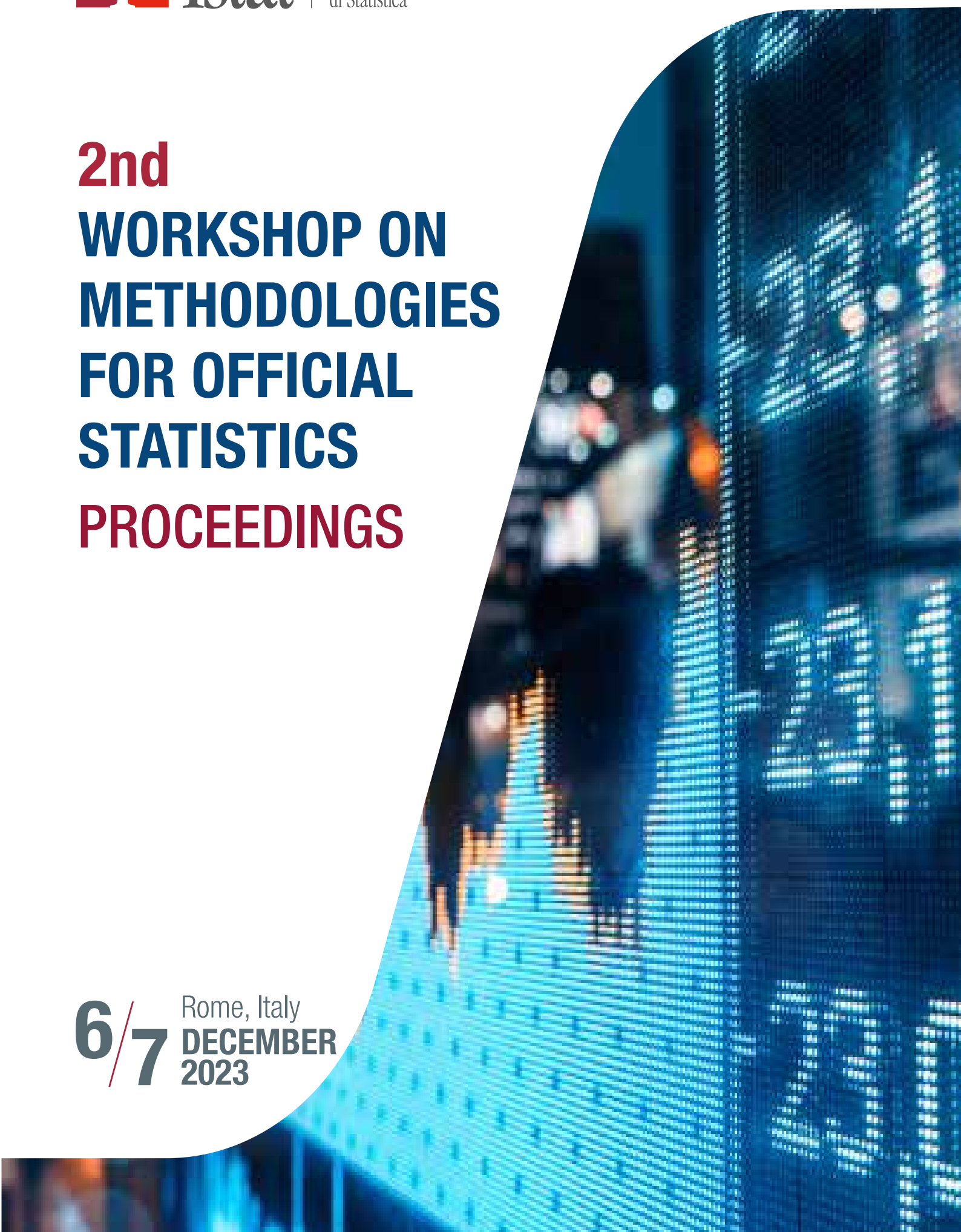
Istat | Istituto Nazionale di Statistica

2nd WORKSHOP ON METHODOLOGIES FOR OFFICIAL STATISTICS

PROCEEDINGS



6/7 Rome, Italy DECEMBER 2023

2nd Workshop on Methodologies for Official Statistics Proceedings

ISBN 978-88-458-2149-3

Istituto Nazionale di Statistica
Via Cesare Balbo, 16 – Roma

Index

Pag.

Page

SESSION 3
Quality for non-traditional sources

SESSION 4
Machine Learning Methods in Survey Statistics

CLOSING SESSION

Welcome by the President of the Italian National Institute of Statistics - Istat

Francesco Maria Chelli[1]

Dear guests, and dear colleagues,

It is a great pleasure for me to welcome all the participants, both in-person and online, in this second Istat's *Workshop on Methodologies for Official Statistics*.

Being focussed on methods, this workshop is of utmost importance for Istat. Good methods make data of good quality, and good quality is the root of the Institute's good reputation. I am confident that the workshop will grant us enough favourable time to go into the details of the problems encountered in the statistical production processes, as well as to explore new approaches for satisfying the informative needs of our stakeholders with reliable solutions.

This is confirmed by the themes selected for discussion: new data sources and new methods, especially in the data-mining context. This is a new frontier for National Statistical Institutes: depending on the results discussed on occasions like this one, the methods discussed today and tomorrow could become common in the future. Hence, our expectations on what will be presented and debated during these two days are high.

This workshop is the result of the efforts of several people. One month ago, I signed the establishment of a new Advisory Committee on Statistical Methods. The Committee members will be ubiquitous in this workshop, as chairs, discussants and invited speakers. The Committee was established in 2017, and since then it supported more than 50 Istat methodological research projects. The first projects are now used for statistical production, as those related to setting up registers and the integrated use of different sources.

I am thankful to all the members of the Advisory Committee on Statistical Methods, who are leader statisticians who work on different methodological research areas in academic institutions and National Statistical Institutes, and I appreciate their support.

I wish to thank the coordinator of this Committee, Professor Daniela Cocchi, who is going to chair this committee for the seventh year and the next three years to come.

Let me mention all the Committee members, starting from those who served in the last term: Professor Natalie Shlomo (University of Manchester), Professor Maria Giovanna Ranalli (University of Perugia), Professor Li-Chun Zhang (University of Southampton, and Statistics Norway), Professor Brunero Liseo (Sapienza University of Rome), Piero Falorsi (formerly at Istat).

A very warm welcome to the new members of the Committee: Professor Marco Alfò (Sapienza University of Rome), Professor David Haziza (University of Ottawa), and Professor Piet Daas (Statistics Netherlands and Eindhoven University of Technology).

I take this opportunity to thank in particular Professor Shlomo, who is the President of the International Association of Survey Statisticians. We are very proud that one of this workshop sessions is jointly organised with IASS and thank you for the help and support in organising this event.

I would like to thank also all those who will present their research results in the different

1 Francesco Maria Chelli (presidente@istat.it), Italian National Institute of Statistics - Istat.

sessions. The programme includes three master classes by Professor Changbao Wu (University of Waterloo, Canada), Professor Stefano Maria Iacus (Harvard University) and Fabio Ricciato (Eurostat).

The workshop will be fuelled by speakers from many different institutions: Istat Italy, U.S. Census Bureau, INSEE France, LUMSA University Italy, Statistik Austria, and Central Bureau of Statistics Ireland.

I am confident that our workshop will consolidate research partnerships, and favour the creation of new teams, to work together on the issues raised in these days.

I wish you a very good workshop.

The path taken by Istat on the use of innovative sources: some methodological issues

Massimo Fedeli[2]

Istat and, more generally, Statistical Institutes are asked nowadays to make an enormous effort to take pictures, and give data and information on a reality that is changing faster and faster. The demand for information from stakeholders is (fortunately) increasingly high and varied in its forms. Parallel to these demands, there is the possibility of having a hitherto unprecedented amount of information available, think for example of all the information on the web, from satellite, or even from smart devices. These ingredients seem to suggest a marriage that would thus enable a positive response to the needs illustrated.

As representative examples of this context in Istat, we can, for instance, mention the studies initiated on the use of remote sensing to provide new information, improve existing information, or even make the production process more efficient. The case of using remote sensing for urban green estimation will be illustrated in this workshop. Another Istat project, in line with Eurostat's innovation agenda, involves using Automatic Identification System (AIS) data to enhance the timeliness of maritime statistics production and enrich information through techniques such as network analysis. Still, Istat is investing in work that exploits sentiment analysis methodologies to give a representation with high timeliness of people's sentiment on some important issues. In addition to the world of Trusted Smart Statistics (TSS), to which the above examples belong, there is also the use of integrated sources, often of a different nature such as administrative sources and sample surveys, which is now inevitable and is of particular relevance in Istat's production processes.

However, for a perfect marriage, several issues must be resolved. Among the others, methodological issues are of particular importance.

The introduction of new elements into the statistical production processes to complement classical ones, based on sample surveys or to be used to provide new solutions, implies a deep reflection on the issues that will be raised and discussed in this workshop.

A first question is concerned with methodologies and designs for multi-source processes with non-probability data. This issue is critical because new data sources, which may be affected by various errors, can often be made statistically more valid when combined with survey data over which there is full control. To this end, however, techniques still need to be developed to allow for these integrations and the necessary corrections for the use of non-sample data.

The generally unstructured nature of big data and their large volume naturally lead to the use of Machine Learning techniques. Their use, however, in the field of Official Statistics needs to be further investigated since they have been developed in similar, yet different contexts. For example, whether and how to take into account the characterising elements of a survey data set in the case of integration with non-probabilistic data is still a sensitive issue.

Another important aspect characterising the production of Official Statistics concerns the quality of the information produced. For a Statistical Institute the situation in which information is disseminated without having assessed the quality of that information cannot exist. It is a

2 Massimo Fedeli (fedeli@istat.it), Italian National Institute of Statistics - Istat.

concept inherent in the word “Official” Statistics. However, the main measures of quality, especially concerning accuracy, have been developed in a context that basically refers to the world of sample surveys. The introduction of innovative sources therefore entails an investment in this issue as well.

The importance of this workshop on Methodologies for Official Statistics stems from these premises. I am confident that we will gain useful information to continue on the path taken by the Italian National Institute of Statistics profitably.

Official Statistics reflect life

Monica Pratesi[3]

Statistical production is experiencing a major change in Official Statistics. The impact of digitisation on the economic and social spheres of society imposes a continuous learning attitude: Official Statistics reflect life.

In a "datafied" world like the one we live in, it is natural to exploit alternative forms of data collection: satellite imagery, data describing transactions (financial transactions: orders, invoices, payments; logistic transactions: deliveries, storage records, travel records), web data, mobile network operator data, electronic invoices (B2B e-invoicing mandatory for companies).

ESS and NSIs are working towards the (re)use in the statistical production processes of new sources of data, including data generated and held by the private sector (privately held data). An example for all: the location data, which are routinely collected by Mobile Network Operators (MNO data), are one of the most appealing candidate sources for (re)use in Official Statistics, but also one of the most challenging to enabling the production of statistics, delivering a dynamic view of population presence and mobility.

However, this mine of information hides pitfalls in the context of official statistical production. They are data of various structuring or rather we could say unstructured, designed data and found data, which therefore require adaptation, transformations and processing aimed at their use for statistical production purposes. Their integrated use with surveys or administrative data is often of fundamental importance to maximise their information potential to support or complement current statistics.

This is the path, which can no longer be postponed.

At the same time, there is another face of the same coin to consider: engaging with respondents in surveys and protecting data collected for statistical purposes under the GDPF (*i.e.* General Data Protection Framework). I do not enter privacy issues here. I only say that we need to better explore the use of Apps and smart devices in smart surveys. Increasingly low response rates can be contrasted by reusing administrative data, to lower the respondent burden, but contact policies need a revision. How can we communicate effectively with respondents in this digitisation era? We need to enhance our response rate by implementing a new contact policy management. This involves better engagement with local institutions and Civil Society Organisations, as well as providing respondents with smart devices, similar to Citizen Science experiments. Smartphones can automate data collection and incorporate many important data-gathering functions - such as capturing images, audio and text - into a single tool that can stamp the date, time and geographic coordinates associated with an observation. Mobile applications for smartphones, tablets and other gadgets can turn just about anyone into a citizen scientist/ co-creator of data.

Coming to the programme of the workshop, I am pleased to note that many of the communications foreseen recall production themes: population census, price index, mobile data, urban green, employment. This combination of methodological research and thematic innovation is fundamental for the health of the Institute, the integration between the skills of

3 Monica Pratesi (monica.pratesi@istat.it), Italian National Institute of Statistics - Istat.

the production sector and those of methodologies is an essential element and we cannot survive one without the other. Alongside these, I also note, with equal pleasure, the presence of master classes held by important international researchers, which are useful for highlighting critical issues and perspectives relating to these frontier issues, and for inserting Istat into a context of comparison and international discussion.

I close by spending some words on quality. The evaluation of data quality in an increasingly multi-source context, in which non-traditional and non-probabilistic sources are also used, and therefore new methods such as Machine Learning, is by no means a foregone conclusion. Once again, theoretical issues such as the definition of the concept of quality have a practical and operational impact in the context of Istat production.

We work producing experimental statistics: the transition from experimental to Official Statistics requires a rigorous evaluation of quality. It is therefore necessary to develop new instruments that help measure and communicate to stakeholders the quality of the official data produced. The main difficulty lies not in producing data, but in measuring the natural uncertainty associated with them.

Concluding, I am sending five messages to contribute to the workshop's discussion:

1. Quality is in the eye of the beholder. My vision is that the Code of Practice needs a revision, the last one was in 2017.
2. Novelties or innovation? New methods are not only statistical methods. The Generic Statistical Business Process Model (GSBPM) with its phases is old-fashioned: Civil Society Organisations are to be included, as well as privately held data, with a better focus on the so-called Citizen Statistics and Citizen Generated Data.
3. Uncertainty is here to stay.
4. Environment, social responsibility of companies, circular economy, pandemics, and climate changes urge for new survey methods and better integration of administrative data files.
5. There is no difference between theory and practice, but in practice there is. This is why research in Official Statistics is in quest of sustainable, user-friendly methods, with a prompt translation in processes to produce meaningful data on current phenomena.

SESSION 1

Methodologies and designs for multi-source processes with non-probability data

Session 1
Master class | Challenges and strategies in dealing with non-probability survey samples

Changbao Wu[1]

Abstract

Statistical analysis of non-probability survey samples faces three major challenges: the unknown sample participation mechanism; the unknown population represented by the sample; and the dearth of suitable internal and external data for valid and efficient estimation. We discuss strategies proposed in the recent literature and the strengths and weaknesses of these methods in dealing with specific challenges.

Keywords: Calibration techniques; Double robustness; Inverse probability weighting; Participation probability; Pseudo maximum likelihood; Undercoverage.

1. Introduction

The term "biased sample" is often associated with non-probability survey samples and is used as an indication that statistical analysis with non-probability survey samples is a difficult task. In real world, almost all samples are biased, including probability survey samples for which the simple sample mean is not a valid estimate of the population mean unless the sample is selected by simple random sampling. The biased nature of probability survey samples, however, has never been a major issue in design-based inference since estimation biases can be corrected through suitable weighting using the known sample inclusion probabilities. The Horvitz-Thompson estimator (Horvitz and Thompson 1952), commonly known as the HT estimator, was developed in survey sampling and has been one of the main pillars of design-based inference for probability samples. The method was also independently proposed by Narain (1951), which led to the argument by Rao (2005) that the "NHT estimator" is a more suitable term. The NHT estimator deserves more explicit credits as a general tool for the broad field of statistics. It has been widely used in missing data analysis and causal inference as the inverse probability weighted (IPW) estimator.

The first major challenge in dealing with "biased" non-probability survey samples is the unknown sample participation/inclusion mechanism. A natural starting point is to assume that there is an underlying probability model, denoted as q, which guides the sample participation process. There is no guarantee that such a model exists but starting with an assumption is what statisticians always do to find approximate solutions to the otherwise unsolvable problems.

[1]Changbao Wu (cbwu@uwaterloo.ca), University of Waterloo, Canada. This research is supported by the Natural Sciences and Engineering Research Council of Canada and the Canadian Statistical Sciences Institute.

Let $\mathcal{U} = \{1, 2, \ldots, N\}$ be the target finite population of N distinct units. Let y_i and $\boldsymbol{x}_i$ be the values of the study variable y and the vector $\boldsymbol{x}$ of auxiliary variables associated with unit i. The population level data can be represented by $\{(y_i, \boldsymbol{x}_i), i \in \mathcal{U}\}$. Let $\mathcal{S}_A$ be the set of n_A units included in the non-probability sample. Let $\{(y_i, \boldsymbol{x}_i), i \in \mathcal{S}_A\}$ be the dataset for the non-probability sample. Let $R_i = I(i \in \mathcal{S}_A)$ be the indicator of unit i being included in the non-probability sample $\mathcal{S}_A$, where $I(\cdot)$ is the indicator function, defined for all units in $\mathcal{U}$, *i.e.* $i = 1, 2, \ldots, N$. We define the probability of unit i participating in $\mathcal{S}_A$ as $\pi_i^A = P(R_i = 1 \mid y_i, \boldsymbol{x}_i) = E(R_i \mid y_i, \boldsymbol{x}_i)$. The π_i^A is termed by some researchers as the "propensity score"; see, for instance, Valliant and Dever (2011), Chen, Li and Wu (2020), Wang, Valliant and Li (2021), Kim and Morikawa (2023), among others. But many recent papers on non-probability survey samples seem to prefer the use of "participation probability" (Beaumont 2020; Rao 2021; Wu 2022a). The joint distribution of $(R_1, R_2, \ldots, R_N)$ is guided by the assumed participation probability model q, and the non-probability sample is uniquely determined as $\mathcal{S}_A = \{i \mid R_i = 1 \text{ and } i \in \mathcal{U}\}$. The dependence of the participation probability on the underlying model q is often explicitly expressed as $\pi_i^A = E_q(R_i \mid y_i, \boldsymbol{x}_i)$. It is apparent that $\sum_{i=1}^N R_i = n_A$.

Wu (2022a) discussed three commonly used assumptions for the model q and the participation probabilities. These assumptions can sometimes be justified in practical applications but none can be rigorously tested using the non-probability survey sample.

- **A1** The sample participation indicator R_i and the study variable y_i are independent given the set of covariates $\boldsymbol{x}_i$, *i.e.* $(R_i \perp y_i) \mid \boldsymbol{x}_i$.
- **A2** All the units in the target population have non-zero participation probabilities, *i.e.* $\pi_i^A > 0, i = 1, 2, \cdots, N$.
- **A3** The indicator variables $R_1, R_2, \cdots, R_N$ are independent given the set of auxiliary variables $(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N)$.

Assumption **A1** is similar to the missing at random (MAR) assumption for missing data analysis. Under **A1**, we have $\pi_i^A = P(R_i = 1 \mid \boldsymbol{x}_i, y_i) = P(R_i = 1 \mid \boldsymbol{x}_i) = \pi(\boldsymbol{x}_i)$ for some unknown function $\pi(\cdot)$. Assumption **A2** can be violated in practice, leading to undercoverage problems. Assumption **A3** is for the convenience of forming a likelihood function and is not crucial for the validity of estimation of participation probabilities.

One of the key features of non-probability survey samples is the absence of sampling design for the selection of units. However, "survey design" remains an important aspect of non-probability samples. Justification of assumption **A1**, for instance, relies heavily on the knowledge of potential sample participation behaviours and on identifying key auxiliary variables to be included for data collection. It is also well understood that estimation of $\pi_i^A = \pi(\boldsymbol{x}_i)$ under assumption **A1** requires auxiliary information from the target population. The ideal scenario is that the complete auxiliary information $(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N)$ is available. A more practical scenario is that auxiliary information can be obtained from an existing probability survey.

- **A4** There exists a probability survey sample $\mathcal{S}_B$ of size n_B with information on the auxiliary variables $\boldsymbol{x}$ (but not on y) available in the dataset $\{(\boldsymbol{x}_i, d_i^B), i \in \mathcal{S}_B\}$, where d_i^B are the design weights for the probability sample $\mathcal{S}_B$.

The $\mathcal{S}_B$ is called the reference probability survey sample. The most crucial part of assumption **A4** is that the set of auxiliary variables $\boldsymbol{x}$ is observed in both the non-probability sample $\mathcal{S}_A$ and the probability sample $\mathcal{S}_B$. A reference probability survey sample is often available in practice, but it is rare that a particular reference probability sample contains all the important auxiliary variables required for assumption **A1**. How to combine information from multiple existing sources is another challenge for analysing non-probability survey samples. The two-sample setup with $\mathcal{S}_A$ and $\mathcal{S}_B$ was first introduced by Rivers (2007) on sample matching. Discussions in the rest of the paper are under assumptions **A1**-**A4** except for Section 3 where assumption **A4** is relaxed and information from multiple existing probability survey samples can be combined.

2. Estimation of Participation Probabilities

The first challenge facing the analysis of non-probability survey samples is the estimation of unknown participation probabilities. There are three parametric methods frequently cited in the recent literature on non-probability survey samples: the method of Valliant and Dever (2011) based on the pooled sample, the pseudo maximum likelihood method of Chen, Li and Wu (2020), and the method of Wang, Valliant and Li (2021) using a two-step computational procedure. The three methods hereafter are referred to as VD2011, CLW2020, and WVL2021, respectively. In this section, we provide theoretical comparisons among the three methods using the general estimating functions theory, and discuss their conceptual and computational differences. It is shown that the method of VD2011 leads to invalid results unless the non-probability sample is a simple random sample or the sampling fraction is negligibly small, and the method of WVL2021 is sub-optimal as compared to the method of CLW2020.

2.1 The method of CLW2020

It was stated in Section 1 that the joint distribution of $(R_1, R_2, \ldots, R_N)$ is guided by the assumed participation probability model q, which further defines the participation probabilities. It is where conceptual differences among the three methods can clearly be identified. Consider a parametric model q with $\pi_i^A = P(R_i = 1 \mid \boldsymbol{x}_i) = \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is the vector of unknown model parameters and $\pi(\cdot, \cdot)$ has a known functional form. The full likelihood function is given by $L(\boldsymbol{\alpha}) = \prod_{i=1}^{N} (\pi_i^A)^{R_i} (1 - \pi_i^A)^{1-R_i}$, which leads to the full log-likelihood function $\ell(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \{R_i \log(\pi_i^A) + (1 - R_i) \log(1 - \pi_i^A)\}$, with the more computationally friendly version given by

$$\ell(\boldsymbol{\alpha}) = \sum_{i \in \mathcal{S}_A} \log\{\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})\} + \sum_{i \in \mathcal{U} \setminus \mathcal{S}_A} \log\{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})\}\,. \tag{1}$$

It is apparent from (1) that estimation of the model parameters $\boldsymbol{\alpha}$ requires $\boldsymbol{x}_i$ for all units i from the entire target population $\mathcal{U}$.

Under the two-sample framework with the availability of a reference probability sample

$\mathcal{S}_B$, the pseudo log-likelihood function of Chen *et al.* (2020) is defined as

$$\ell_1(\boldsymbol{\alpha}) = \sum_{i\in\mathcal{S}_A} \log\left\{\frac{\pi(\boldsymbol{x}_i,\boldsymbol{\alpha})}{1-\pi(\boldsymbol{x}_i,\boldsymbol{\alpha})}\right\} + \sum_{i\in\mathcal{S}_B} d_i^B \log\{1-\pi(\boldsymbol{x}_i,\boldsymbol{\alpha})\}\,. \tag{2}$$

The first key feature of $\ell_1(\boldsymbol{\alpha})$ is that $E_p\{\ell_1(\boldsymbol{\alpha})\} = \ell(\boldsymbol{\alpha})$, where $E_p(\cdot)$ refers to expectation with respect to the probability sampling design for $\mathcal{S}_B$. In other words, the pseudo log-likelihood function $\ell_1(\boldsymbol{\alpha})$ is a legitimate likelihood function with the given samples $\mathcal{S}_A$ and $\mathcal{S}_B$. The pseudo score functions, defined as $\boldsymbol{U}_1(\boldsymbol{\alpha}) = \partial\ell_1(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$, are given by

$$\boldsymbol{U}_1(\boldsymbol{\alpha}) = \sum_{i\in\mathcal{S}_A} \frac{\pi_i'(\boldsymbol{\alpha})}{\pi(\boldsymbol{x}_i,\boldsymbol{\alpha})\{1-\pi(\boldsymbol{x}_i,\boldsymbol{\alpha})\}} - \sum_{i\in\mathcal{S}_B} d_i^B \frac{\pi_i'(\boldsymbol{\alpha})}{1-\pi(\boldsymbol{x}_i,\boldsymbol{\alpha})}\,, \tag{3}$$

where $\pi_i'(\boldsymbol{\alpha}) = \partial\pi(\boldsymbol{x}_i,\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$. The second key feature of the pseudo maximum likelihood approach of CLW2020 is that $E_{qp}\{\boldsymbol{U}_1(\boldsymbol{\alpha}_0)\} = \mathbf{0}$ for any smooth parametric forms $\pi(\boldsymbol{x}_i,\boldsymbol{\alpha})$, where $E_{qp}(\cdot)$ refers to expectation under the joint randomisation of q and p and $\boldsymbol{\alpha}_0$ denotes the true values of the model parameters such as $E_q(R_i \mid \boldsymbol{x}_i) = \pi(\boldsymbol{x}_i,\boldsymbol{\alpha}_0)$. The pseudo score functions are unbiased and are optimal under the current two-sample setup in the same spirit of Godambe (1960) for general estimating functions.

The maximum pseudo likelihood estimator $\hat{\boldsymbol{\alpha}}$ is obtained by solving the score equations $U_1(\boldsymbol{\alpha}) = \mathbf{0}$ and the maximum pseudo likelihood estimators of participation probabilities are computed as $\hat{\pi}_i^A = \pi(\boldsymbol{x}_i,\hat{\boldsymbol{\alpha}})$, $i \in \mathcal{S}_A$. The IPW estimator of the population mean $\mu_y = N^{-1}\sum_{i=1}^N y_i$ is computed as $\hat{\mu}_{yIPW} = \hat{N}_A^{-1}\sum_{i\in\mathcal{S}_A} y_i/\hat{\pi}_i^A$, where $\hat{N}_A = \sum_{i\in\mathcal{S}_A}(\hat{\pi}_i^A)^{-1}$.

2.2 The method of VD2011

The paper by Valliant and Dever (2011) was the first serious attempt in addressing estimation of participation probabilities under the two-sample setup. It inspired several follow-up papers, including Chen *et al.* (2020) and Wang *et al.* (2021). The proposed method was based on fitting a survey weighted logistic regression model to the pooled sample $\mathcal{S}_A \cup \mathcal{S}_B$ with the "binary response variable" defined as $D_i = 1$ if $i \in \mathcal{S}_A$ and $D_i = 0$ of $i \in \mathcal{S}_B$, for $i \in \mathcal{S}_A \cup \mathcal{S}_B$, assuming there are no overlaps between $\mathcal{S}_A$ and $\mathcal{S}_B$. Let $\hat{N}_B = \sum_{i\in\mathcal{S}_B} d_i^B$. Valliant and Dever (2011) defined the survey weights for the pooled sample $\mathcal{S}_A \cup \mathcal{S}_B$ as $w_i = 1$ if $i \in \mathcal{S}_A$ and $w_i = d_i^B(\hat{N}_B - n_A)/\hat{N}_B$ if $i \in \mathcal{S}_B$. It follows that the total weight over the pooled sample is $\sum_{i\in\mathcal{S}_A\cup\mathcal{S}_B} w_i = \hat{N}_B$. Fitting a survey weighted logistic regression model using the dataset $\{(D_i,\boldsymbol{x}_i,w_i), i \in \mathcal{S}_A \cup \mathcal{S}_B\}$ amounts to maximising the objective function

$$\ell_2(\boldsymbol{\alpha}) = \sum_{i\in\mathcal{S}_A} \log\{\pi(\boldsymbol{x}_i,\boldsymbol{\alpha})\} + \sum_{i\in\mathcal{S}_B} w_i \log\{1-\pi(\boldsymbol{x}_i,\boldsymbol{\alpha})\}\,. \tag{4}$$

The objective function $\ell_2(\boldsymbol{\alpha})$ can be viewed as an estimate of the full log-likelihood $\ell(\boldsymbol{\alpha})$ given in (1), where the second term in (4) is conceived as an estimator of the second term in

(1), since $\sum_{i\in\mathcal{S}_B} w_i = \hat{N}_B - n_A$, which matches the total number of units in $\mathcal{U}\backslash\mathcal{S}_A$. Unfortunately, the objective function $\ell_2(\boldsymbol{\alpha})$ is not a valid estimate of $\ell(\boldsymbol{\alpha})$ under general conditions. Furthermore, the functions $\boldsymbol{U}_2(\boldsymbol{\alpha}) = \partial\ell_2(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$ are given by

$$\boldsymbol{U}_2(\boldsymbol{\alpha}) = \sum_{i\in\mathcal{S}_A} \frac{\pi_i'(\boldsymbol{\alpha})}{\pi(\boldsymbol{x}_i,\boldsymbol{\alpha})} - \Big(1 - \frac{n_A}{\hat{N}_B}\Big) \sum_{i\in\mathcal{S}_B} d_i^B \frac{\pi_i'(\boldsymbol{\alpha})}{1-\pi(\boldsymbol{x}_i,\boldsymbol{\alpha})}\,, \tag{5}$$

which further leads to

$$E_{qp}\{\boldsymbol{U}_2(\boldsymbol{\alpha}_0)\} \doteq \sum_{i=1}^{N} \pi_i'(\boldsymbol{\alpha}_0) - \Big(1 - \frac{E_q(n_A)}{N}\Big) \sum_{i=1}^{N} \frac{\pi_i'(\boldsymbol{\alpha}_0)}{1-\pi(\boldsymbol{x}_i,\boldsymbol{\alpha}_0)}\,,$$

where the approximate equal sign amounts to omitting a high order negligible term in the equation. It follows that $E_{qp}\{\boldsymbol{U}_2(\boldsymbol{\alpha}_0)\} \neq \mathbf{0}$ under general conditions and the estimator $\hat{\alpha}$ obtained from solving $\boldsymbol{U}_2(\boldsymbol{\alpha}) = \mathbf{0}$ is not consistent for $\boldsymbol{\alpha}$. However, we do have $E_{qp}\{\boldsymbol{U}_2(\boldsymbol{\alpha}_0)\} \doteq \mathbf{0}$ under two scenarios: (i) n_A is fixed and $\pi(\boldsymbol{x}_i,\boldsymbol{\alpha}_0) = n_A/N$ for all i, *i.e.* the non-probability sample $\mathcal{S}_A$ is a simple random sample from the target population; and (ii) $E_q(n_A)/N = o(1)$. Scenario (i) usually does not occur in practice. Under scenario (ii), the "sampling fraction" is negligibly small, we typically have $\pi(\boldsymbol{x}_i,\boldsymbol{\alpha}_0) = o(1)$ and $1/\{1-\pi(\boldsymbol{x}_i,\boldsymbol{\alpha}_0)\} = 1 + o(1)$, uniformly over all i, which lead to $E_{qp}\{\boldsymbol{U}_2(\boldsymbol{\alpha}_0)\} \doteq \mathbf{0}$.

The choice of equal weights $w_i = 1$ for $i \in \mathcal{S}_A$ for the method of VD2011 implicitly assumes exchangeability among units in the non-probability sample, which is typically untrue for participation in non-probability survey samples. More importantly, the "binary response variables" D_i's are defined with the given $\mathcal{S}_A$ and $\mathcal{S}_B$, and are conceptually different from the sample participation indicators $(R_1, R_2, \ldots, R_N)$. The assumed participation probability model q does not lead to a meaningful interpretation of the joint distribution of $\{D_i, i \in \mathcal{S}_A \cup \mathcal{S}_B\}$. From a pure computational point of view, the estimated participation probabilities $\hat{\pi}_i^A = \pi(\boldsymbol{x}_i, \hat{\boldsymbol{\alpha}})$ with $\hat{\boldsymbol{\alpha}}$ solving $\boldsymbol{U}_2(\boldsymbol{\alpha}) = \mathbf{0}$ do not provide valid results unless it is scenario (i) or (ii) described above.

2.3 The method of WVL2021

The paper by Wang *et al.* (2021) provides a remedy to the method of Valliant and Dever (2011). Instead of pooling the two samples $\mathcal{S}_A$ and $\mathcal{S}_B$ together and rescaling the weights d_i^B to match the size of $\mathcal{U}\backslash\mathcal{S}_A$, the authors first created an artificial enlarged population $\mathcal{S}_A^* \cup \mathcal{U}$, where $\mathcal{S}_A^*$ consists of the same set of units in $\mathcal{S}_A \subset \mathcal{U}$ but these units are viewed differently in the union of $\mathcal{S}_A^*$ and $\mathcal{U}$. The authors then defined the indicator variable $\delta_i = 1$ if $i \in \mathcal{S}_A^*$ and $\delta_i = 0$ if $i \in \mathcal{U}$. The setting leads to the use of $\ell_3(\alpha) = \sum_{i\in\mathcal{S}_A} \log(p_i) + \sum_{i\in\mathcal{S}_B} d_i^B \log(1-p_i)$ as the likelihood function, where $p_i = P(\delta_i = 1 \mid \mathcal{S}_A^* \cup \mathcal{U})$. The authors' most critical argument is that the true participation probabilities π_i^A can be computed through the equation $\pi_i^A = p_i/(1-p_i)$.

The conceptual issues remain for the method of WVL2021, since the participation probability model q, once again, does not lead to meaningful interpretation of the joint distribution of $\{\delta_i, i \in \mathcal{S}_A^* \cup \mathcal{U}\}$. The latter is only conditionally defined with the given

$(R_1, R_2, \ldots, R_N)$. The authors suggested to assume a logistic regression model on the p_i, *i.e.* $p_i = \exp(\boldsymbol{x}_i^T\boldsymbol{\alpha})/\{1+\exp(\boldsymbol{x}_i^T\boldsymbol{\alpha})\}$, where $\boldsymbol{x}^T$ denotes the transpose of $\boldsymbol{x}$, which leads to a log-linear model for the true participation probability, *i.e.* $\pi_i^A = \exp(\boldsymbol{x}_i^T\boldsymbol{\alpha})$, a potential source of concerns for the estimated $\hat{\pi}_i^A = \exp(\boldsymbol{x}_i^T\hat{\boldsymbol{\alpha}})$ due to the range restriction on π_i^A.

The method of WVL2021 can be examined further from a computational point of view. Noting that $\pi_i^A = p_i/(1-p_i)$ leads to $p_i = \pi_i^A/(1+\pi_i^A)$ and $1-p_i = 1/(1+\pi_i^A)$, the method of WVL2021 for estimating the model parameters $\boldsymbol{\alpha}$ in $\pi_i^A = \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})$ is equivalent to maximising the objective function

$$\ell_3(\alpha) = \sum_{i\in\mathcal{S}_A} \log\left\{\frac{\pi(\boldsymbol{x}_i,\boldsymbol{\alpha})}{1+\pi(\boldsymbol{x}_i,\boldsymbol{\alpha})}\right\} - \sum_{i\in\mathcal{S}_B} d_i^B \log\{1+\pi(\boldsymbol{x}_i,\boldsymbol{\alpha})\}\,. \tag{6}$$

The function $\ell_3(\alpha)$ specified in (6) is not a valid replacement of the full log-likelihood function $\ell(\boldsymbol{\alpha})$ given in (1). The final estimator $\hat{\boldsymbol{\alpha}}$ is the solution to $\boldsymbol{U}_3(\boldsymbol{\alpha}) = \partial\ell_3(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha} = \boldsymbol{0}$, where

$$\boldsymbol{U}_3(\boldsymbol{\alpha}) = \sum_{i\in\mathcal{S}_A} \frac{\pi_i'(\boldsymbol{\alpha})}{\pi(\boldsymbol{x}_i,\boldsymbol{\alpha})\{1+\pi(\boldsymbol{x}_i,\boldsymbol{\alpha})\}} - \sum_{i\in\mathcal{S}_B} d_i^B \frac{\pi_i'(\boldsymbol{\alpha})}{1+\pi(\boldsymbol{x}_i,\boldsymbol{\alpha})}\,. \tag{7}$$

It turns out that $E_{qp}\{\boldsymbol{U}_3(\boldsymbol{\alpha}_0)\} = \boldsymbol{0}$ for any smooth parametric forms $\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})$. The estimating functions $U_3(\boldsymbol{\alpha})$ are unbiased under the joint randomisation of q and p, and the estimator $\hat{\boldsymbol{\alpha}}$ obtained from solving $\boldsymbol{U}_3(\boldsymbol{\alpha}) = \boldsymbol{0}$ is consistent for $\boldsymbol{\alpha}$. The method of WVL2021 is successful in correcting the biases in estimators of VD2011. It can be viewed as a special case of estimating equation based methods to be described briefly in Section 3. The final estimator $\hat{\boldsymbol{\alpha}}$, however, is sub-optimal compared to the likelihood based method of CLW2020. The structures of the two equations (6) and (7) for the method of WVL2021 resemble the two equations (2) and (3) for the method of CLW2020 but (6) does not approximate the pseudo log-likelihood function given in (2) and (7) differs from the pseudo score functions given in (3).

3. Calibration and Doubly Robust Estimation

Doubly robust estimators of μ_y are constructed using two working models: an assumed model q for the participation probabilities and an outcome regression model ξ for the response variable y given $\boldsymbol{x}$. Let $\hat{m}_i = m(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}})$ where $m(\boldsymbol{x}_i, \boldsymbol{\beta}) = E_\xi(y_i \mid \boldsymbol{x}_i)$ is the mean function under ξ with a known form $m(\cdot,\cdot)$ and $\hat{\boldsymbol{\beta}}$ is a suitable estimator of $\boldsymbol{\beta}$. The doubly robust estimator of μ_y proposed by Chen *et al.* (2020) is constructed as $\hat{\mu}_{yDR} = (\hat{N}_A)^{-1}\sum_{i\in\mathcal{S}_A}(y_i-\hat{m}_i)/\hat{\pi}_i^A + (\hat{N}_B)^{-1}\sum_{i\in\mathcal{S}_B} d_i^B\hat{m}_i$, where $\hat{N}_A = \sum_{i\in\mathcal{S}_A}(\hat{\pi}_i^A)^{-1}$ and $\hat{N}_B = \sum_{i\in\mathcal{S}_B} d_i^B$. The estimator $\hat{\mu}_{yDR}$ satisfies $E_q(\hat{\mu}_{yDR}) \doteq \mu_y$ under the participation probability model q, regardless of the model ξ, and $E_\xi(\hat{\mu}_{yDR} - \mu_y) \doteq 0$ under the outcome regression model ξ, irrespective of the model q. The estimator is doubly robust in the sense that it is consistent if one of the two working models is correctly specified. The estimator will only fail if both models are misspecified.

Estimation of participation probabilities is the most crucial part of IPW estimators and doubly robust estimators. For a chosen parametric form $\pi_i^A = \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})$, the model parameters

$\boldsymbol{\alpha}$ can be estimated through a set of unbiased estimating functions. Let $\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{\alpha})$ be a user-specified vector of functions with the same dimension of $\boldsymbol{\alpha}$. Let

$$\boldsymbol{G}(\boldsymbol{\alpha}) = \sum_{i \in \mathcal{S}_A} \frac{\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\alpha})}{\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})} - \sum_{i \in \mathcal{S}_B} d_i^B \boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\alpha}) \,. \tag{8}$$

It follows that $E_{qp}\{\boldsymbol{G}(\boldsymbol{\alpha}_0)\} = \mathbf{0}$ for any chosen form of $\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\alpha})$, where $\boldsymbol{\alpha}_0$ denotes the true values of the model parameters such as $E_q(R_i \mid \boldsymbol{x}_i) = \pi(\boldsymbol{x}_i, \boldsymbol{\alpha}_0)$. The $\boldsymbol{G}(\boldsymbol{\alpha})$ is a set of unbiased estimating functions under the joint randomisation of q and p, and the estimator $\hat{\boldsymbol{\alpha}}$ obtained by solving $\boldsymbol{G}(\boldsymbol{\alpha}) = \mathbf{0}$ is consistent under some mild moment and smoothness conditions (Tsiatis, 2006). The maximum pseudo likelihood method of CLW2020 corresponds to the choice of $\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\alpha}) = \{\pi_i'(\boldsymbol{\alpha})\}/\{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})\}$, while the method of WVL2021 amounts to using $\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\alpha}) = \{\pi_i'(\boldsymbol{\alpha})\}/\{1 + \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})\}$.

It should be noted that the $\boldsymbol{G}(\boldsymbol{\alpha})$ given in (8) is slightly different from the $\boldsymbol{G}(\boldsymbol{\alpha})$ presented in Wu (2022a, equation (4.4)). The two versions are equivalent but the current form leads to a clear calibration interpretation. The estimating equations $\boldsymbol{G}(\boldsymbol{\alpha}) = \mathbf{0}$ associated with (8) become

$$\sum_{i \in \mathcal{S}_A} \frac{\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\alpha})}{\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})} = \sum_{i \in \mathcal{S}_B} d_i^B \boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\alpha}) \,. \tag{9}$$

The "weights" $\{\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})\}^{-1}$ for the non-probability sample $\mathcal{S}_A$ are calibrated over $\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\alpha})$, with the population controls $\sum_{i=1}^N \boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\alpha})$ being estimated from the reference probability sample $\mathcal{S}_B$. Chen *et al.* (2022) showed that the model-calibration techniques of Wu and Sitter (2001) can be used to achieve doubly robust estimation with non-probability samples under an assumed working model for the outcome regression.

Under a linear outcome regression model ξ with $E_\xi(y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the vector of unknown regression coefficients, the traditional calibration techniques can be employed to achieve double robustness. Let $\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})$ be the chosen parametric form for participation probabilities, where $\boldsymbol{x}$ and $\boldsymbol{\alpha}$ have the same dimension. Let $\hat{\boldsymbol{\alpha}}$ be the solution to

$$\boldsymbol{G}_1(\boldsymbol{\alpha}) = \sum_{i \in \mathcal{S}_A} \frac{\boldsymbol{x}_i}{\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})} - \sum_{i \in \mathcal{S}_B} d_i^B \boldsymbol{x}_i = \mathbf{0} \,. \tag{10}$$

The form of $\boldsymbol{G}_1(\boldsymbol{\alpha})$ in (10) corresponds to the use of $\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{\beta}) = \boldsymbol{x}$ in (8). The IPW estimator $\hat{T}_{yIPW} = \sum_{i \in \mathcal{S}_A} y_i / \pi(\boldsymbol{x}_i, \hat{\boldsymbol{\alpha}})$ for the population total $T_y = \sum_{i=1}^N y_i$ is doubly robust. Noting that $E_\xi(T_y) = \sum_{i=1}^N \boldsymbol{x}_i^T \boldsymbol{\beta}$, we have

$$E_{\xi p}(\hat{T}_{yIPW}) = E_p\Big\{ \sum_{i \in \mathcal{S}_A} \frac{\boldsymbol{x}_i^T \boldsymbol{\beta}}{\pi(\boldsymbol{x}_i, \hat{\boldsymbol{\alpha}})} \Big\} = E_p\Big(\sum_{i \in \mathcal{S}_B} d_i^B \boldsymbol{x}_i \Big)^T \boldsymbol{\beta} = E_\xi(T_y) \,.$$

The estimator $\hat{\mu}_{yIPW} = \hat{N}_A^{-1} \sum_{i \in \mathcal{S}_A} y_i / \hat{\pi}_i^A$ with $\hat{\pi}_i^A = \pi(\boldsymbol{x}_i, \hat{\boldsymbol{\alpha}})$ obtained through (10) is termed as the *calibrated IPW estimator* (Chen *et al.* 2020). The use of calibration equations (10) was also discussed in Rao (2021), Beaumont and Rao (2021), and Chen *et al.* (2023).

There are two important features of the double robustness of the calibrated IPW estimator through traditional calibration (10) for the estimation of participation probabilities. First, the

estimator does not involve the estimation of the unknown regression coefficients $\boldsymbol{\beta}$. It is a useful tool for dealing with undercoverage problems to be discussed in Section 5. Second, the calibration equations in (10) allow population auxiliary information to be combined from multiple reference probability survey samples. For instance, the non-probability sample $\mathcal{S}_A$ may contain five auxiliary variables, two of them are available in one reference probability sample and three in another reference probability sample. The equations (10) only require the estimated population controls and it does not matter if these estimates are from different reference probability samples.

The question of convergence often arises in practice when one solves (10) to obtain $\hat{\boldsymbol{\alpha}}$. From a theoretical point of view, this is not an issue. Suppose that $\pi(\boldsymbol{x}_i, \boldsymbol{\alpha}) = g(\boldsymbol{x}_i^T\boldsymbol{\alpha})$ for some monotone increasing smooth inverse link function $g(\cdot)$ under a generalised linear model for the binary R_i. The "Hessian matrix" is given by

$$\boldsymbol{H}(\boldsymbol{\alpha}) = \frac{\partial}{\partial \boldsymbol{\alpha}} \boldsymbol{G}_1(\boldsymbol{\alpha}) = -\sum_{i\in\mathcal{S}_A} \frac{g'(\boldsymbol{x}_i^T\boldsymbol{\alpha})}{\{g(\boldsymbol{x}_i^T\boldsymbol{\alpha})\}^2} \boldsymbol{x}_i\boldsymbol{x}_i^T \,,$$

where $g'(\cdot)$ is the derivative of $g(\cdot)$. It follows that $-\boldsymbol{H}(\boldsymbol{\alpha})$ is positive-definite, as long as the data matrix $\{\boldsymbol{x}_i, i \in \mathcal{S}_A\}$ is of full rank. The usual Newton-Raphson iterative procedure for solving $\boldsymbol{G}_1(\boldsymbol{\alpha}) = \boldsymbol{0}$ is guaranteed to converge.

4. Poststratification

A common scenario in practice for non-probability survey samples is that the auxiliary variables $\boldsymbol{x}$ included in the sample are all discrete. The IPW estimators under such scenarios reduce to poststratified estimators. Our discussions in this section are mostly taken from materials presented in Section 5 of Wu (2022).

When the auxiliary variables are all ordinal or categorical, the sample $\mathcal{S}_A$ can be poststratified into $\mathcal{S}_A = \mathcal{S}_{A1} \cup \cdots \cup \mathcal{S}_{AK}$ corresponding to the cross-classification of sampled units using the combinations of levels of the $\boldsymbol{x}$ variables. For instance, if $\boldsymbol{x} = (x_1, x_2)^T$ with x_1 having two levels and x_2 having three levels, we have a total of $K = 2 \times 3 = 6$ subpopulations defined by $\boldsymbol{x}$. Let n_k be the size of $\mathcal{S}_{Ak}$ and N_k be the size of the corresponding subpopulation. Under the assumption **A1**, the participation probabilities $\pi_i^A = \pi(\boldsymbol{x}_i)$ become a constant for all units in the same subpopulation and are given by $\pi_i^A = E_q(n_k)/N_k$ for the kth subpopulation. The IPW estimator $\hat{\mu}_{yIPW} = \hat{N}_A^{-1} \sum_{i\in\mathcal{S}_A} y_i/\hat{\pi}_i^A$, with $\hat{\pi}_i^A = n_k/\hat{N}_k$ for $i \in \mathcal{S}_{Ak}$, reduces to the poststratified estimator

$$\hat{\mu}_{yPST} = \frac{1}{\hat{N}^A} \sum_{k=1}^{K} \sum_{i\in\mathcal{S}_{Ak}} \frac{y_i}{\hat{\pi}_i^A} = \sum_{k=1}^{K} \hat{W}_k \bar{y}_k \,, \tag{11}$$

where $\bar{y}_k = n_k^{-1} \sum_{i\in\mathcal{S}_{Ak}} y_i$, $\hat{W}_k = \hat{N}_k/\hat{N}^A$, $\hat{N}_k$ is the estimated size of the kth subpopulation which requires information from external sources. We also impose that $\hat{N}^A = \sum_{k=1}^K \hat{N}_k$. Under the current setting with the availability of a reference probability sample $\mathcal{S}_B$ on $\boldsymbol{x}$, we form the same poststratifications of $\mathcal{S}_B$ as cross-classified by levels of $\boldsymbol{x}$ and obtain $\mathcal{S}_B =$

$\mathcal{S}_{B1} \cup \cdots \cup \mathcal{S}_{BK}$. We can then use $\hat{N}_k = \sum_{i \in \mathcal{S}_{Bk}} d_i^B$. The estimator $\hat{\mu}_{yPST}$ can easily be constructed when the dimension of $\boldsymbol{x}$ is low and the number K is not large.

It is known to the statistical research community that the IPW estimator $\hat{\mu}_{yIPW} = \hat{N}_A^{-1} \sum_{i \in \mathcal{S}_A} y_i / \hat{\pi}_i^A$ under general settings can be sensitive to small estimated participation probabilities. The poststratified estimator $\hat{\mu}_{yPST}$ given in (11) serves as a robust alternative for scenarios where the dimension of $\boldsymbol{x}$ is not low and/or some components of $\boldsymbol{x}$ are continuous. The K strata are formed based on homogeneous groups in terms of participation probabilities. Suppose that $\hat{\pi}_i^A = \pi(\boldsymbol{x}_i, \hat{\boldsymbol{\alpha}})$, $i \in \mathcal{S}_A$ are computed based on a parametric model, q. Suppose also that $n_A = m_A K$ with a chosen K where m_A is an integer. Let $\hat{\pi}_{(1)}^A \leq \cdots \leq \hat{\pi}_{(n_A)}^A$ be the estimated propensity scores in ascending order. Let $\mathcal{S}_{A1}$ be the set of the first m_A units in the sequence, $\mathcal{S}_{A2}$ be the second m_A units in the sequence, and so on. The poststratified estimator of μ_y is computed as $\hat{\mu}_{yPST} = \sum_{k=1}^K \hat{W}_k \bar{y}_k$, which has the same form of the estimator given in (11). The estimates of the stratum weights, $\hat{W}_k$, $k = 1, 2, \cdots, K$ can be obtained by using the reference probability sample $\mathcal{S}_B$ as follows. Let $b_k = \max\{\hat{\pi}_i^A : i \in \mathcal{S}_{Ak}\}$, $k = 1, 2, \cdots, K-1$. Let $b_0 = 0$ and $b_K = 1$.

(a) Compute $\hat{\pi}_i = \pi(\boldsymbol{x}_i, \hat{\boldsymbol{\alpha}})$, $i \in \mathcal{S}_B$.
(b) Define $\mathcal{S}_{Bk} = \{i \mid i \in \mathcal{S}_B, b_{k-1} < \hat{\pi}_i \leq b_k\}$, $k = 1, 2, \cdots, K$.
(c) Calculate $\hat{N}_k = \sum_{i \in \mathcal{S}_{Bk}} d_i^B$, $k = 1, 2, \cdots, K$.

It is apparent that $\mathcal{S}_B = \mathcal{S}_{B1} \cup \cdots \cup \mathcal{S}_{BK}$ and $\sum_{k=1}^K \hat{N}_k = \hat{N}^B = \sum_{i \in \mathcal{S}_B} d_i^B$. The estimated stratum weights are then given by $\hat{W}_k = \hat{N}_k / \hat{N}^B$.

The choice of K needs to reflect the balance between homogeneity of the units within each post-stratum (in terms of participation probabilities) and the stability of the poststratified estimator (in terms of the stratum sample sizes from $\mathcal{S}_A$). When the sample size n_A is small or moderate, a small number such as $K = 5$ should be used. For scenarios where n_A is large, a larger K should be used such that units within the same poststratified sample $\mathcal{S}_{Ak}$ have similar estimated participation probabilities. A practical guidance for the choice of K is to ensure that $m_A \geq 30$ for the poststratified samples.

5. Dealing with Undercoverage

Assumption **A2**, "All the units in the target population have non-zero participation probabilities, *i.e.* $\pi_i^A > 0$, $i = 1, 2, \cdots, N$", is often referred to as the *"positivity assumption"* on participation probabilities. It is a required condition for valid inference on the target finite population. It is a well-known result in the context of probability sampling that there exists an unbiased linear estimator of the population mean/total if and only if the sample inclusion probabilities are positive for all the units in the target population (Wu and Thompson 2020). When $\pi_i^A = 0$ for certain units, the observed sample $\mathcal{S}_A$ no longer represents the entire target population, leading to undercoverage problems. The discussions below are taken from materials presented in Chen *et al.* (2023).

It should be noted that biases due to undercoverage cannot be fully removed without additional information about the uncovered subpopulation. Some existing methods, developed under the positivity assumption **A2**, perform better than others in terms of mitigating biases.

Chen *et al*. (2023) showed that the calibrated IPW estimator discussed in Section 3 has some promises in simulation studies to have smaller biases than other existing estimators when the auxiliary variables $\boldsymbol{x}$ are strong predictors for y and a linear outcome regression model is adequate. In general, model-based prediction methods deserve more attention in dealing with undercoverage problems (Kim *et al*. 2021).

Chen *et al*. (2023) proposed a split population approach to dealing with undercoverage through a convex hull formulation. Let $\mathcal{U}_0 = \{i \mid i \in \mathcal{U} \text{ and } \pi_i^A > 0\}$. It is apparent that $\mathcal{U}_0 \subset \mathcal{U}$ and $\mathcal{U}_0 \neq \mathcal{U}$ when assumption **A2** is violated. Let $\mathcal{U}_1 = \{i \mid i \in \mathcal{U} \text{ and } \pi_i^A = 0\}$. It follows that $\mathcal{U} = \mathcal{U}_0 \cup \mathcal{U}_1$. Let $N = N_0 + N_1$ where N_0 and N_1 are the sizes of the two subpopulations $\mathcal{U}_0$ and $\mathcal{U}_1$. Let $\mu_{y0} = N_0^{-1} \sum_{i \in \mathcal{U}_0} y_i$ and $\mu_{y1} = N_1^{-1} \sum_{i \in \mathcal{U}_1} y_i$. We have $\mu_y = W_0 \mu_{y0} + W_1 \mu_{y1}$, where $W_k = N_k/N$ for $k = 0, 1$.

The key concept for the proposed approach of Chen *et al*. (2023) is the so-called *accessibility function*. Let $\Phi(\boldsymbol{x}_i)$ be a function of $\boldsymbol{x}_i$ that measures the accessibility of unit i to the survey. An individual with a small value of $\Phi(\boldsymbol{x}_i)$ will have (practically) no chance to access the survey. More formally, we have $\pi_i^A = P(i \in \mathcal{S}_A \mid \boldsymbol{x}_i, y_i) = 0$ if $\Phi(\boldsymbol{x}_i) \leq c$ for an unknown cut-off value c on accessibility. The two subpopulations can alternatively be defined as

$$\mathcal{U}_0 = \{i \mid i \in \mathcal{U} \text{ and } \Phi(\boldsymbol{x}_i) > c\} \text{ and } \mathcal{U}_1 = \{i \mid i \in \mathcal{U} \text{ and } \Phi(\boldsymbol{x}_i) \leq c\}.$$

The truncation on $\Phi(\boldsymbol{x}_i)$ to exclude certain units from the non-probability survey can be viewed as a deterministic process, which motivates the use of the term "*deterministic undercoverage*" by Chen *et al*. (2023). An overly simplified example is when x_i represents the "age" of unit i and all young individuals (*i.e.* $x_i \leq c$ for a chosen c) are excluded from the survey.

The proposed approach requires neither the form $\Phi(\cdot)$ nor the value of the cut-off c to be known. It only assumes the accessibility function to be convex. Let $\mathcal{H}_k$ be the convex hull generated by $\{\boldsymbol{x}_i : i \in \mathcal{U}_k\}$ for $k = 0, 1$. It follows that $\Phi(\boldsymbol{x}) > c$ if $\boldsymbol{x} \in \mathcal{H}_0$ and $\Phi(\boldsymbol{x}) \leq c$ if $\boldsymbol{x} \in \mathcal{H}_1$. There are no overlaps between $\mathcal{H}_0$ and $\mathcal{H}_1$. Let $\mathcal{H}_A$ be the convex hull formed by $\{\boldsymbol{x}_i : i \in \mathcal{S}_A\}$. We have $\mathcal{H}_A \subseteq \mathcal{H}_0$ and the difference between the two becomes negligible when n_A is large. Similarly, the convex hull $\mathcal{H}_B$ formed by $\{\boldsymbol{x}_i : i \in \mathcal{S}_B\}$ approximates $\mathcal{H}_0 \cup \mathcal{H}_1$ when n_B is large since $\mathcal{S}_B$ represents the entire target population $\mathcal{U}$.

The two subpopulations $\mathcal{U}_0$ and $\mathcal{U}_1$ can be identified through a split among units in the reference probability sample $\mathcal{S}_B = \mathcal{S}_{B,0} \cup \mathcal{S}_{B,1}$, where $\mathcal{S}_{B,0} = \{j \mid j \in \mathcal{S}_B \text{ and } \boldsymbol{x}_j \in \mathcal{H}_A\}$ and $\mathcal{S}_{B,1} = \mathcal{S}_B \backslash \mathcal{S}_{B,0}$. Note that verifying $\boldsymbol{x}_j \in \mathcal{H}_A$ is equivalent to checking if there exists a sequence of constants $a_i \geq 0$ for $i \in \mathcal{S}_A$ such that

$$\sum_{i \in \mathcal{S}_A} a_i = 1 \quad \text{and} \quad \sum_{i \in \mathcal{S}_A} a_i \boldsymbol{x}_i = \boldsymbol{x}_j \,.$$

It can be done with existing computational packages. The sizes N_0 and N_1 of the two subpopulations $\mathcal{U}_0$ and $\mathcal{U}_1$ can be estimated by $\hat{N}_k^B = \sum_{i \in \mathcal{S}_{B,k}} d_i^B$, $k = 0, 1$, which satisfy $\hat{N}_0^B + \hat{N}_1^B = \hat{N}^B$.

Estimation of μ_y with the split population amounts to estimating μ_{y0} using $\mathcal{S}_A$ and $\mathcal{S}_{B,0}$ and dealing with μ_{y1} using $\mathcal{S}_{B,1}$. An ideal solution, although not very practical, is to have a subsample of $\mathcal{S}_{B,1}$ and obtain measurements on y. A doubly robust estimator of μ_y can then be constructed using all available information.

6. Concluding Remarks

The use of non-probability survey samples and other non-probability data sources such as administrative records has become more and more common in recent years. Even with probability survey samples with a good sampling frame and a sound sampling design, the ever increasing nonresponse rates render them to be non-probability in nature as the final sample inclusion mechanism becomes unknown. Meng (2022) went one step further to claim that "there is no such thing as probability sample in real life". This is practically true for human populations where "any rigorous rules and precise procedures (for probability samples) are almost surely as aspiration, not prescription" (Wu 2022b).

It has become clear in the recent literature that probability survey samples and design-based theory play crucial roles in dealing with non-probability survey samples. As mentioned in the introduction section, the IPW estimators are rooted in the Horvitz-Thompson estimator. The doubly robust estimators widely used in missing data analysis and causal inference as well as for non-probability samples also have a predecessor in generalised difference estimators (Cassel *et al.* 1976) developed in survey sampling. Calibration techniques (Deville and Särndal 1992) and model-calibration methods (Wu and Sitter 2001) are other examples where methodologies are first developed for probability samples but become general inferential tools for the field of statistics (Wu 2023). Challenges faced by dealing with non-probability survey sample also provide opportunities to develop new strategies for this growing topic in survey sampling and official statistics.

References

Beaumont, J. -F. 2020. "Are probability surveys bound to disappear for the production of official statistics?". *Survey Methodology*, Volume 46, N. 1: 1-28.

Beaumont, J.-F., and J.N.K. Rao. 2021. "Pitfalls of making inferences from non-probability samples: can data integration through probability samples provide remedies?". *The Survey Statistician*, Volume 83: 11-22.

Cassel, C.M., C.-E. Särndal, J.H. Wretman. 1976. "Some results on generalized difference estimation and generalized regression estimation for finite populations". *Biometrika*, Volume 63, N. 3: 615-620.

Chen, Y., P. Li, and C. Wu. 2023. "Dealing with undercoverage for non-probability survey samples". *Survey Methodology*, Volume 49, N. 2: 497-515.

Chen, Y., P. Li, and C. Wu. 2020. "Doubly robust inference with non-probability survey samples". *Journal of the American Statistical Association*, Volume 115, N. 532: 2011-2021.

Chen, Y., P. Li, J.N.K. Rao, and C. Wu. 2022. "Pseudo empirical likelihood inference for non-probability survey samples". *The Canadian Journal of Statistics*, Volume 50, N. 4: 1166-1185.

Deville, J.C., and C.E. Särndal. 1992. "Calibration estimators in survey sampling". *Journal of the American Statistical Association*, Volume 87, N. 418: 376-382.

Godambe, V. P. 1960. "An optimum property of regular maximum likelihood estimation". *Annals of Mathematical Statistics*, Volume 31, N. 4: 1208-1212.

Horvitz, D.G., and D.J. Thompson. 1952. "A generalization of sampling without replacement from a finite universe". *Journal of the American Statistical Association*, Volume 47, N. 260: 663-685.

Kim, J.K., and K. Morikawa. 2023. "An empirical likelihood approach to reduce selection bias in voluntary samples". *Calcutta Statistical Association Bulletin*, Volume 75, N. 1: 8-27.

Kim, J.K., S. Park, Y. Chen, and C. Wu. 2021. "Combining non-probability and probability survey samples through mass imputation". *Journal of the Royal Statistical Society, Series A*, Volume 184, N. 3: 941-963.

Meng, X.-L. 2022. "Comments on *Statistical inference with non-probability survey samples* – Miniaturizing data defect correlation: A versatile strategy for handling non-probability samples". *Survey Methodology*, Volume 48, N. 2: 339-360.

Narain, R.D. 1951. "On sampling without replacement with varying probabilities". *Journal of the Indian Society of Agricultural Statistics*, Volume 3: 169-174.

Rao, J.N.K. 2021. "On making valid inferences by integrating data from surveys and other sources". *Sankhya B*, Volume 83: 242-272.

Rao, J.N.K. 2005. "Interplay between sample survey theory and practice: an appraisal". *Survey Methodology*, Volume 31, N. 2: 117-138.

Rivers, D. 2007. "Sampling for web surveys". In *Proceedings of the Survey Research Methods Section*, *Joint Statistical Meetings*: 1-26. Alexandria, VA, U.S.: American Statistical Association.

Tsiatis, A.A. 2006. *Semiparametric Theory and Missing Data*. New York, NY, U.S.: Springer.

Valliant, R., and J.A. Dever. 2011. "Estimating propensity adjustments for volunteer web surveys". *Sociological Methods & Research*, Volume 40, N. 1: 105-137.

Wang. L., R. Valliant, and Y. Li. 2021. "Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts". *Statistics in Medicine*, Volume 40, N. 24: 5237-5250.

Wu, C. 2023. "Calibration techniques for model-based prediction and doubly robust estimation". *The Survey Statistician*, Volume 88: 86-93.

Wu, C. 2022*a*. "Statistical inference with non-probability survey samples". *Survey Methodology*, Volume 48, N. 2: 283-311.

Wu, C. 2022*b*. "Author's response to comments on *Statistical inference with non-probability survey samples*". *Survey Methodology*, Volume 48, N. 2: 367-373.

Wu, C., and R.R. Sitter. 2001. "A model-calibration approach to using complete auxiliary information from survey data". *Journal of the American Statistical Association*, Volume 96, N. 453: 185-193.

Wu, C., and M.E. Thompson. 2020. *Sampling Theory and Practice*. Cham, Switzerland: Springer.

Introduction to Session 1 invited talks

Piero Falorsi, Maria Giovanna Ranalli[1]

Abstract

The common theme of the papers presented in this session is the production of Official Statistics by integrating information coming from survey data and administrative registers. The papers involve researchers from National Statistical Institutes and University departments from four countries. Two papers present approaches that have already been or are in the process of being implemented in the streamline of the institutes' statistical output, while one illustrates collaborative research on a methodological advancement for a production process already implemented.

Keywords: Data integration, administrative registers, latent variable models, signs of life, audit surveys, employment statistics

1. Overlook of the session and of the papers

The three papers presented in this session deal with statistical methods and approaches for integrating information coming from survey data and administrative registers. In this sense, they all fit in the new paradigm proposed by Citro (2014) for Official Statistics production: "*I argue that we can and must move from a paradigm of producing the best estimates possible from a survey to that of producing the best possible estimates to meet user needs from multiple data sources*" (p. 137). This approach has since been adopted in modernisation experiences of National Statistical Institutes (NSIs, see *e.g.* Istat 2016; UNECE 2024).

The papers involve researchers from NSIs and University departments from four countries. Two papers present approaches that have already been or are in the process of being implemented in the streamline of the institutes' statistical output, while one illustrates collaborative research on a methodological advancement for a production process already implemented. The three papers share several common aspects beyond that of using different sources. One common aspect is the use of statistical models for dealing transparently and reproducibly with problems that administrative data do not allow to be solved independently. Another common characteristic is using Census survey data to validate the quality of admin- istrative record data, *e.g.* by estimating the parameters of statistical models and/or deriving (over or under-coverage) probabilities used for statistical registers. Finally, a relevant point is given by the many commonalities across the four countries in the use of administrative and survey data to construct statistical registers for Official Statistics.

The first paper – *Multi-source data: new approaches for non-standard employment statistics. The Dutch and Italian experience* – by Danila Filipponi, Silvia Loriga, Mauricio Garnier Villarreal, Dimitris Pavlopoulos, and Roberta Varriale is the fruit of collaborative research between Sapienza University of Rome, Vrij University of Amsterdam, CBS, and Istat.

The paper fits into the flow of literature on hidden Markov models that integrates Labour Force Survey data and administrative registers to improve the production of Official Statistics on employment (Filipponi *et al.* 2019; Boeschoten *et al.* 2021), which is at the base of employment

1 Piero Demetrio Falorsi (piero.falorsi@gmail.com), Istat Advisory Committee on Statistical Methods, and Sapienza Università di Roma; Maria Giovanna Ranalli (maria.ranalli@unipg.it), Università degli Studi di Perugia.

variable definition and production at Istat. In particular, by mitigating measurement inaccuracies intrinsic in the primary data sources, hidden Markov models allow to extract the latent phenomenon and exploit longitudinal information. In this paper, the focus is on mobility trends over time, by studying the transition from flexible to permanent employment. With respect to previous studies, here the Authors face the issue of integrating data not only from different sources; but also from two different countries (Italy and the Netherlands) using multiple-group hidden Markov models. This should make it possible to harmonise the statistics generated and to compare data from the two countries more effectively. Extensive model selection is conducted to understand (and hence properly incorporate into the model) the differences between the two countries in terms of the impact of measurement error.

