Investigating the use of machine learning methods in Banff and G-Sam

Current and proposed research

October 14, 2020

Delivering insight through data for a better Canada

Darren Gray
Statistics Canada

Generalized Systems at Statistics Canada

System	Use
G-Link	Record linkage
★ G-Sam	Sampling
G-Code	Automated coding
★ Banff	Edit and imputation
G-Est	Estimation
G-Series	Time series
G-Confid	Disclosure avoidance
G-Tab	Tabulation

Preparing for ML methods

- *Can* we incorporate ML methods into our generalized systems?
  - For Banff, we conducted proof of concept incorporation of missForest package into E&I process flow
- *Should* we incorporate ML methods into our generalized systems?
  - Identify methods with broad application
  - Identify methods that outperform existing tools
  - For Banff, development of the Imputation Assessment and Comparison Tool (ImpACT)

ML methods under investigation

System	ML Method	Application	Status
Banff	Random Forest	Investigation of missForest imputation package	Ongoing
Banff	Feature selection and weighting	Use in Gower distance in nearest neighbour donor imputation	Ongoing
Banff & G-Sam	Clustering	Choice of strata / imputation classes	Pending

missForest package

- Popular R package for imputation; uses random forest trained on observed values to predict missing values
- Testing status:
  - Proof of concept – incorporated into Banff process flow (2018)
  - Tested against Banff imputation methods on synthetic data (2019)
  - Plans to test against production imputation process on survey data (upcoming)

Feature selection and weighting

- Previous investigation into feature selection for donor imputation in CANCEIS (Stelmack, 2018)
- Current research project into Gower distances for donor imputation (Beth Ayres) involves use of feature selection and weighting

Clustering

- Would like to investigate clustering algorithms in two domains:
  - Stratification (sampling)
  - Imputation classes

References

- Stekhoven, D. J. (2015). missForest: Nonparametric missing value imputation using random forest. *Astrophysics Source Code Library*
- Stelmack, A. (2018). On the Development of a Generalized Framework to Evaluate and Improve Imputation Strategies at Statistics Canada, *United Nations Statistical Commission and Economic Commission for Europe – Workshop on Statistical Data Editing*.
- Gray, D. (2019). A Generalized Framework to Evaluate Imputation Strategies: Recent Developments. *In JSM Proceedings, Government Statistics Section. Alexandria, VA: American Statistical Association. 1861-1870.*
- Gray, D. (2020). Evaluating Imputation Methods using ImpACT: First Case Study, *United Nations Statistical Commission and Economic Commission for Europe – Workshop on Statistical Data Editing.*

THANK YOU!

Contact:
darren.gray@canada.ca

The content of this presentation represents the position of the author and may not necessarily represent that of Statistics Canada.