EUROPEAN MASTER IN OFFICIAL STATISTICS
EMOS

Abstract
EMOS Master thesis competition 2025

'Methods for integrating survey data and big non-survey data'

Author: **Elena Viti, University of Pisa & University in Trier**

Keywords: Data integration, probability sampling, big data

1. INTRODUCTION

This thesis addresses the critical challenge of integrating data from diverse sources, focusing on the combination of probability samples and big data. Data integration offers significant advantages, including cost reduction through the supplementation of existing surveys and enhanced accuracy of estimates by mitigating the limitations of individual data sources. Crucially, combining information from multiple surveys can be beneficial for addressing both sampling and non-sampling errors. By strategically using one survey to compensate for information lacking in another, we can improve the precision and reliability of estimates. This is particularly relevant when dealing with finite populations where single surveys may not provide sufficient data to accurately measure certain phenomena. The choice of appropriate integration methods, however, depends heavily on the type of sample considered. Probability samples, with known selection probabilities, allow for rigorous design-based inference, while big data, frequently a non-probability sample, poses distinct challenges concerning representativeness and data quality. This research systematically explores these scenarios: the integration of multiple probability samples, and the integration of probability samples with big data, examining how methodological choices must adapt to these varying data characteristics.

2. OBJECTIVE

The research aims to: 1) Compare existing integration techniques for probability samples using macro and micro approaches; 2) Develop a framework for integrating probability samples with big data, adapting existing calibration methods to handle challenges like measurement error and duplicate data; 3) Evaluate the performance of these approaches through simulation studies using synthetic data, investigating the impact of variable selection and data characteristics on estimation accuracy.

3. METHODS

The study systematically compares two established approaches for probability sample integration: a macro approach, which aggregates summary statistics from multiple surveys, and a micro approach, which creates a single synthetic dataset by combining individual-level data. For the macro approach, the generalized regression estimator (GREG) is employed, exploring both proportional and optimal weighting strategies September 2024based on sample sizes and the inclusion of control variables. The micro approach utilizes multiple imputation techniques to address missing data within the probability samples. The integration of probability samples with big data is addressed using a framework that adapts existing calibration weighting techniques. This framework does not introduce a new calibration method but rather presents a novel application of existing methods to the specific challenges posed by combining probability samples with big data. The framework addresses non-probability sampling by treating the big data as a separate population. The calibration process is modified to accommodate the frequent presence of measurement errors and duplicate entries in big datasets. A regression-based calibration method was implemented, adjusting weights to align sample totals of auxiliary variables with known population totals where possible. The methodology uses the information in big data to improve estimates from the probability sample by creating a new variable (δ_i) which indicates whether a unit i from sample A is also present in big data B. This allows calibration to improve the estimation from probability sample A by incorporating data from big data B, which may contain information not present in sample A. The AMELIA synthetic dataset is used to generate a series of simulations evaluating the performance of these integration approaches across varied conditions, such as differing sample sizes, data quality (presence of duplicates and measurement errors), and the number and type of auxiliary variables employed in the calibration. The simulations also explore scenarios involving monotone and non-monotone missing data patterns in the probability samples.

4. RESULTS

The simulation results demonstrate the relative strengths and weaknesses of the different integration methods. For the integration of multiple probability samples, the macro approach, employing the GREG estimator, proved most efficient when common variables exhibited strong correlations with the target variables. The micro approach, while effective in creating a unified dataset, showed sensitivity to inconsistencies in data quality across the different probability samples. Regarding the integration of probability samples with big data, the results highlight the significant influence of data quality (presence of duplicates and measurement errors) and the choice of auxiliary variables on estimation accuracy. The framework employing calibration weighting generally improved estimates compared to relying solely on the probability sample, particularly when known population totals for auxiliary variables were available. However, scenarios with a high proportion of duplicates or significant measurement error within the big data yielded less precise estimates.

5. CONTRIBUTION

The use of big data sources in estimation is nowadays a relevant topic also in the content of official statistics. However, from a methodological point of view, it is important to evaluate the impact of

the use of big data in the estimation process. In the thesis the focus was on evaluating the impact of the use of big data in context of integrating data from diverse sources. The thesis results highlight that the effectiveness of the integration framework was shown to be highly dependent on the extent to which the auxiliary variables in the big data captured variation in the target variable not already explained by the probability sample. The simulations reveal a September 2024trade-off: including more auxiliary variables improves accuracy when known population totals are available but can lead to instability if correlations become too strong or if totals are unknown. The optimal choice of auxiliary variables and the robustness of the integration method is heavily influenced by the quality and characteristics of the big data itself. Therefore, the thesis results suggest that the including big data in the estimation process should be carefully evaluated.

6. REFERENCES

[1] Yang, Shu and Jae Kwang Kim (2020). “Statistical data integration in survey sampling: A review”. In: Japanese Journal of Statistics and Data Science 3.2, pp. 625–650.

[2] Tam, Siu-Ming and Frederic Clarke (2015). “Big data, official statistics and some initiatives by the Australian Bureau of Statistics”. In: International Statistical Review 83.3, pp. 436–448.

[3] Yang, Shu and Peng Ding (2019). “Combining multiple observational data sources to estimate causal effects”. In: Journal of the American Statistical Association.

[4] Kim, Jae-Kwang and Siu-Ming Tam (2021). “Data integration by combining big data and survey sample data for finite population inference”. In: International Statistical Review 89.2, pp. 382–401