The second and the third paper describe the experience of the French NSI (INSEE) and of the U.S. Census Bureau, respectively, in implementing the production of population statistics by means of several administrative sources. "*Setting up statistical registers of individuals and dwellings in France: approach and first steps*" presented by Aurélien Lavergne illustrates a profound change in the production of Official Statistics in France and describes the program implemented at INSEE to create a system of interconnected registers of individuals, dwellings and households. The latter becomes the reference universe based on the use of a large number of sources, making it more resilient than using tax data only as it is being done currently. This experience represents another example of a paradigm shift where we observe the transformation of statistical operations by gradually replacing survey data with administrative data. The paper gives a comprehensive overall picture of the process that remarkably acknowledges other countries' experience, methods, and software.

"*Producing U.S. Population statistics using multiple administrative sources*" by J. David Brown and Marta Murray-Close describes the several challenges encountered by the U.S. Census Bureau when constructing administrative record-based population estimates for 2020. In particular, 31 sources are combined to create an extended population register to achieve more comprehensive coverage under the principle of redundancy using the Signs-of-Life method. The paper deals with various topics: locational accuracy, person coverage and its consistency across time, coverage of children, distinguishing international migrants from continuous U.S. residents, and the choice of demographic characteristics when multiple ones are reported or when they are missing altogether. For each issue, it provides details on the challenges and solutions adopted to address them.

2. Perspectives

The three papers open the room for deeper insights and further enhancements. The first paper uses an interesting approach that requires careful model selection and validation. The latter can use different tools according to the inferential target of interest. The paper seems to tackle two different inferential aims: the first is more analytical and pertains to understanding the structural differences in the measurement process in the two countries, while the second is more descriptive in targeting population estimates of counts and rates. While the approach used in the paper and based on information criteria is suitable for the former, different tools based on measures of the predictive accuracy of the model should be used for the latter, not only for the overall population estimates but also for geographical and/or socio-demographic subpopulations of particular interest. In addition, the paper opens for the need for methodological tools that allow measurement of the error of the final population estimates.

Infact, classical design-based variance estimates of population estimates must be integrated with (likely bootstrap-based) measures of the variability of model predictions.

The second paper highlights how complex the transition to a new production method might be. The experience carried out at Istat in the past years to face a similar challenge suggests that it should be planned carefully and it should allow for time to test/evaluate/validate the new estimation strategies. As long as France can afford the large Rolling Census it's being conducted now, with approximately five million dwellings and 9.3 million inhabitants involved every year, it should keep doing it, in particular for validating some of the estimation choices to be made in the (near) future such as the evaluation of over/under coverage of the register and the development and assessment of the residency index. In fact, the construction of the residency index calls for many choices, such as the values for the parameters of the convex combination, the weights, and the threshold. A new perspective in this context can be provided by considering residency as a latent variable hidden behind the signs of life. A continuous latent construct can be extracted by means of Item Response Theory models, which may provide data-driven weights linked to the discrimination parameters. Alternatively, a categorical latent variable can be considered to obtain latent residency classes that cluster profiles of signs of life. Locational accuracy is one of the issues faced in the third paper, as well. In particular, a person-place model is used to predict the probability that a given address is the person's address on the reference date using data from the ACS. This approach does not provide allocation of a unit to a single residence, but fractions of a person may be included in multiple locations. This "fuzzy" approach is very sensible and can be further enhanced by the approach currently considered in Italy, where an application of graph sampling (Zhang 2021) of pairs of individuals and addresses from a "redundant" population register enhanced with all addresses of a unit is being studied to assess the prevalent address of a unit. Another issue faced in the paper, that of race and ethnicity discrepancy between the administrative register and the Census, could be well handled using the approach proposed in the first paper for employment and based on hidden Markov models.

References

Boeschoten, L., D. Filipponi, and R. Varriale. 2021. "Combining multiple imputation and hidden Markov Modeling to obtain consistent estimates of employment status". *Journal of Survey Statistics and Methodology*, Volume 9, N. 3: 549-573.

Citro, C. F. 2014. *From multiple modes for surveys to multiple data sources for estimates*. *Survey Methodology*, Volume 40, N. 2: 137-162.

Filipponi D., U. Guarnera, and R. Varriale. 2019. "Hidden Markov Models to Estimate Italian Employment Status". *Conference NTTS*, March 11-13, 2019. Brussels, Belgium.

Istituto Nazionale di Statistica - Istat. 2016. *Istat's Modernisation Programme*. Rome, Italy: Istat. https://www.istat.it/it/files//2011/04/IstatsModernistionProgramme_EN.pdf.

United Nations Economic Commission for Europe - UNECE. 2023. "Modernization of Official Statistics". *UNECE website*. https://unece.org/statistics/modernization-official-statistics.

Zhang, L-C. 2021. *Graph sampling*. Boca Raton, FL, U.S.: CRC Press.

Measurement of contract type from multi-source data. Preliminary results from the Dutch and Italian experience

Danila Filipponi, Mauricio Garnier-Villarreal, Silvia Loriga, Dimitris Pavlopoulos, Reinoud Stoel, Roberta Varriale[1]

Abstract

This paper explores the impact of measurement errors on employment contract data and mobility trends over time, focussing particularly on cross-country comparisons. We compare Italian and Dutch employment data from 2016 to 2021, integrating information from the Labour Force Survey and the Employment Register. Using a multiple-group hidden Markov model, we estimate the impact of measurement errors within each country and facilitate meaningful cross-country comparisons.

Keywords: Measurements errors, multiple-group hidden Markov model, employment career.

1. Introduction

Measurement errors can have a significant impact on data analysis, and the importance of these errors can vary between data sources. When dealing with survey data, measurement errors may be caused by a range of factors, including cognitive processes, social desirability, and design and implementation issues on data collection. Register and administrative data are typically prone to measurement errors deriving from administrative delays, misregistration, or inconsistent administrative procedures beyond differences between definitions adopted for statistical and administrative variables. Regardless of their origin, these errors can introduce substantial bias into the resulting statistical information. While it is reasonable to assume a certain degree of consistency in measurement errors over time from the same data source, challenges arise when comparing indicators among various countries that employ distinct data collection methodologies. In such cases, the impact of measurement errors may hinder meaningful territorial comparisons. This challenge persists even when statistical institutions dedicate significant efforts to standardise the definitions and computation methods for many socio-economic indicators.

Over the past few decades, the incidence of temporary employment became a key socio-economic indicator in the labour market. In both socio-economic research and policy making, understanding the role of this type of employment in the life course is fundamental. Inspired by this challenge, the central question of this research is to understand better when temporary employment serves as a stepping stone to permanent employment and when it devolves into a vicious cycle of precarious, short-term jobs (Latner 2022). Research on employment dynamics

1 Danila Filipponi (dafilipp@istat.it), and Silvia Loriga (loriga@istat.it), Italian National Institute of Statistics - Istat; Mauricio Garnier-Villarreal (m.garniervillarreal@vu.nl), Dimitris Pavlopoulos (d.pavlopoulos@vu.nl), Reinoud Stoel (r.stoel@cbs.nl), Vrije Universiteit Amsterdam and Statistics Netherlands (CBS); Roberta Varriale (roberta.varriale@uniroma1.it), Sapienza Università di Roma, Roma, Italy. The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics - Istat.

has shown that the transition from temporary to permanent employment can be significantly biased due to measurement errors present in the data used for analysis (see, for example, Pavlopoulos and Vermunt 2015; Pankowska *et al.* 2021; Pavlopoulos *et al.* 2023). Even a minor amount of measurement error in the classification of employment contracts can lead to a substantial overestimation of transitions over time, such as moving from a temporary to a permanent contract. Such bias can have profound implications for policymakers.

One possible approach for addressing measurement errors when multiple data sources are available is using latent variable models. When several sources contain information closely related to the target variable, but none can be assumed to be a proper measure of the target variable, latent variable models can be used to predict the true target value given the observed measurements in the data sources. In this context, Latent Class Analysis (LCA) is a method for identifying a latent categorical construct of interest using observed categorical variables. LCA can be used to evaluate and correct measurement errors (Vermunt 2010; Biemer 2011). Hidden Markov models (HMMs) extend LCA when longitudinal data are available. These models have been used to correct for measurement error in mobility between employment states (Bassi *et al.* 2000) and employment contracts (Pavlopoulos and Vermunt, 2015; Pankowska *et al.* 2021; Pavlopoulos *et al.* 2023). Moreover, they have been used to estimate the employment status in Italy (Filipponi *et al.* 2021).

This paper addresses the issue of how to effectively compare data on employment contracts and mobility over time between employment statuses in different countries, each affected by varying levels and types of measurement errors. We compare Italian data on contract types provided by the Italian National Institute of Statistics (Istat) and Dutch data on contract types provided by Statistics Netherlands (CBS). Both datasets combine unit-level information on the employment contract type from the Labour Force Survey and the Employment Register. The Italian data spans 2017 to 2021, while the Dutch data covers 2016 to 2019. To tackle the impact of measurement errors in both countries, we adopt a multiple-group HMM with two indicators for employment contracts and a country factor, using data from the years where the two countries' data overlap, *i.e.* 2017-2019. We aim to estimate the error-corrected distribution of the employment contract type and the error-corrected mobility rates (*e.g.* from temporary to permanent employment) in Italy and the Netherlands.

The rest of the paper is organised as follows: Section 2 describes the data. Section 3 introduces the model employed in our analysis. Finally, Section 4 presents the model results and Section 5 draws some conclusions.

2. The data

Istat and CBS rely on multiple data sources to collect information employment. The primary source for Official Statistics on the labour market is the Labour Force Survey (LFS), which is conducted by both National Statistical Institutes (NSIs) following the European standards outlined in EU Regulation 2019/1700 of the European Parliament and the Council. The LFS provides data on employment and key job market indicators, including occupation, economic sector, hours worked, contract types, and training. Additionally, both NSIs gather and process data from various administrative sources.

In the Italian LFS, the sampling design is a two-stage process with primary units being municipalities and final units being households. Large municipalities with populations exceeding a specific threshold are always included in the sample, while smaller municipalities

are grouped into *strata* and one municipality is selected in each *stratum* with a probability proportional to its population. Households are then randomly selected from the municipal population register. These selected households are interviewed four times within a 15-month period, first in two consecutive quarters, followed by a two-quarters break, and finally two other consecutive quarters of interview. Interviews are spread to all weeks throughout the quarter. The LFS produces monthly, quarterly and yearly estimates for key labour market indicators. These estimates are further segmented by gender, age, and geographical area down to the NUTS3 level (on a yearly basis), with the reference population being residents in Italy aged 15 and over. For further details on LFS content, methodologies, and organisation refer to Istat (2006). The Italian Employment Register (ER), managed by Istat, is built by integrating administrative data collected mainly by social security and tax authorities. Different data sources from these agencies are used depending on the type of employment contract or tax deadlines, leading to data quality and content variations. For this reason, these data go through distinct preprocessing and harmonisation procedures, extensively detailed in Istat (2015) and Baldi *et al.* (2018): harmonised data is organised with an employer-employee linkage structure, forming the basis for extracting information about the "worker" that is coherent with the International Labour Office definitions. The ER data do not contain information on irregular work and do not fully cover jobs whose wages do not exceed a certain threshold (Varriale and Alfó 2023).

In the Dutch LFS, the sample design is a stratified two-stage cluster design based on addresses. *Strata* are the geographical regions, with municipalities as primary units and addresses as secondary sampling units. All households residing at an address, up to a maximum of three, are included in the sample. The Dutch LFS follows a quarterly rotation scheme and is representative of the Dutch population aged 15 and older. Respondents are interviewed in five consecutive panel waves, and interviews are spread to all weeks throughout each quarter. The Dutch ER is administered by the Employee Insurance Agency (Uitvoeringsinstituut Werknemersverzekeringen, UWV) and contains information on the labour market and income for all insured workers in The Netherlands (Bakker *et al.* 2014). The ER is built by collecting and matching information from various sources, including the tax office, declarations from temporary work agencies, and the population register. Notably, there is no missing data, as employers are required to submit tax-reporting statements. However, it is worth noting that the dataset contains monthly information (aggregated from daily data in the source dataset), but employers typically submit relevant data only once or twice a year, which can lead to systematic error between consecutive extractions.

In this work, we consider two different indicators in each country related to the employment contracts that are derived from the LFS and the ER. Both indicators have three different levels: employees with permanent contracts (PE), employees with temporary contracts (TE), and others (OT), such as self-employed and non-employed individuals. The analysis is based on quarterly data and focuses on individuals aged between 25 and 55 to ensure a homogeneous population from a working life perspective. In the Italian data, the period covered spans from 2017 to 2021, *i.e.* 20 data points. The number of LFS interviews carried out by each household during this period may be either 4 if the rotation scheme took place within these years or fewer than 4 if it extended beyond this time frame. The LFS data retains information from all survey waves in which individuals participated, like details about employment status, employment contract types, hours worked during the reference week, educational levels, and whether the interview was conducted via proxy. For the same group of individuals, if their jobs are covered by ER, we observe quarterly information at the unit level, covering all quarters from January 2017 to December 2021. The variables in the dataset include contract type, age, gender,

citizenship, municipalities of residence, and labour income, classified into various income classes. Regarding Dutch data, the period covered spans from 2016 to 2019, for a total of 16 data points. Information from all waves of the LFS in which individuals participated is retained, for a maximum of five consecutive data points. The same individuals receive individual-level information from the ER covering all quarters from January 2016 to December 2019. The available information includes the employment status, the employment contract types, whether the LFS interview was conducted via proxy, age, gender and country of birth. To reduce the size of the dataset, a 10% sample of units was randomly selected from the original data in both countries. The sample was stratified by the month of the first interview in LFS to ensure participation from all LFS cohorts.

2.1 Descriptive statistics of the observed data

The objective of this study is to estimate both the *true* size of temporary employment and the *true* transition rate from temporary to permanent employment, and to provide comparable statistics between the two countries. To this aim, some initial summaries need to be offered.

Table 2.1 - Distribution of employment categories in ER and LFS. Italy. Years 2017-2021

Employment contract LFS \ER	Permanent	Temporary	Other	All
		Overall Percentages		
Permanent	41.06	1.23	2.79	45.08
Temporary	1.85	5.44	1.74	9.03
Others	2.46	1.50	41.93	45.89
All	44.67	8.34	47.00	100.00
		Row Percentages		
Permanent	91.07	2.73	6.19	100.00
Temporary	20.48	60.28	19.24	100.00
Others	5.36	3.26	91.38	100.00
All	44.67	8.34	47.00	100.00
		Column Percentages		
Permanent	90.51	15.08	6.01	45.08
Temporary	4.07	66.59	3.74	9.03
Others	5.42	18.33	90.25	45.89
All	100.00	100.00	100.00	100.00

Source: Istat

Table 2.1 presents a cross-tabulation of employment categories based on data from both the Italian LFS and ER. The diagonal values of the total percentage table show the highest percentages, reflecting cases where both sources agree on the classification. In contrast, off-diagonal values represent discrepancies in classification, indicating potential classification errors in at least one of the data sources. As shown in the table, the two measures do not align for approximately 12% of the total number of cases. Overall percentage in table 2.1 further illustrates the marginal distribution of employment categories as measured by the LFS and ER. The two distributions are very similar, with only slight differences emerging in the percentage of the categories *Temporary* and *Other*. The row and column percentage tables show that the main discrepancy can be observed in the category *Temporary*, highlighting the difficulty in both sources to measure temporary jobs, probably mainly due to the difficulties in correctly identifying the reference period of the information and the respondents' misclassification due to an erroneous understanding of employment categories.

Table 2.2 presents a cross-tabulation of employment categories based on Dutch data from both LFS and ER sources. Here, the overall incidence of off-diagonal values, representing discrepancies in the classification of the two data sources, is approximately 18% of the total number of cases, showing a higher misalignment of the two sources compared with the Italian case. Further, the marginal distributions of employment categories as measured by the LFS and ER are strongly different with an incidence of the temporary workers in the LFS (18.3%) higher than the one in ER (10.8%). In addition, the greater disagreement concerns temporary employment: more than 50% of the cases that are recorded as temporary contracts in the ER are differently classified by the LFS.

Table 2.2 - Distribution of employment categories in ER and LFS. The Netherlands. Years 2016-2019

Employment contract LFS \ER	Permanent	Temporary	Other	All
		Overall Percentages		
Permanent	48.63	8.60	4.45	63.80
Temporary	1.20	9.20	1.38	10.83
Others	0.90	1.08	24.56	25.37
All	47.90	18.27	33.83	100.00
		Row Percentages		
Permanent	78.84	13.95	7.21	100.00
Temporary	10.21	78.09	11.70	100.00
Others	3.38	4.07	92.55	100.00
All	47.90	18.27	33.83	100.00
		Column Percentages		
Permanent	95.86	45.56	14.64	63.80
Temporary	2.37	48.72	4.54	10.83
Others	1.77	5.72	80.82	25.37
All	100.00	100.00	100.00	100.00

Source: CBS

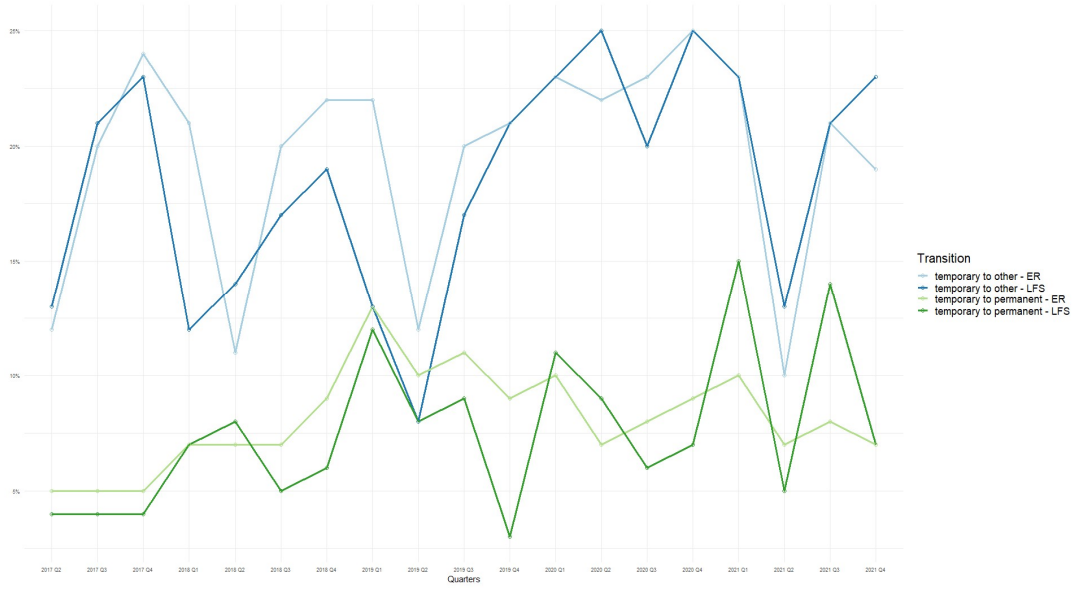
While both countries exhibit discrepancies between LFS and ER, the nature and extent of these discrepancies differ. Understanding these variations is crucial for interpreting and improving the reliability of employment data in each country. Moreover, the labour market structure in the Netherlands and Italy exhibits important differences. The Netherlands has a higher incidence of permanent contracts, reflecting a more stable labour market. In contrast, Italy shows a pattern with a comparatively lower incidence of permanent contracts and a high percentage of non-employed.

Figures 2.1 and 2.2 show, for the two countries, the transition rates from a temporary contract, *i.e.* the transition rates to permanent contracts and other categories (predominantly not employed) between adjacent quarters. For LFS, transitions are evaluated only when two consecutive observations are available. In Italian data (Figure 2.1), minimal disparities in flow patterns between the LFS and ER data sources are observed, the only difference being a slightly smoother transition from temporary to permanent contracts in the ER data. For both sources, the transition rate from temporary to permanent contracts remains consistently low, below 15%. Conversely, the transition from temporary to other categories, mainly indicating not being employed, can reach 25%. Substantial fluctuations become apparent when examining different quarters. A notable trend is the increasing shift from permanent contracts to other categories in the last quarters of the year. This implies that a segment of temporary contracts does not change towards permanent contracts by the end of the year. This pattern remains consistent in LFS and ER, indicating a robust trend across data sources. This trend suggests that the temporal dynamics of these transitions are not independent of the time variable. Dutch data (Figure 2.2) differ

significantly from Italian data, not only in terms of discrepancies between the two sources but also in employment dynamics. There are noticeable differences in transition patterns concerning the Figure 2.1, especially in the shift from temporary to permanent contracts, as measured by the LFS and ER. Fluctuations over quarters are particularly evident in the transition from temporary to permanent contracts. According to the LFS, there is an observed increase during the first quarter, suggesting a transformation of contracts at the beginning of the year. However, this trend is not confirmed by the ER, showing an important discrepancy between the two sources.

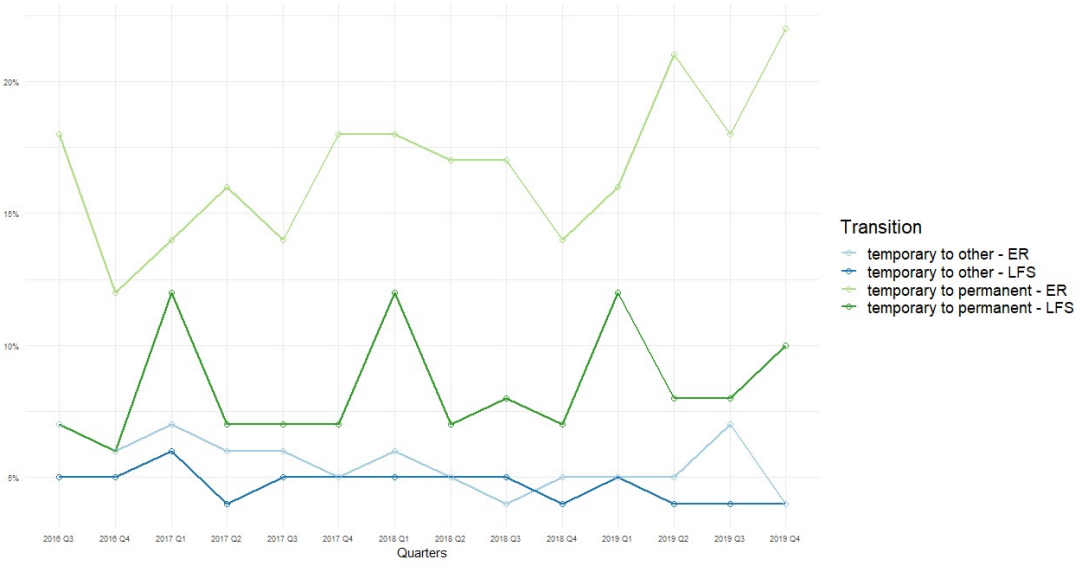
This preliminary analysis highlights the importance of using a statistical model for employment transitions to address the issue of measurement errors in the two data sources. It also emphasizes the need to explore whether certain model parameters vary across countries.

Figure 2.1 - Observed transition flow in LFS and ER data by quarters. Italy, years 2017-2021



Source: Istat

Figure 2.2 - Observed transition flow in LFS and ER data by quarters. The Netherlands, years 2016-2019.



Source: CBS

3. The multiple-group hidden Markov model

In this paper, we employ a multiple-group HMM to examine the impact of measurement error in Italy and The Netherlands. A "basic" hidden Markov model assumes that a latent variable follows a homogeneous first-order Markov process, while measurement errors are conditionally independent given the hidden states. Furthermore, it assumes invariance of model parameters across all first-level units (individuals). With respect to the latter assumption, on the contrary, a multiple-group HMM allows for variations in some model parameters between groups, in this case, between the two countries. Following a fixed-effects approach as delineated by Clifford and Goodman (1984), we incorporate group dummies in the model.

The variables in the multiple-group HMM are *X, C, E*. Let us denote with X_{ikt} the value of the *true* (latent) target variable at time t for subject i in country k, where $t = 0, \ldots, T$, $i = 1, \ldots, N$, and $k \in \{1, 2\}$. In our study, X_{ikt} has three categories, permanent contract (PE), temporary contract (TE), and individuals not participating in paid employment (OT). It's important to note that the last category includes not only unemployed individuals but also people in education and self-employment. The variables C_{ikt} and E_{ikt} represent the two measurements of the target variable: C_{ikt} denotes the observed contract type of person i at time point t and in country k according to the ER, and E_{ikt} according to LFS. C_{ikt} and E_{ikt} assume the three values of the target variable. We refer to a particular category of the observed variables by c_t and e_t, and the latent variable by x_t. As indicated in the previous section, we use quarterly data from 2017 to 2021 for the Italian data and quarterly data from 2016 to 2019 for the Dutch data, and we estimate the model for the years where the two countries' data overlap, *i.e.* 2017-2019. This means that the total period covers three years, and t runs from 0 to $T = 12$.

Under the basic assumption of HMM, the probability of following a certain observed path of C_{ikt} and E_{ikt} over the entire period can be expressed as follows:

$$
\begin{aligned}
P(\mathbf{C}_i = \mathbf{c}_i, \mathbf{E}_i = \mathbf{e}_i) = \sum_{x_0=1}^{3} \sum_{x_1=1}^{3} \cdots \sum_{x_T=1}^{3} & P(X_{ik0} = x_0) \\
& \prod_{t=1}^{T} P(X_{ikt} = x_t | X_{ik(t-1)} = x_{t-1}) \\
& \prod_{t=0}^{T} P(C_{ikt} = c_t | X_{ikt} = x_t) \\
& \prod_{t=0}^{T} P(E_{ikt} = e_t | X_{ikt} = x_t)^{\delta^1_{ikt}}.
\end{aligned}
\tag{1}
$$

In equation (1), $P(X_{ik0} = X_0)$ represents the initial state probabilities, $P(X_{ikt} = X_t | X_{ik(t-1)} = X_{t-1})$ are the transition probabilities from $t-1$ to t, $P(C_{ikt} = C_t / X_{ikt} = {}_{xt})$ are the measurement error probabilities for the *ER*, and $P(E_{ikt} = E_t / X_{ikt} = X_t)$ are the measurement error probabilities for LFS. The indicator variable δ^1_{ikt} takes a value 1 when LFS data is obtainable for country k at time t. Unlike ER data, if the data for country k is accessible for one year, it is only available for the quarter in which individual i in country k is included in the sample.

The most comprehensive version of a multiple-group HMM is achieved when the model parameters governing the initial state probabilities, latent transition probabilities, and measurement error probabilities are considered specific to each group k, corresponding to the unrestricted multiple-group models. A constrained version of the multiple-group HMM model can be obtained by assuming invariant measurement error probabilities, creating a partially

heterogeneous model. Naturally, the flexibility exists to modify this assumption for specific indicators as necessary.

To approximate more realistic scenarios of the labour market, the model in Equation 1 can be extended in different ways. Specifically, we allow the latent transition probabilities to be time-heterogeneous, introducing a dependence on a quadratic specification of time (*i.e.* be conditional on t and t^2, $P(X_{ikt} = x_t | X_{ik(t-1)} = x_{t-1}, t, t^2)$).

We also relax the basic assumption of the Independent Classification Error (ICE) which means that the observed states are independent of one another within and between time points. In our case, this assumption is unrealistic since it has been shown that both survey and register data on the employment contract type contain systematic error (Pankowska *et al.* 2021). We relax the ICE assumption in two ways. We allow the response from the survey E_{ikt} to depend on covariates V_{ikt}, while in other models we introduced direct across-time correlation in the measurement error in both the register and survey indicator. In this case, the observed indicators in both the survey and the register data are allowed to depend on the lagged observed and lagged true contract type. Instead of estimating freely all different sets of error probabilities in the register and survey data, we applied restrictions that correspond to realistic scenarios: we defined a constrained model that estimates an extra error parameter when an error was made in *t-1* (*an error*), or for the case where an error was made in *t-1* and this error can be repeated in *t* (*same error* structure). These assumptions can be introduce in both countries or just in one.

The joint probability of having a particular observed state path considering time-dependent transition probabilities and the systematic measurement error can be expressed as follows:

$$
\begin{aligned}
P(\mathbf{C}_i = \mathbf{c}_i, \mathbf{E_i} = \mathbf{e}_i) = & \sum_{x_0=1}^{3} \sum_{x_1=1}^{3} \dots \sum_{x_T=1}^{3} P(X_{ik0} = x_0) \\
& \prod_{t=1}^{T} P(X_{ikt} = x_t | X_{ik(t-1)} = x_{t-1}, t, t^2) \\
& \prod_{t=0}^{T} P(C_{ikt} = c_t | X_{ikt} = x_t, X_{ik(t-1)} = x_{t-1}, C_{ik(t-1)} = c_{t-1}) \\
& \prod_{t=0}^{T} P(E_{ikt} = e_t | X_{ikt} = x_t, X_{ik(t-1)} = x_{t-1}, E_{ik(t-1)} = e_{t-1})^{\delta^1_{ikt}}.
\end{aligned}
\tag{2}
$$

Summarising, the multi-group HMM is composed by 2 parts: the structural and measurement part. In the structural part, we estimate the initial and transition probabilities, while in the measurement part we estimate the error patterns of both indicators. This can take two structures, random and correlated (systematic), and when correlated it can be of the type *an error* or *same error*.

Based on a sample of independent realisations from the distribution (Equation 2), estimates of the relevant model parameters can be obtained via Maximum likelihood estimation using the Expectation-Maximisation (EM) algorithm, the extent of which is implemented in the software Latent GOLD v.6.0 (Vermunt and Magidson 2016).

4. Model results

We chose to estimate the model for the years where the two countries' data overlap, *i.e.* 2017-2019. This decision aims to exclude the impact of the exceptional year 2020, affected by the *COVID-19* pandemic, which is available only for the Italian dataset. By doing so, we

aim to enhance the precision and reliability of our estimates by minimising the influence of the exceptional circumstances of the pandemic on the model's parameters.

The final model selection occurred in two steps. In the first step, we compared models with different structural and random error invariance structures. In the second step, we focussed on specifying different configurations of the measurement error component: random or correlated. In the following, we use the term (parameter) "invariance" to refer to models in which there is no difference between the values of the parameters estimated in the two countries, while "heterogeneity" describes models in which model parameters are estimated to be different in the two countries.

For the initial step, we considered five models. Model (a) is the baseline model assuming invariance of parameters in both the structural model and the measurement error part, which consists of the random component only. Models (b)-(e) assume heterogeneity in the structural model and: (b) invariance in the measurement error of both indicators (LFS, ER), (c) invariance in measurement error of only ER indicator and heterogeneity in measurement error of the indicator derived from LFS data, (d) invariance in measurement error of only LFS indicator and heterogeneity in measurement error of the indicator derived from ER data, and finally, (e) heterogeneity in measurement error of both indicators. The results of the first step are presented in Table 4.1. Model fit measures suggest that incorporating the country variable in either the structural or measurement component (models (c)-(e)) marginally enhances the fit, but the improvement is not substantial (maximum 1.3%). However, heterogeneity in the structural part of the model can be theoretically justified by the strong differences in the labour markets of the two countries. Consequently, we proceed with model (b) and test, in the second step, whether the measurement error component should incorporate the country covariate.

Table 4.1 - Model Comparison: Measurament Invariance

Model	LL	BIC(LL)	AIC(LL)	AIC3(LL)	Npar
(a) Baseline	-440973.1545	882314.3288	882010.3089	882042.3089	32
(b) Invariance Measur. Error LFS-ER	-436235.1753	873068.3828	872574.3505	872626.3505	52
(c) Invariance Measur. Error ER	-433848.4492	868363.9344	867812.8984	867870.8984	58
(d) Invariance Measur. Error LFS	-433617.9833	867903.0025	867351.9666	867409.9666	58
(e) Unrestricted multiple-group	-430950.7967	862637.6331	862029.5934	862093.5934	64

Source: Own computation

In the second step, we studied the multi-group HMM error structure by comparing 13 models. The results of the second step are presented in Table 4.2. Models (b1)-(b8) were examined with a focus on parameters' invariance across countries in the measurement error component for LFS and ER. In these models, we have different configurations of measurement error, but equal parameters in the two countries, both for LFS and ER indicators. Correlation in the measurement error for both the register and survey indicators was introduced through the estimation of an additional error coefficient. This involved considering whether the error made at time point *t-1* could be repeated in *t*, as described in section 3. As described above, we formulated two constrained models: one accounting for an additional error coefficient only if the same error was present at time *t-1* (*same error*), and a second considering an additional error coefficient if any error was present at time *t-1* (*an error*). Model (b1) corresponds to Model (b) in the first step of model selection, and represents the baseline model of step 2. Therefore, Model (b1) assumes heterogeneity in the structural model and invariance in the measurement random error of both indicators (LFS, ER). In Model (b2), we have correlated

error in LFS determined by age and proxy interviewing and random error in ER. Models (b3) and (b5) assume error correlation in LFS, with the *same error* and *an error* configuration. For ER, Models (b4) and (b6) assume error correlation, with the same error and an error configuration. Models (b7) and (b8) allow for error correlation in both LFS and ER, with the *same error* and *an error* configuration.

As an additional analysis, we studied models with different configurations of the measurement error component. In particular, in Models (b9)-(b13), we studied the presence of correlated errors in LFS and ER, but with different configurations, *i.e.* the *same error* and *an error*, in Italy and The Netherlands. Finally, Model (b13) assumes the same type of measurement error configuration, *same error*, but with different intensity parameters for the two countries. By comparing fit index criteria such as BIC, AIC and AIC3, we opted for model (b13).

Table 4.2 - Model Comparison: error structure

Model	LL	BIC(LL)	AIC(LL)	AIC3(LL)	Npar
Heterogeneity structural, Invariance Measurement LFS - ER					
(b1) Baseline. Uncorrelated Error IT and NL	-436235.175	873068.382	872574.350	872626.350	52
(b2) Corr. LFS (age proxy IT and NL)	-435786.081	872216.197	871684.162	871740.162	56
(b3) Corr. LFS (same error IT and NL)	-416458.682	833584.400	833033.364	833091.364	58
(b4) Corr. ER (same error IT and NL)	-420546.896	841760.828	841209.792	841267.792	58
(b5) Corr. LFS (an error IT and NL)	-421855.053	844377.142	843826.106	843884.106	58
(b6) Corr. ER (an error IT and NL)	-427214.857	855096.750	854545.714	854603.714	58
(b7) Corr. LFS and ER (an error IT and NL)	-418251.661	837239.361	836631.322	836695.322	64
(b8) Corr. LFS and ER (same error IT and NL)	-409936.581	820609.202	820001.163	820065.163	64
(b9) Corr. LFS (an error NL same error IT)	-412822.588	826461.721	825787.177	825858.177	70
(b10) Corr. LFS (an error IT same error NL)	-410697.258	822199.561	821534.517	821604.517	70
(b11) Corr. ER (an error NL same error IT)	-435995.259	872807.062	872132.518	872203.518	70
(b12) Corr. ER (an error IT same error NL)	-436043.501	872892.046	872227.002	872297.002	70
Heterogeneity structural, Heterogeneity Corr Error, Invariance Random Error LFS - ER					
(b13) Corr. ER and LFS (same error)	-417297.090	835472.029	834746.181	834822.181	76

Source: Own computation

Tables 4.3 and 4.4 show, highlighted in italics, the conditional probabilities for error repetition (*same error*) estimated with Model (b13). For example, when the true latent state in *t-1* is temporary and the observed contract type in *t-1* is permanent, the probability of error repetition, *i.e.* observing a permanent contract type in *t* given a temporary latent state in *t*, is equal to 0.81. In Italy, the conditional error probabilities are always smaller than in the Netherlands, and this is especially true for the ER indicator.

Figure 4.1 illustrates the transition probabilities from temporary employment to permanent employment in both observed data and those estimated by model (b13) for both countries. As noted previously by Pavlopoulos and Vermunt (2015), the latent transition probabilities generally appear lower than those observed in both the LFS and ER, with variations between the two countries. In the Italian case, the estimated transition probabilities are noticeably smoother but closely aligned with the observed ones; in contrast, in the Dutch case, they are definitively lower than the observed probabilities. These results reflect the varying levels of coherence between the two countries.

Table 4.3 - LFS indicator. Conditional probabilities for error repetition (same error). Years 2017- 2019

Country	Observed LFS $t-1$	Latent t	Latent $t-1$	Permanent	Observed LFS, t Temporary	Other
Italy	Permanent	Temporary	Temporary	0.811	0.172	0.017
Italy	Permanent	Other	Other	0.816	0.005	0.178
Italy	Temporary	Permanent	Permanent	0.161	0.838	0.001
Italy	Temporary	Other	Other	0.015	0.759	0.226
Italy	Other	Permanent	Permanent	0.303	0.002	0.695
Italy	Other	Temporary	Temporary	0.117	0.325	0.559
Netherlands	Permanent	Temporary	Temporary	0.971	0.027	0.003
Netherlands	Permanent	Other	Other	0.979	0.001	0.020
Netherlands	Temporary	Permanent	Permanent	0.160	0.839	0.001
Netherlands	Temporary	Other	Other	0.014	0.784	0.202
Netherlands	Other	Permanent	Permanent	0.051	0.000	0.949
Netherlands	Other	Temporary	Temporary	0.111	0.308	0.582

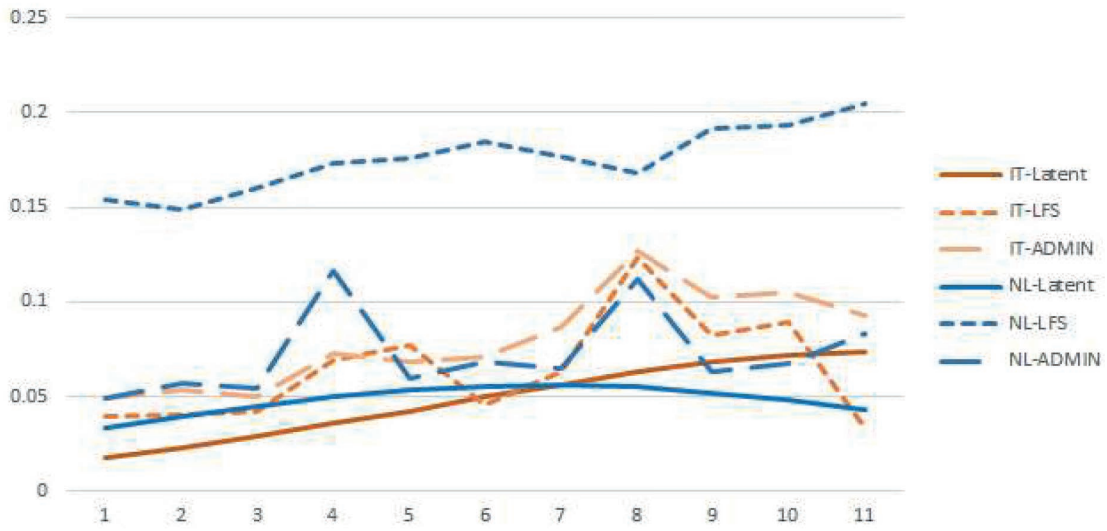
Source: Istat-CBS

Table 4.4 - ER indicator. Conditional probabilities for error repetition (same error).Years 2017-2019

Country	Observed ER $t-1$	Latent t	Latent $t-1$	Permanent	Observed ER, t Temporary	Other
Italy	Permanent	Temporary	Temporary	0.883	0.112	0.006
Italy	Permanent	Other	Other	0.785	0.002	0.214
Italy	Temporary	Permanent	Permanent	0.270	0.728	0.001
Italy	Temporary	Other	Other	0.004	0.279	0.718
Italy	Other	Permanent	Permanent	0.671	0.002	0.327
Italy	Other	Temporary	Temporary	0.022	0.698	0.280
Netherlands	Permanent	Temporary	Temporary	0.870	0.124	0.006
Netherlands	Permanent	Other	Other	0.876	0.001	0.123
Netherlands	Temporary	Permanent	Permanent	0.093	0.907	0.000
Netherlands	Temporary	Other	Other	0.002	0.716	0.282
Netherlands	Other	Permanent	Permanent	0.064	0.000	0.936
Netherlands	Other	Temporary	Temporary	0.017	0.540	0.443

Source: Istat-CBS

Figure 4.1 - Observed transition flows from temporary to permanent contracts in LFS and ER data and estimated flow by quarters. Italy-CBS, years 2017-2019



Source: Istat

5. Conclusions and future work

In this paper, we explored the impact of measurement errors on cross-country differences in the distribution of the employment contract type and mobility between different types of employment contracts. In particular, we compared Italian and Dutch employment data integrating information from the Labour Force Survey and the Employment Register using a multiple-group hidden Markov model on data from 2017 to 2019. The findings point to the necessity for awareness of measurement error when carrying out analyses on labour market statistics. In particular, there is a need to estimate error-corrected transition rates to both evaluate the impact of measurement errors within each country and facilitate meaningful cross-country comparisons.

Our work on this important topic should be considered preliminary. In the future, we will contemplate several enhancements to the model. Firstly, it is crucial to disaggregate the variable "contract type" into more detailed subcategories, especially for the "other" level, distinguishing between non-employed and self-employed individuals. Furthermore, the structural part of the model could be expanded by incorporating covariates and addressing unobserved heterogeneity. The measurement component of the model could be enriched by testing additional specifications of systematic errors. Additionally, conducting sensitivity analyses of the model's assumptions through Monte Carlo-type simulations is part of our planned improvements.

6. Acknowledgements

This paper is part of the project DYNANSE that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 864471).

References

Amuedo-Dorantes, C., and R. Serrano-Padial. 2007. "Wage Growth Implications of Fixed-Term Employment: An Analysis by Contract Duration and Job Mobility". *Labour Economics,* Volume 14 , N. 5: 829-847.

Bakk, Z., and J. Kuha. 2018. "Two-Step Estimation of Models Between Latent Classes and External Variables". *Psychometrika*, Volume 83, N.4: 871-892.

Bakker, B.F.M., J. van Rooijen, and L. van Toor. 2014. "The System of Social Statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics". Amsterdam, the Netherlands: IOS Press, *Statistical Journal of the IAOS*, Volume 30, N. 4: 411-424.

Baldi, C., C. Ceccarelli, S. Gigante, S. Pacini, and F. Rossetti. 2018. "The Labour Register In Italy: The New Heart Of The System Of Labour Statistics". *Rivista Italiana di Economia, Demografia e Statistica*, Volume LXXII, N. 2: 95-105.

Bassi, F., J.A. Hagenaars, M.A. Croon, and J.K. Vermunt. 2000. "Estimating True Changes when Categorical Panel Data are Affected by Uncorrelated and Correlated Classification Errors: An Application to Unemployment Data". *Sociological Methods & Research*, Volume 29, N. 2: 230-268.

Baum, L.E., T. Petrie, G. Soules, and N. Weiss. 1970. "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains". *The Annals of*

Mathematical Statistics, Volume 41, N. 1:164-171.

Becker, G.S. 1993. *Human Capital*. New York, NY, U.S.: Columbia University Press.

Biemer, P.P. 2011. *Latent Class Analysis of Survey Errors*. Hoboken, NJ, U.S.: John Wiley & Sons.

Clifford, C., and L. Goodman. 1984. "Latent Structure Analysis of a Set of Multidimensional Contingency Tables". *Journal of the American Statistical Association*, Volume 79, N. 388: 762-771

Cox, D.R., M. Jackson, and S. Liu. 2009. "On square and ordinal contingency tables: a comparison of social class and income mobility for the same individuals". *Journal of the Royal Statistical Society: Series A,* Volume 172, N. 2: 483-493.

Enders, C.K. 2010. *Applied Missing Data Analysis*. New York, NY, U.S.: Guilford Press.

Fabbris, L. 2010. "Dimensionality of scores obtained with a paired-comparison tournament system of questionnaire item." In Palumbo, F., C.N. Lauro, and M.J. Greenacre (*eds.*). *Data Analysis and Classification. Proceedings of the 6th Conference of the Classification and Data Analysis Group of the Società Italiana di Statistica*: 115-162. Berlin/Heidelberg, Germany: Springer.

Filipponi, D., U. Guarnera, and R. Varriale. 2021. "Latent Mixed Markov Models for the Production of Population Census Data on Employment". In C. Perna, C., N. Salvati, and F. Schirripa Spagnolo (*eds.*). *Book of Short Papers*: 112-117. *SIS 2021 - Statistical Learning, Sustainability and Impact Evaluation.* Milano, Italy: Pearson Italia.

Gash, V., and F. McGinnity. 2006. "Fixed-Term Contracts-The New European Inequality? Comparing Men and Women in West Germany and France". *Socio-Economic Review*, Volume 5, N. 3: 467-496.

Gebel, M. 2010. "Early career consequences of temporary employment in Germany and the UK". *Work, Employment and Society*, Volume 24, N. 4: 641-660.

Istituto Nazionale di Statistica - Istat. 2015. *Atti del 9 Censimento generale dell'industria e dei servizi e Censimento delle istituzioni non profit. Fascicolo 2 - Il Censimento delle imprese*. Parte I. Roma, Italy: Istat.

Istituto Nazionale di Statistica - Istat. 2006. "La rilevazione sulle forze di lavoro: contenuti, metodologie, organizzazione". *Metodi e Norme*, Volume 32: 173-196.

Latner, J.P. 2022. "Temporary employment in Europe: stagnating rates and rising risks". *European Societies*, Volume 24, N. 4: 383-408.

Latner, J.P., and N. Saks. 2022. "The wage and career consequences of temporary employment in Europe: Analysing the theories and synthesizing the evidence". *Journal of European Social Policy*, Volume 32, N. 5: 514-530.

Manzoni, A., J.K. Vermunt, R. Luijkx, and R. Muffels. 2010. "Memory Bias in Retrospectively Collected Employment Careers: A Model-Based Approach to Correct for Measurement Error." *Sociological Methodology,* Volume 40, N. 1: 39-73.

Masyn, K.E. 2013. "Latent Class Analysis and Finite Mixture Modeling". In Nathan, P.E., and T.D. Little (*eds.*). *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2: Statistical Analysis*. Oxford, UK: Oxford University Press.

Mooi-Reci, I., and R. Dekker. 2015. "Fixed-Term Contracts: Short-Term Blessings or Long-Term Scars? Empirical Findings from the Netherlands 1980-2000". *British Journal of Industrial Relations*, Volume 53, N. 1: 112-135.

Organization for Economic Co-operation and Development - OECD. 2023. *OECD website.* https://stats.oecd.org/#.

Pankowska, P.K., B.F.M. Bakker, D. Oberski, and D. Pavlopoulos. 2021. "Dependent interviewing: a remedy or a curse for measurement error in surveys?" *Survey Research Methods*, Volume 15, N. 2: 135-146.

Pankowska, P.K., B.F.M. Bakker, D. Oberski, and D. Pavlopoulos. 2018. "Reconciliation of Inconsistent Data Sources by Correction for Measurement Error: The Feasibility of Parameter Re-use". *Statistical Journal of the IAOS,* Volume 34, N. 3: 317-329.

Pavlopoulos, D. 2013. "Starting Your Career With a Fixed-Term Job: Stepping-Stone or "Dead End"?" *Review of Social Economy*, Volume 71, N. 4: 474-501.

Pavlopoulos, D., M. Garnier-Villarreal, and R. Varriale. 2023. "Patterns of flexible employment careers. Does measurement error matter?" In Chelli, F.M., M. Ciommi, S. Ingrassia, F. Mariani, and M.C. Recchioni (*eds.*). *Book of the Short Papers*: 985-990. *SIS 2023 - Statistical Learning, Sustainability and Impact Evaluation.* Milano, Italy: Pearson Italia.

Pavlopoulos, D., and J.K. Vermunt. 2015. "Measuring Temporary Employment. Do Survey or Register Data Tell the Truth?" *Survey Methodology*, Volume 41, N. 1: 197-214.

Siegfrid, T. 2010. "Odds are, it's wrong: Science fails to face the shortcomings of statistics". *Science News*, Volume 177, N. 7: 26-29.

Van Lissa, C.J., M. Garnier-Villarreal, and D. Anadria. 2023. "Recommended Practices in Latent Class Analysis Using the Open-Source R-Package tidySEM". *Structural Equation Modeling: A Multidisciplinary Journal*, Volume 31, N. 3: 526-534.

Varriale, R., and M. Alfó. 2023. "Multi-source statistics on employment status in Italy, a machine learning approach". *METRON*, Volume 81: 37-63.

Vermunt, J.K. 2010. "Longitudinal Research Using Mixture Models". In van Montfort, K., J.H.L. Oud, and A. Satorra. *Longitudinal Research with Latent Variables*: 119-152. Berlin, Germany: Springer.

Vermunt, J.K. 2003. "Multilevel latent class models". *Sociological Methodology*, Volume 33: 213-239.

Vermunt, J.K., and J. Magidson. 2016. *Technical guide for Latent GOLD 5.1: Basic, Advanced, and Syntax*. Belmont, MA, U.S.: Statistical Innovations Inc.

Setting up statistical registers of individuals and dwellings in France: Approach and first steps

Aurélien Lavergne[1]

Abstract

In order to rationalise a more massive use of administrative data in the production of demographic and social statistics, the French national statistical institute (INSEE) has decided to launch the Résil program (Individual and dwellings statistical registers). The goal of this program is to create a system of individual, dwellings and household registers based on the mobilisation of external data, particularly administrative data, in strict compliance with the conditions of individual data protection. Thus, through these registers, INSEE will have a reference universe that will allow:

- *to constitute the sampling frames as well as the calibration margins for household surveys;*
- *to measure the quality of coverage of sources;*
- *to match different datasets: surveys with administrative data, administrative data to administrative data in order to provide richer information.*

To set up this information system, INSEE will draw on international experiences, as it plans to use methodologies already implemented in several national statistical institutes, including:

- *deterministic matching methods for identifying and comparing different administrative sources. These methods are essential insofar as France does not have a unique and shared identifier;*
- *the Signs-of-Life method to define the reference population from the presence of individuals in several input data;*
- *the Dual System Estimation to measure the quality of coverage of the registers, taking into account the annual data of the population census (it is a real opportunity to have these annual points of comparison in order to assess the quality of the registers, especially during the set-up period).*

This paper will therefore present the stakes of the implementation of these registers, then describe the main methodological principles and the proofs of concept planned for its implementation.

Keywords: Administrative data; statistical registers; Signs of Life; dual system estimation; data linking.

1. Introduction

In order to rationalise a more massive use of administrative data in the production of demographic and social statistics, the French national statistical institute (INSEE) has decided to launch the Résil programme (Statistical Register of Individuals and Dwellings). The goal of this program is to create a system of individual, housing and household registers based on the mobilisation of external data, particularly administrative data, in strict compliance with the conditions of individual data protection. Thus, through these registers, INSEE will have a reference universe that will allow:

1 Aurélien Lavergne (aurelien.lavergne@insee.fr), Institut National de la Statistique et des Études Économiques - INSEE.

- to constitute the sampling frames as well as the calibration margins for household surveys;
- to measure the quality of coverage of sources;
- to match different datasets: surveys with administrative data, administrative data to administrative data in order to provide richer information.

This paper briefly presents the context of the Résil programme before focussing on the definition of the reference population envisaged, which will be based on administrative data. The results of the first simulations will be presented and analysed, highlighting the progress already made and the difficulties still to be overcome.

1. The target and the state of play of Résil

1.1 The French context

We have no population register and no individual identifier shared with all government departments. The subject of registers is very sensitive in France and a project of shared ID was stopped 50 years ago after a big polemic, some people still remember that.

Another key contextual point, not specific to France, is the need to go further in the use of administrative sources and to develop new tools for this (processing, linking, ...).

So in 2020, we launched a project called Résil, to be deployed in 2025, aiming at building statistical registers of individuals and dwellings, updated with several administrative sources, and used for statistical purposes only, for example to facilitate record linkages or providing sampling frames.

1.2 In concrete terms, what will Résil look like and how will it be fed and used?

The Résil (Statistical Register of Individuals and Dwellings) programme aims to build a system of registers which will consist more specifically of:

- Two registers (individuals and dwellings) updated on an ongoing basis and essentially containing identifiers (see the paragraph below on the content of Résil);
- A reference universe consisting of three annual databases (individuals, dwellings and households) containing all the individuals present in the area on a given date, all the dwellings (ordinary and community) and all the households to which these individuals belong;
- Three services: reception of external sources, production of the reference universe and production of files enhanced by matching.

These various components are described below.

1.2.1 A statistical register of individuals and dwellings

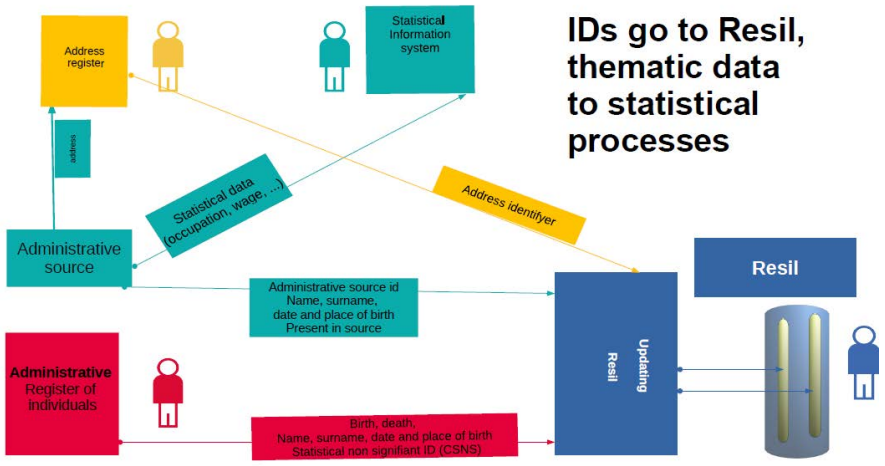
Schematically, Résil will consist of two statistical registers, one for individuals, the other for dwellings, with many observations (all individuals, all dwellings suitable to live in) but very few variables (identifiers, signs of life, links between individuals and dwellings).

These registers are continuously updated from several administrative sources, annually or monthly, to ensure the best possible coverage of the population, because no source is completely exhaustive, or compliant with the concepts of usual residence. It is also updated by the administrative register of individuals, which enables it to be up to date on births and deaths on French soil. Beyond the benefits in terms of compliance, using several sources is important because it allows us to be more resilient in the event of the failure or transformation of a source. We saw recently in France, with the suppression of one of our fiscal sources, which was the spine of several statistical processes, that it was not a virtual risk.

To feed Résil from these sources, we need only IDs (individual, dwelling and addresses); "statistical variables" will be led to other databases, in order to feed statistical processes (incomes, employment and wages, dwelling descriptions...). This is why, in addition to the registers, the project has built a shared reception service for administrative data, which is described in detail below.

Resil needs to receive and process several administrative sources in order to identify people who are present in the data source to feed its Sign-of-Life model (Chapter 4). That is why a tool of integration of administrative data was developed.

Figure 1.1 - A shared reception service for administrative data



The administrative input files (blue rectangle in the diagram above) will be split into 3:

- identifying data which will supply Résil;
- localisation data (address) which will supply the French address register. In return, the address register will provide Résil with the non-meaningful address identifier corresponding to the unencrypted address;
- the other data contained in these sources. These pseudonymised data will be available to other INSEE information systems to produce statistics based on these administrative data.

It allows us to progress in this step, which was previously managed in silos and will now be processed with a unique tool, allowing ambitious investments toward better performances, scalability, security and metadata processing.

At this step, we will be able to replace names, surnames and addresses by non-significant codes, ensuring a pseudonymisation (suppression of direct IDs as names, surnames and social security numbers, but it is still possible to recognise persons) of statistical datasets. In this

way, Résil enables the minimisation principle to be implemented even further, as identification variables will be kept in a single location and no longer replicated across different systems.

1.2.3 The "reference universe"

From these registers, we will take three photos, updated annually, to produce "reference universes": a list and localisation of all individuals present on the French territory at the reference date (using the Signs-of-Life method), a list of inhabitable dwellings at the reference date, a list of households (all individuals living in the same dwelling) at the reference date. For each individual, we will define the usual residence, choosing if necessary between the different addresses we picked up from different sources.

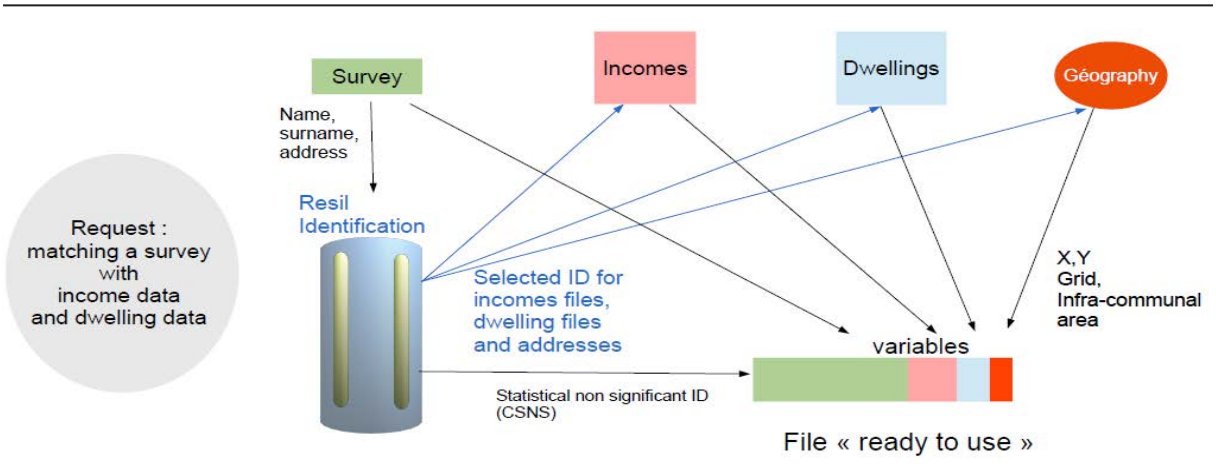
These reference universes will be used to evaluate the coverage of administrative sources, which represents significant progress. At the moment, our scope of interest corresponds to the scope covered by the source and we have no point of comparison to do so, we take the source "as it comes", doing comparisons with previous years or months, but it is not sufficient to have a complete evaluation of its coverage. Thanks to these "reference universes", we will be able to calibrate the scope of the administrative data on our theoretical scope.

1.2.4 The production of enhanced files

These "reference universes" will also be used as spines for the construction of enriched files (which is the subject of the second service that we are going to propose to our users), by picking up statistical variables in statistical datasets and linking by common IDs. It can provide, for example, sampling frames, or enriched surveys with administrative data.

The diagram below gives an example of the enrichment of a survey with administrative data.

Figure 1.2 - Example of the enrichment of a survey with administrative data



The various stages of enrichment are as follows.

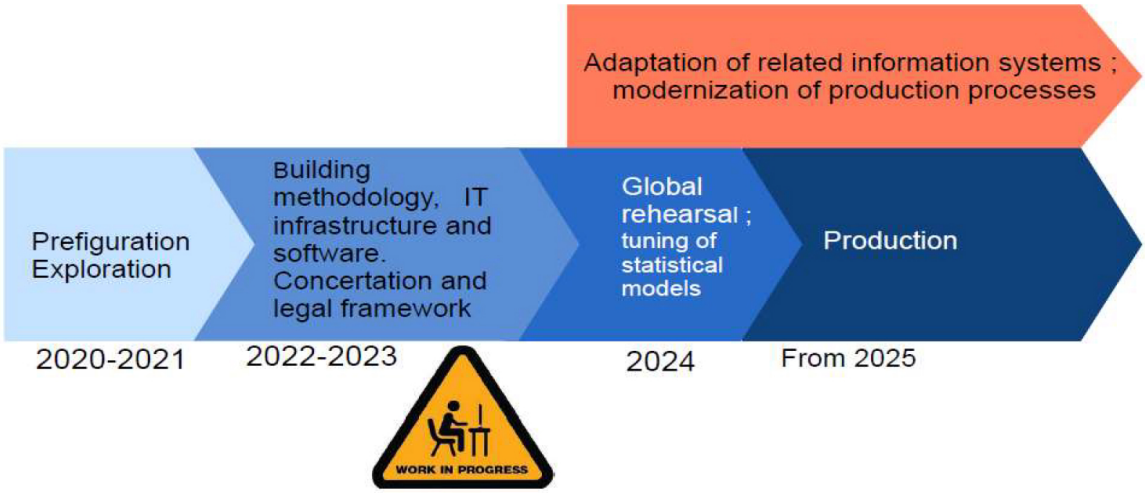
The survey manager provides Résil with the identity details of the respondents to his survey and the list of administrative data he needs to complete the variables he has to collect. This makes it possible to reduce the survey collection workload by not asking questions whose answers are in an administrative file.

Résil begins by identifying the individuals that belong to the sample. It then retrieves the desired variables from the raw administrative data or variables in other INSEE information systems (such as the information system in charge of incomes or dwellings or geographical data), and creates a file containing the values of the desired variables for each survey respondent.

1.2 The time schedule

Résil will be deployed in 2025; the time schedule is as follows:

Figure 1.3 - Résil deployment in 2025



2. The definition of the reference universe

2.1 Processing the reference universe: technical steps

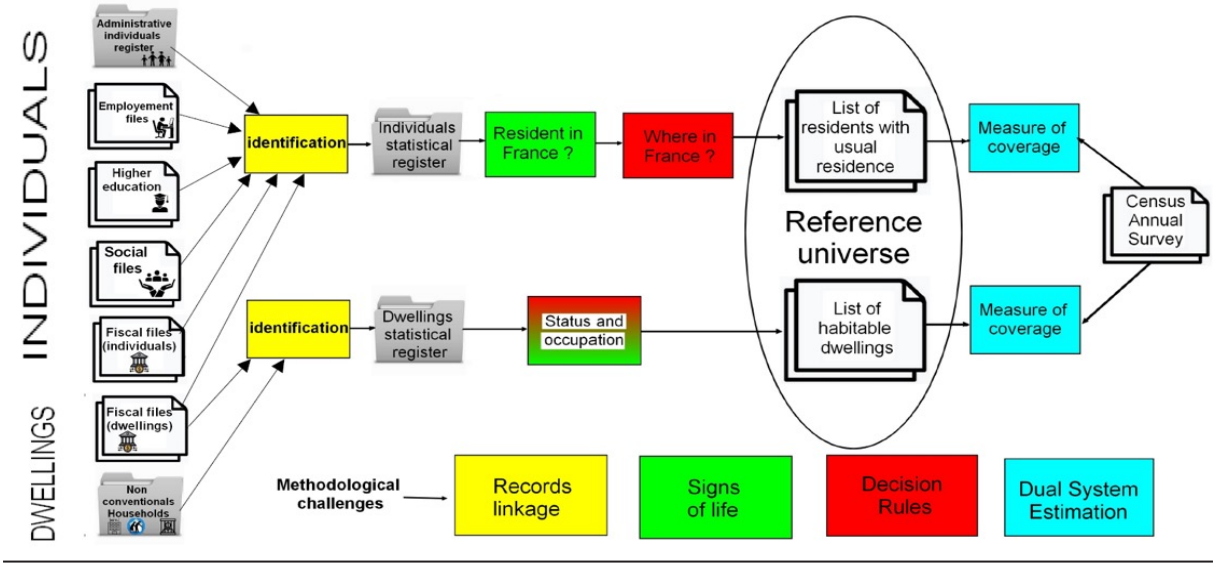
At the left of Figure 2.1, you have administrative files, for individuals and dwellings, that will be used to initialise and update the register. The administrative sources that will provide Résil with data are tax sources (on individuals and housing), social sources (social benefits), employment sources (salaries) and a source on students (to improve their location, as they are often wrongly located with their parents in tax files).

The administrative register for identification of individuals will be used to initialise the register, and to update it with data on births, deaths, immigrants and changes of identity. It is exhaustive for people born in France. It contains people not born in France, but living or having lived in France and needing social coverage. The main drawback of this register is that it does not contain any address, neither indicate if people are still living in France. That is why it overestimated a lot the French population and Résil must use other administrative data in order to identify people of this register who have still their usual residence in the French territory at the reference date, and where this usual residence is located.

In concrete terms, other sources are used to update the register with two types of information:

- is the individual present in the source?
- if yes, at what address or in what dwelling?

Figure 2.1 - Reference universe in Résil



In addition, there may be a delay (of more than several months) in registering immigrants in the administrative individuals' register.

Thus, an individual born abroad may be present in an administrative source before being known to the administrative individuals' register. In that condition, we will create an observation.

Another difficulty we face is the lack of a single identifier shared by all administrations.

Matching between an external source and Résil can generate matching errors, which is why the quality of the match is taken into account when weighting the Signs-of-Life model.

Furthermore, if an individual from an administrative unit is not found in Résil, two scenarios are considered:

- if the individual not found was born abroad, he or she is created in the directory (see above).
- if not, we will assume that it is a bad identification because every person born in France is in the administrative individuals' register, so in Résil, and we won't take this individual in our Signs-of-Life model.

Updating the register seems easier for dwellings, since we have only two sources: a register of dwellings used for local taxation and a register of non-conventional households, set up to manage data collection of the census surveys. The difficulty will be to isolate the cases of non-conventional households registered as well in fiscal files, to avoid double counts.

To obtain the reference universes, we have some methodological challenges, the quality of which will affect the result:

- identification to the register: if we make a mistake, we can wrongly generate a sign of life in a source (false positive) or we can miss updating an observation (false negative);
- decision about reference address for the individuals that have several addresses (in France or not, if several addresses in France, localisation at infra-national level);
- use of a Signs-of-Life model (is the person still living in France?);
- measuring coverage by comparison with census surveys or census sampling frames (a dedicated register of buildings in bigger municipalities); using Dual System (DSE) or Trimmed Dual System Estimation (TDSE) method.

2.2 The technical challenges

In order to obtain the most exhaustive reference universe possible, and thus avoid missing individuals who were resident in France on the 1st of January of a given year, or taking into account individuals who have left French territory, we have to face the following issues:

- Areas or populations difficult to grasp in administrative sources (some overseas, homeless or informal settlements, …). We are indeed aware that the quality of administrative sources is insufficient in a large part of the overseas territories (overseas territories have more than 2 million inhabitants), and we will likely need to implement specific coverage surveys. Our objective is to minimise coverage defects in Résil as much as possible. However, we must be cautious not to overcorrect for one population at the expense of another. Overcorrection could introduce a coverage bias, rendering Résil unusable as a sampling frame.
- Quality of record linkages. For each match, we will draw a control sample from the accepted pairs. A visual analysis of this sample will allow us not only to estimate the false positive rate but also, if necessary, to propose new matching rules in the event of systematic errors.
- Tuning of the Signs-of-Life model (and robustness through years).
- Decision rules in cases of multiple addresses to get the usual residence: to determine an individual's usual residence, high priority is given to fiscal files; however, in certain specific cases, other administrative sources may be used. For example, in the case of young individuals, their usual residence will be considered where they live for their studies, not with their parents, even if they are still registered with them in the fiscal files. For individuals who juggle between two places of residence – one near their workplace, where they stay, for example, from Monday to Friday, and the other during the weekend with their family in their home city – it will be considered that their usual residence is where they live with their family, not the place where they reside during the week for work.

