U.PORTO

MASTER
MODELLING, DATA ANALYSIS AND DECISION SUPPORT SYSTEMS

Challenges in Official Statistics: Combining non-probability and probability sampling

A theoretical perspective and application to the Global Entrepreneurship Monitor survey

Joana Martins

2026

U.PORTO
FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

CHALLENGES IN OFFICIAL STATISTICS: Combining non-probability and probability sampling. A theoretical perspective and application to the Global Entrepreneurship Monitor survey.

Joana Filipa Costa Martins

Dissertation

Master in Modelling, Data Analysis and Decision Support Systems

Supervised by
Patrício Costa

2026

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude towards my supervisor, Prof. Patrício Costa, who guided and was present in the steps that led to the development of this study, gave important feedback and guidance and showed a willingness to help me whenever necessary. His help was crucial to the completion of this dissertation.

I would also like to thank STATEC for the three-month internship placement in Luxembourg and for the proposal of this intriguing and challenging topic. Being present in the office, having contact with other international colleagues working in official statistics, and presenting my work at economic seminars was very insightful. the opportunity to join STATEC through the European Master's in Official Statistics (EMOS) program was the springboard I needed. Working daily at a national institute tested my analytical skills beyond the confines of theory and prepared me, in practice, for the demands of the profession.

A special acknowledgment to my internship supervisor, Guillaume Osier, and to the person responsible for the Global Entrepreneurship Monitor at STATEC, Cesare Riillo, who contributed to the progress of this work through weekly reunions and feedback during my internship, and through daily support and monitoring, with an emphasis on the practical component of this dissertation.

I would also like to thank DevStat for supporting and funding my internship as part of the EMOS program. Without this support, I would not have been able to undertake this internship at STATEC, which was crucial in sparking my interest in a career in official statistics.

Special thanks go to Maziek Beręsewicz and Lukas Chrotowski, some of the creators of the `nonprobsvy` package, who, in addition to creating and developing this essential tool, which has simplified the practical implementation of the methods studied, attended one of my economics seminars in STATEC where I had the opportunity to present the progress of my work to date. There, they provided important feedback and suggestions for the further development and improvement of this work and expressed their willingness to offer future support.

In addition, I would like to thank all the professors and teachers who were present

throughout my academic journey at FEP and shared their knowledge, which proved essential in many ways and made many stages of my work easier, allowing me to put those skills into practice. In fact, I would like to thank the FEP community for allowing me to undertake this master's degree; it has been my great dream to study Data Analysis after completing my bachelor's degree, and to do so at a faculty as prestigious as FEP is truly a privilege.

Last but not least, I would like to thank my family for their support in every way possible over the last five years I have been studying, and for making so much effort so that I could pursue my dreams and finish my studies. They are the most important figures in my life.

The completion of this dissertation represents, far more than the culmination of an academic cycle, a definitive bridge to the job market. The structure of my academic journey proved fundamental to this process of professional maturation. On the one hand, the Master's degree in Modelling, Data Analysis, and Decision Support Systems (MADSAD) equipped me with the algorithmic and analytical tools necessary to handle the current complexity of information. On the other hand, my participation in the EMOS program introduced the dimension of formal statistical rigor. This synergy is invaluable: in today's era of data science, where computational complexity sometimes overshadows the theoretical foundation, rigorous statistics is often underestimated, yet it is the only anchor that guarantees the true validity and representativeness of the inferences we produce.

This project allowed me to bring an important learning journey full circle. My initial exposure to sampling theory took place in a classroom, where I acquired the theoretical foundations and learned the classic distinctions between probabilistic and non-probabilistic methods. However, this work demonstrated how that academic foundation was merely the tip of the iceberg. Seeing how current literature delves deeper into the integration of these sampling methods was fascinating. Developing and dissecting these concepts while surrounded by STATEC's outstanding professionals and my advisor, Patrício Costa, provided me with a level of statistical literacy and analytical intuition that independent study could never have replicated. The constant exchange of perspectives, the expert guidance, and the iterative discussion of every detail of the models were irreplaceable.

ABSTRACT

Non-probability surveys have become widely used in research due to their efficiency and low cost, but suffer from selection bias, becoming unreliable for direct generalization to the broader population. To address this, as part of the European Master in Official Statistics (EMOS), the National Institute of Statistics of Luxembourg (STATEC) proposed evaluating a more robust framework for correcting selection bias. This research integrates the CAWI non-probability component of the 2024 Luxembourg Adult Population Survey (APS) from the Global Entrepreneurship Monitor (GEM) with EU Statistics on Income and Living Conditions (EU-SILC) reference data through common covariates. The methodology compares Inverse Probability Weighting (IPW), Mass Imputation, and Doubly Robust estimators, focusing on Total early-stage Entrepreneurial Activity (TEA) estimates and weight/variance diagnostics. It was found that these IPW weights are competitive, providing slight improvements over GEM's current calibration method, whereas the alternative approaches proved restrictive and variable dependent. While the validity of any weighting scheme relies heavily on strong structural assumptions, primarily Missing at Random (MAR) and correct model specification, this research shows that the IPW framework offers a viable path to improve statistical reliability and can be readily implemented using existing tools like the `nonprobsvy` package in R.

KEYWORDS

Survey, Non-probability sampling, Inverse Probability Weighting, Mass Imputation, Doubly Robust, Variable selection, `nonprobsvy` package

RESUMO

Os inquéritos não-probabilísticos têm sido amplamente utilizados em investigação devido à sua eficiência e baixo custo, mas sofrem de viés de seleção, tornando-se pouco fiável para generalização direta para a população em geral. Para dar resposta a este problema, no âmbito do programa *European Master in Official Statistics* (EMOS), o Instituto Nacional de Estatística do Luxemburgo (STATEC) propôs a avaliação de um quadro metodológico mais robusto para a correção deste viés. Esta investigação integra a componente não-probabilística CAWI do *Adult Population Survey* (APS) do Luxemburgo de 2024, do projeto *Global Entrepreneurship Monitor* (GEM), com os dados de referência do inquérito *EU Statistics on Income and Living Conditions* (EU-SILC) através de covariáveis comuns. A metodologia compara a Ponderação pelo Inverso da Probabilidade, a Imputação em Massa e os estimadores Duplamente Robustos, com foco nas estimativas da variável TEA (*Total early-stage Entrepreneurial Activity*) e nos diagnósticos de pesos e variância. Os resultados indicam que estes pesos IPW são competitivos, proporcionando ligeiras melhorias face ao atual método de calibração do GEM, ao passo que as abordagens alternativas se revelaram restritivas e dependentes da variável em análise. Embora a validação de qualquer esquema de ponderação dependa fortemente de pressupostos estruturais rígidos, nomeadamente a condição *Missing at Random* (MAR) e a correta especificação do modelo, esta investigação demonstra que a abordagem IPW oferece um caminho viável para melhorar a fiabilidade estatística, podendo ser facilmente implementada através de ferramentas existentes, como o pacote `nonprobsvy` em R.

PALAVRAS-CHAVE

Inquérito, Amostragem não probabilística, Ponderação por Probabilidade Inversa (PPI), Imputação em Massa, estimador Duplamente Robusto, Seleção de variáveis, pacote `nonprobsvy`

TABLE OF CONTENTS

1. INTRODUCTION 1

I. LITERATURE REVIEW 4

2. THE UNIVERSE OF SAMPLING AND SURVEY DATA COLLECTION 5

2.1. Probability Sampling Designs 6

2.1.1. Simple Random Sampling 6

2.1.2. Random-digit Dialling as a telephone sampling/contact frame 7

2.2. Non-probability Sampling Approaches 7

2.2.1. Quota Sampling 8

2.2.2. Online Access Panels as non-probability sample sources 8

3. SURVEY ERRORS 10

3.1. Selection Bias 10

3.2. Non-response Error 10

3.3. Non-coverage Error 12

4. RESURGENCE OF NON-PROBABILITY SAMPLING 14

5. INFERENCE APPROACHES UNDER NON-PROBABILITY SAMPLING 15

5.1. Propensity scores-based approach 16

5.2. Model-based prediction approach 20

5.3. Doubly robust estimation 24

II. CASE STUDY 25

6. SURVEYS 26

6.1. Non-probability sample: Adult Population Survey 26

6.1.1. Methodology of APS in Luxembourg 2024/2025 27
6.2. Reference probability sample: EU-SILC 29
7. DATA HARMONIZATION 31
7.1. Candidate common variables 31
7.2. Missing values 34
8. VARIABLE SELECTION 39
8.1. Stack (selection) model 40
8.2. Outcome model 42
9. APPLICATION OF THE METHODS 43
9.1. R nonprobsvy package 43
9.2. Inverse Probability Weighting 44
9.3. Mass Imputation 47
9.4. Doubly Robust 49
9.5. Comparison between the models 49
10. ANALYSIS OF THE WHOLE GEM's APS 52
10.1. Weight Sharing Method 52
10.2. Analysis of indicators 53
10.3. Bootstrap variance 56
11. STUDY AND COMPARISON WITH DIFFERENT CASE STUDIES 58
11.1. Statistics Canada's crowdsourcing data 58
11.2. US COVID-19 Trends and Impact Survey (CTIS) 59
12. CONCLUSION AND DISCUSSION 61
13. REFERENCES 64
14. APPENDIX 69

1. INTRODUCTION

The study of populations has been a concern since ancient times. Although population counts existed in earlier Eastern civilisations, it was in the Greco-Roman world that censuses became institutionalised as a fundamental tool of governance. These may be considered prohibitively technical because they focus on quantifying births, deaths, and migratory flows. Still, they were linked to the pragmatic needs of state administration, which leads us to view demography not only as numbers but also as a means of understanding our past (Scheidel, 2009).

The census has established itself as the primary and oldest method of data collection. According to the United Nations Statistical Division (2017), the census is a holistic process involving the planning, collection, evaluation and analysis of demographic, economic and social data at a specific point in time and place. This approach is fundamental because it is universal, covering all individuals in a territory and operating at the most detailed geographical level. Due to its ability to provide benchmark statistics, the census is the basic infrastructure for wealth distribution and socio-economic development planning, serving as an indispensable foundation for both public administration and modern scientific research.

With the emergence of new fields of study, the need to conduct more and varied research on the population arises. It is impossible to rely on the census as the sole data collection method in surveys, since it is exhaustive and expensive, to the point that we do not have sufficient resources to conduct a census often enough or for every specific study (United Nations Statistical Division, 2017). Given these considerations, survey sampling emerges as a technically robust and economically viable solution for surveys (Thompson, 2012).

Survey sampling is the process of selecting a sample from a population for observation, enabling inference and parameter estimation about the entire population (Thompson, 2012). According to the United Nations Statistical Division (2017), integrating sampling methods not only reduces costs but can also improve the quality of the collected information. By operating on a smaller scale, it becomes possible to employ teams with specialised training and to use more complex questionnaires, which tends to minimise non-sampling errors. Thus, survey sampling does not replace the census, but complements it, using the census structure as a sampling frame for further investigations that require agility and detail that a total count cannot ensure.

A survey is more than a mere technique; it is a systematic method for gathering information from entities to construct quantitative descriptors of the attributes of a large population. This systematic nature, often reflected in the French term *enquête*, implies a rigorous quest for information that serves as a bridge between theoretical constructs and empirical measurement. Under this paradigm, the quality of the data is no longer measured solely by the size of the observation, but by its 'fitness for use', the accuracy and timeliness required to achieve a specific research purpose (Wolf *et al.*, 2016).

Probability sampling began to gain acceptance among statisticians in the 1940s and is based on the existence of a list in which each element of a finite population has a unique identifier; it involves selecting samples using mathematical rules that allow the probabilistic characteristics of the sample to be rigorously established. This approach has become the most advantageous option and the preferred choice for statisticians seeking sound, valid statistical inference, as it offers broad applicability and requires minimal assumptions. The robustness of this method lies in the fact that it treats unknown population values as fixed, ensuring that all statistical statements and the creation of 'unbiased' estimators derive exclusively from the probabilities established by the sample design itself, thereby providing a formal basis for design-based inference, subject to coverage, response, measurement, and processing errors (Fuller, 2011).

The very success of probability sampling led to a new problem: surveys became more ambitious, with long, sophisticated questionnaires, which, together with privacy concerns and the burden of responding, led to substantial declines in response rates in many surveys and polling contexts. This difficulty, coupled with the 'ever-increasing costs' of probabilistic sampling and technological advances, forced a paradigm shift. Now, government agencies, research institutes, and industrial organisations are seeking cheaper, faster alternatives for data collection. These alternatives include administrative data, social networks, and non-probabilistic samples. Although the need for non-probability sampling methods has increased in this context, they have long been used as an economical alternative when resources were scarce. These methods, on their own, had lost momentum, especially in "official statistics", due to the lack of a "theoretical foundation" that enabled statistical inference and the success and dominance of widely accepted probabilistic sampling methods (Wu & Thompson, 2020).

This dissertation aims to provide an alternative to address the lack of a theoretical framework for statistical inference in non-probability sampling by complementing it with a reference probabilistic sample that shares common variables with the non-probability sampling. The methods tested by (Wu, 2022) are Inverse Probability Weighting (IPW), Mass Imputation (MI), and Doubly Robust (DR). This study focused primarily on the first, specifically in the context of the Adult Population Survey (APS) of the Global Entrepreneurship Monitor (GEM) of 2024 in Luxembourg. The methods will be implemented in R, primarily using the specialised `nonprobsvy` package. The work consist of applying and comparing different statistical adjustment approaches, including weighting and imputation methods, to determine which technique best improves statistical inference in non-probabilistic data sources. This dissertation evaluates whether probability-sample-assisted methods can reduce selection bias in the CAWI component of the Luxembourg APS when estimating the total entrepreneurial activity (TEA) among adults aged 18-64.

To assess the empirical performance of the methods, estimates were compared using diagnostic criteria and available benchmark distributions. The main research question is: "How do IPW, MI, and DR estimators compare with the current GEM calibration in terms of TEA estimates, standard errors, design effects, weight stability, and agreement with official benchmark distributions?"

I. LITERATURE REVIEW

2. THE UNIVERSE OF SAMPLING AND SURVEY DATA COLLECTION

Statistical inference is at the heart of data science, defined as the ability to analyse data and draw conclusions about unknown populations (Wu & Thompson, 2020). More than just a description, inference provides the essential tools for processing information and, crucially, quantifying remaining uncertainty through the formulation of statistical models (Held & Sabanés Bové, 2014). In the context of social sciences, this inference does not exist in a vacuum, it is the core of Survey Methodology. This field acts as an interdisciplinary bridge, where statistical rigour meets the preoccupations of sociology, psychology, and economics to transform raw data into "quantitative descriptors" of society (Groves *et al.*, 2009). Survey sampling is the process of selecting a sample from a population for observation, with the primary objective of estimating the population's characteristics. This discipline is not limited to data collection but covers the entire spectrum of how data is selected, whether the process is under the direct control of the researcher or determined by chance, and how that information is used to make inferences about the population of interest (Kish, 1995).

In a typical setup, the population is defined as a finite and known number of N units (such as individuals or companies). Each unit is associated with a value y (the variable of interest), which is treated as a fixed quantity rather than a random variable. The effectiveness of this process depends on the sampling design and the procedure used to select the sample from the population (Thompson, 2012). The application of this design depends on the availability of a sampling frame, ideally a complete and exhaustive list of units in the target population. In practice, it is often difficult to obtain a frame that exactly matches the population of interest. Even when sampling goes beyond finite lists and uses geographical units, the fundamental principles of inference on random selection and unbiased estimates remain crucial for scientific validity (Groves *et al.*, 2009).

This structure establishes the basis for distinguishing sampling methods by their selection mechanisms. In probabilistic sampling, the inclusion probabilities are known and defined by the sample design; in non-probabilistic sampling, this mechanism is unknown and not controlled by the researcher (Kish, 1995).

Although there is a wide variety of approaches, this section does not aim to review all available methods; rather, it presents the core concepts necessary for the empirical study. A conceptual distinction is made between sampling designs (the rules defining how individuals are selected) and contact strategies (the practical infrastructure used to reach them, such as Random Digit Dialling and Access Panels). These strategies are highlighted because the APS survey uses both for its sample collection.

2.1. Probability Sampling Designs

At the heart of probabilistic sampling is random selection, which reduces the likelihood of obtaining an unrepresentative sample and provides a reliable way to mitigate unforeseen bias. The selection process must be based on a random mechanism that assigns each unit a specified probability of selection. The rigour of this method is based on the prior definition of accuracy requirements, which are decisive in calculating the sample size (United Nations, 2008).

The validity of inference in probabilistic sampling requires that all selected units be properly enumerated. In the analysis and estimation of population parameters, scientific rigour extends to data processing, where each unit or element must be weighted according to its probability of selection. This weighting requirement ensures that the conclusions drawn from the sample can be generalised to the whole with a quantifiable level of uncertainty. What distinguishes the different types of probabilistic sampling is, fundamentally, the design structure and the way randomness is applied in selecting units (United Nations, 2008).

2.1.1. Simple Random Sampling

According to Wu & Thompson (2020), Simple Random Sampling (SRS) is the most basic probabilistic sampling design and serves as the baseline for all other more complex methods. Its core principle is the equiprobability of selection: every possible sample of size n has exactly the same probability of being selected from the sampling frame. This design ensures that the sample mean remains a **design-unbiased estimator** of the population parameters.

This method can be implemented through two main procedures:

- **SRS without replacement:** Units are selected one by one and are not picked again. This is the preferred method in survey research as it is more statistically efficient, incorporating a finite population correction that reduces the variance of the estimates.
- **SRS with replacement:** A unit may be selected more than once. Although easier to handle, in practice, it is less efficient, especially when the population N is small relative to the sample n.

2.1.2. Random-digit Dialling as a telephone sampling/contact frame

The Random Digit Dialling (RDD) method involves conducting surveys by calling telephone numbers generated at random by a computer. This methodology is used because it supports probability-based sampling of telephone numbers, although coverage and nonresponse errors may still affect inference, and it reaches both listed and unlisted numbers. RDD has been associated with high response rates, a crucial factor in minimising bias. However, the method presents efficiency and cost challenges, as the random mechanism often generates numbers that are irrelevant to the sample, such as invalid numbers, resulting in a significant expenditure of interviewer time and resources on unproductive calls (Schonlau *et al.*, 2004).

According to Groves *et al.* (2009), mobile phones are replacing landline services, resulting in greater coverage loss among the younger population and those moving out of their parents' homes. The major advantage of this replacement is that mobile phones are usually associated with a single person. In contrast, landlines are often associated with households (clusters of persons), making the sampling frame closely aligned with the target population. On the other hand, if the researcher wishes to target a specific region, landlines have the advantage of indicating respondents' regions of residence.

2.2. Non-probability Sampling Approaches

In non-probability sampling, the selection of units does not rely on a random mechanism, meaning the inclusion probabilities remain unknown. Unlike the framework of probability sampling, non-probability methods require a theoretical basis or an explicit set of modelling assumptions to enable statistical inference from the sample to the target population. Without such framework, these methods can be considered insufficient for scientific inference, as

they lack the necessary mechanisms to quantify uncertainty or mitigate selection bias (Baker *et al.*, 2013).

2.2.1. Quota Sampling

According to Moser (1952), quota sampling is a foundational non-probability method in market research and opinion polls. In quota sampling, the choice of final units is left to the interviewer's discretion, who must fill previously defined quotas.

The process is divided into two phases (Moser, 1952):

- **Quota Definition:** Based on census data or official statistics, the sample structure is determined (e.g., how many men/women, what age groups).
- **Field Selection:** The interviewer receives their "target" and is free to choose whom to approach to fill these cells.

There are three standard controls used by almost all organisations: Gender, Age, and Social Status. Interviewers frequently determine a respondent's social class based on observable characteristics like occupation and appearance, introducing what is known as interviewer bias (Moser, 1952).

Also, quota surveys often take place on the street or in workplaces. This leads to a location bias, as people interviewed at the door of a factory will be different from people interviewed in a park or an office (Moser, 1952).

2.2.2. Online Access Panels as non-probability sample sources

Online access panels are groups of recruited respondents who agree to participate in web-based surveys. This research has grown due to technological advances and the increasing growth of the Internet in everyday life. Recruitment usually occurs through non-probability methods, such as self-selection. Institutions can use online channels (pop-ups, web ads) or offline channels (invitations in traditional surveys) to attract new members. The use of these panels offers logistical benefits compared to traditional methods such as face-to-face interviews, including speed of collection, cost efficiency, and automation (rapid analysis of

results due to automation). However, one of the challenges of access panels is the extent to which the results obtained can be extrapolated to the population (Loosveldt & Sonck, 2008).

