Master Class on Predictive Inference of Machine Learning for Official Statistics by Li-Chun Zhang

David Haziza

Department of Mathematics and Statistics
University of Ottawa

4th Workshop on Methodologies for Official Statistics
Rome, Italy

December 1, 2025

Machine learning in surveys

- Advantages:
  - Unlike traditional parametric methods, ML models are inherently robust in capturing complex non-linear relationships, even when specific polynomial terms (e.g., quadratic, cubic) or interaction terms are not included in the model.
  - ML often require minimal manual effort.
  - Many ML algorithms are known for their ability to produce high-quality predictions.
- Drawbacks:
  - Because of their flexibility, ML procedures generally require larger sample sizes than parametric procedures, especially when the number of predictors is large $\longrightarrow$ Their rate of convergence is slower than the parametric rate of $1/\sqrt{n}$.
  - ML models require the specification of hyperparameters $\longrightarrow$ the model's architecture. The performance of ML procedures can be highly sensitive to the choice of hyperparameters.
  - Many ML algorithms are known for their ability to produce high-quality predictions.

Machine learning: Prediction vs. Estimation

- Machine learning can be applied to both prediction and the estimation of finite population parameters
- Design-based prediction/conformal prediction:
  - Toth and McConville (2024), Nalenz et al. (2024), Bersson and Hoff (2022), Wieczorek (2023)
- Estimation of finite population parameters:
  - Model-assisted estimation: Toth and McConville (2019), Sande and Zhang (2020), Dagdoug et al. (2023).
  - Treatment of unit nonresponse: Gelein (2018), Kern et al. (2019), Larbi et al. (2025); Riddles and Opsomer (2025);
  - Treatment of item nonresponse: Dagdoug et al. (2022, 2023); Dagdoug and Haziza (2025).
  - Data integration: Lee et al. (2022), Chen et al. (2023); Seaman (2025)
  - Small area estimation: Krennmair and Schmid (2023), Beaumont et al. (2023), Michal et al. (2023), Viljanen et al. (2023); Parker (2024); Wang et al. (2024); Krennmair et al. (2025); Parker and Eideh (2025).