3. Quality assessment: a measure of coverage of the universe of reference

3.1 Assessing the quality with census surveys: opportunity and difficulties

In the case of France, the existence of an annual census survey is a very important asset for measuring the quality of Résil. It will enable measuring the degree of coverage of the register and of the census every year for a part of the French territory (that which is counted).

The census has the advantage to be a large-scale survey that counts 5 million dwellings and 9 million of individuals each year with the targeted concepts of residence. However, it also presents several difficulties:

- It is not exhaustive in large municipalities (the sample size in municipalities over 10,000 inhabitants is 8% of dwellings each year, so 40% in the 5-year cycle) and only one-fifth of small municipalities are surveyed each year;
- The identification data quality in the census is not very good for some of them (mainly for those responding by paper questionnaire) because of some data capture errors or missing values for variables used in the identification process, such as names, surnames, dates and place of birth, ...

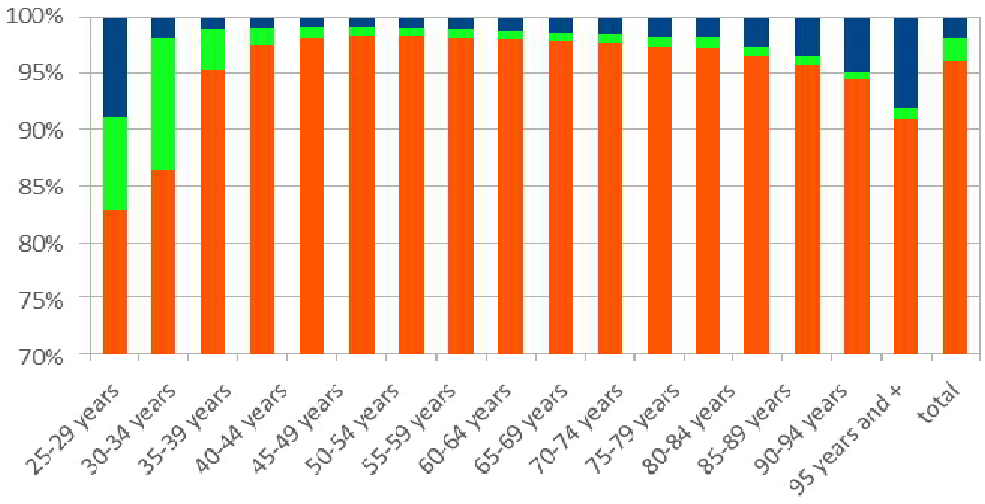
3.2 Several sources: opportunity and difficulties

By using several sources, we have a better coverage of the population, especially on youngest people.

The graph 3.1 is quite expressive: the green bars are representing people added with other sources than tax data; they really complete the red bars and assure a more homogeneous coverage by age. The blue bars represent people who have been registered but who are not found in any administrative source.

It should be noted that the scope of the graph is limited to individuals who have reached the age of majority. For our study, we did not have the identity details of minors, but we will eventually have them in Résil.

Figure 3.1 - Proportion by age of individuals counted found in fiscal files (orange), found in other sources (green); not found in any file (blue). Year 2020.



Source: Own computation

Some sources may have problems with coverage or location for certain populations (for example, the tax source does not allow us to find all young people between the ages of 18 and 25, nor to locate them correctly at the time they leave the parental home but remain financially dependent on their parents). For this population, the benefit of the enrolment in higher education is significant (+ 1,5 pt. of coverage).

The results confirm the benefits of mobilising each of these sources, in terms of population coverage, compared with the tax source alone. The total overall gain in coverage is 2 points for individuals aged over 18 (from 95.4 % to 97.5 %), but rises to 10 points for 21-25-year-olds (from 85.9% to 95.6%). This means that the coverage by age is much more homogeneous than with the tax source alone (see Figure 3.1).

For people living in institutions, the coverage rate rises by 10 points, from 80% to 90%.

It should be noted that in some cases it is not possible to find individuals who have been collected by the census in the administrative sources because the quality of their civil status in the census is too poor. The 7% of individuals aged between 18 and 20 who are missing from administrative sources do not systematically reflect a lack of coverage by these sources. It may be partly due to a matching problem linked to the poor quality of identity features in the census.

But, the reverse of the medal, we have over-coverage when compiling administrative sources, as shown in the following schema. At least 14% are in the basic compilation of sources and are not found in the census survey, due to several factors:

- missing values or bad quality of identification variables in the census, so people exist in the dataset but we do not recognise them (for example we have 3% of imputed answers for non-responses in the census survey for the considered municipalities);
- a different localisation of people between census and Résil (we don't search in the right municipality);
- "false negatives" in id tool, so leading duplicates in the compilation of sources;
- at least, over-coverage of the administrative sources (some sources contain people who live abroad).

4. A reference universe based on the Signs-of-Life method

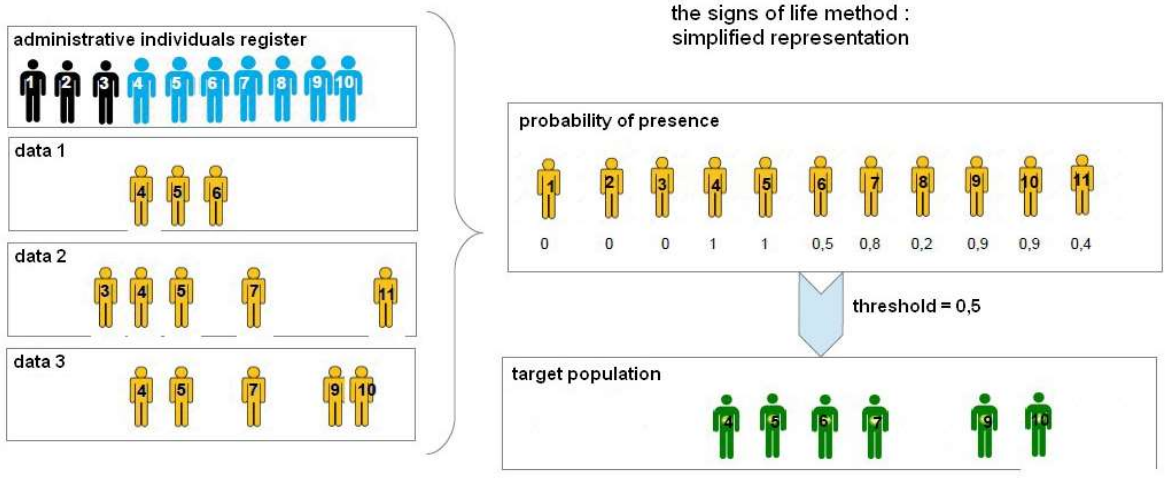
4.1 Constitution of a reference universe

With reference to Figure 4.1:

- Individuals in black are recorded as deceased in the administrative register (RNIPP) for the year in question.
- Individuals in blue are alive in the RNIPP. It should be noted that some of these are potentially deceased abroad but no official death certificate has been recorded in France.
- Individuals in yellow are present in the administrative sources and in the Résil databases.
- Individuals in green are the reference population for the year in question.
- Individuals in red are individuals present in Résil but who are not part of the reference population for the year in question.

To account for births, deaths, and individuals arriving in France from abroad who need social security, we rely on the administrative register of individuals. However, this register does not provide information on individuals leaving the French territory.

Figure 4.1 - Simplified representation of the Signs-of-Life method



Therefore, we will attempt to estimate the probability of residence in French territory for each individual by analysing their presence in the various administrative sources at our disposal.

For example, if an individual who is not deceased in the administrative register (RNIPP) is not present in any of the administrative sources, there is a high probability that he or she no longer resides in France. If he or she appears in only some of the sources in which he or she would normally be found, we can assume that there is a non-zero probability that he or she has left the country. Each source can be weighted according to its quality and relevance to the individuals concerned (for example, if we use the student file, it will be only relevant for 18-25-year-olds). The probability of presence thus calculated for each individual on the basis of presence in external sources is then compared with a threshold (which will have been defined on the basis of comparisons with EAR data). If the probability exceeds the threshold, the individual will be considered to be part of my reference population for the year in question. If not, the individual will remain in the Résil databases (for at least 10 years) but will not be part of the reference population.

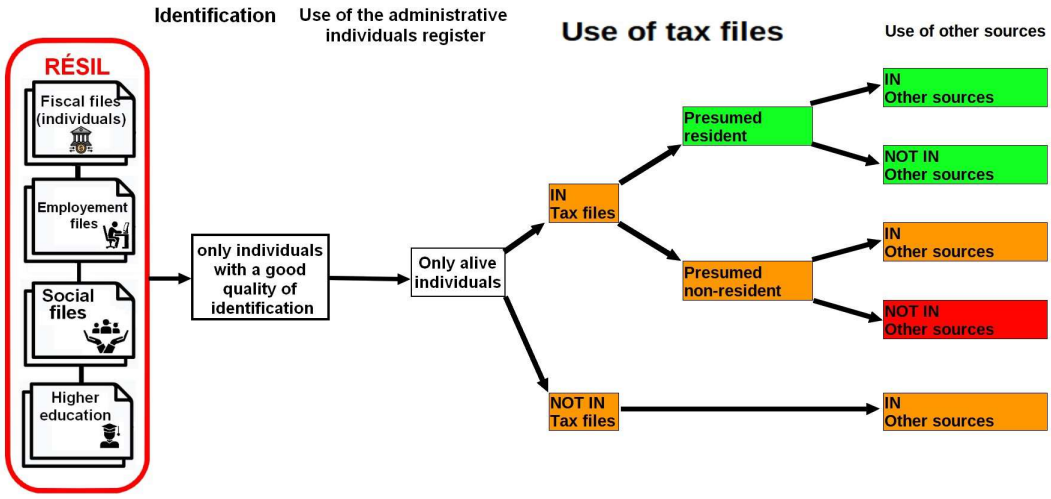
The first step is a simpler model, described in Section 4.2. It shows the interest of the method, but is still limited.

4.2 First tests of implementation of the "signs of life" method

4.2.1 First rules of decision

The first step consists of retaining an individual in the final population only if they are alive and if their identification quality is good in at least one source.

Figure 4.2 - Simplified representation of the Signs-of-Life method implementation



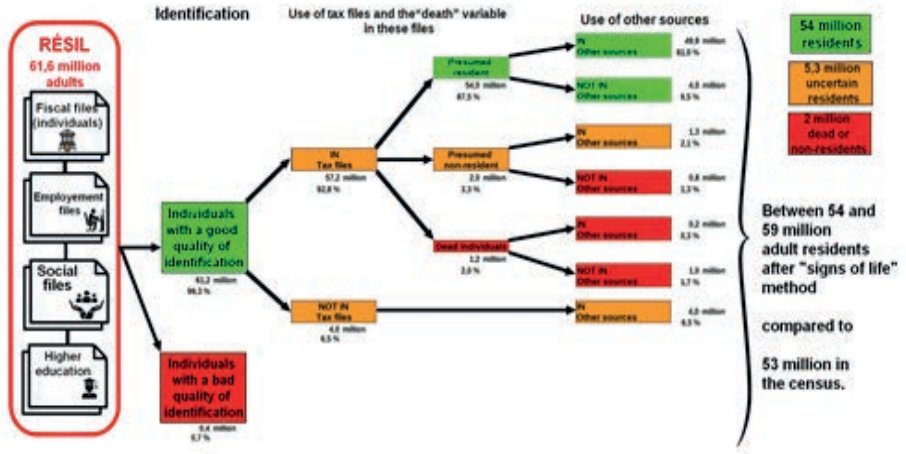
Then, a high priority is given to fiscal files because for now, tax files is the only source for which we have determined if an individual resides or not in the French territory.

The schema shows a segmentation of the population into two groups based on the tax source:

- Individuals residing in France according to the tax source;
- those presumed non-residents based on several variables directly present in the tax source.

For other administrative sources apart from fiscal sources, the presence of an individual in the reference population is uncertain.

Figure 4.3 - Results due to the use of the Signs-of-Life method



Source: Own computation

4.2.2 The Signs-of-Life method: first results

For this initial experimentation and due to technical reasons, it was not possible to use the administrative directory of individuals to account for deaths. Therefore, we use the "death" variable directly present in the tax files. As a result, the vital status of individuals available only in other administrative sources but absent from fiscal sources is unknown.

This first implementation of the signs of life method counts between 54 and 59 million resident adults on the French territory, compared to the 53 million resident adults, census recorded in 2020.

For the 5.3 million individuals with uncertain resident status, we need to explore further and establish new decision criteria based on other administrative sources available to us, excluding fiscal sources.

5. The outlook to improve the quality of the coverage of the universe of reference

5.1 Use a more complex Signs-of-Life method to reduce the over-coverage

In Figure 5.1, we can observe that using multiple sources improves coverage but also introduces a risk of over-coverage. The graph illustrates:

- In blue: the population of Résil segmented by age group, where all individuals present in administrative sources are retained.
- In red: the Résil population after applying the Signs-of-Life method.
- In yellow: the population measured by the census.

This initial application shows the interest of the method, as it reduces over-coverage across all age groups. However, it is not yet sufficient to match the levels of the census population, indicating the need to go further. Our methodological challenges are as follows:

- a better identification tool, using address and composition of dwellings in addition to name, surname, date and place of birth;
- more precise rules for the determination of usual residence, to be fully compliant with the concept used for the census: for example, in the case of young people we will consider that their usual residence is where they live for their studies, and not with their parents, even if they are still registered with them in the fiscal files;
- the application of a more sophisticated model for the signs of life, with different weighting for each source, and taking into account the presence the previous year, like in this model used in Estonia:

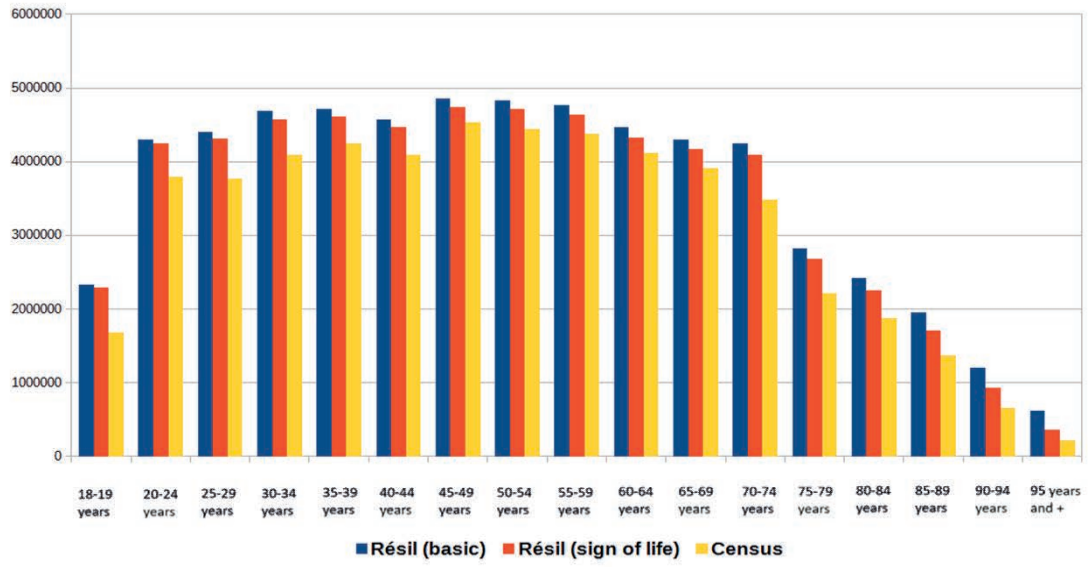
$$I(i,t) = \alpha \cdot I(i,t-1) + \beta \cdot \sum_{k=1}^{k=n} a_k(i,t) \cdot E_k(i,t)$$

- $I(i,\ t)$ is the index of residence of the individual I, for year t.
- $E_k\ (I,\ t)$ is the sign of life of individual I, in source k, for year t; a_k is the weight of source k in the model.
- α and β are the respective weights of the residence index for the previous year and the synthetic sign of life in the considered year.

The population of reference of the year t contains the set of individuals for whom $I(I,t) \geq S$ where S is a threshold defined a priori. To define this threshold, several tests will be carried out using data from administrative records and the annual census survey.

The reference index may be calculated multiple times for a given reference year. Indeed, for a given year, not all sources will be available at the same time and the determination of several reference populations is envisaged (*e.g.* provisional, semi-definitive and definitive reference population). The threshold for the definition of residents from the residence index may vary according to the type of population. The residency index will therefore be dated.

Figure 5.1 - Estimation of the French population in 2020, by age group, based on the census and several Signs-of-Life models in Résil



Source: Own computation

5.2 Application of dual system estimation in Résil, based on the rolling census to have a better measure of the quality of the coverage

5.2.1 Methodology

The initial implementation of the DSE model was applied to all small French municipalities surveyed in 2020. A Résil prototype was built over the year 2020 using various data sources. To build this prototype, data sources were matched together using the non-significant individual identifier. The initial decision rules implemented in the Signs-of-Life model described in paragraph 3 were applied to this prototype. Thus:

- Only living individuals were considered.
- Only individuals with good identification quality in at least one source were retained.
- Individuals presumed non-residents in fiscal sources and unknown in other sources were not retained.

The individuals derived from this prototype were located at their usual residences. For this purpose, high priority was given to the location provided by the tax files and only individuals located in a small municipality surveyed in 2020 were retained.

Finally, the population of the Résil prototype, restricted to small municipalities surveyed in 2020, amounts to 6,048,702 individuals aged 16 and older. In addition, 5,210,396 individuals of 16 and older were surveyed in 2020 in the small French municipalities

5.2.2 Application of the DSE model to small municipalities, surveyed in 2020

Among these 6,048,702 individuals:

- 5.17 million individuals are found in both the EAR and Résil (in green), accounting for 85.5% of the individuals (green part in Figure 5.2);
- 13.9% are only found in Résil (in yellow);
- and only 0.6% are exclusive to the EAR (in blue).

The matching rate is 99.3% for the EAR and 86.1% for Résil.

The 13.9 % ratio of non-matched Résil individuals is the result of several difficulties

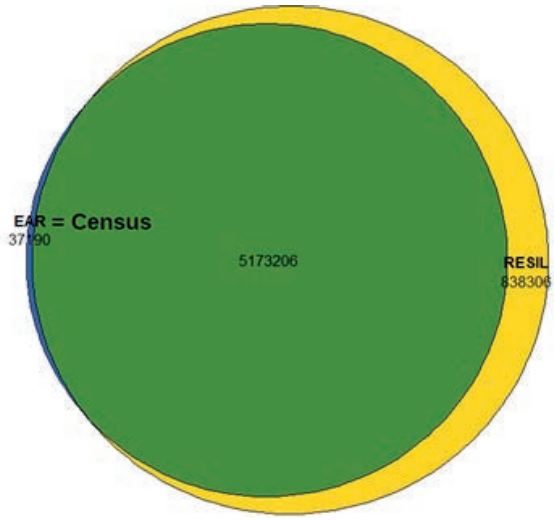
- impossibility to link a part of them with census survey, due to a total imputation of census data (matching errors);
- a bad localisation of people in Résil's prototype, some people being localised wrongly in municipality surveyed in census (over-coverage);
- over-coverage of Résil prototype.

It is necessary to deal with the ongoing problem of over-coverage in the Résil database. Indeed, the implementation of the DSE model has several application conditions, the most restrictive for us being the absence of over-coverage in each of the sources. In the census, the risk of over-coverage is very limited (some double accounts); in Résil, minimising over-coverage is the main challenge of the Signs-of-Life model. And good localisation rules will as well minimise the "local over-coverage", at the scale of a municipality.

Due to over coverage of the Résil prototype, the results of DSE are not significant for the moment.

If over-coverage remains significant, the implementation of the Trimmed Dual System Estimation model will be necessary.

Figure 5.2 - Proportion by age of individuals counted found in fiscal files (orange), found in other sources (green); not found in any file (blue). Year 2020



Source: Own computation

References

Durr, J.-M., O. Haag, F. Dupont and O. Lefebvre. 2022. "Setting up statistical registers of individuals and dwellings in France: approach and first steps". *Statistical Journal of the IAOS*, Volume 38, N. 1: 215-223.

Producing U.S. population statistics using multiple administrative sources

J. David Brown, Marta Murray-Close[1]

Abstract

We identify several challenges encountered when constructing U.S. administrative record-based (AR-based) population estimates for 2020. They include locational accuracy, person coverage and its consistency over time, filtering out non-residents and people not alive on the reference date, uncovering missing links across person and address records, and predicting demographic characteristics. We discuss several ways to address these issues. Regression results illustrate how the challenges and solutions affect the AR-based county population estimates.

Keywords: administrative records, population estimates, record linkage

1. Introduction

Administrative records (AR) have important advantages over survey-collected data when making population estimates. Some people missed in surveys appear in AR. Our AR-based U.S. population estimate for 2020 (hereafter the 2020 AR census) is higher than the 2020 Census count by 2.3 percent (339,200,000 compared to 331,400,000), and the difference in estimates is especially large for historically undercounted populations like young children, Blacks, Hispanics, and non-citizens. It is also much cheaper to produce AR-based estimates than to conduct a survey-style census, facilitating more frequent production of AR-based estimates.

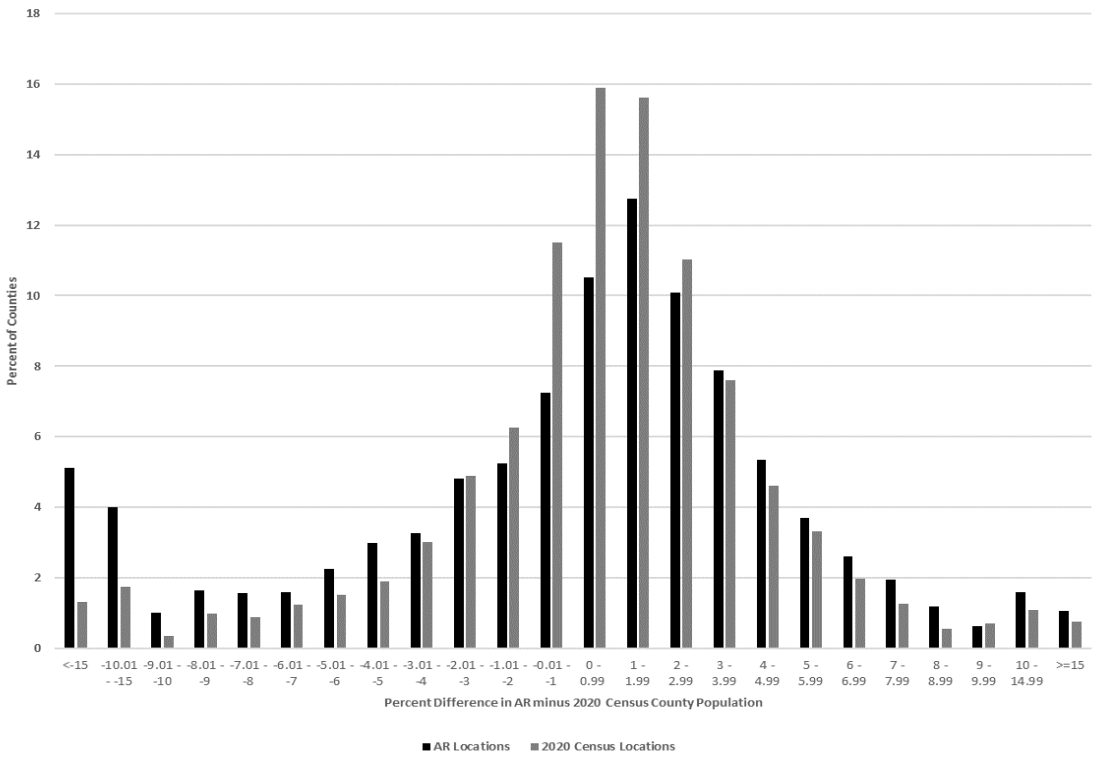
Producing AR-based estimates involves some challenges, however. To take one example, though AR population coverage is quite comprehensive at the national level, Figure 1.1 shows that coverage is very uneven across counties[2]. The AR census estimate is more than 15 percent lower than the 2020 Census count in 5 percent of counties. If AR people who can be linked to the 2020 Census are placed in the county where they were enumerated in the 2020 Census, only about 1 percent of counties have a more than 15 percent lower AR estimate. The share of counties with estimates that are between -1 and 1 percent different rises by 9.6 percentage points. AR estimates may thus be less accurate at lower levels of geography.

Besides locational accuracy, other challenges encountered when creating AR population estimates include person coverage completeness and its consistency across time, coverage of children in particular, distinguishing international migrants from continuous U.S. residents appearing infrequently in AR, record linkage, and the choice of demographic characteristics when multiple ones are reported or when they are missing altogether. In this paper we discuss these challenges and ways to address them.

1 J. David Brown, (j.david.brown@census.gov), Martha Murray, (marta.murray.close@census.gov), U.S. Census Bureau. Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the U.S. Census Bureau. The U.S. Census Bureau has ensured appropriate access and use of confidential data and has reviewed these results for disclosure avoidance protection (Projects 7516813 and 7516814: CBDRB-FY23-0253 and CBDRB-FY23-014-054). This paper draws heavily from Brown *et al.* (2023), and details about the methodology and further analysis can be found there.

2 The United States has 3,143 counties.

Figure 1.1 - Percent difference between AR and 2020 Census County Population, AR and 2020 Census Locations (a)



Source: U.S. Census Bureau, P.L. 94-171 Redistricting Data and 2020 AR census
(a) AR location is the county where the AR data record the AR person. 2020 Census location is the county where the AR person was enumerated in the 2020 Census, for the subset of AR people who can be linked to the 2020 Census. The percent differences are calculated using the mean of the two estimates as the denominator. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY23-0253).

2. Achieving comprehensive coverage

Thorough coverage is essential for making high-quality population estimates. No single U.S. AR source covers the entire population, however. Table 2.1 reports the percent of the 2020 AR census population covered by each individual AR source and the percent only in that source. Since some of the sources come from individual states, we also report the percentages among states included in that source. The table shows that if we were to use a single source, we could cover at most 76.4 percent of people in 2020, using Internal Revenue Service (IRS) 1040 individual tax return data. No other source comes close to covering the entire population.

To achieve more comprehensive coverage, we have combined 31 sources. Many of the sources cover a segment of the population. Three sources are for retirees (IRS 1099-R, Medicare, and Social Security Administration Master Beneficiary Record (SSA MBR)), covering about 15 percent of the population. Several sources (Department of Housing and Urban Development housing assistance programmes and state welfare programs) cover low-income people who may not need to file taxes, some cover parts of the prison population (Bureau of Prisons and U.S. Marshals Service), and others cover immigrants (*e.g.* U.S. Citizenship and Immigration Services (USCIS), Immigration and Customs Enforcement (ICE), and Customs and Border Protection (CBP)).

Table 2.1 - Percent of AR Census People Observed in each AR data source (a)

Data Source	All AR census people		AR census people in states covered by data source	
	Percent in this source	Percent only in this source	Percent in this source	Percent only in this source
Internal Revenue Service (IRS) 1040 forms	76.37	21.28	76.37	21.28
IRS 1099 forms	60.37	4.36	60.37	4.36
IRS 1099-R forms	14.08	0.00	14.08	0.00
Any IRS (1040/1099/1099-R)	88.52	36.92	88.52	36.92
Medicare Enrolment Database	15.11	0.09	15.11	0.09
Federal Housing Administration mortgage insurance	3.40	0.04	3.40	0.04
Department of Housing and Urban Development Public and Indian Housing Information Center, Tenant Rental Assistance Certification System, and Computerized Homes Underwriting Management System	0.91	0.08	0.91	0.08
Indian Health Service Patient Registration File	0.38	0.02	0.38	0.02
Social Security Administration (SSA) Master Beneficiary Record	14.99	0.04	14.99	0.04
SSA Supplemental Security Record and Special Veterans Benefits	2.56	0.14	2.56	0.14
Selective Service System	2.64	0.04	2.64	0.04
U.S. Postal Service National Change of Address	4.66	0.17	4.66	0.17
State Department passports	5.24	0.25	5.24	0.25
State Department Worldwide Refugee Admissions Processing System	0.01	0.01	0.01	0.01
Customs and Border Protection Arrival and Departure Information System (ADIS)	1.16	0.97	1.16	0.97
Immigration and Customs Enforcement (ICE) Enforcement and Removal Operations	0.24	0.19	0.24	0.19
ICE Student and Exchange Visitor Information System	0.24	0.10	0.24	0.10
U.S. Citizenship and Immigration Services (USCIS) naturalizations, lawful permanent residents, refugees, and asylees	3.59	0.19	3.59	0.19
USCIS people thought to be without lawful status on April 1, 2020	0.27	0.05	0.27	0.05
USCIS Temporary Protected Status	0.05	<0.01	0.05	<0.01
Department of Defense's (DoD) Defense Manpower Data Center	0.03	<0.01	0.03	<0.01
Department of Interior Law Enforcement Management Information System and Incident Management Analysis and Reporting System	<0.01	<0.01	<0.01	<0.01
Bureau of Prisons	0.11	0.06	0.11	0.06
U.S. Marshals Service	0.02	0.02	0.02	0.02
Veterans Service Group of Illinois (VSGI)	41.62	4.86	41.62	4.86
Alaska Permanent Fund Division (state source)	0.12	<0.01	58.10	0.35
State driver's licenses	2.84	0.15	58.41	3.03
Supplemental Nutrition Assistance Program, Temporary Assistance for Needy Family, and Women, Infants, and Children programs (state sources)	3.76	0.85	8.82	2.00
Census Household Composition Key	1.06	1.06	1.06	1.06
2016 Medicaid	0.33	0.33	0.33	0.33
SSA Numerical Identification file (NUMIDENT) (ages 0-1)	0.43	0.43	0.43	0.43

Source: U.S. Census Bureau, 2020 AR census

(a) The denominator for the percentages in the first two columns is the total number of people in the AR census, and the denominator in the last two columns is the number of people in the AR census in states covered by the particular source. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY23-0253).

Since unevenness of AR coverage at lower levels of geography is a particular concern, we study how the challenges and potential solutions discussed in this report affect it. Table 2.2 reports regressions showing associations between county characteristics representing challenges/solutions and the percent difference between the 2020 AR census and 2020 Census county population estimates. In the first two columns, we show regressions including one factor at a time, and the last two columns show a pooled regression. Counties, where higher shares of people in the AR census come only from state welfare programme (SNAP, TANF, and WIC) data, tend to have higher AR census estimates. In counties where the AR non-citizen share (especially those with unknown legal status) is higher, the overall AR estimate is relatively higher than the 2020 Census. This could be evidence that the addition of sources focussed on low-income people and non-citizens paid off in achieving better coverage than that achieved in survey-style data collection, or it alternatively could reflect errors in their placement by geography in the AR data. We will refer to results from these regressions throughout the remainder of the paper.

Table 2.2 - OLS regression estimates of associations between county characteristics and the percent difference between the AR Census and 2020 Census County Population Estimates (a)

	Regressions with variables representing single county characteristic		Multivariate regression	
	Coefficient	Standard error	Coefficient	Standard error
SNAP, TANF, and WIC	0.614	0.147	1.063	0.114
Naturalised citizens	0.308	0.203	-0.296	0.058
Legal non-citizens	-0.225	0.267	0.715	0.051
Non-citizens with unknown legal status	0.741	0.158	0.765	0.049
CHCK add	2.004	0.147	0.813	0.110
NUMIDENT ages 0 to 1 add	-1.391	0.323	-0.566	0.140
Driver's license	-1.075	0.148	0.683	0.139
2020 Census group quarters population	-0.950	0.055		
AR EPIK	0.654	0.169		
AR ITIN	0.669	0.124		
AR lacks MAFID	-0.520	0.050	-0.399	0.034
VSGI records without PIK	-0.383	0.040		
USPS does not recognise the address	-0.184	0.013	-0.116	0.014
USPS commercial address	3.053	0.740	0.391	0.396
USPS does not deliver mail to address	-0.029	0.025	-0.141	0.014
No vintage-2020 AR source	-0.254	0.071	-0.144	0.041
2019 Medicaid	-2.922	0.346	-1.105	0.104
Late 2019 IRS 1040	-1.701	0.463		
Two AR sources	0.312	0.158	-0.126	0.055
Three AR sources	0.266	0.056	0.286	0.044
Four AR sources	0.654	0.126	0.130	0.052
Five or more AR sources	-0.062	0.076	-0.161	0.037
Mean person-place probability	0.276	0.050		

Source: U.S. Census Bureau, P.L. 94-171 Redistricting Data and 2020 AR census

(a) These are OLS regressions with a dependent variable of the percent difference between the 2020 AR census and 2020 Census county population estimates, calculated using the mean of the two estimates as the denominator. The first two columns of regression results come from regressions run separately for each explanatory variable or group of explanatory variables, separated by a blank row. Specification selection for the multivariate regression was performed using a backward stepwise procedure, which dropped some of the variables. The standard errors are robust. Variable definitions are in Table 2.3. This table does not show the results for some of the variables in the multivariate regression. The complete results are in Tables 70 and 71 of Brown et al. (2023). The data presented in this table are approved for dissemination by the DRB (CBDRB-FY23-0253).

It is especially tricky to cover children using AR. Most sources contain only adults. IRS 1040 returns include dependent children, but they only appear with about a one-year lag after birth. We have addressed this in two ways. The Census Household Composition Key (CHCK) links children to one or both parents. When a child aged 18 or under does not appear directly in any AR source near the reference date, but one or both parents do, the child is placed at the same location(s) as their mother when present, and their father otherwise. With the CHCK-provided children, the AR census estimate for children aged 0 to 17 is just 0.04 percent below the 2020 Demographic Analysis (DA) estimate, while it is 4.97 percent below DA without them (Table 2.4)[3]. Table 2.2 shows that the AR census estimates tend to be higher relative to the 2020 Census in counties where more children are included through CHCK. Second, U.S.-born children under age 2 found only in the Social Security Administration Numerical Identification file (NUMIDENT) are included in the AR census in their city and state of birth. We do not include older children found only in the NUMIDENT, because most children aged 2 or above are dependents in IRS 1040 returns, and the likelihood that a child moved from their birth city increases over time. The AR census estimate for children aged 2 or under is 14.72 percent below DA before including children from the NUMIDENT and 0.97 percent below after doing so (Table 2.4). Table 2.2, though, shows a negative association between the share of children aged 0 to 1 found only in the NUMIDENT and the percent difference between the AR census and 2020 Census estimates. This could reflect low AR coverage in general in counties with more NUMIDENT adds[4].

3 The Census Bureau's Demographic Analysis estimates, which use comprehensive birth and death records for U.S.-born people, are a benchmark for decennial census quality and are often used when evaluating coverage of children (methodological details in Jensen *et al.*, 2020).

4 A difference between CHCK and NUMIDENT adds that could explain the contradictory results is that CHCK adds require that the parents be found in current AR sources, while NUMIDENT adds occur when the parents are not in current AR sources.

Table 2.3 – Definitions of county-level explanatory variables

Explanatory variable	Definition
Percent difference	100 * (AR census county population – 2020 Census county population)/[(AR census county population + 2020 Census county population)/2]
SNAP, TANF, and WIC	The percentage of AR census people in the county who are found only in SNAP, TANF, or WIC data.
Naturalised citizens	The percentage of the county's 2020 AR census people who are naturalised citizens.
Legal non-citizens	The percentage of the county's 2020 AR census people who are legal non-citizens.
Non-citizens with unknown legal status	The percentage of the county's 2020 AR census people who are non-citizens with unknown legal status.
CHCK add	The percentage of AR census people in the county who have no AR source with an address, but who are found in CHCK.
NUMIDENT ages 0 to 1 add	The percentage of AR census people in the county who are in the NUMIDENT, but no other AR source, and who were under the age of 2 on April 1, 2020.
Driver's license	The percentage of AR census people in the county who are found only in driver's license data.
2020 Census group quarters population	The percentage of 2020 Census people in the county who were enumerated in group quarters.
AR EPIK	The percentage of AR census people in the county who have an EPIK.
AR ITIN	The percentage of AR census people in the county who have an ITIN.
AR lacks MAFID	The percentage of AR census people in the county who do not have a MAFID.
VSGI records without PIK	The percentage of person records in VSGI with an address in the county that do not have a PIK.
USPS does not recognise the address	The percentage of addresses in the county that are not in the USPS Spring 2020 Delivery Sequence File. The omitted category for does not recognize address, commercial, and does not deliver mail to address is residential address with USPS delivery.
USPS commercial address	The percentage of addresses in the county that are classified as commercial in the USPS Spring 2020 Delivery Sequence File.
USPS does not deliver mail to address	The percentage of addresses in the county that are recognized by the USPS, but do not receive mail delivery according to the USPS Spring 2020 Delivery Sequence File.
No vintage-2020 AR source	The percentageof AR census people in the county who have no AR sources from 2020.
2019 Medicaid	Here 2016 Medicaid records are dropped from the AR census, and 2019 Medicaid records are added. This is the percent of AR census people in the county who are found only in 2019 Medicaid data.
Late 2019 IRS 1040	The percentage of people in the county who are in a late 2019 IRS 1040 return with an address in the county. A late return is one that was not delivered to the Census Bureau until 2021. The denominator is the number of people in the AR census in the county after adding the late 2019 IRS 1040 returns. People already in the AR census are placed only in their late 2019 IRS county when calculating this variable.
Two (three, four, five or more) AR sources	The percentage of AR census people in the county who have two (three, four, five or more) AR sources. The omitted category for number of AR sources is one AR source.
Mean person-place probability	The mean of the person-place probabilities from the random forest model among all person-address records in the AR census in the county.

Despite the use of so many sources, we undoubtedly omit some U.S. residents. Inclusion of more sources could remedy this. Table 2.2 shows a generally positive association between the number of AR sources and AR county-level coverage relative to the 2020 Census. About 3.0 percent of the population is found only in driver's license data in the five states providing those data (Table 2.1). The association between the share of people in the county contributed only by driver's license data and AR coverage is positive, though only in the multivariate regression (Table 2.2). About 2.0 percent of the population of the 22 states providing state welfare program data are found only in those sources (Table 2.1), and as discussed above, contributions from these sources are positively associated with AR coverage. Having these state sources in all 50 states and the District of Columbia could thus increase coverage. We include data on federal prisons, but data on state and local prisons and jails and other types of group quarters are not currently available. Brown *et al.* (2023) show that many of the people enumerated in group quarters in the 2020 Census have housing unit addresses in AR, which is an issue of locational accuracy rather than coverage, but the available AR sources surely miss some of the group quarters population. Table 2.2 shows a negative association between the share of the county population that is in group quarters in the 2020 Census and AR coverage. Some unemployment insurance recipients may be missed by the sources in the AR census, while they are included in the National Directory of New Hires, which is not currently authorised for Census Bureau use. The homeless population is unlikely to be well covered in available AR data (*e.g.* they may not file taxes). Local agencies working with the homeless could potentially provide data for that group.

Table 2.4 – AR census and demographic analysis child population estimates (a)

	Number of children	Percent difference with 2020 Demographic Analysis
	Age 0-2	
2020 Demographic Analysis middle estimates	11,420,000	0.00
2020 AR census without age 0-1 additions from NUMIDENT	9,854,000	-14.72
2020 AR census with age 0-1 additions from NUMIDENT	11,310,000	-0.97
	Age 0-17	
2020 Demographic Analysis middle estimates	74,660,000	0.00
2020 AR census without CHCK	71,040,000	-4.97
2020 AR census with CHCK	74,630,000	-0.04

Source: U.S. Census Bureau, 2020 Census Edited File, 2020 AR census, 2020 Demographic Analysis, 2020 Population Estimates Program, and 2020 Post-Enumeration Survey

(a) The 2020 Demographic Analysis middle estimates are from Jensen *et al.* (2020). The percent differences are calculated using the mean of the two estimates as the denominator. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY23-0253).

Another issue affecting coverage is record linkage. The Census Bureau's Person Identification Validation System (PVS) assigns a unique identifier called a Protected Identification Key (PIK) to personal records that can be linked to a set of AR reference files. We only include people who have been assigned a unique identifier, because this allows us to unduplicate AR to ensure that each person is included just once in the estimates, to verify that the person is eligible to be included in the estimates, and to link in locational and demographic characteristics about them. For data of vintages before 2009, PVS covered only people with Social Security numbers (SSNs), *i.e.* citizens and work-eligible non-citizens. For 2009 and subsequent vintages, this was expanded to include people with Individual Taxpayer Identification Numbers (ITINs)[5]. In an attempt to cover people without SSNs or ITINs in the 2020 AR census, AR with sufficient personally identifiable information (PII) to be linkable, but who did not receive a PIK, were unduplicated, assigned a unique identifier called an Enhanced Protected Identification Key (EPIK), and placed in a new set of reference files[6]. The PVS process was then re-run on any AR without a PIK but with sufficient PII to be linkable so that EPIKs could be assigned to individual AR. The 2020 AR census includes 6.7 million people with ITIN PIKs and 4.5 million with EPIKs, illustrating the value of expanding the PVS record linkage process. The additional linkage improves AR coverage at the local level as well, as the associations between AR ITIN and EPIK shares in the county and AR coverage are positive in Table 2.2.

When AR are ingested by the Census Bureau, an attempt is made to link the addresses to the Master Address File (MAF) of U.S. addresses and assign a MAFID, a unique address identifier. Not all addresses can be linked to the MAF. When initially putting together the 2020 AR census estimates, we excluded any AR without a MAFID[7]. The resulting estimates were below the Census Bureau's Population Estimates Program (PEP) estimates, and in some areas the differences were large. We then decided to add people lacking MAFIDs but who have a state of residence, increasing the population estimate by 8.7 million people[8]. The association between the share of AR people in the county without a MAFID and AR coverage is negative in Table 2.2, which could reflect difficulty in assigning PIKs to AR when address linkage is difficult, which we discuss next.

5 All citizens and work-eligible non-citizens can have SSNs. ITINs are nine-digit numbers in a publicly known range found in the SSN field of administrative records. They are issued by the Internal Revenue Service to people needing to pay taxes, but who are ineligible for an SSN.

6 Brown *et al.* (2023) provide details of the procedure and show descriptive statistics about people with SSN PIKs, ITIN PIKs, and EPIKs.

7 Every person in the 2020 Census is assigned a MAFID. If a person's address is not initially found in the MAF, follow-up activity is done to either match the address to an existing MAFID or to create a new one. We did not have the resources to do this for AR without a MAFID in the 2020 AR census project.

8 The 8.7 million figure includes the 1.5 million children under the age of 2 found only in the NUMIDENT shown in Table 2

The inability to assign PIKs to AR can cause omissions, so comprehensive person linkage is important for coverage. Veterans Service Group of Illinois (VSGI) third-party data provide a good example of this, as 34.7 percent of the records cannot be assigned a PIK and are thus not used in the AR census. VSGI records containing SSNs are assigned PIKs at high rates. When an SSN is not present, the ability to assign a PIK varies with the specificity of the address. Post Office (P.O.) boxes may be reused by people with different residential addresses. Rural routes may not always refer to a specific housing structure. Linkage is particularly difficult when no address is available. Table 2.5 shows that the percent of VSGI records receiving a PIK ranges from 72.6 percent for people with blank addresses to 88.3 percent for those with street addresses, for an overall average of 87.8 percent when an SSN is present. The variation across address types is much larger for records without an SSN (2.3 percent for a blank address compared to 51.1 percent with a street address), with a significantly lower average of 48.1 percent. P.O. boxes and rural routes are more common in rural areas, leading to lower AR coverage there. Table 2.2 shows that a county's share of VSGI records that lack PIKs is negatively associated with AR coverage. In addition, the table shows that AR coverage is lower in counties where the U.S. Postal Service (USPS) delivers mail to a lower percentage of addresses, whether they recognise the address or not. These areas are likely to be ones where P.O. box and rural route mailing addresses are more common.

Table 2.5 – Percent of third-party person records without SSNs receiving a PIK by address type (a)

Address type	Percent of records with SSN receiving a PIK	Percent of records without SSN receiving a PIK
Street address	88.26	51.12
Post Office box	86.39	33.40
Rural route	80.96	9.21
Blank street address	72.64	2.27
Total	87.79	48.05

Source: Veterans Service Group of Illinois (VSGI) third-party records, July 2020
(a) These percentages are among records that do not contain a SSN. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY23-0253).

One potential solution to the P.O. box linkage issue is to use a newly obtained file from USPS containing P.O. box numbers and the residential addresses of their owners, among people who obtain their P.O. box for free[9]. We plan to test the usefulness of this file.

People almost never change SSNs over their lifetime, so there should be little remaining duplication after unduplicating by SSN PIK. In contrast, a person can change ITINs, which can later be reused by others, or they can switch from using an ITIN to a SSN[10]. This could lead to some duplication in the estimates, and it could be one reason for the positive association between the AR ITIN share and relative AR coverage in Table 2.2[11]. Further research is needed to investigate the extent of duplication among people with ITIN PIKs and how it can be minimised.

Ongoing Census Bureau research is developing methods to accurately assign PIKs to a higher share of AR, which would improve AR population statistics coverage.

9 These people are not provided with mail delivery to their residences and are instead given free Group E P.O. boxes.

10 ITIN to SSN switches can happen when a person becomes eligible for an SSN.

11 Suppose a person uses one ITIN in AR in year *t-1* and a different ITIN or SSN in AR in year t. Since we use both year *t-1* and *t* AR sources in the estimates for year *t,* we include the person twice.

3. Excluding people not eligible to be counted

Though we aim to maximise coverage of people who should be included in population estimates on the reference date, we try to exclude people who are ineligible. One criterion is being alive on the reference date. We identified people who were born after April 1, 2020 (the 2020 Census reference date), using birth dates from the fourth quarter 2020 NUMIDENT. We identified people who died before April 1, 2020, using death dates from the fourth quarter 2020 NUMIDENT and information from the following administrative data sources: Bureau of Prisons, Department of Defense's Defense Manpower Data Center, IRS 1040 returns, Medicare, Selective Service System, and VSGI. In addition, we excluded people who lacked a death date in the NUMIDENT but who, based on their birth date, were 115 or older[12]. Table 3.1 shows the number of people excluded from the 2020 AR census because of death information from each AR source and the number and share of them who can be linked to a person in the 2020 Census. A tiny fraction of people identified as deceased in the NUMIDENT can be linked to 2020 Census records, while significant shares of those in Medicare, Bureau of Prisons, and VSGI can be linked, suggesting that more research is needed about the quality of death information coming from sources other than the NUMIDENT.

Table 3.1 - Source contributions to identification of deceased person records (a)

Source	This source's number of AR census people dropped because deceased	This source's number of AR census people dropped because deceased, linked to 2020 Census person	Percent linked to 2020 Census people
NUMIDENT	52,500,000	183,700	0.35
IRS 1040	404,000	365,900	90.57
Medicare Death	4,300	821	19.09
Medicare Part A Termination	94,000	72,290	76.90
Medicare Part B Termination	1,206,000	829,300	68.76
Medicare Parts A & B Termination	659,000	473,200	71.81
SSS	80	(D)	(D)
DMDC; IMARS	(D)	(D)	(D)
BOP	350	75	21.43
VSGI	72,500	60,850	83.93
Age over 114	19,000	324	1.71

Source: U.S. Census Bureau, 2020 Census Unedited File and 2020 AR census
(a) "(D)" signifies that the cell is suppressed because of disclosure avoidance. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY23-0253).

Very limited death information is available for people with ITIN PIKs or EPIKs, so erroneous inclusion of people who are not alive is more of a concern for those groups. Presence in or absence from current AR sources (signs of life) is the best information available on whether they are alive or not.

Filtering out people who could be alive, but who are not U.S. residents on the reference date, as well as people whose deaths were omitted from our death sources is discussed below.

4. Coverage consistency

Consistency of coverage over time is a second important characteristic of population estimates. Otherwise, population growth estimates would be inaccurate. Some sources are not available every year. For example, the SSA MBR data are not available before 2015 or between

12 As of April 1, 2020, the oldest known person in the United States was 114. A list of the oldest people in the United States can be found at https://en.wikipedia.org/wiki/List_of_the_verified_oldest_people, viewed on July 29, 2022.

2016-2018, and the available years for state sources vary by state. Inconsistent availability could lead to variable coverage.

People appearing in AR sources in 2019 but not 2020 is a sign of inconsistent AR coverage. Table 2.2 shows that the share of such people in the 2020 AR census in the county is negatively associated with AR coverage.

The data-sharing agreements for several sources used in the 2020 AR census are currently inactive[13]. Table 4.1 shows that the 2020 AR census total population estimate drops by 1.7 percent (5.7 million) when excluding sources no longer accessible because of inactive data-sharing agreements. The declines are quite uneven across demographic groups. They are larger for people under age 45, Hispanics, non-Hispanic Asians, and non-Hispanic Some Other Race (SOR).

Table 4.1 - Percent difference in population estimates without no longer accessible sources and with late-arriving sources (a)

	Percent difference in population when dropping sources no longer accessible	Percent difference in population when adding late-arriving sources
Total population	-1.69	0.71
Male	-1.73	0.77
Female	-1.66	0.64
Age 0-2	-2.24	1.06
Age 3-5	-1.71	0.88
Age 6-14	-1.88	0.92
Age 15-17	-2.36	0.93
Age 18-24	-2.81	0.48
Age 25-34	-2.39	0.53
Age 35-44	-1.87	0.70
Age 45-54	-1.34	0.67
Age 55-64	-0.97	0.74
Age 65-74	-0.87	0.62
Age 75 and over	-0.67	0.76
Hispanic	-4.80	1.08
Non-Hispanic American Indian and Alaska Native	-0.89	2.76
Non-Hispanic Asian	-4.47	1.22
Non-Hispanic Black	-0.75	1.00
Non-Hispanic Native Hawaiian and Pacific Islander	-0.78	1.54
Non-Hispanic White	-0.51	0.41
Non-Hispanic Some Other Race	-4.89	1.06
Non-Hispanic Two or More Races	-1.16	0.45

Source: U.S. Census Bureau, 2020 AR census
(a) The data presented in this table are approved for dissemination by the DRB (CBDRB-FY23-0253).

Some sources are not received in time to be included in estimates needed by a certain date. Two sources arrived too late to be included in the first 2020 AR census estimates that were needed by December 31, 2020. One is 2019 Medicaid data, and the other is the subset of the Tax Year 2019 IRS 1040 tax returns that had been received by IRS in 2020 but were not yet processed[14]. If they are included, the total population rises by 0.7 percent (2.4 million). The gain is larger for children and people who are neither non-Hispanic White nor non-Hispanic Two or More Races (Table 4.1). Having more AR people found only in 2019 Medicaid data or late Tax Year 2019 IRS 1040 returns is associated with lower county-level AR coverage when using only sources available in 2020 (Table 2.2).

The coverage of any given source varies across years. For example, the number of people in IRS 1040 tax returns typically changes little from year to year (Table 4.2). The IRS number does

13 All data obtained with the assistance of Executive Order 13880 were incorporated into the analytic files for the 2020 AR census by January 12, 2021.

14 Medicaid is a government health insurance program for people with low income. Before receiving 2019 data, the most recent Medicaid data available were from 2016. We included the 2016 data in our estimates. We replaced 2016 with 2019 Medicaid data when making the estimates including late-arriving sources.

not track well with Census Bureau population estimates, however, it sometimes declinesF when the population estimates are increasing. The IRS number jumps by 10 million between 2019 (Tax Year 2018) and 2020 (Tax Year 2019), then falls by 7.8 million between 2020 (Tax Year 2019) and 2022 (Tax Year 2021), while the population estimates rise each year. The number of people with ITINs appearing in IRS 1040 returns steadily declines after 2012. The number in 2022 represents only 45.7 percent of the 2012 peak. In contrast, official estimates suggest there was little change in the size of the unauthorised immigrant population during the 2010s (Baker 2021*a*)[15].

Table 4.2 - People in IRS 1040 returns compared to Census Bureau population estimates (a)

Year	People in IRS 1040's	ITINs in IRS 1040's	Vintage 2019 Population Estimates	Vintage 2022 Population Estimates
2010	272,500,000	8,527,000	309,300,000	N.A.
2011	275,600,000	9,109,000	311,600,000	N.A.
2012	277,600,000	9,480,000	313,800,000	N.A.
2013	277,000,000	8,924,000	316,000,000	N.A.
2014	278,200,000	8,491,000	318,300,000	N.A.
2015	278,900,000	8,145,000	320,600,000	N.A.
2016	280,900,000	7,848,000	322,900,000	N.A.
2017	279,800,000	7,097,000	325,000,000	N.A.
2018	281,600,000	6,465,000	326,700,000	N.A.
2019	281,500,000	5,160,000	328,200,000	N.A.
2020	291,500,000	4,481,000	329,900,000	331,500,000
2021	289,500,000	4,559,000	N.A.	332,000,000
2022	283,700,000	4,334,000	N.A.	333,300,000

Source: IRS 1040 returns; U.S. Census Bureau, vintage-2019 and vintage-2022 Population Estimates Program estimates
(a) The data presented in this table are approved for dissemination by the DRB (CBDRB-FY23-014-054).

One way to improve consistency is to use enough sources so that as many people as possible are covered by multiple sources. Consistent with this, Table 2.2 shows that the share of people with multiple AR sources is generally positively associated with AR coverage[16]. People in multiple sources are much less likely to exit the data from one year to the next for a reason other than death or to appear in the data for a reason other than birth (Table 4.3). For some groups the AR census has such redundancy. Though IRS 1099-R forms, SSA MBR, and Medicare data each cover 14-15 percent of the population, each source uniquely covers only <0.01, 0.04, and 0.09 percent of the population, respectively, because the three sources cover nearly the same group of people (Table 2.1)[17].

Table 4.3 - Percent distribution of number of AR Sources for people in both 2020 and 2021 estimates compared to others (a)

Number of sources	In 2020 and 2021 data	In 2020, not in 2021, not death	Not in 2020, in 2021, not birth
1	20.34	88.71	70.86
2	42.47	8.53	14.95
3	25.84	1.75	8.26
4	7.79	0.64	2.85
5	2.48	0.25	1.43
6	0.73	0.09	1.02
7	0.30	0.03	0.57
8	0.04	0.01	0.05
9	0.01	<0.01	0.01
10	<0.01	(D)	<0.01
11	<0.01	0.00	(D)
12	(D)	0.00	0.00

Source: U.S. Census Bureau, administrative records data for 2020 and 2021 population estimates
(a) "(D)" signifies that the cell is suppressed because of disclosure avoidance. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY23-014-054).

15 Most ITIN holders are likely to be undocumented immigrants.

16 The people with at least one 2019 or 2020 source are included in the 2020 estimates. However, if a large fraction of the county's people in the AR census are in just one source, that may make it more likely that many people in the county are in no sources.

17 The SSA MBR data include retirees receiving social security payments, the IRS 1099-R forms report retirement savings account distributions, and Medicare is a government health insurance program for people aged 65 or over.

5. Distinguishing infrequently appearing U.S. residents from International migrants

An important issue to address when attempting to maximise coverage of U.S. residents and maintain consistency over time is how to treat people who do not appear in AR consistently over time or who have both U.S. and foreign addresses. We would like to include continuous U.S. residents in population estimates even if they have not appeared recently in any AR source. International migrants, however, should be excluded on reference dates when they are not U.S. residents. We use the Customs and Border Protection Arrival and Departure Information System (ADIS) dates of entry and exit from the United States among non-immigrant visa holders (*e.g.* student and temporary work visa holders), the group most likely to enter and exit the country frequently, to exclude people who were not in the United States on the reference date[18]. Second, we require that people have a U.S. address in an AR source between January 2019 and October 2020, with a few exceptions[19]. This is not perfect, as some continuous U.S. residents may have last appeared in AR prior to 2019. Some people who are not covered by the ADIS data that we accessed may have emigrated before the reference date despite having appeared with U.S. addresses in 2019-2020 AR sources.

It would be preferable to use ADIS data on people with all immigration statuses so that we could filter out more non-residents[20]. ADIS alone would still not be sufficient to identify all non-residents, though, because it misses some entry and exit activity along the U.S.-Mexico border (Baker 2021*b*).

We are developing models to predict the probability that a person is a U.S. resident on the reference date. The models predict presence in the decennial census, which we assume to be strong evidence that a person is a living U.S. resident on Census Day. We consider the lack of a link between a person in AR to the decennial census to be weaker evidence that a person is not a living U.S. resident, because the person could also have a decennial census record without a PIK or be a living U.S. resident who was omitted from the decennial census. One set of models is for people who have both a U.S. and foreign address in the reference year. Predictors include factors such as the number of AR sources with U.S. addresses for the person in the reference year and the number in the previous year, whether the person had a foreign address in the previous year, birthplace, citizenship, and other demographic characteristics. A second set of models is for people who appeared in AR prior to the reference year, but not during it. Explanatory variables include how long ago the person last appeared in AR, the number of sources the person appeared in during that year, whether the person had a foreign address in the meantime, birthplace, citizenship, and other demographic characteristics[21]. We plan to adjust the probabilities to take into account the possibility that the person is in the decennial census with a missing link or was erroneously omitted from it.

18 Brown *et al.* (2023) describe the rules used to exclude non-immigrant visa holders not appearing to be U.S. residents on the reference date.

19 The exceptions are children added from CHCK or the NUMIDENT discussed above and people in the 2016 Medicaid file.

20 The ADIS data that we accessed contain only non-immigrant visa holders. The data sharing agreement is inactive, so we currently have no access to ADIS data.

21 Note that these models not only address emigration, but also deaths missed by the death information we currently use in the filtering process. Death is a reason for being omitted from current AR sources.

6. Achieving locational accuracy

Not only do we aim to include all U.S. residents in population estimates, but we also try to place them in the locations where they reside on the reference date. Some people have addresses that cannot be geocoded at lower levels of geography, while others have multiple addresses in AR. The share of people in the 2020 AR census with no MAFID (and thus no subcounty geocodes) is 2.6 percent, 51.6 percent have one MAFID, and 45.8 percent have more than one. In addition to the 8.7 million people without a MAFID, another 418,000 people have MAFIDs that lack subcounty geography in the MAF, for a total of 9.1 million without subcounty geography[22].

MAF coverage emphasizes physical residential addresses, which are most relevant for censuses and household surveys. AR contain mailing addresses, which are not always where a person lives. For example, a person could have a Post Office (P.O.) box mailing address in AR. The MAF contains very few P.O. boxes, however. About 59.9 percent of AR addresses lacking MAFIDs have P.O. boxes. Among AR people linked to the 2020 Census, 19.6 percent of those without AR MAFIDs are in multiunit buildings with 2 to 4 units, mobile homes, and group quarters in the 2020 Census, compared to just 8.0 percent of people with AR MAFIDs. Such structures may not have clear addresses[23].

Unexpected characters in the address field can cause an AR address to not have a MAFID. Many AR addresses in Hawaii and Queens, New York lacking MAFIDs have dashes in their house numbers. Our project team standardised these addresses in such a way that they could be linked to the MAF, and many were assigned MAFIDs.

We are working on a method to impute tract-level geography for people without subcounty geography[24]. The model predicts whether any given potential tract the AR person could be in is the tract where they were enumerated in the 2020 Census. For each tract that an AR person could potentially be in, we construct a set of tract-level predictors, including shares of people by race/ethnicity, age group, and sex, as well as interactions between those variables and the person's own race/ethnicity, age group, and sex. We also include the shares of addresses in the tract that are of different types (*e.g.* street addresses, P.O. boxes, and rural routes) and interactions with the person's own address type. The rationale is that people may be more likely to live in tracts where others have similar demographics and address types to them.

To handle multiple addresses, we estimate a person-place model to predict the probability that a given address is the person's address on the reference date. Predictors include variables such as AR source indicators, the time between an administrative record's vintage and the reference date, and the number of AR addresses the person has. The test dataset is the American Community Survey (ACS), a continuously running survey of about 3.5 million U.S. households per year. The model predicts the probability that a particular AR address for a person is their address in the ACS, using the ACS tabulation month as the reference date. The coefficients are applied to all AR person-address pairs. Each person's modelled probabilities are normalised to sum to one, so that the total weight for each person in the estimates is one. We include each of the person-place pairs in the estimates, weighted by the normalised person-place probabilities. Fractions of a person may be included in multiple locations.

22 About 300,000 of the 8.7 million people without a MAFID have only state geography.

23 We have found discrepancies in how group quarters addresses are handled between AR and the MAF. The Bureau of Prisons data provide a single address for each prison complex, while the MAF has individual entries for each building in the complex.

24 Counties are divided into tracts, each with a population ranging from 1,200 to 8,000 people (U.S. Census Bureau 2023).

The average person-place probability (prior to normalisation) among people with an address in the county in AR (*i.e.* confidence in locational placement) is positively associated with AR coverage (Table 2.2).

7. Achieving demographic characteristic accuracy

We assign age-group, sex, race, and ethnicity probabilities to each person in the 2020 AR census using a combination of AR data and previously collected decennial census and household survey data. A key challenge in compiling accurate demographic information is missing data: some people lack demographic records from AR data as well as demographic reports from census and survey data. The challenge of missing demographic data is most pronounced for people with ITIN PIKs and EPIKs, who are less likely than people with SSN PIKs to be included in our demographic data sources, and for race and ethnicity, which are missing at higher rates than age and sex even among people who are included.

Age is well measured, because the date of birth from which age is calculated is time-invariant, and reported values are available for nearly everyone. For SSN PIKs, the date of birth comes from the NUMIDENT. For EPIKs, it comes from the AR source where the person-place pair was found, the 2010 Census, or the ACS. Date of birth is a key variable in the PVS process, so few records without date of birth receive a PIK or EPIK. For ITIN PIKs, age comes from the same sources as it does for EPIKs, but it is missing at higher rates. We model the probability a person is in different age groups when age is missing, using first name and local area characteristics as predictors.

Sex is also well-measured because it rarely changes or is missing. Data on sex comes from the same sources as data on age. The model for sex adds a middle name to the characteristics used in the age model.

In contrast to sex and age values, race and ethnicity values are often missing. For SSN PIKs, race and ethnicity data come from the Census Best Race File (Ennis *et al.* 2018), which consolidates information from AR data, household surveys, decennial census responses, and third-party data using a set of business rules. For EPIKs, they come from the AR source where the person-place pair was found, the 2010 Census, or the ACS. This is also true for ITIN PIKs, with the exception that race and ethnicity are obtained from the Best Race File when available. The race and ethnicity models use last name and local area characteristics.

An additional challenge in assigning accurate race and ethnicity values is that reporting of these characteristics can change over time (Liebler *et al.* 2017). Table 7.1 shows the percent differences between the AR census and the 2020 Census by age group, sex, and race/ethnicity when the AR census values are constructed in two different ways. The first uses previously collected values from Census Bureau decennial censuses and surveys and from AR sources, and modelled imputations for missing values. The second uses 2020 Census values for the 265,600,000 people (78.3 percent of the AR census) who can be linked between the AR census and 2020 Census. There is virtually no difference in the sex and age group results when switching to 2020 Census values, reflecting a high degree of consistency in those variables. The differences for race/ethnicity are quite large, exceeding 10 percentage point points for non-Hispanic Two or More Races, non-Hispanic American Indian and Alaska Native, non-Hispanic Asian, and non-Hispanic Some Other Race.

Table 7.1 - Percent difference between 2020 AR Census and 2020 Census by Sex, Age, Race, and Ethnicity (a)

	2020 AR census using previously collected values and imputations	2020 AR census using 2020 Census values or linked people	Percentage point difference between using 2020 Census values and previously collected/imputations
Male	3.74	3.74	0.00
Female	0.94	0.94	0.00
Age 0-2	5.64	5.64	0.00
Age 3-5	5.56	5.40	-0.16
Age 6-14	0.82	0.71	-0.11
Age 15-17	-0.62	-0.70	-0.08
Age 18-24	-0.71	-0.64	0.06
Age 25-34	5.70	5.70	0.00
Age 35-44	4.68	4.77	0.09
Age 45-54	4.87	4.92	0.05
Age 55-64	2.34	2.32	-0.02
Age 65-74	-2.79	-2.79	0.00
Age 75+	-1.38	-1.38	0.00
Hispanic	14.96	10.77	-4.19
Non-Hispanic American Indian and Alaska Native	54.43	17.50	-36.93
Non-Hispanic Asian	-18.91	-2.22	16.69
Non-Hispanic Black	6.61	4.67	-1.94
Non-Hispanic Native Hawaiian and Pacific Islander	3.79	2.38	-1.41
Non-Hispanic White	1.76	-0.37	-2.13
Non-Hispanic Some Other Race	24.28	11.86	-12.42
Non-Hispanic Two or More Races	-71.42	-4.84	66.58

Source: U.S. Census Bureau, 2020 Census Demographic and Housing Characteristics file and 2020 AR census
(a) The percent differences are calculated using the mean of the two estimates as the denominator. The percentage point differences in column three are calculated as the column two value minus the column one value. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY23-0253).

There are several potential reasons for the large race and ethnicity differences with the 2020 Census for the same people. The 2020 Census questions and processing method are different than what was used previously (Jones *et al.* 2021), giving more opportunities to mention multiple races. Different respondents may answer another way about a person's characteristics across surveys (*e.g.* a parent, roommate, or neighbour may have reported about the person in the 2010 Census, while the person may have self-reported in the 2020 Census). The person may change how they view their race and ethnicity. In addition, AR sources use fewer race categories than Census Bureau decennial census and household surveys.

We plan to test different race and ethnicity modelling approaches. For example, rather than using business rules to decide which value to use, we could estimate a hidden Markov model or a latent class model. Bycroft *et al.* (2023) used the latter for their New Zealand AR population estimates.

8. Conclusion

This paper discusses several challenges encountered when constructing AR-based U.S. population estimates, including how to achieve and maintain comprehensive coverage, filter out people who should not be in the estimates, place people where they reside on the reference date, and attach accurate demographic characteristics.

Our analysis suggests that the estimates benefit from using multiple sources in several ways. No one source covers everyone, so combining sources achieves more comprehensive coverage. Since a source's coverage changes over time and may not be available in all years, building in redundancy with multiple sources covering a person can improve estimate consistency over

time. Having multiple sources can improve prediction about where a person resides on the reference date. If, for example, multiple sources place the person at the same address at the same time, that improves confidence that the address is their residence at the time. Building in redundancy with multiple sources mitigates the operational risk of dependency on outside data providers to continue providing data in a timely fashion.