The validity of access panel data is affected by factors that compromise its comparability with traditional probabilistic methods. Limited Internet access and self-selection in recruitment are the primary sources of bias. The panel ends up being composed of individuals with specific digital skills and availability to respond to surveys. In addition, the absence of an interviewer and the way questions are presented influence responses. In online mode, respondents control the pace of the interview, which can lead to "satisficing" effects. The social distance between the respondent and the computer reduces the pressure to provide socially acceptable responses, as seen in sensitive topics such as attitudes towards immigrants, where online panels reveal more negative opinions than face-to-face interviews (Loosveldt & Sonck, 2008).

To bring the panel data closer to the reality of the general population, post-stratification and propensity score adjustment techniques are often applied. However, the study by Loosveldt & Sonck (2008), showed that post-stratification corrects demographic proportionality but does not necessarily guarantee the representativeness of substantive attitudes.

3. SURVEY ERRORS

Across the studies conducted so far, it has been shown that errors and biases are inherent to the survey process, regardless of the specific design employed. The statistician could adopt a rigid stance of total opposition to bias, as if it were a sin, and use only mathematically unbiased estimators. However, the approach we should adopt does not seek the 'absence of error,' but rather the intelligent management of the errors (Kish, 1995).

The effects of biases and variable errors differ depending on the statistic being analysed. Biases that have large effects on the mean of a total sample may have a negligible impact on comparisons of means or regression statistics. Conversely, variable errors that are insignificant for large sample means may have substantial effects on comparisons of small subclasses or in regression analyses (Kish, 1995). The next points present the most relevant examples of survey errors in this study.

3.1. Selection Bias

Selection biases are directly linked to problems in sample selection procedures. The most striking feature of this error is that it can distort results across all sample sizes with the same intensity. Generally, this bias arises when statistical calculations are not adjusted to account for unequal selection probabilities. For example, if one adult is selected from each household, people who live alone are more likely to be chosen than those in large families; if the statistician does not correct for this, the final data will be biased. Fortunately, these errors can often be corrected through weighting during data analysis (Groves *et al.*, 2009).

3.2. Non-response Error

Nonresponse is one of the most critical sources of failure in obtaining data from elements that were correctly selected for the sample. Kish (1995) argues that reporting and accurately measuring nonresponse rates are fundamental responsibilities of the researcher, as they are indispensable for assessing the quality of the survey and predicting the impact of systematic errors in future studies.

Non-response bias is determined not only by the number of people who did not respond, but by the systematic difference between them and those who did respond. While this bias can be devastating for estimating population totals (aggregates), it may have less effect on averages or comparisons between subclasses if non-response patterns are similar between the groups being compared. Willingness to respond is influenced by deep demographic and social factors. Residents of large cities and higher socioeconomic classes tend to be more difficult to contact and have higher refusal rates. On the other hand, families with young children are easier to find, while single or divorced people are more elusive. Older people have higher refusal rates, while young newlyweds are more cooperative. Success depends on the respondent's motivation, the guarantee of anonymity, and the interviewer's ability to spark interest from the very first moments of contact (Kish, 1995).

Kish (1995) proposes several methodologies to mitigate the problem:

- **Focus on Eligible Respondents:** Non-response rate should be calculated on respondents who belong to the target population, excluding ineligible units.
- **Sample Adjustment:** It is possible to anticipate non-response and non-coverage by increasing the initial sample size to ensure that the final number of interviews reaches the planned target.
- **Criticism of Substitution:** Kish warns that substituting respondents is often a technical error, as substitutes tend to resemble those who have already responded more than the profile of those who refused.
- **Optimism about Postal Surveys:** Despite the common belief that postal surveys have low response rates, the author cites examples (such as the British Social Survey) where rates of over 90% were achieved through persistent procedures and advance notices.

While Kish's strategies remain important for survey administration, contemporary literature highlights that the response rate alone is no longer the sole primary indicator of survey quality. The American Association for Public Opinion Research (AAPOR) provides standard definitions that distinguish between response and cooperation rates, enabling consistent calculations across different surveys. However, AAPOR also emphasizes that evidence of nonresponse bias must be examined even in surveys with acceptably high response rates, recommending that survey outcomes be assessed through additional indicators, such as insignificant levels of bias and low levels of missing data (*Response Rates Calculator - AAPOR*, 2022). This modern perspective is strongly supported by Groves & Peytcheva (2008), whose

meta-analysis found that the nonresponse rate, by itself, is a poor predictor of nonresponse bias. Emphasizing that nonresponse bias can vary across different estimates within the same survey, as it depends heavily on whether the specific survey variable is correlated with the respondent's propensity to participate.

Call-backs (new attempts to make contact) are considered the most effective method for reducing the non-response rate, and the number of recontacts does not need to be uniform. The first visit is crucial not only to secure the interview but also to gather information about the respondent's schedule, which makes the second and third visits more efficient in terms of 'success per call'. Although initial calls absorb the fixed costs, the unit cost per response tends to rise sharply from the fourth or fifth attempt onwards, due to longer journeys to find the few remaining respondents (Gendall & Davis, 1993).

3.3. Non-coverage Error

Noncoverage is considered one of the most complex and silent challenges in sampling. It is classified as a non-observation error, resulting in the failure to obtain data from parts of the population that should, by right, be included in the sample. Its fundamental characteristic is that excluded elements have a zero probability of being selected, meaning they are technically "invisible" to the researcher from the outset. It is crucial to distinguish this error from non-response: while in non-response, the individual was selected but did not respond, in non-coverage, they were never included in the sampling frame to be chosen (Lee, 2006).

Kish (1995) identifies that this problem arises from two sources: incomplete sampling frames (errors in preparing the list) and missing units (failures in executing field procedures.

Non-coverage has different effects depending on what you want to measure. It is devastating for estimating the total size of a population, because if people are missing from the list, the final total number will inevitably be underestimated. On the other hand, the impact may be less on average if the excluded people are similar to those included. However, if the excluded have very different characteristics (e.g., poorer or nomadic families), the sample average will be biased (Kish, 1995).

The biggest problem with non-coverage is that it is very difficult to measure internally. Unlike non-response, where we know who did not respond, with non-coverage, we do not know who is missing. Measurement sometimes requires an external check, comparing sample data with independent and reliable sources, like national censuses or official records (Kish, 1995).

To reduce these errors, Kish (1995) suggests several operational improvements, such as:

- Using larger, more compact geographic segments to facilitate the identification of boundaries on the ground.
- Implementing linking procedures (such as the half-open interval) to capture new units that have emerged since the list was created.
- Performing independent quality checks by different teams ("blind listings").
- Applying statistical weighting when exclusion rates in certain areas are known.

4. RESURGENCE OF NON-PROBABILITY SAMPLING

The landscape of survey methodology is currently undergoing a significant paradigm shift. For decades, probability sampling and its design-based inference remained the gold standard; however, escalating costs and a dramatic decline in response rates have forced a re-evaluation of data collection strategies. In this context, non-probability survey samples have emerged as an alternative (Wu & Thompson, 2020).

However, this resurgence presents a fundamental methodological paradox. As previously established, non-probability samples lack the theoretical foundation required for design-based inference, often leading to biased estimates due to unknown selection probabilities. This raises a critical question: if these methods are statistically inferior regarding representativeness, how can their widespread adoption be justified in rigorous research? A practical solution does not lie in substituting probability sampling, but rather in complementing it. This necessity has spurred the development of advanced mixed-mode designs.

While the concept of 'mixed-mode' is frequently used to describe survey designs that offer respondents multiple participation channels (e.g., offering both web and telephone options), driven by the dual need to resolve coverage issues, reaching populations inaccessible via a single frame, and to mitigate non-response bias, the focus of this work lies on a distinct form of integration. Specifically, this study addresses the integration of a probability sampling reference sample with a non-probability sample by leveraging common variables shared between both datasets. The primary objective of this integration is to overcome the limitations regarding statistical inference and validity estimates inherent in non-probability designs by complementing them with a probabilistic component.

5. INFERENCE APPROACHES UNDER NON-PROBABILITY SAMPLING

The central purpose of any statistical study is to obtain a reliable estimate of a parameter that characterises the total population. However, in contexts where data are generated by non-probabilistic mechanisms, such as online access panels, the inclusion probabilities of individuals referred to as propensity scores (π_i^A) in the sample are unknown (Wu, 2022). They can be represented by the following equation

$$\boldsymbol{\pi_i^A} = P\,(\, i \in S_A \mid x_i, y_i) = \; P\,(\, R_i = 1 \mid x_i, y_i\,)\,, \qquad i = 1, 2, \ldots, N$$

where:

- $\boldsymbol{i}$**:** index for population units ($i = 1, 2, \ldots, N$);
- $\boldsymbol{S_A}$: the non-probabilisty sample ($S_A \subset \{1, 2, \ldots, N\}$);
- $\boldsymbol{x_i}$**:** vector of auxiliary or confounding variables (covariates) associated with unit i;
- $\boldsymbol{y_i}$**:** study variable or outcome of interest for unit i;
- $\boldsymbol{R_i}$**:** response or participation indicator variable. $R_i = 1$ if the individual is part of the non-probability sample, $R_i = 0$ if they belong to the population but were not sampled.

Since these probabilities are often determined by the individuals themselves through self-selection, they remain unknown to the researcher and can be estimated using the propensity score. The initial formulation allows participation to depend on both x_i and y_i; the subsequent MAR assumption is precisely what permits modelling participation as a function of observed auxiliary variables only (Wu, 2022).

These probabilities are unknown, and in order to calculate them and correct for sample bias, we need to assume the following conditions (Wu, 2022):

- **Missing at random (MAR):** The sample participation indicator R is independent of the study variable y, conditional on the auxiliary variables x;
- **Positivity:** All individuals in the population must have a non-zero probability of being included in the sample;
- **Independence:** Individual participations are independent of each other;

- **Reference sample:** The probabilistic sample must contain the same auxiliary variables x as the non-probabilistic sample.

In this section, the focus will be on the following three main approaches from the study (Wu, 2022) available in the `nonprobsvy` package.

5.1. Propensity scores-based approach

In the propensity scores-based approach, the goal is to estimate the propensity for an individual to be selected in the non-probabilistic sample S_A. Under the MAR assumption, we can assume that these probabilities are a function of the auxiliary variables (x_i) as so

$$\boldsymbol{\pi_i^A} = P\left(R_i = 1 \mid x_i\right) = \pi(x_i)$$

However, since the exact mechanism of participation is unknown, three popular parametric forms are assumed to model this function.

a. Inverse Logit function

$$\boldsymbol{\pi_i^A} = 1 - \{1 + \exp(x_i'\alpha)\}^{-1} = \frac{\exp(x_i'\alpha)}{1 + \exp(x_i'\alpha)}$$

b. Inverse Probit function

$$\boldsymbol{\pi_i^A} = \Phi\left(x_i'\alpha\right)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the normal distribution with mean 0 and variance 1.

c. Inverse complementary log-log function

$$\boldsymbol{\pi_i^A} = 1 - \exp\{-\exp(x_i'\alpha)\}$$

After selecting one of the parametric forms for modelling the participation mechanism, the next step is to estimate the unknown parameters α of the model, a process that is usually performed using the pseudo-maximum likelihood method, written as follows:

$$\boldsymbol{\ell}(\boldsymbol{\alpha}) = \log\left\{\prod_{i=1}^{N}\left(\pi_i^A\right)^{R_i}(1-\ \pi_i^A)^{1-R_i}\right\} = \sum_{i\epsilon S_A}\log\left(\frac{\pi_i^A}{1-\ \pi_i^A}\right) + \sum_{i=1}^{N}\log(1-\ \pi_i^A)$$

In this formula, the first term adds only the units that participated in the non-probability sample (S_A), while the second term adds all units in the population (N).

As complete population information is rarely available, the probabilistic reference sample (S_B) is used to estimate the population sum. Thus, the log-likelihood is estimated by:

$$\boldsymbol{\ell}^*(\boldsymbol{\alpha}) = \sum_{i\epsilon S_A}\log\left(\frac{\pi_i^A}{1-\ \pi_i^A}\right) + \sum_{i\epsilon S_B} d_i^B\log(1-\ \pi_i^A)$$

In this version, the population sum is estimated using the sample S_B, using its respective design weights d_i^B, which represent the inverse of the probability of selecting each individual in a probability sample, indicating how many units of the population are represented by each sample element, allowing the sample data to be expanded to the population level. The estimator $\hat{\alpha}$ is the value that maximises the pseudo-likelihood function, obtained by solving the pseudo-score function $U(\alpha) = \frac{\partial \ell^*(\alpha)}{\partial \alpha} = 0$. If we use the inverse logit function, the pseudo-score function is

$$\boldsymbol{U}(\boldsymbol{\alpha}) = \sum_{i\epsilon S_A} x_i - \sum_{i\epsilon S_B} d_i^B \pi(x_i,\alpha)x_i$$

This estimator $\hat{\alpha}$ is considered consistent, which means the estimator gets closer to the true population value as the sample size increases. This consistency is a large-sample property that holds under the assumed participation mechanism and standard regularity conditions.

The pseudo-score equations $U(\alpha) = 0$ can be replaced by a system of general estimation equations (GEE), where $h(x,\alpha)$ is a user-specified vector of functions with the same dimension as α. The general form of these equations is defined by:

$$\boldsymbol{G}(\boldsymbol{\alpha}) = \sum_{i\epsilon S_A} h(x_i,\alpha) - \sum_{i\epsilon S_B} d_i^B \pi(x_i,\alpha)h(x_i,\alpha)$$

Under the required regularity and identification conditions, suitable choices of $h(x, \alpha)$ the estimator $\hat{\alpha}$ obtained by solving $G(\alpha) = 0$ can yield consistent estimators. Despite this, the method has some disadvantages compared to the maximum likelihood method, such as lower efficiency, since the estimator $\hat{\alpha}$ is typically less efficient than that based on pseudo-score functions, due to the theoretical optimality of the maximum likelihood estimator, and instability, since limited empirical results suggest that the solution for $G(\alpha) = 0$ may become unstable depending on the choice of functions $h(x, \alpha)$.

If the user specifically chooses $h(x, \alpha) = x/\pi(x, \alpha)$, the equation reduces to:

$$\boldsymbol{G(\alpha)} = \sum_{i \in S_A} \frac{x_i}{\pi(x_i, \alpha)} - \sum_{i \in S_B} d_i^B x_i$$

This formula is described as a 'distorted' version of the original pseudo-score functions. Its practical importance lies in the fact that the equation requires only the estimated population totals of the auxiliary x.

On the other hand, non-parametric approaches offer a more flexible alternative by not imposing strict restrictions on the relationship between auxiliary variables and the probability of participation. In this approach, the propensity scores (π_i^A) are supposed to be related to the variable x_i but we don't impose any explicit formula.

- **Kernel regression estimator:** If we had information for the finite population, and if we let $K_h(t) = K(t/h)$ be a chosen kernel with a bandwidth h, we can define the Nadaraya-Watson for $\pi(x)$ as

$$\widetilde{\boldsymbol{\pi}}(\boldsymbol{x}) = \frac{\sum_{j=1}^{N} K_h(x - x_j) R_j}{\sum_{j=1}^{N} K_h(x - x_j)}$$

- **Nonparametric kernel regression estimator:** As we do not have complete information for the entire population, the formula is adapted. The numerator uses only the non-probabilistic sample (S_A) (since R_j=0 outside it) and the denominator is estimated using the reference sample (S_B) with $i \in S_A$:

$$\widehat{\boldsymbol{\pi}}_{\boldsymbol{i}}^{\boldsymbol{A}} = \hat{\pi}(x_i) = \frac{\sum_{j \in S_A} K_h(x_i - x_j)}{\sum_{j \in S_B} d_j^B K_h(x_i - x_j)}$$

This estimator is consistent and very flexible due to its non-parametric nature. However, it may face computational and precision difficulties when the dimension of $\boldsymbol{x}$ is high.

After obtaining the propensity score estimates ($\hat{\pi}_i^A$) using the parametric or non-parametric methods discussed above, it is possible to estimate the population mean (μ_y) using the Inverse Probability Weighting (IPW) technique. There are two main versions of the estimator constructed based on the estimate $\hat{\pi}_i^A = P(i \in S_A | x_i)$

- **Horvitz-Thompson estimator ($\widehat{\boldsymbol{\mu}}_{\boldsymbol{IPW1}}$):** Uses the real size of the population

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{IPW1}} = \frac{1}{N} \sum_{i \in S_A} \frac{y_i}{\hat{\pi}_i^A}$$

- **Hájek estimator ($\widehat{\boldsymbol{\mu}}_{\boldsymbol{IPW2}}$)**: Uses the estimated size of the population ($\widehat{N}^A$) where $\widehat{N}^A$ is the sum of the inverse weights ($\widehat{N}^A = \sum_{i \in S_A} (\hat{\pi}_i^A)^{-1}$)

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{IPW2}} = \frac{1}{\widehat{N}^A} \sum_{i \in S_A} \frac{y_i}{\hat{\pi}_i^A}$$

Although the Horvitz-Thompson estimator is fundamental in sampling theory, the Hájek estimator usually performs better even in cases where the total population size (N) is known.

To establish the large-sample properties of these estimators, first must be examined an ideal population-level version based on the true propensity scores, defined as $\tilde{\mu}_{IPW} = N^{-1} \sum_{i \in S_A} y_i / \pi_i^A$. In practie, it cannot be calculated because it uses real propensity scores and not estimates, but allows us to demonstrate that $E_q(\tilde{\mu}_{IPW}) = \mu_y$ and $V_q(\tilde{\mu}_{IPW}) = O(n_A^{-1})$. Under a correctly specified propensity scores model, $\tilde{\mu}_{IPW}$ converges to μ_y when the sample size $n_A \to \infty$.

5.2. Model-based prediction approach

From a different perspective, the researcher gives up trying to understand why people responded or did not respond. Instead, he focuses on creating a strong predictive model. He uses the non-probabilistic sample S_A to train a model of how characteristics (x) affect the response (y). Then he applies this rule to the entire probabilistic sample S_B to "predict" what the entire population would respond. Using semiparametric outcome regression models, the goal is to develop a formula that predicts the response y based on the person's characteristics x. The model is defined by two parts:

- **Conditional Mean:** $\boldsymbol{E(y_i|x_i) = m(x_i, \beta)}$: This mean, given this person's profile (x_i), what response do we expect from them on average?" The function m is the mathematical formula and β is the vector of the model coefficients.
- **Conditional Variance:** $\boldsymbol{V(y_i|x_i) = v(x_i)\sigma^2}$**:** This measures the margin of error, that is, how much the actual responses vary around our mathematical prediction.

To estimate β (the model coefficients), a quasi-maximum likelihood estimator is used, where the estimate of β is a solution to the following equation:

$$\boldsymbol{S(\beta)} = \sum_{i \in S_A} \frac{\partial m(x_i, \beta)}{\partial \beta} \{v(x_i)\}^{-1}\{y_i - m(x_i, \beta)\} = 0$$

The process begins with the sum applied exclusively to individuals who are part of the non-probabilistic sample (S_A). The fundamental component of the equation is the residual, defined as the difference between the actual observed value (y_i) and the value predicted by the model $m(x_i, \beta)$, to minimise this overall error. For the estimate to be accurate, the formula incorporates a partial derivative ($\frac{\partial m(x_i, \beta)}{\partial \beta}$), which acts as the gradient (or Jacobian). It defines the sensitivity of the mean function ($m(x_i, \beta)$) with respect to the parameters (β). Essentially, it projects the error (residual) back onto the parameter space, indicating the exact direction and magnitude by which β must be adjusted to optimise the model. The inverse variance weight ($\{v(x_i)\}^{-1}$) is used to address heteroscedasticity, thereby increasing the estimator's robustness by assigning less weight to observations with higher variance. By equalling this sum to zero, it adjusts the values of β until this error is minimised. This

mathematical framework forms the basis of Generalized Linear Models (GLMs), allowing for flexible estimation depending on the nature of the outcome variable.

After training the model on the non-probabilistic sample S_A, we can use one of these two estimators:

i.

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{y1}} = \frac{1}{N}\sum_{i=1} \widehat{m}_i$$

This estimator is based entirely on the model prediction $\widehat{m}_i$ for all individuals i in the population. It is calculated as the weighted average of the predictions ($\widehat{m}_i$) that represents the predicted (or imputed) value of the variable of interest y_i for individual i, calculated by the model based on that person's characteristics (x_i) made for each of the units in the probabilistic sample S_B. In the context of a linear regression model where $m(x, \beta) = x'\beta$. This estimator simplifies to the product between the vector of weighted means of the auxiliary variables ($\hat{\mu}_x$) and the estimated coefficients of the model ($\hat{\beta}$):

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{y1}} = \hat{\mu}'_x\hat{\beta}$$

ii.

