Project title: Interpretable machine learning (iML) for official statistics

Supervisor: Katarzyna Reluga

Project description

Accurate estimation and prediction of socioeconomic indicators are essential for implementing policies effectively and optimizing resource allocation. While the population-level estimation of these indicators (e.g., for countries or large demographic groups) is well established in the literature [1, 2, 3, 5], there is ongoing research focused on developing efficient methods to estimate these indicators at more granular geographical and socio-economic levels, such as local communities and districts [3, 4].

This project seeks to fully integrate interpretable machine learning (iML) into official statistics, offering a comprehensive framework for data-driven policymaking. The prospective PhD student will work on both developing new theories around interpretable ML in official statistics and building user-friendly software (e.g., R packages, Shiny apps, Python libraries) for use by policymakers. The emphasis on each component will depend on the student's preferences, but both aspects are essential to the project. The aim is to develop "case-by-case fine-tuned" procedures for estimating socioeconomic indicators by combining state-of-the-art machine learning algorithms – such as adapted versions of gradient boosting, random forests, Bayesian regression trees, and deep neural networks – with traditional statistical methods, including linear and generalized linear mixed models. The innovation of this approach lies in its flexibility and independence from any single statistical or ML method. Instead, the focus is on delivering the most appropriate solution based on a pre-defined optimality measure (e.g., mean squared error), tailored to the specific characteristics of the data.

References

[1] Hájek, J. (1971). Comment on “An essay on the logical foundations of survey sampling, part one”. *The foundations of survey sampling*, 236.

[2] Horvitz, D. G. and Thompson, D. J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, 47(260):663– 685.

[3] Morales, D., Esteban, M. D., Pérez, A., and Hobza, T. (2021). *A course on small area estimation and mixed models*. Springer.

[4] Rao, J. N. K. and Molina, I. (2015). *Small area estimation*. John Wiley & Sons.

[5] Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer, Berlin.