Using multi-source data leads to choices about how to handle discrepancies across the sources. We have developed models to assign probabilities to each of the person's AR addresses. The Census Bureau has developed business rules to select demographic characteristics when they are discrepant across sources. We are considering a latent class model as an alternative to the business rules.

The effort to achieve comprehensive coverage increases the risk of erroneous inclusions. We are developing models to predict the probability that a person is a living U.S. resident on the reference date.

We find that address linkage is difficult in rural areas where the USPS does not deliver mail directly to residences. In such places there are significant discrepancies in placement of people in AR and the 2020 Census, as well as an inability to assign PIKs to some AR. The latter results in omissions of people from the AR-based estimates, contributing to lower estimates than the 2020 Census counts in those areas. This calls for more effort to link AR people in rural areas to their residential addresses.

References

Baker, B. 2021*a*. "Estimates of the Unauthorized Immigrant Population Residing in the United States: January 2015-January 2018". *Population Estimates*. Washington, D.C., U.S.: Department of Homeland Security, Office of Immigration Statistics. https://ohss.dhs.gov/sites/default/files/2023-12/unauthorized_immigrant_population_estimates_2015_-_2018.pdf.

Baker, B. 2021*b*. "Population Estimates of Nonimmigrants Residing in the United States: Fiscal Years 2017-2019". *Population Estimates*. Washington, D.C., U.S.: Department of Homeland Security, Office of Immigration Statistics. https://ohss.dhs.gov/sites/default/files/2023-12/ni_population_estimates_fiscal_years_2017_-_2019v2.pdf.

Brown, J.D., S.R. Cohen, G. Denoeux, S. Dorinski, M.L. Heggeness, C. Lieberman, L. McBride, M. Murray-Close, H. Qin, A.E. Ross, D.H. Sandler, L. Warren, and M. Yi. 2023. *Real-Time 2020 Administrative Record Census Simulation*. Washington, D.C., U.S.: U.S. Census Bureau. https://census.gov/programs-surveys/decennial-census/decade/2020/planning-management/evaluate/eae/2020-admin-record-census-simulation.html.

Bycroft, C., J. Elleouet, and H. Tran. 2023. *Harmonising Ethnicity from Multiple Administrative Sources Using Latent Class Models*.Wellington, New Zealand: Stats NZ. https://stats.govt.nz/research/harmonising-ethnicity-from-multiple-administrative-data-sources-using-latent-class-modelling/.

Ennis, S.R., S.R. Porter, J.M. Noon, and E. Zapata. 2018. "When Race and Hispanic Origin Reporting are Discrepant Across Administrative Records and Third Party Sources: Exploring methods to Assign Responses". *Statistical Journal of the IAOS*, Volume 34, N. 2: 179-189.

Jensen, E.B., A. Knapp, H. King, D. Armstrong, S.L. Johnson, L. Sink, and E. Miller. 2020. *Methodology for the 2020 Demographic Analysis Estimates*. Washington, D.C., U.S.: U.S. Census Bureau. https://www2.census.gov/programs-surveys/popest/technical-documentation/methodology/2020da_methodology.pdf.

Jones, N., R. Marks, R. Ramirez, and M. Rios-Vargas. 2021. "2020 Census Illuminates Racial and Ethnic Composition of the Country". *America Counts: Stories Behind the Numbers*, Washington, D.C., U.S.: U.S. Census Bureau. www.census.gov/library/stories/2021/08/improved-race-ethnicity-measures-reveal-united-states-population-much-more-multiracial.html.

Liebler, C.A., S.R. Porter, L.E. Fernandez, J.M. Noon, and S.R. Ennis. 2017. "America's Churning Races: Race and Ethnicity Response Changes Between Census 2000 and the 2010 Census". *Demography*, Volume 54, N. 1: 259-284.

U.S. Census Bureau. 2023. "Glossary". *U.S. Census Bureau website*. https://census.gov/programs-surveys/geography/about/glossary.html#par_textimage_13.

SESSION 2

Innovative data for Official Statistics: Methodological challenges

Introduction to Session 2 invited talks

Brunero Liseo, Li-Chun Zhang[1]

Abstract

The main theme of this session is the exploration of the potential of new statistical techniques to improve the production of official data. In particular, some experiments in progress in Istat will be discussed for the integration of data from interviews with data collected by sensors and the use of photographic information for the quantification of green areas in urban areas. The session ends with a general reflection on the use of Machine Learning techniques in the context of Official Statistics.

Keywords: Sensor data, Machine Learning, vegetation indices, smart survey, explainability

1. Synapse of the session

The evolution of statistical methodology in Data Science has been supported by the availability of an extremely huge computational power not comparable with the past. It has made the use of modern Data Science methods absolutely essential for the production of Official Statistics. Data Science allows automated collection, processing, and analysis of large amounts of data and timely reporting are now easier to produce (De Boom and Reusens 2023).

However, the quality of official data provided by Data Science and Machine Learning methods rely on the accuracy of the data sources and the statistical sense of the underlying procedures.

In this session two different possibilities will be described. Both of them are object of experimental trials in Istat. The first paper is delivered by C. De Vitiis and it deals with the use of data collected from sensor receptors (smartphone, tablet, etc.). Here the goal is twofold: on one hand the automatic collection of data alleviates the burden of people involved in the survey. On the other hand, the integration of those new data with information provided by light questionnaires might significantly improve the accuracy of the information. It is obvious that the participation of people to this kind of surveys implies a different role of the respondent who are now requested to provide an informed consent to sharing rather than a mere participation.

The second paper is presented by F. De Fausti and it deals with the use of digital information to quantify the amount of green areas in an urban context. It is performed using the *Normalised Difference Vegetation Index* (NDVI) defined as:

$$NDVI = \frac{NIR - RED}{NIR + RED}$$

where NIR and RED are the percentage of Near Infra-Red and Red components of images taken with a resolution of 20 cm pixel on the ground (urban areas) and 50 cm for the extra-urban areas. Other indices, also involving the green and blue parts of the spectrum are also available.

The values of NDVI are then used as an input for obtaining a classification of urban areas in terms of abundance of green parts. Several statistical methods are considered, including

1 Brunero Liseo (brunero.liseo@uniroma1.it), Sapienza Università di Roma, Italy; Li-Chun Zhang (L.Zhang@soton.ac.uk), Statistisk Sentralbyrå, Norway and University of Southampton, UK.

kernel density estimation and cluster analysis in order to better estimate the threshold level that characterises a green area.

Our opinion here is that the natural model for this kind of data would be a mixture with an unknown number of components and where the actual parameters of interest are not the parameters of the components of the mixture but - rather - the fractions of units falling in each category. This opens the way to a sort of non (semi-) parametric mixture model which could be implemented either in a classic or a Bayesian framework.

The third paper is a discussion on the potential benefits that the Machine Learning (ML) philosophy can bring in the field of Official Statistics. The ongoing debate is centred on the fact that ML algorithm are often perceived as black boxes. As such, they may conceal implicit assumptions that jeopardise the objectivity of the released data. The Author illustrates his positive viewpoint.

2. Further discussions

2.1 Smart surveys

Unlike survey data that arise from probing the respondents for the required information, smart surveys can enable statistical data to be derived from automatically generated and recorded digital footprints. "Such a *content-orientated* approach can have advantages compared *unit-orientated* surveys, provided the target measurement is factual and the digital records can form a reliable basis of the responses that one ideally could have obtained by surveying the subjects" (Zhang and Haraldsen 2022). The required transition from informed consent of participation (by answering the questions directly) to informed consent of sharing (of ones' digital footprints) would have many far-ranging implications.

Sharing can be achieved either actively or passively. For example, a form of active sharing is data donation (*e.g.* Boeschoten *et al*. 2022), whereby a participant willingly undertakes to obtain personal digital records from the relevant platform (or service provider) and deliver them to the analyst (or statistical agency). Clearly, this would generate its own challenges, with respect to response rate and selection bias, which need to be handled.

The data to be shared passively can either reside on a platform operated by the statistical agency or the third parties (such as mobile phone operator, Google, Facebook). In the latter case, the statistical agency would need to obtain the data from a third party on behalf of those individuals who have consented to sharing their data. Many new and difficult issues would need to be overcome before this can become a viable approach, including the necessary legal framework, the agency's technical ability and capacity to handle the raw data, beyond the statistical issues of representation and measurement.

Research and development efforts in these directions deserve attention and resource.

2.2 Organic data

Both orthophoto and satellite images are so-called organic data that require appropriate transformations before they become ready-to-use features (or statistical data). To choose between different algorithms or pipelines that generally lead to different results, some form

of supervised learning is necessary. Take, for instance, Figure 2.7 in Mugnoli *et al.* (2024), reproduced here in Figure 2.1, where two different green-area measurements result from the same image, it is obviously necessary to establish the relative merits of the two measurements, in order for the adopted statistics to be treated as trustworthy and not merely some 'evocative impressions'.

Figure 2.1 - Ortho-image of Ravenna (left), identification of Green Areas shown in Black for K- Medians (centre), and Advanced pipeline (right). Advanced pipeline selects stronger green areas, see red square detail

Source: Mugnoli *et al.* (2024)

For instance, in familiar notations, let the target of interest be

$$Y = \sum_{i \in U} y_i$$

over all the pixels $U = \{1, ..., N\}$, where $y_i = 1$ if green area or 0 otherwise. Each image like the leftmost one in Figure 2.1, contains a cluster of pixels, given the partition of the total area of interest by a set of images. Let s be a probability sample of images from this set, yielding a cluster sample of pixels from U. To illustrate, let the leftmost image in Figure 2.1 belong to s, let y_i be the classification of the pixels by algorithm A (middle of Figure 2.1), and let z_i be those by algorithm B (rightmost of Figure 2.1).

Now, the necessary task for supervised learning based on s is to quantify the relative merits of $\{y_i\}$ vs. $\{z_i\}$. This requires additional resource and attention, which is however unavoidable if one wishes to make heads or tails of the value of algorithms A and B. In the one extreme, one may visit the field (depicted by the image) and carry out the time-consuming ground measurement of green area; in the other extreme, one may simply ask experts to decide whether the middle or rightmost image is a better transformation of the original leftmost image, such that each sample image κ yields $\delta_\kappa = 1$ if algorithm A is better or $\delta_\kappa = 0$ if algorithm B is better.

By the theory of Sanguiao-Sande and Zhang (2021), one can *e.g.* obtain an unbiased estimator of $\sum_\kappa \delta_\kappa$ from s, which only depends on the known sampling design of the images, regardless the assumptions or models underlying y_i or z_i (*i.e.* algorithm A or B). This would allow one to choose between the two algorithms, in terms of their performances for the given finite population U, even if the choice may vary for another finite population U'.

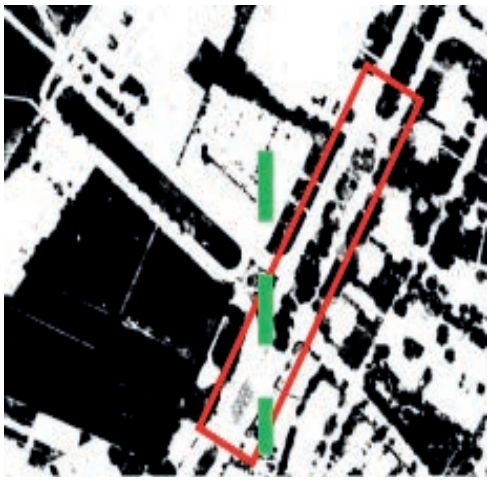
Moreover, by the theory of Zhang *et al.* (2024), one can make design-unbiased inference of y_i by algorithm A, either in terms of the total Y or the pixel (or image) level classification performance; similarly for z_i by algorithm B. The inference is valid over repeated sampling (of images) and evaluation (of given algorithm), irrespective of the assumptions or models underlying the algorithm.

When working with organic or other new forms of data, one must not only be occupied with the actual data transformations but ignore the potential errors. Bringing valid statistical inference to Machine Learning is necessary in order to satisfy the high quality demand of Official Statistics, where there are often no other tangible cost or loss apart from how close one can get to the descriptive truth (as it is defined).

2.3 Explainability

As it is stated in the paper, "Ensuring the transparency, interpretability, and ethical use of Machine Learning models in Official Statistics is a pressing concern". As possible means, one may consider using "white-box models" (such as linear regression, classification tree or support vector machine) to explain the local behaviour of "black-box models" (such as random forest, boosting trees or neural networks).

Figure 2.2 - LIME by support vector classifier (green, dashed) of algorithm B classification



Source: Own elaboration on Figure 2.7 in Mugnoli *et al.* (2024)

Heuristically, let us suppose that Figure 2.2 illustrates a local interpretable model-agnostic explanation (LIME) for algorithm B (rightmost in Figure 2.1 discussed above), showing how the algorithm can be approximated by a support vector classifier at the given point. Whether or not the reader may be satisfied with such an explanation, a question that remains unanswered is to how well algorithm B serves the purpose of classifying green area, and whether a statement on this property can be made validly. To put in another way, instead of explaining 'how an algorithm works', it may seem more relevant to explain 'how the given algorithm is being used and how we assess its performance'. As Box says,"All models are wrong, some are useful." That is, it seems more fruitful to focus on how a given model (with unavoidable shortcomings) is applied in a given context, rather than examining endlessly every aspect of the model (which will become more and more disappointing as the examination deepens).

As discussed above, one can obtain, say, an estimated MSE of the green area total in Rome calculated by algorithm B, and the MSE-estimator is *unbiased over repeated sampling* of s and y_i used for developing algorithm B. Isn't this an explanation more relevant to the users of the statistics generated by algorithm B, compared to LIME in Figure 2.2?

References

De Boom, C., and M. Reusens. 2023. "Changing Data Sources in the Age of Machine Learning for Official Statistics". *Paper presented at UNECE Machine Learning for Official Statistics Workshop 2023*. Geneva, Switzerland, 5-7 June 2023. https://unece.org/statistics/documents/2023/05/ml2023s2belgiumdeboompaperpdf.

Boeschoten, L., J. Ausloos, J. Möller, T. Araujo, and D. Oberski. 2022. "A framework for privacy preserving digital trace data collection through data donation". *Computational Communication Research*, Volume 4, N. 2: 388-423.

Mugnoli, S., A. Sabbi, F. De Fausti, G. Lancioni, and F. Sisti. 2024. "Quantification of urban green areas: An innovative remote sensing approach for official statistics". *2nd Workshop on methodologies for official statistics- Proceedings*, Session 2. Roma, Italy: Istat.

Sanguiao-Sande, L., and L.-C. Zhang., 2021. "Design-Unbiased Statistical Learning in Survey Sampling". *Sankhya A*, Volume 83: 714-744.

Zhang L.-C., and G. Haraldsen. 2022. "Secure Big Data Collection and Processing: Framework, Means and Opportunities". *Journal of the Royal Statistical Society Series A: Statistics in Society*, Volume 185, N. 4: 1541-1559. https://doi.org/10.1111/rssa.12836.

Zhang, L.-C., L. Sanguiao-Sande, and D. Lee. 2024. "Design-based predictive inference". *Journal of Official Statistics*. https://doi.org/10.1177/0282423X241277719.

Smart Surveys: Methodological issues and challenges for Official Statistics

Claudia De Vitiis, Fabrizio De Fausti, Francesca Inglese, Monica Perez[1]

Abstract

In smart surveys respondents employ mobile devices to acquire information through active and passive data collection and can share existing data collected by trusted third parties. Smart data can supplement or replace self-reports, improve the quality of social surveys and reduce respondent burden and non-response. Smart surveys lead, inevitably, to new sources of representation errors and measurement errors. The European Statistical System (ESS) have been financed two projects on this topic: ESSNet Smart Surveys 2020-2022, delivered preparatory work to create a European wide methodological and architectural framework; ESSNet Smart Surveys Implementation 2023, to test micro services, solution/ component for Time Use and Household Budget surveys. The focus of the paper is describing the main methodological and data collection issues addressed by the projects.

Keywords: Smart survey, smart device, sensor data.

1. Introduction

The Smart Surveys are surveys in which respondents are asked to employ smart devices (*e.g.* smartphones, tablets, activity trackers) to collect survey data through active and passive data collection of questionnaire and/or sensor data. The concept of smart surveys goes well beyond the mere use of web-based (online) data collection that essentially transform the paper questionnaire into an electronic version. Smart surveys involve dynamic and continuous interaction with the respondent and with her personal device(s). They combine data collection modes based on input from the data subjects (active data) with data collected passively by the device sensors (*e.g.* accelerometer, GPS, microphone, camera, etc.) (Struminskaya *et al.* 2020).

In Trusted Smart Surveys (TSS) the respondents are also asked to share existing data collected by trusted third parties, like government authorities and larger, stable enterprises willing to establish data delivery agreements. Constituent elements of a trusted smart survey are the strong protection of personal data based on privacy-preserving computation solutions, the full transparency and auditability of processing algorithms.

Smart surveys offer new opportunities for developing social surveys, especially those based on burdensome compilation of diaries (Household Budget Survey, henceforth HBS, and Time Use Survey, TUS), as they aim to collect new data sources through devices (smartphones, tablets, wearables) that use sensors to provide information about themselves or their surroundings. The measurement capabilities of mobile devices can supplement or potentially even replace self-reports in surveys: sensor data collected passively (*e.g.* location, motion, activity trackers) and respondents' activities on smartphones (*e.g.* taking pictures, scanning receipts) increase available data sources.

1 Claudia De Vitiis (devitiis@istat.it), Fabrizio De Fausti (defausti@istat.it), Francesca Inglese (fringles@istat.it), Monica Perez (perez@istat.it), Italian National Institute of Statistics - Istat. The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics - Istat.

Smart surveys are smart because they demand a change in how surveys are designed and executed. First, the design phase is either more experimental or relies on already developed methods or analytical algorithms. The design is different, because smart surveys also use proxies for information: a sensor reading and its classification is not the same as an answer to a direct question or a cluster of questions. It is also expected to be more complex, time and cost-intensive, due to utilising data sources that were not intended to serve Official Statistics and could quickly change over time (Ricciato *et al.* 2019*a*, p. 593; Ricciato *et al.* 2019*b*; Ricciato *et al.* 2020).

These innovative ways of data collection offer new challenges to improve significantly the quality of the statistics produced by social surveys in the National Statistical Institutes (NSIs), while aiming at reducing the burden for the respondents. However, the acquisition of information through new forms of data and a different involvement of participants constitute aspects of smart surveys that have consequences in terms of both representation (selection) and measurement errors. Participant selectivity, (un)willingness to provide sensor data or perform additional tasks, privacy concerns and ethical issues, data quality and usefulness, etc., are important challenges faced in smart survey projects (Keusch *et al.* 2019).

The introduction of smart features in the survey can have a strong impact on the Generic Statistical Business Process Model (GSBPM) especially in the design, data collection and data processing phases. The design phase of GSBPM needs to be redesigned considering new data collection strategies for reducing representation error and for controlling measurement error (metadata, paradata, contextual data), new methods in the processing phase (Machine Learning algorithms used for prediction, in data cleaning and imputation steps).

Data collection strategies concern the use of contact and reminder strategies, recruitment materials and incentive approaches. The type of interviewer assistance needs to be defined: interviewers can be involved in recruiting interviewees and keeping them motivated; interviewers can be involved from the start or only after a non-response. The recruiting material should be prepared taking into account the activities in which the interviewee is involved. As app data collection requires downloading and registering an unknown app, the recruiting material can include instructions, an overview of basic screens, a landing page, a brief tutorial on how to navigate and possibly a brochure explaining what data is collected and for what purpose (to ensure data confidentiality).

Incentives of different nature (monetary, gamification, feedback) can be used to increase participation or avoid dropping out, but also to counteract the privacy intrusiveness of passive (sensor) data collection. The choice of incentives depends on the burden of the survey for the respondents and on the privacy intrusiveness of sensors. The satisfaction of one type of incentive over another depends on the characteristics and the smartphone skills/habits of the respondents involved in the survey. Providing in-app feedback to respondents that can be instantaneous or postponed to the end of the data collection might motivate more people to participate in the survey, and participants might be more motivated to provide accurate data, so that the feedback is more useful to them. However, personalised feedback might lead participants to change the behaviour that is being measured with the app, is costly to implement, and also constrains other design decisions for the data collection (ESSNet Smart Survey 2020-2022).

Despite the implementation of the best strategies to maximise participation, it may still be necessary to adopt mixed mode data collection, smart and traditional techniques. In these cases, experimental studies for the evaluation of the mode effect must be carried out.

New methods for processing data, not yet explored in the traditional surveys, are relevant goals of the smart surveys. Machine Learning algorithms play an important role in smart data acquisition and in sensor data processing, but an important issue in the context of smart surveys is how to use ML and how to achieve levels of accuracy of results consistent with the quality standards required for Official Statistics. ML can be used for structuring unstructured data or to classify objects acquired from the images, or physical activities using accelerometer data, or leisure activities using GPS data matched with street maps. Generally, the model accuracy for classification prediction can be improved using in the loop experts, but often, above all, the feedback of respondent (Benedikt *et al.* 2020).

The methodological and data collection issues addressed in the paper can be traced back to the objectives and activities in which the Italian National Institute of Statistics was and is involved in two ESSNET projects on smart surveys (ESSNet on Smart Surveys 2020-2022).

This article is structured as follows: in Section 2 a brief overview of the ESSNet projects on smart surveys and their main objectives is given; in Section 3 the main methodological aspects are presented and discussed; finally, in Section 4, two pilots testing data collection issues for smart surveys in Italy are outlined.

2. Smart survey ESSNet Projects

The ESSNet on Smart Surveys, which carried out its activities between 2020 and 2022, constitutes a contribution towards many important achievements foreseen within the European Statistical System (ESS): (i) testing and developing (trusted) smart surveys within the ESSNet, based on the use of innovative data collection tools (ii) the conceptualisation, development and implementation of a new reference architecture for trusted smart statistics as well as the evolvement of new skills within the ESS. The ESSNet delivered preparatory work towards a European wide system to share und re-use smart survey solutions and components, agnostic to particular application domains, implementing a set of common functions and configurable services that can be used to build particular instances of trusted smart surveys for specific application domains and/or target areas.

In this context, the work-package 3 worked on the latter goal through two main tasks: (i) conceptualisation and development of a general platform for trusted smart surveys, following a top-down design approach; (ii) development of proofs-of-concept (PoCs) in the form of modular prototype elements for essential aspects of the architecture.

The design of the framework aimed at highlighting the relationships between many different aspects, related to three main dimensions: i) Architectural: concerning the design and development of software solutions for smart data collection and processing; ii) Methodological: regarding the available privacy preserving techniques, and the different methods for smart data gathering, validation and processing (*e.g.* data collection strategy, design of the user interface, edit checks and data quality, Machine Learning models); iii) Technological: dealing with the design of the technical infrastructure according to the privacy preserving requirements, as well as the interaction between the platform components.

The work-package 2 (WP2), aiming at the first goal, tested existing solutions (tools for smart surveys, applications and survey settings) through four pilots in four different survey topics: Time Use, Household Budget, Health and Living condition. The results of the pilot tests were presented and discussed in terms of shareability, modularity and country specific issues of the considered solutions (ESSNet on Smart Surveys 2020-2022).

The project currently underway, the Smart Survey Implementation project (SSI 2023-2025), has the goal to implement and demonstrate the concept of Trusted Smart Surveys, realising a proof of concept for a complete, end-to-end, data collection process and demonstrating solution combining: 1. Involvement and engagement of citizens as active contributors; 2. Acquiring, processing and combining data collected from smart devices and other appliances; 3. Contributing to the trustworthiness by guarantying strong privacy safeguards. The SSI project is carrying out i) Implementation of use cases and/pilot surveys with smart methodologies ii) Smart surveys in a mixed-mode data collection environment iii) Experimentation in Household Budget Survey (HBS) and Time Use (HETUS) surveys.

In this context, the methodological work-package, WP2, aims to define general methodological elements trusted smart surveys should have so that they can be used in statistical production by European NSIs. The four sub-tasks focus on either an 'opportunity' or 'threat' that was identified in the framework produced in the ESSNet on Smart Surveys 2020-2022. The four sub-tasks are: i) the successful recruitment of participants for smart surveys; ii) using Machine Learning to improve human-computer interaction in smart surveys; iii) respondent involvement and human-computer interaction in smart surveys; iv) integrating smart surveys with traditional survey methods by estimating the mode effect.

In particular, the task 2.2 on Machine Learning, is developing methodological standards around the use of Machine Learning models. The main issues are under what circumstances results from ML models can be used directly as statistical data, and under what circumstances data should be fed back to respondents. What to do when the quality of the Machine Learning outcome is too low? When should respondents be asked to provide new input (a picture or open-text) because no meaningful information could be extracted? Under all these issues, a fundamental role is played by the training datasets used in the ML: how and when should these be updated/improved? Case studies are the ML methods used in HBS and TUS.

The current project foresees also several large field tests (pilots in some of the participating countries, see Section 4) implementing different recruitment strategies and different modes of data collection for HBS and TUS. Moreover, in some countries a "perception survey" is carried out to investigate the perception and the opinion of the general population about the introduction of smart surveys in the official statistical production.

3. Methodological aspects of a Trusted Smart Surveys

In social surveys, a connection of new data sources – sensor and app data – with self-reports represents an added value of smart surveys compared to traditional surveys or digital data. In fact, the integration of different data sources can mitigate the surveys and digital data weaknesses. Smart surveys form a bridge between primary (survey) data collection and secondary (big) data collection.

However, there are multiple challenges to collecting sensor and app data: participant selectivity, (non) willingness to provide sensor data or perform additional tasks, privacy concerns and ethical issues, quality and usefulness of the data, etc. These aspects have consequences in terms of both representation (selection) and measurement errors. The application of the total survey error framework (Biemer *et al.* 2017) can provide a useful tool to guide methodological and practical decisions in sensor-app-based data collection. However, it needs to be redefined taking into account the hybrid forms of data collection and the device features for collecting, linking or processing data (device intelligence, internal and external sensors, data donation).

Inevitably, smart surveys lead to new sources of representation and measurement errors, that need the development of new strategies aiming to prevent and control possible sources of error and new methodologies in the assessment and correction of different types of error associated to sensor data (noise, outliers, missing, etc.). The quality of the data becomes an important concern of smart surveys which requires, on the one hand, a careful look at the definition of a framework for sensor data, the identification of paradata and contextual data useful in monitoring data collection processes, but also to improve the prediction achieved through ML algorithms. Prediction can be improved using appropriate ML methods that are capable of querying the respondent about data quality or to acquire missing data (label) to improve its performance when accuracy degrades.

3.1 Representation and measurement errors

Representation errors are determined by the availability or not of a smartphone or other mobile devices by the individuals selected in the sample (coverage error), or by their willingness to participate (non-response). Participation is influenced by technological barriers, topic of the survey, duration of data collection, respondent characteristics including privacy and security concerns, and respondent ability with smartphone and its tasks (Keusch *et al.* 2019). Consent to participate is required for legal and ethical reasons, but willingness to consent varies per type of sensor and depends on the context and purpose of the measurements. Increased intrusiveness of a sensor measurement can seriously affect the response.

Non-response can occur at many stages, not only from the consent to participate, to download and install an app or device, but also to use the app (whether actively or passively), to capture and transmit data, often repeatedly over a period. The question of non-response becomes more complex as the additional tasks that can be performed increase. Activities can vary in the degree of involvement of the participants, the level of burden, the sensitivity of the data collected, the technical requirements (*e.g.* battery usage or data transmission volume).

Downloading an app for data collection might involve further potential self-selection effects, as it requires additional steps from participants. Mechanisms of respondents' willingness to share sensor data depend on control over data collection, smartphone ability and privacy concerns. Willingness to share may be greater for activities where participants have control over what data are collected and when.

The growing rate of smartphone usage does not solve the coverage problem, if those who use smartphones are different from those who do not in the characteristics of interest, and if the respondents who are willing to engage in specific tasks (*e.g.* install apps, share sensor data) differ from non-willing smartphone users.

Furthermore, there are differences between those who use smartphones, due to the existence of different operating systems. iPhone owners differ significantly from other smartphone owners in their attitudinal and behavioural characteristics, and these differences cannot be corrected by weighting based on socio-demographic information (Bähr *et al.* 2020).

Sensor data introduce significant changes to measurement, starting from the definition of the concepts themselves. In fact, measurement errors can be caused by incorrect starting concepts, and/or by the inadequate operational definition of the variables. The conversions of the theoretical concepts do not adequately measure the concept that was to be analysed, or only partially measure it.

Measurement errors in sensor data can occur during the collect phase and in the processing phase. Within the data collection process, measurement errors are generated by different sources, by the respondents' behaviour or by the sensors themselves. Operational errors are determined by the respondents who may incorrectly initialise the measurements or use the devices wrongly. Sensor measurement from smartphone differs by operating system. While iPhones and Android devices usually have the same or very similar embedded sensors, the way these sensors interact with the operating system (OS) (*e.g.* how often measurements are taken with a sensor), and whether and how external apps are allowed to interact with the sensors, differs by OS. In practice, it is very difficult to develop research apps that work exactly the same across all brands of devices. Similarly, it is difficult to standardise in-browser sensor measurement. Different sensor-equipped devices can produce different results, raising the issues of comparability. The speed of innovation in sensor measurement poses further threats to comparability of measurement over time.

The quality of sensor measurement can be affected by sensor inaccuracy (imprecision, time inequivalence, device inequivalence). Depending on sensor quality and age, sensors may produce systematic and random measurement errors. Systematic errors occur when the sensor measurements deviate from known absolute levels over time (drift). Periodic recalibration is needed to avoid time-dependent systematic errors, but incorrect calibration can produce systematic errors themselves. Instead, random deviations of sensor measurements over time produce noise.

Reporting or data acquisition errors are measurement errors caused by both technologies and humans. During the processing phase, specification errors may be introduced when sensor data are manipulated, to search for patterns or to explore the accuracy and precision of data, as well as when different sensors are combined. The processing of sensor data is made complicated by the volume of data and the need to adopt processing strategies (such as aggregation or sampling) before their use. In addition, the evaluation and adjustment phase of measurement errors - outliers, noise, missing data - can be time-consuming, especially in the search for appropriate methodological solutions, as in the case of the treatment of missing data.

Sensor data can be missing for short periods of time, due to communication loss or technical issues but, also, for longer periods. The entity of missing data may vary due to smartphone batteries running empty, or a particular sensor, an app, or the device itself when it is turned off by the participant. Measurement challenges can exasperate the missing data problem, and the collected data will not reflect the true behaviour of an individual. This is the case in which participants install the apps but fail to carry a smartphone everywhere. The strategies of dealing with missing items are very complex because data vary across sensors, depending on the extent and nature of the missingness patterns, and the phenomena under study (Bähr *et al.* 2020).

3.2 Smart data quality

3.2.1 Paradata and contextual data

An important concern of data collection in smart surveys is data quality. While it is true that sensor data acquired passively can lead less measurement errors than self-reports, it is also true that these data are not free from biases. The heterogeneity in sensor quality across smartphone types and the variations in availability of data affect the measurements. Additionally, in case of

participatory sensing, the biases that are generated for traditional surveying have to be taken into account. During the data collection phase it is very important to implement quality checks. Soft and hard checks of plausibility of entered data and notifications of missing data implemented in an interactive and dynamic model that offers insight into the process operation and improved monitoring are needed.

The acquisition of paradata in smart surveys must be designed considering the methods adopted for data collection (active or passive), the functionality of the app developed, the type of device used (smartphone, wearable) and other features performed. The choice of indicators to assess the overall smart survey performance is a complex process, requiring more empirical evidence about the relevance of the information that can be acquired from a device.

Paradata can mitigate survey errors as they are useful: to detect no activity signal or to get information on each contact over time; for tracking, through logs, information on how certain functionalities of the app were used (*e.g.* how often did the respondents open the insights page); to detect insight in technical difficulties in using the survey app related to the device. The implementation of log files through which an app records and stores events is a complex task, because all the problems that may arise during the collection phase should be taken into account in advance. Information concerning which browser is being used, what version, and on which operating system can be acquired through a browser's user agent string (UA).

Furthermore, paradata can be used to have control over what is measured in the app, to perform comparison of expected results and observation over time (diversity in reports, verification of rule-based category, etc.).

To assess data quality for a smart survey, contextual data on the app usage and on performance of sensors are needed. Users' behaviour with apps may vary from user to user, according to their contextual information in different dimensions such as temporal context, work status in workday or holiday, spatial context, their emotional state, Wi-Fi status, or device related status etc. App usage pattern can be collected from built-in sensors and application programming interfaces. By processing sensor data (consistency validation, metadata enrichment), context information are generated for extracting behaviour patterns or a subject's activity.

The contextual data should assist all likely types of representation and measurement errors that one would like to analyse and/or adjust. For representation and sensor data, it is useful to know if the respondent has access to the sensor, has the ability to use the sensor, if the sensor produce missing data, or if there were problems in data transmission. Here, context information is intended to capture the respondent's behaviour/ability (ability to operate the sensor, to handle the sensor according to instructions), the performance of sensor itself (reliability, deterioration, anomalies) and the problems occurred during reading the sensor data. Advanced analytic techniques to discover information, hidden patterns, and unknown correlations among the contexts are necessary.

In defining a general data collection framework for a smart survey, several dimensions must be taken into account that can affect participants' concerns and data quality (*e.g.* criteria for sensor selection related to research objectives and logistics, to the evaluation of sensor characteristics, to participant engagement, to human participant protection). Minimise the risk and burden on participants while maximising the quantity and quality of data is of primary importance. The set of the sensors used can play an important role in the outcome of a survey, as data quality is intrinsically constrained by the characteristics of the sensors and the interactions of the participants with those sensors.

Data quality needs to be analysed considering the type of sensor and analytic goals involved, but also the specific features of a smart survey. Indeed, a smart survey can employ device intelligence and internal sensors as well as others smart features, such as access to external sensors (*e.g.* activity trackers) and personal and public online data, linkage consent. In the smart survey design, many aspects must be considered, such as: the trade-off between passive and active data to obtain, for example, a high and balanced response and data quality; the right boundary between respondent burden, respondent engagement and data quality; not least the integration of data from different sources and with different quality levels.

3.2.2 Data quality framework

The presence/combination in a smart survey of traditional data (provided directly by the respondent using a questionnaire) with big/sensor data (provided in active or passive way) deriving from different sources (internal sensors, external sensors, public online data, personal online data) forces us to look at quality as a requirement that needs the definition of new concepts and metrics and the development of new approaches for the validation analysis.

Ideally, a general data quality (DQ) framework should be declined right from the data source, the type of data and sensor, but also considering new sources of representation (coverage, participant selectivity, non-willingness to provide sensor data) and measurement (sensor) errors. For sensor data, characteristics and properties of sensors, and the quality of measurements must be considered in defining a DQ framework. Sensor measurements can be affected by limitations of the sensor itself (inaccuracy, time inequivalence), the heterogeneity of devices (inequivalence), the behaviour of the participants in the survey, etc. Different aspects should be considered if sensor data derive from third parties, but in this case the flow of the data acquisition would be different and also the quality would be carried out in a different perspective.

Quality for sensor data, can be represented with internal and objective metrics (intrinsic characteristics of sensor data) and with context-based metrics. By following this approach, it is possible to identify two types of data quality estimation: (i) DQ assessment, which estimates the quality of the raw data; (ii) DQ evaluation, which estimates the quality of processed data considering context-based metrics.

Data quality assessment implies many dimensions: believability (comparison with the correct operating bounds), completeness (missing values), free-of-error (erroneous values), consistency (over time), timeliness (delay), accuracy (deviation from true value) and precision (granularity of readings). For assessing and evaluating data quality, it is particularly important to acquire information during the data collection phase, referred to elementary units. Paradata can offer information on several statistical parameters of the measured smartphone sensors and insights into their performance, while contextual data can be useful to characterise users' day-to-day situations that have an influence on their smartphone and app usage, and consequently on data quality. Users' behaviour with smartphone and apps may vary from user to user, according to their contextual information, such as temporal context, work status in workday or holiday, spatial context, emotional state, Wi-Fi status, or device related status etc.

Data quality evaluation involves some steps: the selection of appropriate metrics; the use of methods for the integration of data quality metrics (*e.g.* data correctness: consistency, completeness, sensor accuracy) and then for integrating security and privacy metrics (*e.g.* Machine Learning techniques) as data security may influence elements of data correctness (Immonen *et al.* 2015); the development of evaluation methods that include many data quality components integrated into the unified overall data quality score.

3.3 Smart surveys and Machine Learning

Machine Learning (ML) algorithms play an important role in smart surveys but how ML is to be used in this context is the key question. The level to which automation can replace the direct acquisition of information or replace manual processes without degrading data quality and/or increasing respondent burden, is the crucial point in the use of ML.

In smart surveys, ML is generally used for structuring unstructured data or algorithms aimed at classification: in HBS OCR is used for reading images of receipts and receipts and classification algorithms are necessary to trace the products declared or acquired from the images to the COICOP classification. In TUS ML algorithms can be used to support the respondent who must fill in the activity diary, providing suggestions based on the prediction of activities based on locations (data from GPS matched with contextual information from map services).

In sensor data applications, models seldom reach 100% accuracy. Certain population subgroups or certain survey statistics may require manual inspection. In most ML classification problems, it takes little effort to achieve close to 80% accuracy, but it is increasingly difficult to push for the last 20%. This is a significant challenge for Official Statistics that require high precision and accuracy. Acceptable error rates are usually agreed between survey teams and their end users, typically less than 5% (Benedikt *et al.* 2020).

In such cases of improvement of the ML models accuracy, human interventions must be envisaged in order to assign correct labels. The new labelled item is used to retrain the model to make it more up-to-date. Over time, the machine learns from humans and becomes more and more accurate.

Furthermore, ML methods require continuous updating. Updating can be done fully automated through online learning or semi-automated through active learning. Retraining is ideally done based on incoming datasets while preserving the privacy of the respondents. In practice, when respondents provide data for which processing performance falls below specified thresholds, then this data should be used for retraining ML model.

Active learning is the subset of ML in which a learning algorithm can query a user interactively to label data to obtain the desired outputs. In active learning, the algorithm selects the subset of examples to be labelled from a set of unlabelled data. These algorithms represent a key component in Human-in-the-Loop where human and machine intelligence combine to create more accurate models (Benedikt *et al.* 2020).

For the HBS survey, a fundamental task to improve products classification, concerns the measures that must be adopted to ensure that the level of accuracy of the ML algorithms over time remains constant and at pre-established levels. Such actions become necessary as the cases of unlabelled products increase. In this survey, active learning ("sequential design" in statistics) may be the most appropriate ML algorithm, since the situations where the classification procedure fails have to be managed during the data collection phase. The involvement of the respondent is necessary to collect labels that train a more accurate model. In the interactive learning procedure, it is necessary to develop the decision-making process for automatically determining when to queries or to stop. In a survey context, two aspects must be taken into account that may be in conflict with each other, the burden on respondents and the high level of classification accuracy. Therefore, a stopping criteria must find the right trade-off between annotation and ML performance.

For TUS and the geotracking domain (GPS - location data), the main problems to be faced for the implementation of ML algorithms concern processing of measurement errors (outliers, noise, missing data) in location data and the optimal choice of the features and contextual data (OpenStreetMap) that are functional for prediction of the daily activities or transport modes. All these choices have a significant impact on the quality/accuracy of the predictions.

4. Testing data collection issues for Trusted Smart surveys

For Official Statistics, there are numerous challenges that smart surveys pose. Smart surveys have methodological implications on data quality, for instance "selection effect" and the "accuracy" of the collected data (lack of expert supervision and interaction) are crucial issues to address. Istat is beginning to address them by experimenting on the field some test and pilot surveys - mainly devoted to the Time Use survey - that address new conditions and in which the respondents are called upon to experiment with new tools to respond and provide the requested information.

Smart surveys need to evaluate and analyse the measurement error that derives mostly from the use of different instruments (apps, sensors, etc.) but not only. In fact, smart surveys can introduce innovative elements in different phases of the GSBPM, such as for the data collection phase (strategies for recruitment of respondents, preliminary actions for the survey promotion, data collection with or without technical assistance for the use of the apps, etc.) and data processing phase (classification and/or coding internal or external to the app, etc.). All this involves the assessment of measurement errors that can arise and affect estimates.

To understand which characteristics of smart surveys are attractive or dissuasive for respondents, it is advisable to verify directly involving the population and verifying any critical issues directly in the field. This approach must make it possible to identify and profile the population target groups that are most similar or, conversely, most reluctant to participate in smart surveys.

To this end, as part of the ESSnet SSI 2023-2025 project, Istat is going to conduct a so called "perception survey" (named in Italian "New methods of data collection for statistical surveys") aims at knowing how the population would react to a smart survey, how much people are attracted or reluctant to share with an NSI the information captured by the sensors and how capable they are in assuming a "smart" behaviour that could be asked to adopt if they are called upon to participate in a smart survey. Within the project, a similar survey is conducted in the Netherlands and Slovenia as well.

The survey considers four areas of data acquisition through smart devices: i) information on one's geographical position provided through GPS/geolocation activation; ii) information on food expenditure provided through images of receipts; iii) information on physical activity captured by pedometers; iv) data on the energy consumption of your home through images of energy meters (gas, electricity, water).

The survey sample, whose size is 4,000 resident citizens, is a two- stage stratified sample. The sample is representative of the resident population in Italy (first stage units are municipalities; second stage units are resident citizens).

The survey uses two questionnaires and two data collection methods: a first self-reported paper questionnaire (PAPI) distributed through a network of interviewers; a second self-reported online questionnaire (CAWI) to be completed by respondent only after the Papi questionnaire

has been completed. The two methods are not one the alternative of the other but they are both required to take part the survey. The two questionnaires, in fact, must be understood as two sections of a single questionnaire, administered with different methods. The first questionnaire, the paper one, embeds login credentials to go into the online "smart" questionnaire. Online questionnaire is web responsive and respondents are required to filled in the form by smartphone or tablet (PC or laptop are strongly not recommended) in order to test the abilities to reply by using smart device and doing "basic" operations by using sensors embedded into the device (geo-reference and photo-camera). The data collected consist of a set of questions and a set of smart data.

The choice of the PAPI method, which may seem to be in contrast with the innovative characteristics of a smart survey, is motivated by the need to have a questionnaire accessible to every target population. With the paper questionnaire distributed through interviewers we want to reach every target population and not only people who have advanced technological skills or who have no reticence in sharing their information via the web, but also with those who have opposing opinions.

The topics dealt with the paper questionnaire are:

- preferences to the information channels and reasons for being/not being recruited;
- attitude to use smart devices;
- experience in using apps and internet;
- opinions on advantages on using apps to conduct surveys;
- willingness to share data collected by apps with the NSI;
- reasons for unwillingness to share app data with the NSI (data security, privacy, etc.);
- importance of informed consent (the need to know in advance which data will be collected/shared, need to control collected/shared data).

In the online questionnaire, respondents must answer some questions for each of the four above-mentioned topics (travel, food and beverage expenditure, physical activity, energy meters). In each of these subsections, respondents are required also to carry out actions – for instance, take its own geolocation or take a picture and share it – by using the sensors available on the smart device. For each operation, sensors can be opt-in or opt-off.

Therefore, the survey collects both ordinary and special personal data. Ordinary personal data are attributes related to a series of attitudes and opinions, a series of mobile/smart device usage variables, shopping receipts characteristics, energy meter data, step counts. Special data are location measurements (one measurement per respondent). Issues related to data processing and their protection have been described and assessed through a DPIA (Data Protection Impact Assessment) which describes in detail the data collection process, tools and risk assessment.

The results of the survey will be used to identify, with greater accuracy within the national context, as well as comparatively with other countries, what may be the most suitable communication strategies towards respondents, as well as recruitment strategies of interviewees and information/support for the use of sensors to conduct smart surveys. Keeping in mind the quality of the data, the objective is to evaluate the response rates and the degree of "trust" towards these new methods of data collection overall and separately by profile of the interviewees.

In this way, it will be possible to evaluate whether the transition towards smart surveys of existing data collection processes requires mixed-mode data collection strategies and which population targets require them.

Another decisive aspect for the success of smart surveys concerns the decisions that must be made in defining the investigation design with respect to the ethical-legal component and the privacy-by-design and privacy-by-default[2] choices in the development of the apps used within the data collection process.

Therefore, the results of the perception survey should help guiding choices on all these issues.

A second step of the experimental research aimed at implementing smart surveys refers to a field test based on the TUS pilot survey carried out by an app for compiling the daily activity diary. It must be noted that the TUS survey, usually made in Italy every 5 years, still in its latest 2022-2023 edition (in progress at the time of writing this paper) is traditionally carried out with paper questionnaires. The option to make possible compilation of the diary by an app would be an important step forward in the field of reducing respondent statistical burden and make the survey process more efficient for the NSI.

We must recall that TUS is an important observation tool on how people organise and use the time and on the relationships between the daily schedules of the various family members[3]. In fact, the main peculiarity of this survey lies in the fact that by compiling a daily diary it is possible to know the way in which people divide the 24 hours (divided into 144 10-minute intervals) between the various daily activities, journeys, places frequented and the people with whom he/she spent them. That is, it is information that presents an extremely high level of detail, not comparable with that obtainable from traditional questionnaires with fixed questions.

Given this detailed information advantage, the filling burden on respondents is very high and is no longer easily sustainable nowadays, highlighting for the TUS survey reductions in response rates in filling out the diary. The option to make possible compilation of diaries by using an app (also having suggestions on locations visited during the day and activities carried out in each visited place) would be an important step forward in the field of reducing respondent statistical burden, increase response rate and make the survey process more efficient for the NSI.

In the experiment, the respondent's use of the support provided by a geolocation sensor may be evaluated. Whether the respondent authorises the geolocation, the sensor can provide suggestions about the movements and places where the person went during the day, making it easier for them to remember not only the places visited but also the activities carried out with details and less memory effort, enhancing the quality of the data collected as well as reducing statistical burden.

The pilot survey aims to evaluate the effect mode between the traditional paper diary and the use of an app for compiling the diary of the TUS survey activities. It is planned to be carried out in autumn 2024 by involving approximately 3,000 residents in Italy. Although the design is still in progress and some changes in the test design could occur, the test would aim to evaluate outcomes according to: i) respondents compile diary of daily activities by using suggestions provided by geolocation embedded in the app; ii) respondents filling out diary by app without any suggestion; iii) interviewers involvement with an active role in supporting respondents in the technical operations related to the download and use of the app, as well as providing information and reassurances on privacy and the risks of sharing information.

2 Privacy by default principle (art 25. GDPR) provides, in fact, that -by default- only personal data should be processed to the extent necessary and sufficient for the intended purposes and for the period strictly necessary for such purposes. It is therefore necessary to design the data processing system ensuring that the data collected is not excessive, so that each data subject receives a high level of protection even if he does not take action to limit data collection.

3 The survey is regulated by the Italian law n. 53 of 2000, art.16.

The effect mode would be assessed from another point of view, that is taking into account results from the Time use survey carried out in 2022-2023 by using traditional survey design based on paper diary and the outcomes of the experimental pilot test.

References

Bähr, S., G.-C. Haas, F. Keusch, F. Kreuter, and M. Trappmann. 2020. "Missing Data and Other Measurement Quality Issues in Mobile Geolocation Sensor Data". *Social Science Computer Review*, Volume 40, N. 1: 212-235. https://doi.org/10.1177/0894439320944118.

Benedikt, L., C. Joshi, L. Nolan, N. de Wolf, and B. Schouten. 2020. "Optical Character Recognition and Machine Learning Classification of Shopping Receipts". *ESSnet SEP-2105369. @HBS>An app-assisted approach for the Household Budget Survey.* https://ec.europa.eu/eurostat/documents/54431/11489222/6+Receipt+scan+analysis.pdf.

Biemer, P. P., E. de Leeuw, S. Eckman B. Edwards, T. Kreuter, L.E. Lyberg, N.C. Tucker, and B.T. West (*eds*). 2017. *Total Survey Error in Practice*. Hoboken, New Jersey: John Wiley & Sons.

ESSNet on Smart Surveys 2020-2022. Available at https://wayback.archive-it.org/12090/20231227175044/https://cros-legacy.ec.europa.eu/content/essnet-smart-surveys_en.

Immonen A., P. Pääkkönen, and E. Ovaska. 2015. "Evaluating the Quality of Social Media Data in Big Data Architecture". *IEEE Access*, Volume 3: 2028-2043. https://doi.org/10.1109/ACCESS.2015.2490723.

Keusch, F., B. Struminskaya, C. Antoun, M.P. Couper, and F. Kreuter. 2019. "Willingness to Participate in Passive Mobile Data Collection". *Public Opinion Quarterly*, Volume 83, N. S1: 210-235.

Ricciato, F., K. Giannakouris, A. Wirthmann, and M. Hahn. 2020. "Trusted Smart Surveys: a possible application of Privacy Enhancing Technologies in Official Statistics". In Pollice, A., N. Salvati, and F. Schirripa Spagnolo (*eds*). *Book of Short Papers SIS 2020. 50th Scientific Meeting of the Italian Statistical Society (SIS 2020)*: 53-58. Milano, Italy: Pearson Italia.

Ricciato, F., A. Wirthmann, K. Giannakouris, F. Reis, and M. Skaliotis. 2019*a*. "Trusted smart statistics: motivations and principles". *Statistical Journal of the IAOS*, Volume 35, N. 4: 589-603.

Ricciato, F., A. Bujnowska, A. Wirthmann, M. Hahn, and E. Barredo-Capelot. 2019*b*. "A reflection on privacy and data confidentiality in official statistics". *Conference paper ISI World Statistics Congress (ISI 2019)*. Kuala Lumpur, Malaysia, 18-23 August 2019. https://www.bis.org/ifc/events/isi_wsc_62/ips177_paper3.pdf.

Struminskaya, B., Lugtig, P., Keusch, F., Hohne, J.K. 2020. "Augmenting Surveys with Data from Sensors and Apps; Opportunities and Challenges". *Social Science Computer Review*. https://doi.org/10.1177/0894439320979951.

Quantification of urban green areas: An innovative remote sensing approach for Official Statistics

Stefano Mugnoli, Alberto Sabbi, Fabrizio De Fausti, Giuseppe Lancioni, Francesco Sisti[1]

Abstract

One of the most studied 'objects' in remote sensing is certainly vegetation. There are numerous spectral indices developed by specialists aiming to highlight certain aspects of the vegetation cover (i.e. water stress, biomass quantification, fire damages, etc.). In our analysis, starting with high-resolution remote sensed images (AGEA ortho-images with 20 and 50 cm pixel resolution), some of the most used vegetation indices are calculated to extract statistics related to the total vegetation cover in the major Italian urban centres.

Keywords: Remote sensing, vegetation indices.

1. Introduction

Remote sensing is a branch of applied sciences aimed at obtaining qualitative and quantitative information by investigating objects without direct contact. This is achieved through sensors installed on planes, satellites, and drones, which measure the electromagnetic wave radiation emitted or reflected by the objects under study.

For many decades, Earth observation through satellites has been a well-established procedure for monitoring our planet and conducting valuable surveys to study various environmental and territorial aspects. These aspects include vegetation condition, water pollution, hydrogeological instability, land cover, soil consumption, and more (Chiocchini *et al.* 2018).

The advantages of the remote sensed images are considerable, starting with acquiring territorial information very easily compared to other ways; furthermore, the possibility of having images continuously, allows the study of phenomena that would be impossible to investigate in other ways.

The detected parameters by sensors are electromagnetic ones, *i.e.* radiation emitted, phase, polarisation, amplitude of the electromagnetic field; all these parameters determine the so-called 'spectral signature' of all detected objects.

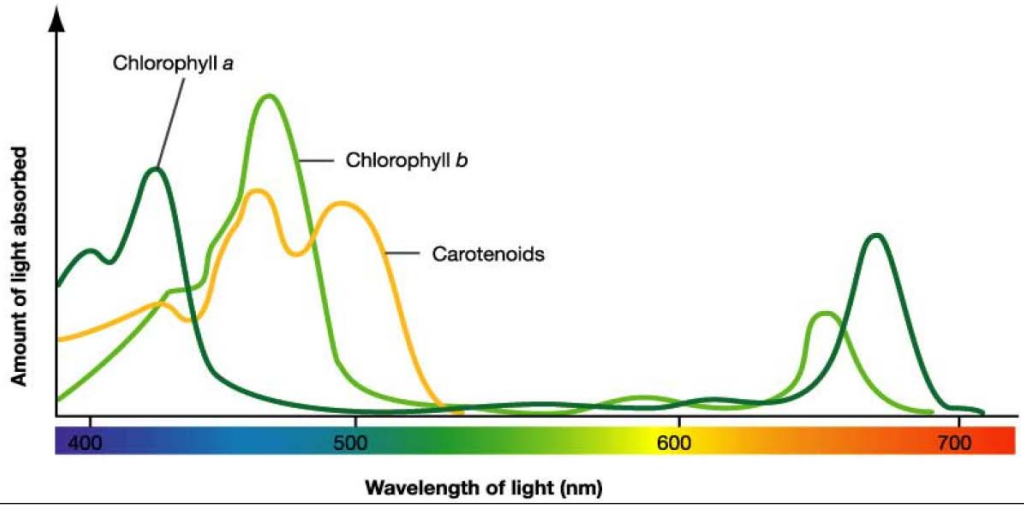
The spectral signature of an object is, in practice, its peculiar behaviour in respect of incident radiations to the different wavelengths; so knowing the spectral signature of an object we are able to uniquely identify it.

One of the most studied 'object' by remote sensing is certainly vegetation; the spectral indices developed by specialists who want to point out some aspects of the vegetation cover (*i.e.* water stress, biomass quantification, fire damages, etc.) are indeed numerous.

The whole thing is based on the chlorophyll behaviour (Figure 1.1) in relation to its ability to absorb light radiation at various wavelengths.

1 Stefano Mugnoli (mugnoli@istat.it), Alberto Sabbi, (sabbi@istat.it), Fabrizio De Fausti (defausti@istat.it), Giuseppe Lancioni (lancioni@istat.it), Francesco Sisti (francesco.sisti@istat.it), Italian National Institute of Statistics - Istat. The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics - Istat.

Figure 1.1 – Absorption spectrum of plant pigments



Source: Pigment Exctraction Lab, https://ressources.unisciel.fr/tp_virtuels/Pigment_Extraction_Lab/co/module_Virtual%20Experiment_1.html

In our analysis, starting with high-resolution remote sensed images (AGEA ortho-images with 20 and 50 cm pixel resolution) some of the most used vegetation indices (Xue and Boafeng 2017; Kriegler *et al.* 1969; Pristeri *et al.* 2021) are calculated in order to extract statistics linked to the total vegetation cover in the major Italian urban centres.

1.1 Importance of the quantification of the total vegetation cover inside urban areas

Measures of the approximated green cover in urban areas represent a very important data for the analysis and the development of many indicators linked to various aspects of the cities life.

For example, the 'quality of life', in some cases, is closely related to the possibility to enjoy of the public and private green areas (parks, gardens, historic villas, sports facilities, etc.).

In addition, the environmental quality is based on the presence and the health of the vegetation cover in each place.

Those represented above are some of the macroscopic aspects influenced by vegetation to take into account: there are many others more cryptic and maybe more difficult to investigate, but surely no less important (air and water quality, biodiversity, environmental impact assessment, soil consumption, etc.).

The proposed statistical analysis aims at providing the calculated data as a basic instrument for further investigation of dynamics that regulate the big cities.

1.2 Vegetation indices

The point of reference are the very high spatial resolution ortho-images released by AGEA (Agency for the Agricultural Supply) characterised by a 20 cm pixel on the ground (urban areas) and 50 cm for the extra-urban areas.

These images cover, over a three years period, the entire Italian territory and starting from 2012, they have been released to Istat at four spectral bands:

Red (R) λ 650 nm;
Green (G) λ 550 nm;
Blue (B) λ 450 nm;
Near Infrared (NIR) λ 800 nm.

By some simple python scripts, we are able to calculate the following indices:

NDVI (Normalised Difference Vegetation Index)[2]: It is surely the most famous and used vegetation index. Its conventional formula, that is based on the behaviour of the chlorophyll a and b, is:

NDVI=(NIR-RED)/(NIR+RED)

From all the above, NDVI uses the RED wavelength as absorption channel and the NIR as reflection channel.

ENDVI (Enhanced Normalised Difference Vegetation Index)[3]: to obtain better results it is possible to use the GREEN wavelength as reflectance channel; it is important to remember that a plant, in optimum condition, reflects both GREEN and NIR. So, the NDVI formula can be implemented as follows:

ENDVI=([(NIR+GREEN)-(2*BLUE)])/([(NIR+GREEN)+(2*BLUE)])

This formula sums up GREEN and NIR channels to calculate the reflectance. BLUE channel is multiplied by 2, just to compensate the sum NIR+GREEN.

GLI (Green Leaf Index)[4]: The Green Leaf Index was designed for use with digital cameras capturing only true-colour bands (RGB), scaled 0 to 255. As such, it is useful as a vegetation index that can be applied in the absence of NIR information.

GLI= [(Green - Red) + (Green - Blue)] / [(2 * Green) + Red + Blue]

From all the above, we obtain continuous images (float type) and so, these files must be re-classified to extract just the pixels related to areas covered by vegetation. To do this, it is necessary to carefully study the histogram of the image obtained, in order to evaluate the threshold value beyond which there is a high probability that the pixel is 'green'.

1.3 The 'Threshold problem'

By extracting a vegetation index from an image, in our case a mosaic representing the entire urban area, we obtain a continuous raster file from which it is possible to identify four land cover classes. To better illustrate this concept, the subsequent figures demonstrate how the pixels from the vegetation index image are grouped within the image histogram (Figure 1.3). The reference image, containing all four 'land cover' classes, is presented below (Figure 1.2).

It is clear that it is quite difficult to detect the exact value of the pixel that uniquely identifies a 'green pixel'. Therefore, we decided to classify the vegetation index image using the ISODATA classification algorithm[5] of the ERDAS Imagine© (version 2022) software, by setting to four the number of the classes.

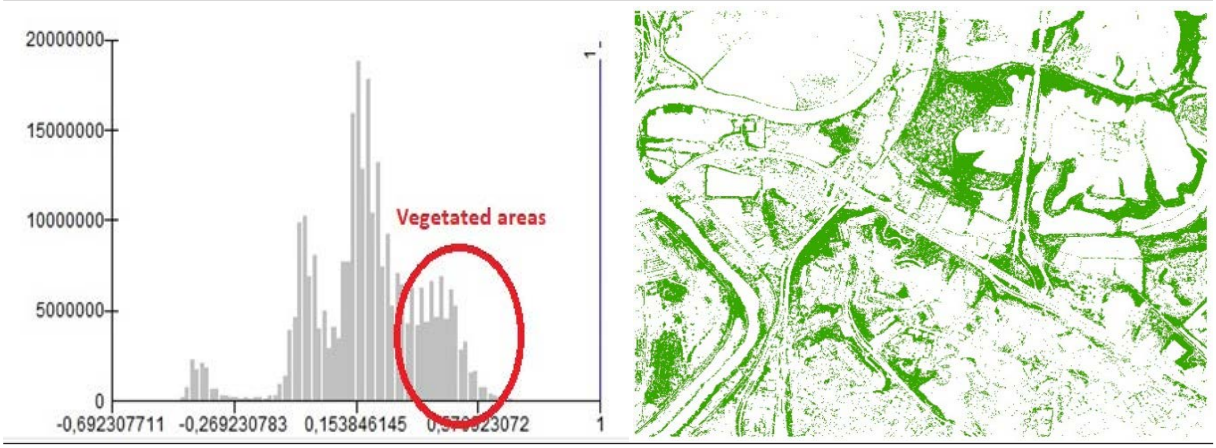
The last cluster is formed by green pixels.

2 For more information: https://it.wikipedia.org/wiki/Normalized_Difference_Vegetation_Index.
3 For further information https://maxmax.com/endvi.htm.
4 For more information: http://localhost:8080/imaginehelp/html/#/home/670421/10/11.
5 ERDAS IMAGINE uses the ISODATA algorithm to perform an unsupervised classification.
Click http://localhost:8080/imaginehelp/html/#/home/unsupervised_isoframe/10/11 to find out more

Figure 1.2 - Image R,G,B of a portion of a urban areas of Rome (Castel Giubileo, Villa Spada, Serpentara, Colle Salario)

Source: AGEA ortho-image

Figure 1.3 - Part of the histogram of the image shown in Figure 1.2, related to the 'green pixels'



Source: Author's processing

1.4 First results

In Table 1.1, a summary of the obtained results for the major Italian urban centres.

The number shown in Table 1.1 put in evidence the variability that exists within the Italian localities; it seems that the inhabitants of Padua and Reggio Calabria are much luckier than Torino and Napoli ones, because of a greater amount of green urban m^2 per capita.

Certainly, the problematic issues related to calculation of the green inside urban areas are much more complex and they depends of many different aspects.

However, it is also the case that the image processing techniques can help us to solve the problem significantly.

Table 1.1 - Obtained results for the major Italian urban centres

Urban area	Av. green area (Ha)	Av. m² per capita	Flight year	Green area (Ha)	Flight year	Green area (Ha)	Flight year	Green area (Ha)	Flight year	Green area (Ha)
Torino	2.119,8	24,3			2015	1.834,8	2018	2.404,7		
Milano	3.440,5	27,8	2012	3.315,1	2015	3.565,9				
Verona	1.055,9	48,1			2015	1.111,7	2018	1.000,1		
Padova	1.664,5	81,7			2015	1.658,8	2018	1.670,2		
Mestre	1.072,6	72,6	2012	1.045,8	2015	1.054,5	2018	1.117,4		
Trieste	726,6	38,8	2011	767,1	2014	716,1	2017	696,7		
Genova	1.780,1	30,6	2010	1.726,9			2016	1.833,2		
Bologna	2.843,5	77,6	2011	2.789,6	2014	2.892,2	2017	2.827,3	2020	2.864,7
Firenze	1.244,9	35,6					2016	1.386,7	2019	1.103,1
Roma	9.854,9	42,5			2014	10.012,8	2017	9.697,1		
Napoli	2.372,8	24,7	2011	2.473,2	2014	2.522,9	2017	2.276,2	2020	2.218,7
Bari	1.047,9	37,7	2010	1.053,6	2013	996,5	2016	1.148,0	2019	993,4
Reggio di Calabria	1.331,2	78,7	2012	1.258,3	2015	1.404,0				
Messina	655,9	29,8			2013	553,9	2016	697,1	2019	716,8
Catania	764,7	26,5			2013	790,6	2016	738,8		
Palermo	1.911,2	29,4	2010	1.859,0			2016	1.963,4		
Sassari	563,4	61,3							2019	563,4
Cagliari	473,8	32,7					2016	459,2	2019	488,5

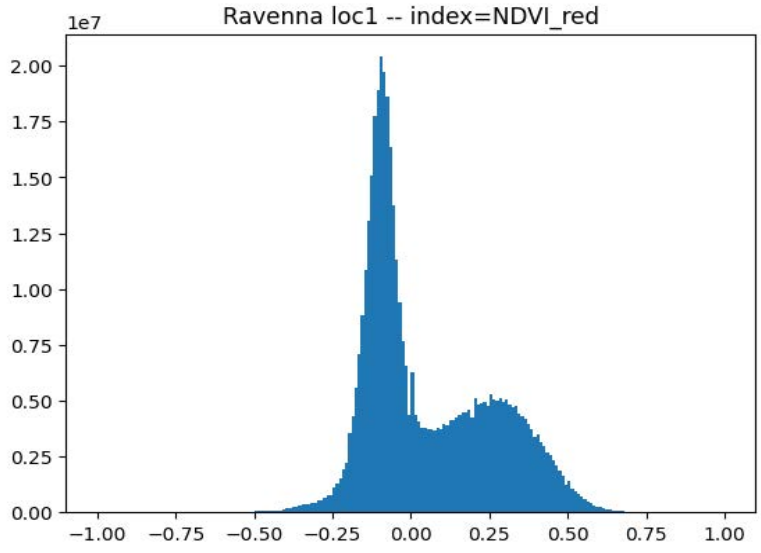
Source: Istat elaboration on Istat and AGEA data

2. *Ad hoc* Machine Learning algorithms

Although there are strong theoretical arguments for expecting four distinct and well-defined regions in the NDVI distribution, these are seldom observed in actual data.

Sometimes one of the classes overshadow the one next to it, and in some extreme cases we observe distributions with only two or even just one clear maxima (Figure 2.1).

Figure 2.1 - NDVI histogram for the city of Ravenna; two humps are clearly visible

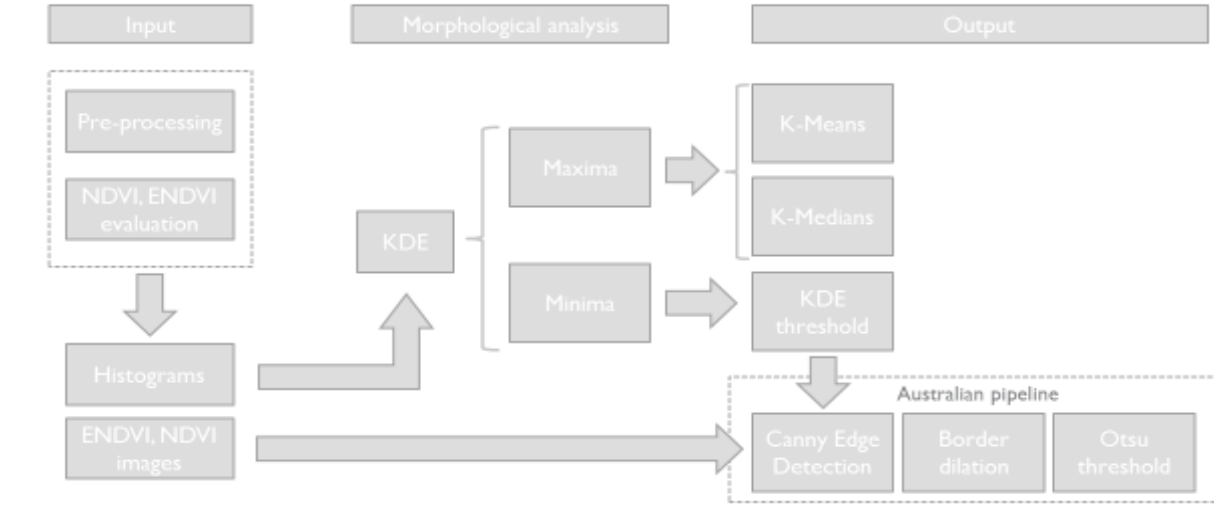


Source: Author's processing

To overcome this issue and get a more reliable result we therefore considered tailored Machine Learning approaches to identify an urban green threshold even for general cases of unknown numbers of clusters.