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{y2}} = \frac{1}{N}\left\{\sum_{i\epsilon S_A} y_i - \sum_{i\epsilon S_A} \widehat{m}_i + \sum_{i=1} \widehat{m}_i\right\}$$

The construction of this second estimator follows a model-assisted imputation framework, partitioning the target population into observed and unobserved domains. To mitigate bias arising from potential model misspecification, the estimator directly integrates the empirically observed values (y_i) from the non-probabilistic sample (S_A). Concretely, it computes the baseline population estimate by projecting the model predictions over the probability reference sample and subsequently applies a bias-correction term derived from the empirical residuals (y_i - $\widehat{m}_i$) of S_A. Consequently, the statistical model is utilised exclusively to impute outcomes for the unobserved population fraction, ensuring the final estimate is robustly calibrated to the collected data. Using the structure of a linear regression model, this estimator can be rewritten as follows

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{y2}} = \frac{n_A}{N}\left(\bar{y}_A - \bar{x}'_A\hat{\beta}\right) + \mu'_x\hat{\beta}$$

where

- $\bar{\boldsymbol{y}}_{\boldsymbol{A}}$**:** is the sample mean of the variable of interest in the non-probability sample.
- $\bar{\boldsymbol{x}}_{\boldsymbol{A}}$**:** is the vector of the means of the auxiliary variables in the non-probability sample.

In this context, the Mass Imputation (MI) estimator consists of the practical application of this modelling logic, in which the parameters are estimated in the non-probabilistic sample (S_A) are used to predict the values of the variable of interest for all individuals in a probabilistic reference sample (S_B). The premise is to treat the study variable y as missing data for all units i belonging to the reference probability sample S_B. The MI approach uses the same logic as model-based prediction methods but allows the use of more flexible techniques developed for missing-data problems. The effectiveness of this approach is based on the assumption of model transportability, meaning that the model was constructed based on the non-probabilistic sample (S_A) can be validly used for prediction in the probabilistic reference sample (S_B).

Let y_i^* be the imputed value to replace the missing y_i in the S_B sample. The MI estimator is defined by the following equation

$$\hat{\mu}_{yMI} = \frac{1}{\widehat{N}_B}\sum_{i\in S_B} d_i^B\, y_i^*$$

where

- $\boldsymbol{d_i^B}$**:** are the design weights of the probabilistic reference sample.
- $\widehat{\boldsymbol{N}}_{\boldsymbol{B}}$**:** It is the estimate of the population size, defined as the sum of the design weights. ($\widehat{N}_B = \sum_{i\in S_B} d_i^B$).

There are two main ways to generate the imputed values y_i^*:

- **Parametric Approach- Deterministic Regression Imputation:** The imputed value is defined as $y_i^* = m\left(x_i, \hat{\beta}\right)$, where $\hat{\beta}$ is a consistent estimator obtained from

the non-probabilistic sample S_A. Under a linear model where $y_i^* = x_i'\hat{\beta}$, the estimator $\hat{\mu}_{yMI}$ reduces to $\hat{\mu}_x'\hat{\beta}$.

- **Nonparametric Approach- Sample Matching Method (Nearest Neighbour):** For each $i \in S_B$, the value of y is imputed as $y_i^* = y_j$, where $j \in S_A$ is a "donor" selected for being the closest neighbor to i in terms of distance between their auxiliary variables x_i and x_j.

Another approach described in Czerniawska *et al.* (2026) is the Predictive Mean Matching (PMM) estimator, a technique that combines both NN (Nearest Neighbour) and parametric (or nonparametric) imputation estimators. Instead of calculating the direct distance between covariates, PMM uses a distance function weighted by the estimated regression coefficients ($\hat{\beta}$). In the case of GLMs, this distance is based on the inverse link function of the predicted values. According to the literature, despite the algorithmic similarity to the NN method, the properties of its imputations differ significantly. The two main advantages of PMM over NN are that it does not suffer from the "curse of dimensionality," since the assumptions for pairing predicted values are, in fact, more easily met with a larger number of covariates, and it does not require the strict positivity assumption. By projecting a high-dimensional space into a single predictive score, PMM prevents the erratic matching that inflates variance in standard NN. Furthermore, because the model's prediction is used solely as a distance metric to select an actual observed donor, it inherently provides robustness to model misspecification and ensures stability.

In addition to the MI, the literature presents alternative approaches and extensions that aim to optimise the use of available information or deal with different types of data. One such approach is reverse sample matching, which arises from the need to take advantage of the actual observed values (y_i) in the non-probabilistic sample, rather than using them solely to calibrate the model; this method reverses the traditional logic by transferring the design weights from the reference sample to the units in the biased sample via nearest-neighbour matching or kernel weighting. Other approaches include fractional imputation, recommended for binary or ordinal variables, and multiple imputation (MI^2), which proposes creating multiple sets of imputed data to preserve the conditional distribution of the study variable and enable more general statistical inferences from a fully completed reference sample.

5.3. Doubly robust estimation

Although the IPW technique depends solely on the validity of the propensity model, and mass imputation depends on the accuracy of the outcome regression model, Doubly Robust (DR) estimation combines both approaches to protect against incorrect specification of either model. It simultaneously uses the propensity score model and the outcome regression model. The general theoretical form of the estimator is defined by

$$\tilde{\mu}_{DR} = \frac{1}{N}\sum_{i\epsilon S_A}\frac{y_i - m_i}{\pi_i^A} + \frac{1}{N}\sum_{i=1}^{N} m_i$$

The first term is an adjustment based on the propensity scores using the prediction error $\epsilon_i = y_i - m_i$ and the second term represents the prediction based on the model μ_y. The estimator $\tilde{\mu}_{DR}$ is consistent if one of the two models is correctly specified.

There are two adapted estimators when using the probabilistic reference sample (S_B):

- **Estimator DR1:** Uses the population size N

$$\hat{\mu}_{DR1} = \frac{1}{N}\sum_{i\epsilon S_A}\frac{y_i - \hat{m}_i}{\hat{\pi}_i^A} + \frac{1}{N}\sum_{i\epsilon S_B} d_i^B \hat{m}_i$$

- **Estimator DR2:** Uses the estimated population sizes. It's recommended to use this since it has better performance in terms of bias and mean squared error (MSE)

$$\hat{\mu}_{DR2} = \frac{1}{\widehat{N}_A}\sum_{i\epsilon S_A}\frac{y_i - \hat{m}_i}{\hat{\pi}_i^A} + \frac{1}{\widehat{N}_B}\sum_{i\epsilon S_B} d_i^B \hat{m}_i$$

where d_i^B are the design weights for the probability sample S_B, $\widehat{N}_A = \sum_{i\epsilon S_A}\left(\hat{\pi}_i^A\right)^{-1}$ and $\widehat{N}_B = \sum_{i\epsilon S_B} d_i^B$.

II. CASE STUDY

6. SURVEYS

To implement and test the previous methods, we need at least two surveys: a non-probabilistic sample, which, by its very nature, leads us to conclude that it lacks the properties required for accurate statistical inference, and a reference probabilistic sample that shares common variables with the non-probabilistic sample and will be used as a benchmark to overcome this inferential limitation. For this matter, we decided to use the main study of the GEM, the APS, as our non-probabilistic sample, and the EU-SILC survey as our reference probabilistic sample, both in 2024 and in Luxembourg.

6.1. Non-probability sample: Adult Population Survey

Founded in 1999, GEM is the largest and longest-running longitudinal study on entrepreneurship worldwide. The project is owned by the Global Entrepreneurship Research Association (GERA), a UK-based non-profit funded by sponsoring institutions and fees from National Teams. Operationally, these National Teams represent the project locally, managing annual data collection, producing "National Reports," and handling communications within their territories. Conceptually, GEM posits that entrepreneurship is not an isolated choice driven solely by income, but the result of an "ecosystem" where social attitudes often outweigh access to capital (Hill *et al.*, 2025).

According to Hill *et al.* (2025), at the heart of the GEM methodology are two annual surveys: the APS, which collects primary data from at least 2,000 respondents aged 18–64 per country, and the National Expert Survey (NES). This study will focus on the APS and its conduct in Luxembourg. The APS is the primary instrument of the GEM project, designed to capture the entire spectrum of entrepreneurial activity. Unlike official statistics, which often count only registered companies, the APS surveys a representative sample of the adult population, enabling measurement of the entrepreneur's perspective within the wider ecosystem. The questionnaire has an architecture based on Blocks (segmentation logic where questions are filtered by respondent profile) and Modules (*GEM Global Entrepreneurship Monitor*, n.d.).

While operational execution is decentralised and managed by National Teams or independent vendors, technical control remains strictly centralised. To ensure international

comparability, the process follows a rigid protocol overseen by the GEM Data Team (*GEM Global Entrepreneurship Monitor*, n.d.):

- **Sampling:** Surveys must cover all geographic regions (urban and rural) and include non-participants in the labour market.
- **Timing:** Fieldwork occurs simultaneously worldwide (typically April–June) to neutralise seasonal variations.
- **Quality Control:** To minimise non-response bias, a strict call-back protocol is enforced. Furthermore, rigorous translation and back-translation procedures ensure semantic consistency across languages.

The core metric of the GEM is the Total Entrepreneurial Activity (TEA). The classification of entrepreneurial status is a dynamic process that focuses on the phase combining the period prior to the launch of a new business (nascent entrepreneurship) and the phase immediately following its creation (managing a new business). Essentially, the TEA variable serves as a binary indicator of whether an individual is actively involved in the initial phase of the entrepreneurial process. Additionally, the survey also identifies managers of established firms, that is, businesses that have already surpassed the critical 42-month (3.5-year) threshold. This makes it possible to clearly distinguish between individuals at the early stages of developing their ideas and those who already manage mature, sustainable firms (*GEM Global Entrepreneurship Monitor*, n.d.).

6.1.1. Methodology of APS in Luxembourg 2024/2025

The survey employed a mixed-mode, multi-frame sampling approach combining Computer-Assisted Web Interviewing (CAWI) and Computer-Assisted Telephone Interviewing (CATI) (GEM Luxembourg, 2025):

- **CAWI:** The web-based portion utilised a non-probability opt-in panel ("mypanel.lu"), which contains over 17,000 registered members. The selection was stratified by age, gender, nationality, and region of residence.
- **CATI:** The telephone portion relied on a validated database containing over 500,000 numbers. This sample was drawn probabilistically using RDD and Random Last Digit (RLD) methodologies.

The distribution across modes was 60% CAWI (1,233 interviews), 25% CATI mobile (597 interviews), and 15% CATI landline (203 interviews). Although during fieldwork 1,233 responses were collected via CAWI, the team excluded 12 responses deemed inaccurate due to short completion times and nonsensical answers, leaving a final sample of 2,201. To maximise contact rates, the fieldwork protocol included up to five call attempts or reminders per case. It is important to note that no statistical imputation was applied to address item non-response (GEM Luxembourg, 2025).

The unweighted data revealed an overrepresentation of older respondents, particularly males, with imbalances varying by survey mode. To correct this, post-stratification weights were computed using auxiliary variables from official population statistics, including employment status, gender by country of birth, age by gender, and education level. Mode-specific weights were calculated first, then calibrated to the total population (GEM Luxembourg, 2025).

Table 1 critically highlights the disparity in efficiency between the collection modes, emphasising the methodological "cost" of RDD versus the online panel. Mobile mode has the highest response rate (RR3 of 3.9%), which is statistically explained by the large volume of cases with unknown eligibility (UH = 11,285) combined with a very high estimated eligibility factor (e = 84%). This severely penalises the denominator of the RR3 formula, as it is assumed that the vast majority of these non-contacts were valid people who simply did not answer. In contrast, the Landline mode, despite having an operationally inefficient ineligibility rate (6,516 invalid/commercial numbers, reflecting the degradation of fixed lists), achieves a higher RR3 (7.2%) in part because its e factor is low (22%), "protecting" the response rate by assuming that most unknowns are deadlines. On the other hand, the online mode served as the backbone of the study, ensuring the robustness of n (1,233 complete interviews, more than double the sum of the phones) with a participation rate (PR) of 25.4%. However, PR is not directly comparable to RR3 of probabilistic modes, as it does not deal with the uncertainty of eligibility (UH=0) typical of RDD (GEM Luxembourg, 2025).

Regarding ethical standards, participation in the survey was entirely voluntary, with all respondents explicitly informed of their rights and the study's objectives prior to interview. All procedures were conducted in full compliance with the General Data Protection Regulation (GDPR), and data storage adheres to the strict statistical confidentiality provisions mandated by Luxembourg's national statistical law (GEM Luxembourg, 2025).

Finally, regarding data accessibility, fully anonymised datasets are released to the public three years after collection. GEM member teams are granted earlier access, one year following the publication of the Global GEM Report, while access to the specific Luxembourg APS dataset is available at STATEC premises (GEM Luxembourg, 2025). For this study, STATEC provided confidential access to the 2024 APS survey and questionnaire.

The selection of this survey as the non-probabilistic sample for applying statistical integration methods is justified by its data-collection architecture and the methodological comparison it enables. Within the scope of this research, the first empirical stage will focus exclusively on the CAWI subsample, since the CATI component is based on a probability sampling design, whereas the CAWI component constitutes a non-probabilistic sample and is inherently subject to coverage and self-selection biases. By isolating the CAWI component, the study creates the ideal setting for testing the effectiveness of the proposed methods. Subsequently, the CATI component will be reintroduced when constructing the combined APS**.** A key aspect of this choice is that the GEM team already produces and applies standard calibration weights (post-stratification weights) to correct biases in the original sample. Thus, using this dataset not only mitigates selection bias through integration methods with EU-SILC but also establishes a rigorous empirical test: directly comparing the performance and robustness of these more ambitious methodologies and data combinations with the conventional weighting methodology already implemented and validated by the GEM team.

6.2. Reference probability sample: EU-SILC

The European Union Statistics on Income and Living Conditions (EU-SILC), launched in 2003, is a tool anchored in the European Statistical System (ESS) that aims to collect, in a timely and comparable manner, multidimensional microdata. This data focuses primarily on income distribution, poverty, and social exclusion, while also covering various European policies related to living conditions, such as child poverty, access to health care and other services, housing, over-indebtedness, and overall quality of life (Eurostat, 2024).

The structure of this survey provides two key types of data: cross-sectional data, which capture variables on income, poverty, and living conditions at a specific point in time or over a specific period; and longitudinal data, which allow for the observation of changes at the individual or household level over time, with individuals being tracked periodically over a

minimum period of four years (Eurostat, 2024).

The collection of this information is structured at two levels of observation. The core component of the survey, the detailed income, is collected at individual and household levels. Information on social exclusion and housing conditions, on the other hand, is obtained primarily at the household level, while variables on employment, education, and health are collected for individuals aged 16 and older. Thanks to this level of detail, EU-SILC has been the instrument of choice for providing data for structural indicators of social cohesion (such as the at-risk-of-poverty rate and the S80/S20 ratio) and for supporting the Open Methods of Coordination in the areas of social inclusion and pensions (Eurostat, 2024).

The choice of EU-SILC as the probabilistic reference is justified by its methodological rigour and multidimensional depth. As an official operation of the European Statistical System, EU-SILC ensures a probability sample design, consistent design weights, and population representativeness. Furthermore, the success of methods such as IPW and MI requires that the reference sample share a broad and rich set of covariates with the non-probability sample. Since EU-SILC collects comprehensive information on sociodemographic, labour, educational, and income characteristics (at both the individual and household levels), it provides the ideal auxiliary variable.

For this study, STATEC provided a subset of the EU-SILC survey of 2024, consisting of 5,000 respondents. This subset was drawn using simple random sampling, a procedure that aligns with STATEC's statistical protection policies while preserving the overall structure of the original sample. While the original sampling design features, the strata and clusters, were retained, the survey weights were readjusted. Because this subset represents only a fraction of the full sample, applying the original weights would no longer represent the total population of Luxembourg. The decision to use a subsample rather than the full dataset was a pragmatic choice made at the beginning of the practical implementation, primarily to reduce the computational burden. Since this work serves mainly as an initial exercise to test the feasibility of this methodological approach, the specific scale of the data was not the primary concern. Although choosing a complete sample or an alternative reference sample was a viable option to complement the research, time constraints led the focus to shift toward exploring and testing all the integration methods proposed in the package. Using the complete sample is a suggestion for expanding and deepening this research in the future.

7. DATA HARMONIZATION

The data harmonisation step serves as the fundamental foundation for any empirical study, ensuring that information extracted from different sources is consistent, comparable, and structurally coherent prior to modelling. This process involves standardising concepts, units of measurement and coding formats, thereby eliminating discrepancies inherent in the original sources. The importance of this stage lies in the fact that statistical algorithms assume strict uniformity in the input data. Without prior harmonisation, any noise or methodological inconsistency between the databases will inevitably propagate, introducing biases and compromising the validity and reliability of all subsequent conclusions.

In the specific context of integrating probabilistic and non-probabilistic samples, a sufficiently strict harmonisation of common variables is required for the estimators to rely on comparable covariate information across datasets. The proposed methods depend entirely on exact comparability and overlap between the datasets. As such, the common covariates must be harmonised into comparable categories (e.g., ensuring that binary variables share the exact same classification), and item nonresponse must be handled consistently before model estimation. If there are structural discrepancies or **unaddressed** missing data, the calculation of selection propensities or the estimation of predictive models becomes biased or computationally unfeasible. This process was carried out entirely in R, starting with data reading. Within the packages used at this stage, the most important were `readxl`, `dplyr`, `mice` and `ggplot2` (*A Grammar of Data Manipulation*, n.d.; *Create Elegant Data Visualisations Using the Grammar of Graphics*, n.d.; *R: The R Project for Statistical Computing*, n.d.; *Read Excel Files*, n.d.; Buuren & Groothuis-Oudshoorn, 2011). The filtering of the non-probabilistic side of the APS was intentionally performed only after the multiple imputation phase to maximise the available auxiliary information and increase the predictive power and stability of the imputation.

7.1. Candidate common variables

Rather than processing the entire datasets, we isolated the relevant data to reduce computational burden. Specifically, we retained the common variables, identifiers (IDs), and survey weights in both datasets. Additionally, for the APS, we retained potential variables of interest, such as TEA and Fear of Failure, as well as the interview mode to filter out non-

probabilistic respondents (CAWI). For the common variables, the strategy adopted looking into the 2024 questionnaire of the APS and looking into possible candidates for common variables that could also exist in the EU-SILC, that is, a larger dataset with more variables.

In this process, it was crucial to aggregate small categories within the variables. The aggregation was guided by ex-ante criteria: ensuring minimum category frequencies, mitigating the risk of sparse cells, and maintaining strict conceptual comparability between the APS and EU-SILC datasets. Because the IPW method relies on the inverse of these estimated propensity scores, near-zero denominators often caused by those sparse cells, inevitably generates extreme weights for a small subset of observations. These inflated weights disproportionately increase the variance of the estimators, leading to high instability. By aggregating these underrepresented categories into broader, conceptually meaningful groups based on these predefined criteria, it provides a more balanced distribution. This controls the expected impact on the final weights, thereby stabilising the weight estimation. In the APS data, responses like 'refused' or 'don't know' were grouped into a single 'NA' category and treated as item nonresponse. This step is essential to clearly separate genuine missing data that requires imputation from cases that are simply non-applicable.

It should be noted that, although the process of harmonising and aggregating variables is presented as the initial stage of the workflow preceding variable selection and modelling, its execution was not purely linear, but rather highly iterative. This step was revisited and refined on multiple occasions throughout the study, enabling a controlled assessment of the direct impact of different aggregation configurations on the final analytical results of the models.

To ensure consistency across both datasets, the following variables were matched and harmonized. The setup presented below corresponds to the final configuration retained:

- **Age:** "RB082" in the EU-SILC and "AGE7C" in the APS. This variable was changed in the EU-SILC to match the APS categorical format ('18-24', '25-34', …, '55-64'), including removing respondents who fell outside the age range 18-64. In this step, the EU-SILC size was reduced from 5000 to 3183 whereas the size of the APS remained the same, bearing in mind that the sample already consists solely of individuals in that age range.

- **Gender:** "RB090" in the EU-SILC and "GENDER" in the APS. They were in the

same categorical format in both datasets.

- **Household size**: "HB120" in the EU-SILC and "HHSIZE" in the APS. The APS had categorised from 1 to 9 household size and the EU-SILC had in numerical format. Due to categorical sparsity, the new categories created are '1', '2', and '3+'.

- **Country of birth:** "RB280" in EU-SILC and "LUBORNCT" in APS. This variable was in SCL GEO code in the EU-SILC and in categorical format in the APS (e.g. '0' born in Luxembourg, '4' in Portugal, etc.). The package `countrycode` allows converting country codes to their respective country names. Due to categorical sparsity, the newly created categories are 'Luxembourg', 'Portugal', 'Other countries in Europe' and 'Rest of the World'.

