CANSSI Distinguished Postdoctoral Fellows Projects List

Project Name: Aggregation of machine learning procedures for small area estimation

Supervisor:

- Name: David Haziza
- Affiliation: University of Ottawa
- Email: dhaziza@uottawa.ca
- Website: www.davidhaziza.com

Co-supervisor:

- Name: Mehdi Dagdoug
- Affiliation: McGill University
- Email: mehdi.dagdoug@mcgill.ca
- Website: https://mehdidagdoug.github.io/

Location/University:

- University of Ottawa
- McGill University

Project abstract

The objective is to develop practically relevant and theoretically sound aggregation methodologies for small area estimation based on machine learning procedures. Small area estimation refers to a class of estimation procedures used to estimate characteristics of specific subpopulations in the presence of small sample sizes. A popular area level is the so-called Fay-Herriot (FH) model that is based on a linear link function. Machine learning methods may be useful when the underlying model assumptions are violated. Which machine learning algorithm to use is a question of practical interest. Aggregation procedures combine the outputs of several machine learning procedures to produce a suitable aggregate, which can lead to an increase efficiency. This project aims at establishing the theoretical properties of aggregation procedures, applying the proposed methods to real data sets and implementing an R-package.

Plan for interdisciplinary/applied experience

In the last two decades, small area estimation has become essential in national statistical offices (e.g., Statistics Canada) because it helps address the increasing demand for localized high-quality estimates. The applications of small area estimations are quite diverse and include the estimation of poverty rates, food insecurity, and welfare program participation in small regions to target anti-poverty efforts more effectively, the estimation of educational attainment and student performance in specific school districts, the estimation of regional economic indicators, such as GDP and unemployment rates, and the estimation of pollution levels, air quality in specific geographical areas, among others. We expect the PDF to spend time at Statistics Canada during the second year of the fellowship. This will allows us to assess the proposed methodology using real data collected at Statistics Canada. David Haziza is an expert in survey sampling, and M. Dagdoug is an expert in statistical learning procedures. The success of this project hinges on the utilization of techniques from both of these fields, and the combined expertise of the mentors offers an ideal blend to meet the project's objectives.

Plan for teaching/training/education

Teaching: The postdoc will teach one three-credit course at McGill University in Year 1 and one three-credit course at the University of Ottawa in Year 2. In Year 1, the PDF will be supported by the Faculty and Teaching Development team from the Teaching and Learning Services at McGill University, who provides the PDF with professional development training, guidance, and support with course design, including effective teaching and assessment strategies. In Year 2, the PDF will be supported by the Teaching and Learning Support Services at the University of Ottawa. Support: The fellow will be encouraged to participate in a CANSSI-sponsored summer training program. In Year 1, the PDF will be encouraged to participate in workshops offered by the McGill Career Planning Service, which assists PDFs in career development, project management, networking, and job search. In Year 2, the fellow will be encouraged to take workshops related to career development provided by the Career Development Center at the University of Ottawa.

Plan for mentoring

There will be bi-weekly Zoom meetings with both supervisors and the PDF to present research updates. Given the short distance between Montreal and Ottawa, we will meet jointly in person at least four times a year. The PDF will be involved in the (informal) co-supervision of a graduate student and a summer undergraduate student (NSERC-USRA). In Year 2, we expect the PDF to spend some time at Statistics Canada with the small area estimation group. This collaboration will introduce the PDF to practical challenges, enhance their communication skills, and give them opportunities to share their expertise with non-specialists (data users at Statistics Canada). The PDF fellows will showcase their research at both national (e.g., the annual conference of the statistical society of Canada) and international conferences (e.g., the joint statistical meetings, the small area estimation conference), which will also serve as platforms for networking. Supervisor: David Haziza (Full Professor, University of Ottawa) will provide guidance on small area estimation. Co-supervisor: Mehdi Dagdoug (Assistant Professor, McGill University) will provide guidance on aggregation of machine learning procedures. There are weekly statistics seminars at McGill University and at the University of Ottawa. Further, there are monthly colloquiums held jointly between the four Montreal universities and the other Quebec universities. The PDF will be encouraged to attend the seminars. The PDF will also have the chance to participate to the ongoing reading groups within each department. M. Dagdoug is closely involved in the organisation of these events.

Proposed schedule

Research. In their first year, the PDF will work on extending the suite of population alarm functions.

Year 1:

The PDF will spend the first year of the fellowship at McGill University in Montreal, QC. Research: The PDF will focus on the development of aggregation procedures in the context of the Fay-Herriot model. During the first year, the postdoc will be expected to prepare at least one peer-reviewed research article focusing on the methodology of the proposed technique. The PDF will also attend conferences and workshops. Teaching: Teach a statistics course in the Department of Mathematics and Statistics at McGill University.

Supervision: The PDF will help co-supervise an undergraduate summer student at McGill University and/or a research project through undergraduate project courses. Development: EDI training, career planning training, teaching development seminars offered by Teaching and Learning Services at McGill University.

Year 2:

The PDF will spend the second year of the fellowship at the University of Ottawa in Ottawa, ON. Research: The PDF will focus on the estimation of the mean square error of the aggregate predictor. The PDF will also apply the proposed methods using data from Statistics Canada. The research outcomes are expected to result in 1-2 additional peer-reviewed research articles focusing on theoretical, computational, and applied issues. The PDF will present their work at conferences. Teaching: Teach a statistics course in the Department of Mathematics and Statistics at the University of Ottawa.