The various approaches represent a sequence of progressive refinement in the determination of the green threshold. The overall schema of our methodology is shown in Figure 2.2.

Figure 2.2 - General pipeline of the process



Source: Author's processing

2.1 Data preprocessing

In order to correctly and extensively evaluate the methodology described in this chapter, we use two different data sources: ortho-images, and satellite images. In both cases, sources images are related to the city of Ravenna. A third source is represented by the shape files of the same city and is used to crop the raster images to the correct borders.

Ortho-images. Very high spatial resolution images released by AGEA characterised by a 20 cm pixel on the ground (urban areas) and 50 cm for the extra-urban areas. Taken by 2020 flight; refer to Section 1.2 for further details.

Ortho-images need pre-processing to be used as input in the following methodology; in detail:

- original images are delivered in ecw file format; they are converted in GeoTiff;
- images are cropped as per the shape file, with the aim of correctly selecting the urban areas;
- NDVI is evaluated; note that the input are multi-band images, while the output are single-band ones;
- multiple images are combined to form a mosaic: a single image of the full city.

All elaboration steps are executed using python public libraries.

Satellite images. Unlike ortho-images, we have developed the advanced pipeline described in 2.4 on satellite images using the services made available on the Google Earth Engine (GEE) cloud platform. We used data acquired from the Sentinel-2 mission produced by Copernicus, the European Union's Earth observation program. Sentinel-2 is a wide-swath, high-resolution, multi-spectral imaging mission supporting and monitoring earth observation studies. We used the Sentinel-2 L2 (S2-L2) product from ESA and data available from GEE (https://docs.sentinel-hub.com/api/latest/data/sentinel-2-l2a/) containing 12 UINT16 spectral bands representing Surface Reflectance (SR) scaled by 10000. The data are available from 2017-03-28, to calculate NDVI vegetation index we used the bands B2, B3, B4, B8 with resolution of 10 meters.

To perform the analysis with satellite images, we carry out a pre-processing different from that of the ortho-images, which exploits the availability of numerous satellite passes during an observation period.

- Acquisition of a collection of S2-L2 images for a specific time period (from 2022-03-01 to 2022-07-01). 24 images of S2 have been acquired.
- For each one, an average is calculated. In order to eliminate outliers and mitigate artefacts that can create shadows and cloud cover, the average is performed on values that fall within an interval between the 15th and 40th percentile (https://ttgeospatial.com/2020/07/22/tracking-deforestation-of-the-amazon-with-google-earth-engine/)
- Bicubic resampling is performed. This makes the images smoother and allows for better tracking of the contours of vegetation patterns.
- Images are cropped as per the shape file, with the aim of correctly select the urban areas;
- NDVI is evaluated; note that the input are multi-band images, while the output are single-band ones;

2.2 Histogram smoothing

In the realm of satellite imagery analysis, particularly when discerning vegetated areas in urban landscapes, a significant challenge arises from the inherently noisy nature of histogram depicting the frequency distribution of NDVI values.

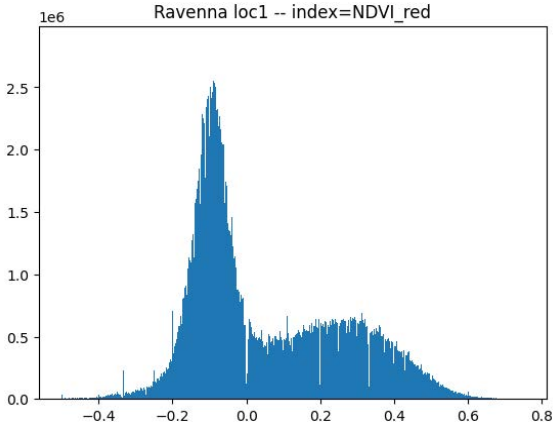
The observed noisy (fluctuating) frequencies often obscure the meaningful maxima and minima, as shown in Figure 2.3. This makes difficult to discern clear patterns.

To address this issue, we need to uncover the underlying data distribution by means of the Kernel Density Estimation (KDE). KDE works by estimating the distribution value for each point as the sum of the contributions of a kernel centred in each of the other points. Kernel can be one of well-known probability density functions; we use a Gaussian kernel with bandwidth related to the standard deviation of data points.

KDE acts as a smoothing mechanism for our data. The smoothed curve represents the frequency distribution function of our data (Figure 2.4); it is crucial in our analysis, since it makes possible to estimate local maxima and minima.

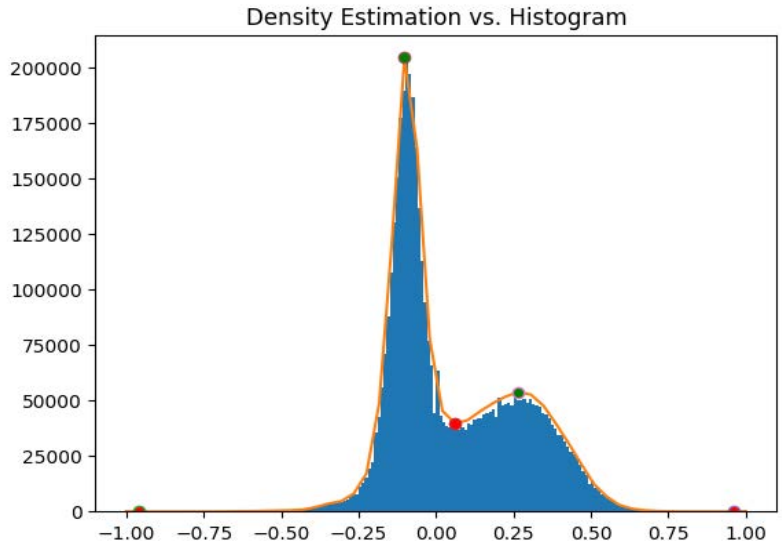
The local maxima, highlighted by KDE, serve a dual purpose: they indicate the likely number of clusters and provide the starting positions for the centroids in subsequent clustering algorithms. Both parameters are indeed crucial for a proper application of the subsequent clustering algorithms.

Figure 2.3 - Example of a histogram with noisy fluctuations around a clear bimodal distribution



Source: Author's processing

Figure 2.4 - Kernel Density Estimation KDE (orange curve) vs. histogram. Maxima (green dots) and minima (red dots) are shown



Source: Author's processing

Note that this approach lets us identify a first order estimate of the green areas: the threshold is the rightmost minimum in the distribution. This estimate does not require any additional parameter, *i.e.* the number of expected clusters.

2.3 Clustering

We experimented two different approaches for data point clustering: K-Means and K-Medians. Both are effective in grouping data into distinct clusters, which is essential in our study for pinpointing the exact positions of the green threshold within the NDVI histograms.

Leveraging the local maxima identified by KDE as the initial cluster centroids, K-means assigns each data point to the nearest one in terms of 2-norm Euclidean distance. Then a new version of the centroids is evaluated as the average of the positions of its associated points, and again points are assigned to the nearest centroid. The process eventually converges when positions are stable, that is when the algorithm minimises the sum of the squared distances between each data point and its assigned centroid.

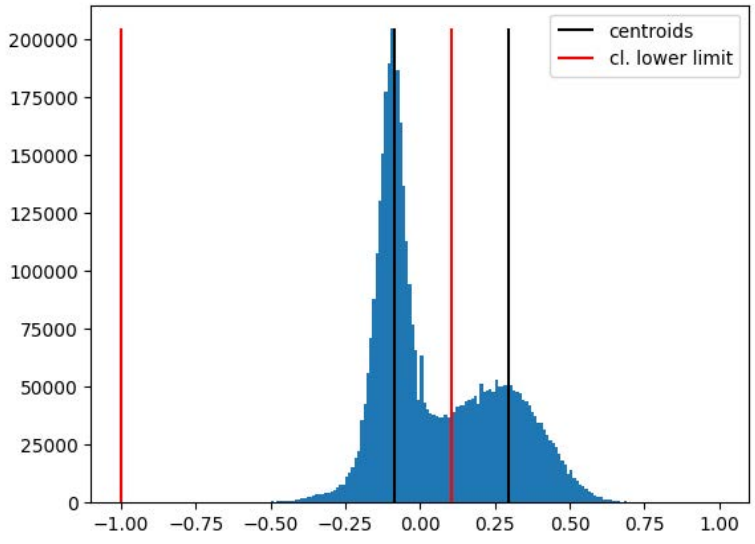
The urban green threshold is then given by the lower limit of the rightmost cluster, *i.e.* the cluster related to the centroid with the highest NDVI value.

K-Means enhances the precision of the estimate of the green threshold since it connects near points, which means points with analogous NDVI value (Figure 2.5).

We also explored the K-medians clustering approach. This because K-means is efficient for segmenting one-dimensional data, but it can be sensitive to outliers. K-medians addresses this issue effectively.

The algorithm works just the same as K-Means, with one relevant difference: it leverages on median to update positions of the centroids. In this way, it minimises the sum of the 1-norm absolute distances between each data point and its assigned centroid. Median is a positional statistical description and is robust with respect to the outliers. Centroid initialisation is the same as K-Means, so is the determination of the threshold as the lower value of the rightmost cluster.

Figure 2.5 - Clusters centroids (black lines) and clusters lower limits (red lines) found by K-Means



Source: Author's processing

Incorporating K-Medians into our methodology allowed us to validate and enhance the clustering results obtained from K-means. By using this complementary approach, we added a layer of robustness to our analysis, improving the overall accuracy and reliability of our conclusions about vegetation distribution in urban environments.

One of the challenges we encountered arises in histograms with only one maximum and no clear local minima. In such cases, our current strategy is to default to a two-cluster model, where the centroids initial positions are: the maximum and a fictitious point halfway the maximum and the value +1, the right full-scale for NDVI. This heuristic approach has the benefit to focus on the right-hand side of the histogram.

2.4 Advanced Pipeline

We aim to enhance our analysis of green vegetation detection in urban satellite imagery, refining the thresholds evaluated via the clustering methods. We focus particularly on two challenging aspects that are commonly encountered in this field. These issues are broadly present in green detection analysis and require innovative solutions to improve the accuracy and effectiveness of our approach.

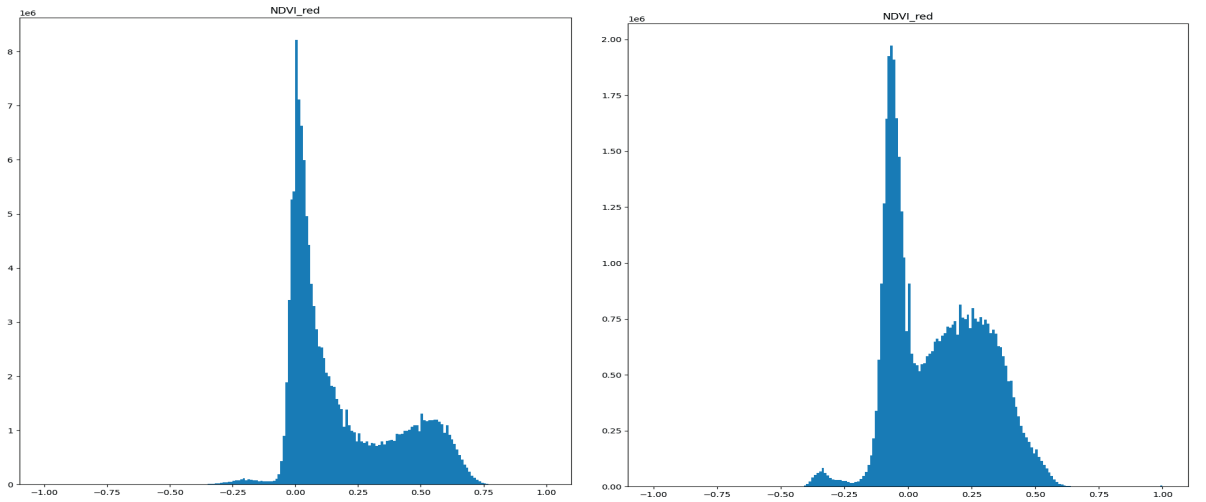
First, we encountered the challenge of varying green nuances across different images. One significant factor contributing to this variation consists of seasonal effects, which can alter the type of green detected in each image. For instance, the lushness of vegetation in spring presents a different shade of green compared to those of a later period (Eastman *et al.* 2013). The most relevant issue is a shift of the meaningful features of the histogram (Figure 2.6).

Second, we faced the challenge of the excessive presence of non-green areas in urban landscapes, which often overshadow the green clusters in our analyses. We aim at addressing this issue in an unsupervised fashion making the process as automatic as possible.

In doing so, we adopt the approach of Donchyts *et al.* (2018). This was originally applied in water-index detection and offers valuable insights for our context. We customised the methodology for vegetation analysis.

The method consists in a complete pipeline, as illustrated in the following (Figure 2.2).

Figure 2.6 – The number and positions of maxima and minima varies through different images



Source: Author's processing

- Initially, we employ the KDE algorithm, as previously explained, to establish a preliminary threshold for green pixels in the image. Pixels below the KDE threshold are turned black, while those above are kept unchanged. This masking enables the following steps to focus on green areas.
- We apply the Canny Edge Detection algorithm on the masked image. It performs a segmentation of the input image based on the variation of the index values, and is itself a sequence of algorithms: first, the image is convoluted with a Gaussian filter; second, gradients are evaluated; third, pixels with higher gradients are selected as parts of the segment, being the ones with higher variation in the index value. This step generates segments along the KDE threshold since there is a steep variation between green and non-green areas; and in the full green areas above KDE threshold, where it can separate faint-green by strong-green pixels.
- A buffering technique is applied around the detected edges. This step narrows the focus to the immediate areas surrounding the vegetation, providing a more targeted region for analysis. Buffered image is used as a mask on the original image to resample pixels in the regions of high index variation. This reduces noise from irrelevant regions.
- Finally, Otsu clustering is applied to the resampled histogram, to determine an optimal threshold for binary separation. It works on single tone images (such as the NDVI images are) and splits the values in two classes with a threshold, by minimising the intra class variance of the points. It efficiently distinguishes non-green from green areas. Further, if the green areas are non-homogeneous and present tones of green, it shifts the threshold towards the higher NDVI, so selecting areas with more intense vegetation.

2.5 Preliminary results

The illustrated methodology has been applied to the determination of the urban green threshold for the city of Ravenna, in Northern Italy.

Quantitative results are summarised in Table 2.1. For the ortho-images, the total number of pixels is 457,439,517 which gives 18,297,580.68 m^2 (1,829.76 Ha)

There is a general agreement for the first three approaches: KDE, K-Means and K-Medians. For a variation of the threshold in the range 0.06/0.11, a much smaller change in the percent of green area is observed: from 39.86% to 36.25%. This is because the transition between non-green and green in the histogram is localised in the minimum (KDE) or in its vicinity (clustering algorithms, Figure 2.5), so a shift in the threshold means a small change in the number of green pixels. Indeed, KDE and K-Medians are close, while K-Means is farther: this is due to the less robustness of the algorithm to the outliers, as mentioned above.

Table 2.1 - Results of the different approaches for the determination of the urban green threshold. Ortho-images for Ravenna, and satellite (Advanced pipeline S2)

Approach	threshold	Green pixels	Green area (Ha)	Green area (%)
KDE	0.06	182,322,979	729.29	39.86
K-Means	0.11	165,830,738	663.32	36.25
K-Medians	0.08	174,698,659	698.79	38.19
Advanced pipeline	0.18	136,061,208	544.24	29.74
Advanced pipeline S2	0.29	152,774,397	611.60	33.40

Source: Author's processing

A different trend is remarked by the advanced pipeline. In this case, there is a strong segmentation inside the green area, meaning a great variety of vegetal structures with highly varying NDVI, and the Otsu clustering determines a shift in the threshold to discard fainter green areas. This is a desired effect: if clustering allows us to split non-green from green, advanced pipeline makes us able to split soft green from strong green, refining the previous threshold determinations and enhancing the overall quality of the approach. This is an expected behaviour displayed by the ortho-images and is related to the high spatial resolution.

By a qualitative point of view, we can illustrate the effectiveness of our pipeline where the algorithm-detected green areas are distinctly marked. For instance, one of our images displays a detailed view of trees, correctly identified as green vegetation (Figure 2.7).

Figure 2.7 - Ortho-image of Ravenna (left), identification of Green Areas shown in Black for K-Medians (centre), and Advanced pipeline (right). Advanced pipeline selects stronger green areas, see red square detail

Source: Author's processing on AGEA ortho-image

Additionally, an intriguing aspect of our results is the accurate exclusion of non-vegetative green areas, such as a green-coloured football field. This area, despite its green hue, was correctly classified as non-vegetation due to its lack of infrared emission, a key component factored into the NDVI (Figure 2.8).

Figure 2.8 - Synthetic soccer field correctly excluded as green vegetation in the S2 advanced pipeline

Source: Author's processing on AGEA ortho-image

5. Conclusion

From the above, it can be inferred that ortho-images are particularly suitable for extracting and quantifying total greenery within urban areas, surpassing satellite images due to their lower spatial resolution. However, the timely and frequent information provided by satellite images is a valuable source. Further studies on the production of greenery statistics within urban areas in large cities are needed. Indeed, their use as a proxy in periods when ortho-images are not available is desirable. Moreover, methodological practices utilising Machine Learning techniques can help us overcome certain challenges, such as the threshold problem, in an automatic and objective manner.

The outcomes are positive and extremely encouraging, suggesting that future work will aim to improve methods, extend the area of investigation, quantify urban green areas more precisely, and classify them based on well-known land cover categories.

A project of this nature is reliant on both methodological and thematic expertise; indeed, specialists in both fields compose the team that conducted this research (the authors of this paper).

References

Chiocchini, R., A. Ferrara, and S. Mugnoli. 2018. "Quantificazione e analisi territoriale delle infrastrutture verdi sulla base di indicatori statistici: risultati della sperimentazione per l'area del IX municipio del comune di Roma Capitale". In Arcidiacono, A., D. Di Simine, S. Ronchi, and S. Salata (*eds*). *Centro di Ricerca sui consumi di suolo, Rapporto 2018*: 156-162. Roma, Italy: Istituto Nazionale di Urbanistica - INU Edizioni.

Donchyts, G., J. Schellekens, H. Winsemius, E. Eisemann, and N. van de Giesen. 2016. "A 30 m resolution surface water mask including estimation of positional and thematic differences using Landsat 8, SRTM and OpenStreetMap: A case study in the Murray-Darling Basin, Australia". *Remote Sensing*, Volume 8, N. 5: 386.

Eastman, J.R., F. Sangermano, E.A. Machado, J. Rogan, A. Anyamba. 2013 "Global Trends in Seasonality of Normalized Difference Vegetation Index (NDVI), 1982–2011". *Remote Sensing*, Volume 5, N. 10: 4799-4818.

Kriegler, F., W. Malila, R. Nalepka, and W. Richardson. 1969. "Preprocessing Transformations and Their Effects on Multispectral Recognition". In *Proceedings of the 6th International Symposium on Remote Sensing of Environment*: 97-131. Ann Arbor, MI, U.S.: University of Michigan.

Pristeri, G., F. Peroni, S.E. Pappalardo, D. Codato, A. Masi, and M. De Marchi. 2021. "Whose Urban Green? Mapping and Classifying Public and Private Green Spaces in Padua for Spatial Planning Policies". *ISPRS International Journal of Geo-Information*,Volume 10, N. 8: 538. https://doi.org/10.3390/ijgi10080538.

Xue, J., and S. Boafeng. 2017. "Significant Remote Sensing Vegetation Indices: a Review of Developments and Applications". *Journal of Sensors*, Volume 2017, N. 1: 1-17.

Machine Learning in Official Statistics: is explainability an issue?

Maurizio Naldi[1]

Abstract

The growing use of Machine Learning raises questions about its fruitful use in Official Statistics. An essential feature of Official Statistics is its transparency, both in terms of sources and data processing, a feature that is not always guaranteed by Machine Learning techniques, due to the lack of explainability of the results. In this work, after having identified the possible applications of Machine Learning in Official Statistics and the need for explainability, we show how explainability techniques are necessary only for some ML techniques and that solutions, based mainly on local and model-independent approaches, have already been proposed in the literature for many applications.

Keywords: Machine Learning; artificial intelligence; axplainability; quality; Official Statistics.

1. Introduction

Before the advent of Machine Learning (ML), Official Statistics were primarily shaped by well-established methodologies and manual data processing techniques, which conferred the results a high degree of reliability. However, the collection, analysis, and interpretation of data were often time-consuming, labour-intensive, and prone to human error. While these methods provided a solid foundation, they faced limitations in handling the increasing volume and complexity of modern datasets.

With the introduction of Machine Learning into the realm of Official Statistics, a seismic shift is going to occur. Machine Learning algorithms have the capacity to autonomously identify patterns, correlations, and trends within massive datasets, allowing statisticians to extract insights more efficiently and accurately. Machine Learning comes into play when we talk about big data. The possibility of collecting data autonomously, continuously, and through a data-driven approach and extracting statistics in real time has even led to coining the term smart statistics (Vichi and Hand 2019). The availability of data in large volumes and high frequency allows us to feed machine-learning algorithms and obtain accurate results, more than what model-based classification and regression tools allowed us to obtain in the past. An example is given by the adoption of decision trees and random forest algorithms to predict the individual employment status in Italy (Varriale and Alfò 2023).

The integration of Machine Learning techniques has revolutionised various facets of Official Statistics. In data collection, automated processes and advanced sampling strategies have streamlined survey design, reducing costs and improving response rates. Data cleaning and imputation benefit from Machine Learning's ability to handle missing or noisy data, enhancing the overall quality of statistical outputs. Moreover, forecasting and predictive modelling have reached new heights of sophistication, enabling statisticians to anticipate trends and changes with unprecedented precision.

1 Maurizio Naldi (m.naldi@lumsa.it), Università LUMSA, Roma, Italy.

The wide adoption of Machine Learning in Official Statistics was further self-evident in 2018, when the survey by Beck, Dumpert, and Feuerhake (2018) listed 136 machine-learning projects in 25 national statistical institutes, with classification and imputation (which is anyway a form of classification) being the major tasks and regression lagging far behind.

With classification in mind as the most popular task, both Meertens *et al.* (2022) and Kloos (2021) suggested ways of improving the statistical quality by reducing the misclassification bias that may mar Machine Learning performances.

However, this transition is not without challenges. Ensuring the transparency, interpretability, and ethical use of Machine Learning models in Official Statistics is a pressing concern. Striking a balance between the innovative capabilities of Machine Learning and the traditional principles of statistical integrity is crucial to maintaining public trust in the accuracy and reliability of official statistical information. At the same time that their role in Official Statistics was recognised, some worries about their wide use emerged since the inception of their usage (Braaksma and Zeelenberg 2015). In particular, concerns were expressed about their accuracy, the exact identification and stability of their coverage (which casts some doubt about the correspondence of the observed data with the statistical phenomenon that the National Statistics Institute wishes to describe). However, their availability cannot be ignored, and they can anyway be used as a (rich) additional source of data, however affected by misses and biases, which can help improve overall estimates. Those conclusions were written in 2015, but the suggestion by Braaksma and Zeelenberg (2015) is yet more convincing today. Solutions for the integration of new sources of data to reduce the selection bias often associated with those new sources have been proposed, *e.g.* by Righi *et al.* (2019) and D'Orazio (2023).

In this paper, we wish to assess whether the most doubt-provoking issue in the use of Machine Learning in Official Statistics, *i.e.* explainability, is really an issue. In Section 2, we provide a panorama of possible uses of ML in Official Statistics and enumerate the reasons why explainability is important in that realm in Section 3. We close in Section 4 by highlighting how explainability techniques are being proposed and tested for all the potential uses of ML in Official Statistics.

2. Machine Learning and Official Statistics

Incorporating Machine Learning into Official Statistics processes can lead to more accurate, timely, and relevant data, ultimately improving decision-making, policy development, and public understanding of various economic, social, and environmental trends.

Machine Learning can be applied in either a supervised or an unsupervised way. In supervised Machine Learning, we feed the ML algorithm with a dataset, which we use for training and testing the algorithm, where we know the ground truth, *i.e.* the output we expect from the algorithm. Supervised tasks are classification and regression, the difference between the two lying in the type of output we expect. The output will be a class, *i.e.* a discrete-value variable, in classification, and a continuous numeric variable in regression. The information for either classification or regression comes from a set of features that we expect to know as a proxy for the target variable (class or numeric value) of interest. In unsupervised tasks, we do not know the ground truth and we just aim at finding common patterns, *i.e.* clustering the data into groups that are as homogeneous as possible (based on our knowledge of the associated features as in supervised Machine Learning).

What is the use of ML in Official Statistics? A useful, though rough statement, is that Statistics is the art of counting (Stone 2020), but we cannot or do not want to always count within the whole population. Sometimes, we cannot do that because the data we are interested in are not available. Sometimes, we do not want to do it, because it would be too costly. Statisticians are well versed in solving this problem and resort to sampling, *i.e.* counting over a sample and projecting the results over the whole population. The problem there lies mainly in correctly choosing the sample so that it is really representative of the whole population or introducing corrections to account for the discrepancies between the sample and the population.

However, how can ML help there? A problem with counting is that you have to know when some instance is to be counted. For example, if you wish to know how many buildings are equipped with roof solar panels, you have to be able to recognise when a building roof hosts a solar panel, so that you can include that building in the count (see the example employed by Kloos 2021). If you know that just for a small proportion of buildings, because you were able to inspect just a few buildings, you can train a Machine Learning algorithm to recognise solar panels based on the relatively small dataset you have manually labelled and some features (*e.g.* the pixels making a photo snap of the building roof) and then use that algorithm to automatically label a very large number of building roofs or even the whole population of building roofs.

We can then draw a most general answer by saying that ML helps when it can replace manual labelling over the whole population by machine-based labelling for the whole population (or a very large proportion of it, so that sampling bias is reduced as much as possible). But ML can be of help in more cases than just improving our counting. In the following, we try to list the major applications where ML can be of help.

Data Imputation: Machine Learning can be used to fill in missing or incomplete data in Official Statistics. Algorithms can analyse existing data and make predictions to impute missing values, improving the overall data quality. Examples of methods for this task are described by Batista and Monard (2003) and Poulos and Valle (2018).

Data Quality Assurance: ML algorithms can help identify and correct errors, inconsistencies, or outliers in datasets, ensuring that the data used for statistical analysis is accurate and reliable.

Survey Sampling: Machine Learning techniques can optimise survey sampling methods to select representative samples more efficiently. This helps in reducing the cost and time associated with data collection.

Data Classification: ML can be used to automatically classify and categorise data, making it easier to organise and analyse information. For example, it can categorise products into different industries or services into various sectors.

Anomaly Detection: Official Statistics often involve the detection of unusual patterns or anomalies. Machine Learning can help identify outliers or irregularities in data, which may indicate errors or significant changes in trends.

Time Series Forecasting: Machine Learning models, particularly time series forecasting algorithms, can predict future trends and values based on historical data. This is valuable in economic forecasting and population projections.

Sentiment Analysis: Analysing social media data and other unstructured sources using natural language processing and sentiment analysis can provide insights into public sentiment, which is valuable for understanding public opinion and potential biases.

Geospatial Analysis: Machine Learning can enhance geospatial data analysis, aiding in mapping and monitoring trends and spatial patterns, such as population distribution, land use, and environmental changes.

Data Linkage: ML techniques can be used to link datasets from different sources, allowing Official Statistics agencies to integrate data from multiple domains, leading to more comprehensive insights.

Fraud Detection: Official Statistics agencies often deal with financial and economic data. Machine Learning models can help detect fraudulent activities and transactions within these datasets.

Natural Language Processing: NLP techniques enable automated text analysis, which can be valuable for extracting structured information from unstructured text data like reports, news articles, and survey responses. Automated Reporting is a particularly useful feature, whereby ML systems can generate automated reports and summaries based on statistical analysis, reducing the manual effort required to produce official reports and publications. NLP itself can be seen as an aid to explainability, offering a text-based input/output interface that helps familiarity with non-expert users of ML software.

Data Visualisation: Machine Learning can be used to create advanced data visualisation tools that help present statistics in a more understandable and compelling manner, making it easier for the public to interpret the data.

3. Explainability in Official Statistics

Though explainability is a notion of general applicability, it has received great attention in the context of Machine Learning. If we look at the results on Scopus for explainability[2], we get 6652 results, but 6293 (94.6%) are associated with either Machine Learning or artificial intelligence. In this Section, we review the reasons why explainability is relevant in the field of Official Statistics.

First, we have the need for transparency. Official Statistics play a crucial role in informing public policy, decision-making, and public understanding. To maintain public trust, it is essential that the methods and algorithms used in the production of Official Statistics are transparent and explainable. Explainable models provide insights into how statistical conclusions are reached, making it easier for stakeholders to trust and scrutinise the data. Transparency appears to have been first introduced in the context of Official Statistics in version 2.0 of the Quality Assurance Framework of the European Statistical System (Eurostat 2019), where it is stated that the statistical authorities have to document their production processes to achieve transparency of processes. The need for transparency in Official Statistics has later been recognised in the general guidelines for Official Statistics set by the US (National Academies of Sciences and Medicine 2017) and expressly addressed in the consensus study by a panel within the National Academies of Sciences and Medicine 2022. the relationship between transparency and algorithms is further investigated by D'Acquisto (2022) in Section V.3 of his book.

Government agencies responsible for Official Statistics are also to be accountable for the accuracy and reliability of the data they produce, as mentioned in the preamble of the Fundamental Principles of Official Statistics stated by the United Nations[3]. Explainability

2 Data retrieved on November 15th, 2023.

3 The full resolution can be read at https://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf.

ensures that agencies can account for their methods and defend the validity of their statistical results. Reference to those principles has been made for several national statistics institutes, *e.g.* in Greece (Michalopoulou 2015). The need for the accountability of those agencies must also be traced along the funding chain from the government so that such financing is seen as legitimate, as highlighted by Pullinger (2022). In many countries, there are legal and regulatory requirements governing the production and dissemination of Official Statistics, and explainability is often a requirement to ensure that the methods used comply with these regulations. The feature of explainability may also help in addressing the presence of biases or unfairness that may be inadvertently introduced in the data (*e.g.* the problem of preferential sampling in surveys addressed by Vedensky, Parker, and Holan (2023)). An explainable model allows for the identification of potential sources of bias and provides a basis for addressing these issues to ensure that Official Statistics are fair and unbiased.

Explainability also aids in quality control by enabling statisticians and data analysts to understand the behaviour of models and algorithms. This understanding allows them to identify and correct errors or inconsistencies in the data more effectively. The relevance of explainability to quality is so outstanding that it is mentioned as one of the five dimensions of quality by Yung *et al.* (2022).

Since Official Statistics are meant to be used by a broad audience, including policymakers, researchers, and the general public, explainable statistics are more accessible to these users, enabling them to interpret and use the data more effectively for their specific needs. This is especially important, when Official Statistics inform wide-ranging decisions, from economic policies to public health strategies. Decision-makers rely on the credibility of the data and the reasoning behind it. Explainable statistics provide a clear rationale for the numbers, making it easier for decision-makers to trust the data in their policy and strategy formulation. This is particularly true in the presence of economic and social shocks as that experienced during the recent COVID pandemics, as analysed by Oleński (2023). The presence of explanations for the generation of Official Statistics also helps build public trust in Official Statistics. When the public understands how the Statistics are generated, including the methods used and the data sources, they are more likely to trust the information and use it for various purposes. Mistrust and distrust in Official Statistics have been recognised as a growing threat, but they may also as a spur to build trust (Lehtonen 2019).

Finally, explainability also allows us to address ethical issues. The ethical use of data is paramount in Official Statistics, and explainable models help identify ethical concerns such as data privacy violations, and they make it easier to uphold ethical principles in data collection, processing, and dissemination.

4. Closing the circle: is explainability an issue?

We have seen in Section 2 that Machine Learning is going to be more and more used in Official Statistics. We have seen in Section 3 that explainability is a desirable (or even required) feature in Official Statistics, and it is certainly a growing issue in Machine Learning. We are now in a position to investigate the possibility of fulfilling the explainability needs of Official Statistics with Machine Learning tools.

We can roughly divide the panorama of Machine Learning algorithms into two groups as far as explainability is concerned. We can borrow the classification put forward by Thampi (2022) for interpretability. In his book, Thampi distinguishes between white-box models, which are inherently transparent, and black-box ones, which are not. Examples of white-box models are linear and logistic regression, decision trees, and generalised additive models (GAMs). We could add Naive Bayes and Support Vector Machines (SVM). In these models, tracking the reasons for classification is quite easy. We can understand how the input features are transformed into the target variable and can also identify the features playing the most significant role in predicting the target variable. On the other hand, the path from input to output is quite less transparent in black-box models. Examples of black-box models are all the algorithms based on ensemble techniques and neural networks. In the former group, we can include Random Forests, bagging techniques and boosting techniques (*e.g.* AdaBoost, Gradient Boosting, and XGBoost). Even less transparency is guaranteed with neural networks, *e.g.* Convolutional neural networks (CNNs) and Recurrent neural networks (RNNs), and deep learning neural networks.

Since white-box models can be considered easily explainable, we have to focus on black-box models to identify the threats they pose to explainability and look for possible workarounds. Guidotti *et al.* (2018) provided a thorough survey of explainability techniques for black-box models.

Hereafter, we review the tasks identified in Section 2 and see how they have been approached in the literature to solve the explainability issue. It is to be noted that most references do not concern applications in Official Statistics, but their approach may be borrowed and employed in similar tasks rising in Official Statistics.

As to data imputation, a local post-hoc explainability approach has been proposed by Cinquini *et al.* (2022) to handle missing values. Başağaoğlu *et al.* (2022) have proposed SHaply Additive eXplanation (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) to add explainability to boosting techniques (Extreme Gradient Boosting, Light Gradient Boosting, and Categorical Boosting), Extremely Randomised Trees, and Random Forest in the context of hydroclimatic applications.

For the help that Machine Learning techniques can give in data classification, all the explainability techniques adopted for general classification may be used.

Several approaches have been proposed to address explainability issues in anomaly detection tasks. For example, Nguyen *et al.* (2023) again adopt either SHAP or LIME.

For time-series forecasting, the combination of numeric association rules between input (past values) and output (prediction) and visual explanation techniques is employed by Troncoso García *et al.* (2023) to explain and interpret multi-step time series forecasting models. Saliency maps are instead used by Saadallah, Jakobs, and Morik (2019) to properly prune models in ensemble learning from a combination of deep learning neural networks. Freeborough and Zyl (2022) use ablation, permutation, random noise, and integrated gradients to explain the output of RNNs and LSTMs.

For sentiment analysis, solutions based on LIME, SHAP, and model attention have been proposed by Yadav, Kaushik, and McDaid (2023) to explain the identification of conversational hate speech. Again, LIME, SHAP, and Eli5 are employed to explain sentiments about vaccination by de Camargo *et al.* (2023). Shapley values are exploited to extract the important features from tourism online reviews by De Nicolò *et al.* (2023). In their survey of customer reviews of food delivery services, Adak, Pradhan, and Shukla (2022) found that the two most used explainability techniques were LIME and SHAP.

The explainability of geospatial analyses for monitoring land characteristics is also addressed by Zhang *et al.* (2023), who provide an explanatory framework for landslide susceptibility evaluation models based on the SHAP-XGBoost (eXtreme Gradient Boosting) algorithm. The poor explainability of most AI models for geospatial analysis has been stressed by Gonzales-Inca *et al.* (2022) in hydrological and hydraulic modelling, water quality modelling, and fluvial geomorphic and morphodynamic mapping, and by McCord *et al.* (2022) in tax mass appraisal.

Again, SHAP is employed for credit card fraud detection by Biswas *et al.* (2023). Psychoula *et al.* (2021) analyse the explainability performance of several ML models before and after the application of SHAP and LIME. The same SHAP plus LIME and WIT (What-If Tool) are instead employed by Buyuktepe *et al.* (2023) for food frauds.

The application of XAI in NLP has been reviewed by Mathews (2019) and El Zini and Awad (2022). Occhipinti, Rogers, and Angione (2022) have analysed 12 Machine Learning models using SHAP. The use of NLP as an aid in explainability is advocated for by Dessureault and Massicotte (2023), where graphs, tables, and text are employed to accompany ML results.

5. Conclusions

Explainability is an essential feature of Official Statistics. Though the use of Machine Learning may raise some doubt about the possibility of achieving such a feature with black-box models (typically based on ensemble or deep learning approaches), explainability has been shown to be possible for most of the foreseeable applications of ML in Official Statistics. The approaches adopted in most cases rely on SHAP and LIME, *i.e.* local model-agnostic models. The application of such model-agnostic methods allows us to expect a wide applicability. We can expect to be able to borrow explainability-oriented algorithms from other application fields into the more cautious world of Official Statistics.

References

Adak, A., B. Pradhan, and N. Shukla. 2022. "Sentiment analysis of customer reviews of food delivery services using deep learning and explainable artificial intelligence: Systematic review". *Foods*, Volume 11, N. 10: 1500.

Başağaoğlu, H., D. Chakraborty, C. Do Lago, L. Gutierrez, M.A. Sahinli, M. Giacomoni, C. Furl, A. Mirchi, D. Moriasi, and S.S. Sengör. 2022. "A review on interpretable and explainable artificial intelligence in hydroclimatic applications". *Water* , Volume 14, N. 8: 1230.

Batista, G., and M.C. Monard. 2003. "An analysis of four missing data treatment methods for supervised learning". *Applied artificial intelligence*,Volume 17, N. 5-6: 519-533.

Beck, M, F. Dumpert, and J. Feuerhake. 2018. *Machine Learning in Official Statistics*. Available at *ArXiv*. https://doi.org/10.48550/arXiv.1812.10422.

Biswas, J., A.A. Mridha, M.S. Hossain, A.S. Trisha, M.S. Ahmed, and M.I. Hossain. 2023. "Interpretable credit card fraud detection using machine learning leveraging SHAP". *Conference paper* in *IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT)*: 1206-1211. Qingdao, China.

Braaksma, B., and K. Zeelenberg. 2015. "Re-make/Re-model: Should big data change the modelling paradigm in official statistics?". *Statistical Journal of the IAOS* , Volume 31, N. 2: 193-202.

Buyuktepe, O., C. Catal, G. Kar, Y. Bouzembrak, H. Marvin, and A. Gavai. 2023. "Food fraud detection using explainable artificial intelligence". *Expert Systems,* Volume 31, N. e13387.

Cinquini, M., F. Giannotti, R. Guidotti, and A. Mattei. 2022. "Handling missing values in local post-hoc explainability". In *World Conference on Explainable Artificial Intelligence*: 256-278. Heidelberg, Germany: Springer.

D'Acquisto, G. 2022. *Decisioni algoritmiche.* Torino, Italy: Giappichelli.

D'Orazio, M. 2023. "Statistical learning in official statistics: the case of statistical matching". *Statistical Journal of the IAOS*, Volume 35, N. 3: 435-441.

de Camargo, L.F., J. da Costa Feitosa, E. Bonatti, G. Ballminut Simioni, and J. Remo Ferreira Brega. 2023. "Explainable artificial intelligence-a study of sentiments about vaccination in Brazil". In *International Conference on Computational Science and Its Applications*: 617-634. Heidelberg, Germany: Springer.

De Nicolò, F., L. Bellantuono, D. Borzı, M. Bregonzio, R. Cilli, L. De Marco, A. Lombardi, E. Pantaleo, L. Petruzzellis, and A. Shashaj. 2023. "The verbalization of numbers: An explainable framework for tourism online reviews". *International Journal of Engineering Business Management*, Volume 5: 1-16.

Dessureault, J-S, and D. Massicotte. 2023. "AI²: the next leap toward native language-based and explainable machine learning framework". *Automated Software Engineering*, Volume 30, Article 32.

El Zini, J., and M. Awad. 2022. "On the explainability of natural language processing deep models". *ACM Computing Surveys*,Volume 55: 1-31.

Eurostat, European Commission. 2019. *Quality Assurance Framework of the European Statistical System*. *Version 2.0*. Luxembourg: Publication Office of European Union.

Freeborough, W., and T. van Zyl. 2022. "Investigating explainability methods in recurrent neural network architectures for financial time series data". *Applied Sciences*, Volume 12, N. 3: 1427.

Gonzales-Inca, C., M. Calle, D. Croghan, A. Torabi Haghighi, H. Marttila, J. Silander, and P. Alho. 2022. "Geospatial artificial intelligence (GEOAI) in the integrated hydrological and fluvial systems modeling: review of current applications and trends". *Water*, Volume 14, N. 14, Article 2211.

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. 2018. "A survey of methods for explaining black box models". *ACM computing surveys (CSUR)*, Volume 51, N. 5: 1-42.

Kloos, K. 2021. "A new generic method to improve machine learning applications in official statistics". *Statistical Journal of the IAOS*, Volume 37, N. 4: 1181-1196.

Lehtonen, M. 2019. "The multiple faces of trust in statistics and indicators: A case for healthy mistrust and distrust". *Statistical Journal of the IAOS*, Volume 35, N. 4: 539-548.

Mathews, S.M. 2019. "Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review". In Arai, K., R. Bhatia, and S. Kapoor (*eds*). *Intelligent Computing. CompCom 2019. Advances in Intelligent Systems and Computing*, Volume 998. Cham, Switzerland: Springer.

McCord, M.J., P.T. Davis, P.E. Bidanset, and L.D. Hermans. 2022. "Prediction accuracy for property tax mass appraisal: A comparison between regularised machine learning and the eigenvector spatial filter approach". *Journal of Property Tax Assessment & Administration*, Volume 19, N. 2: 539-548.

Meertens, Q.A., C.G.H. Diks, H.J. van den Herik, and F.W. Takes. 2022. "Improving the output quality of official statistics based on machine learning algorithms". *Journal of Official Statistics*, Volume 38, N. 2: 485-508.

Michalopoulou, C. 2015. "Professional independence and accountability of statistical agencies are crucial: A brief history of the Greek official statistics". *Statistical Journal of the IAOS*, Volume 31, N. 4: 507-512.

National Academies of Sciences, Engineering, and Medicine. 2022. *Transparency in Statistical Information for the National Center for Science and Engineering Statistics and All Federal Statistical Agencies*. Washington, DC, U.S.: the National Academies Press.

National Academies of Sciences, Engineering, and Medicine. 2017. *Principles and Practices for a Federal Statistical Agency*. Washington, DC, U.S.: the National Academies Press.

Nguyen, M.D., A. Bouaziz, V. Valdes, A.R. Cavalli, W. Mallouli, and E. Montes De Oca. 2023. "A deep learning anomaly detection framework with explainability and robustness". *ARES '23: Proceedings of the 18th International Conference on Availability, Reliability and Security*, Article 134: 1-7.

Occhipinti, A., L. Rogers, and C. Angione. 2022. "A pipeline and comparative study of 12 machine learning models for text classification". *Expert Systems with Applications*, Volume 201, Article 117193. https://doi.org/10.1016/j.eswa.2022.117193.

Oleński, J. 2023. "Quality of the global information environment: The prerequisite of social and economic sustainability". In *Handbook of Research on Socio-Economic Sustainability in the Post-Pandemic Era*: 334-347. Hershey, PA, U.S.: IGI Global.

Poulos, J., and R. Valle. 2018. "Missing Data Imputation for Supervised Learning". *Applied Artificial Intelligence*,Volume 32, N. 2: 186-196.

Psychoula, I., A. Gutmann, P. Mainali, S.H. Lee, P. Dunphy, and F. Petitcolas. 2021. "Explainable Machine Learning for Fraud Detection". *Computer*, Volume 54, Issue 10: 49-59.

Pullinger, J. 2022. "Financing official statistics: Some reflections". *Statistical Journal of the IAOS*, Volume 38, N. 2: 439-441.

Righi, P., G. Bianchi, A. Nurra, and M. Rinaldi. 2019. "Integration of survey data and big data for finite population inference in official statistics: Statistical challenges and practical applications". *Statistica & Applicazioni*,Volume XVII, N. 2: 135-158.

Saadallah, A., M. Jakobs, and K. Morik. 2019. "Explainable online ensemble of deep neural network pruning for time series forecasting". *Machine Learning*, Volume 111, N. 9: 3459-3487.

Stone, D. 2020. *Counting: How we use numbers to decide what matters*. New York, NY, U.S.: Liveright Publishing.

Thampi, A. 2022. *Interpretable AI: Building explainable machine learning systems*. New York, NY, U.S.: Simon & Schuster.

Troncoso-García, A.R., M. Martínez-Ballesteros, F. Martínez-Álvarez, and A. Troncoso. 2023. "A new approach based on association rules to add explainability to time series forecasting models". *Information Fusion*, Volume 94: 169-180.

Varriale, R., and M. Alfò. 2023. "Multi-source statistics on employment status in Italy, a machine learning approach". *Metron*, Volume 81: 37-63.

Vedensky, D., P.A. Parker, and S.H. Holan. 2023. "A Look into the Problem of Preferential Sampling through the Lens of Survey Statistics". *The American Statistician*, Volume 77, N. 3: 313-322.

Vichi, M., and D. Hand. 2019. "Trusted smart statistics: the challenge of extracting usable aggregate information from new data sources". *Statistical Journal of the IAOS*, Volume 35, N. 4: 605-613.

Yadav, S., A. Kaushik, and K. McDaid. 2023. " Understanding Interpretability: Explainable AI Approaches for Hate Speech Classifiers". In Longo, L. (*ed*). *Explainable Artificial Intelligence. xAI 2023. Communications in Computer and Information Science*, Volume 1903: 47-70. Cham, Switzerland: Springer.

Yung, W., S-M. Tam, B. Buelens, H. Chipman, F. Dumpert, G. Ascari, F. Rocci, J. Burger, and I. Choi. 2022. "A quality framework for statistical algorithms". *Statistical Journal of the IAOS*, Volume 38, N. 1: 291-308.

Zhang, J., X. Ma, J. Zhang, D. Sun, X. Zhou, C. Mi, and H. Wen. 2023. "Insights into geospatial heterogeneity of landslide susceptibility based on the SHAP-XGBoost model". *Journal of Environmental Management*, Volume 332.

SESSION 3

Quality for non-traditional sources

Session 3
Master class | Quality for new data sources: Progress, challenges and directions for the European Statistical System

Fabio Ricciato[1]

Abstract

In this short contribution we reflect on the implications of (re)using privately-held data in Official Statistics from the perspective of "quality". Starting from the notion of "quality" as defined in the context of European statistics, we show that the regular production of Official Statistics based on privately-held data entails a combination of potential quality benefits and quality costs. Statistical authorities should carefully assess and select use-cases and data sources for which the cost-vs-benefit balance is positive. In the context of the European Statistical System, we highlight the factors that motivate the development of a common methodological framework, open-source tools and share infrastructures at the European level.

Keywords: Official Statistics; quality; non-traditional data sources; privately-held data.

1. Quality in Official Statistics

In the context of Official Statistics the term "quality" has assumed a peculiar meaning, wide in scope and embracing multiple dimensions. For the European Statistical System (ESS), the quality dimensions are defined in the European Statistics Code of Practice[2] (CoP) and in the Quality Assurance Framework[3] (QAF). Taken together, these documents define the self-regulatory framework that distinguishes *Official Statistics* from other sources of statistical information like, *e.g.* commercial statistics and other public statistics.

The notion of quality as defined in COP/QAF spans three areas, namely *institutional environment, statistical processes* and *statistical output*. This approach is motivated by the consideration that the characteristics of the final statistical figures (*What statistics* are produced) depend on the underlying production process (*How* they are produced), and the latter in turn depends on the surrounding production environment (*Who* produces them). For each of these areas, the COP and QAF define principles and indicators. These documents are at the same time *prescriptive and aspirational*: they set minimum conditions to be fulfilled, but also high-level objectives to be pursued in a perspective of continuous improvement.

The notion of *quality* in Official Statistics is very articulated, as we have just seen, but at the same time *dynamic*: it is indeed a continuously evolving concept. This is not surprising when one considers that the whole system of *Official* Statistics keeps evolving, and is itself embedded in an ever-changing societal context. The evolution of quality in Official Statistics is reflected in the temporal flow of norms at different levels: the EU Regulation 223/2009 on European Statistics was adopted in 2009, amended first in 2015 and then revised again in March 2024; the CoP was published first in 2005 with a second and third version in 2011 and 2017, respectively; the QAF was published first in 2011 and revised in 2019. It is natural

1 Fabio Ricciato (fabio.ricciato@ec.europa.eu), European Commission, Eurostat. The views and opinions expressed are those of the author and do not necessarily reflect the official policy or position of the European Commission.

2 European Statistics Code of Practice - revised edition 2017.https://ec.europa.eu/eurostat/documents/4031688/8971242/KS-02-18-142-EN-N.pdf.

3 Quality Assurance Frameworkof the ESS-version 2.0.2019. https://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V2.0-final.pdf.

to expect that both CoP and QAF will be revised again following the adoption of the new regulation on European Statistics.

2. New data sources in Official Statistics: quality benefits and costs

Official Statistics are traditionally produced based on a combination of primary statistical data, from censuses and surveys, and administrative records. Such *traditional data* sources are becoming insufficient to meet the growing demands and expectations by statistical users for more, better, richer, and timelier statistics.The ESS is now preparing to extend (future) production processes to leverage also other types of *new* "non-traditional" data sources, including data generated and held in the private sector, *i.e.* Privately-Held Data (PHD).

It is important to remark that PHD, like other new data sources, are set to *augment, not replace* traditional data sources (Baldacci *et al.* 2021). As discussed below, a new generation of statistical products can be developed based on the integration of non-statistical PHD with statistical data, going beyond what would be possible with each of the two in isolation.

Looking through the glasses of *quality* as defined in CoP and QAF, the perspective of (re) using PHD for Official Statistics brings opportunities but also major challenges.

On the one hand, the motivation for considering PHD in the first place is rooted in the expectation that they will *increase the quality of statistical output*, by enabling the production of new, more, better, richer, and timelier statistics, mapping to the relevant CoP/QAF Principles11-14, namely *Relevance, Accuracy and Reliability, Timeless and Punctuality, Coherence and Comparability*. In the hypothetical scenario where the prospected new statistics (with comparable characteristics in terms of statistical output) were to be produced without recurring to PHD, the resulting cost and burden on respondents would be unbearable: in this sense leveraging PHD may be considered instrumental to *preserve the quality of the statistical processes* in terms of the CoP/QAF Principles 9-10, namely *Non-excessive Burden on Respondents* and Cost *effectiveness*.

On the other hand, bringing PHD into the statistical production requires finding new solutions to issues that touch almost all items in the CoP/QAF. Furthermore, some issues that are relevant for PHD go beyond the scope of the current CoP/QAF and require new extensions, for example in the direction of ensuring compliance with non-statistical legislation (*e.g.* telecom legislation for location data sourced from mobile networks or mobile phones); public acceptance for the secondary (re)use of highly granular personal data that were primarily collected for other purposes; sustainability of partnership models between statistical authorities and private dataholders; reproducibility, maintainability, explainability and sensitivity of softwarised statistical methods (Ricciato 2022). These are examples of issues and additional dimensions that may be expected to make their way into the future version of COP/QAF.

Taking a bird's-eye view over the quality opportunities vis-à-vis the quality challenges shows clearly that the (re)use of PHD in Official Statistics entails a cost-vs-benefit balance: the expected *quality benefits* associated to improved statistical output must be high enough to offset the *quality costs* incurred in fulfilling minimum conditions on almost all quality dimensions (*the gain must be worth the pain*).

It cannot be assumed that the cost-vs-benefit balance will be always positive: we must accept that in some cases the quality costs may prevail over the quality benefits. This simple consideration

should drive statistical authorities to be selective and assess very carefully the use-cases and data sources for which the anticipated quality benefits justify the prospective quality costs. In doing so, they should consider not only the initial costs required to *achieve* the stage of regular statistical production (*e.g.* research and experimentation; initial development of methodologies, software, and data interfaces; negotiation and establishment of partnership agreements with data holders; deployment of business processes) but also the operational costs to *sustain* regular statistical production (*e.g.* maintenance and continuous update of methodologies, software, and data interfaces; maintenance of partnerships and business processes).

Such cost-vs-benefit assessment, and the consequent selection of use-cases and data sources, should be conducted at different stages and start as early as possible in the statistical development path. Even a qualitative assessment should suffice to identify upfront those data sources for which the prospective quality costs, to achieve and/or sustain regular statistical production, are likely too high vis-à-vis the expected benefits. The technological maturity and market structure of the business sector where data are generated should be considered as key dimensions for the assessment. For instance, low levels of technological maturity and penetration, and high levels of market fragmentation and data heterogeneity, are elements that contribute to drive upwards several cost factors. Such analysis exercise should be preferably carried out in the perspective of statistical production at European scale, and should guide the allocation of resources and investments in methodological development and experimentation at the ESS level.

3. New data sources and new processes for new statistics

The secondary (re)use for statistical purposes of data generated primarily for nonstatistical purposes requires the establishment of new organisational processes. In many cases, the new processes will involve also the data holder(s) to some extent. The data holders must provide access to the data and to the associated meta-data, with modalities that need to be agreed with statistical authorities. But there is more information to provide: as statistical production extends to PHD, the (primary) data generation process enters the equation, therefore knowledge about the business and technological aspects that drive *how the data are produced, and therefore determine what information they carry*, becomes an essential component of the (secondary) reuse process in Official Statistics. Communicating such knowledge and information, called *para-data* in the ESS Handbook for quality and metadata reports[4], is required not only in the methodological development phase, to enable proper design and implementation of data processing methods, but also in the operational production stage. For example, anomalies, interruptions, and errors affecting the data generation process, hence the quality of the data that will eventually enter the statistical production pipeline, must be properly and promptly communicated by the data holder(s) to the statistical authority. Such dialogue will be unavoidably bi-directional: the statistical authority may detect unreported anomalies or implausible patterns, possibly based on comparison with other data sources, that are to be reported back to the data holder(s) to be correctly interpreted, possibly corrected or anyway mitigated. These examples imply that rules and roles on the side of both organisations, statistical authority and data holder(s), must be defined to ensure that events and incidents affecting statistical quality are properly detected and communicated. This requires the definition of agreed criteria (What information is relevant and should be communicated? What exactly should be considered "anomalous"?), functions and

4 ESS Handbook for quality and metadata reports – 2021 re-edition.https://ec.europa.eu/eurostat/documents/3859598/13925930/KS-GQ-21-021-EN-N.pdf.

policies (Who is in charge? Who shall communicate to whom?), interfaces and templates (How shall the communication take place?), etc. All these aspects must be encoded into a quality system designed specifically for PHD.

The deployment and execution of such processes will unavoidably consume resources and create additional burden on the side of the data holders as well as on the side of the statistical authorities. Receiving and processing information is not less resource consuming than preparing and transmitting it, and both the transmitter and the receiver share the common goal of minimising the amount of transferred information. Statistical authorities and data holders could cooperate to define communication processes that are not only effective but also efficient for each side, keeping the cost and burden down to the minimum possible level without jeopardising statistical quality.

4. Conditions for successful partnerships with data holders

The High-level Expert Group on facilitating the use of new data sources for Official Statistics states that the (re)use of PHD for Official Statistics should be based on fair and effective partnerships between businesses and statistical authorities, underpinned by a legal framework setting out clear requirements and safeguards for private data holders[5]. When the additional costs incurred by private data holders to enable data reuse for Official Statistics are substantial, they should receive financial compensations based on a fair reference cost model. Furthermore, on top of legislative and financial measures, the Expert Group recommends that statistical authorities put in place non-financial incentives to motivate data holders to cooperate with statistical authorities (for additional details see Eurostat 2022).

In the ideal scenario, the partnership model is designed in a way to let private data holders benefit not only from the act of cooperating with the statistical authorities (*e.g.* improved corporate reputation) but also from the statistical products (or by-products) deriving from such cooperation. If the data holders see their interest in that the final statistical product to which they contribute is of highest possible quality, then their cooperation efforts would go beyond the necessity to fulfill compliance requirements.

Translating this abstract goal (or wish) into concrete operational terms is admittedly very difficult, and in some cases devising a convincing system of incentives for the businesses will not be possible. However, statistical authorities should at least explore this direction and attempt to identify such set of incentives. The efforts may be successful in some business sectors. Hereafter we outline a possible approach for one specific kind of data, namely Mobile Network Operator (MNO) data, that may inspire similar reasoning for other business sectors.

Private businesses are already leveraging location data derived from the operation of mobile networks (MNO data) to deliver commercial statistics and "mobile analytics" services (terminology borrowed from Eurostat 2023). The success of this line of business should be seen positively by statistical authorities, as it sustains the business investments necessary to ensure MNO data availability in general (*e.g.* technical infrastructure, organisational processes), and specifically for reuse in Official Statistics. In other words, the use of MNO data for commercial analytics purposes is indirectly supportive of (rather than detrimental to) the prospective reuse

5 Empowering society by reusing privately-held data for Official Statistics – A European approach. Final Report of the Expert Group on facilitating the use of new data sources for Official Statistics, 2022. https://ec.europa.eu/eurostat/web/products-statistical-reports/-/ks-ft-22-004

of such data for Official Statistics. In the reverse direction, perhaps we can imagine a system where the public release of Official Statistics based on MNO data would not be detrimental but rather beneficial, at least indirectly, to the market demand for commercial analytics based on the same data. Such a hypothetical system would need to fulfill at least three necessary conditions.

First, Official Statistics based on MNO data should be sufficiently differentiated from commercial statistics. Businesses must be reassured that the publication by statistical authorities of Official Statistics based on MNO data will not cannibalise the market demand for mobile analytics offered on commercial terms. This is possible by differentiating the two lines of products along dimensions such as spatial and temporal granularity, timeliness, level of detail and variables, as argued already in Ricciato *et al*. (2018). Therein, the authors proposed to consider the possible analogies with the so-called "freemium" model that has proved successful in several business sectors, whereby making available to the public some "free" version of service or product does not reduce but rather increases the market demand for "premium" versions thereof. Along the same reasoning, the public release by statistical authorities of certain Official Statistics in aggregate form (*e.g.* average number of foreign tourists in mid-to-large towns during the previous quarter, delivered the next month) could potentially increase the appetite for more detailed analytics offered on commercial basis by businesses (*e.g.* the daily number of tourists and the number of nights spent in a particular town, disaggregated by the visitor's country of origin, delivered the next day). In other words, Official Statistics and commercial analytics would serve different purposes and would cover different segments of the information space.

Second, as elaborated by the ESS Task Force on use of Mobile Network Operator data for Official Statistics in their recent position paper (Eurostat 2023), Official Statistics based on MNO data must be based on a standardised and fully open *reference methodological framework*. Once developed and deployed operationally to serve statistical purposes, such methodological framework would then represent a natural standard also for the industry: producers of commercial analytics would have the opportunity to align (partly or fully) their basic definitions and methods to the standard ESS reference, and in this way increase transparency, comparability and credibility of their commercial figures towards their customers. This would not jeopardise their ability to compete among themselves in offering to potential customers on a commercial basis more advanced commercial analytics, *e.g.* based on proprietary improved methods and/or additional definitions. We tend to believe that also in this field, likewise other business sectors, a certain degree of standardisation is not detrimental but rather beneficial to market competition.

Third, within their information segment, Official Statistics produced by statistical authorities must deliver some added value going beyond what would be achievable by the commercial "mobile analytics" providers in the same area. The key added value may be increased accuracy through the integration of data from multiple MNOs and with statistical data. In fact, data from business companies typically refer to their specific customer bases that cannot be considered representative of the general population (as stated already in Baldacci *et al.* 2021). To counteract non-representativeness and bias (*e.g.* in population coverage or geographical coverage) of customer the base seen by each individual MNO, statistical authorities should aim to produce Official Statistics that integrate information sourced from multiple MNOs (an approach called "Multi-MNO orientation" in Eurostat 2023). Furthermore, they may integrate data from (multiple) MNOs with other kinds of non-MNO data, *e.g.* statistical data from ad-hoc sample surveys or censuses, to further improve stability and representativeness of the final figures. Such data integration can and must be done in full compliance with data protection legislation, possibly

leveraging advanced privacy-preserving technologies (Ricciato 2024), with no derogation to the established principle that statistical data cannot be used for non-statistical purposes, and without interfering with business competition dynamics among data providers (level-playing field). The final Official Statistics, produced and released publicly by statistical authorities, may then serve as reference for calibrating the commercial analytics developed independently by MNOs and their partner companies specialised in mobile analytic services (Eurostat 2023).

The perspective of combining data from multiple MNO with statistical data carries important strategic implications. In this scenario, statistical authorities would not be in the position of merely data *consumers* vis-à-vis the private data *providers*, but they would rather position themselves as *partners* of data holders, contributing with statistical data and methodologies to produce Official Statistics that are indirectly beneficial also for the contributing businesses. Moreover, they would further reassert the role of statistical data, in this case as a means to "fertilise" MNO data, and more in general the vast stock of so-called "big data", enabling the production of multi-source statistics that inherit the best of both worlds, namely the timeliness and richness of big data with the reliability and representativeness of statistical data.

5. The role of the European Statistical System

In the vision outlined insofar statistical authorities are required to address a number of important challenges along multiple dimension. The enterprise may overwhelm the resources, capacities and capabilities of any single statistical authority. The good news is that, for all these challenges, the terms of the problem are very similar if not identical for all European countries. Therefore, if a good solution can be found, it is almost certainly a good solution for all countries. We can identify two main reasons for this fortunate condition. First, the business and technological processes that generate PHD tend to be rather uniform across European countries (*e.g.* mobile network technologies do not change from one country to another). Second, contrary to administrative data that feature a certain degree of heterogeneity across different countries due to historical legacies in the development of national public administrations, dealing with PHD does not involve any historical legacy. Therefore, in the development of new methodologies for PHD, statistical authorities can enjoy the luxury of starting from scratch (clean-slate design).

These considerations reinforce the motivation for ESS members to join forces and pool resources to address these common challenges collectively at the European level. For each PHD source, the methodology and quality frameworks can be defined, developed and maintained at the ESS level (and then implemented at national level). For aspects where national peculiarities require some degree of customisation, the necessary flexibility can be incorporated into the design of a common methodological framework. To the extent that the methodologies need to be implemented in software tools, the latter can be developed open-source and maintained at the ESS level (and then used at the national level). If some complex infrastructure is required, it may be designed, built and operated at the European level as a shared ESS infrastructure, and then used on-demand by ESS members.

The coordinated adoption of common methodological standards, shared tools and infrastructures that are developed collaboratively by the ESS members at the European level is not in contradiction with the choice by the individual statistical authorities to implement them (possibly with some customisations) at the national level, and in this way remain in direct control of the production processes.

The benefits of addressing these challenges at the European level are manifolds. First, it obviously prevents duplication of costs and efforts. Second, the *ex ante* adoption of common definitions and detailed methods greatly reduces, and possibly eliminates altogether, the need for ex-post reconciliation and assessment of comparability of the final figures. Third, when personal data are involved, the adoption of a common European methodological framework opens the possibility of defining the necessary data protection measures, and specifically the supplementary technical and organisational measures required by GDPR Art. 89, directly at the European level, through a dialogue with the European Data Protection Supervisor (EDPS) and European Data Protection Board (EDPB). Furthermore, adopting a common European approach (as opposite to heterogeneous national approaches) to the quality challenges posed by PHD may trigger positive reinforcement mechanisms (network effects). The following analogy illustrates the point: when some open-source tool becomes popular and gets used and developed by many entities, it will more likely attract further users and developers, in a cycle of positive reinforcement that will eventually *reduce the cost* and at the same time *improve the quality* of the software. Analogously, the adoption of common and open methodologies, tools and shared infrastructures by the ESS members could trigger reinforcement mechanisms vis-à-vis other actors, including other public entities and business companies, resulting in greater adoption and promotion of the same methodologies, tools and infrastructures, with a positive return for the ESS.

References

Baldacci, E., F. Ricciato, and A. Wirthmann. 2021. "A Reflection on The Re(Use) of New Data Sources for Official Statistics". *Indice. Revista de Estadística y Sociedad*, N. 83. http://www.revistaindice.com/numero83/p8.pdf.

Eurostat, ESS Task Force on the use of Mobile Network Operator (MNO) data for Official Statistics. 2023. "Reusing mobile network operator data for official statistics: the case for a common methodological framework for the European Statistical System - 2023 edition". *Statistical Reports*. Luxembourg: Publications Office of the European Union. https://ec.europa.eu/eurostat/web/products-statistical-reports/w/ks-ft-23-001.

Eurostat, Expert Group on facilitating the use of new data sources for official statistics. 2022. "Empowering society by reusing privately-held data for official statistics - A European approach". *Statistical Reports*. Luxembourg: Publications Office of the European Union. https://ec.europa.eu/eurostat/web/products-statistical-reports/-/ksft-22-004.

Ricciato, F. 2024. "Steps Toward a Shared Infrastructure for Multi-Party Secure Private Computing in Official Statistics". *Journal of Official Statistics*, Volume 40, N. 1. https://doi.org/10.1177/0282423X241235259.

Ricciato, F. 2022. "A reflection on methodological sensitivity, quality and transparency in the processing of new "big" data sources". In *European Conference on Quality in Official Statistics* (*Q2022*). Vilnius, Lithuania, 8-10 June 2022. https://zenodo.org/records/10246419.

Ricciato, F., F. De Meersman, A. Wirthmann, G. Seynaeve, and M. Skaliotis. 2018. "Processing of Mobile Network Operator data for Official Statistics: the case for public-private partnerships". In *104th DGINS Conference*. Bucharest, Romania, 10-11 October 2018. https://zenodo.org/records/10246468.

Introduction to Session 3 invited talks

Li-Chun Zhang, Natalie Shlomo[1]

Abstract

Total survey error has long been established as a quality framework for survey sampling in Official Statistics. Total error frameworks have also been developed for multisource statistics based on integrating administrative registers or other non-survey data. New data sources, however, keep raising source-specific issues and challenges that require further elaboration, enhancement or development of either the quality framework or the associated processes. In particular, the papers in this session deal with data originated from Mobile Network Operators, web scraping and supermarket retail transactions, respectively.

Keywords: Mobile Network operator, quality framework, retail transaction, web scraping.

1. Introduction to the papers

Total survey error has long been established as a quality framework for survey sampling in the context of Official Statistics; see *e.g.* Groves *et al.* (2004). Total error frameworks have also been developed for multisource statistics based on integrating administrative registers or other non-survey data (Zhang 2012; Reid *et al.* 2017; Rocci *et al.* 2022). The increasing uptake of various new data sources, however, keeps generating many source-specific issues and challenges that require continued elaboration, enhancement or development of either the quality framework or the associated processes. One might consider this as an overarching theme for the papers presented in this session.