- **Parents' country of birth:** "PB230" and "PB240" in EU-SILC and "LUBORNFACT" and "LUBORNMOCT" in APS. Due to categorical sparsity, the newly created categories are 'Luxembourg' if both parents were born in Luxembourg and 'Abroad' if at least one of them was born in a different country.

- **Educational level:** "PE041" in EU-SILC and "LUREDUC" in the APS. This variable was in different formats across both datasets, with many specific categories, leading to categorical sparsity. The newly created categories are 'Low' (from primary to lower secondary levels), 'Medium' (Upper and post-secondary levels) and 'High' (Tertiary or equivalent and higher levels).

- **Activity status:** "RB211" in EU-SILC and "OCCUP" in APS. Due to categorical sparsity, the newly created categories are 'Employed', 'Unemployed', 'Retired' and 'Inactive'.

- **Employment status:** "PL040" in EU-SILC and "OCCUP" in APS. Due to categorical sparsity, the newly created categories are 'Self-employed', 'Employee' and 'Others'. Even though in the EU-SILC family workers are not considered employees, in this analysis, they were considered so that the sum of 'Self-employed' and 'Employees' equals the total of Employers in Activity status. It was crucial to distinguish between genuine missing values and structural gaps by recoding cases in which the absence of a response logically stemmed from the individual not being in

employment and treating as missing data only those situations in which the employment status was genuinely unknown or had not been reported.

- **Full or Part-time:** "PL145" in EU-SILC and "OCCUP" in APS. The categories created are "Full-time", "Part-time" and "Others". To distinguish between true and false NAs, the approach mentioned in the previous point was applied.

Since the last three mentioned variables derive from the same variable "OCCUP" from APS, it was included a mapping table that indicates how each variable was categorized (**Table 2**).

Tables 3 and 4 detail the distribution within the categories of each variable for both datasets. It is important to note that the weighting in each survey differs. SILC uses expansion weights because its objective, as an official probability sample, is to estimate totals and reflect the actual size of the population. In contrast, the APS uses rescaled weights (whose sum equals the sample size) to correct for internal representativeness biases. These weights reflect the current APS calibration weights and will be used for a future comparison with the results of the effect distributions generated by the IPW.

7.2. Missing values

After analysing **Tables 3** and **4**, we can detect a few missing values in some categories. As previously noted, the models won't work properly with their existence. When dealing with missing values, the most straightforward and intuitive strategy would be simply to exclude all incomplete observations. Although this approach is tempting due to its extreme simplicity and ease of implementation, its analytical cost is too high to ignore. The indiscriminate removal of rows not only reduces the sample size, thereby compromising the study's statistical power, but also risks introducing substantial bias into the results by erroneously assuming the omissions are Missing Completely At Random (MCAR). The loss of this information would distort the representativeness of the samples, making it imperative to adopt a more sophisticated and rigorous solution. To overcome these limitations and preserve the data's structural integrity that the decision was made to implement the Multiple Imputation by Chained Equations (MICE) algorithm. The citizenship variable was retained in the reference sample solely as an auxiliary variable for imputation; it was not considered a candidate for a common variable for sample integration. Its inclusion was justified by its

strong correlation with the countries of birth of individuals and their parents, providing the MICE algorithm with crucial structural information to detect patterns and improve the accuracy of missing-value estimation.

The MICE algorithm, also referred to in the literature as Fully Conditional Specification (FCS), is a multiple imputation technique considered the approach of choice for complex problems involving incomplete data. This tool is specifically designed to address the challenge of missing values across multiple variables. Rather than attempting to model all variables simultaneously through joint modelling (JM), MICE operates on a variable-by-variable basis. The method constructs an imputation model from a set of conditional densities, creating a separate model for each variable with missing data. Starting from an initial imputation, the algorithm generates final imputations iteratively based on these conditional densities, with a small number of iterations (usually 10-20) sufficient to stabilise the results. MICE stands out as a highly advantageous and attractive tool because it offers a robust alternative when no suitable global multivariate distribution can be found for the dataset (Buuren & Groothuis-Oudshoorn, 2011). The specific choice of the MICE algorithm over simpler imputation methods stems from the strict need to preserve the original statistical relationships between the variables. Basic methods, such as imputation by the mean or the mode, distort the variance and destroy the natural correlations within the sample. Furthermore, it was also due to its ease of implementation using the `mice` package in R.

To apply the MICE algorithm correctly, the missing data should be assumed to be MAR. Unlike the MCAR scenario, in which missingness is entirely stochastic and independent of the data, MAR posits that the probability of a value being missing can be systematically explained by the other observed variables in the dataset. In practice, it is impossible to prove beyond doubt that the data are MAR, given that the information required to validate this assumption is, by definition, missing, thereby preventing a direct empirical distinction between MAR and MNAR (Missing Not At Random). While it is possible to carry out statistical tests to formally reject the MCAR mechanism, this does not demonstrate that the data are MAR. Instead, the rejection of MCAR, together with observed associations between missingness and measured covariates, supports treating MAR as a working assumption, although MNAR mechanisms cannot be ruled out.

Little's test is a statistical procedure used to evaluate whether missing data follow an MCAR

mechanism. A significant test result ($p < .05$) leads to the rejection of the MCAR null hypothesis, suggesting that the data are instead MAR or MNAR (Li, 2013). **Table 5** present the results of this test in both datasets. The Little test yielded statistically significant results ($p < .001$) for both databases, leading to the rejection of the null hypothesis that omissions occur in a strictly random manner (MCAR). It is worth noting, however, that significance tests often reject the null hypothesis in large samples due to high statistical power. Consequently, while rejecting MCAR does not prove MAR or MNAR, this diagnostic establishes MAR as the most robust and reasonable working hypothesis moving forward.

To further investigate and complement Little's test, missing indicator variables were created for each of the incomplete variables in both datasets. These indicator variables assumed a value of 1 when the data was missing and 0 when it was observed. Using these binary indicators as dependent variables, a series of logistic regression models was fitted for each incomplete variable. To determine if a respondent's profile influenced their probability of non-response, fully observed sociodemographic characteristics were included as predictors. As summarized in **Table 6**, the models reveal that missingness is not entirely random but driven by several observed characteristics across both datasets. In the APS dataset, identifying as female is associated with a significantly higher probability of omitting information regarding perceived entrepreneurial skills ($\beta = 0.56$, $p = 0.030$), start-up intentions ($\beta = 0.55$, $p < 0.001$), and perceived opportunities ($\beta = 0.50$, $p < 0.001$). Additionally, respondents aged 25 to 34 were more likely to have missing data for perceived opportunities ($\beta = 0.68$, $p = 0.010$). Similarly, within the EU-SILC dataset, the models yield robust evidence of non-random missingness. For instance, respondents born outside of Luxembourg (e.g., Rest of the World) exhibit a significantly higher likelihood of having missing data for Educational Level ($\beta = 1.18$, $p < 0.001$) and Full/Part-Time status ($\beta = 1.01$, $p = 0.012$). Conversely, being an active employee systematically decreases the log-odds of omissions across multiple variables, such as Parents' Country of Birth ($\beta = -1.32$, $p = 0.024$), Educational Level ($\beta = -0.55$, $p = 0.033$), and Full/Part-Time status ($\beta = -1.29$, $p < 0.001$). Larger household sizes also had a significant negative effect, reducing the probability of missing values for variables such as Parents' Country of Birth and Citizenship.

Following this preliminary analysis, the MICE algorithm was implemented. The predictor matrix was carefully configured: the diagonal was set to zero to prevent variables from predicting themselves, while fully observed sociodemographic covariates were retained as

predictors to maximise the models' predictive power. The selection of estimation methods strictly adhered to the structural nature of each variable across the datasets. Proportional odds models ('polr') were applied to ordinal variables, specifically Educational Level, to preserve their natural hierarchy. Binary logistic regressions ('logreg') for binary covariates, including Gender, Parents' Country of Birth, Fear, Opportunity, Start, and Skill. For nominal variables with multiple categories namely Full/Part-Time, Activity Status, Employment Status, Household Size, Country of Birth, and Citizenship multinomial logistic regressions ('polyreg'). To support statistical stability and minimise Monte Carlo variance in the integrated results, the algorithm was configured to generate 30 imputed datasets ('m = 30') over 15 iterations ('maxit = 15'). A fixed random seed (seed = 20260212) was used for the exact reproducibility of the procedure. Finally, post-imputation convergence diagnostics were conducted. It is important to note that while these diagnostics strongly support the plausibility of the imputation models and the stability of the iterations, they do not formally guarantee absolute validity.

The configuration of MICE to generate 30 imputed sets served purely diagnostic and validation purposes, to map the uncertainty of the estimates and to generate the variability required for the construction of visual convergence diagnostics (trace plots). These diagnostics suggest that the imputation process was stable, although they do not formally validate the missing-data mechanism. The analysis of the trace plots in **Figures 1 and 2** indicate good convergence and mixing of chains for all variables. Note that, because they are categorical variables, these "means" and "sd" serve only to assess stability, not for substantive interpretation. Overall, the diagnostics suggest that the imputations are stable and plausible, with no signs of non-convergence. The strip plots (observed vs. imputed) in **Figures 3 and 4** indicate good overlap between imputed and observed values, with no modal collapse or "empty" categories. In summary, the graphs show no evidence of non-convergence problems, severe bias, or undue restriction of imputations to a few categories.

To obtain a complete dataset compatible with the subsequent analytical pipeline, one of the 30 datasets generated by MICE was randomly selected. Due to the high computational complexity and specific technical constraints of the downstream statistical integration methods, which involve complex propensity score estimation and calibration steps, applying Rubin's rules to pool the results across all 30 imputed datasets proved unfeasible within the current framework. While the primary focus of this research is the evaluation of data

integration techniques rather than the multiple imputation process itself, this decision is acknowledged as a methodological limitation. Relying on a single imputed dataset prevents the full propagation of imputation uncertainty and may lead to underestimated standard errors in the final estimates. Consequently, adapting this integration pipeline to incorporate full multiple imputation pooling remains a valuable direction for future research.

Figures 5 and **6** present a visual comparison of the distributions of each category before and after imputation in each dataset. The graphical assessment suggests that the application of the MICE was highly effective in preserving the original structure of the data. For example, the proportion of individuals with a high educational level in the EU-SILC sample changed from 48.1% in the original data to 48.8% after imputation. There is a greater variation in the distributions of the variables 'EasyStart', 'Opportunity', 'FearofFailure' and 'Skill' – opinion questions regarding entrepreneurial activity – as these have a higher number of missing values compared with the other variables. However, based solely on graphical comparison, no substantial visual differences were observed in the overall distributions.

After this process, a final overview of both datasets can be made. **Table 7** presents only the analysis of the distributions for the variables that contained missing values in the EU-SILC ('EducationalLevel', 'ParentsCountryBirth' and 'FullPartTime') which changed following imputation, as the distributions of the remaining variables remained the same as in **Table 3.** In the case of the APS (**Table 8**), the analysis was carried out again for all variables, not just those that had been imputed, because following imputation, it was possible to filter out the non-probabilistic side. Therefore, **Table 8** differs from **Table 4** not only because it was compiled after imputing missing values, but also because it was based solely on the non-probabilistic side, not on the entire sample. Although the citizenship variable in the EU-SILC was included and imputed in the MICE procedure due to its high predictive value as an auxiliary variable, it was omitted from this distributional analysis because it is not a common variable eligible for final integration of the samples.

8. VARIABLE SELECTION

The transition from the data harmonisation phase to effective integration using IPW, MI, and DR estimators requires a systematic variable selection process. Rather than indiscriminately including all common variables, the available covariates were evaluated to support a rigorous model-building decision. Initially, the full set of harmonised variables was assessed for multicollinearity. Subsequently, exploratory stacking models and outcome models were applied to evaluate the predictive power of each covariate and determine its inclusion in the final integration steps. For the propensity score models (IPW), selection focused on variables demonstrating a substantive association with the selection mechanism, specifically, predictors that affect the probability of an individual belonging to the non-probability sample. Conversely, for the outcome imputation (MI) models, the criteria prioritized covariates that showed a strong predictive association with the main variable of interest, the binary indicator of TEA. The DR estimator benefits greatly from this dual specification, as it simultaneously employs both the optimally selected propensity and outcome variables. This concurrent integration ensures consistent estimates if at least one of the two underlying models is correctly specified (Wu, 2022).

Furthermore, model specifications should not be limited to main effects alone; it is crucial to assess the inclusion of interaction terms. The inclusion of these interactions allows us to capture complex patterns and heterogeneities in the probability of selection or in the propensity for entrepreneurship that would otherwise remain hidden. However, this modelling is subject to a strict bias-variance trade-off. Over-specification of the model, through the excessive addition of covariates and complex interactions, drastically increases the risk of overfitting and numerical instability. Given the binary nature of both the sample belonging and the TEA variable, excessively complex models are highly susceptible to mathematical perfect separation, which prevents algorithms from converging. Thus, variable selection must be parsimonious, balancing the predictive power needed to correct biases with the computational stability required by the estimators (Elliott & Valliant, 2017).

As a preliminary step, multicollinearity among the candidate variables was identified and addressed because the simultaneous inclusion of highly correlated variables (predictors that provide redundant statistical information by explaining the same dimension of the data) compromises the analytical stability of the models (Daoud, 2017). The diagnostic test for

multicollinearity revealed a strong linear dependence among the variables 'ActivityStatus', 'EmploymentStatus' and 'FullPartTime' in the APS. This was formally evidenced by a QR decomposition of the model matrix, which detected exact linear dependence (aliasing) and rank deficiency. This structural redundancy arised from the data harmonisation process itself: whilst in the EU-SILC these three indicators were constructed from distinct source variables, in the APS they were derived from a single variable. Although this original variable was subject to different recategorisations to emulate each of the three employment concepts, the underlying explanatory variance remained shared, resulting in perfect multicollinearity.

To solve this overdetermination and prevent the algorithms from collapsing, 'ActivityStatus' and 'FullPartTime' were excluded, retaining only 'EmploymentStatus'. However, this methodological decision carries a substantive analytical cost. By eliminating these two variables, the integrated models forfeit the ability to capture work intensity (differentiating between full-time and part-time) and lose the distinction of non-working individuals, as the unemployed, retired, and inactive populations are structurally absorbed into broader, less detailed categories. Despite this loss, the decision was based on the criterion that an individual's employment status has the greatest explanatory power for the TEA. This choice is supported by the GEM Luxembourg (2025) report. According to the document, official data regarding self-employment are specifically used to 'measure individuals labour market status' and provide 'complementary perspectives on entrepreneurship'. Furthermore, the report notes that 'self-employed workers represent only a subset of all entrepreneurs', establishing a direct theoretical link between employment categories and entrepreneurial activity. Finally, the graphical analysis of **Figure 7** shows that despite methodological differences, these official employment indicators share a 'long-term upward trend' with GEM measures and they 'capture the same broad dynamics as official statistics'. Therefore, retaining the 'EmploymentStatus' variable ensures that the models capture these validated labour market dynamics while avoiding the mathematical collapse caused by multicollinearity.

8.1. Stack (selection) model

To determine the optimal variables and interactions influencing inclusion in the non-probability sample, a stacked dataset was constructed by pooling the APS and the EU-SILC,

with a binary indicator to distinguish the origin of each observation. To identify the most relevant predictors before applying the IPW estimator, a base exploratory logistic regression was specified, incorporating the main effects of the remaining harmonized covariates.

To capture complex, non-linear selection mechanisms while avoiding overfitting and data dredging, a set of interactions was evaluated, including interactions between demographic characteristics (such as age and gender) and labour market variables (employment status), driven by the hypothesis that socio-economic profiles differentially impact survey response propensities. Model comparison was based on Akaike Information Criterion (AIC), deviance (likelihood-ratio tests), and weight diagnostics. Among those tested, empirical results demonstrated that the interaction between 'EmploymentStatus' and 'EducationalLevel' yielded the optimal balance of fit and stability. This result is consistent with Human Capital Theory within the entrepreneurship literature. This interaction captures the intersectionality of work arrangements and educational attainment, acknowledging that the profile, motivations, and survey responsiveness of a highly educated self-employed professional often differ from those of a self-employed individual with lower formal education, reflecting the distinction between opportunity-driven and necessity-driven entrepreneurship (Marvel *et al.*, 2016). **Table 9** presents a comparative analysis of the statistical metrics between the model with the interaction term (Model 1) and the base parsimonious model (Model 2):

- **Residual Deviance:** This metric quantifies the model's lack of fit with the observed data with lower values indicate a better fit (Hosmer *et al.*, 2013). The transition from the Base Model (Model 2) to the model with interaction (Model 1) resulted in a reduction in Residual Deviance from 4971.9 to 4963.0, suggesting that capturing the combined effect of education and employment status reduces the prediction error arising from the selection mechanism.
- **AIC:** Assesses the relative quality of statistical models through a trade-off between goodness of fit and complexity. As the AIC calculation penalises the addition of new parameters, a reduction in its value indicates that the gain in information outweighs the loss of degrees of freedom (Wagenmakers & Farrell, 2004). Given the slight decrease in the AIC from 5003.944 to 5003.016, Model 1 holds a better result.
- **Deviance Test ($Pr > \chi 2$):** Likelihood Ratio Test evaluates whether the reduction in error between the two nested models is statistically significant (Hosmer *et al.*, 2013). The inclusion of the interaction consumed 4 degrees of freedom, resulting in

a p-value of 0.06293. Although this value is not significant at the 5% level, it is marginally significant at a 10% confidence level.

The statistical evidence supporting the inclusion of the interaction term is weak, providing only a marginal improvement in fit at the 10% significance level. As indicated by an AIC reduction of less than 1 and a deviance test p-value of .063, the interaction model does not demonstrate clear superiority. Therefore, both specifications were retained for sensitivity comparison in the subsequent implementation of the IPW procedure.

8.2. Outcome model

As the final stage of the variable selection, the predictive models for the outcome were specified for the subsequent application of the MI. To ensure consistency, the approach previously adopted in the stack model was replicated. Initially, a base model was fitted including only the main effects of the retained common variables. Next, an expanded model was evaluated with the interaction term between Educational Level and Employment Status.

To assess the empirical impact of including the interaction term in the predictive model for the outcome, a direct comparison was made between the base specification and the interactive specification (**Table 10**).

- The transition from the Model 2 to Model 1 resulted in a slight decrease in Residual Deviance, which fell slightly from 613.64 to 610.01.
- However, the AIC worsened, rising from 645.638 in the base model (Model 2) to 650.005 in the interaction model (Model 1).
- The Deviance Analysis corroborates this lack of fit: the reduction in error at the cost of 4 additional degrees of freedom resulted in a *p*-value of 0.458, a result that is markedly non-significant.

The empirical evidence clearly demonstrates that the additional complexity is not justified in this predictive scenario. The sharp increase in the AIC indicates that the negligible improvement in data fit does not offset the penalty imposed by adding the new parameters. Given the complete lack of statistical significance, retaining the interaction would result in overfitting. Consequently, for the implementation phase of the MI, only the Base Model was studied.

9. APPLICATION OF THE METHODS

9.1. R `nonprobsvy` package

After the selection and preparation of variables, the central and most critical part of this study follows: the practical application of data integration models. In this phase, the aim is to apply and test IPW, MI and DR and their respective algorithmic variants. Until recently, the manual coding and implementation of all these methodological combinations would have been a computationally intensive, fragmented and difficult-to-replicate process. Fortunately, to automate and facilitate this process, the statistical package `nonprobsvy` has been and continues to be actively developed by the researchers Łukasz Chrostowski, Maciej Beręsewicz and Piotr Chlebicki, contributing for the realisation of this work. Without going into detail about its internal mechanisms just yet, its main methodological achievement lies in unifying these complex statistical methods into a single, agile tool. This allows the researcher to focus purely on methodological and analytical decisions.

According to Chrostowski *et al.* (2025), the `nonprobsvy` package in R implements the previous approaches like analytical and bootstrap variance estimation, and can be integrated with the `survey` or `sampling` packages. All analyses were reproduced using the package versions reported in the rendered R outputs (specifically, `nonprobsvy` version 0.2.3 and `survey` version 4.5). The package is built around the `nonprob()` function, which features several arguments that dictate not only the choice of the statistical methods but also their specific operational mechanics. Among these, the parameters used in the final implementations are the following:

- **`data:`** for the non-probability sample. Must include all variables in selection, outcome and target.
- **`selection:`** a formula for the selection equation that contains the auxiliary variables and defines the propensity score model for the IPW estimator.
- **`outcome:`** a formula for the outcome equation that contains the target variables under the same prediction model for the MI and DR estimator.
- **`target:`** a formula with the target variables.
- **`svydesign:`** for the probability sample that is used for calibration and validation of the non-probability estimates.

