Paving the way forward for a modern and responsive national statistical office

Keywords: non-probabilistic, quality, confidentiality, small area estimation, machine learning.

1. INTRODUCTION

Faced with today's data revolution, many national statistical offices (NSO) are modernising their programs to be more efficient and responsive to users' needs. A strong statistical research and development (R&D) program is one way to fuel evidence-based decision-making and a key enabler of modernisation. Our agency's methodology R&D program is geared towards providing solutions to current issues and challenges, identifying or developing sound theoretical frameworks and exploring new areas which could be beneficial in the medium to long term future. This paper will present a high-level view of our current research priorities and recent achievements. In particular, concrete examples in five specific areas will be showcased. The paper will also briefly discuss how the R&D program is evolving to address the ever-changing demand landscape, and how this impacts the forthcoming research needs.

2. RESEARCH PRIORITIES AND RECENT ACHIEVEMENTS

2.1. On the use of non-probabilistic data source in a scientifically rigorous framework

Supported by the fundamental piece by Neyman (1934) [1] and its continuous development in the 20th century described in Rao (2005) [2], the use of probabilistic surveys has been the preferred tool of most NSOs to answer information needs. In light of response rate declines, high collection costs and the proliferation of alternative data sources, a trend towards the exploration of non-probabilistic approaches is seen throughout the world. In 2018, our agency conducted a review [3] of the use of non-probabilistic data sources in a scientifically rigorous framework. The goal of this review was aimed at addressing the following questions:

(a) In which context can data from a non-probabilistic source replace adequately a probabilistic survey?

(b) In cases where full replacement is not advisable, how can the non-probabilistic data source be used to minimise response burden and collection cost of a probabilistic survey within a valid statistical inference framework?

Current solutions, potential leads and food for thought presented in [3] are split into two groups; design-based approaches and model-based approaches. A high level summary of those approaches will be presented.

2.2. Defining, measuring and communicating quality in a multi-source environment

The methods and language for measuring and communicating data quality mostly originate from sampling theory. While the basic concepts were extended to take into consideration non-sampling errors in the Total Survey Error Framework, the question of how to fully define and measure quality in the new/big data paradigm in a multi-source environment is still open. Our research work into this foray is multi-fold. It includes (a) exploring the theoretical framework (some of which is addressed by the work covered in the project described in (2.1)) and (b) addressing the immediate needs of data users and producers. Inspired by the nutritional label seen of food items, our agency is drafting a quality informational label to help users assess the quality of a wider scope of statistical and experimental products. The research work builds on the well-known dimensions of quality in the context of official statistics and, amongst others goals, aims to develop a whole-of-government data quality framework. To the extent possible, the results of consultations on data users' desires and data producers' current practices will be shared in the presentation, along with the initial quality informational label.

2.3. Access, privacy and confidentiality

As is the case in many other NSOs, our agency is actively involved in various activities to increase the access to information to feed a data-driven society while preserving the values of privacy that citizens cherish. Some solutions revolve around digital technologies, operational or legislative approaches. This part of the presentation will rather focus on recent methodological achievements. Two explicit use-cases will be highlighted where more data could be made available to the public while preserving the confidential nature of the data. The first example makes use of a disclosure control model based largely on Bayesian decision theory to provide an alternative to suppression in aggregated tabular data [4]. The method is scheduled to be used in production for the first time in December 2018. In the second example, a fully synthetic data set with appropriate analytical value was created and used for a special production project in September 2018 [5]. Lessons learned and future work involved with the two projects will be shared.

2.4. High-definition data and small area estimation

The need for data at local or small targeted domains is well documented in a decision-driven society. Small area estimation techniques are one way to address this challenge. The reader is referred to [6] for a great overview of the methods. In practice, the model from [7] is amongst the most used for real-life applications. In the recent years, our agency has been developing the tools to apply such models in our production environment. With this hands-on experience, various strategies and best practices related to small area estimation were developed and will be shared.

2.5. Use of machine learning to improve the statistical production process

As part of our agency's modernisation initiative, machine learning approaches are explored to enhance or replace various steps in the statistical business process if efficiency or accuracy can be improved. The use of machine learning techniques in automatic coding appears to be a natural fit. In the last few months, a large number of exploratory projects aiming to incorporate machine learning in our processes have been initiated; many of these are in the context of text classification. A Community of Practice (CoP) for text classification and natural language processing was launched in January 2018. This CoP offers a place for practitioners to share their results, successes and challenges. Current

challenges and on-going exploratory work [8] to address them will be presented. They revolve around the creation of appropriate training sets for supervised machine learning, tuning of model hyperparameters in the context of limited computing environment and quality monitoring in real-life ongoing production. This work is also used as a platform to explore the use of open-source software in our production environment.

3. CONCLUSIONS

This paper and presentation will showcase recent progress and achievements in the development and use of new techniques for the production of official statistics in our organisation. In each case, concrete high-level examples will be summarised and/or referenced.

REFERENCES

[1] J. Neyman, On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society 97 (1934), 558-625.

[2] J.N.K. Rao, Interplay between sample survey theory and practice: an appraisal, Survey Methodology 31, (2005), 117-138.

[3] *Omitted for blind review*

[4] *Omitted for blind review*

[5] *Omitted for blind review*

[6] J.N.K. Rao and I. Molina, Small Area estimation (2015), Second Edition, Wiley, Hoboken, NJ.

[7] R.E. Fay, and R.A. Herriot, Estimates of income for small places: An application of James-Stein procedures to census data. Journal of the American Statistical Association 74, (1979), 269-277.

[8] *Omitted for blind review*