The first project, presented by G. Simeoni, investigates the quality aspects of using Mobile Network Operator (MNO) data for Official Statistics. Despite the obvious potentials, there are currently perhaps fewer than a handful Official Statistics that would not have existed without the MNO data. One central obstacle is the lack of micro data access due to confidentiality restrictions and business interests. Another issue is the technical nature and large amount of the event and network data in their raw state, for which the statistical agencies lack knowledge as well as capacity (had access to such data been possible).

It is thus natural that the authors distinguish whether access to raw data is granted or not when considering the quality of input MNO data. Built further from previous works in the literature, the main components of a relevant quality framework are identified, and quality aspects with respect to both data and metadata are studied. Novelties can be expected concerning pre-processed MNO data, in particular, how to measure, maintain or improve the quality of pre-processing that is complex but entirely in the hands of the MNOs.

The second paper, presented by M. Six and A. Kowarik, deals with web scraping. Online prices provide a typical example from the "early days" of web scraping at statistical agencies, *e.g.* pertaining to air flight, hotel accommodation or groceries. Job vacancy and enterprise

1 Members of the Istat Advisory Committee on Statistical Methods: Li-Chun Zhang (L.Zhang@soton.ac.uk), University of Southampton, United Kingdom, and Statistisk Sentralbyrå, Norway; Natalie Shlomo (Natalie.Shlomo@manchester.ac.uk), University of Manchester, United Kingdom, and President of the International Association of Survey Statisticians (IASS).

characteristics have received attention in the ESSnet Big Data project. Although the operations may still lag behind in terms of scale, compared to those outside Official Statistics such as The Billion Prices Project[2], the approach to web scraping is maturing in various other respects; see *e.g.* Daas and van der Deof (2021).

Here Six and Kowarik highlight their relevant experiences. In particular, a structured approach is presented and elaborated for "landscaping", which can be understood as cataloguing and measurement of all web-based data sources relevant for the topic of interest, analogous to address canvassing in advance of a population census. The property of representativeness and completeness are considered and discussed in greater details.

The so-called scanner data arising from retail transactions are perhaps the "oldest" big data adopted for Official Statistics. For example, the sub Consumer Price Index (CPI) of food and non-alcoholic beverages has been based exclusively on scanner data since 2005 in Norway, which are delivered weekly and organised according to the unique item identifier (*i.e.* barcode) variously known as GTIN, EAN or UPC. Scanner data is also an important source for National Account and other statistics; see *e.g.* Zhang (2021) on combining retail and payment transactions for the Consumer Expenditure Survey and CPI.

In the third paper of this session, Dawson and O'Brien present the work with scanner data for the purpose of CPI, which was accelerated in Ireland due to the disruptions to traditional price surveying caused by the COVID-19 pandemic. Special attention is given to the decision around whether to use the new data source and the practical developments undertaken to mitigate the risks associated with their inclusion in production.

2. Discussion points following the presentations

The discussant thanked the authors of the three papers in Session 3 as they all included sound methodology, excellent examples and case studies under the agenda of incorporating new forms of data into the production pipeline for Official Statistics. This common objective would benefit from more interactions and cross-collaborations between NSIs, for example through the ESSnet programme.

2.1 Quality aspects using Mobile Network Operators

The first paper focusses on Mobile Network Operators (MNO) data for the production of Official Statistics. The authors aim is to develop a quality assessment of MNO data. The quality assessment needs to include different stages of the preparation of MNO data: (*i.*) the pre-processing that is carried out within the MNO, and (*ii.*) the processing carried out by the NSI to transform the data into Official Statistics. The proposed quality framework is in-line with the standard European Statistical System (ESS) Quality Framework, but is extended to include quality dimensions at both the input level and the ingested NSI level, with a focus on hyper-dimensions as developed for the quality framework of administrative data (Daas *et al.* 2011). One quality dimension that needs to be included in the framework is the transparency of the process for transforming this data into Official Statistics.

2 https://thebillionpricesproject.com.

The quality framework is largely conceptual and includes qualitative descriptions of quality at different levels of processing the data, but it would be beneficial to develop quantitative metrics such that a dashboard of metrics can be produced. In addition, a steady streaming of MNO data from the private Mobile Phone Operators is needed if they are to be used in the production of Official Statistics. Therefore, the quality framework needs to include longitudinal checks over time that can flag any shocks to the data streaming, such as duplication and missing data. One approach for conceptualising the qualitative descriptions in each of the proposed quality dimensions is to turn them into Likert-type scales. This can then be followed by a Principle Component Analysis (PCA) to obtain an overall quantitative score for the product.

Another concern when relying on MNO data for the production of Official Statistics is that they are private companies and can suddenly stop sending their data. Therefore, legal aspects such as contracts and legislation are required prior to large-scale dependence on this data.

2.2 Landscaping web data

The second paper presents ongoing work in the ESSNet Web Intelligence Network (WIN) project and focusses on different adaptations of web scraping depending on the project objectives. The approaches listed in the paper are:

1. Gathering information through web scraping for an existing frame of websites: a sample of URLs can be selected randomly within the frame for web scraping. This approach is demonstrated in the test case presented in the paper for web scraping online job advertisements according to a frame provided by Eurostat.
2. Identifying websites that belong to a target population for web scraping: This approach will likely suffer from selection bias, particularly as the sample of websites is largely determined by quota or cut-off sampling. This approach is demonstrated in the test case presented in the paper for selecting websites according to a checklist of characteristics. Those websites that do not have the necessary meta-information or have a 'captcha' requirement are automatically rejected from the web scraping.

The paper did not set out a standard quality framework as the other two papers in the session and this will eventually have to be carried out prior to developing the use of web-scraped data in Official Statistics. The authors make good use of Machine Learning methods to facilitate the selection of URLs for web scraping as well as developing sound quantitative scores to judge the quality of the websites.

2.3 Quality of transaction data for use in the Consumer Price Index

The third paper details the work of assessing the quality and fitness-for-use of scanner data (digital transaction data) for the Consumer Price Index at the Ireland Central Statistics Office. This would replace the manual gathering of prices within selected stores. There have been rapid experiences and case studies using scanner data across many NSIs, particularly due to the impact of the pandemic, and therefore more cross-collaborations would benefit this work.

The proposed quality assessment of the scanner data are in-line with the standard ESS quality framework, but again were largely qualitative descriptors. See section 2.1 on how a quantitative score card can be developed from a qualitative assessment of quality. Similar to MNO data, there needs to be a steady stream of scanner data ingested into the NSI and therefore longitudinal

checks over time are essential for detecting shocks (duplicated data or missing data). There needs to be a clear distinction between one-off quality checks and ongoing longitudinal quality checks that can lead to a time-series of quality metrics for detecting shocks.

Since the scanner data is reliant on private stores, there needs to be risk mitigation in place if the stream of scanner data suddenly stops, for example reverting back to a manual price collection if necessary. The authors also state that there are "no issues in accuracy" given that all scanner data is ingested, but the data still needs to be quality assessed for accuracy given the occurrence of duplicated or systematic (informative) missing data. The authors present a test where both manual and scanner data are collected at the same time and it would be useful to report on the quality issues arising from this work.

References

Daas, P., and S. van der Deof. 2021 "Using Website texts to detect Innovative Companies". *Working paper,* N. 01-21. Heerlen, the Netherlands: Center of Big Data Statistics - CBS.

Daas, P., S. Ossen, M. Tennekes, L.-C. Zhang, C. Hendriks, K. Foldal Haugen, F. Cerroni, G. Di Bella, T. Laitila, A. Wallgren, and B. Wallgren. 2011. "Reports on methods preferred for the quality indicators of administrative data sources". *Deliverable 4.2 BLUE-ETS Project SSH-CT-2010-244767*.

Groves, R.M., F.J. Jr. Fowler, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourrangeau. 2004. *Survey Methodology*. New York. NY, U.S.: John Wiley & Sons.

Reid, G., F. Zabala, and A. Holmberg. 2017. "Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ". *Journal of Official Statistics*, Volume 33, N. 2: 477-511.

Rocci, F., R. Varriale, and O. Luzi. 2022. "Total Process Error: An Approach for Assessing and Monitoring the Quality of Multisource Processes". *Journal of Official Statistics*, Volume 38, N. 2: 533-556.

Zhang, L.-C. 2021. "Proxy Expenditure Weights for Consumer Price Index: Audit Sampling Inference for Big-Data Statistics". *Journal of the Royal Statistical Society Series A: Statistics in Society*, Volume 184, N. 2: 571-588.

Zhang, L.-C. 2012. "Topics of statistical theory for register-based statistics and data integration". *Statistica Neerlandica*, Volume 66, N. 1: 41-63.

Quality aspects using Mobile Network Operators data for Official Statistics

Gabriele Ascari, Erika Cerasti, Cristina Faricelli, Paolo Mattera, Sara Piombo, Roberta Radini, Giorgia Simeoni and Tiziana Tuoto[1]

Abstract

The potentialities of data from Mobile Network Operator (MNO data) for the production of Official Statistics are well-recognised. MNO data can support traditional statistical production as well as cover new topics. Several experiences have been carried out on specific use-cases, but the need of a further step towards standardisation is emerging as well as the need to explore and systematise quality issues in the use of this data. In this paper we approach the definition of a quality framework for Official Statistics based on MNO data. We identify the main components of the quality framework, relying also on previous existing work, and we develop the quality aspects related to the institutional environment, input data and first reflections on throughput quality.

Keywords: Quality framework, MNO data, input quality, trusted smart statistics.

1. Introduction

The exploitation of data generated by the mobile devices for the production of Official Statistics has received an increasing interest in the last decade. Several experiments have been carried out in various areas, both in low and high income countries, to study the density of present population at different daytime, for mobility, tourism, migration, and disaster displacements. Most recently, a significant boost came from the situation generated by the COVID-19 pandemic, during the acute phase of which the possibility of conducting traditional statistical surveys was severely limited if not possible at all, resulting in the search for alternative, less expensive, more timely data sources able to provide "good" proxies on the topics of interest for Official Statistics (Santamaria *et al.* 2020).

Moving beyond explorative activities, research projects and one-off case studies, experimental statistics based on Mobile Network Operator (MNO) data require several enabling conditions: the establishment of sustainable models of data access; the definition of adequate technical and organisational measures to ensure protection of personal data and business sensitive information; the development of methodological aspects to ensure that the statistical figures produced using MNO data comply with the principles of Official Statistics. In this paper we aim to explore these requirements and the conditions under which they can be met by statistics derived from MNO data; in other words, the degree to which such data can be considered fit-for-purpose of National Statistical Offices. The most comprehensive way to investigate this

1 Gabriele Ascari (gabascari@istat.it), Erika Cerasti (erika.cerasti@istat.it), Cristina Faricelli (cristina.faricelli@istat.it), Paolo Mattera (paolo.mattera@istat.it), Sara Piombo (sara.piombo@istat.it), Roberta Radini (radini@istat.it), Giorgia Simeoni (simeoni@istat.it), Tiziana Tuoto (tuoto@istat.it), Italian National Institute of Statistics - Istat.
The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics - Istat.
Though the article is the result of a joint work, paragraphs are attributed as follows: paragraph 1 to Tiziana Tuoto, 2 to Cristina Faricelli e Paolo Mattera, 3 to Gabriele Ascari, 4.1 to Giorgia Simeoni, 4.2 to Tiziana Tuoto and Gabriele Ascari, 5 to Giorgia Simeoni, 5.1 to Giorgia Simeoni and Sara Piombo, 5.2 and 5.2.1 to Gabriele Ascari and Sara Piombo, 5.2.2 to Giorgia Simeoni, 5.2.3 to Tiziana Tuoto, Cristina Faricelli and Paolo Mattera, 5.3 to Erika Cerasti and Roberta Radini, 6 to Giorgia Simeoni.

question is to develop a quality framework for using MNO data in Official Statistics. This task requires the analysis of several different aspects, from the identification of specific quality requirements at Institutional level to the set up of a dedicated quality layer in the new production process. This quality layer has peculiar elements of input and output quality compared to the more traditional the ones based on surveys and administrative. Nevertheless, we believe the investigation of such a quality framework is now a piece of work in a wider set of international and European actions devoted and convergent to the full exploitation of MNO data in Official Statistics. The high-level requirements for the definition of a methodological standard, open and transparent, where the methodological choices are fully supported by quality assessment is one of the key messages of the recent position paper prepared by the ESS Task Force on the use of Mobile Network Operator data for Official Statistics (Eurostat 2023). The spirit is also at the basis of the EU funded projects "Development, Implementation and Demonstration of a Reference Processing Pipeline for the Future Production of Official Statistics Based on Multiple Mobile Network Operator Data" (whose acronym is TSS Multi-MNO) to which Istat is currently actively participating and "Trusted Smart Statistics: methodological developments based on new data sources" (with the acronym TSS-METH-TOOLS), coordinated by Istat and participated by 9 European National Statistical Institutes (NSIs).

The paper is organised as follows: Section 2 is devoted to describe what is intended by MNO data and what are their specificities; Section 3 after it highlights the differences between the traditional statistical processes and the ones based on MNO data. Then, before introducing our proposal, we take stock of the already existing works by considering, on the one hand, consolidated quality frameworks for more traditional sources; on the other, the results of recent international projects on the quality of MNO data. Section 5 introduces the overview and the components of the quality framework we are proposing. Subsection 5.1 develops the Institutional level aspects of the framework, while section 5.2 analyses input quality dimensions; finally Section 5.3 introduces our first considerations related to process quality. Concluding remarks reported the ideas that we would like to develop for completing the process quality and analysing output quality issues.

2. What do we mean by MNO data?

Before analysing the quality of the statistical output based on MNO data, or even the quality of the production process that transform MNO data into statistics, we have to clarify what we intend as MNO data. As in Eurostat (2023), we use the term "MNO data" to refer generically to all location data collected on the side of the network, hence, we mean a set of different types of data produced by MNOs and here briefly described:

- *Event data*: generated by the mobile devices directly due to their activities: calling, receiving a call, sending and receiving text messages (call detail records or CDR), connecting to the internet (data detail records or DDR) connecting to the telco network (signalling data). Each event data corresponds to a particular device which contains a determined Subscriber Identity Module (SIM). These data are characterised by high temporal and spatial coverage, in particular signalling data, which are not voluntary activities of the device user, but they are an information exchange concerning the establishment and control of the communication and the management of the network. Signalling data are, if available, the most promising data for our purposes since we are here interested mainly in positioning devices at a certain time (how many, where,

when). Other aspects, such as networking among devices, technological preferences and attitudes, are neglected. For the same reason we neglect machine-to-machine activities.

- *Network data*: technical data referring to the kind of technologies, the technicalities and the state of the antennas and network. They also contain the position of the antennas. The network is very dynamic, a frequent update of its configuration is generally needed. This kind of data is at the base of all algorithms used to retrieve the position of the device.

Both Event and Network data share similar characteristics that differentiate them from other data. They are actually "big" in terms of volume; we can register hundreds of events per device per day, with implications, *e.g.* in the IT infrastructure necessary to manage them. In addition, these data, remain personal data even after pseudonymisation, discouraging their transmission to preserve confidentiality. Other types of data, particularly those generated on the side of the mobile device, *e.g.* GPS data points collected by apps or by the operating system and then delivered to platform operators, sometimes called Mobile Phone Data (MPD), as well as Business data, inclusive of information on customer contracts, are currently not considered in this paper.

3. Peculiarities in producing Official Statistics based on MNO data

Event data and network data can be used together with other auxiliary data as input of a complex process to produce statistics of interest.

There are many differences between a traditional statistical process and a process that involves the use of MNO data. One of those differences concerns the data holder of information. MNO data fall in the dimension of privately-held data, that is data that are not collected by the NSI itself but are gathered by a private, third-party organisation. In this regard, MNO data differ from other sources that are commonly used by NSIs (*e.g.* administrative data) not only for their original not statistical purpose but also because the data holders are entities that do not belong to the national (or international) statistical system. Of course, protocols for the exchange of information between NSIs and private organisations are already in place for many scenarios. The general principles and good practices still apply, such as: continuous feedback on the provided data, pursuit of agreements so that the flow of data is not at risk of interruption, metadata completeness.

However, general principles aside, in the case of MNO data the situation appears more complex than the acquisition of administrative data. This complexity arises from two peculiarities of these data as mentioned in the previous section: due to their "Big data" and confidential nature, the raw data processing usually goes through off-premise, outside of the direct control of the NSIs. A NSI cannot access microdata but only obtain aggregated data that have already been processed by the MNOs.

The off-premise custody of the data may be seen under both a negative and a positive light by the NSIs: on one hand, statistical institutes may know little about or, worse, completely ignore the processing procedures that took place on the data; on the other side, storage and processing of MNO data require a high-performance, advanced technological infrastructure that very few NSIs, if any, actually own or would be able to afford. Even if the process takes place in the MNOs premises, Statistical Authorities pursue an active collaboration with data providers in order to obtain clear documentation of the processing carried out by MNO (transparency), and that the latter is carried out, as much as possible, according to Statistical Authorities indications. Another specific aspect of this type of process that can impact the quality of the output is the sequence of processing steps and the algorithms that model them and the version of the software

that implements them. These three distinct elements must be appropriately documented or, where possible, made available in open mode so that it is possible to evaluate their quality.

4. Existing experiences on quality frameworks

4.1 Consolidated quality frameworks in Official Statistics

As well known, the European Statistical System (ESS) can rely on a consolidated quality framework, the ESS Common Quality Framework, primarily based on the principles and indicators the European Statistics Code of Practice (ES CoP). The ES CoP is a self-regulatory instrument adopted by the ESS National Statistical Authorities and Eurostat that is based on 16 principles organised in 3 areas: the institutional environment, statistical processes and statistical outputs. For each principle there are indicators that represent best practices to implement the principle. For the institutional environment the principles are institutional level fundamental requirements that should be respected to make the statistical authority able to produce reliable statistics, such as professional independence, mandate for data collection, impartiality and objectivity; in the statistical output area, principles coincide with traditional output quality criteria, like relevance, accuracy, timeliness and so on. For the statistical process area general principles for the conduction of statistical processes are given, like the use of sound methodology.

This high-level framework is then complemented by specific frameworks that define and describe the sources of errors, arising in different statistical process types that can have an impact on the quality of the output (and in particular, but not only, on Accuracy). As an example, the Total Survey Error model, well described in Groves *et al*. (2009), identifies the common error sources that can affect traditional sampling survey, such as sampling, coverage, non-response, measurement and processing errors. For each type of error is clear which type of impact they could have on estimates in terms of bias or variability.

Then with the increasing use of administrative data in Official Statistics another model was developed. Administrative data are created outside the control of the statistical authority with a not-statistical purpose, thus, the introduction of an evaluation of the input quality was deemed necessary. A model to represent dimensions of the input quality was developed (Daas *et al.* 2009) and the sources of errors identified in the Total Survey Error model were extended/ adapted for the statistics based on administrative data (Zhang 2012; Reid *et al.* 2017). Very often, the use of such data involves the integration of different data sources to reach the needed coverage of the population or to collect all the necessary variables. Thus, a specific quality framework for multisource statistics was developed (ESSnet KOMUSO, 2019).

Most of these frameworks are adopted at Istat as a theoretical basis for quality assessment activities; even if sometimes the models are adapted to better represent the characteristics of Istat's production, especially when considering the internal System of Statistical Registers or the evaluation of input administrative data sources.

4.2 Existing proposals for assessing quality of new data sources

The next challenge to be faced is the definition of a quality framework for statistics based on new data sources (*e.g.* big data or trusted smart statistics) and Istat has recently set up a Task Force with this aim. A high-level proposal in this respect has been developed by Amaya *et al.*

(2020): they tried to extend Total Survey Error to big data. As we will see, some research has been carried out in this field under the umbrella of some ESSnet projects.

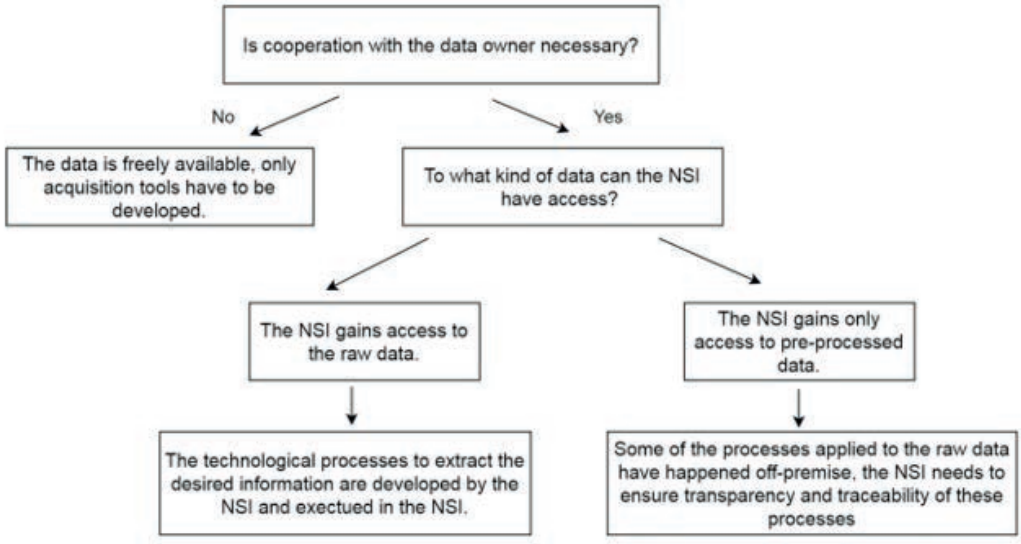
One of the most structured results regarding quality aspects in accessing, processing, and using new data sources for official statistical purposes was developed as part of the ESSnet Big Data II project. Guidelines were formulated based on quality related experiences within the project on diverse new data sources and data-specific processes. However, in this paper we are not considering all the possible types of big data but we are only focussing on MNO data. The latter are one of the data sources considered in the guidelines, as a particular case of privately-held data.

The structure of the guidelines follows a production process logic, with a focus on the phases and processes that are affected by new data sources. The most obvious change happens in the input phase, where the acquisition and the recording of the data can look completely different than in the case of survey data or administrative data, especially in case of privately owned or held data. The throughput phase receives great attention, formally introducing a distinction between a lower processing level, where potentially unstructured raw data is processed into well-structured intermediate ("statistical") data, and an upper layer in which the statistical data is used to produce statistical output. In this way the guidelines re-elaborate and systematise - in a quality context - some considerations originally proposed by Eurostat (2019) for methodological, technical and governance aspects and further developed in Ricciato *et al.* (2019). On the contrary, the output phase analysis in terms of quality was quite limited, since it was considered that generally the usage of new data sources does not alter the typical processes of the output phase like dissemination and evaluation. In the same reasoning, some well-known and traditional quality dimensions - mainly output related - like relevance, reliability, timeliness and punctuality were barely discussed and analysed, under the assumption that statistical output has to be relevant, reliable and be published on time, regardless of its sources. For other traditional quality dimensions like comparability and coherence, additional aspects become relevant: is a data source comparable (in the sense of stable) over time? Is a statistical output for which new data sources are used coherent with a statistical output produced on the basis of traditional data sources?

The guidelines devoted specific attention to the acquisition and recording of privately owned and held data, and in particular to the need of a negotiation for accessing (part of) the potentially pre-processed data. In the case of MNO data, the guidelines also discuss the complexity related to distinguish between the input phase and the throughput phase.

Indeed, as already mentioned, with MNO data often the NSI only gains access to data pre-processed by the data source, and processes normally happening at the premise of the NSI ("on-premise") take place at the data source. There, typical processes include selection of units and variables, some form of aggregation, but also some form of validation can happen at the premise of the data source ("off-premise"). Depending on the differentiation between on-premise and off-premise, the NSI has different insights in the processes applied and thus, different guidelines become relevant. This whole topic of pushing computation out, with technical processes happening on-premise of private data-owners/holders, was recognised as completely new and not covered by any of the traditional quality categories focussing on output quality. ESSnet Big Data II Project also proposes a template for reporting quality and metadata, taken from the widely known SIMS (Single Integrated Metadata Structure). The definitions and guidelines of the template are based on the updated version of the ESS Handbook for Quality and Metadata Reports (Eurostat 2021). Sub concepts of the SIMS considered as not relevant, were deleted. When the existing sub concepts did not cover all relevant quality aspects for new data sources, new sub concepts were introduced.

Figure 4.1- Illustration of the decision tree to classify different scenarios of data access to a new data source



Source: ESSnet Big Data II 2020*a* Deliverable K3: Revised Version of the Quality Guidelines for the Acquisition and Usage of Big Data

Here, one comes across the problem that the SIMS is generally output-oriented, this means it is a standard to report the quality of Official Statistics. When it comes to new data sources, the output so far is almost never an Official Statistics, sometimes it is not even publishable. At the time of the ESSnet Big Data II, the "output" has often more the form of a throughput data set, which could further be used and processed. To avoid a problem with wording, the term "statistical output" in this analysis does not mean that it has to be a publishable statistical product. The changes made by the ESSnet Big Data II experts were highlighted in colour compared to the original template. This quality template was tested by the ESSnet Big Data II and, the filled-in quality templates for the MNO data are available as examples.

Meanwhile, efforts for a better understanding of the potential uses of big data in Official Statistics were not limited to the European Statistical System: indeed, different task teams under the coordination of the United Nations explored the use of mobile phone data (MPD) for statistical purposes. Methodological guides for specific statistical domains (tourism, migration, information society, population mapping and disaster statistics) were published in 2022. Most of them address the issues behind quality assurance, some in more detail than the others.

As an example, it may be interesting to look at the guidelines on tourism and their proposal for quality assurance: a systematic approach is recommended, since "all aspects of MPD should be examined and evaluated with certain principles and standards" (UNSD 2022). The main feature of the proposed approach consists in the identification of quality gates at the most relevant parts of the process, namely an input, a throughput and an output quality gate. Of course, this is the same segmentation of the statistical process that intervenes in the aforementioned ESSnet guidelines. The quality gates are a general approach in quality assurance systems. Each gate operates as a checkpoint for the data to be allowed to go through the next steps of the process. Indeed, the guidelines show a list of quality indicators and the potential errors that can be encountered at the different gates.

It also should be kept in mind that the guidelines just described are related to a specific domain, tourism statistics, while the quality considerations presented in this paper are meant to be general enough to capture multiple domains with the proper adaptations. At the same time,

since the nature of the data is the same, the indicators used in the guidelines for tourism statistics could be extended to a more general framework. Indeed, a lot of the indicators among the quality checks for raw data could be used for multiple domains other than tourism: for example, the number of daily records or subscribers or the cell occupancy rate, deemed as "critical" in the checks of the input quality gates, are measures that could (and should) be monitored in other domains and applications related to MNO data. Lastly, the output quality gate is described in less detail than in other phases, as output quality issues for MNO data processes do not differ much from those of processes deriving from traditional sources, as already affirmed in the ESSnet guidelines.

5. Overview of the proposed quality framework

Taking into account the consolidated quality frameworks for traditional data sources, the results already achieved by previous international projects and the peculiarities of MNO data, we propose a structured quality framework including the identification of specific quality requirements at Institutional level and the set-up of a quality layer that follows the production process.

Quality requirements at Institutional level are based on the analysis of the ES CoP principles of the Institutional Environment area and their application in the case of MNO data.

Then, as suggested by ESSnet Big data II the quality layer to be associated to the statistical process and statistical product is decomposed in:

- input quality, for which the idea is to start from the approach developed for the quality of administrative data, and evaluate its applicability and possible extension or adaptation to the case of MNO data; for sake of clarity we specify that here we refer as input data the original Event and Network data before they are processed by the MNO, and not the intermediate aggregated output that is transmitted to NSIs;
- lower throughput quality, basically coinciding with the processing made by MNO on their premises. Together with the input data analysis, this part of the most peculiar component of the production process and consequently of the quality framework;
- upper throughput quality, the further processing and analysis steps carried out by NSIs on pre-aggregated data, including also integration with auxiliary data and other methods to improve their quality, and application of validation techniques; in this phase should also be considered the integration of data from different MNO and the related challenges.
- output quality, for which the main reference remains the traditional quality criteria of statistical products, eventually enriched to inform the final user of the novelty of the production process.

The development of such quality layer is still in progress, and in this paper we will develop the quality aspects at Institutional level, the input quality issues and some first reflections on the lower throughput quality.

5.1 Quality at Institutional level

Despite the ES CoP principle and indicators are high-level indications, they also need to be periodically revised and updated to accommodate developments in Official Statistics. In fact, the ES CoP has already undergone two revisions, and the latest revision has also started to address the use of privately-held data for the production of European Statistics. Here, we focus

on the principles of the institutional environment area, which may be relevant when using MNO data, in order to understand if the needed quality requirements are already present or it would be useful to adapt or integrate them.

As already mentioned, a specific characteristic in the context of data provided by MNOs, is that NSIs can neither control the raw data generation nor directly monitoring the data processing for quality purposes. These considerations lead to promote the cooperation between Statistical Authorities and MNOs. Principle 1bis of the ES CoP is about coordination and cooperation, and in particular Indicator 1bis.3 promote the cooperation of the ESS with its advisory bodies, as well as with the members of the European System of Central Banks, academic institutions and other international bodies. We think that the principle could be strengthened by considering cooperation and collaboration between Statistical Authorities and MNOs (or more in general with private holder of data), for the provision of data and metadata for their correct interpretation and use, as well as all activities related to the development, production, and dissemination of statistics. In the institutional context, it is crucial to promote cooperation with MNOs by jointly seeking collaboration. This shared effort, especially at the ESS level, is instrumental in establishing data standardisation rules that enhance the comparability of subsequently produced statistics.

Obviously, the first step for the use of MNO data is to have access to them. Principle 2 of the ES CoP is precisely on Mandate for Data Collection and Access to Data. Indicator 2.1 states that the mandate of the statistical authorities to collect and access information should be specified in law. Indicator 2.4 is addressed to facilitate the access to other data, such as privately-held data, for statistical purposes, while ensuring statistical confidentiality and data protection. According to Indicator 2.1, the use of MNO data for Official Statistics should be based on a coherent legislative framework. If access to this type of privately-held data were guaranteed by law, data acquisition would be streamlined, and the quality of input data would also improve, as it would become possible to fix rules regarding data transmission, data formats, and so on. To this aim a modification of the current legislation would be envisaged. Indeed, a modification of European statistical law (EU regulation 223/2009) is currently under discussion, and one of the main changes proposed is related to facilitate the access to privately-held data for the production of European Statistics. In the absence of the law agreements, collaborative protocols or contracts can be tools to facilitate the cooperation and the access to data in the appropriated way. It is advisable that such a contract/agreement is based on a template agreed upon by the NSIs and harmonised with Eurostat coordination to ensure the best collaboration with MNOs, as well as the comparability at ESS level of the statistics produced through MNO data, starting from the data collection.

In accordance with the Principle 3 of the ES CoP, in order to manage MNO data, the NSIs should have the appropriate human and technological resources for the storage and processing of large amounts of data, ensuring their confidentiality and integrity. It is less clear what should be assured to data providers, in the scenario, like the one of MNO data, in which the data provider is also asked to carry out part of data processing. Also the MNO should have adequate technical and human resources to facilitate data provision to NSIs, aligning with specified settings and timeliness while minimising the need for excessive workforce effort and expenses on the part of the MNO itself.

Principle 4 of the ES CoP is about Commitment to quality, and all the indicators are relevant, but it should be considered that the input data generation as well as part of the processing take place outside of the NSI's direct control. Therefore, the NSI does not have direct governance over the tools implemented to ensure the management of their quality. *Ad hoc* methods and tools for quality management should be developed, *e.g.*:

- it will be necessary to develop specific methods for monitoring processes (Indicator 4.2). They could be for example based on the definition of quality indicators or measures that can be calculated by MNOs and used as a basis for implementing specific improvement actions. Of course, this should be carried out by MNOs in a collaborative environment, with the aim of providing NSIs with the necessary information to evaluate the process, even if it is conducted by the MNO externally;
- documentation from the MNO becomes vital for quality assessment (Indicator 4.3). When a third party provides data to an NSI, transparency is a factor to which should be given even more importance in the quality assessment then in the case all the process is under the NSI control. If the NSI cannot access the specifics of the data generation process, it will be difficult to ensure the quality of statistics derived from that data and whether the data can be provided continuously and in a comparable manner over time. Furthermore, the NSI should establish a post-processing evaluation system for the intermediate aggregated data it receives to assess its quality and provide feedback to the MNO for improving future deliveries.

Principle 5 regards Statistical Confidentiality and Data Protection. Statistical confidentiality is guaranteed by law for data providers and it must also be ensured against MNOs. It is advisable to include appropriate safeguard clauses on this topic when establish agreements with the MNOs. The receiving Statistical Authority should commit not to share the data with third parties and ensure that the statistics produced and disseminated using the data provided by the MNO will not disclose sensitive or confidential company information. Conversely, MNOs will supply NSIs with aggregated data that are valuable for statistical production while obeying to the GDPR.

Summarising, relevant aspects for quality at institutional level refer to: improve the cooperation with MNOs, facilitate data access to NSI through legislation or defining harmonised templates for agreements that should include requirements for documentation and transparency and also clauses to assure confidentiality, develop *ad hoc* methods for quality monitoring and assessment.

5.2 Input quality aspects

The production of statistics from private data sources, such as location data provided by MNOs, is also highly dependent on the quality of these sources. It is therefore important to determine the quality of such input data systematically, objectively, and in a standardised way. First of all, we should identify the most relevant aspects, or dimensions, to be taken into account. For this purpose we decided to take inspiration from the quality framework for administrative data sources, originally developed by Daas *et al.* (2009[2]) for CBS (Statistics Netherlands) and subsequently taken up and further developed by the international BLUE-ETS project (Daas *et al.* 2011), composed of three distinct categories or hyper dimensions. The three hyper dimensions - Source, Metadata, and Data - are used to assess the statistical usability of a data source originated from MNOs for other (non statistical) purposes. Each hyper dimension consists of several dimensions, each of which consist of several quality indicators. Below, we provide an adaptation of the dimensions for the Source, Metadata, and Data hyper dimensions in the context of MNO data.

2 The same approach was also further developed by MIAD - Methodologies for an Integrated Use of Administrative Data in the Statistical Process (https://cros-legacy.ec.europa.eu/content/miad-methodologies-integrated-use-administrative-data-statistical-process_en).

5.2.1 Hyperdimension source

1. *Dimension Supplier*

Especially during the experimental phase of using MNO data, it will be essential to have all information about the supplier and the contact reference into the organisation that creates the data source, as there is a higher probability of needing clarification or explanations regarding the construction and interpretation of the data provided. This is also in view of the need for ongoing collaboration that will certainly arise in the initial experimental periods. While the purposes for which MNOs collect and store the data are well known, information on the differences between the purposes of different data managed (CDR, signalling, GPS...) could be useful to be considered.

2. *Dimension Relevance*

Given the complexity of the procedures to collect, store and process MNO data, the usefulness of such effort could be questioned; in other words, can these data be so relevant for Official Statistics that the additional effort in terms of costs, expertise, and infrastructure is justified? Incidentally, this is also one of the questions that arise following the section "Source" of the CBS checklist on administrative data sources. The answer, of course, is "yes", with some caveats. Indeed, although we are still in the experimental stages of implementing MNO data processes into statistical production, it is already possible to foresee some uses of this new source with the help of the checklist. Arguably, at the moment, is difficult to imagine the use of MNO data to replace an existing statistical product entirely. However, one more realistic use for MNO data is to supplement or check current sources or products, meaning that statistics produced from MNO data could be used to validate or enhance an existing data product. For example, traditional estimates of the present population could be compared to the ones produced with MNO data, which can also be useful to explore the same theme in more detail, computing for example the present population at specific times of the day. Given this, it could be expected that, in some cases, the burden on the respondents could be mitigated thanks to the use of MNO data. Again, this is an advantage that will probably not materialise immediately, but during the course of years where MNO statistics will be processed, tested and put into production alongside traditional ones. In this perspective, these data could be also used to produce statistics on topics still not covered by official statistical production, and answer to information needs that are still unmet.

3. *Dimension Privacy and security*

Applying this dimension to original input data is related to the data management internal to the MNO, that should obviously assure the respect of legislation and it is one of the main reasons for which microdata cannot be provided directly to NSI. From the point of view of the NSI the dimension could be applied to the provision of data to the NSI. As already pointed out in Section 5.1, in the absence of a legislative framework to rule data provision from MNOs, preparing a detailed contract becomes crucial. This contract should not only outline the contents, structure, and format of the data (see next dimension) but also establish clear agreements on the privacy protection in accordance with the GDPR. It is essential to define secure methods for data transmission within the contract. These agreements should be harmonised at the European level to ensure consistency and comparability of statistics starting from the input phase within the European Statistical System. European standardisation of the contract will not only promote consistency in produced statistics but also need to specify the minimum level of confidentiality that datasets must guarantee, balancing it with the content requirements for NSIs. To ensure an adequate level of information in the data, aiming to meet the need for producing detailed and

sufficiently high-quality statistics (with a low level of uncertainty of the estimations), it will be necessary to find an appropriate level of aggregation for the microdata. The level of aggregation must be balanced: not excessively high to ensure data sufficient informative and not overly detailed to preserve data confidentiality. This balance should be achieved with the support and collaboration of privacy authorities, along with a commitment to confidentiality from the NSI receiving the data.

4. *Dimension Delivery*

Input data are not delivered to NSIs, thus this dimension is not straightforward applicable to the MNO case, but it is very relevant for the aggregated data that are provided to NSIs. Specific formal agreements should be developed. Such agreements should still cover the usual aspects of frequency, punctuality, format of the delivery and so on. Frequency of deliveries can be expected to be greater than the case of administrative data sources; also, a certain degree of automation will be probably involved and this can facilitate the respect of the expected punctuality.

It should be also remembered that NSIs will deal with multiple MNOs. In order to avoid difficulties, the delivery should be standardised across them and probably adopt a centralised platform. Data formats should be as far as possible harmonised for every delivery and every provider. To obtain this, strict collaboration and specific guidelines for the involved MNOs are required.

Finally, among the indicators proposed in the checklist, the costs of the data source are listed. Costs will be estimated more accurately once MNO data processes will be carried out regularly, but it is important to include also indirect costs in the assessment. In particular, the costs of training the staff or employing new experts, the costs related to new IT infrastructures and so on.

5. *Dimension Procedures*

Transparency is fundamental, and the MNO should provide all relevant information to the NSI about the data generation process and communicate any changes in data content promptly. The NSI needs to ensure the MNO's collaboration in terms of timeliness and consistency in data supplies. In the event of any delays, the assurance that such delays will be communicated to the NSI as soon as possible is essential. However, the NSI cannot substitute the supply with other providers.

5.2.2 Hyperdimension metadata

1. *Dimension Clarity*

The clarity dimension, similarly to the clarity quality criteria of the statistical output, is related to the availability of information useful to correctly interpret the data. In this broad sense it is absolutely applicable to MNO data, for which the information provided by the MNO are necessary. In particular, structural metadata (units, variables, classifications, etc.) that describe the input data (as well as the aggregated data provided) should be clearly described. MNO data essentially deal with events that are referred to mobile devices, and this implies that to derive Official Statistics on individuals some transformations according to determined hypotheses should be applied. Some combination of characteristics of the event and of the device can support the identification of specific individual subpopulations (tourists, commuters...). Also, the definition of all the variables, together with their format and set of admissible values, are necessary metadata. Since the stream of MNO data is potentially continuous, it is very important to clarify the time interval to which data refer. But the time reference in this kind of data is more than a metadata: it is part of the data and should be considered in the Data dimension. An

additional relevant metadata is related to the space dimension. The geographical limits to which the data refer are necessary parameter to use and interpret the data.

Whatever variation in the metadata should be timely communicated to NSI.

2. *Dimension Comparability at the metadata level*

This is a crucial dimension, as these data are not only generated with a completely different purpose from Official Statistics, but the definitions of the main concepts are originally very far from the Official Statistics definitions and the comparability between metadata related to MNO data and Official Statistics requirements could be minimal if the original MNO data are considered. But the comparability of metadata between original MNO data and Official Statistics is not something that should be pursued. The input MNO data will be transformed according to specific hypotheses to answer to Official Statistics information needs.

3. *Dimension Unique keys*

The integrability of the data is a main issue as well the possibility to follow a device over time, so unique keys should be present at microdata level.

4. *Dimension Data treatment* (*by data source keeper*)

Since part of the processing of MNO data is made directly by MNO, this dimension assumes particular relevance, even if it could, at least partially overlap with the first part of throughput quality. Agreements with MNOs should assure a high level of transparency of the treatment that they made on data before providing them to NSI. If MNO has already prepared documentation on their data processing, *e.g.* for ISO 9001 certification, such documentation can be also useful for the NSI. A further envisaged improvement is that NSI provide clear indications to the MNO on how to manage the data as well as requiring quality measures related to the application of such methods to the data.

5.2.3 Hyperdimension data

Quality checks on raw events and network data are performed by MNOs differently from each other. In general, they include initial checks on syntactic correctness (erroneous formed strings) and completeness (analysis of missing data). In addition, MNOs perform technical and instrumental checks. Many of those controls are not interesting from a statistical perspective, while other controls can have a significant impact on the quality of the data. It is important to identify those relevant for the process.

1. *Dimension Technical checks*

This dimension, in the case of administrative data, refer to the first checks of technical usability of the file and data in the file applied by the NSI to the data received. In the context of the MNO data, these checks are carried out directly by the MNO, so the perspective is a bit different, even if the checks are similar. In the best case, the NSI could have provided indication to the MNO on the checks to apply, *e.g.*:

- controls on file corruption should be implemented at first, since data files must be readable;
- checks on missing data are also implemented, possibly marking missing variables with a flag dividing them between, for example, “necessary”, “important” and “nice to have”;
- other technical controls include syntactic checks at various levels, format and range validity checks for many variables (*e.g.* date, time, time zones, geographical coordinates)

and checks on the correctness of the geographical reference system and its variables (*e.g.* datum, projection parameters) and data duplication controls;

- since NSIs are interested in studying human being behaviour, several checks are performed at this stage, to exclude events generated by machine-to-machine connection activities.

2. *Dimension Accuracy*

Another critical quality dimension is the accuracy of the data: correctness, reliability and certification of the information reduce the risk of errors. Incorrect or unreliable input data are potentially invalidating for the entire elaboration process, since the error in the original data propagates during the following elaborations, generally resulting in an amplification of the absolute error.

In the case of MNO data, an instance in which low accuracy can lead to a substantial error is the device positioning estimated from the geographical position of the network cells it is connected to. In this situation, the position of a network cell could be incorrect, leading to an erroneous estimation of the position of the device. Indeed, the geolocation of the device with the desired precision is not always possible, many kinds of errors and technical issues may occur, causing a wrong estimation of the device position. Erroneous network data may produce, for example, jumping of hundreds of kilometres in a few seconds. However, the largest contribution to positioning error is given by the methods implemented by MNOs to estimate the so called "cell coverage". Even if this is part of the process quality component, it is worth mentioning that checks and indicators must be developed to control those phenomena and above all the quality of the methods used by different MNOs.

3. *Dimension Completeness*

The concept of completeness can have different interpretations for MNO data. First of all, the mobile network coverage is not uniform on a given territory, presenting a complete and redundant service on urban areas, and several zones with scarce or absent service, mainly in rural areas. Each mobile network cell includes many antennas, covering the area around the cell with different characteristics and angles. Additionally, MNOs can supply temporary coverage to needed areas, due to highly populated events or cell malfunctions or maintenance. The resulting network coverage is time-dependent, irregular and different for each MNO, presenting a complex environment to handle. All these characteristics are crucial for the quality of the input data, adding a layer of complexity relevant for the overall result. Coverage data, in form of cell availability and structure, antenna details and regular update procedures are not always disclosed by the MNOs, posing an upper threshold to the data quality. Another problem is that MNO events carry information only about the device and the SIM card it contains. There is not a one-to-one correspondence between SIM subscribers, device owners and population unit. It is possible, for the same person, to have more than one device and more than one SIM into the same device. In addition, there are age classes, children and older people, which do not use mobile phones at all, resulting in a undercoverage of the population.

Finally, there could be a temporal discontinuity in the data, referring to missing timeframes that could last hours, days or weeks, for example caused by device hardware malfunctions, software crashes or battery drainage. These occurrences can impact the data, and specific measures are needed to mitigate erroneous data interpretation. A common example is the specific night behaviour of the user: if a user is accustomed to switch off his/her device during the night, the data will present gaps at night time.

4. *Dimension Time-related dimension*

The time indicators of the data, commonly in form of timestamps preponed to event lines in log files, are a vital source of information for referencing the time at which an event occurs. Each MNO, and at some extent each system that is referencing the time during the operations, adopts a format for the timestamps, that includes a combination of information order (date before time or reverse), words or numerals for the month and weekday, separators (usually hyphens or slashes) and spaces. Apart from the timestamp format, for international operations there are more details to take into account, such as time zone and daylight saving time that could be different for different countries.

5. *Dimension Integrability*

A key quality dimension for data is their integrability, referring to the extent to which a data source is capable of undergoing integration or being integrated with other data sources. A highly integrable data source is one that can seamlessly blend into an existing database without requiring much elaboration. For example, the data source could be coded in a compressed format or include internal checks that limit or complicate the possibility to integrate the information with previously acquired data.

5.3 First reflections on the quality of the lower throughput

The technological progress occurred in the recent years made a great amount and variety of data available. NSIs are asked to provide timely relevant data to support policy making hence they need to explore the possibility to use new sources of additional data, held in some cases by the private sector, as it occurs with Mobile phone data. As previously said, such data are held and managed by MNOs, they cannot be released and they are analysed directly by the holders. As indicated by recent projects, like the Big Data ESSnet, a convenient solution for NSIs to benefit from the MNO data value could be a collaboration that see the MNO as the principal player in the raw data processing, and the NSI as responsible for the definition of production indicators in different domains of interest, thus splitting the production process in a "lower throughput" stage carried out by MNO in which original microdata are transformed in intermediate aggregate data and an "upper throughput" stage, managed by NSIs for the production of statistics.

The quality of the lower throughput stage is probably the most critical and peculiar aspect of MNO data processing for statistical purposes. The organisational model of the relationship between MNOs and NSIs has a great impact, since the case in which the MNO process the data autonomously has different implications with respect the case in which the NSIs can give indications on how to manage data in a continuative and collaborative communication. In both cases, a detailed documentation of the process can be fundamental for the quality assessment, in addition, in the second scenario MNOs can receive guidance directly from NSIs for the development of standardised algorithm and quality assurance system that could be integrated in the data processing.

Indeed, the contribution of the NSI in the elaboration process design and in its implementation is to promote a standardisation of the process and to realise an Open Algorithm by defining a standardised documentation of the used algorithms. The aim is to encourage the reproducibility, reusability, verifiability and sharing of the process (Grazzini *et al.* 2018) by a community of MNO data analysts who can contribute to the code improvement and verification. The openness of the algorithm is a controversial aspect. The OPAL (OPen ALgorithm) project[3] deals with

3 Cfr. OPen ALgorithm (OPAL). https://www.opalproject.org.

the principle of algorithmic openness, identifying the maximum level of openness with the publication of an open source code used by MNOs and the lower level of openness with the verbal description of the algorithms. Being private companies, MNOs are not willing to make the code public, so it become crucial for the NSI to supervise the entire process in order to ensure the quality of the output, by measuring quality in all implemented methods. To ease the quality assessment, reproducibility, reusability and verifiability, the raw data processing by the MNOs should be arranged in sequential modules, possibly containing sub-processes and sub-functions. Versioning managing is crucial for the code evolution and modularity can help in this. Moreover, a structure made of independent modules with different sub-functions can allow the use of the same process design for different domains and application (production of indicators for different domains), improving flexibility and generalisation. This type of modelling is facilitated by the use of the Enterprise Architecture framework along with a specific standard language[4], able to manage the levels modelling (including the implementation level and the technological level) using a high-level language and to highlight the role of actors in the process. The adoption of a standard language not directly linked to a technical choice extends the accessibility to a non-technical audience and makes module modelling independent of the used development language.

Standard process modelling can also be useful to analyse the process from a more statistical perspective to identify the main issues to be taken into account in the quality framework definition. It is important to analyse the different process steps trying to identify the most relevant sources of errors that could affect the statistical results. Just as an example, without GPS information, in the lower throughput stage the MNO applies methods to geolocalise the device that implies a certain level of uncertainty of the result. Once identified the main sources of errors, ad-hoc quality measures can be proposed to monitor and assess them. This analysis is still to be carried out in a systematic way.

6. Next steps

In this paper we described the first part of the long and complex journey we have undertaken to define a quality framework for statistics based on MNO data. It may seem like we are not even halfway there, but the systematic analysis of the applicability of Institutional level principles and input data quality dimensions gave already us an insight of the main peculiarities and quality issues that should be faced.

The next step will be the analysis of the lower throughput steps, to identify error sources arising, possible mitigating actions and quality measures. It is already clear that the uncertainty that can be introduced in the data by the different methods adopted by the MNO to manage spatial and time dimension will be relevant. Given the importance of transparency for this phase, a possibility is also to develop a specific template for documentation that could simplify and reduce the burden for MNO.

The analysis of applicability of input data quality dimensions proposed for administrative data revealed that many of them are more tailored to be applied to the intermediate aggregated data that are provided by MNO to NSI. This suggests that the first step of the development of the quality layer for the upper throughput quality can start with the analysis of the quality characteristics of this intermediate product. For this purpose, the SIMS template adapted for MNO data in the ESSnet Big Data II (see Section 4) can be the right tool to begin.

4 Cfr. Archimate. https://pubs.opengroup.org/architecture/archimate3-doc/.

Then, being the upper throughput stage under the control of NSIs, it could be easier, once Sound methodologies[5] and Appropriate statistical procedures[6] have been set up for the treatment of these data, to build a quality assurance system oriented to the monitoring of the methods applied and to the evaluation of their impact on the statistics produced. In this context, particular attention should be posed to the integration of data from multiple MNOs.

Finally concerning output quality, the evaluation should follow the traditional quality criteria[7], with a special focus on Comparability and coherence for the validation of the results produced, on Clarity for *ad hoc* quality reporting and, obviously, on Accuracy, to understand the possible impact on final estimates of the errors arisen during the process.

References

Amaya A., P. Biemer, and D. Kynion. 2020. "Total Error in a Big Data world: adapting the TSE framework to Big Data". *Journal of Survey Statistics and Methodology*, Volume 8, N. 1: 89-119.

European Commission, Directorate-General for Communications Networks, Content and Technology. 2020. *Towards a European strategy on business-to-government data sharing for the public interest – Final report prepared by the High-Level Expert Group on Business-to-Government Data Sharing*. Luxembourg: Publication Office of European Union.

Daas P., S. Ossen, R. Vis-Visschers, and J. Arends-Tóth. 2009. *Checklist for the Quality evaluation of Administrative Data Sources*. The Hague/Heerlen, the Netherlands: Statistics Netherlands.

Daas, P., S. Ossen, M. Tennekes, L.-C. Zhang, C. Hendriks, K. Foldal Haugen, F. Cerroni, G. Di Bella, T. Laitila, A. Wallgren, and B. Wallgren. 2011. "Reports on methods preferred for the quality indicators of administrative data sources". *Deliverable 4.2 BLUE-ETS Project SSH-CT-2010-244767.*

ESSnet Big Data II, Quaresma, S., J. Maślankowski, D. Salgado, G. Ascari, G. Brancato, L. Di Consiglio, P. Righi, T. Tuoto, P. Daas, M. Six, and A. Kowarik (*eds*). 2020*a. Deliverable K3: Revised Version of the Quality Guidelines for the Acquisition and Usage of Big Data*. Luxembourg: Eurostat.

ESSnet Big Data II, J. Maślankowski, D. Salgado, S. Quaresma, G. Ascari, G. Brancato, L. Di Consiglio, P. Righi, T. Tuoto, P. Daas, M. Six, and A. Kowarik (*eds*). 2020*b Deliverable K6: Quality report template*. Luxembourg: Eurostat.

ESSnet KOMUSO. 2019. "Quality guidelines for multisource statistics (QGMSS)*"*. *Technical report*. Luxembourg: Eurostat. https://wayback.archive-it.org/12090/20231216134723/https://cros-legacy.ec.europa.eu/system/files/qgmss-v1.1_1.pdf.

Eurostat, ESS Task Force on the use of Mobile Network Operator (MNO) data for Official Statistics. 2023. "Reusing mobile network operator data for official statistics: the case for a common methodological framework for the European Statistical System – 2023 edition". *Statistical Reports*. Luxembourg: Publications Office of the European Union. https://ec.europa.eu/eurostat/web/products-statistical-reports/w/ks-ft-23-001.

Eurostat 2021. "ESS handbook for quality and metadata reports EHQMR, 2021 re-edition". *Manuals and guidelines*. Luxembourg: Publications Office of the European Union. https://ec.europa.eu/eurostat/documents/3859598/13925930/KS-GQ-21-021-EN-N.pdf.

5 Principle 7 of the ES Code of Practice, in the area of Statistical Processes.
6 Principle 8 of the ES Code of Practice in the area Statistical Processes.
7 Relevance, Accuracy and reliability, Timeliness and Punctuality, Coherence and Comparability, Accessibility and Clarity.

Eurostat. 2019. "Integrating alternative data sources into official statistics: a system-design approach*". 67th plenary session of the Conference of European Statisticians*. Paris, France, 28-29 June 2019.

Eurostat. 2017. *European Statistics Code of Practice*. Luxembourg: Publications Office of the European Union. https://ec.europa.eu/eurostat/web/products-catalogues/-/ks-02-18-142.

Grazzini, J., P. Lamarche, J. Gaffuri, and J-M. Museux. 2018. "Show me your code, and then I will trust your figures: Towards software-agnostic open algorithms in statistical production". *European Conference on Quality in Official Statistics (Q2018).* Kraków, Poland. 26-29 June 2018.

Groves, R.M., F.J. Jr. Fowler, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourrangeau. 2004. *Survey Methodology*. New York, NY, U.S.: John Wiley & Sons.

Iovan, C., A.M. Olteanu-Raimond, T. Couronné, and Z. Smoreda. 2013. "Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies". In Vandenbroucke, D., B. Bucher, and J. Crompvoets (*eds*). *Geographic Information Science at the Heart of Europe. Lecture Notes in Geoinformation and Cartography*. Cham, Switzerland: Springer.

Reid, G., F. Zabala, and A. Holmberg. 2017. "Extending TSE to administrative data: A quality framework and case studies from Stats NZ". *Journal of Official Statistics*, Volume 33, N. 2: 477-511.

Ricciato, F., A. Wirthmann, K. Giannakouris, F. Reis, and M. Skaliotis. 2019. "Trusted smart statistics: Motivations and Principles". *Statistical Journal of the IAOS*, Volume 35, N. 4: 589-603.

Santamaria, C., F. Sermi, S. Spyratos, S.M. Iacus, A. Annunziato, D. Tarchi, and M. Vespe. 2020. *Measuring the impact of COVID-19 confinement measures on human mobility using mobile positioning data.* Luxembourg: Publications Office of the European Union.

Stodden, V. 2014. "The reproducible research movement in statistics". *Statistical Journal of the IAOS*, Volume 30, N. 2: 91-93.

UNSD 2022. "Methodological guide on the use of mobile phone data: tourism statistics". *UN Statisitcs Wiki.* https://unstats.un.org/wiki/display/MPDTS.

Zhang, L.-C. 2012. "Topics of statistical theory for register-based statistics and data integration". *Statistica Neerlandica*, Volume 66, N. 1: 41-63.

Navigating quality challenges in landscaping web data: New aspects and source stability

Magdalena Six, Alexander Kowarik[1]

Abstract

This paper delves into the challenges of landscaping web data, specifically focussing on the development of quality aspects for new data sources. It highlights the limitations of traditional quality dimensions when working with web-scraped data and emphasises the need for additional considerations along the data processing pipeline. It explores the process of website selection, emphasising the importance of a standardised assessment tool to ensure comparability between different countries. Moreover, it discusses the impact of source stability on data quality, illustrating how unstable access to data sources can block accurate analysis and limit the reliability of statistical indicators. Real-world examples showcase the complexities of interpreting observed web data, further emphasising the significance of reliable and stable data sources.

Keywords: Web scraping, quality, relevance of sources, landscaping.

1. Landscaping of websites for web scraping with focus on selection models

Within a company or organisation, the term "landscaping" refers to cataloguing and measurement of all the data in the company or organisation.

Similarly, in the world of web-based data, landscaping could be understood as cataloguing and measurement of all web-based data sources relevant for the topic of interest.

It is worth noting that no general definition of landscaping in case of web-based data for Official Statistics has emerged yet. There seems to be a common understanding that "landscaping" refers to the process(es) before the actual ingestion of data from the websites starts.

Informally speaking, "landscaping" can be interpreted as "getting an overview of the relevant sources". Once one knows about all relevant or all potentially relevant sources, one can gather information in a further step about these websites. Based on this information one can select the sources out of the potentially relevant sources, which are afterwards actually used for web scraping.

We are not aware of a precise definition clarifying if the term "landscaping" refers only to the first step, namely the cataloguing of potential sources, or if it also comprises the measurement of the sources and based on this measurement, the selection of sources.

Following examples such as the data pipeline for online job vacancies where landscaping seems to include all processes before the data ingestion starts, we give our own definition as follows:

- landscaping comprises all process steps necessary to catalogue all relevant sources for a specific topic of interest, to measure the quality and technical viability of the catalogued sources and to select the sources, which are actually used, based on the measured criteria.

1 Magdalena Six (Magdalena.six@statistikgv.at), Alexander Kowarik, (Alexander.kowarik@statistikgv.at), Statistics Austria. The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of Statistics Austria. This work has been part of the European project "2020-PL-SmartStat Web Intelligence Network". Part of the text has been released published elsewhere, especially as deliverables for the mentioned project

The three sub-processes cataloguing, measurement and selection build and depend on each other. Starting from the first cataloguing, this can be an iterative process between measurement and selection and even cataloguing.

Depending on the topic of interest, the difficulty of the landscaping exercise as a whole and of each sub-process can vary enormously.

There are two dimensions that are central for understanding the complexity of the landscaping for a certain/topic:

- if all websites should be captured or representatives should be selected;
- if additional information is available or not.

The most common kind of additional information is information from the statistical agency itself, *e.g.* names of businesses in a certain branch from the business register or relevant companies for online-shopping of clothes.

If a representative subset of websites should be selected, the concept of representativity, which is not defined mathematically, but mere as an idea, has to be operationalised in some way, *e.g.* by a random sample or a cut-off sample.

1.1 Cataloguing

This process is easier if the starting point is the list of enterprises being part of a certain population of interest. If you have the companies names you can use them specifically in your online search, which will lead to more specific results.

It can also be the aim to actually identify the population of enterprises/units active in a certain field, based on their enterprise websites. In this situation you have to search for keywords instead of enterprise names, which will generally lead to a higher variety of search results.

In both cases – with or without additional information about the target population - the result of the cataloguing process is a list of URLs which can, but do not have to belong to the respective statistical unit / respective topic of interest. Especially, when you want to capture all websites of a large target population, the sheer number of URLs makes it impossible to visit each website "manually" to decide if the URL belongs to the target population. You therefore need an automated mechanism, which estimates for each URL in your catalogued URLs if it belongs to the target population.

1.2 Selection of websites

In the case of starting from a population of enterprises/units, you have a list of enterprises from the Statistical Business Register (SBR), for which you want to find information on the web. On the other hand, you have a catalogue of retrieved websites after searching for the respective enterprises, which might or might not belong to the respective enterprises in the SBR. So, the selection step is selecting the valid URLs that correspond to statistical units, mostly via linking procedures. Additionally, you have to check technical criteria if a website can be scraped, *e.g.* by looking at the robots.txt. The validity check often contains checking unique identifiers that (by law) must be present on a company's website such as the value added tax or company registration number.

If no additional information is available for the population, selection mainly corresponds to classifying websites if they belong to your target population, *e.g.* selling or developing a certain product. This can be done for example with:

- word-based methods: counting the occurrences of certain keywords;
- semi-supervised learning: a set of positives examples needs to be available;
- supervised learning: a training data set with positive and negative examples must be available.

1.2.1 Selection of websites (as Multi-Criteria Decision Making problem)

When it is necessary to select representatives from all catalogued websites, it is of utter importance that this selection process does not happen arbitrarily. Especially, when the comparability between different countries has to be guaranteed, the need for a standard tool for the assessment of websites becomes obvious.

A generic answer to the question "Which websites should be selected?" would probably involve answers such as: "The most important ones", "The ones with the highest quality", "The most representative ones" or all of the beforementioned. But quantifying importance or quality can be rather tricky.

The need to decide for/rank several websites with respect to specific criteria such as the popularity of the website, the trustworthiness of the website owner, the structure of the data on the website etc. shows that the task fits well to the framework of Multi-Criteria Decision Making (MCDM) . MSCM is well known as subdiscipline of Operations Research. The field of MSCM offers manifold tools and methods for determining the best alternative by considering more than one criterion in the selection process.

MCDM problems are characterised by three ingredients:

1. the alternatives which should be ranked (the websites);
2. the criteria based on which the alternatives should be ranked (the characteristics of the job portals);
3. the model, which determines - based on the criteria and the values of each alternative - the ranking of the alternatives (the selection model).

There are in general three groups of criteria that can be used in the selection process: 1) information from the website itself, 2) information about the website and 3) information from previous scraping of a specific website. The inclusion of all three categories of information might be costly but leads to the most trustworthy selection of websites suitable for scraping.

2. Quality indicators to measure the relevance and stability of selected OJA sources

In this chapter, we focus on a specific aspect of the whole production process – the relevance and the stability of the selected sources based on the examples of job portals. We do this from the perspective of a user of the pre-processed data from the central Eurostat platform – the web intelligence hub, who has limited insight and control in Eurostat's landscaping decisions and advanced selection models. We believe that this is a very meaningful exercise as it is close to the current actual situation when using the centrally scraped OJA (Online Job Advertisement) data– it shows if the producers' efforts in landscaping OJA sources fulfil indeed the users' expectations about the relevance and the stability of the sources.

2.1 Indicators for relevance of selected sources

If a producer of Official Statistics is not in charge of the landscaping process itself, it is crucial to check if the included sources are indeed the most important ones and correspond to the sources that would have been considered by NSI domain experts as well. For this we propose to consider the following indicators:

- if your NSI scrapes OJA data itself, compare the included sources from your own scraping processes with the included sources on the Web Intelligence Platform (WIP);
- if your NSI does not scrape, consult the labour market experts in your NSI and ask them to name the *x* most important job portals in your country and compare this list with the sources on the WIP for your country.

2.2 Indicators for the stability of existence of the included sources

The general goal when working with OJA data is to capture dynamics in the labour market with the indicator 'number of vacant positions advertised online'. It is impossible to scrape every existing job portal (source) per country. Already in the landscaping process, specific sources (websites) per country were selected by Eurostat. Unfortunately, the stability of the time series is impacted by changing sources included in the OJA data.

Creating a time series by simply adding up all unique OJAs over an instable number of sources probably won't capture effects from the labour market, but effects reflecting the inclusion of certain sources at a certain point in time. *E.g.* if a formerly included source falls away and the number of all scraped OJAs falls, one does not know if this decrease is due to the excluded source or due to a decrease of advertised open positions. If the number of existing sources is instable, more advanced methodological tools such as chaining need to be considered to construct a meaningful time series over the aggregated sources. As a first step, it is important to get an overview over the stability of the existence of the sources. For this we propose to look at the following indicators:

- determine if it is always the same sources in the course of the time span considered;
- determine for several points in time (*e.g.* at the beginning of the time series, the middle and the end of the time series) the *x* (*e.g.* 5 or 10) most important sources w.r.t. to the volume of OJAs scraped of each of the sources. Are the thereby found important sources included in the list of scraped sources over the whole time series?

2.3 Indicators for the stability of the popularity of the included sources

Even if the number of scraped sources stays stable over time and all of the most important sources are included in the whole time series, the following can happen: the popularity of one source increases, leading to more OJAs on this portal, uncorrelated with a general increase of vacant positions in the job market. The following indicators can give you hints if such a phenomenon occurs in the data:

- calculate the ranking of the most important sources w.r.t the OJA volume and observe this ranking over the course of time;
- determine the number of OJAs per source and check (*e.g.* via a plot of the individual time series) if the dynamics of the individual time series per source are similar.

2.4 Indicators for the stability of sources over different versions of data

When a new version of data is available centrally, the data should not change for time intervals which were already covered by older versions of available data. Most important, this is true for the sources - It can be annoying if a source which was included in former years disappears for the present year. But it completely makes your analysis unusable if the sources in former years disappear for former years in a new version of data.

To measure the stability between different versions of data, we propose the following steps:

- load an old version of OJA data as well as the most recent version from the WIP. For the overlapping years, calculate for relevant sources the number of OJAs per year for the old data version and the most recent data version. Calculate the difference in absolute numbers as well as well as in relative numbers;
- of course, you can do this for several versions of old data.

References

Kowarik, A., P. Daas, M. Bruno, M. Six, O. ten Bosch, G. Ruocco, C. de Maricourt, V. Chavdarov. 2021. “Minimal guidelines and recommendations for Implementation”. *Deliverable 4.1 ESSnet Trusted Smart Statistcs – Web Intelligence Network, Grant Agreement Number: 101035829 – 2020-PL-SmartStat*. Luxembourg: Eurostat. https://cros.ec.europa.eu/system/files/2023-12/deliverable_4_1_minimal_guidelines_and_recommendations_for_implementation_essnet_tss_win.pdf.

Assessing the quality of transaction data for use in the Consumer Price Index

Anthony Dawson, Calvin O'Brien[1]

Abstract

This paper details the work of the Consumer Prices area of the Central Statistics Office (CSO) to assess the quality of an alternative data source for quality and fitness of use in the Consumer Price Index. It details the methodology used to assess the alternative data source, how that assessment is factored into the decision around whether to use the new alternative data source and how to mitigate risks associated with the inclusion of an alternative data source. This paper is accompanied by a practical assessment of one of the alternative data sources used in the Consumer Price Index in Ireland which is based off of the Voorburg Group Paper on Guidelines for Incorporating Alternative Data Source in Official Statistics[2].

Keywords: Scanner data, alternative data, quality assessment, Consumer Price Index.

1. Introduction

The emergence of new alternative data sources has provided the Central Statistics Office (CSO) with excellent new opportunities to meet new data requirements, fill existing data gaps or improve the quality of the existing statistical outputs. However, new alternative data sources can provide challenges for assessing the quality of the new data sources that are not met by existing standards or procedures within a survey area. It is therefore important for the CSO to have a structured and considered evaluation process which will help to mitigate risks and determine the suitability of a new data source.