- **`method_selection:`** function for the propensity scores (`c("logit", "probit", "cloglog")`).
- **`method_outcome`**: method for the MI (`c("glm", "nn", "pmm", "npar")`).
- **`control_out:`** function used to specify the estimation process including variable selection methods.
- **`control_sel:`** function that provides control parameters for the selection fitting.

Using the `svydesign()` function in the `survey` package before this function was a mandatory step because it declares and formalises the sampling design of the reference probability sample. It tells R exactly how that sample was collected, incorporating the sampling units (`ids`), the sample weights (`weights`) that expand the sample to the total population size, and the stratification variables (`strata`). Without this step, the model would treat your reference data as a simple random sample, which would undermine calibration or propensity score calculations, leading to biased estimates.

In practice, formalising the sampling structure for the EU-SILC sample required defining crucial parameters. Firstly, the sampling units were identified by combining the household identifier and the personal identifier (representing the hierarchical level of data collection). Secondly, and of the utmost importance, the sample expansion weights were specified but with a critical methodological nuance. Instead of using the original cross-sectional weights, a new developed calibration weight variable was applied. To ensure valid inferences, it was imperative to recalibrate the weights for this specific subsample because they no longer reflect the actual size of the population. This process was based on Luxembourg's official demographic figures using data on the resident population by age and sex as at 1 January 2024 (*LUSTAT Data Explorer • Population by Age and Sex on 1st January*, n.d.). By adjusting the sub-sample weights to match these known and updated population totals, the data's structural representativeness was restored. The subsample was recalibrated after restricting the reference sample to adults aged 18-64 and implemented after data preparation and imputation, so that the reference design object reflected the target analytical population.

9.2. Inverse Probability Weighting

The first data integration approach applied was the IPW estimator. To this end, two distinct predictive specifications were evaluated: Model 1, which includes the interaction term

between Educational Level and Employment Status, and Model 2, which is limited exclusively to the main effects of the common variables. In both specifications, the estimation of propensity scores was based on a logistic ('logit') regression model. The logit specification exhibited the greatest convergence stability and the best fit to the data compared with alternatives such as the 'probit' or 'cloglog' models.

A crucial methodological nuance in the specification of the `nonprob()` function was the inclusion of a SCAD (Smoothly Clipped Absolute Deviation) regularisation term. Although the prior variable selection process established the core structure of the models, the application of SCAD acts as an essential algorithmic safeguard. This penalty mechanism suppresses residual noise and reduces the impact of marginal parameters that were not captured in the previous selection (Fan & Li, 2001). Finally, the estimation method used was GEE, as discussed theoretically in Section 5.1., which uses the inverse of the calculated probabilities to expand and reweight the non-probability sample, forcing its structure to align and match with the reference sample.

Table 11 show the performance and impact of the two specifications. Model 1 projects a TEA of 10%, with a standard error of 1.4%. Model 2, on the other hand, estimates a slightly higher proportion of 10.8%, with a standard error of 1.6%. This divergence of demonstrates how the structure of the propensity score model influences the reweighting of the sample and, consequently, the final result. In terms of the mathematical distribution, Model 1 generated a higher maximum weight (2497.5) than Model 2 (2275.6), whilst also exhibiting greater weight variance (64714.6 versus 64366.3). The impact of higher weight variance is one of the most critical issues in IPW. Greater dispersion means that there are extreme disparities in the weight assigned to each individual. When the model estimates probabilities very close to zero for certain profiles, the inverse of the probabilities generates astronomical weights. Although this constraint combats selection bias, it destabilises the estimation. High variance makes the final result dependent on specific observations, sacrificing the model's overall robustness and increasing the margin of error. Weight trimming was not applied to bound extreme values in the current estimation. Therefore, implementing such techniques remains a valuable recommendation for future sensitivity analyses.

To formally quantify the loss of statistical efficiency, **Table 11** also presents the design effect (deff) and design factor (deft) metrics. Deff expresses the ratio between the variance of the

estimate obtained using IPW pseudo-weights and the variance that would be expected if a SRS of the same size had been drawn. In practice, deff tells us the extent to which our decision to apply weights has inflated the uncertainty of the estimate. The deft is the square root of deff and converts this penalty to the standard error scale (Salganik, 2006). The deff and deft were calculated using the mathematical relationship between the sample size and the Effective Sample Size (ESS). Analysing the values, both models exhibit identical penalties. Model 1 has a deff of 1.4886 (and a deft of 1.2201), whilst Model 2 has a deff of 1.4860 (and a deft of 1.2190). This means that the complexity increased the variance of the estimates by almost 49% (and the standard errors by approximately 22%), when compared to a random sampling design. In short, the inclusion of the interaction term in Model 1 did not improve the efficiency of the estimation (identical deff and deft values), and the greater weighting variance justifies the preference for the more parsimonious model.

In **Figure 8** there is an analysis of the boxplots of the IPW weights of each model. The box plot is excellent for visualising where most of the data is concentrated (the central 'box', which contains 50% of the observations) and for identifying outliers (red dots). The Log10 scale on the x-axis is used to visually compress the distances, since the weights range from low values to the thousands. The black vertical line inside the box (the median) is roughly in the same region for both models (slightly below 300). This means that, for the 'typical' individual in the sample, both models assign a similar weight. From the long tail on the right and the red dots, both models generate extreme weights. However, as seen in the previous table, Model 1 has an outlier that extends even further to the right than that of Model 2. In IPW, these extreme observations are problematic because a single individual can exert disproportionate influence on the final TEA estimate, thereby inflating the variance.

Figure 9 shows the density plots of the IPW weights for each model. This representation acts as a smoothed histogram, showing the 'mass' of the data: where the curves are highest, there is the greatest concentration of individuals assigned that weight value. Model 2 (Yellow Area) exhibits a much smoother distribution, concentrated in a single large peak (unimodal) around 200–300. It has a long tail on the right, but the decline is more gradual and stable. On the other hand, Model 1 (Blue Area) reveals the most telling aspect of this analysis. Model 1 displays a bimodal distribution (two well-defined 'humps' or peaks: one just past 100 and another near 400). This visual separation is the direct result of the interaction (EducationalLevel: EmploymentStatus). By forcing the model to examine the cross-

categorised subcategories, the algorithm identifies a very specific socio-demographic group within the non-probability sample that needs to be heavily reweighted (the second peak near 400), isolating it from the rest of the sample (the lower peak).

These plots provide further evidence that the interaction specification (Model 1) produced less stable weight diagnostics. Although the model with interaction attempts to provide a finer mapping (creating a bimodal separation in the density to account for specific subgroups), this mapping results in a fragmentation of the data. The direct consequence is the emergence of more extreme maximum weights (visible at the tail of the boxplot) and increased variability in the overall distribution. Since the aim of IPW is to stabilise the sample in order to reduce bias while preserving precision (addressing the design effect constraints, deff and deft), the more parsimonious base specification (Model 2) offers a more balanced and methodologically robust choice.

9.3. Mass Imputation

The second data integration approach was Mass Imputation (MI). As concluded during the variable selection phase for the outcome, the interaction term did not add any value. Consequently, the MI modelling was based exclusively on the main effects model. To ensure robust analysis, two distinct imputation strategies were tested: Model 1, a Generalised Linear Model (GLM) with a logit link function for direct probability imputation; and Model 2, the Predictive Mean Matching (PMM) algorithm.

In the variable selection integrated into the `nonprob()` function, a divergence from the IPW was observed. While propensity score estimation benefited from SCAD regularisation, the outcome imputation models performed better under LASSO. In the IPW, SCAD was used to regularise the selection model and reduce instability in the estimated weights (Fan & Li, 2001). In contrast, the MI, particularly in its PMM variant, prioritises minimising prediction error and maintaining the stability of overall variance (section 5.1.). LASSO, by shrinking the coefficients and introducing a tolerable bias, sacrifices a small amount of bias in exchange for exceptional stability in prediction (variance reduction) (Tibshirani, 1996).

The results in **Table 12** indicate that the PMM approach produced a higher estimate and a lower standard error compared to the GLM model in this exercise. There is a noticeable

difference in the estimated proportion of the population engaged in TEA. The GLM model estimates that 10.5% of the population is engaged in TEA, while the PMM method projects a higher proportion at 11.5%. Theoretically, PMM tends to better preserve the natural variability and distribution of the original data by imputing the exact values observed in the non-probability sample to the most similar individual in the reference sample. Regarding precision, the GLM model yielded a standard error of 1.2%, whereas the PMM method resulted in a substantially lower standard error of 0.5%. However, since the true population parameter for TEA is unknown, these metrics must be interpreted with caution. The plausibility of such a low standard error can suggest potential underestimation of variance. If the variance estimation process does not fully incorporate the uncertainty inherent in the imputation mechanism, the standard error can become artificially deflated, leading to overconfidence in the precision of the PMM estimate. The lower standard error should be interpreted as model-based precision within this implementation, not as definitive evidence that PMM is more reliable. Finally, it is crucial to make a methodological note: unlike IPW, this analysis does not present diagnostic metrics for weights. In MI, we do not reweight the non-probability sample using pseudo-weights generated by the model. Instead, the model is used to impute the values of the variable of interest directly into the reference probability sample. Consequently, the weighting structure remains unchanged, using exclusively the design weights of the reference sample for population expansion.

Figure 10 illustrates the distribution of predicted probabilities estimated by the base model underlying MI. A visual analysis confirms the specification's consistency, as evidenced by the strong positive skewness and the concentration of the probability mass in the left tail. This behaviour accurately reflects the empirical structure of the data, confirming that the model correctly captured TEA as a rare and minority phenomenon in the general population.

Figure 11 shows the Receiver Operating Characteristic (ROC) curve, a tool that evaluates the discriminatory power of a binary classification model (in this case, distinguishing between those who are in TEA and those who are not). The ROC curve plots sensitivity against 1-specificity across classification thresholds. Sensitivity on the Y-axis is the true positive rate, interpreted as the model's ability to correctly identify true entrepreneurs, while 1-specificity on the X-axis is the false positive rate, representing the proportion of non-entrepreneurs incorrectly classified as entrepreneurs. The AUC (Area Under the Curve) summarizes this curve as a number between 0 and 1. A model without discriminatory capacity has an AUC

of 0.5 while a perfect model has an AUC of 1 (Hoo *et al.*, 2017). An AUC of 0.724 is considered to represent acceptable to good discriminatory power. This means that if an individual in TEA and an individual not in TEA are chosen at random from the sample, there is a 72.4% probability that the model ranked a higher probability score to the true entrepreneur. This indicates that the model had acceptable discriminatory ability, rather than behaving like a random classifier. The curves and the AUC of both models are identical because they evaluate the predicted continuous probability rather than the final imputation.

9.4. Doubly Robust

As a final step, to test the DR model, the propensity model without interactions was combined with the outcome model using GLM. In this case, as of now, there is no feature in the `nonprobsvy` package that allows us to include the PMM model within the DR outcome. Therefore, although it was the model with the lowest standard error in the MI analysis, the only option for testing a DR model was to use the GLM.

Table 13 presents the results of the DR estimator, which estimates a population-level TEA of 10.9% with a standard error of 1.5%. It is important to note that all weighting metrics, maximum weight, variance of weights, deff, and deft, are strictly identical to those of IPW Model 2. This result is a natural consequence of the model's design: since the DR's selection component used exactly the same base specification and the same penalty (SCAD) to estimate the propensities, the algorithm replicated the IPW weights. This finding is also supported by the DR density plot (**Figure 12**), which has the same shape as the IPW 2 model.

Figure 13 shows the ROC curve for the DR estimator, which achieved an AUC of 0.727. Although DR uses the same predictive model as the MI, the resulting AUC is virtually identical (0.727 versus 0.724). This minimal difference of 0.003 is negligible and does not represent a substantive or statistically significant increase in discriminatory power.

9.5. Comparison between the models

To provide a final comparison of all models, **Table 14** presents the metrics side by side. In addition, results for the APS without weighting and for the APS with the current calibration have been included. The unweighted APS produces a lower TEA estimate (8.3%) than all

adjusted estimators**,** illustrating the impact of ignoring the sampling design, even though it yields a low standard error (0.8%). On the other hand, the current APS weighting adjusts this estimate to 10.6% but at the cost of a sharp increase in variance, as evidenced by a standard error (SE) of 1.4% and a high deff of 1.9419. This deff value serves as a benchmark: any new weighting model should ideally fall below this threshold to be considered more efficient than the current approach.

The IPW 1 and IPW 2 models demonstrate a clear improvement compared to the current APS weighting method. IPW Model 2, although it has an SE of 1.6%, achieves a deff of 1.4860, which is significantly lower than the current deff of 1.9419. This means the variance penalty from the weights is much less severe. More importantly, one of the greatest advantages of the IPW is its universality (it is not variable-specific). The propensity score model focuses exclusively on the mechanism by which individuals are selected for the sample, ignoring the outcome variable. If, in future studies, the objective is to analyse other variables (such as Fear of Failure or Perception of Opportunities), one needs only use the same column of weights. This feature makes IPW extremely lightweight and operationally efficient for multivariate analyses.

Despite their potential appeal as outcome-specific estimators, MI and DR come with a high cost: they are inherently variable-specific. Unlike IPW, the predictive model built for MI requires that covariates be specifically adjusted to maximise the prediction of that exact outcome (in this case, TEA). If the researcher wishes to analyse a second variable from the GEM, they must start the process from scratch: train a new model, select new variables, validate new assumptions, and perform a new imputation. In large-scale surveys with dozens of metrics, this variable-specific nature makes these approaches exhaustive and computationally expensive.

The DR estimator serves as the final synthesis, merging the two approaches. Looking at the table, it is clear that it inherits the exact weighting structure of IPW Model 2 (sharing the maximum weight of 2275.55 and the deff of 1.4860), stabilizing the point estimate at 10.8%. The inclusion of the predictive model, however, helps "stabilize" the variance, reducing the standard error from 1.6% (in the pure IPW model) to 1.5%. However, this efficiency gain, like the MI, is limited to being a variable-specific model.

In conclusion, the comparative evaluation of these models must be disaggregated across distinct methodological criteria. Regarding proximity to the current GEM weighting estimate (TEA 10.6%), both the DR estimator and IPW Model 2 (10.9%) align closely with the established baseline, as does MI with GLM (10.5%). When evaluating variance and standard error (SE), MI with PMM yields the lowest apparent SE (0.5%); however, this metric should be interpreted with caution, as it may underestimate imputation uncertainty. Among the weighted approaches, the DR estimator (1.5%) offers a slight stabilisation in variance compared to the pure IPW Model 2 (1.6%). In terms of deff and weight stability, the IPW and DR models demonstrate clear superiority over the current calibration. Both IPW Model 2 and the DR estimator achieve a lower maximum weight (2275.6) and a deff of 1.4860, significantly outperforming the current APS baseline (deff 1.9419) and proving highly efficient at minimising the variance penalty.

Finally, regarding scalability and dependence on specific models, a critical methodological trade-off emerges. MI and DR offer robust point estimates but are inherently variable-specific, requiring computationally expensive, bespoke predictive models for each individual outcome. Conversely, if the objective is to produce a scalable, general-purpose dataset for cross-sectional analysis of multiple entrepreneurial dimensions, IPW stands out as the most operationally viable approach, since its weighting structure relies solely on the selection mechanism and is independent of the outcome variables.

10. ANALYSIS OF THE WHOLE GEM's APS

This stage of the research marks the transition from an isolated assessment of the non-probabilistic sample to a holistic analysis of the entire APS data. The central objective is now to reconstruct a single dataset, reuniting the non-probabilistic and probabilistic components. The goal is to directly compare this new global dataset with the calibration methodology currently used by the GEM team. The decision to proceed with this integration using the winning IPW model is based on the weighting's cross-sectional universality. Reconstructing the global APS dataset using MI or DR would require estimating, fine-tuning, and validating a separate predictive model for each survey variable. In addition to being analytically impractical and computationally intensive, the individualised and massive imputation of each metric could distort the empirical association structure among survey variables. Any subsequent analytical cross-examination of variables would result in an artificial relationship.

The reconstruction of the APS dataset cannot be achieved by simply concatenating this subsample with the original probability-based subsample due to an incompatibility in the scales of the weights. In the initial APS, the probabilistic and non-probabilistic samples were calibrated together to reflect the correct demographic proportions in the sample. Applying IPW to the non-probability subsample disrupted this scale, as it generated new weights cross-referencing data with the external reference sample, with weights expanded to the population scale. The weights on both sides are no longer commensurable, which destroys the dataset's proportionality and produces biased estimates. For this reason, reconstructing the full dataset requires harmonising the scales.

10.1. Weight Sharing Method

The Multiple-Base Composite Estimation approach was adopted. As demonstrated by Favre-Martinoz & Gros (2017), this approach is consistent with the principles of the Weight Sharing Method, allowing the information from both sources to be combined through a linear combination governed by the parameter θ:

$$\hat{Y}_{combined} = \sum_{j \in S_{NP}} \theta \frac{1}{\hat{\pi}_j^A} y_j + \sum_{k \epsilon S_B} (1-\theta) w_k y_k$$

Where

- $\boldsymbol{S_A}$**:** Non-probabilistic sample;
- $\boldsymbol{S_B}$: Probabilistic sample;
- $\boldsymbol{\theta}$: Composition factor (weight given to the non-probabilistic sample);
- $\boldsymbol{\hat{\pi}_j^A}$**:** Estimated propensity score;
- $\boldsymbol{w_k}$**:** Current calibrated weights;
- $\boldsymbol{y_j}$**,** $\boldsymbol{y_k}$: Outcome variable observed value

The structure of the equation reflects a linear combination of two population estimates. The first part of the formula contains the non-probabilistic component, where the observed variable (y_j) for each individual is rescaled by the weight generated via IPW ($\frac{1}{\hat{\pi}_j^A}$), expanding the sample to reflect the target population, with its total sum modulated by the composition factor θ. The second part contains the probabilistic component, with original calibration weights from the sampling design (w_k) applied to the respective responses (y_k). The resulting estimate is weighted by the complementary proportion of the composition factor (1- θ).

The choice of the composition factor (θ) constitutes the main strategic decision in this estimator, requiring a trade-off between variance and bias. Adopting an θ proportional to the sample size, although optimised for minimising the variance, would prove counterproductive in this context: it would assign excessive weight to the non-probabilistic component, thereby marginalising the reference sample and transferring greater degree of the residual selection bias into the final estimate. For this reason, the value chosen for θ was 0.5. Setting the parameter to 0.5 contributes to a balanced management of this trade-off. While it does not entirely neutralise selection bias, this equal weighting mitigates its impact, ensuring that the composite estimator structurally anchors its properties to the design-unbiased nature of the probabilistic component (Favre-Martinoz & Gros, 2017).

10.2. Analysis of indicators

To assess the alignment of the new global APS dataset, this segment provides a direct comparison with official statistics provided by STATEC as benchmarks. It should be noted that this comparison serves as a diagnostic of alignment rather than a full external diagnostic

of alignment. This analysis focuses on three indicators—Country of Birth, Employment Status, and Household Size—and compares three scenarios: the APS without weights, the APS with the current GEM calibration, and the new dataset integrated with the IPW model. Two metrics were used to measure which of these approaches most closely approximates to STATEC's data: the Sum of Absolute Errors (SAE) and the Mean Absolute Error (MAE). The SAE calculates the sum of all differences between the proportions of the survey and the STATEC values for all categories, measuring the total cumulative deviation for that profile. The MAE divides this total by the number of categories in the indicator, providing the average deviation per category. These metrics identify which weighting method best corrects the demographic structure of the APS. Although the current GEM calibration is based on some of these indicators, as mentioned in the APS methodology, this comparison was made using benchmarks from different datasets. For example, for Country of Birth, the GEM team used 2023 data to calibrate the 2024 APS because, as of that date, there was no complete population data available. In this comparison, datasets more closely aligned with the year of the study were used. Similarly, the new IPW calibration was performed independently of the benchmark datasets, ensuring that this assessment provides a distinct and external evaluation of demographic alignment.