Supervision: The PDF will help in the co-supervision of an MSc student at the University of Ottawa. Development: career development and job search support -- career advising appointments and workshops at the Career Development Center at the University of Ottawa.

List of qualifications of suitable candidates

- PhD in Statistics or closely related field
- Knowledge and experience in one (or more) of the following areas: statistical learning, survey sampling, small area estimation
- Proficiency with programming languages including R or Python
- Ability to teach courses in statistics at the undergraduate level in English (or French as the University of Ottawa is a bilingual university)
- Strong communication skills and willingness to learn.

Research description

Research project Small area estimation (SAE) refers to a class of estimation procedures used to estimate characteristics or parameters of specific subpopulations, geographical areas, or "small area" within a larger region when direct survey data for that area are limited or unavailable. Small areas often have insufficient sample sizes to provide accurate estimates through the customary design-based estimation procedures. SAE procedures borrow strength from information available at higher levels of aggregation. In other words, data from a larger region or the overall population are used to help make more accurate estimates for smaller areas. This can involve incorporating auxiliary data, such as data from censuses or administrative records, to supplement survey data. A popular area level is the so-called Fay-Herriot (FH) model (Fay and Herriot, 1979) that is based on a linear link function. The FH model also requires the direct variance estimates, which are often unstable due to the small sample sizes. As a result, these variance estimates are often smoothed. In addition to point estimates, SAE aims to provide estimates of their precision to reflect the uncertainty associated with these estimates; see Rao and Molina (2015) for a comprehensive treatment of SAE. In recent years, machine learning (ML) methods have attracted some interest in the SAE field to enhance estimates' accuracy and efficiency.

These techniques leverage advanced algorithms and predictive models to extract valuable insights from auxiliary data sources and large-scale surveys. In SAE, ML methods can be used to identify complex relationships between covariates and the characteristics of small areas, allowing for more accurate estimates; see, e.g., Viljanen et al. (2022), Krennmair and Schmid (2021). ML methods are attractive if the classical assumptions underlying the FH model are not met. For a given SAE problem, we may consider several ML candidates such as regression trees, boosting, random forests, etc. Each of these candidates would lead to a different SAE estimator. Selecting one estimator from the list of candidates may be challenging as no ML algorithm is universally superior to the others in all the scenarios. The choice of a candidate must, therefore, be made in a data-dependent fashion. We may choose one candidate from the list (model selection) or construct a new estimator based on several candidates (model aggregation). This project focuses on model aggregation, whereby the estimators produced by each of the ML methods are combined using a convex combination. The problem of aggregation has been studied for a wide variety of statistical problems, such as density estimation (e.g., Yang, 2000), classification (e.g., Guedj and Alquier, 2013) and regression (e.g., Wegkamp, 2000). To the best of our knowledge, aggregation procedures in SAE have yet to be investigated. Research goals and possible approaches The objective of the present proposal is to develop practically relevant and theoretically sound aggregation methodologies for small area estimation. Aggregation procedures are typically implemented by using either linear or exponential weighting. We plan to explore both methodologies.

Theoretically, we will work towards establishing (exact) oracle inequalities stating that the linear (or, convex) aggregated small area (SA) estimator has an L2 loss lower or equal than that of the best linear (respectively, convex) combination of the candidates, up to a small remainder term. These inequalities have the advantage of being non-asymptotic and typically require using concentration inequalities. We also plan to investigate to investigate the issue of mean square error prediction. This may require using resampling procedures such as the bootstrap. The proposed procedures will implemented in an R-package that will include various aggregation procedures and predictors of their mean square error, among others. Potential applied and theoretical impact With the large number of new statistical learning algorithms available to practitioners, it has become increasingly challenging to decide which one to use in practice. New learning algorithms often also require the choice of several hyperparameters. Modelling has thus become a challenging task in practice, as a wrong choice of hyper-parameters or estimators may lead to a loss of efficiency. Developing a theoretically sound estimation strategy based on aggregation will allow end-users to fit several candidates they deem appropriate and directly use an aggregation procedure to build an efficient estimator, thus reducing the burden of modelling and the risk of invalid inferences. Also, as an aggregated estimator aims to be at least as efficient as the best combination of the candidates, it will likely result in improved efficiency of SAE procedures. Resources needed, location, and role of the supervisors The PDF will spend one year working at McGill University (working with M. Dagdoug) and one year in uOttawa (working with D. Haziza). David Haziza is an expert on survey sampling, and Mehdi Dagdoug has expertise in statistical learning procedures. We will provide the PDF with a workstation in both universities. Moreover, if some parts of the project become computationally intensive, the PDF will be granted access to the computing facilities available at McGill University and uOttawa.

References Guedj, B., & Alquier, P. (2013). PAC-Bayesian Estimation and Prediction in Sparse Additive Models. Electronic Journal of Statistics 7, 264-291. Fay, R.E. and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. Journal of the American Statistical Association 74, 269-277. Krennmair, P. and Schmid, T. (2022). Flexible domain prediction using mixed effects random forests. Journal of the Royal Statistical Society C 71, 1865-1894. Rao, J.N.K. and Molina, I. (2015). Small Area Estimation. Wiley & Sons. Viljanen, M., Meijerink, L., Zwakhals, L and van de Kassteele, J. (2022). A machine learning approach to small area estimation: predicting the health, housing and well-being of the population of Netherlands. International Journal of health Geographics 21, 1-18. Wegkamp, M. (2003). Model selection in nonparametric regression. The Annals of statistics 31, 252-273. Yang, Y. (2000). Mixing strategies for density estimation. The Annals of statistics 28, 75-87.