1.1 Challenges associated with new data sources

Depending on the data source, the National Statistical Institute (NSI) has differing levels of control over the variables available, and the methodology associated with collecting and processing the raw data. For example, with traditional administrative data sourced from other government departments, an NSI may have input into how the data is processed and transmitted to the NSI. However, with other data sources such as transaction or scanner data from a retailer, there may be no way of assessing methodology prior to transmission to the NSI.

In addition to methodological issues, it is also important to consider ethics from the start of the acquisition process. For structured datasets sourced from the private sector, it is especially important to keep in mind the incentives of each party involved in the discussion of acquisition. This is in order to maintain the impartiality of the NSI and avoid unforeseen impacts, intentional or otherwise.

1 Anthony Dawson, (Anthony.dawson@cso.ie), Calvin O'Brien, (calvin.obrien@cso.ie), Central Statistics Office, Ireland. The views and opinions expressed are those of the author and do not necessarily reflect the official policy or position of the Central Statistics Office, Ireland.

2 https://voorburggroup.org/Documents/2022%20Ottawa/Papers/1006.pdf.

1.2 Assessing quality of new data sources

There is a desire when a new data source becomes available to immediately insert the new data source into the existing processes. However, there are two factors that should be accounted for before a decision is made to insert new data sources into a statistical process.

1.2.1 Quality of the data

The new data should be assessed to determine the overall quality of the data. These are checks to assess:

- missing variables;
- missing observations;
- other measurable metrics for assessing quality of the data.

1.2.2 Fitness of use of the data

The data should also be assessed to determine it suitable for use in the survey area it is required. This will include checks such as:

- the number of variable present and that the variables are suitable for use in the survey area;
- the number of observations in a typical dataset and whether the volume of data is suitable for the survey area;
- consistency checks between existing data sources and the new data source;
- if the data is a regular transmission, the consistency across time periods of the data.

1.3 Ongoing assesment

Once a decision has been made to include a new data source into a process, it is important to maintain regular assessments of a data source to ensure that the quality of the data does not change over time. The NSI has no control over how the data is collected and processed and therefore any change in the providers' methodology may have unintended impacts on the data that is provided to the NSI.

2. Central Statistics Office assessment of scanner data for use in the Consumer Price Index

2.1 Price collection in the Consumer Price Index (CPI)

Historically, in the CPI, pricers went into shops around the Republic of Ireland and collected prices from a number of shops. These pricers use a specially developed mobile phone app to record prices of products in shops and transmit these prices back to the CSO for use in the CPI. Before the COVID-19 pandemic there were around 100 pricers located around the country collecting prices. This method of pricing was paused during the COVID-19 pandemic and alternative methods of collecting price data had to be implemented quickly to ensure the continued production of the CPI.

The CSO moved to automated and manual web scraping as well as requesting data directly from retailers when pricers were no longer able to enter shops and these became the predominant data sources during the COVID-19 pandemic. It was clear that once the pandemic had finished that a new hybrid method of price collecting was the way forward and that the number of pricers could be reduced. Since the return of pricers to shops, there are only 30 pricers going into stores around the country and the short fall has been replaced by manual web scraping done by in office staff and new scanner data that is transmitted directly from retailers. Scanner data are digital transaction data recorded at the cash registers of retail shops which inform about turnover, sales and the type of item sold. This data can be used to calculate and average priced based on the turnover and sales and allows an NSI to factor in special offers such as two for one and promotional prices. While manual web scraping, which is searching online for prices and recording the prices, is very similar to what the pricers were originally doing in store, scanner data comes with completely new methodological and organisational challenges. While some of these were ignored in favour of getting the data in for use during the pandemic, as the CSO expands its use across the Consumer Price Index, more rigorous methods of assessing the quality of the data and whether a particular retailers' data is suitable for use in the CPI have to be developed.

2.2 Developing an assessment method for scanner data

There are well developed guidelines in place for survey design and monitoring however, alternative data sources can provide challenges for assessing quality and fitness of use that are not met by existing documentation. Usually with a standard survey, the NSI would be collecting the data, processing, and disseminating that data. Scanner data does not give the NSI any control over how the data is collected and processed on the retailer side, therefore some methods could potentially be hidden but included in the received data product. This is an example of a processing error but may not be accounted for by standard methods of assessing the quality.

The topic of how to assess alternative data sources for quality and fitness of use has been a topic at the Voorburg Group for a number of years and recently a focus group was set up to develop guidelines for the assessment of alternative data sources. This group presented their work at the 2022 Voorburg Group which consisted of a paper along with a fitness of use questionnaire which makes use of existing standards to evaluate data. This questionnaire was developed based on existing standards such as the Generic Statistical Business Process Model (GSBPM). It also takes a lot of information from existing literature for reporting on administrative data, with papers such as the Stats New Zealand Guide to Reporting on Administrative Data Quality. This questionnaire was intended to help with the consideration of where quality data can be assessed and controlled.

It was determined that this questionnaire was the right tool to allow the CSO to do a full assessment of the data based on the GSBPM and to determine whether any new scanner data was suitable for use in the CPI and how ongoing quality metrics should be defined.

2.3 Assessing a new scanner data source

The CSO implemented the questionnaire in determining whether a new source of scanner data was suitable for use in the CPI as discussed previously. While there were pricers already attending the store, the option to receive a data transmission on a weekly basis became available from the retailer. The decision was made that this should be investigated fully before a commitment was

made to move to the using the scanner data in the CPI. This investigation was based on receiving the data for a number of months to allow for longitudinal analysis as well as the individual data. Below is a quick summary of the different aspects of the GSBPM and how they factored into a final decision being made on the quality of the data and its use in the CPI. A full assessment using the questionnaire is detailed in Appendix A.

2.3.1. Specifying needs

In the specifying needs stage, it is important to determine the necessity of acquiring a new data source. While the retailer was already being surveyed using traditional methods, the provision of scanner data would free up pricers to possibly price other smaller store, increasing the scope and quality of the CPI. As there is a statutory requirement for retailers to provide this type of information if requested, there is no cost associated with the acquisition of this data source, therefore the cost is minimal and only related to the time cost of setting up the automated transfer procedure.

2.3.2 Design

In the design phase, detailed examination of the metadata, coverage and variables will determine whether the scanner data source is fit for purpose for use in the CPI. The CSO had multiple meetings with the data provider to determine that the data was suitable. This included discussions regarding whether the data covered the whole of the Republic of Ireland and if it covered everything sold in the shops. It was also an opportunity to discuss what variables are available, any variables that could be added to improve the coding of the data and what derived variables needed to be created on the CSO's side. The consultation period during the design phase of the assessment allows the CSO to plan the build phase, as well as processing and analysis at a later date once data has been received from the retailer.

2.3.3 Build

Based on the discussions in the design phase with the retailer, it was agreed that the retailer would transmit data using the same Secure File Transfer Protocol system that had been used for other retailers who had provided data during the pandemic. This shows the value of good communication during the design phase. As this data was being used to replace an existing data source, once the data had been ingested and processed the data was able to fit into already existing statistical programmes.

2.3.4 Collect

Again, the collect phase is based off of constructive conversations with the data provider in the design phase. It is important to get a good understanding of how the data is collected by the data supplier. In the case of scanner data, the "collection" of data is done via the scanners in the supermarket and is a census of every item that is sold on a weekly basis. This method of collecting data is suitable for use in the CPI so there were no concerns with how the data was being collected. The variables available from this data collection method were also suitable.

The collection method used by the retailer also had some benefits compared to the existing data collection method used by the Central Statistics Office. The ability to calculate a unit price means that the CSO is able to capture special offers that may occur throughout the month and

factor these into price calculations. This was unreliable under the old methodology as pricers would only be in a shop on one day and may miss special offers on a specific day. From the consultation during the design phase, it was determined that the scanner data has a constant flow of products entering and leaving the data. There is no record on the dataset of new items into the data set or what items have left the dataset. It is important to keep track of the number of new and old products so that any issues can be quickly identified. It was decided that quality metrics should be set up to track new and missing items when the data arrives.

2.3.5 Processing the new scanner data

The process phase of the evaluation is the most detailed and important aspect of assessing the quality of the data and whether it is fit for use in the CPI. It is important to address issues such as how to deal with missing variables and observations, and determining whether the process of incorporating the new data source matches up with the existing processes, or if new processes need to be developed. As the CSO are already receiving scanner data from other data sources, there was already a process in place and the new scanner data fit into this process without much need for any adjustment of data or outputs. The questionnaire gave the CSO a good idea of the type of questions that should be asked of the scanner data source and after consideration, it was determined that while some the existing processes were adequate, there was not enough attention paid to the ongoing measurement of quality of the data. This included further analysis on missing data, outlier detection, significant changes in price or coherency issues. It was determined that these quality metrics would need to be tracked for a number of months before a final decision was made on whether to include or exclude the new data source into the CPI index as a replacement for manual pricing.

2.3.6 Analysis of derived data

The scanner data is combined with data from a number of other sources such as manual pricing, web scraping and other scanner data sources. Therefore, there is no need for any statistical disclosure control. While it would be possible to create other derived variables from the data, these are not within the scope of the CPI and do not need to be assessed.

2.3.7 Dissemination

The dissemination phase is a chance to evaluate the dissemination of the data. Questions to be considered here are whether the data will be a replacement for existing data sources or will be a new statistical product and whether the quality is sufficient to be considered for use in Official Statistics or whether they should go out as experimental statistics.

In this case, the scanner data is being used to replace an existing data collection method, in store pricing, and will supplement a number of other data sources. The data will form part of the Consumer Price Index, a key national economic indicator so the quality of the data has to be of the highest quality to be considered for use.

2.3.8 Evaluation

The evaluation phase is used throughout the previous categories of the GSBPM assessment to assess the performance of the data source relative to the initial need. However, there are two major evaluations

that were carried out, one after the initial consultation period as part of the collect phase and one once the data had been collected for 8 weeks and tested as part of the process and analyse phase.

The initial evaluation was used to determine what sort of initial tests were needed in order for the data to be deemed suitable for use in the CPI. As already discussed, an initial set of tests were required to determine the consistency of the data over multiple weeks. The consistency would be measured by linking products using their in-store identifiers. Prices would need to be tracked to check for any significant changes week on week, using upper and lower thresholds to determine outliers. It was also decided that a system to track the number of new and removed products from the data. The data would also be compared to the existing data collected manually to ensure that there was a degree of coherence between the two stores.

The final evaluation process was conducted once 8 weeks of data had been collected from the retailer. This evaluation consisted of assessing the answers to the questionnaire that had been answered at the beginning and updating the responses based on the experience of dealing with the data over a period of time. The metrics set out in the initial evaluation were key in determining the quality and suitability of the scanner data for use in the CPI.

3. Conclusions

Assessing the quality of alternative data is a new challenge that many statistical offices are facing as data collection moves away from traditional methods. While alternative data sources bring a number of new opportunities to either improve existing statistical products or create brand new products, the NSI has to be wary of a number of factors including possible survey errors and possible issues around the ethics of using data that wasn't initially collected for statistical purposes.

Scanner data, while being a very useful data source due to the lack of resourcing required compared to traditional price collecting method, is not designed as a statistical data collection. This has meant the CSO has had to do a thorough evaluation of the data before it is considered for use in the Consumer Price Index. While there is a lot of experience of utilising scanner data across the European Statistical System, it is important that a country carries out a thorough review of specific data sources before incorporating them into their processes.

The CSO, using the fitness of use questionnaire developed by the Voorburg Group, has evaluated scanner data at two stages: initially upon consultation with the retailer and after a number of weeks of data provision. This two-stage evaluation allowed the CSO to determine early if their acquisition stage was worth continuing with and then to ensure that the data provided is to the standard expected over a prolonged period of time. This also allowed for analysis of more quantitative metrics to be carried out on the data such as missing observations, data coherence over time and to other data sources and consistency of the data from the retailer.

After the final evaluation, it was determined that the scanner data provided by the retailer was suitable for use in the Consumer Price Index. The metrics set out in the evaluation stage showed that there was longitudinal consistency with the data provided and that issues with missing values were not significant enough to cause issues. A minor issue with the data transmission was flagged by these metric analysis programmes and was quickly rectified with the retailer. The decision was made that the scanner data source will be incorporated into the CPI from the start of the New Year.

Appendix A

Assessing Supermarket X Scanner Data's Fitness for Use in the Consumer Price Index

Introduction

With the emergence of new alternative data sources, the requirement to assess quality is becoming increasingly critical. These alternative data sources however can provide challenges for assessing quality that are not present with administrative data. Issues such as inability to specify or design data sources, lack of control over the data and biases from the private sector need to be accounted for. This questionnaire aims to assess the quality of Supermarket X and determine whether it is fit for use in the CSO.

Specifying needs

What are the intended goals and future uses of this information?

The goal of acquiring this data is for the data to be used in the production of the Consumer Price index. The data will be used for grocery prices on the direct pricing side of CPI and replaces the existing method of collecting data which was via in-store pricing.

The data could possibly be used by other areas including retail sales.

Is there a fee for acquiring this data? Does the need outweigh the cost?

As the provision of transaction data to the CSO for the purpose of use in the CPI is covered by the Statutory Instrument, there is no cost attached to the acquisition of the data.

Will the timeliness of the data be in line with the needs of the key users and stakeholders?

The data arrives on the Monday following the reference week, providing plenty of time for use in the production of the CPI.

Does the data exhibit the characteristics of an admin dataset or alternative dataset?

The data structure is consistent with every transmission so there is no issues around comparison from week to week. The data is ingested in to the CSOs internal data system, where metadata is applied to the datasets.

Design

Is the Geographical Coverage of the data adequate for your purposes?

The data contains total quantity of product sold in every shop in the Republic of Ireland. This means the data is of adequate geographical coverage.

Does the population covered by the data align with your needs?

The data is a census of all items sold in Supermarket X shops in the given time frame. This meets the needs of the CSO.

Build

What new components may be required as a result of the updating of the process? Are these components designed, built and tested with the alternative data source?

The data is transmitted to the CSO automatically by Supermarket X. This process was executed and tested by the CSO's IT department in collaboration with Supermarket X, ensuring smooth data transmission. This is a process already in place for multiple similar data sources and therefore was simple enough to set up.

All internal data ingestion and transformation is done in conjunction with the internal data team. Again, this process was already in place for another source of transaction data so the process could be replicated with minor amendments to match with the data structure.

Implementation into the existing grocery price process is also easy to replicate so again, minor changes to match the data structure need to be done but most of the work is repetition of existing processes.

Collect

Does the collection process (performed by the data provider) have any impact on the intended use? Any means that can be used to mitigate this impact?

The collection method is from Supermarket X's own internal database. It is a census of all items that pass through the tills. The CSO gets a weekly summary of this which includes the quantities of items sold that week and the total revenue per item. This means there is no impact on the intended use.

Is the data available at the level of granularity that is required?

Although the shop offers promotions like discounts such as sales and '2 for 1' offers, the CSO receives the total quantity of each item sold per week along with total revenue per week. From these unit price per item can be calculated, meaning the data is available at the level of granularity that is required.

Are accuracy indicators available for the variables that are most important? If so are they within a range that is acceptable and if not is there a plan to address it.

The data is a census of all transactions so there are no issues with accuracy of the data.

Do the variables that are most important have enough valid values for the purpose of the data need?

The data is a census of all transactions so there are no issues with lack of data.

Is there sufficient consistency across records in the file to meet your needs?

Scanner data has a constant flow of products entering and leaving the data. There is no record of new items or replacement in the dataset, this must be tracked by the CSO. It is probably worth tracking the number of new items in a dataset from week to week and querying with Supermarket X if there are large number new products or items replaced. These types of metric analysis are currently being developed.

Process

How will missing values in the data be handled?

This will depend on what values are missing:

- A missing variable should be immediately queried with Supermarket X and the data not used until a resolution has been made.
- Missing observations are to be expected with scanner data, however if there is a product that had a large quantity of sales in one period and then is missing from another period then this is something that would probably need to be queried. This is a metric that needs to be tracked and has been added to the metric reports.

Are there any circumstances associated with the chosen reference and collection periods that might cause issues in the quality of completeness of the data? If yes, how can they be addressed?

Extreme weather events could possibly affect the data as the stores may be closed, however this is very unlikely to happen nationwide.

What types of response errors are expected and what is the likelihood of their occurrence (reporting error, incorrect information)? How will risks be mitigated?

As Supermarket X are reliant on this data, there is very little chance that the data would be incorrect. However, checks are in place to check for any obviously incorrect data (negative sales, prices etc.) and depending on their importance they can be queried or left out of the data.

Is there evidence of bias in the data? Does the NSI have the ability to maintain independence of their statistical outputs with respect to the objectives of the data provider or the originally intended use of the data?

There is no evidence of bias in the data as the data is a census of all transactions that pass through the tills at Supermarket X. While there is a possibility for the provider to provide biased data, there is no benefit to them to do this so it is unlikely that this would happen. It is possible that if doubts were to occur around the data, web scraping or spot checks could be done to assess coherency.

Are standard concepts and/or classifications being used in the data files? If not how will this be addressed?

Supermarket X have their own classification system of product classification. Manual classification is performed in the office on new products to assign them to the correct CPI coding. Currently, a Machine Learning solution is being developed with the aim of automating this process which will be verified by staff.

Will established statistical method be used to create the indirect estimates, direct tabulation or analysis for official release?

As the data will be used in CPI production for areas such as direct pricing, it will be used for official releases.

Will the product derived from the data be compared with historical data?

No, the product derived from the data will replace historical data in CPI production.

Has the mechanism for data transmission been identified, built and tested?

Due to the CSO already having processes in place for multiple similar data sources, identifying and building the mechanism for data transmission was a simple process.

Data is collected at the tills in Supermarket X, where a weekly census of all items sold is generated. This data is transmitted automatically to the CSO and internal data ingestion is done in conjunction with the internal data team. The data is then processed and prepared for analysis, after which it enters CPI production.

This mechanism has been tested and operates correctly. There was an issue with the data provided being the same for a number of weeks. This was picked up and rectified with the supplier. Checks have been added to the process to ensure this is picked up as soon as possible from now on.

Have measures been identified for monitoring the quality of data transmitted on a ongoing basis?

Yes, automated monthly reports are in place which displays information such as new and removed items along with irregularities in the data such as negative prices or significant changes in item prices month-on-month.

Analysis

Are there obligations to the data provider or the constituent target population on the dissemination of data derived from the alternative data source? Do specific disclosure control measures need to be put in place?

As the data is combined with data from other sources including other supermarket scanner data, manually collected prices and web scraped prices, there is no requirement to put disclosure control measures in place.

There are no obligations to the data provider or constituent target audience as a result of moving to the scanner data. While other information such as market share of the supermarkets could be produced from the data, this is not within the scope of the CPI and the statutory instrument used to collect the data.

Disseminate

Will the final data products replace existing data products or will they be new to the NSI?

The data products will replace the previous method of manual data collection for this supermarket.

Will the final data products be considered as "official statistics"? Or will they be released as "experimental" statistics?

As the data products are replacing the previous grocery file that was used from formation data, they will be considered as Official Statistics.

Evaluate

Each section of this questionnaire provides an opportunity to evaluate the statistical process as well as questions on data ethics. The NSI should review this questionnaire and record their

reflections at various intervals (i.e. before acquiring or implementing an alternative data source but also during data development and periodically after implementation) to ensure that expectations are realised and/or re-evaluated.

To evaluate whether Supermarket X data was suitable for use in CPI, an initial set of tests to determine the consistency of the data were performed. These tests were performed over multiple weeks of data that had been received by the CSO.

The consistency of products in the datasets was tested by linking products using in-store IDs. The variables of these products were investigated over several weeks to see if they remained consistent. The results showed that most product variables did remain consistent. Changes that did occur were expected, such as reducing product size and renaming products.

The product prices were investigated to identify any significant changes week on week. An upper and lower threshold was set, and products which exceeded this threshold were further investigated to determine whether the change was realistic or an error. The findings revealed that very few price changes were classified as significant, and those that were, were deemed realistic.

New and removed products were documented and checked each week to determine whether product launches and removal were performed appropriately. It was found that new items represented new products in store, while removed items and their in store IDs were not reintroduced to the data.

Missing values in the dataset were listed for each week to determine whether any significant data was missing. Although there were initial issues with this, it was found this was due with issues in how the data was processed and once this was resolved there was no missing data.

Irregularities such as negative sales and quantities were investigated in the data to see if these arose from errors in the data or if any issues could potentially arise from these. The irregularities were investigated to see if they were appropriate. It was found that there was no significant irregularity that was inappropriate, so the data was determined consistent in this regard.

One issue that arose in assessing the consistency was one week's data started to reoccur rather than new weeks being ingested. This stemmed from an issue from Supermarket X and was resolved quickly once they were informed.

From these initial checks, the data was deemed suitable for use in CPI production and these checks are currently ongoing on a weekly basis to determine any issues that may arise in the future.

References

Dawson, A., R. Draper, S. Kilbey, M. Beaulleu, and K. Virgin - Voorburg Group. 2022. "Guidelines for Incorporating Alternative Data Sources in Official Statistics". *37th Meeting of Voorburg Group.* Virtual meeting, Ottawa, Canada, 13-22 September 2022.

SESSION 4

Machine Learning Methods in Survey Statistics

Introduction to Session 4 invited talks

Natalie Shlomo[1]

Abstract

The International Association of Survey Statisticians (IASS) is one of seven associations of the International Statistical Institute (ISI). Its aim is to promote good survey theory and practice around the world. Given new and emerging methods and tools that have the potential to create new content in the development of survey statistics, it is important for the IASS to support research towards the modern data ecosystem, new sources of data and their integration with survey data. This has led to the IASS promoting Data Science and applications of Machine Learning in survey research. We are proud to sponsor Session 4 of the 2023 Istat Second Workshop on Methodologies for Official Statistics, titled: Machine Learning Methods in Survey Statistics. This article provides an overview of Machine Learning approaches in survey statistics and introduces the session.

Keywords: International Association of Survey Statisticians, data science, quality framework.

1. Overview

The International Association of Survey Statisticians (IASS) is one of seven associations of the International Statistical Institute (ISI). Its aim is to promote good survey theory and practice around the world. Given new and emerging methods and tools that have the potential to create new content in the development of survey statistics, it is important for the IASS to support research towards the modern data ecosystem, new sources of data (including nonprobability sampling) and their integration with survey data. This has led to the IASS supporting research in Data Science where we use analytical and statistical methods to analyse large amounts of data to extract knowledge and understanding. In particular, the IASS is engaged in promoting applications of Machine Learning in survey research and are proud to sponsor Session 4 of the 2023 Istat Second Workshop on Methodologies for Official Statistics, titled: Machine Learning Methods in Survey Statistics. The use of Machine Learning techniques has become more prevalent in the area of Official Statistics, particularly as we move towards more use of 'found' data and administrative data in our statistical production systems.

Machine Learning techniques have been applied to the survey research pipeline under two broad headings: (1) survey data collection and (2) survey adjustments and post-processing. Under the survey data collection heading, some examples of the use of Machine Learning include optimising data collection under adaptive or responsive survey designs, predicting web breakoffs in web-based internet surveys and transforming input data, such as satellite imagery or price bar codes, into usable flat data. Under the survey adjustments and post-processing heading, some examples of the use of Machine Learning techniques have been applied to statistical data editing, nonresponse and weighting classifications, unit and item imputations for missing data, automatic coding, data integration and small area estimation.

1 Natalie Shlomo (natalie.shlomo@manchester.ac.uk), University of Manchester, United Kingdom, and President of the International Association of Survey Statisticians (IASS).

Buskirk *et al.* (2018) provide an overview of Machine Learning in survey research. Machine Learning techniques can be supervised (training with labelled data) or unsupervised (training with unlabelled data). Supervised learning is typically used to produce a prediction for some dependent variable while unsupervised learning might focus on pattern detection, for example, cluster analysis. The authors point out that Machine Learning techniques are algorithmic and data-driven and require tuning parameters, for example, the number of clusters, penalty parameter (amount of shrinkage) in LASSO, and the number of nodes in tree-based methods. There needs to be a distinction between inference and exploratory/prediction purposes where the latter is generally the focus of Machine Learning applications and their utility can be maximised. The authors also make a distinction on how predictive models in Machine Learning are assessed for goodness of fit, particularly through cross-validation where a sub-sample of the data is used as the training sample and the remaining sample as the test sample to evaluate accuracy. One clear conclusion that can be drawn from Buskirk *et al.* (2018) is that supervised learning requires high-quality training data that needs to be continually updated and revised to avoid selection and algorithmic biases over time. Kern *et al.* (2023) discuss the impact of the annotation instrument when producing training data and the impact on downstream model performance and predictions.

In an article by Puts and Daas (2021), the authors discuss Machine Learning techniques in the context of Official Statistics stating that "applying Machine Learning algorithms to produce Official Statistics is still challenging". The focus of this article is on the quality standards framework employed at National Statistical Institutes (NSIs), and how to adapt them to Machine Learning applications. For example, on the quality standard of "Accessibility and Clarity", it is particularly challenging to explain how results are obtained when the Machine Learning algorithm is a 'black-box' method, such as a neural network and deep learning processes. The authors mention the challenges that still need to be resolved before widespread usage of Machine Learning can be employed in Official Statistics, citing:

- methodology concerning the human annotation of data:
- sampling the population to obtain representative training sets;
- using stratification in the context of Machine Learning algorithms;
- data structure engineering and selection to increase the transparency of models;
- reducing spurious correlations;
- methodology for studying causation;
- correcting the bias caused by the Machine Learning model;
- dealing with concept drift (representatively over time).

In the next two sections of this overview, I outline two case studies that highlight an exemplary use of a Machine Learning application, one case study in survey data collection and the other case study in survey post-processing. I conclude in Section 4 with some final thoughts on adapting Machine Learning into statistical systems for the production of Official Statistics at NSIs and introduce the papers presented in Session 4 of the Workshop.

2. Case study I

In the paper by Chen *et al.* (2022), Machine Learning techniques were used in a study to predict web breakoff in a repeated, cross-sectional non-probability online web survey administered to members of the 'Lightspeed Panel', an opt-in web panel in the United States.

The first wave was conducted between September and October 2019 while the second wave was collected in October 2020. Both waves performed similarly with about a 17% rate of breakoff, hence the first wave was used as the training data and the second wave the test data evaluated through cross-validations. The survey included information on the last question that the respondent completed, making it appropriate to analyse under a survival model. The Machine Learning technique under the survival model approach was the LASSO-Cox survival model and this was compared to the standard Cox survival model. The authors also ignored the clustering of questions within individuals to use more standard Machine Learning prediction models, including the LASSO-logistic regression model, random forest, gradient boosting, and support vector machine and these were compared to the standard logistic regression model. The authors also looked at different sets of covariates in the models: Demographics (age, education, ethnicity, student status, marital status); concurrent (responding device, item missing, matrix question, open-ended question, question topic, and question word count); and cumulative (as above but aggregated across questions, and included the number of times the respondent logged into the survey).

Results showed that under the survival models, the traditional Cox model performed better than the Machine Learning LASSO-Cox model in predicting breakoffs. Ignoring the clustering effects and using Machine Learning prediction models, the gradient boosting method provided the best prediction performance across all evaluation metrics that were used: Sensitivity, AUC, Accuracy, Specificity and Precision. Comparing the gradient boosting to the traditional Cox survival model, the gradient boosting performed slightly better in the AUC metric, showing that accounting for the clustered data structure did not necessarily translate into a significant improvement in break-off prediction. The findings also showed that using values of time-varying predictors concurrent to the breakoff status was more predictive of breakoff, compared to aggregating their values from the beginning of the survey, implying that respondents' breakoff behaviour is more driven by the current response burden.

3. Case study II

In a paper by Evans and Oyarzum (2021) and internal communication, the authors study the process of automatic coding to predict occupation, economic activity and other classifications at Statistics Canada. Some NSIs, including Statistics Canada, are investigating the use of the neural network library for word embeddings and text classification created by Facebook's AI Research lab called 'fastText'. The neural network fastText has the advantage that it works on word and n-gram embeddings and provides a score on the prediction confidence that allows for the planning of a thorough quality assurance process based on drawing samples proportional to the level of confidence in the coding. The aim is to move to 100% automatic coding together with optimal sampling methods specifically designed for quality assurance where the samples are presented to human coders for verification. One proposed sampling design is stratified sampling where samples are drawn from bins defined by the prediction confidence scores with an optimal sample allocation constrained to the desired level of accuracy, costs and maximum workload. The verification of the samples provides an output prediction error rate and also allows for updating the labelled training data to mitigate risks of future algorithmic biases. In a simulation by Statistics Canada on the coding of occupation, out of a 121,000 workload, there was a savings of about 25% in the manual coding needed under the new approach compared to the traditional approach with approximately the same level of prediction error rates.

Nevertheless, some caveats and lessons learnt based on experiences at NSIs need to be considered prior to moving to production platforms:

- the use of the 'black box' fastText requires a good understanding of how the algorithm works, for example, how does sorting the labelled data impact on the quality of the coding;
- there is a need to maintain the skills of human coders at NSIs and preserve this knowledge since new quality assurance approaches rely heavily on these skills;
- processes need to be put in place regarding how to ensure high-quality data streams and up-to-date training data sets, how to monitor model decay once deployed, and how to develop user interfaces;
- there is a need to assess and adapt the quality framework for automatic coding, particularly with respect to five criteria: explainability (understanding what causes a model to make particular decisions), accuracy, reproducibility, timeliness, cost effectiveness.

4. Final thoughts and introduction to Session 4

Given the growing prevalence of Machine Learning in survey research, it is imperative for statistical societies, such as the International Association of Survey Statisticians (IASS), to explore both scientific and ethical uses of these new and emerging techniques in our statistical production systems and how they can be applied. The IASS has a leadership role to play towards the creation of a survey culture for Trusted Smart Statistics under current complex data systems. Hence, the IASS has been monitoring cutting-edge Data Science research, and the understanding and application of Machine Learning techniques.

More specifically, the IASS is reviewing and advising on current best practices, when Machine Learning techniques are applicable and how to avoid pitfalls when these techniques may be misused. Our current recommendations around Machine Learning in the survey production pipeline for Official Statistics include:

- NSIs need to develop skills training and capacity building in Data Science with a focus on Machine Learning techniques among their workforce;
- there should be in-house expertise within the NSI to evaluate emerging Machine Learning techniques and the NSIs need to take the lead in their development;
- at the moment, current applications of Machine Learning are generally well suited for data management, visualisation, exploration and predictions, but research is only in the development stages for allowing statistical inferences;
- when discussing potential applications of Machine Learning within the NSI, all parties need to be engaged in these discussions: data scientists, methodologists, and subject matter experts;
- NSIs should start with small projects demonstrating proof of concept and the willingness of the organisation to try new methods. These can then be followed by taking Machine Learning applications into the production pipeline;
- NSIs should get involved with international research collaborations;
- NSIs need to develop new and robust quality frameworks for the use of Machine Learning applications in Official Statistics.

With this overview, I am pleased to introduce Session 4 sponsored by the IASS of the Second Istat Workshop on Methodologies for Official Statistics, titled: *Machine Learning Methods in Survey Statistics*. The session includes the following presentations:

1. *On the use of Machine Learning methods for the treatment of unit nonresponse in surveys* by David Haziza, John Tsang, Khaled Larbi and Mehdi Dagdoug with a discussion;
2. *State of play and perspectives on Machine Learning at Istat* by Marco Di Zio;
3. *Machine Learning in Official Statistics: Towards statistical based Machine Learning* by Marco Puts and Petrus J.H. Daas with a discussion.

References

Buskirk, T. D., A. Kirchner, A. Eck, and C. S. Signorino. 2018. "An Introduction to Machine Learning Methods for Survey Researchers". *Survey Practice.* Volume 11, N.1. https://doi.org/10.29115/SP-2018-0004.

Chen, Z., A. Cernat, and N. Shlomo. 2022. "Predicting Web Survey Breakoffs Using Machine Learning Models". *Social Science Computer Review*, Volume 41, N. 2: 573-591.

Evans, J. and J. Oyarzun. 2021. "Need for Speed: Using fastText (Machine Learning) to Code the Labour Force Survey". *Proceedings of Statistics Canada Symposium 2021, Adopting Data Science in Official Statistics to Meet Society's Emerging Needs.* https://www150.statcan.gc.ca/n1/en/pub/11-522-x/2021001/article/00013-eng.pdf?st=W0-FxeD7.

Kern, C., S. Eckman, J. Beck, R. Chew, B. Ma, and F. Kreuter. 2023. "Annotation Sensitivity: Training Data Collection Methods Affect Model Performance". In *Findings of the Association for Computational Linguistics: EMNLP 2023*: 14874-14886. Stroudsburg, PA, U.S.: Association for Computational Linguistics (ACL). https://aclanthology.org/2023.findings-emnlp.992.pdf.

Puts, M. J. H. and P. J. H. Daas. 2021. "Machine Learning from the Perspective of Official Statistics". *The Survey Statistician,* Volume 84: 12-17. http://isi-iass.org/home/wp-content/uploads/Survey_Statistician_2021_July_N84_02.pdf.

On the use of Machine Learning methods for the treatment of unit nonresponse in surveys

Khaled Larbi, John Tsang, David Haziza, Mehdi Dagdoug [1]

Abstract

In recent years, there has been a significant interest in Machine Learning in national statistical offices. Thanks to their flexibility, these methods may prove useful at the nonresponse treatment stage. In this article, we conduct an empirical investigation in order to compare several Machine Learning procedures in terms of bias and efficiency. In addition to the classical Machine Learning procedure, we assess the performance of ensemble approaches that make use of different Machine Learning procedures to produce a set of weights adjusted for nonresponse.

Keywords: Aggregation procedure; efficiency; nonresponse bias; propensity score estimation.

Introduction

In the last two decades, response rates have been steadily declining in medium to large-scale surveys conducted by National Statistical Offices. Consequently, there is growing concern regarding the potential nonresponse bias. Unit nonresponse, where no information is available for any of the survey variables, is typically treated by some form of weight adjustment procedure. The underlying principle behind weight adjustment is to inflate the weight of respondents in such a way that they effectively represent the nonrespondents. The inflation factor is defined as the inverse of the estimated response probability.

The treatment of unit nonresponse starts with formulating a nonresponse model, describing the relationship between the response indicators (equal to 1 for respondents and 0 for nonrespondents) and a vector of explanatory variables. Determining a suitable model also consists of selecting of a vector of explanatory variables that are both predictive of the response indicators and related to the survey variables; see Haziza and Beaumont (2017) for a discussion.

In recent years, there has been a growing interest within National Statistical Offices in the application of Machine Learning techniques in the context of weighting for unit nonresponse. Some reasons for the popularity of Machine Learning procedures include: (i) Machine Learning models can automatically learn and adapt from data, reducing the need for manual intervention. (ii) They can capture complex, non-linear relationships between variables that

[1]Khaled Larbi (khaled.larbi@insee.fr), INSEE, France; John Tsang (john.tsang@uottawa.ca), University of Ottawa, Canada; David Haziza (dhaziza@uottawa.ca), University of Ottawa, Canada; Mehdi Dagdoug (mehdi.dagdoug@mcgill.ca), McGill University, Canada.

may be difficult to model using traditional parametric procedures such as logistic regression. (iii) A number of Machine Learning algorithms are known for their excellent predictive performance. However, one should exercise some caution when Machine Learning procedures are used for the treatment of unit nonresponse because the survey statistician faces an estimation problem rather than a prediction problem. If the aim lies in estimating a finite population total/mean, the most predictive nonresponse model may not necessarily yield to the best estimator in terms of mean square error. This is somewhat different from what is encountered in the context of imputation for item nonresponse, whereby highly predictive procedures are expected to produce accurate estimates of population totals/means.

In this article, we investigate the use of Machine Learning procedures for estimating the response probabilities. We illustrate through an empirical study that a highly predictive procedure may lead to poor estimates in terms of mean square error; see Section 2. In Section 3, we conduct an extensive simulation study to assess the performance of adjusted estimators in terms of bias and efficiency. Other empirical investigations on the use of Machine Learning in the context of unit nonresponse for survey data can be found in (Phipps and Toth 2005; Lohr *et al.* 2015; Gelein 2017; Kern *et al.* 2019). In Section 4, we describe a number of aggregation procedures whereby the predictions produced by multiple Machine Learning procedures are combined to construct a suitable aggregate. The performance of aggregation procedures is assessed in terms of bias and efficiency. Finally, make some final remarks in Section 5.

1. Preliminaries

Consider a finite population $\mathcal{U}$ of size N; i.e., $\mathcal{U} = \{1, \ldots, k, \ldots, N\}$. The aim is to estimate the population total of a survey variable y, $t_y := \sum_{k \in \mathcal{U}} y_k$. To that end, we select a sample S, of size n, according to a sampling design, $P(S \mid \mathbf{Z})$, with first-order inclusion probabilities $\pi_k, k \in U$, where $\mathbf{Z}$ denotes the matrix of design information. In the absence of nonsampling errors, a design-unbiased estimator of t_y is the well-known Horvitz–Thompson estimator:

$$\widehat{t}_{y,\pi} = \sum_{k \in \mathcal{S}} d_k y_k, \tag{1}$$

where $d_k = 1/\pi_k$ denotes the design (basic) weight attached to unit k.

In the presence of unit nonresponse, the survey variable y is collected for a subset $\mathcal{S}_r \subset \mathcal{S}$. Let R_k be a response indicator attached to unit k such that $R_k = 1$ if unit k responds to the survey, and $R_k = 0$, otherwise. Let $p_k \equiv P(R_k = 1 \mid y_k, \mathbf{x}_k, k \in \mathcal{S})$ denote the response probability associated with unit k, where $\mathbf{x}_k$ denotes a vector of fully observed variable attached to unit k. We make the following assumptions: (i) The response indicators R_k are mutually independent, $k = 1, \ldots N$; (ii) The response indicators R_k are independent of the sample selection indicators I_k, where $I_k = 1$ if $k \in \mathcal{S}$, and $I_k = 0$, otherwise. This assumption implies that the response probability of a unit is essentially determined by fixed respondent characteristics. In the context of adaptive collection designs (Groves and Heeringa 2006), this assumption may be violated. (iii) The positivity assumption is satisfied;

i.e., $\pi_k > 0$ for all k and $p_k > 0$ for all k.

An unadjusted estimator of t_y is given by:

$$\widehat{t}_{y,un} = N \frac{\sum_{k \in \mathcal{S}} d_k R_k y_k}{\sum_{k \in \mathcal{S}} d_k R_k} \equiv N \widehat{\overline{Y}}_r. \tag{2}$$

The nonresponse error of $\widehat{t}_{y,un}$ defined as the difference between the unadjusted estimator and the full sample estimator, can be expressed as:

$$\widehat{t}_{y,un} - \widehat{t}_{y,\pi} = N \left\{ \frac{\widehat{N}_m}{\widehat{N}_\pi} \left(\widehat{\overline{Y}}_r - \widehat{\overline{Y}}_m \right) \right\}, \tag{3}$$

where $\widehat{N}_m = \sum_{k \in S} d_k (1 - R_k)$, $\widehat{N}_\pi = \sum_{k \in S} d_k$, and

$$\widehat{\overline{Y}}_m = \frac{\sum_{k \in S} d_k (1 - R_k) y_k}{\sum_{k \in S} d_k (1 - R_k)}$$

denotes the (unfeasible) mean of the nonrespondents. The term $\widehat{N}_m / \widehat{N}_\pi$ in (3) can be viewed as an estimate of the nonresponse rate. Alternatively, the population size N in (2) may be replaced by the estimated population size $\widehat{N}_\pi$. When the data are Missing Completely At Random (MCAR), we have $\mathbb{E}\left(\widehat{\overline{Y}}_r - \widehat{\overline{Y}}_m \right) \approx 0$, and $\widehat{t}_{y,un}$ would be virtually unbiased. However, the bias may be significant if the nonresponse rate is high and/or the behaviour of the respondents differ systematically from that of the nonrespondents in terms of the y-variable.

Turning to adjusted estimators, assuming that the response probabilities p_k are known, an unbiased estimator of t_y is the so-called double expansion estimator (Särndal *et al.* 1992):

$$\widehat{t}_{y,DE} = \sum_{k \in \mathcal{S}} \frac{d_k}{p_k} R_k y_k. \tag{4}$$

In practice, the p_k's are unknown and are replaced with estimated response probabilities $\widehat{p}_k$. More specifically, we start by postulating the following nonresponse model:

$$\mathbb{E}(R_k \mid y_k, \mathbf{x}_k) = p(\mathbf{x}_k), \tag{5}$$

where $p(\cdot)$ is given function. In the case of a parametric procedure (*e.g.* logistic regression), the function $m(\cdot)$ is predetermined, whereas it is left unspecified in the case of nonparametric and Machine Learning procedures.

An adjusted estimator of t_y is the propensity score-adjusted estimator given by:

$$\widehat{t}_{y,PSA} = \sum_{k \in \mathcal{S}} \frac{d_k}{\widehat{p}(\mathbf{x}_k)} R_k y_k, \tag{6}$$

where $\widehat{p}(\mathbf{x}_k)$ denotes the fitted value attached unit to $k \in S_r$. The weights adjusted for nonresponse are denoted by $w_k^* = d_k/\widehat{p}(\mathbf{x}_k)$, $k \in S_r$. The nonresponse error of $\widehat{t}_{y,PSA}$ can be expressed as:

$$\widehat{t}_{y,PSA} - \widehat{t}_{y,\pi} = (\widehat{t}_{y,DE} - \widehat{t}_{y,\pi}) - \sum_{k\in\mathcal{S}} \frac{d_k}{\widehat{p}(\mathbf{x}_k)} R_k y_k \left(\frac{\widehat{p}(\mathbf{x}_k) - p_k}{p_k}\right). \tag{7}$$

Since $\mathbb{E}(\widehat{t}_{y,DE} - \widehat{t}_{y,\pi}) = 0$, the estimator $\widehat{t}_{y,PSA}$ is virtually unbiased for t_y if

$$\mathbb{E}\left\{\sum_{k\in\mathcal{S}} \frac{d_k}{\widehat{p}(\mathbf{x}_k)} R_k y_k \left(\frac{\widehat{p}(\mathbf{x}_k) - p_k}{p_k}\right)\right\} \approx 0.$$

An alternative adjusted estimator of t_y is the so-called Hájek estimator:

$$\widehat{t}_{y,H} := N \frac{\sum_{k\in\mathcal{S}} \frac{d_k}{\widehat{p}(\mathbf{x}_k)} R_k y_k}{\sum_{k\in\mathcal{S}} \frac{d_k}{\widehat{p}(\mathbf{x}_k)} R_k}. \tag{8}$$

If the nonresponse model is correctly specified, we expect that $\mathbb{E}(\sum_{k\in\mathcal{S}} \frac{d_k}{\widehat{p}(\mathbf{x}_k)} R_k) \approx N$, which implies that both $\widehat{t}_{y,PSA}$ and $\widehat{t}_{y,H}$ would exhibit the same asymptotic bias. However, they may differ significantly in terms of variance, even in the absence of bias.

2. Estimation vs. prediction

In this section, we illustrate empirically that the most predictive model does not necessarily yield the best estimator of t_y in terms of mean square error. Indeed, including predictors that are highly predictive of R_k may lead to very small estimated response probabilities $\widehat{p}_k$, which may result in extreme adjusted weights w_k^*. In this case, both (6) and (8) may be inefficient. How then to choose the $\mathbf{x}_k$ variables to incorporate in the model? A common recommendation is to include the variables $\mathbf{x}_k$ that are related to both the indicator variable R_k and the survey variable y; *e.g.* Little and Vartivarian (2005), Beaumont (2005) and Kim *et al.* (2019). When an x-variable exhibits a strong correlation with R_k but is unrelated to y, excluding it from the nonresponse model is advisable. Indeed, including such a variable would not effectively mitigate nonresponse bias, but could potentially lead to a significant increase in the variance of the adjusted estimator.

To illustrate this point, we conducted a limited simulation study. We generated a finite population $\mathcal{U}$ of size $N = 10,000$ with seven variables: one survey variable y and six auxiliary variables $x_1, x_2, \ldots, x_6$. We first generated the x-variables according to the following distributions: $x_1 \sim$ Gamma$(5, 1)$; $x_2 \sim$ Gamma$(1, 5)$; $x_3 \sim$ Gamma$(1, 6)$; $x_4 \sim$ Gamma$(1, 10)$; $x_5 \sim$ Gamma$(1, 20)$; $x_6 \sim$ Gamma$(0.5, 50)$. Given x_1-x_6, we generated the y-variable according to the linear regression model:

$$y_k = 2 - 2x_{1k} + 4x_{2k} + \epsilon_k,$$

where the errors ϵ_k were generated from a normal distribution with mean equal to zero and variance equal to 225. This led to a model R^2 approximately equal to 0.64.

From the population, we selected $10,000$ samples, of size $n = 1,000$, according to simple random sampling without replacement. In each sample, each unit was assigned a response probability p_k:

$$p_k = 0.05 + \frac{0.95}{1 + \exp\left(-0.05x_{1k} + 0.05x_{2k} - 0.05x_{3k} + 0.05x_{4k} - 0.05x_{5k} + 0.02x_{6k}\right)}.$$

This led to a response rate of about 55% in each sample. The response indicators R_k were generated using a Bernoulli distribution with probability p_k.

Our goal was to estimate the population total of the y-values, $t_y = \sum_{k \in \mathcal{U}} y_k$. In our experiment, the variables x_1-x_6 were fully observed, while the y-variable was prone to missing values.

In each sample, we computed two estimators of t_y:

(i) the naive estimator given by (2);

(ii) the propensity score-adjusted estimator, $\widehat{t}_{y,PSA}$ given by (6), where $\widehat{p}(\mathbf{x}_k)$ was obtained using (i) the score method (see Section 2.1) based on different subsets of x_1-x_6, and regression trees (see Section 2.2) based on different subsets of x_1-x_6.

As a measure of bias of an estimator $\widehat{t}$, we computed the Monte Carlo percent relative bias:

$$\text{RB}_{MC}(\widehat{t}) = 100 \times \frac{1}{10,000} \sum_{b=1}^{10,000} \frac{(\widehat{t}_{(b)} - t_y)}{t_y}, \tag{9}$$

where $\widehat{t}_{(b)}$ denotes the estimator $\widehat{t}$ in the bth sample, $b = 1, \ldots, 10,000$. We also computed the Monte Carlo relative efficiency of $\widehat{t}$, using the full sample estimator $\widehat{t}_{y,\pi}$ given by (1), as the reference:

$$\text{RE}_{MC}(\widehat{t}) = 100 \times \frac{\text{MSE}_{MC}(\widehat{t})}{\text{MSE}_{MC}(\widehat{t}_{y,\pi})}, \tag{10}$$

where

$$\text{MSE}_{MC}(\widehat{t}) = \frac{1}{10,000} \sum_{b=1}^{10,000} \left(\widehat{t}_{(b)} - t_y\right)^2$$

and $\text{MSE}_{MC}(\widehat{t}_{y,\pi})$ is similarly defined.

In each sample, we also computed the Monte Carlo percent coefficient of variation of the adjusted weights $w_k^* = d_k/\widehat{p}(\mathbf{x}_k)$:

$$\text{CV}_{MC}(w_k^*) = 100 \times \frac{1}{B}\sum_{b=1}^{B}\frac{s_{w^*(b)}}{\overline{w}^*_{(b)}},$$

where

$$s_{w^*} = \sqrt{\frac{1}{n_r - 1}\sum_{k\in S_r}(w_k^* - \overline{w}^*)^2}$$

with $\overline{w}^* = n_r^{-1}\sum_{k\in S_r} w_k^*$. Finally, we computed the Monte Carlo mean square error of the predictions defined as:

$$\text{MSE}_{MC}(\widehat{p}) = 100 \times \frac{1}{B}\sum_{b=1}^{B}\frac{1}{n_r}\sum_{k\in S_r}\left(\widehat{p}_{(b)}(\mathbf{x}_k) - p_k\right)^2,$$

where $\widehat{p}_{(b)}(\mathbf{x}_k)$ denotes the estimated response probability attached to unit k in the bth sample.

2.1 The score method

The score method (Little and Vartivarian 2005; Eltinge and Yansaneh 1997; Haziza and Beaumont 2007) may be described as follows:

Step 1: Obtain preliminary estimated response probabilities, $\widehat{p}^{LR}(\mathbf{x}_k)$, $k \in S$, from a logistic regression.

Step 2: Form C classes based on the estimated response probabilities, $\widehat{p}^{LR}(\mathbf{x}_k)$, using an equal quantile method. We set $C = 20$, which led to classes of size 50.

Step 3: Adjust the weight of the respondents within a class by multiplying their design weight d_k by the inverse of the response rate observed within the sane class.

Table 2.1 - Monte Carlo measures for several estimators of t_y: The score method

Estimator	$\widehat{t}_{y,naive}$	$\widehat{t}_{y,PSA}$ x_1	$\widehat{t}_{y,PSA}$ x_1-x_2	$\widehat{t}_{y,PSA}$ x_1-x_3	$\widehat{t}_{y,PSA}$ x_1-x_4	$\widehat{t}_{y,PSA}$ x_1-x_5	$\widehat{t}_{y,PSA}$ x_1-x_6
$\text{RB}_{MC}(\widehat{t})$ in (%)	-13.4	-12.2	-0.2	-0.8	-0.3	-1.0	-0.4
$\text{RE}_{MC}(\widehat{t})$	623	561	134	141	142	161	206
$\text{CV}_{MC}(w*)$ in (%)	0	13	16	19	30	50	84
$\text{MSE}_{MC}(\widehat{p})$	4.7	5.0	4.9	4.6	4.1	1.3	0.4

Source: Own computation

The results for the score method, displayed Table 2.1, can be summarised as follows:

- As expected, the naive estimator was biased with a relative bias of -13.4%. This is not surprising as the naive estimator makes no use of the variables x_1 and x_2, which are related to both R_k and y.
- The propensity score estimator $\widehat{t}_{y,PSA}$ based on the variable x_1 exhibited a smaller bias than the naive estimator, which can be explained by the fact that it incorporated the variable x_1. The remaining bias is due to the non-inclusion of x_2 in the nonresponse model.
- The propensity score estimator $\widehat{t}_{y,PSA}$ based on the variable x_1 and x_2 was nearly unbiased bias (-0.2%) as it included both x_1 and x_2 in the nonresponse model. In terms of relative efficiency, this estimator was the best, with a value of RE equal 134. It is worth noting that the other propensity score estimators were nearly unbiased but were less efficient than $\widehat{t}_{y,PSA}$ based on x_1 and x_2. In other words, adding x_3 to x_6 to the model did not impact the bias but did lead to an increase in variance.
- The most predictive model of R_k is the one that included the variables x_1-x_6. However, except for $\widehat{t}_{y,PSA}$, based on x_1 only, the estimator $\widehat{t}_{y,PSA}$ based on x_1-x_6 was the worst in terms of relative efficiency, with a value of RE equal to 209. In comparison with $\widehat{t}_{y,PSA}$, based on x_1 and x_2, this corresponds to a 55% increase in terms of mean square error. This result suggests that the most predictive model may not necessarily translate into the best estimator of t_y. In fact, a quick look at the values of $\text{MSE}_{MC}(\widehat{p})$ shows that the model that incorporates the variables x_1-x_6 led to the smallest value of $\text{MSE}_{MC}(\widehat{p})$ (about 0.4), whereas the model that incorporated x_1 and x_2 led to a value of $\text{MSE}_{MC}(\widehat{p})$ of 4.9, which is about 12 times larger.
- A large dispersion of the adjusted weights w_k^* led to estimators with a large variance. This is why, in practice, limiting the dispersion of the adjusted weights w_k^* is desirable.

2.2 Regression trees

We repeated the simulation experiment with regression trees using the same setup described in Section 2.1. The simulation study was conducted using the R package rpart. Regression trees require the specification of some hyper-parameters such as the complexity parameter, denoted by c_p, and the minimal number of observations per terminal node, denoted by n_0. We used different values of c_p: 0; 0.001; and 0.01 (the default value). We also used two values for n_0: 10 and 25. With of value of c_p set to 0.001 (say), any split that does not decrease the overall lack of fit by a factor of 0.001 is not attempted. Large values of c_p will thus lead to shallower trees.

Results for $n_0 = 10$ and $n_0 = 25$ are shown in Table 2.2 and Table 2.2, respectively. They can be summarised as follows:

- for $n_0 = 10$, we note that the estimator $\widehat{t}_{y,PSA}$, based on x_1 and x_2, was nearly unbiased for $c_p = 0$ and $c_p = 0.001$. However, the bias of $\widehat{t}_{y,PSA}$ increased as more variables were incorporated in the tree procedure. For instance, for $c_p = 0$, the estimator $\widehat{t}_{y,PSA}$, based on x_1 and x_2, showed a value of relative bias of about -0.6%, whereas the estimator $\widehat{t}_{y,PSA}$, based on x_1-x_6 showed a relative bias of about -6.5%. The same was true

for all values of c_p. This can be explained by the fact that, as the number of predictors increased, the fraction of splits that involved either x_1 or x_2 (the variables associated with both R_k and y) diminished. For instance, for $c_p = 0$ and only x_1 and x_2 were used as predictors, 100% of the splits used either x_1 or x_2. But when all the variables x_1-x_6 were included, only 16.8% of the splits used x_1, and 13.5% of the splits used x_2. In other words, above 70% of the splits did not use either x_1 or x_2;

- with an increasing value of c_p, the tree became progressively shallower, which led to larger biases. For instance for $c_p = 0$, the estimator $\widehat{t}_{y,PSA}$ based on x_1 and x_2, showed a value of RB equal to -0.6%, whereas it was equal to -8.0% for $c_p = 0.01$. Fewer terminal nodes limit the tree's ability to capture local behaviour effectively;
- results for $n_0 = 25$ followed similar patterns as those obtained for $n_0 = 10$, except that the propensity score estimator was biased in all the scenarios;
- like the score method, the value of $\text{MSE}_{MC}(\widehat{p})$ decreased as more predictors were incorporated in the model. Similarly, the dispersion of the adjusted weights w_k^* increased as more predictors were included.

2.3 Discussion

In Sections 2.1 and 2.2, we performed propensity score estimation based on the score method and regression trees, respectively. For regression trees, the bias of $\widehat{t}_{y,PSA}$ increased as more predictors were included in the model. This pattern was not observed for the score method. This can be explained by the fact that, for the score method, the weighting classes were based on the preliminary score $\widehat{p}^{LR}(\mathbf{x}_k)$ that can be viewed as a scalar summary of all the information contained in x_1-x_6. Therefore, the sample partitions obtained through the score method implicitly made use of all the predictors, and in particular x_1 and x_2. This is why $\widehat{t}_{y,PSA}$ was virtually unbiased as long as at least both x_1 and x_2 were included. For regression trees, the situation is more intricate. Indeed, when all the predictors x_1-x_6 were included, we ended up with trees that made use of x_1 and x_2 for a fraction of the splits. As a result, we were not able to eliminate the nonresponse bias as effectively.

These results suggest we should exercise caution if variable selection is performed prior to nonresponse adjustment. Indeed, if the variable selection method resulted in the elimination of some important predictors (which are those that are related to both R_k and y) in the presence of other predictors that are highly related to R_k but not to y, the propensity score-adjusted estimator may likely suffer from an appreciable bias.

3. Simulation study

We conducted an extensive simulation study to assess the performance of several Machine Learning procedures (see Section 3.2 below) in terms of bias and efficiency.

Table 2.2 - Monte Carlo measures for several estimators of t_y: regression trees with $n_0 = 10$

	RB$_{MC}(\widehat{t})$ in (%)	RE$_{MC}(\widehat{t})$ in (%)	MSE$_{MC}(\widehat{p})$	CV$_{MC}(w*)$ in (%)
		$c_p = 0$		
$\widehat{t}_{y,PSA}$ x_1	-11.1	572	4.0	29
$\widehat{t}_{y,PSA}$ x_1-x_2	-0.6	116	4.3	36
$\widehat{t}_{y,PSA}$ x_1-x_3	-1.7	140	3.9	43
$\widehat{t}_{y,PSA}$ x_1-x_4	-2.6	162	3.8	48
$\widehat{t}_{y,PSA}$ x_1-x_5	-4.1	206	3.4	53
$\widehat{t}_{y,PSA}$ x_1-x_6	-6.5	318	2.9	62
		$c_p = 0.001$		
$\widehat{t}_{y,PSA}$ x_1	-11.2	577	3.9	29
$\widehat{t}_{y,PSA}$ x_1-x_2	-0.7	117	4.2	36
$\widehat{t}_{y,PSA}$ x_1-x_3	-1.8	142	3.8	43
$\widehat{t}_{y,PSA}$ x_1-x_4	-2.8	164	3.7	48
$\widehat{t}_{y,PSA}$ x_1-x_5	-4.1	209	3.3	53
$\widehat{t}_{y,PSA}$ x_1-x_6	-6.6	322	2.9	62
		$c_p = 0.01$		
$\widehat{t}_{y,PSA}$ x_1	-13.7	802	3.0	5
$\widehat{t}_{y,PSA}$ x_1-x_2	-8.0	414	3.0	14
$\widehat{t}_{y,PSA}$ x_1-x_3	-7.3	360	2.9	23
$\widehat{t}_{y,PSA}$ x_1-x_4	-7.3	341	2.8	33
$\widehat{t}_{y,PSA}$ x_1-x_5	-7.8	364	2.6	39
$\widehat{t}_{y,PSA}$ x_1-x_6	-10.0	519	2.4	49

Source: Own computation

Table 2.3 - Monte Carlo measures for several estimators of t_y: Regression trees with $n_0 = 25$

	$RB_{MC}(\widehat{t})$ in (%)	$RE_{MC}(\widehat{t})$ in (%)	$MSE_{MC}(\widehat{p})$	$CV_{MC}(w*)$ in (%)
		$c_p = 0$		
$\widehat{t}_{y,PSA}$ x_1	-11.6	608	3.1	15
$\widehat{t}_{y,PSA}$ x_1-x_2	-3.1	168	3.1	20
$\widehat{t}_{y,PSA}$ x_1-x_3	-4.6	210	2.8	26
$\widehat{t}_{y,PSA}$ x_1-x_4	-5.9	263	2.7	29
$\widehat{t}_{y,PSA}$ x_1-x_5	-7.4	337	2.5	33
$\widehat{t}_{y,PSA}$ x_1-x_6	-10.0	514	2.2	41
		$c_p = 0.001$		
$\widehat{t}_{y,PSA}$ x_1	-11.8	625	3.1	14
$\widehat{t}_{y,PSA}$ x_1-x_2	-3.4	174	3.1	19
$\widehat{t}_{y,PSA}$ x_1-x_3	-4.7	214	2.8	26
$\widehat{t}_{y,PSA}$ x_1-x_4	-6.0	268	2.7	29
$\widehat{t}_{y,PSA}$ x_1-x_5	-7.4	341	2.5	33
$\widehat{t}_{y,PSA}$ x_1-x_6	-10.1	517	2.2	41
		$c_p = 0.01$		
$\widehat{t}_{y,PSA}$ x_1	-14.0	824	3.1	2
$\widehat{t}_{y,PSA}$ x_1-x_2	-9.2	489	3.0	9
$\widehat{t}_{y,PSA}$ x_1-x_3	-8.2	403	2.8	17
$\widehat{t}_{y,PSA}$ x_1-x_4	-8.7	419	2.7	24
$\widehat{t}_{y,PSA}$ x_1-x_5	-9.2	447	2.5	30
$\widehat{t}_{y,PSA}$ x_1-x_6	-11.6	632	2.3	38

Source: Own computation

Table 3.1 - Descriptive statistics of percent RE across the 36 scenarios: the propensity-score-adjusted estimator

ML procedure	Min	Q1	Median	Q3	Max	Mean
bart 1	144	194	280	635	1845	489
rf 2	130	211	281	660	2799	561
rf 1	131	213	282	657	2781	560
xgb 2	132	197	295	621	2054	515
rf 5	154	207	304	717	2331	576
xgb 1	172	215	326	653	2253	552
rf 4	157	212	329	782	2359	579
rf 3	158	213	330	784	2351	579
xgb 3	171	231	336	837	2227	589
xgb 4	178	238	338	719	2574	607
knn 1	174	243	346	778	2174	576
bart 2	169	215	359	853	2087	628
knn 2	157	219	360	740	3543	693
cart 20	132	255	490	716	1904	611
cart 50	139	242	504	867	2185	602
cart 30	130	240	508	704	1924	608
cart 40	132	238	509	785	2050	605
logit	145	216	521	1233	4948	952
logit lasso	149	221	553	1242	4556	898
mob	146	254	579	1355	5287	1037
cubist 2	128	339	614	1642	37936	3128
cubist 5	151	290	648	1368	24764	1978
cubist 4	151	290	655	1396	25358	2010
cubist 1	156	323	708	1612	29335	2287
score	318	746	1236	1811	20307	2495
svm 2	251	673	2188	11525	140425	20169
svm 1	251	669	2327	9823	96179	10414
cubist 3	312	4034	10242	35640	13988674	445022

Source: Own computation

Table 3.2 - Descriptive statistics of percent RE across the 36 scenarios: the Hájek estimator

ML procedure	Min	Q1	Median	Q3	Max	Mean
xgb 4	180	221	304	732	2912	599
bart 1	158	200	306	556	1710	478
bart 2	176	205	307	656	1743	522
xgb 1	175	209	307	643	2457	547
rf 4	174	205	314	729	2355	569
rf 3	173	205	315	729	2347	568
xgb 3	175	206	324	709	2447	577
xgb 2	159	199	325	572	2057	517
rf 5	167	215	326	770	2074	581
rf 2	170	203	328	657	2462	558
rf 1	170	204	330	656	2453	557
knn 1	179	223	337	628	1867	534
cart 50	148	211	368	602	2195	514
cart 40	141	216	380	621	2040	512
knn 2	202	238	385	818	3379	714
cart 30	140	220	400	629	1905	512
cart 20	146	237	402	621	1889	522
logit lasso	145	201	414	1031	1811	613
mob	141	213	456	1054	1793	648
logit	139	201	457	953	1903	607
cubist 2	147	293	522	882	3857	768
cubist 5	151	254	525	799	3262	713
cubist 4	152	256	527	799	3276	715
cubist 1	153	261	546	800	3348	729
score	224	505	723	1353	8356	1332
cubist 3	224	582	812	1183	4528	1106
svm 2	189	358	910	1401	5024	1161
svm 1	189	357	952	1482	4884	1122

Source: Own computation

3.1 The setup

We generated several finite populations of size $N = 50,000$. Each population consisted of a survey variable Y and seven auxiliary variables, four of which were continuous and the remaining being discrete. First, the continuous auxiliary variables were generated as follows: $X^{(s)} \sim$ Gamma$(3,2)$, $X^{(c_1)} \sim \mathcal{N}(0,1)$; $X^{(c_2)} \sim$ Gamma$(3,2)$ and $X^{(c_3)} \sim$ Gamma$(3,2)$. The discrete auxiliary variables were generated as follows: $X^{(d_1)} \sim \mathcal{MN}(N, 0.5, 0.05, 0.05, 0.1, 0.3)$; $X^{(d_2)} \sim$ Ber(0.5) and $X^{(d_3)} \sim$ UD$(1;5)$, with UD denoting the uniform discrete distribution. Two configurations for these predictors were used: (i) The predictors were independently generated; (ii) The predictors were generated through Gaussian copulas to produce a level of correlation among them.

Given the values of the auxiliary variables, we generated several y-variables according to the following two models:

$$y_k = \gamma_0 + \gamma_1^{(s)} X_{1k}^{(s)} + \gamma_1^{(c)} X_{1k}^{(c)} + \gamma_2^{(c)} X_{2k}^{(c)} + \gamma_3^{(c)} X_{3k}^{(c)} + \sum_{j=2}^{5} \gamma_{1j}^{(d)} (1_{\{X_{1k}^{(d)}=j\}}) + \gamma_2^{(d)} X_{2k}^{(d)} + \sum_{k=2}^{5} \gamma_{3j}^{(d)} (1_{\{X_{3k}^{(d)}=j\}}) + \varepsilon_k \tag{11}$$

and

$$y_k = \delta_1 X_{2k}^{(c)} + \delta_2 (X_{2k}^{(c)})^2 (1 - 1_{\{X_{3k}^{(d)}=2\} \cup \{X_{3k}^{(d)}=3\}}) + \log(1 + \delta_3 X_{2k}^{(c)}) (1_{\{X_{3k}^{(d)}=2\} \cup \{X_{3k}^{(d)}=3\}}) + \varepsilon_k, \tag{12}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. Model (11) is linear in the regression coefficients, whereas Model (12) is nonlinear.

Each population was partitioned into ten strata on the basis of the auxiliary variable $X^{(s)}$ using an equal quantile method. From each population, we selected $B = 5,000$ samples according to stratified simple random sampling without replacement of size $n = 1,000$ based on Neyman's allocation.

For the populations generated according to the linear model (11), we simulated the case of both a (virtually) non-informative sampling design and an informative sampling design. For the non-informative sampling design, the correlation between the y-variable and the design weights d_k was equal to 0.02, whereas it was equal to approximately 0.3 for the informative sampling design. For the non-informative sampling design, the vector of coefficients

$$\left(\gamma_0, \gamma^{(s)}, \gamma_1^{(c)}, \gamma_2^{(c)}, \gamma_3^{(c)}, \gamma_{12}^{(d)}, \gamma_{13}^{(d)}, \gamma_{14}^{(d)}, \gamma_{15}^{(d)}, \gamma_{22}^{(d)}, \gamma_{32}^{(d)}, \gamma_{33}^{(d)}, \gamma_{34}^{(d)}, \gamma_{35}^{(d)}\right)$$

was set to $(-0.2, 5.0, 5.0, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)$ and to $(-10, 5.0, 5.0, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)$ for the informative sampling design. Finally, for the nonlinear model (12), the vector of coefficients $(\delta_0, \delta_1, \delta_2, \delta_3)$ was set

to $(4, 4, 4, 4)$. This led to 6 different survey variables y.