A joint analysis of the Country of Birth indicator (**Figure 14**) illustrates the impact of weighting strategies on correcting sampling bias. The GEM Unweighted reveal a profound mismatch with STATEC's official population records. There is an underrepresentation of younger age groups in all regions and overrepresentation of older individuals (particularly evident in the peak of the 55–64 age group born in Luxembourg). This bias is quantified by an SAE of 41.24 and an MAE of 3.57. The application of weighting methodologies mitigates most of these distortions. Visually, both the current calibration method (Current Weighting Scheme) and the proposed new approach (Shared Weights) bring the estimated proportions closer to the true distribution of the official population (dark grey bars). However, evaluation using error metrics show the lowest overall discrepancy, reducing the SAE to 2.31 and the MAE to 0.82 with the new methodology, surpassing the calibration currently in use, which has an SAE of 2.61 and an MAE of 0.90. These results demonstrate that the IPW not only improves alignment on the evaluated benchmark dimensions but also manages to map the cross-demographic matrix of the population with a lower degree of overall distortion than the traditional system.

Figure 15 shows the distribution of household size, revealing a distinct statistical pattern. Unlike observed in the previous indicator, the unweighted sample exhibits the highest precision, with the lowest SAE=9.54 and MAE=3.18. Applying any of the weighting schemes results in an increase in the absolute error. This finding warrants critical attention and serves as an important diagnostic limitation of the weighting process. The fact that the unweighted sample approximates household size better than the weighted methods clearly demonstrates that data integration involves a structural trade-off: adjusting weights to improve the accuracy of certain core indicators can inevitably worsen the fit of others. When propensity models and calibration margins are optimized to correct for primary selection vectors (such as gender, age, or education), secondary variables not included in the model may undergo marginal distortions resulting from the readjustment of weights. However, since this specific variable was included in the IPW model, this misalignment must be justified by the asymmetry in the reference database. While the official benchmark values reflect the entire EU-SILC dataset, the IPW algorithm was trained on a random subsample of that survey. Consequently, the weights were optimized to replicate the distribution of that subsample, capturing its sampling variance rather than being strictly anchored to the overall population totals. Also, the present analysis focuses on the overall dataset after applying weight sharing, which comes at the cost of some of the fit achieved by IPW. Nevertheless, despite this inherent methodological trade-off, the key point to note is the consistency of results: the shared-weights model produces estimates that are slightly closer to official data than those of the current GEM method.

Figure 16 shows the distribution of employment status, illustrating the effectiveness of the calibration methodologies. The unweighted data reveals a SAE of 4.06 and a MAE of 1.51. An evaluation of the metrics reveals a trade-off between the two models. On the one hand, the Current Weighting scheme minimized the overall SAE, achieving an SAE of 0.94 compared to 1.10 for the new approach. On the other hand, the Shared Weights performs better in minimizing the MAE (0.62 versus 0.65). This divergence in metrics suggests that, although traditional calibration produces a slightly lower sum of deviations, the Shared Weights methodology penalizes individual categories less, distributing the margin of error more homogeneously and evenly across the different categories.

10.3. Bootstrap variance

To estimate the empirical variance and construct confidence intervals, a Bootstrap resampling approach was implemented. To accurately reflect the dual nature of the data, the resampling was divided into two different strategies: for the probabilistic reference sample (EU-SILC), Rao-Wu bootstrap weights were used to preserve its complex survey design, while the non-probabilistic sample (APS) and the probabilistic GEM component were resampled with replacement using simple random sampling**.** This iterative procedure requires that, in each of the 500 valid replicates, the IPW model and the composite estimate be re-estimated from scratch. This is the only analytically feasible formulation to capture not only the baseline sampling error but also the cumulative propagation of uncertainty generated throughout the selection and composite weighting process, accurately reflecting the model's structural stability in estimating the binary TEA indicator.

To approximate the empirical variance and propagate uncertainty across the main integration steps, the resampling strategy was applied across all components of the datasets. For the EU-SILC reference sample, the Rao-Wu bootstrap replication method was utilised, as it properly accounts for the complex survey design of the probabilistic data. Conversely, both components of the GEM survey—the probability (CATI) and non-probability (CAWI) samples—were resampled using an independent and identically distributed (i.i.d.) bootstrap approach (simple random sampling with replacement). Applying an i.i.d. bootstrap to the CAWI component is acceptable for non-probability samples, as they lack a formal sampling frame, while its application to the CATI sample simplifies the replication process prior to the weight combination and population expansion steps. Instead of randomly selecting individuals in isolation, the Rao-Wu algorithm resamples the primary sampling units (PSUs) within their original strata and applies a rescaling factor to the base calibration weights (Rao *et al.*, 19l92). The resampling model generates matrices of replicate weights that replace raw data in the simulation. At each iteration, the IPW model is fitted to a resampled version of the population that preserves the complexity and rigor of the official sampling design, ensuring that the trade-off between bias and variance is measured.

Table 15 demonstrates high stability throughout the integration pipeline, as evidenced by the Bootstrap results. The original sample mean estimate, set at 9.99, is in good agreement with the mean resulting from the Bootstrap process (9.80), yielding a small standard error of

just 0.92%. This precision is reflected in the narrow range of the 95% confidence intervals, which demonstrate an almost perfect overlap between the percentile-based measurement (]8.08; 11.59[) and the approximation to the normal distribution (]7.99; 11.61[). The model records a Coefficient of Variation (CV) of 9.41%, which is a standardized measure that expresses the standard error as a proportion of the estimate itself. Finally, the Monte Carlo's stability is -1.94%, which measures the distance between the estimate and the mean of all simulations. This metric evaluates the empirical bias and internal consistency of the resampling procedure, rather than indicating whether the method over- or underestimates the true population parameter.

Figure 17 illustrates the distribution of the TEA indicator from the Bootstrap simulations. The density curve exhibits a bell-shaped profile. The bootstrap distribution appears approximately symmetric, supporting the use of normal-approximation intervals as a sensitivity check alongside percentile intervals (delimited by the red dashed lines). In practice, this visual distribution indicates that our final estimate is balanced and that the model's uncertainty is correctly mapped, without severe distortions at the extremes.

Figure 18 demonstrates the computational efficiency of the method by tracking the trajectory of the cumulative standard error across simulations. As is to be expected in resampling processes, there is high volatility in the initial iterations. However, as the algorithm progresses, the curve clearly stabilizes, with its variance narrowing significantly and gradually converging towards its reference line as it approaches the 500 runs. From a methodological standpoint, this stabilization suggests that the chosen number of replicates was appropriate and sufficient for this exercise. It indicates that the overall variability generated by the integration of the databases was adequately captured, providing confidence that the reported standard error is stable.

11. STUDY AND COMPARISON WITH DIFFERENT CASE STUDIES

11.1. Statistics Canada's crowdsourcing data

Crowdsourced data is defined as non-probability survey data collected through the voluntary participation of visitors. Participation is self-selected, inclusion probabilities are unknown, and representativeness is not guaranteed by design. While crowdsourcing offers speed, low cost, and operational flexibility, it is prone to participation bias. Consequently, crowdsourced data cannot support design-based inference on its own. Statistics Canada decided to use the Labor Force Survey (LFS), through shared auxiliary variables, to apply the IPW on the crowdsourcing data collected during the COVID period (Beaumont *et al.*, 2024).

The strategy is centred on modelling the participation mechanism using a logistic model estimated by MLE, where the probability sample contributes through survey weights. Variable selection is performed using a stepwise procedure guided by a modified AIC that accounts for uncertainty from estimating population quantities via a probability sample. However, the selected models sometimes include many interaction terms, which may substantially improve the fit to the participation indicator while simultaneously yielding small estimated probabilities and extreme inverse probability weights (Beaumont *et al.*, 2024)

A key approach is how they handle model instability after the estimation is complete. Instead of applying constraints within the model, they rely on homogeneous grouping techniques, like Frank method or nppCART, to aggregate individual predictions into broader, group-level participation rates. While this stabilises the extreme weights, it comes with a trade-off. It severs the direct link between the original participation model and the final IPW estimator. Ultimately, their framework relies on fixing the problem after through discretisation, rather than preventing it at the root by regularising the model itself (Beaumont *et al.*, 2024).

The approach of the present report differs from this strategy along three interconnected dimensions. First, extreme weights and data separation were not treated as inconveniences to be smoothed over later, but as symptoms of model overfitting that had to be fixed directly at the estimation stage. Second, opted for GEE-based estimation instead of MLE for the participation model. This shift allowed for embedding inverse probability weighting directly

within a calibration framework. Rather than relying on probabilities to implicitly match the population, the GEE design mathematically constrains them to do so. The central difference, therefore, is that Beaumont *et al.* (2024) rely on MLE to estimate participation probabilities and subsequently smooth their instability through grouping, whereas this approach enforces population alignment directly via calibration during estimation and uses penalisation to prevent instability from arising.

Yet, despite taking two very different methodological paths, it is striking how closely final conclusions align. Both analyses confirm that while IPW is an effective tool for reducing participation bias, it rarely eliminates it completely. Also, share the observation that main effects are the true drivers of participation. Complex, interaction-heavy models simply do not justify their use; any marginal gains they offer are overshadowed by the instability they introduce. Ultimately, both studies lead to the same realisation: the success of IPW is less about achieving a perfect model fit and entirely about controlling how the weights behave.

11.2. US COVID-19 Trends and Impact Survey (CTIS)

The US COVID-19 Trends and Impact Survey (CTIS), conducted by the Delphi Group at Carnegie Mellon University in partnership with Facebook, is a continuous, cross-sectional internet survey and the largest public health survey conducted in the United States to date. Launched in April 2020, it invites a daily random sample of active Facebook users to report on a wide range of pandemic-related metrics, including symptoms, testing, preventive behaviours, and vaccination. With over 20 million responses in its first year, the CTIS was designed to supplement official public health surveillance. Its massive scale provides real-time indicators that bypass the delays and backlogs of traditional reporting systems, allowing researchers to track epidemiological trends (Salomon *et al.*, 2021).

The methodological design of the study's weighting scheme takes a different approach to selection bias from that used by Salomon *et al.* (2021) for the CTIS. The CTIS draws its sample directly from active Facebook users and relies on a sequential, two-step weighting procedure. First, it calculates the inverse probability of survey completion using internal platform data. Second, it applies a conventional post-stratification adjustment to align the sample with macro-level population totals, restricting this adjustment to age, gender, and geography. Instead, this study methodology introduces a pseudo-design-based framework

that directly links the non-probability sample to a highly calibrated micro-data reference survey (EU-SILC). By embedding an SCAD penalty within the IPW optimisation, it selects and regularises a multidimensional set of covariates. Rather than relying on a rigid set of basic demographic margins, this approach allows the model to actively account for complex confounding variables.

Naturally, this design difference leads to markedly different empirical outcomes and methodological trade-offs, particularly regarding bias and variance. Salomon *et al.* (2021) acknowledge that their limited post-stratification targets left their estimates vulnerable to persistent biases. Most notably, a severe overrepresentation of college-educated respondents whose weights could not be resolved. They accept this baseline bias, arguing that the true value of the CTIS lies in tracking trends over time rather than achieving absolute point accuracy. The SCAD-penalised IPW approach, on the other hand, actively targets this exact type of residual confounding by enforcing balance across a much broader selection of variables. However, as demonstrated in the integrated application, this rigorous reduction in bias does not come for free: it comes at the cost of increased variance inflation. It recognises and accepts that, in order to achieve this level of rigorous regularisation, an increase in standard errors is a necessary and justified consequence.

12. CONCLUSION AND DISCUSSION

This dissertation evaluated the integration of non-probability survey data with a probability reference sample to estimate the TEA. Testing IPW, MI, and DR estimators showed that these adjustments can reduce selection-related discrepancies under strong assumptions but require careful management of the bias-variance trade-off. This case study highlights the ongoing evolution of survey sampling in official statistics. Previously, the limitations of non-probability methods were viewed as strict barriers to their use; today, emerging approaches offer new methodological perspectives. The framework discussed here represents just one of many potential applications and leaves room for deeper exploration. A major advantage of this approach is its high adaptability to other non-probability studies, allowing researchers to experiment with different reference samples and design strategies. Furthermore, this empirical application successfully validated the operational effectiveness of these methods within the `nonprobsvy` package.

However, it is important to note that these methods are based on a set of restrictive theoretical assumptions. In empirical practice, these assumptions are often difficult or even impossible to verify, which requires a cautious interpretation of the final results. The main structural limitations fall into four key categories:

- **Missing-at-Random (MAR):** The theoretical foundation of data integration requires that all factors influencing participation in a non-probability sample (or the selection mechanism) be observed and included in the models. However, it is mathematically impossible to use the observed data to prove the absence of latent variables that may have biased the selection. Factors such as intrinsic motivation or digital literacy are key determinants of participation but are rarely captured by demographic covariates. If unobserved confounders exist, all three methods fail: in IPW, the calculated weights do not correct for the actual bias; in MI, the imputation model will generate distorted values; and in DR, the double robustness breaks down, since both the participation and prediction model share the same lack of information.
- **Positivity:** Any integration method assumes the positivity condition, which dictates that every individual in the target population has a probability greater than zero of being included in the non-probability sample. In reality, this assumption is often violated due to the existence of subgroups that are impossible to identify (e.g.,

severely excluded populations). The presence of these "structural zeros" affects all techniques: in IPW, it results in extreme and unstable weights; in MI, it forces the model to make dangerous extrapolations outside the observed covariate space; and the DR estimator, although it combines both approaches, is unable to synthesize information about profiles that are structurally absent from the data.

- **Independence of Observations and Network Effects:** The statistical modelling underlying IPW, MI, and DR typically assume that observations are i.i.d.. However, in non-probability samples, the reality is characterized by social contagion and clustering phenomena. The sharing of surveys among peer networks or geographic clusters creates structural dependencies. Ignoring this underlying correlation among participants compromises the fit of the models (whether the logistic model in IPW or the predictive model in MI) and systematically leads to an underestimation of the true variance, creating a false perception of high statistical precision.
- **Perfect Reference Sample:** Ultimately, the entire integration paradigm assumes the existence of a high-quality probability sample that serves as the population "ground truth." However, this premise overlooks the fact that official reference sources face growing challenges. If the probability sample suffers from structural nonresponse bias or severe measurement errors, the reliability of the entire integration is compromised. By forcing the non-probability sample to align with the reference sample, IPW, MI, and DR run the risk of simply transferring and propagating the latent biases from the official source into the final estimate.

From an analytical standpoint, the IPW offers a decisive operational advantage over MI and DR due to its flexibility. By focusing on modelling the participation mechanism, the IPW produces a universal weighting solution that can be applied to any variable of interest in the survey. Although this versatility often comes at the cost of higher variance inflation, it contrasts with the MI and DR methodologies, which are limited by the need to construct a specific predictive model for each variable to be analysed. It was precisely the recognition of this cross-cutting flexibility of the IPW, in close collaboration with and at the encouragement of the STATEC team, that motivated allocating most of the practical component's time to its development and algorithmic fine-tuning. As this work represents the first methodological initiative to actively explore this topic of data integration within STATEC, prioritising the IPW was deemed a strategic necessity to establish a solid foundation. The results discussed

here reflect the methodological level achieved to date; with a broader timeframe in the future, the IPW model may undergo further refinements to achieve more rigorous control of its variance, while also opening the door for the MI and DR methods to be developed and explored with the same technical depth.

The complexity of this work did not lie in the technical application or computational testing of the models themselves. Thanks to recent advances in the statistical community and the development of tools such as the `nonprobsvy` package, much of the computational burden associated with estimation is now considerably optimised and automated. The true and most demanding methodological challenge lay in the structural definition of the covariate space and the precise tuning of penalties appropriate to the behaviour of each model. This stage proved extremely meticulous, requiring an exhaustive, iterative analysis to understand how changes in seemingly marginal details influenced the distribution of weights and the stability of the final estimates. Consequently, although this dissertation presents the results of five specifications, the underlying empirical process involved testing, evaluating, and discarding several intermediate iterations. This empirical finding underscores that the critical core of data integration relies heavily on analytical sensitivity during the parameterisation phase. If there were a longer time frame to explore this practical component, this very complexity would open doors for an even more in-depth study of selection and penalisation, which would undoubtedly result in even more refined, robust, and well-controlled estimates.

In addition, the robustness of all these approaches depends on the correct specification of their underlying models: while IPW and DR require modelling of the participation mechanism, MI and DR require modelling of the outcome itself. Despite the short time devoted to the practical component of this work, the results suggest that the proposed integration approach can produce estimates comparable to, and in some diagnostics more favourable than, the current GEM calibration. With more time and continued development, this methodology holds strong potential to become a robust and scalable framework for the integration of non-probabilistic data.

13. REFERENCES

A Grammar of Data Manipulation. (n.d.). Retrieved 26 August 2026, from https://dplyr.tidyverse.org/

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013). Summary Report of the AAPOR Task Force on Non-probability Sampling. *Journal of Survey Statistics and Methodology*, *1*(2), 90–143. https://doi.org/10.1093/jssam/smt008

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J., & Chu, K. (2024). Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. *Survey Methodology*, *50*(1). http://www.statcan.gc.ca/pub/12-001-x/2024001/article/00004-eng.htm

Buuren, S. V., & Groothuis-Oudshoorn, K. (2011). **mice**: Multivariate Imputation by Chained Equations in *R*. *Journal of Statistical Software*, *45*(3). https://doi.org/10.18637/jss.v045.i03

Chrostowski, Ł., Chlebicki, P., & Beręsewicz, M. (2025). *nonprobsvy—An R package for modern methods for non-probability surveys* (arXiv:2504.04255). arXiv. https://doi.org/10.48550/arXiv.2504.04255

Create Elegant Data Visualisations Using the Grammar of Graphics. (n.d.). Retrieved 26 August 2026, from https://ggplot2.tidyverse.org/

Czerniawska, A., Chlebicki, P., Chrostowski, Ł., & Beręsewicz, M. (2026). *Data integration of non-probability and probability samples with deterministic predictive mean matching* (arXiv:2403.13750). arXiv. https://doi.org/10.48550/arXiv.2403.13750

Daoud, J. I. (2017). Multicollinearity and Regression Analysis. *Journal of Physics: Conference Series*, *949*, 012009. https://doi.org/10.1088/1742-6596/949/1/012009

Elliott, M. R., & Valliant, R. (2017). Inference for Nonprobability Samples. *Statistical Science*, *32*(2). https://doi.org/10.1214/16-STS598

Eurostat. (2024). *Methodological guidelines and description of EU-SILC target variables: 2024 operation (Version 7)* [Methodological guidelines]. European Commission.

Fan, J., & Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. https://doi.org/10.1198/016214501753382273

Favre-Martinoz, C., & Gros, E. (2017). *The weight sharing method (Version 1)* [Technical Report]. Institut national de la statistique et des études économiques (Insee), Département des méthodes statistiques.

Fuller, W. A. (2011). *Sampling Statistics*. John Wiley & Sons.

GEM Global Entrepreneurship Monitor. (n.d.). GEM Global Entrepreneurship Monitor. Retrieved 26 August 2026, from https://www.gemconsortium.org/about/wiki

GEM Luxembourg. (2025). *Global Entrepreneurship Monitor Luxembourg 2024/2025* [Report]. STATEC.

Gendall, P., & Davis, P. (1993). Are callbacks a waste of time? *Marketing Bulletin*, *4*(1), 53–57.

Groves, R. M., Fowler, F. J., Couper, M., Lepkowski, J. M., Singer, E., & Tourangeau, R. (Eds). (2009). *Survey Methodology* (2nd ed). Wiley.

Groves, R. M., & Peytcheva, E. (2008). The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. *Public Opinion Quarterly*, *72*(2), 167–189. https://doi.org/10.1093/poq/nfn011

Held, L., & Sabanés Bové, D. (2014). *Applied Statistical Inference: Likelihood and Bayes*. Springer. https://doi.org/10.1007/978-3-642-37887-4

Hill, S., Boutaleb, F., Coduras, A., Friedl, C., Guerrero, M., Ionescu-Somers, A., Menipaz, E., Meyer, N., Sahasranamam, S., Samsami, M., Shay, J., & Zbierowski, P. (2025). *GEM 2024/2025 Global Report: Entrepreneurship Reality Check*. Global Entrepreneurship Research Association (GERA).

Hoo, Z. H., Candlish, J., & Teare, D. (2017). What is an ROC curve? *Emergency Medicine Journal*, *34*(6), 357–359. https://doi.org/10.1136/emermed-2017-206735

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (1st edn). Wiley. https://doi.org/10.1002/9781118548387

Kish, L. (1995). *Survey sampling*. Wiley.

Lee, S. (2006). An Evaluation of Nonresponse and Coverage Errors in a Prerecruited Probability Web Panel Survey. *Social Science Computer Review*, *24*(4), 460–475. https://doi.org/10.1177/0894439306288085

Li, C. (2013). Little's test of missing completely at random. *The Stata Journal*, *13*(4), 795–809.

Loosveldt, G., & Sonck, N. (2008). An evaluation of the weighting procedures for an online access panel survey. *Survey Research Methods*, *2*, 93–105.

