statistical Society: Series C (Applied Statistics)*, *71* (5), 1865–1894.

Mendez, G., & Lohr, S. (2011). Estimating residual variance in random forest regression. *Computational Statistics & Data Analysis*, *55* (11), 2937–2950.

Toth, D., & Eltinge, J. L. (2011). Building consistent regression trees from complex sample data. *Journal of the American Statistical Association*, *106* (496), 1626-1636.