In each sample, nonresponse to the survey variable Y was generated according to six nonresponse mechanisms. That is, for each $k \in \mathcal{S}$, we assigned a response probability p_k according to the following six models:

NR1: $p_k^{(1)} = \text{logit}^{-1}(-0.8 - 0.05X_{1k}^{(s)} + 0.2X_{1k}^{(c)} + 0.5X_{2k}^{(c)} - 0.05X_{3k}^{(c)} + \sum_{k=2}^{5} 0.2(1_{\{X_{1k}^{(c)}=k\}}) + 0.2X_{2k}^{(d)} + \sum_{k=2}^{5} 0.3(1_{\{X_{3k}^{(d)}=k\}}))$;

NR2: $p_k^{(2)} = 0.1 + 0.9\,\text{logit}^{-1}(0.5 + 0.3X_{1k}^{(s)} - 1.1X_{1k}^{(c)} - 1.1X_{2k}^{(c)} - 1.1X_{3k}^{(c)} + \sum_{k=2}^{5} 0.8(1_{\{X_{1k}^{(c)}=k\}}) + 0.8X_{2k}^{(d)} + \sum_{k=2}^{5} 0.8(1_{\{X_{3k}^{(d)}=k\}}))$;

NR3: $p_k^{(3)} = 0.1 + 0.9\,\text{logit}^{-1}\left(-1 + \text{sgn}\,(X_{1k}^{c})\,(X_{1k}^{c})^2 + 3 \times 1_{\left\{X_{1k}^{(d)}<4\right\}\cap\left\{X_{2k}^{(d)}=1\right\}}\right)$;

NR4: $p_k^{(6)} = 0.1 + 0.6\,\text{logit}^{-1}(0.85X_{1k}^{(s)} + 0.85X_{2k}^{(c)} - 0.85X_{3k}^{(c)} - \sum_{k=2}^{5} 0.2(1_{\{X_{1k}^{(c)}=k\}}) + 0.2X_{2k}^{(d)} - \sum_{k=2}^{5} 0.3(1_{\{X_{3k}^{(d)}=k\}}))$;

NR5: $p_k^{(4)} = 0.55 + 0.45\tanh\,(0.05y_k - 0.5)$;

NR6: $p_k^{(5)} = 0.1 + 0.9\,\text{logit}^{-1}\,(0.2y_k - 1.2)$.

The parameters in each nonresponse model were set so as to obtain a response rate approximately equal to 50% in each sample. The response indicators $R_k^{(j)}$ were generated from a Bernoulli distribution with probability $p_k^{(j)}$, $j = 1, \ldots, 6$. Note that the nonresponse mechanism NR1-NR4 involved x-variable only. Below, they will be referred to as ignorable mechanisms. The nonresponse mechanism NR5 and NR6 involved the y-variable. Below, they will be referred to as non-ignorable mechanisms. Overall, we ended up with $6 \times 6 = 36$ scenarios, each corresponding to a given survey variable and a given nonresponse mechanism. Out of the 36 scenarios, 24 were of the ignorable type, and 12 were of the non-ignorable type.

To estimate the response probabilities p_k, we used the following Machine Learning procedures based on the set of explanatory variables, $X^{(s)}$, $X_1^{(c)}$, $X_2^{(c)}$, $X_2^{(c)}$, $X_1^{(d)}$, $X_2^{(d)}$ and $X_3^{(d)}$:

(a) logistic regression:
   - `logit`.

(b) logistic regression with variable selection based on LASSO; *e.g.* see Hastie *et al.* (2001):
   - `logit_lasso` (the amount of penalisation λ was obtained using a 10-fold cross validation).

(c) classification and regression trees; see Breiman *et al.* (1984):
   - `cart20`: Unpruned trees, $c_p = 0$, at least 20 observations in each leaf.
   - `cart30`: Unpruned trees, $c_p = 0$, at least 30 observations in each leaf.
   - `cart40`: Unpruned trees, $c_p = 0$, at least 40 observations in each leaf.

  - `cart50`: Unpruned trees, $c_p = 0$, at least 50 observations in each leaf.

(d) Random forests; *e.g.* see Breiman (2001):
  - `rf1`: Probabilities estimation trees, at least 10 observations in each leaf, 100 trees.
  - `rf2`: Probabilities estimation trees, at least 10 observations in each leaf, 500 trees.
  - `rf3`: Probabilities estimation trees, at least 30 observations in each leaf, 100 trees.
  - `rf4`: Probabilities estimation trees, at least 30 observations in each leaf, 500 trees.
  - `rf5`: Probabilities estimation trees, at least 30 observations in each leaf, 500 trees, variable used for the allocation is selected with probability 1 at each split.

(e) k-nearest neighbors;
  - knn: k determined by 10-fold cross validation with $k \in \{3, 12\}$;
  - knn_reg: k determined by 10-fold cross validation with $k \in \{3, 30\}$.

(f) Bayesian additive regression tree; *e.g.* see Chipman *et al.* (2010).
  - `bart` Bart as a classification method with parameters described in Chipman *et al.* (2010) for all priors;
  - `bart_reg`: Bart as a regression method with parameters described in Chipman *et al.* (2010) for all priors.

(g) Extreme Gradient Boosting (XGBoost); see Chen and Guestrin (2016):
  - `xb1`: 500 trees, $\Gamma = 10$, proportion for subsets : 75 %, learning rate : 0.5, max depth: 2;
  - `xgb2`: 2000 trees, $\Gamma = 2$, proportion for subsets : 100 %, learning rate : 0.5, max depth : 2;
  - `xgb3`: 1000 trees, $\Gamma = 2$, proportion for subsets : 75 %, learning rate : 0.01, max depth : 1;
  - `xgb4`: 500 trees, $\Gamma = 10$, proportion for subsets : 75 %, learning rate : 0.05, max depth : 3.

(h) Support vector machine:
  - `svm1`: $\nu-$SVM with a Gaussian kernel, $\nu = 0.7$, $\gamma = 0.025$;
  - `svm2`: $\nu-$SVM with a linear kernel, $\nu = 0.7$.

(i) Cubist algorithm; see Quinlan (1992) and Quinlan (1993):
  - `cb1`: Unbiased, 100 rules, with extrapolation, 10 committees;
  - `cb2`: Unbiased, 100 rules, without extrapolation, 10 committees;
  - `cb3`: Biased, 100 rules, with extrapolation, 10 committees;
  - `cb4`: Unbiased, 100 rules, with extrapolation, 50 committees;
  - `cb5`: Unbiased, 100 rules, with extrapolation, 100 committees.

(j) Model-based recursive partitioning; see Zeileis *et al.* (2008):
  - `mob`: logit model fitted, $X^{(s)}$ for stratification.

This led to 28 Machine Learning procedures. In certain scenarios and with particular Machine Learning methods, we encountered situations where the estimated response probabilities either became exceedingly small or exceeded 1. To address this, we implemented

a truncation procedure, ensuring that these estimated response probabilities fell within the range of $[0.025, 1]$. The estimates that did not undergo truncation were then adjusted, so the sum of estimated response probabilities before truncation equivalent equal to the sum after truncation.

In each sample, we computed two estimators: (i) the propensity score-adjusted estimator, $\widehat{t}_{y,PSA}$ given by (6) and (ii) The Hájek estimator, $\widehat{t}_{y,H}$ given by (8). As a measure of bias of an estimator $\widehat{t}_y$, we computed its Monte Carlo percent relative bias given by (9). As a measure of efficiency, we computed the Monte Carlo relative efficiency, using the complete data estimator $\widehat{t}_{y,\pi}$, as the reference; see Expression (10).

3.2 Simulation results

Tables 3.1 and 3.2 show some Monte Carlo descriptive regarding the relative efficiency (RE) for the PSA and Hájek estimators, respectively, over all the 36 scenarios: the minimum (Min), the first quartile (Q1), the median (Median), the third quartile (Q3) and the maximum (Max). In Tables 3.1 and 3.2, the Machine Learning procedures are ordered from the best to the worst with respect to the median percent RE (the median of the 36 values of RE).

From Table 3.1, we note that three procedures stood out in terms of relative efficiency: BART, random forests, and XGboost. The commonly employed score method did not yield impressive results, with a median percent RE of about 1236. In the best-case scenario, it exhibited a minimum RE of 318, which was significantly higher than that of the best procedures that exhibited a minimum RE between 130 and 160. Similarly, in the worst case scenario, it exhibited a value of a maximum RE of 20307, which was considerable. In contrast, the best procedures exhibited a maximum RE ranging between 1800 and 2300 approximately. Finally, the procedures mob, cubist, and support vector machines performed the least favourably in our experiments. While we were unable to find a set of hyper-parameters for which they will work well, this does not mean that these methods would perform as poorly as they did for other sets of hyper-parameters.

Results for the Hájek estimator in Table 3.2 were similar to those for the PSA estimator. Again, the best Machine Learning procedures were: XGboost, BART, and random forests. These procedures had similar performances in terms of median percent RE. BART was especially good in the worst scenario with values of maximum percent RE equal to 1710 and 1743, which was significantly smaller than the corresponding values for XGboost and random forests. Again, the score method was outperformed by these three procedures in virtually all the scenarios.

4. Aggregation procedures

Aggregation procedures refer to techniques used to combine the predictions from multiple models into a single, more robust, and accurate prediction. These methods are commonly

Table 4.1 - Descriptive statistics of percent RE across the 24 ignorable scenarios: the propensity score estimator

ML procedure	Min	Q1	Median	Q3	Max	Mean
rf 3	158	208	227	338	1037	298
	(0.1)	(2.7)	(5.3)	(17.9)	(31.8)	(10.3)
Exponential weighting: $\mathcal{L}_{mis}$ (with splitting)	160	182	234	292	1143	294
	(0.5)	(4.0)	(11.7)	(20.5)	(38.4)	(13.2)
Exponential weighting: $\mathcal{L}_{mis}$ (without splitting)	159	182	235	292	1114	293
	(0.6)	(4.0)	(11.6)	(19.8)	(37.8)	(13.0)
Exponential weighting: $\mathcal{L}_{cross}$ (with splitting)	160	183	235	292	1169	296
	(0.5)	(4.0)	(11.3)	(19.4)	(37.3)	(12.8)
Exponential weighting: $\mathcal{L}_{cross}$ (without splitting)	159	182	236	292	1080	291
	(0.3)	(4.0)	(11.9)	(21.1)	(38.8)	(13.4)
xgb 1	172	210	245	332	775	288
	(0.8)	(2.9)	(7.6)	(16.9)	(23.8)	(9.7)
Linear weighting (with splitting)	170	207	246	329	889	308
	(0.0)	(2.2)	(6.9)	(14.6)	(22.0)	(8.6)
Linear weighting (without splitting)	159	181	250	349	2130	383
	(0.6)	(3.4)	(17.2)	(24.5)	(64.3)	(18.8)
knn 2	172	211	266	379	2192	410
	(3.1)	(6.3)	(18.2)	(31.6)	(66.9)	(21.1)
cart 50	170	226	348	515	901	381
	(0.0)	(0.5)	(3.0)	(5.1)	(25.9)	(4.4)
score	318	489	930	1329	11111	1712
	(0.6)	(3.9)	(14.0)	(21.8)	(44.3)	(15.7)

used in ensemble learning, where the goal is to improve a model's performance by leveraging multiple models' strengths (Nemirovski 2000). In the context of unit nonresponse, multiple Machine Learning procedures are used to obtain a set of estimated response probabilities for each sample unit. These probabilities are then combined in some way to obtain an aggregate score. Why use an ensemble method? In general, there is no Machine Learning procedures that outperform all the other competitors in all the scenarios. Indeed, Machine Learning procedures may do well in a particular scenario but not in another scenario. However, one cannot tell in advance which procedure will perform well for a specific scenario. An aggregation procedure may outperform a single procedure in terms of bias and efficiency; *e.g.* see Tsybakov (2003).

We describe two aggregation procedures for combining predictions from multiple models. Let $\widehat{p}_k^{(m)}(\mathbf{x}_k)$ be the estimated response probability attached to unit k obtained through the mth Machine Learning procedure $m = 1, \ldots, M$. For both aggregation procedures, the aggregate score for unit k is defined as:

$$\widehat{p}_k^{agg} = \sum_{m=1}^{M} \omega_m \widehat{p}_k^{(m)}(\mathbf{x}_k), \tag{13}$$

Table 4.2 - Descriptive statistics of percent RE across the 12 nonignorable scenarios: the propensity score estimator

ML procedure	Min	Q1	Median	Q3	Max	Mean
Exponential weighting: $\mathcal{L}_{cross}$ (without splitting)	150	573	765	1410	2335	1054
	(3.1)	(33.5)	(51.1)	(66.8)	(111.8)	(52.9)
Exponential weighting: $\mathcal{L}_{mis}$ (without splitting)	152	571	768	1423	2371	1060
	(3.3)	(34.2)	(51.6)	(66.4)	(111.9)	(53.1)
Exponential weighting: $\mathcal{L}_{mis}$ (with splitting)	157	576	773	1449	2425	1070
	(3.8)	(35.2)	(52.5)	(65.9)	(111.9)	(53.4)
Exponential weighting: $\mathcal{L}_{cross}$ (with splitting)	161	578	776	1465	2474	1078
	(4.2)	(35.2)	(53.1)	(65.5)	(112.1)	(53.7)
Linear weighting (without splitting)	158	555	792	1549	2913	1151
	(4.6)	(34.0)	(55.6)	(63.5)	(120.4)	(55.2)
Linear weighting (with splitting)	180	641	858	1333	2082	1046
	(7.4)	(33.9)	(51.9)	(68.5)	(108.3)	(53.4)
xgb 1	184	610	883	1348	2253	1080
	(7.8)	(34.0)	(52.3)	(70.5)	(113.4)	(54.9)
rf 3	204	762	904	1444	2351	1141
	(10.2)	(40.3)	(55.3)	(71.8)	(111.1)	(56.7)
knn 2	157	399	919	1711	3543	1260
	(2.4)	(24.9)	(58.7)	(64.5)	(128.6)	(56.3)
cart 50	139	783	971	1219	2185	1043
	(2.8)	(25.4)	(43.2)	(73.5)	(104.7)	(47.8)
score	767	1630	1816	3148	20307	4062
	(19.6)	(49.9)	(68.7)	(87.0)	(137.6)	(71.9)

such that $\omega_m \geq 0$ for all $m = 1, \ldots, M$, and $\sum_{m=1}^{M} \omega_m = 1$. That is, the aggregate score $\widehat{p}_k^{agg}$, can be viewed as a convex combination of the individual predictions obtained from each of the M models. Assuming that the estimated response probabilities $\widehat{p}_k^{(m)}(\mathbf{x}_k), m = 1, \cdots, M$, all lie between 0 and 1, the convex combination (13) ensures that the aggregate score $\widehat{p}_k^{agg}$ also lies between 0 and 1. Machine Learning procedures that perform well will be assigned a larger weight ω_m in the weighted average (13). The resulting aggregated PSA estimator is defined as:

$$\widehat{t}_{PSA,agg} := \sum_{k \in S} \frac{d_k}{\widehat{p}_k^{agg}} R_k y_k.$$

Next, we described two standard weighting procedures: linear weighting (Bunea *et al.* 2007, 2006) and exponential weighting (Buckland *et al.* 1997):

(1) *Linear weighting* The aggregate score $\widehat{p}_k^{agg}$ attached to unit k is obtained by fitting a linear regression model with the response indicator R_k as the dependent variable and $\widehat{p}_k^{(1)}(\mathbf{x}_k), \ldots, \widehat{p}_k^{(M)}(\mathbf{x}_k)$, as the set of explanatory variables.
Let $\widehat{\beta}_1, \ldots, \widehat{\beta}_M$, denote the resulting estimated regression coefficients. Under linear

Table 4.3 - Descriptive statistics of percent RE across the 24 ignorable scenarios: the Hájek estimator

ML procedure	Min	Q1	Median	Q3	Max	Mean
rf 3	173	200	215	334	558	277
	(0.2)	(3.1)	(5.2)	(14.1)	(35.8)	(9.7)
Exponential weighting: $\mathcal{L}_{mis}$ (with splitting)	177	198	220	330	534	273
	(0.6)	(3.2)	(5.8)	(13.9)	(38.9)	(10.8)
Exponential weighting: $\mathcal{L}_{cross}$ (with splitting)	178	199	220	331	535	273
	(0.7)	(3.3)	(5.9)	(14.3)	(39.3)	(10.9)
Exponential weighting: $\mathcal{L}_{mis}$ (without splitting)	175	197	220	326	535	272
	(0.6)	(3.1)	(5.6)	(13.6)	(38.5)	(10.6)
Exponential weighting: $\mathcal{L}_{cross}$ (without splitting)	174	196	221	323	535	272
	(0.6)	(3.1)	(5.5)	(13.3)	(38.1)	(10.5)
Linear weighting (with splitting)	175	200	223	324	493	271
	(0.2)	(2.5)	(5.7)	(11.0)	(26.5)	(7.9)
xgb 1	175	191	228	323	493	266
	(0.0)	(2.3)	(5.1)	(13.2)	(31.9)	(8.7)
Linear weighting (without splitting)	180	200	231	392	765	325
	(1.4)	(4.0)	(7.3)	(19.8)	(57.1)	(15.8)
knn 2	202	234	241	411	848	359
	(1.5)	(5.6)	(7.5)	(21.5)	(66.2)	(17.7)
cart 50	161	201	255	379	569	298
	(0.2)	(1.1)	(2.1)	(7.2)	(24.5)	(5.2)
score	224	351	532	736	4629	842
	(0.2)	(2.6)	(8.5)	(21.1)	(33.6)	(12.0)

weighting, the aggregation weights ω_m in (13) are defined as:

$$\omega_m = \widehat{\beta}_m^2 / \sum_{j=1}^{M} \widehat{\beta}_j^2. \tag{14}$$

(2) *Exponential weighting* Let $\mathcal{L}(\cdot)$ denote a loss function. The exponential weights ω_m are given by:

$$\omega_m := \frac{\exp\left\{-n \cdot T \cdot \mathcal{L}\left(\widehat{p}_m\right)\right\}}{\sum_{j=1}^{M} \exp\left\{-n \cdot T \cdot \mathcal{L}\left(\widehat{p}_j\right)\right\}}, \qquad m = 1, 2, ..., M, \tag{15}$$

where $T > 0$ is a hyper-parameter, often referred to as the temperature. When $T \longrightarrow 0$, the weights ω_m in (13) tend to be uniform, whereas $T \longrightarrow \infty$ will assign non-zero weights to the Machine Learning procedures exhibiting a small loss. For a discussion about the choice of the temperature, see Leung and Barron (2006) and Lecué (2007). We consider the following two loss functions:

(a) The misclassification error:

$$\mathcal{L}_{mis}\left(\widehat{p}_m\right) := \frac{1}{n} \sum_{k \in S} \mathbb{1}_{\widehat{R}_m(\mathbf{x}_k) \neq R_k},$$

Table 4.4 - Descriptive statistics of percent RE across the 12 nonignorable scenarios: the Hájek estimator

ML procedure	Min	Q1	Median	Q3	Max	Mean
cart 50	148	653	835	1051	2195	947
	(3.4)	(26.3)	(43.3)	(66.2)	(105.5)	(47.1)
Exponential weighting: $\mathcal{L}_{cross}$ (without splitting)	249	689	914	1281	2410	1108
	(13.4)	(34.0)	(53.4)	(70.9)	(115.3)	(56.3)
Exponential weighting: $\mathcal{L}_{mis}$ (without splitting)	255	702	916	1297	2419	1117
	(13.8)	(34.3)	(53.7)	(70.9)	(115.6)	(56.6)
Linear weighting (without splitting)	287	764	924	1404	2769	1240
	(16.0)	(37.8)	(55.2)	(70.1)	(129.9)	(60.1)
Exponential weighting: $\mathcal{L}_{mis}$ (with splitting)	273	731	924	1326	2420	1132
	(14.9)	(34.9)	(54.6)	(70.7)	(115.7)	(57.2)
Linear weighting (with splitting)	235	687	930	1258	2252	1065
	(12.3)	(32.0)	(53.3)	(70.5)	(110.6)	(54.8)
Exponential weighting: $\mathcal{L}_{cross}$ (with splitting)	288	761	932	1346	2433	1146
	(15.8)	(35.3)	(55.2)	(70.6)	(116.1)	(57.6)
xgb 1	231	669	948	1256	2457	1108
	(12.0)	(32.4)	(53.2)	(73.5)	(116.5)	(56.5)
rf 3	286	743	961	1423	2347	1150
	(16.3)	(36.5)	(56.3)	(68.6)	(113.7)	(57.4)
knn 2	391	813	985	1589	3379	1423
	(21.6)	(42.7)	(56.8)	(67.6)	(144.3)	(64.4)
score	656	1264	1628	2300	8356	2313
	(22.5)	(49.4)	(60.3)	(86.4)	(121.9)	(66.2)

where $\widehat{R}_m(\mathbf{x}_k) := \mathbb{1}_{\widehat{p}_m(\mathbf{x}_k)\geqslant 1/2}$.

(b) The cross-entropy loss:

$$\mathcal{L}_{cross}\left(\widehat{p}_m\right) := \frac{1}{n}\sum_{k\in S}\left\{-R_k \log\left(\widehat{p}_m(\mathbf{x}_k)\right) - (1-R_k)\log\left(1-\widehat{p}_m(\mathbf{x}_k)\right)\right\}.$$

To prevent the issue of overfitting, we consider a sample-splitting scheme that involves training/aggregation. More specifically, the aggregation procedures are implemented as follows:

Step 1: Shuffle the units in $D_S := \{(\mathbf{x}_k, R_k)\ ;\ k \in S\}$ and select a fitting proportion $\rho \in (0;1)$. Let $n_{fit} := n \times \rho$. For simplicity, we assume that n_{fit} is an integer.

Step 2: Partition the data D_S into a fitting set, D_{fit}, of size n_{fit}, and an aggregation set D_{agg}, of size $n_{agg} := n - n_{fit}$.

Step 3: Fit the M models based on D_{fit} to obtain the estimated response probabilities $\widehat{p}_1\left(\cdot, D_{fit}\right), \widehat{p}_2\left(\cdot, D_{fit}\right), \cdots, \widehat{p}_M\left(\cdot, D_{fit}\right)$.

Step 4: Determine the aggregation weights $\omega_m, m = 1, \ldots, M$, on the aggregation set D_{agg}, where ω_m is either given by (14) or (15). That is, the weights ω_m are computed with the loss $\mathcal{L}(\cdot)$ computed on D_{agg} with predictors $\widehat{p}_m\left(\cdot, D_{fit}\right)$ fitted on D_{fit}, $m = 1, \ldots, M$.

Step 5: Output the aggregated response probabilities estimator $\widehat{p}_{agg}\left(\cdot\ , D_{fit}, D_{agg}\right) \equiv \widehat{p}_{agg}$

given by

$$\widehat{p}_{agg} := \sum_{m=1}^{M} \omega_m(D_{agg}) \cdot \widehat{p}_m\left(\mathbf{x}_k, D_{fit}\right), \qquad k \in S_r.$$

To assess the performance of aggregation procedures, we used the same setup as the one described in Section 3.1. Again, we had $6 \times 4 = 24$ ignorable scenarios and $6 \times 2 = 12$ nonignorable scenarios. The aggregation procedures were based on the following $M = 5$ Machine Learning procedures: Xgboost1, cart50, rf3, knn2, and Score; see Section 3.1. The fitting proportion was set to 0 (without splitting) and to 0.7 (with splitting). The temperature T was set to $1/\mathbb{E}(n_{agg}) = 1/300$. We used both linear weighting, whereby the aggregation weights ω_m are given by (14) and exponential weighting based on both $\mathcal{L}_{mis}$ and $\mathcal{L}_{cross}$, whereby the weights ω_m are given by (15).

Tables 4.1 and 4.2 show some Monte Carlo descriptive statistics regarding the relative efficiency (RE) for the PSA estimator for the 24 ignorable scenarios and the 12 nonignorable scenarios, respectively. Tables 4.3 and 4.4 show the Monte Carlo descriptive statistics for the Hájek estimator.

We begin by discussing the results pertaining to the PSA estimator. From Table 4.1, we note that the aggregation procedures based on exponential weighting performed almost as well as the best procedure, here rf3. For the 12 nonignorable nonresponse mechanisms, Table 4.2 shows that all the aggregation procedures outperformed each Machine Learning procedure individually. Similar observations can be made about the Hájek estimator; see Tables 4.3 and 4.4. In our experiments, exponential weighting was slightly more efficient than linear weighting. The effect of aggregating original predictors or their split versions had limited effect when applied with exponential weighting. On the other hand, a careful examination of Tables 4.1-4.3 and 4.4 suggests that, for linear aggregation, aggregating split predictors drastically reduced the efficiency of the aggregated estimators in the worse scenarios. For instance, from Table 4.1, we note that that linear weighting exhibited a value of RE of about 2130 in the worst case when splitting was omitted as opposed to 889 when splitting was performed. Tables 4.2-4.4 also exhibit the same phenomenon. Exponential weighting, however, does not follow this patters: both the splitting and non-splitting versions exhibited similar performances in all our scenarios. The difference between the performance of linear with and without splitting seemed to be caused by significant differences in median absolute RB: for instance, in Table 4.1, the absolute RB in the worse case was equal to 22% for linear weighting with splitting, against 64% for linear weighting without splitting. Further research is needed to investigate this difference in behaviour in more depth. Finally, except for Table 4.4, the best method with respect to the average RE, was an aggregation procedure in all the procedures. Overall, the performance of aggregation procedures seems promising. They allow for a data-driven "automatic" aggregation of several estimated response probabilities and, as suggested by our results, aggregation often leads to good efficiency in comparison to the individual Machine Learning procedures.

References

Beaumont, J-F. 2005. "On the use of data collection process information for the treatment of unit nonresponse through weight adjustment". *Survey Methodology*, Volume 31, N. 2 : 227-231.

Breiman, L. 2001. "Random forests". *Machine Learning*, Volume 45: 5-32.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone. 1984. *Classification and Regression Trees. 1st edition.* London, UK: Routledge. https://doi.org/10.1201/9781315139470.

Buckland, S. T., K.P. Burnham, and N.H. Augustin. 1997. "Model Selection: An Integral Part of Inference". *Biometrics*, Volume 53, N. 2: 603-618.

Bunea, F., A.B. Tsybakov, and M.H.Wegkamp. 2007. "Aggregation for Gaussian regression". *The Annals of Statistics*, Volume 35, N. 4: 1674-1697. https://doi.org/10.1214/009053606000001587.

Bunea, F., A.B. Tsybakov, and M.H. Wegkamp. 2006. "Aggregation and Sparsity Via ℓ1 Penalized Least Squares". In Lugosi, G., and H.U. Simon (*eds*). *Learning Theory. COLT 2006. Lecture Notes in Computer Science*, Volume 4005. Berlin, Germany: Springer.

Chen, T., and C. Guestrin. 2016. "Xgboost : A scalable tree boosting system". In *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 785-794. https://doi.org/10.1145/2939672.293978.

Chipman, H.A., E.I. George, and R.E. McCulloch. 2010. "BART: Bayesian additive regression trees". *The Annals of Applied Statistics*, Volume 4, N. 1: 266 - 298.

Eltinge, J.L., and I.S. Yansaneh. 1997. "Diagnostics for Formation of Nonresponse Adjustment Cells, With an Application to Income Nonresponse in the U.S. Consumer Expenditure Survey". *Survey Methodology*, Volume 23, N. 1: 33-40. https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997001/article/3103-eng.pdf?st=ricg-FnF.

Gelein, B. 2017. "Handling missing data with superpopulation model, design-based approach and machine learning". *PhD diss.* Paris, France: *École National de la Statistique et de l'Analyse de l'information*.

Groves, R.M., and S.G. Heeringa. 2006. "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs". *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Volume 169, N. 3: 439-457.

Hastie, T., R. Tibshirani, and J.J. Friedman. 2001. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. New York, NY, U.S.: Springer.

Haziza, D., and J.-F. Beaumont. 2017. "Construction of Weights in Surveys: A Review". *Statistical Science*, Volume 32, N. 2: 206-226.

Haziza, D., and J.-F. Beaumont. 2007. "On the Construction of Imputation Classes in Surveys". *International Statistical Review*, Volume 75, N. 1: 25-43.

Kern, C., T. Klausch, and F. Kreuter. 2019. "Tree-based Machine Learning Methods for Survey Research". *Survey Research Methods* Volume 13, N. 1: 73-93.

Kim, J.K., S. Park, and K. Kim. 2019. "A note on propensity score weighting method using paradata in survey sampling". *Survey Methodology*, Volume 45, N. 3: 451-463.

Lecué, G. 2007. “Méthodes d’agrégation: optimalité et vitesses rapides”. *PhD diss.* Paris, France: Université Pierre et Marie Curie - Paris VI.

Leung, G., and A.R. Barron. 2006. “Information Theory and Mixing Least-Squares Regressions”. *IEEE Transactions on Information Theory*, Volume 52, N. 8 : 3396-3410.

Little, R.J.A., and S. Vartivarian. 2005. “Does weighting for nonresponse increase the variance of survey means?”. *Survey Methodology*, Volume 31, N. 2: 161-168.

Lohr, S., V. Hsu, and J. Montaquila. 2015. “Using Classification and Regression Trees to Model Survey Nonresponse”. In *Proceedings of the Survey Research Methods Section*: 2071-2085.

Nemirovski, A. 2000. “Topics in Non-parametric Statistics”. In Emery, M., A. Nemirovski, and D. Voiculescu. *Lectures on Probability Theory and Statistics. Ecole d'Ete de Probabilites de Saint-Flour XXVIII - 1998*. New York, NY, U.S.: Springer.

Phipps, P., and D. Toth. 2005. “Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data”. *Annals of Applied Statistics*, Volume 6, N. 2: 772-794.

Quinlan, J.R. 1993. “Combining instance-based and model-based learning”. In *ICML'93: Proceedings of the Tenth International Conference on International Conference on Machine Learning*: 236-243. Amherst, MA, U.S, 27-29 July 1993.

Quinlan, J.R. 1992. “Learning with Continuous Classes”. In *Proceedings of Australian Joint Conference on Artificial Intelligence*: 343-348. Hobart, Australia, 16-18 November 1992.

Särndal, C.-E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York, NY, U.S.: Springer.

Tsybakov, A. B. 2003. “Optimal Rates of Aggregation.” In Schölkopf, B., and M.K. Warmuth (*eds*). *Learning Theory and Kernel Machines. Lecture Notes in Computer Science*, Volume 2777: 303-313. Berlin, Germany: Springer.

Zeileis, A., T. Hothorn, and K. Hornik. 2008. “Model-Based Recursive Partitioning”. *Journal of Computational and Graphical Statistics*, Volume 17, N. 2: 492–514.

State of play and perspectives on Machine Learning at Istat

Marco Di Zio[1]

Abstract

This paper discusses the use of Machine Learning in Istat. It broadly illustrates the road taken by the Institute starting from the first works published on the use of neural networks in Official Statistics in '90s, to the current situation in which the use of big data in Istat has brought a great acceleration on this topic. In order to show the reasons why Machine Learning is a useful tool for these applications, some activities concerning Trusted Smart Statistics are briefly illustrated. In addition to this favourable context, studies have been conducted for the application of Machine Learning when using administrative and survey data. In this regard, some experimental results and future developments are discussed.

Keywords: Official Statistics, quality, multisource data, Trusted Smart Statistics.

1. Introduction

Machine Learning activities in Istat and more generally in statistical agencies begin in the 1990s, although the term Machine Learning was not directly used. Some early work was stimulated by publications in the area of editing and imputation (Nordbotten 1995, Nordbotten 1996, Roddick 1996). Those papers show the potentiality of Machine Learning but end up with the problem of computational feasibility. In the late 1990s, two international projects involving institutes of statistics and universities were launched. The first was named Autimp (Chambers *et al.* 2001) in which tree based methods for imputation are studied. As part of this project, a software named "WAID" was also produced (de Waal 2001). In the early 2000s, the Euredit project supported under the 5th Framework Programme of the European Union was launched. Twelve members among statistical institutes and universities participated in this project (Charlton 2004). In Euredit, multilayer perceptron techniques, correlation matrix memories, self-organising maps, support vector machines are compared to traditional methods for editing and imputation. The comparison is carried out on four different sets of data provided by National Statistical Institutes (NSIs) that cover different Official Statistics domains. The results involving the new techniques show good potentialities and indicate new studies to be carried out (Di Zio *et al.* 2004), though, one more time, problems are highlighted related to the computational feasibility of some algorithms, especially when applied to data such as the census ones. Several research papers have been produced after this project, but the big boost in Istat comes from the Scheveningen Memorandum "Big Data in Official Statistics" in 2013 and Bucharest Memorandum "Official Statistics in a datafied society - Trusted Smart Statistics" in 2018. Thanks to these solicitations, Istat has been more concretely committed to the use of big data for Official Statistics. In this area, data are generally unstructured and with large volume. These characteristics naturally lead to the use of Machine Learning techniques, also taking advantage of the favourable software development environment in recent years. In addition to TSS, under the 2019 UNECE HLG-MOS Machine Learning Project (UNECE 2021), Istat delved into the study of Machine Learning for imputation in the case of multisource data. This

1 Marco Di Zio (dizio@istat.it), Italian National Institute of Statistics - Istat. The views and opinions expressed are those of the author and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics - Istat.

issue is particularly relevant in the context of the Institute's modernisation process, which is based on building an integrated system of statistical registers that provides the spine of the statistics produced by the Institute. The system is designed to collect statistical units and their main characteristics for the various statistical domains of interest (Istat 2016*a*). Imputation may need to be used when integrating administrative data and sample surveys. In this multisource context, the use of MLP for the imputation of the attained level of education for the Italian population has been studied (De Fausti *et al.* 2022*a*; De Fausti *et al.* 2022*b*). In order to assess the validity of the method, the results of the MLP application are compared with those produced by officially adopted methods.

2. Some relevant Istat applications

2.1 Machine Learning in Trusted Smart Statistics at Istat

A relevant application context of Machine Learning techniques concerns the use of remote sensing for Official Statistics. A first application considered the use of deep learning methods for land cover classification (Bernasconi *et al.* 2022), followed by the study of high resolution remote sensed images (Orthophotos with 20 and 50 cm pixel resolution) for quantifying vegetation in urban centres (Mugnoli *et al.* 2024). Vegetation area classification is performed by analysing the Normalised Difference Vegetation Index (NDVI) indicator that exploits the spectral reflectance measurements acquired in the red (visible) and near-infrared regions. The statistical problem is how to determine the threshold that demarcates the green areas from the rest. Unfortunately, although reference threshold values can be found in literature, due to various issues related to the measurement in practice, it is difficult to stably adopt a value that is valid for all geographic areas. The authors therefore study an automatic algorithm based on unsupervised cluster techniques that allows isolating the part that can be classified as vegetation. The inspected techniques are Kmeans, Kmedians and also methods using kernel density estimation. Moreover, it may be useful to assess the intensity of green (vegetation). To this aim, the use of the algorithm known as Canny edge detection and the Otsu thresholding (Donchyts *et al.* 2016) is tested. In addition to green classification, it may be of interest to reconstruct the shape of figures potentially classified as green in order to distinguish different types of vegetation. To this end, it is useful to combine the previous techniques with image pattern recognition methods.

Another application, found also in the Eurostat's innovation agenda, is the use of data produced by the vessel's Automatic Identification System (AIS) for the improvement of the quality of maritime statistics in terms of timeliness, accuracy and relevance. AIS data are produced from signals sent by vessels' transponders at intervals of some minutes. Data include the ship's identification code, its characteristics such as for instance tonnage, and the ship's position. To use those data for statistical purposes, however, some questions need to be addressed. One major problem concerns the treatment of missing data, since in some cases the signal is interrupted. Imputation, that is the reconstruction of the missing information, is an approach that can be adopted. Given the particularly high volume of data and their nature, *i.e.* a signal in space and time, it comes naturally to resort to Machine Learning techniques. Several Machine Learning methods for imputation are studied. Although more experiments are still needed, from the first results the most promising are Xgboost and deep learning methods. In particular, for the latter, the focus is on the TrAISformer (Nguyen and Fablet 2021).

Sentiment analysis is another field on which Istat has started working on. In this context, we move away from the classical production of Official Statistics and look into the Institute's production in the field of experimental statistics. By means of sentiment analysis, Istat provides a timely glimpse into the sentiment about specific issues. In this area, the social mood on economy index has been produced since 2016 as an experimental statistic (Istat 2016*b*; Catanese *et al.* 2022). It is a daily index computed from the Italian Twitter/X's public stream aimed at representing the evolution of the feelings on economics topics. Further studies are currently being conducted on gender-based violence, hate speech and tourism. In these applications, text of messages must be understood and automatically classified into sentiment states. Natural language processing techniques are used: they are concerned with giving computers the ability to understand and elaborate texts. In this context as well, the choice of Machine Learning models appears the most natural one.

In addition to the previous applications, studies using Machine Learning are underway for the use of web scraping and automatic classification techniques, *e.g.* for enterprise automatic classification, for automatic classification of the economic activity, for the estimation of characteristics of enterprises through the use of the notes to the financial statements, for the automatic categorisation of the requests received by the Istat contact centre (Bianchi *et al.* 2022; Bruni *et al.* 2023).

Finally, it is important to mention the Istat engagement in the ESSNet Smart Surveys Implementation of 2023 project (De Vitiis *et al.* 2024). Here, the use of smart devices (*e.g.* smartphones, tablets, activity trackers) to collect data through sensors and mobile applications is explored. These data collection techniques can combine an active approach on input from the data subjects with data collected passively by the device sensors (*e.g.* accelerometer, GPS, microphone, camera, etc.). Machine Learning is studied for structuring unstructured data and to classify objects acquired from the images, or physical activities using accelerometer data, or leisure activities using GPS data matched with street maps.

To conclude, many applications are mentioned and certainly the list is not exhaustive, but it seems clear that the introduction of big data in Official Statistics has opened the door to the use of Machine Learning within Istat.

2.2 Machine Learning for imputation with multisource data

Imputation is an application area of Machine Learning because it is essentially a prediction problem. As so far introduced, there are several studies on the use of Machine Learning for imputation both in Official Statistics and with reference to big data. However, the use of Machine Learning techniques with survey data needs further investigations. In Istat, a particularly relevant area of application is the case in which survey data are integrated with administrative data. In the Italian population census, the attained level of education is obtained by integrating Ministry of Research data with survey data. Administrative data has a lag in time with respect to the reference period, so mass imputation techniques are developed to estimate the level of education at the time of interest (Di Zio *et al.* 2019). Multi-layer perceptron models (MLP) are applied to the 2018 census data for the province of Lombardy, and the results are compared with those obtained with the officially adopted procedure (De Fausti *et al.* 2022*a*; De Fausti *et al.* 2022*b*). The goal is to evaluate the possible improvement in the accuracy of estimates, along with making the process more automatic. To this end, MLPs are applied in different experimental situations: with the same setting as the one adopted for the current procedure, and in a setting where raw data are provided without pre-processing. Moreover, analyses are conducted by both using and not using sampling weights.

The results of the two methods are comparable, so we can say that there is not an improvement from the point of view of accuracy, and this is probably due to the informative power and structure of explanatory covariates. However, it should be pointed out that the comparability of the results is still an important result because two important problems are dealt with, namely, how to make random imputation and how to use sampling weights with MLP. In fact, one of the objectives of the study was precisely to understand how to cope with these issues. There is not a large literature on those topics since Machine Learning methods are mainly developed in other contexts. A recent reference work on this topic (Dagdoug *et al.* 2023) notes that it is not always easy to introduce weights in the packages developed for Machine Learning, so two alternative strategies are proposed (De Fausti *et al.* 2022b): 1) weights are introduced into the loss function adopted by the algorithm, 2) the sample is expanded according to the weights, thus obtaining a pseudo-population on which MLP is applied.

3. Lessons learned and future studies

This paper reports the current state of application of Machine Learning methods as well as the relevant past projects in Istat. What emerges is that these are the main techniques for TSS (Daas 2023). This is due to the nature of the data that are generally 'big' in terms of volume, and often unstructured such as for instance images, signals, texts. It is well known that all of these aspects can be usefully exploited by Machine Learning methods. Furthermore, TSS are generally referred to prediction and classification problems that are the relevant statistical contexts for Machine Learning techniques. Finally, no less important is the fact that most of the scientific literature dealing with big data refers to Machine Learning methods, and this is a relevant aspect for a National Statistical Institute that mainly works in the area of applied research.

The application of Machine Learning to the more classic cases of Official Statistics is a bit different. The typical Official Statistics context makes use of survey samples, and Machine Learning methods must be tested and adapted to deal with the elements characterising the sample and more in general to the inferential context of finite populations. One aspect of fundamental importance is that of dealing with the elements of the sampling design. If the design is not ignorable, it must be taken into account in the model. If the variables that define the sample design are not included as predictors in the Machine Learning model, then the sampling weights should be included in the model to avoid bias. In the case of imputation, another element to keep in mind is that the ultimate goal is not to predict the individual value of each unit, but rather the prediction is aimed at obtaining estimates of aggregated values of the variable of interest, *e.g.* a total, or quantiles. To avoid a biased estimator of distributions when computed on imputed variables, imputation is made by adding an appropriate random residual to the value predicted by the model in order to preserve variability (Chen and Haziza 2019). The use of sampling weights and random imputation is discussed in some papers (Dagdoug *et al.* 2020; De Fausti *et al.* 2022*b*) where the authors show the results of an application on real data. Nevertheless, further studies should be devoted to these questions.

There are other elements that we need to focus on for an application of Machine Learning to data from sample surveys. In the context of statistical register development in NSIs, the use of administrative sources frequently leads to longitudinal data, that is, units observed repeatedly at different times, for example for the level of education we have the individual's entire study path. It may be important for improving the Machine Learning predictions/imputations to

take into account the story of a statistical unit. Another issue that needs to be explored is the possibility in some cases that units occur in clusters. This happens for example in the case of a sampling design where the unit of observation is the household. This implies dependencies for observations within one cluster, leading to violations of independent and identically distributed assumptions, biased estimates, and false inference (Kilian *et al.* 2023). Given the importance of such aspects in Istat, we have planned to do research studies on such topics.

Finally, an aspect of particular importance for a national statistical institute, that must produce official data, concerns quality assessment. Most quality measures developed in Machine Learning refer to the assessment of goodness of prediction. In NSIs, accuracy generally is calculated with respect to aggregations of the data, and is generally disseminated with measures expressing statistical uncertainty, for instance confidence intervals. Machine Learning methods are more developed with the main aim of prediction, and thus accuracy measures reflect these objectives, in fact they are focussed on prediction accuracy. Further studies are needed to move quality evaluations towards inference (Larbi *et al.* 2024; Daas 2023). Some answers might come from resampling techniques, *e.g.* bootstrapping, but further investigation needs to be done on these to see their applicability in the Official Statistics contexts, in fact they may be computationally prohibitive, or need special accommodations in the case of the presence of finite populations (Chen *et al.* 2019).

References

Bianchi, G., G. Bellini, P. Bosso, and P. Papa. 2022. "A machine learning based help-desk approach for units involved in official surveys". *UNECE Expert Meeting on Statistical Data Collection* "*Towards a New Normal*". Roma, Italy, 26-28 October 2022.

Bernasconi, E., F. De Fausti, F. Pugliese, M. Scannapieco, and D. Zardetto. 2022. "Automatic Extraction of Land Cover Statistics from Satellite Imagery by Deep Learning". *Statistical Journal of the IAOS*, Volume 38, N. 1: 183-199.

Bruni, R., G. Bianchi, and P. Papa. 2023. "Hyperparameter Black-Box Optimization to Improve the Automatic Classification of Support Tickets". *Algorithms*, Volume 16, N. 1: 46.

Catanese, E., M. Scannapieco, M. Bruno, and L. Valentino. 2022. "Natural language processing in official statistics: The social mood on economy index experience". *Statistical Journal of the IAOS,* Volume 38, N. 4: 1451-1459.

Chambers, R.L., J. Hoogland, S. Laaksonen, D.M., Mesa, J. Pannekoek, P. Piela, P. Tsai, and T. De Waal. 2001. *The AUTIMP-project: Evaluation of Imputation Software*. Voorburg, the Netherlands: Statistics Netherlands.

Charlton, J. 2004. "Editorial: Evaluating Automatic Edit and Imputation Methods, and the EUREDIT Project". *Journal of the Royal Statistical Society*. *Series A* (*Statistics in Society*), Volume 167, N. 2: 199-207.

Chen, S., and D. Haziza. 2019. "Recent Developments in Dealing with Item Non-response in Surveys: A Critical Review". *International Statistical Review*, Volume 87, N. S1: S192 - S218.

Chen, S., D. Haziza, C. Léger, and Z. Mashreghi. 2019. "Pseudo-population bootstrap methods for imputed survey data". *Biometrika,* Volume 106, N. 2: 369-384.

Daas, P. 2023. *Big Data in Official Statistics*. Eindhoven, the Netherlands: Eindhoven University of Technology. https://research.tue.nl/files/296764797/Rede_Daas_26_5_2023.pdf.

Dagdoug, M., C. Goga, and D. Haziza. 2023. "Imputation Procedures in Surveys Using nonparametric and Machine Learning Methods: An Empirical Comparison". *Journal of Survey Statistics and Methodology*, Volume 11, N. 1: 141-188.

De Fausti, F., M. Di Zio, R. Filippini, S. Toti, and D. Zardetto. 2022*a*. "Multilayer perceptron models for the estimation of the attained level of education in the Italian Permanent Census". *Statistical Journal of the IAOS*, Volume 38, N. 2: 637-646.

De Fausti, F., M. Di Zio, R. Filippini, S. Toti, and D. Zardetto. 2022*b*. "The imputation of the "Attained Level of Education" in the base register of individuals through Neural Networks using sampling weights". *UNECE Conference of European Statisticians, Expert Meeting on Statistical Data Editing*. *Online*, 3-7 October 2022. https://unece.org/statistics/events/SDE2022.

Di Zio, M., R. Filippini, and G. Rocchetti. 2019. "An imputation procedure for the Italian attained level of education in the register of individuals based on administrative and survey data". *Rivista di Statistica Ufficiale*, N. 2-3: 143-174.

De Vitiis, C., F. De Fausti, F. Inglese, and M. Perez. 2024. "Smart Surveys: Methodological issues and challenges for Official Statistics". In *2nd Workshop on Methodologies for Official Statistics - Proceedings*, Session 2. Roma, Italy: Istat.

De Waal, T. 2001. "WAID 4.1: a computer program for imputation of missing values". *Research in Official Statistics*, Volume 2: 47-64.

Di Zio, M., M. Scanu, L. Coppola, O. Luzi, and A. Ponti. 2004. "Bayesian Networks for Imputation". *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, Volume 167, N. 2: 309-322.

Donchyts, G., J. Schellekens, H. Winsemius, E. Eisemann, and N. Van de Giesen. 2016. "A 30 m Resolution Surface Water Mask Including Estimation of Positional and Thematic Differences Using Landsat 8, SRTM and OpenStreetMap: A Case Study in the Murray-Darling Basin, Australia". *Remote Sensing*, Volume 8, N. 5: 386. https://doi.org/10.3390/rs8050386.

Kilian, P., S. Ye, and A. Kelava. 2023. "Mixed effects in machine learning - A flexible mixedML framework to add random effects to supervised machine learning regression". *Transactions on Machine Learning Research*, N. 2.

Istituto Nazionale di Statistica - Istat. 2016*a*. *Istat's Modernisation Programme*. Roma, Italy: Istat. https://www.istat.it/it/files//2011/04/IstatsModernistionProgramme_EN.pdf.

Istituto Nazionale di Statistica - Istat 2016*b*. *Social Mood on Economy Index*. *A daily measure of the italian sentiment on the economy based on X data*. Roma, Italy: Istat. https://www.istat.it/en/experimental-statistic/social-mood-on-economy-index-2/.

Larbi, K., J. Tsang, D. Haziza, and M. Dagdoug. 2024. "On the use of Machine Learning methods for the treatment of unit nonresponse in surveys". *2nd Workshop on Methodologies for Official Statistics - Proceedings*, Session 4. Roma, Italy: Istat.

Mugnoli, S., A. Sabbi, F. De Fausti, G. Lancioni, and F. Sisti. 2024. "Quantification of urban green areas: An innovative remote sensing approach for official statistics". *2nd Workshop on Methodologies for Official Statistics - Proceedings*, Session 2. Roma, Italy: Istat.

Nguyen, D., and R. Fablet. 2021. *TrAISformer - A Transformer Network with Sparse Augmented Data Representation and Cross Entropy Loss for AIS-based Vessel Trajectory Prediction*. arXiv preprint arXiv:*2109.03958*.

Nordbotten, S. 1996. "Neural network imputation applied to the Norwegian 1990 population census data". J*ournal of Official Statistics*, Volume 12, N. 14: 385-401.

Nordbotten, S. 1995. "Editing Statistical Records by Neural Networks". *Journal of Official Statistics,* Volume 11, N. 4: 391-411.

Roddick, L.H. 1996. "Data editing using neural networks". *Data Editing Workshop and Exposition*, 199-205.

United Nations Economic Commission for Europe - UNECE. 2021. *Machine Learning for Official Statistics*. Geneva.Switzerland: United Nations. https://unece.org/sites/default/files/2022-09/ECECESSTAT20216.pdf.

Machine Learning in Official Statistics: Towards statistical based Machine Learning

Marco J.H. Puts, Piet J.H. Daas[1]

Abstract

The paper discusses the important difference between techniques and methods and how this is particularly important when Machine Learning techniques are applied within the context of Official Statistics. This is illustrated by discussing four examples of Machine Learning work performed by the authors which exemplify the importance of a methodological sound approach when including Machine Learning techniques. The need for training data that represents the target population studied illustrates this without a doubt. This is essential to obtain models that can be applied to the unseen part of the population, particularly when found training data is used.

Keywords: Data science, methodology, Official Statistics, representativity.

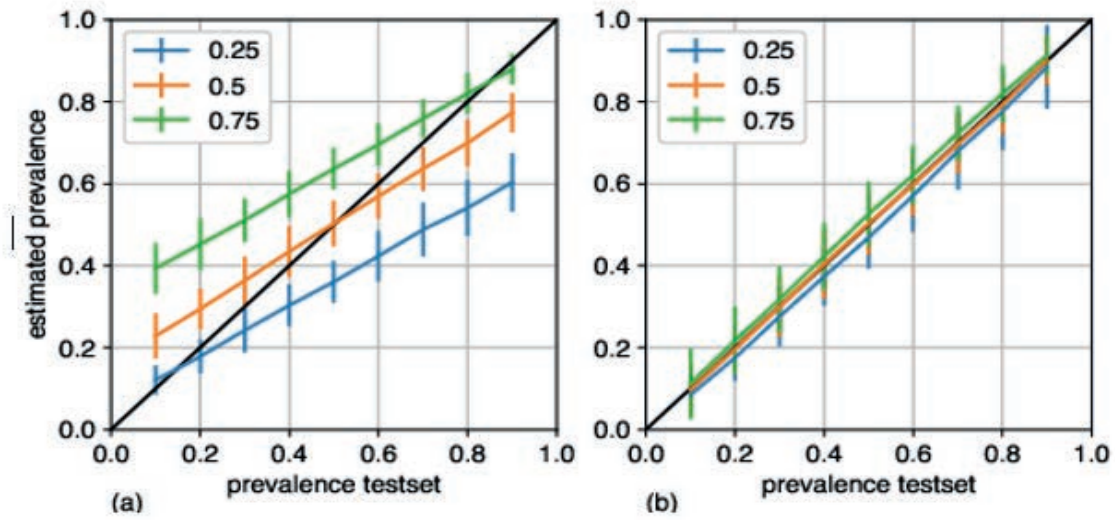
1. Introduction

The recent onset of the academic field of Data Science has greatly stimulated the use of new ways to produce statistics (Daas 2023). This relatively new area of science was first mentioned in the 1960s as a subfield of statistics with a focus on "learning from data", also referred to as data analysis (Donoho 2017), and really took off around 2011. Nowadays, topics such as artificial intelligence, Machine Learning (ML), and data visualisation, just to name a few, are or can be included under the umbrella term for computational data analysis that Data Science has become. In this paper we will focus specifically on the subfield of Machine Learning, also described as 'developing algorithms that learn from data' (Murphy 2012). The overall importance of Machine Learning for Official Statistics lies in i) its focus on learning from data in an evidence-based manner and ii) the efficient implementation of the algorithms (Daas 2023).

ML techniques work because they are able to detect patterns in the data they are trained on. As such, ML-based approaches are examples of working in a data-driven way (Adler and Rips 2008). The ultimate goal of such studies is to find patterns that are applicable to the whole population (UNECE 2021). The latter generalisation is essential in the context of Official Statistics. This is precisely the major concern when applying ML, or any other data-driven approach, within an official statistical context (Puts and Daas 2021*b*). But before we discuss this topic in more detail, the essential difference between techniques, methods, and methodology needs to be explained.

1 Marco Puts (m.puts@cbs.nl), Statistics Netherlands; Piet H. H. Daas (pjh.daas@cbs.nl), Statistics Netherlands and Eindhoven University of Technology (p.j.h.daas@tue.nl). The authors thank Luuk Gubbels for his excellent contribution to the work described in this paper and Yvonne Gootzen for stimulating discussions.

Figure 2.1 - Effects of training a model on a certain ratio of positive items on the classification of datasets with different ratios of positive items: Blue 25%, Orange 50%, and Green 75% positives (side a). After applying a Bayesian correction method developed (side b)



Source: Puts and Daas 2023*b*

1.1 Techniques, methods, and methodology

A technique is basically a way of carrying out a particular task. By using a technique, a person, such as a cook, is able to perform a particular (practical) task; such as 'peeling potatoes'. In the context of ML, a task can be 'discern websites as online platforms or not based on a set of features'. Techniques are often highly specialised. By combining different techniques and ingredients in a particular way, a cook is able to prepare a meal. The entire procedure followed by the cook, including steps that include techniques, is essentially a method. As such, a method defines a particular procedure for accomplishing a particular goal. Everything involved in, for instance, creating a trained ML model that is able to detect online platforms, based on a set of data, is an example of a method. A methodology is an entire system of methods used in a particular area of study. For instance, survey methodology, the study of survey methods and techniques, is an example of such a system. In the remainder of this paper, we will make clear that methods, and hence a methodology, needs to be developed when applying ML in the context of (official) statistic. This is to ensure high quality and reproducible results.

2. The need for Machine Learning methodology

The discipline of ML is predominantly composed of techniques. One only has to look at the index of an ML-book and this becomes obvious (Murphy 2012). These books are filled with descriptions of various algorithms that can be used to study data. Hence, practitioners of ML are used to apply various techniques to the data at hand, such as training a model, and drawing conclusions from that (Sarker 2011). However, because the datasets used by ML practitioners are often composed of 'found' data and – subsequently - may not represent the target population of the study well, there is a need to focus on a more scientific and rigorous way when applying ML. This is certainly the case when ML is applied in the context of Official Statistics (Puts and Daas 2011*b*). Particularly in this scientific area, the findings must be generalisable to the target population studied. Hence, using ML to develop a model that is able to identify online platforms

(based on web data), needs to be done in such a way that the trained model performs well, not only on the train and test data but also on the (unseen) data of the entire population of websites (Daas *et al.* 2023*a*). Here, a method needs to be applied that ensures that a well-performing model is obtained. These and any other important issues identified during the work performed by the authors (Daas and Puts 2023) revealed that there is a need to develop a methodology when applying ML. This is discussed in the following four subsections, each including an example.

2.1 Creating a good training and test set

The dataset used on which ML techniques will be applied is essential. As an example, we will describe the approach developed by the authors and some Statistics Netherlands colleagues, to identify online platform businesses based on the texts on their websites (Daas *et al.* 2023*a*). For this purpose, a model was developed based on a dataset containing known examples of both positive (platform) and negative (non-platform) cases. Because you need to start somewhere, experts from Statistics Netherlands were asked to provide a list of around 500 online platform businesses. The negative cases were obtained by taking an equal-sized random sample of all websites linked to businesses in the Business Register of Statistics Netherlands. The latter cases were thoroughly checked to ensure that no online platforms were accidentally included. Here, we assumed that the experts provided us with a representative sample of online platforms, but this does not have to be the case (but we were aware of that). One can imagine that these positive cases contain many more examples of businesses active in one (or more) particular branch(es). As a result, the model trained on those examples may miss specific groups of platform businesses active in other branches, which may have different features. Iteratively developing the model, extensive manual checking, and paying specific attention to the results obtained for various branches are ways to reduce this form of bias in the model. These and other potential approaches have been investigated (Gubbels 2023). There is a definitive need to develop a method that is able to produce representative datasets for the target population under study.

2.2 Internal and external validity

The next section is about internal and external validation. When one develops a model in a setting where there is a dataset with known outcomes (*e.g.* platform and non-platform cases), an 80% random sample is often drawn of such a dataset on which the model is subsequently trained. During training, the algorithm 'learns' the difference between the two cases (platform and non-platform) in the best possible way. The remaining 20% of the original dataset is used as an independent test set. This test set is used to independently determine how well the model is able to discern between the two cases as the test set contains examples that are entirely new to the model. We refer to this as the internal validation of the model's performance. However, for Official Statistics, we are predominantly interested in the performance of the model on the target population. In other words, for the online platform model, we want to know how well the model performs on totally new, unseen cases included in 'real-world' data. We refer to this as external validation. This requires data (if possible, with known outcomes) from a substantially larger dataset, ideally a representative part of the target population. A manual inspection of a sample of 'real-world' classified data by several experts is a way to determine the external validity of the model (Daas *et al.* 2023*b*). There is a definite need to develop a method that ensures that ML models are both internally and externally valid.

2.3 Bias correction

The third section has to do with the bias introduced when the ratio of the positive and negative cases used in the training (and test set) differs from that in the 'real-word' (target) population. In a supervised setting, a specific percentage of positive and negative examples are included in the dataset on which an ML model is trained. Quite often, 50% positive and 50% negative cases are used. However, these percentages may not have anything to do with the percentages to which these cases occur in 'real world' data. For instance, our best estimate of the percentage of online platforms in the Dutch Business Register suggests 0.25% positive cases (Daas *et al.* 2023*a*). It is nearly impossible to train a useful model on a dataset with such a low number of positive cases simply because a model that always identifies a case as a non-platform will be correct in 99.75% of the cases. So, we need to use a different percentage of positive cases to obtain a model that is able to do that well, but what percentage is best? While looking at that, we observed that a model trained on a particular percentage of positive items introduces a bias when applied to datasets with different (known) percentages of positive items (Puts and Daas 2023*a*). These findings are shown in Figure 2.1.

Figure 2.1 (side a) reveals that models tend to be biased towards the outcome of the percentage of positives on which they are trained. Along the x-axis, we see the true fraction of positive items, whereas the estimated fraction of positive items is shown along the y-axis. The grey line indicates the situation in which the true and estimated values are equal and the bias is thus zero. As one can see, the estimated value is only correct at one point: the fraction on which the algorithm was actually trained. We call this the *intrinsic prevalence*, the model's assumption about the fraction of positives in the dataset. Since online platforms are rare, there is a risk that a model developed on a dataset with increased prevalence will overestimate the number of online platforms when applied to 'real-world' data. This is actually what happened, but the effect was reduced by careful manual checking and validation of the outcome (of the model) by sending companies a questionnaire (Daas *et al.* 2023*a*). Currently, each step in this approach is being studied with the aim of improving it from the viewpoint of automating the selection process as much as possible. This work has resulted in the development of a new metric that can be used to improve the training of a model as it is less affected by high and low ratios of positive and negative examples (Gubbels 2023). It has also resulted in the development of a Bayesian adjustment method (Puts and Daas 2021*a*) to correct for the bias of a specific group of Machine Learning classifiers that produce (pseudo-)probabilities as their outcome. The correction is illustrated in Figure 2.1 (side b). This method also corrects for any bias resulting from a difference between the ratio of falsely classified negative and positive cases caused by the model (Meertens 2021). There is a definite need to develop a method that reduces the bias in ML-based estimates as well as possible.

2.4 Features and more

ML models select features (variables) in the training set that are related to the target variable. During our studies, we observed that different models trained on the same dataset contained varying numbers of features and also different features. Creating multiple models enables one to observe which features are the most important ones and detect and remove (some) of the non-relevant features. Some of them are accidentally included, while others are associated with some of the features selected. In addition, work from Gubbels (2023) revealed that using multiple models in the estimation process of detecting online platforms, produces a much less

biased estimate. What is observed here is that the bias introduced by each model, even after Bayesian correction (see previous Section), averages out when combined. There is a definite need to develop a method that assures that the bias in ML-based estimates is as low as possible.

3. Discussion

From the above, it is obvious that correctly applying ML techniques within the context of Official Statistics is not an easy task. One can, even unconsciously, introduce faults when these techniques are applied in a haphazard way. There is a definitive need for procedural assistance when ML is used. The mere fact that official statisticians want to develop and use models that can be applied - with confidence - to the unseen part of the target population, preferably the complete target population, makes this perfectly clear. Other examples of this need have been described above. Hence, ML methodology, a procedure to correctly apply ML techniques within the area of Official Statistics, is needed. The Bayesian correction method developed (Puts 2023) is an example of a single step in this methodology. More work has to be done to create a whole framework of methods needed. In fact, we can only envy, be proud, and learn from the entire Survey Methodology framework that has been developed by others (Groves *et al.* 2009).

References

Adler, J.E., and L.J. Rips (*eds*). 2008. *Reasoning. Studies of Human Inference and Its Foundations*: 187-370. Cambridge, UK: Cambridge University Press.

Daas, P. 2023. *Big Data in Official Statistics*. Eindhoven, the Netherlands: Eindhoven University of Technology. https://research.tue.nl/files/296764797/Rede_Daas_26_5_2023.pdf.

Daas, P., and M. Puts. 2023. "Lessons learned when applying Machine Learning in Official Statistics: Why it helps to be a survey statistician and a data scientist!". *Paper for the UNECE Machine Learning for Official Statistics Workshop 2023*. Geneva, Switzwerland, 5-7 June 2023. https://unece.org/sites/default/files/2023-04/ML2023_S2_Netherlands_Daas_A.pdf.

Daas, P., W. Hassink, and B. Klijs. 2023*a*. "On the Validity of Using Webpage Texts to Identify the Target Population of a Survey: An Application to Detect Online Platforms". *Journal of Official Statistics*, Volume 40, N. 1: 190-211.

Daas, P., B. De Miguel, and M. De Miguel. 2023*b*. "Identifying Drone Web Sites in Multiple Countries and Languages with a Single Model". *Journal of Data Science*, Volume 21, N. 2: 225-238.

Donoho, D. 2017. "50 Years of Data Science". *Journal of Computational and Graphical Statistics*, Volume 26, N. 4: 745-766.

Groves, R.M., F.J. Fowler Jr., M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey Methodology*. Hoboken, NJ, U.S.: Wiley.

Gubbels, L. 2023. "The sample Pearson Correlation Coefficient for classification models and identifying platform economy businesses from web-scraped data". *Master's thesis Applied and Industrial Mathematics*. Eindhoven, the Netherlands: Eindhoven University of Technology.

Meertens, Q.A. 2021. "Misclassification bias in statistical learning". *PhD thesis*. Amsterdam, the Netherlands: University of Amsterdam. https://hdl.handle.net/11245.1/4b031bbd-5a46-4181-b0f1-52b38a3b63a6.

Murphy, K.P. 2012. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, U.S.: MIT Press.

Puts, M. 2023. "BayesCCal: Bayesian Calibration of classifiers". *Code on Github*. https://github.com/mputs/BayesCCal.

Puts, M.J.H., and P.J.H. Daas. 2021*a*. "Unbiased Estimations Based on Binary Classifiers: A Maximum Likelihood Approach". *Abstract for the 2021 Symposium on Data Science and Statistics, Machine Learning session*. Online, 2-4 June 2021. https://arxiv.org/abs/2102.08659.

Puts, M., and P. Daas. 2021*b*. "Machine Learning from the Perspective of Official Statistics". *The Survey Statistician*, Volume 84: 12-17.

Sarker, I.H. 2021. "Machine Learning: Algorithms, Real-World Applications and Research Directions". *SN Computer Science*, Volume 2: 160.

United Nations Economic Commission for Europe - UNECE. 2021. "Machine Learning for Official Statistics". Geneva, Switzerland: United Nations. https://unece.org/sites/default/files/2022-09/ECECESSTAT20216.pdf.

Closing SESSION

Final considerations and perspectives on the second Workshop on Methodologies for Official Statistics

Orietta Luzi[1]

We are at the end of this second Istat Workshop on Methodologies for Official Statistics, dedicated to the methodological challenges and opportunities opened by using non-traditional data sources and Machine Learning methods to produce Official Statistics.

During the past two days, we have discussed the importance of NSIs taking advantage of the increasing amount and variety of new data sources available in the data ecosystem. This is to respond to society's ever-increasing information needs while keeping survey costs and statistical burden under control, and ensuring the highest levels of quality, privacy preservation, relevance, independence and transparency of Official Statistics.

We discussed the methodological pros and cons associated with this new perspective, we shared ideas, applications, possible approaches, and methods to use these new data, either alone or in combination with traditional ones, to transform them into statistical information, to assess the quality (the trustiness) of statistics obtained using new data sources and new methods like Machine Learning.

Istat is investing significant resources in these research areas, including human resources. Some research infrastructures have been created, such as an Innovation Lab, a Centre for Trusted Smart Statistics and a specific unit to manage the production of experimental statistics under the supervision of the Istat Research Committee.

Istat researchers are involved in several European projects; in this respect, I want to underline the key role of Eurostat in supporting research projects on topics related to the use of non-probabilistic data sources in Official Statistics.

Actually, for Istat and the other NSIs, it is essential not only to continue investing in these issues but also to collaborate to develop common solutions to common problems, together with the academic world and other research institutions. In addition, finding discussion spaces like this Workshop, in my opinion, represents a useful opportunity to align ourselves with each other on the state-of-the-art and current methodological advancements in this increasingly important field of public research.

These comparisons should continue in the future if we want to engage in a progressive and constant evolution towards new production systems. Official Statistics are to be produced based on new methodological paradigms, always in compliance with the Official Statistics' requirements, following the related technological, legal, and communication developments mentioned so frequently over the past two days.

Before concluding this meeting, on behalf of Istat, I would like to thank again Professor Daniela Cocchi, coordinator of the Istat Advisory Committee on Statistical Methods, who chaired this Workshop, all the other Committee members, for chairing the various sessions, and for their very stimulating discussions: Professor Maria Giovanna Ranalli, Professor Li-Chun Zhang, Professor Brunero Liseo, Piero Demetrio Falorsi, and Professor Natalie Shlomo (also President of the International Association of Survey Statisticians).

1 Orietta Luzi (luzi@istat.it), Italian National Institute of Statistics - Istat.

I want to express my gratitude once again to Professor Changbao Wu, Professor Stefano Maria Iacus, and Fabio Ricciato for their insightful and valuable master classes. I also want to thank all the speakers who contributed to the various sessions, including the new members of the Istat Advisory Committee on Statistical Methods: Professor Marco Alfò, Professor David Haziza, and Professor Piet Daas.

I would like to extend my thanks to the Istat Programme Committee for their enormous effort in preparing and managing the scientific part of this Workshop, and to the Istat Communication and IT team for their excellent organisation and technical support.

Finally, yet importantly, I want to thank all of you for attending this Workshop, whether in person or online, and I look forward to seeing you next year for the third edition of this event, once again here at Istat.