LUSTAT Data Explorer • Population by age and sex on 1st January. (n.d.). Retrieved 26 August 2026, from https://lustat.statec.lu/vis?fs[0]=Topics%2C1%7CPopulation%20and%20employment%23B%23%7CPopulation%20structure%23B1%23&pg=0&fc=Topics&bp=true&snb=9&lc=en&df[ds]=ds-release&df[id]=DF_B1102&df[ag]=LU1&df[vs]=1.0&dq=A.A03%2BA02.SSL20%2BSSL19%2BSL13%2BSL12%2BSL11%2BSL10%2BSL09%2BSL08%2BSL07%2BSL06%2BSL05&pd=2024%2C2024&to[TIME_PERIOD]=false

Marvel, M. R., Davis, J. L., & Sproul, C. R. (2016). Human Capital and Entrepreneurship Research: A Critical Review and Future Directions. *Entrepreneurship Theory and Practice*, *40*(3), 599–626. https://doi.org/10.1111/etap.12136

Moser, C. A. (1952). Quota Sampling. *Journal of the Royal Statistical Society. Series A (General)*, *115*(3), 411. https://doi.org/10.2307/2980740

R: The R Project for Statistical Computing. (n.d.). Retrieved 26 August 2026, from https://www.r-project.org/

Rao, J. N. K., Wu, C. F. J., & Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, *18*(2), 209–217.

Read Excel Files. (n.d.). Retrieved 26 August 2026, from https://readxl.tidyverse.org/

Response Rates Calculator—AAPOR. (2022, November 17). https://aapor.org/response-rates/

Salganik, M. J. (2006). Variance Estimation, Design Effects, and Sample Size Calculations for Respondent-Driven Sampling. *Journal of Urban Health*, *83*(S1), 98. https://doi.org/10.1007/s11524-006-9106-x

Salomon, J. A., Reinhart, A., Bilinski, A., Chua, E. J., La Motte-Kerr, W., Rönn, M. M., Reitsma, M. B., Morris, K. A., LaRocca, S., Farag, T. H., Kreuter, F., Rosenfeld, R., & Tibshirani, R. J. (2021). The US COVID-19 Trends and Impact Survey: Continuous real-time measurement of COVID-19 symptoms, risks, protective behaviors, testing, and vaccination. *Proceedings of the National Academy of Sciences*, *118*(51), e2111454118. https://doi.org/10.1073/pnas.2111454118

Scheidel, W. (2009). Population and Demography. In A. Erskine (Ed.), *A Companion to Ancient History* (1st edn, pp. 134–145). Wiley. https://doi.org/10.1002/9781444308372.ch13

Schonlau, M., Zapert, K., Simon, L. P., Sanstad, K. H., Marcus, S. M., Adams, J., Spranca, M., Kan, H., Turner, R., & Berry, S. H. (2004). A Comparison Between Responses From a Propensity-Weighted Web Survey and an Identical RDD Survey. *Social Science Computer Review*, *22*(1), 128–138. https://doi.org/10.1177/0894439303256551

Thompson, S. K. (2012). *Sampling* (3rd ed). John Wiley & Sons. https://doi.org/10.1002/9781118162934

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

United Nations. (2008). *Designing Household Survey Samples: Practical Guidelines*. United Nations. https://doi.org/10.18356/f7348051-en

United Nations Statistical Division. (2017). *Principles and recommendations for population and housing censuses*. United Nations.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196. https://doi.org/10.3758/BF03206482

Wolf, C., Joye, D., Smith, T., & Fu, Y. (2016). *The SAGE Handbook of Survey Methodology*. SAGE Publications Ltd. https://doi.org/10.4135/9781473957893

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, *48*(2). http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00002-eng.htm

Wu, C., & Thompson, M. E. (2020). *Sampling Theory and Practice*. Springer International Publishing. https://doi.org/10.1007/978-3-030-44246-0

14. APPENDIX

Table 1: Disposition Summary and Response Rates by Survey Mode (APS 2024)

Mode	I	P	R	NC	Inel.	UH	Calls	*e*	RR3	*PR*
Online	1233	153	3157	318	0	0	4861	–	–	25.4%
Fixed	203	32	1557	0	6516	4692	13000	22%	7.2%	–
Mobile	597	86	5362	0	1170	11285	18500	84%	3.9%	–

Notes:
I = Completed interviews; P = Partial completes; R = Refusals; NC = Non-contacts but eligible; Inel. = Ineligible cases; UH = Unknown eligibility because non contact; Calls = Total contact attempts (phone numbers called or invitations sent); e = Estimated proportion of cases of unknown eligibility that are eligible; PR = *Participation rate*.
Estimated eligibility rate is calculated as the observed proportion of eligible cases among all known eligibility outcomes:

$$e = \frac{\text{Eligible}}{\text{Eligible} + \text{Ineligible}}$$

This rate is then applied to the unknown eligibility cases (UH) to estimate how many are likely to be eligible.
RR3 is calculated as:

$$RR3 = \frac{I}{I + P + R + NC + (UH \times e)}$$

Participation rate is defined as: $I/(I + R)$. Online mode has no unknown eligibility cases; thus no estimation was applied.

Source: *Global Entrepreneurship Monitor Luxembourg Report 2024/2025, Page 140.*

Table 2: Harmonisation Mapping Table (APS Dataset)

Harmonised Variable	Harmonised Category	APS Variable "OCCUP" Categories
Activity Status	Employed	Employed by others in full-time work, Employed by others in part-time work, Self-employed / entrepreneur
	Unemployed	Unemployed
	Retired	Retired
	Inactive	Student, Homemaker, Caretaking of family member, Not seeking employment, Other
Employment Status	Self-employed	Self-employed / entrepreneur
	Employee	Employed by others in full-time work, Employed by others in part-time work

Harmonised Variable	Harmonised Category	APS Variable "OCCUP" Categories
Employment Status	Others	Unemployed, Retired, Student, Homemaker, Caretaking of family member, Not seeking employment, Other
Full or Part-time	Full-time	Employed by others in full-time work, Self-employed / entrepreneur
	Part-time	Employed by others in part-time work
	Others	Unemployed, Retired, Student, Homemaker, Caretaking of family member, Not seeking employment, Other

Table 3: Distribution of variables in the EU-SILC – before imputation

Variable	Category	Count	Percentage (%)	Weighted Count	Weighted Percentage (%)
Age	18-24	323	10.1	61,183	14.1
	25-34	592	18.6	96,169	22.2
	35-44	783	24.6	97,400	22.5
	45-54	730	22.9	90,282	20.9
	55-64	755	23.7	87,619	20.3
Gender	Male	1,575	49.5	217,255	50.2
	Female	1,608	50.5	215,398	49.8
Household Size	1	503	15.8	70,720	16.3
	2	882	27.7	100,374	23.2
	3+	1,798	56.5	261,559	60.5
Country of Birth	Luxembourg	1,421	44.6	179,155	41.4
	Portugal	300	9.4	53,573	12.4
	Other countries in Europe	1,088	34.2	147,959	34.2
	Rest of the World	374	11.7	51,965	12.0
Parents' Country of Birth	Luxembourg	861	27.0	98,265	22.7
	Abroad	2,300	72.3	328,492	75.9
	NA	22	0.7	5,896	1.4
Educational Level	Low	408	12.8	52,304	12.1
	Medium	1,035	32.5	145,373	33.6
	High	1,532	48.1	201,076	46.5
	NA	208	6.5	33,901	7.8

Variable	Category	Count	Percentage (%)	Weighted Count	Weighted Percentage (%)
Activity Status	Employed	2,179	68.5	299,134	69.1
	Unemployed	126	4.0	16,237	3.8
	Retired	266	8.4	26,901	6.2
	Inactive	612	19.2	90,381	20.9
Employment Status	Self-employed	198	6.2	29,913	6.9
	Employee	1,981	62.2	269,221	62.2
	Others	1,004	31.5	133,519	30.9
Full/Part-Time	Full-time	1,757	55.2	249,413	57.6
	Part-time	375	11.8	43,603	10.1
	Others	1,004	31.5	133,519	30.9
	NA	47	1.5	6,118	1.4

Table 4: Distribution of variables in the whole APS of the GEM – before imputation

Variable	Category	Count	Percentage (%)	Weighted Count	Weighted Percentage (%)
Age	18-24	195	9.6	241	11.9
	25-34	378	18.7	480	23.8
	35-44	457	22.6	477	23.6
	45-54	501	24.8	433	21.4
	55-64	490	24.2	390	19.3
Gender	Male	1,052	52.1	1,031	51.0
	Female	967	47.8	988	48.9
	NA	2	0.1	2	0.1
Household Size	1	261	12.9	248	12.3
	2	559	27.7	525	26.0
	3+	1,196	59.2	1,239	61.3
	NA	5	0.2	9	0.4
Country of Birth	Luxembourg	1,202	59.5	791	39.1
	Portugal	195	9.6	274	13.6
	Other countries in Europe	497	24.6	624	30.9
	Rest of the World	119	5.9	325	16.1
	NA	8	0.4	8	0.4

Variable	Category	Count	Percentage (%)	Weighted Count	Weighted Percentage (%)
Parents' Country of Birth	Luxembourg	729	36.1	489	24.2
	Abroad	1,292	63.9	1,532	75.8
Educational Level	Low	316	15.6	393	19.4
	Medium	670	33.2	641	31.7
	High	995	49.2	954	47.2
	NA	40	2.0	33	1.6
Activity Status	Employed	1,502	74.3	1,411	69.8
	Unemployed	57	2.8	90	4.4
	Retired	219	10.8	185	9.2
	Inactive	224	11.1	314	15.6
	NA	19	0.9	20	1.0
Employment Status	Self-employed	154	7.6	109	5.4
	Employee	1,348	66.7	1,302	64.4
	Others	500	24.7	590	29.2
	NA	19	0.9	20	1.0
Full/Part-Time	Full-time	1,259	62.3	1,183	58.5
	Part-time	243	12.0	229	11.3
	Others	500	24.7	590	29.2
	NA	19	0.9	20	1.0
TEA	No	1,841	91.1	1,822	90.1
	Yes	180	8.9	199	9.9
Easy Start	Agree	692	34.2	771	38.1
	Otherwise	995	49.2	929	46
	NA	334	16.5	321	15.9
Fear of Failure	Agree	1,014	50.2	976	48.3
	Otherwise	941	46.6	989	48.9
	NA	66	3.3	56.5	2.8
Opportunity	Agree	471	23.3	550	27.2
	Otherwise	1,236	61.2	1,167	57.7
	NA	314	15.5	304	15
Skill	Agree	806	39.9	831	41.1
	Otherwise	1,150	56.9	1131	56
	NA	65	3.2	59.2	2.9

Note: This distribution analysis was made in the whole APS sample. The filtering of the non-probabilistic side was made only after the imputation of missing values (Table 8).

Table 5: Little's test results

Dataset	χ2 Statistic	df	p-value	Missing Patterns	Conclusion
SILC	197.0	83	< 0.001	13	Reject MCAR
APS	458.0	342	< 0.001	34	Reject MCAR

Table 6: Significant predictors of missingness for incomplete variables

Dataset	Incomplete Variable	Significant Predictor	Coefficient (β)	p-value
SILC	Parents' Country of Birth	Age (55-64)	-2.31	0.013
		Employment Status (Employee)	-1.32	0.024
		Household Size (2)	-2.77	< 0.001
		Household Size (3+)	-1.99	< 0.001
	Educational Level	Country of Birth (Portugal)	0.53	0.039
		Country of Birth (Other Europe)	0.51	0.005
		Country of Birth (Rest of World)	1.18	< 0.001
		Employment Status (Employee)	-0.55	0.033
	Citizenship	Country of Birth (Portugal)	1.14	0.028
		Country of Birth (Other Europe)	0.85	0.032
		Household Size (3+)	-1.25	0.004
	Full / Part-Time	Gender (Female)	0.62	0.031
		Country of Birth (Rest of World)	1.01	0.012
		Employment Status (Employee)	-1.29	< 0.001
		Employment Status (Others)	-4.66	< 0.001
APS	Skill	Gender (Female)	0.56	0.030
	Easy Start	Gender (Female)	0.55	< 0.001
	Opportunity	Age (25-34)	0.68	0.010
		Gender (Female)	0.50	< 0.001

Note: The Coefficient (β) represents the log-odds of a value being missing. A positive coefficient indicates a higher probability of missing data for that category compared to the baseline, while a negative coefficient indicates a lower probability. Variables without statistically significant predictors ($p \geq 0.05$) were omitted.

Figure 1: Analysis of trace plots of imputed missing values dataset of the EU-SILC

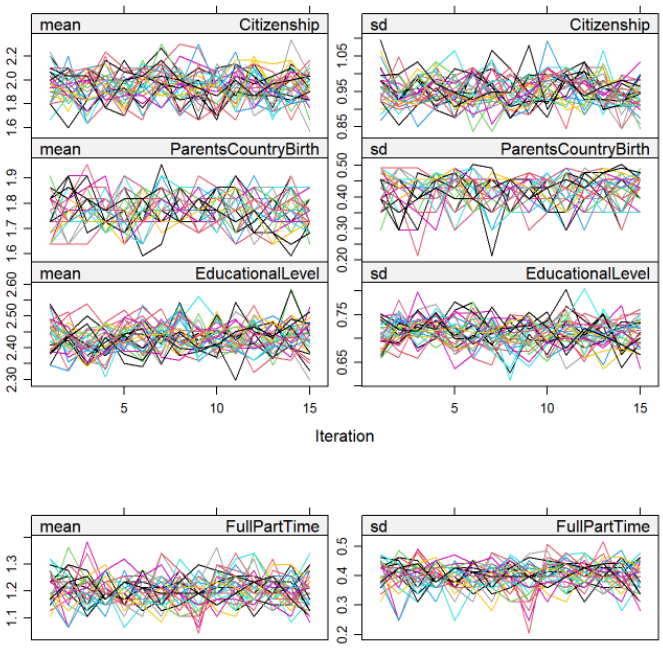


Figure 2: Analysis of trace plots of imputed missing values dataset of APS of the GEM

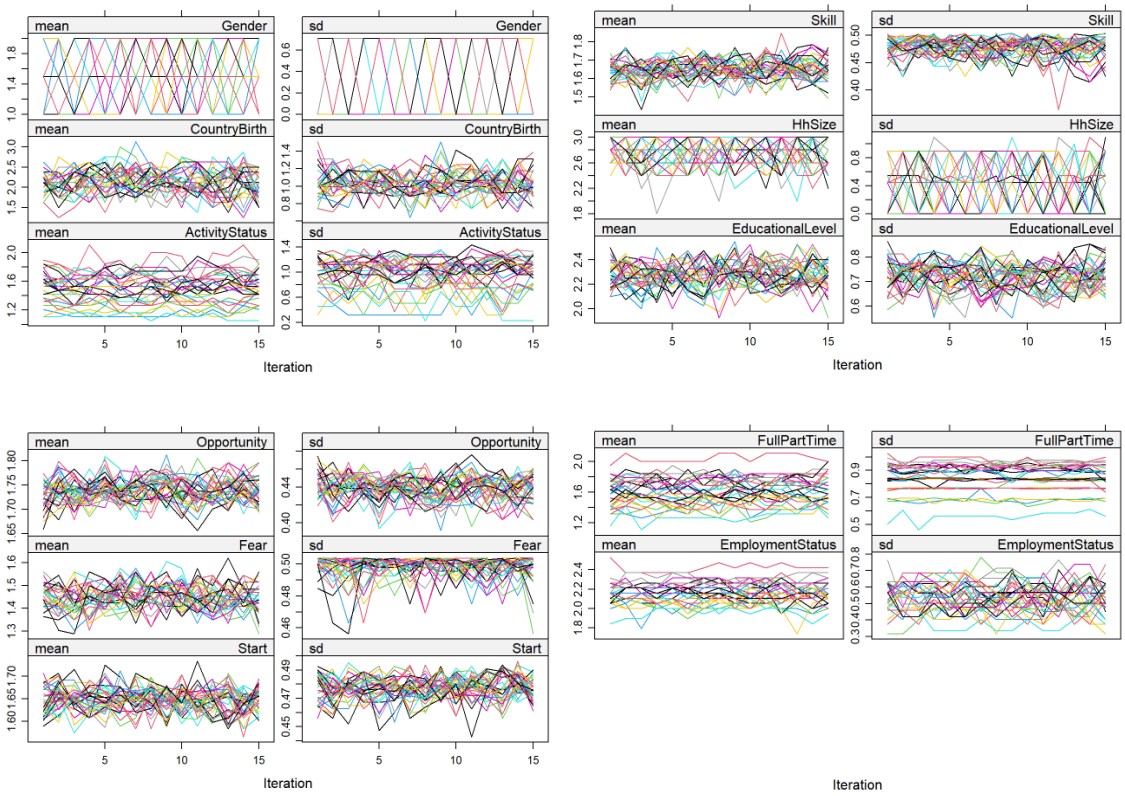


Figure 3: Analysis of strip plots of imputed missing values dataset of the EU-SILC

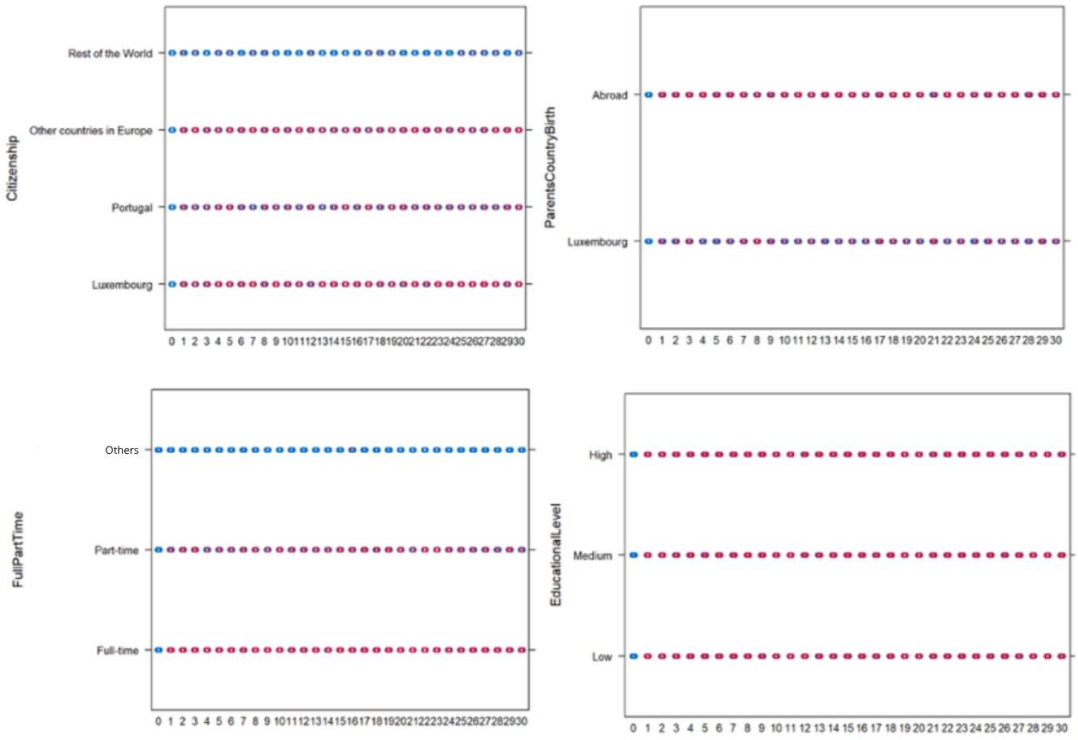


Figure 4: Analysis of strip plots of imputed missing values dataset of APS of the GEM

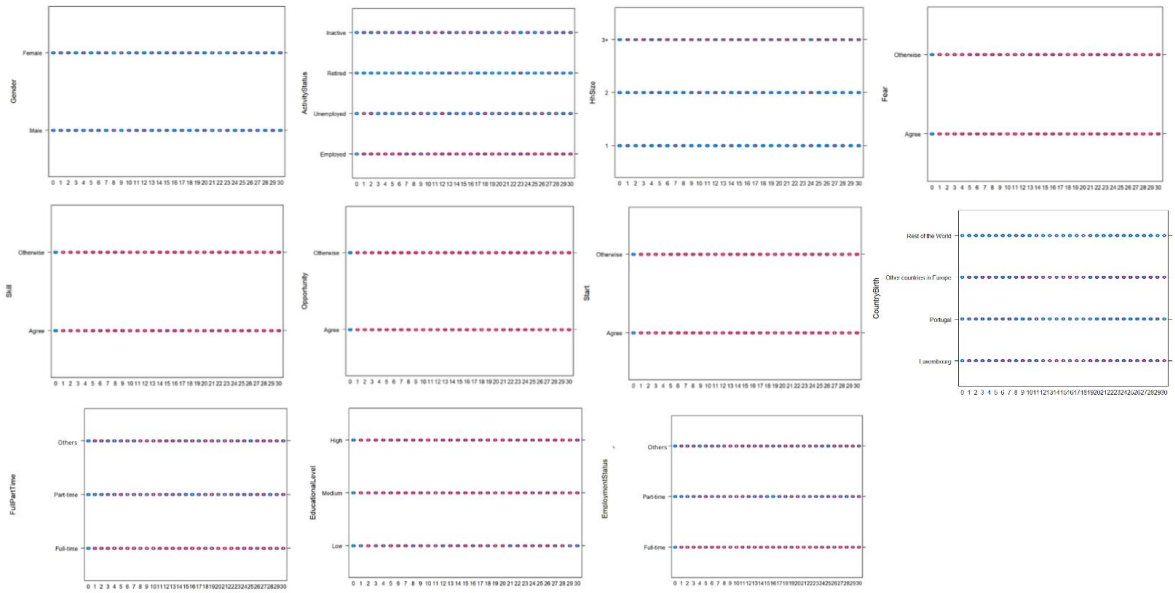


Figure 5: Frequency distributions of variables before and after MICE imputation – Visual representation in the EU-SILC

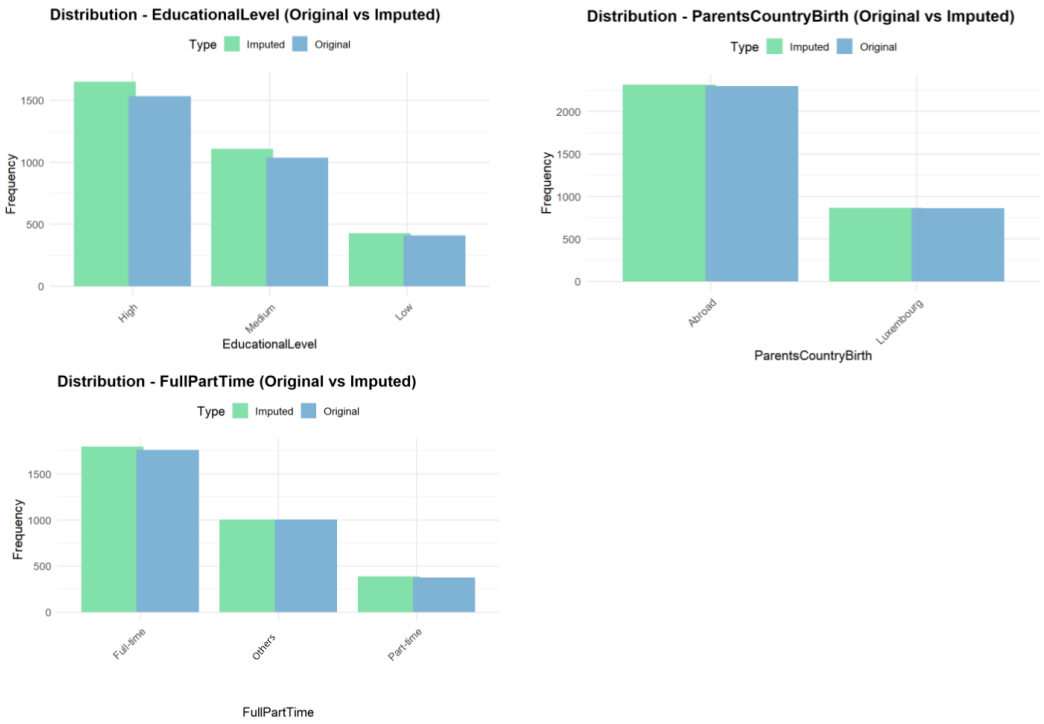


Figure 6: Frequency distributions of variables before and after MICE imputation – Visual representation in the APS of the GEM

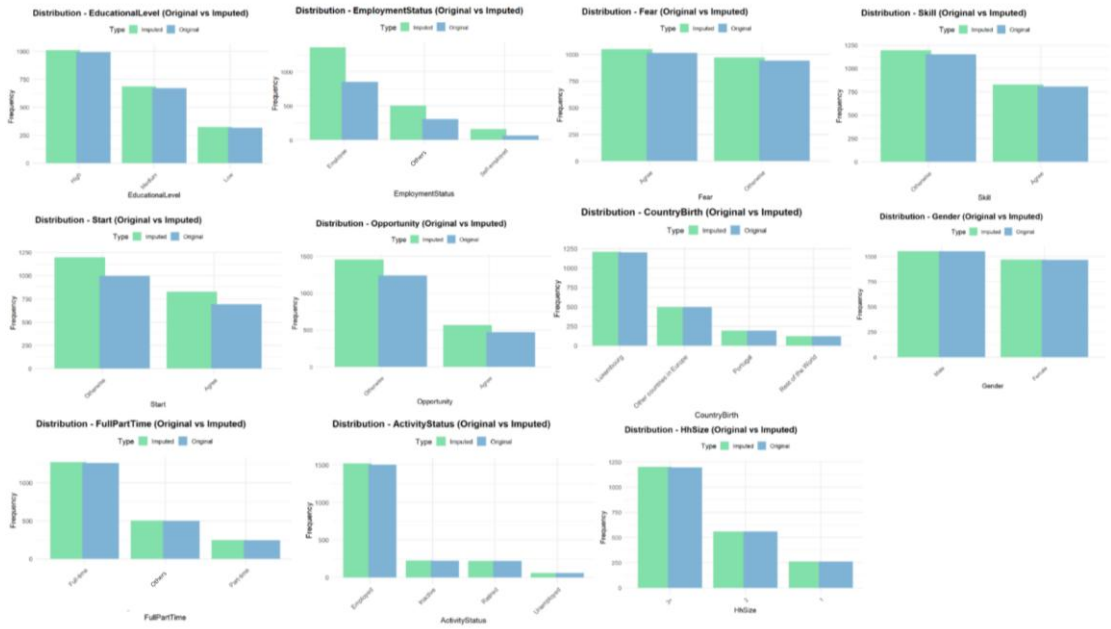


Table 7: Distribution of variables in the EU-SILC – after imputation

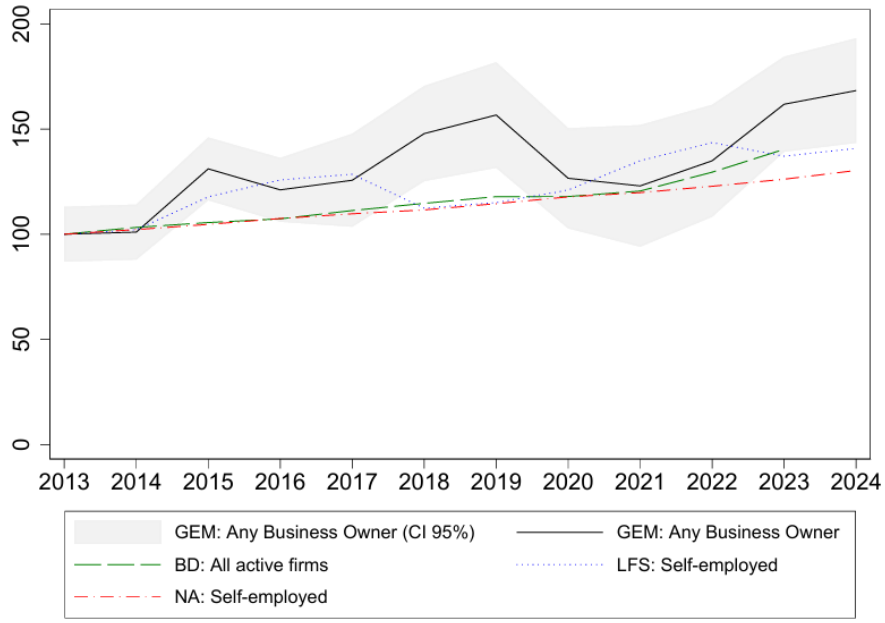
Variable	Category	Count	Percentage (%)	Weighted Count	Weighted Percentage (%)
Parents' Country of Birth	Luxembourg	867	27.2	102,076	23.6
	Abroad	2,316	72.8	330,057	76.4
Educational Level	Low	426	13.4	54,812	12.7
	Medium	1,107	34.8	160,016	37
	High	1,650	51.8	217,825	50.3
Full/Part-Time	Full-time	1,793	56.3	254,097	58.7
	Part-time	386	12.1	45,037	10.4
	Others	1,004	31.5	133,519	30.9

Table 8: Distribution of variables in the non-probabilistic side of the APS of the GEM – after imputation

Variable	Category	Count	Percentage (%)	Weighted Count	Weighted Percentage (%)
Age	18-24	138	11.3	118	11.7
	25-34	280	22.9	238	23.6
	35-44	280	22.9	240	23.8
	45-54	279	22.9	217	21.5
	55-64	244	20	195	19.3
Gender	Male	612	50.1	513	50.9
	Female	609	49.9	496	49.1
Household Size	1	140	11.5	117	11.5
	2	354	29	282	27.9
	3+	727	59.5	611	60.5
Country of Birth	Luxembourg	772	63.2	402	39.8
	Portugal	115	9.4	132	13.1
	Other countries in Europe	293	24	326	32.3
	Rest of the World	41	3.4	150	14.8
Parents' Country of Birth	Luxembourg	426	34.9	235	23.3
	Abroad	795	65.1	774	76.7

Variable	Category	Count	Percentage (%)	Weighted Count	Weighted Percentage (%)
Educational Level	Low	182	14.9	191	18.9
	Medium	443	36.3	327	32.4
	High	596	48.8	491	48.6
Activity Status	Employed	915	74.9	712	70.5
	Unemployed	30	2.5	42	4.2
	Retired	116	9.5	90	8.9
	Inactive	160	13.1	166	16.4
Employment Status	Self-employed	64	5.2	55	5.4
	Employee	851	69.7	657	65
	Others	306	25.1	298	29.5
Full/Part-Time	Full-time	772	63.2	597	59.2
	Part-time	143	11.7	114	11.3
	Others	306	25.1	298	29.5
TEA	No	1,120	91.7	903	89.4
	Yes	101	8.3	107	10.6
Easy Start	Agree	423	34.6	406	40.2
	Otherwise	798	65.4	603	59.8
Fear of Failure	Agree	716	58.6	572	56.7
	Otherwise	505	41.4	437	43.3
Opportunity	Agree	285	23.3	265	26.2
	Otherwise	936	76.7	745	73.8
Skill	Agree	434	35.5	370	36.7
	Otherwise	787	64.5	639	63.3

Figure 7: Entrepreneurship trends in APS. (Index, 2013 = 100)



Notes: BD = Business Demography; NA = National Account; LFS = Labor Force Survey
Source: STATEC BD 2013–2023, STATEC NA; STATEC LFS and GEM APS data 2013-2023.

Source: *Global Entrepreneurship Monitor Luxembourg Report 2024/2025, Page 143.*

Table 9: Performance tests on stack models

Metric	**Model 1 (Interaction Model)**	**Model 2 (Base Model)**
Specification	Main effects + Interaction (EducationalLevel:EmploymentStatus)	Main variables (Age + Gender + CountryBirth + ParentsCountryBirth + EducationalLevel + EmploymentStatus + HhSize)
Model Degrees of Freedom (df)	20	16
Residual Df	4384	4388
Residual Deviance	4963.0	4971.9
AIC	5003.016	5003.944
Deviance Test ($\boldsymbol{Pr} > \chi 2$)	0.06293	-

Table 10: Performance tests on outcome models

Metric	Model 1 (Interaction Model)	Model 2 (Base Model)
Specification	Main effects + Interaction (EducationalLevel:EmploymentStatus)	Main variables (Age + Gender + CountryBirth + ParentsCountryBirth + EducationalLevel + EmploymentStatus + HhSize)
Model Degrees of Freedom (df)	20	16
Residual Df	1201	1205
Residual Deviance	610.01	613.64
AIC	650.005	645.638
Deviance Test ($\boldsymbol{Pr} > \chi 2$)	0.458	-

Table 11: Results of IPW models

Model	Proportion of TEA (%)	Standard error (%)	Maximum weight	Variance of weights	Deff	Deft
1	10	1.4	2497.5	64714.6	1.4886	1.2201
2	10.8	1.6	2275.6	64366.3	1.4860	1.2190

Figure 8: Boxplots of IPW weights by model

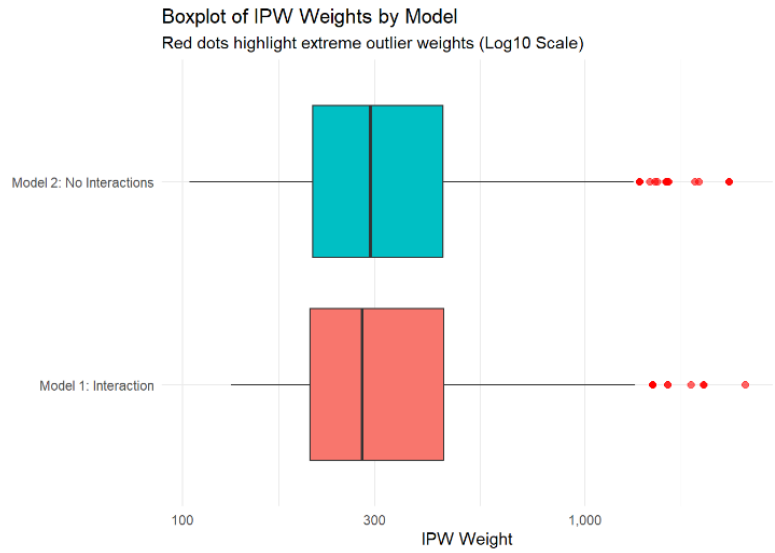


Figure 9: Density plots of IPW weights by model

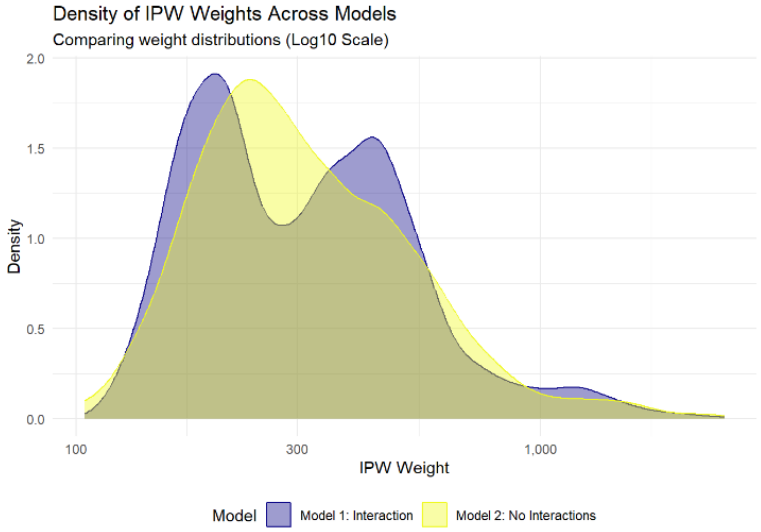


Table 12: Results of MI models

Model	Proportion of TEA (%)	Standard error (%)
GLM	10.5	1.2
PMM	11.5	0.5

Figure 10: Distribution of imputed probabilities in the EU-SILC

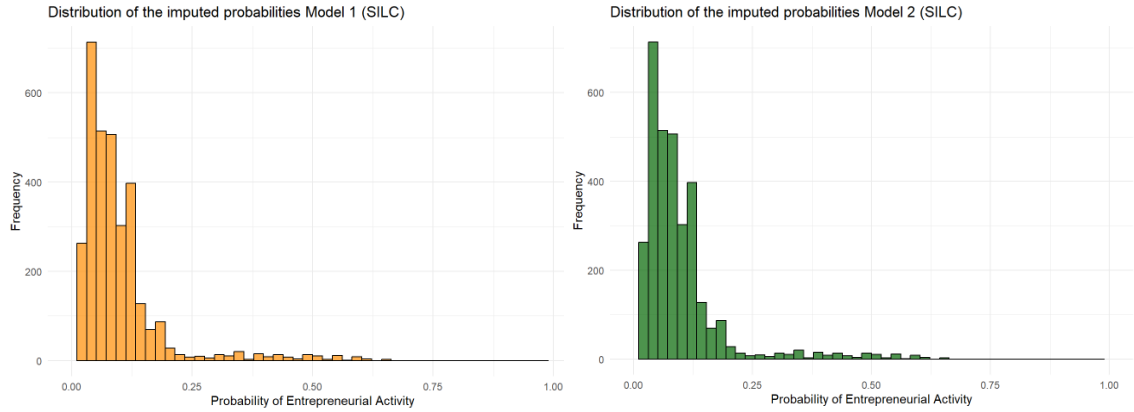


Figure 11: ROC curve of MI Models

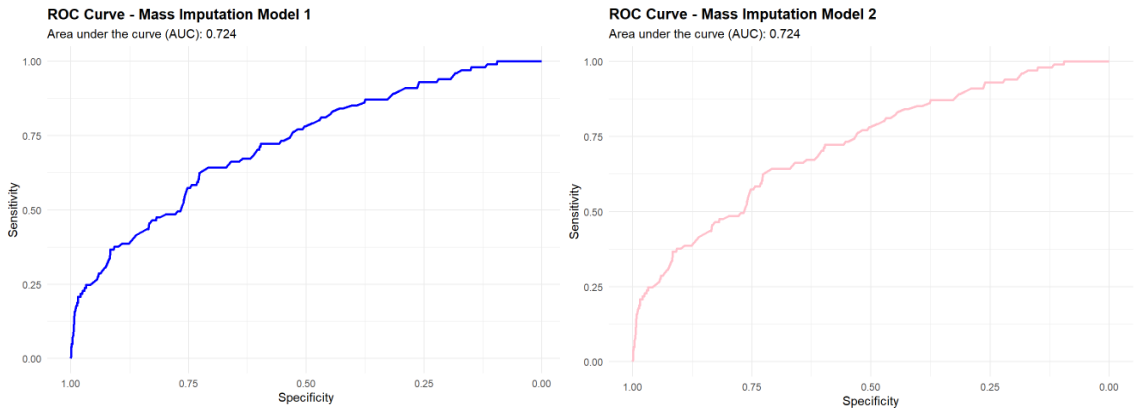


Table 13: Results of DR model

Proportion of TEA (%)	Standard error (%)	Maximum weight	Variance of weights	Deff	Deft
10.9	1.5	2275.6	64366.3	1.4860	1.2190

Figure 12: Density of DR weights

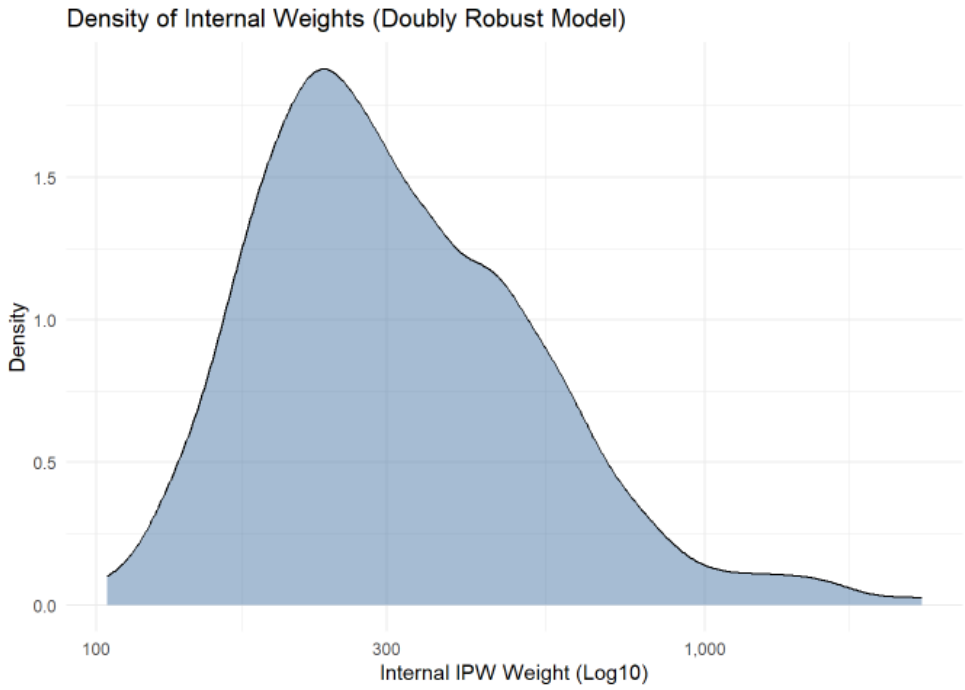


Figure 13: ROC curve of DR

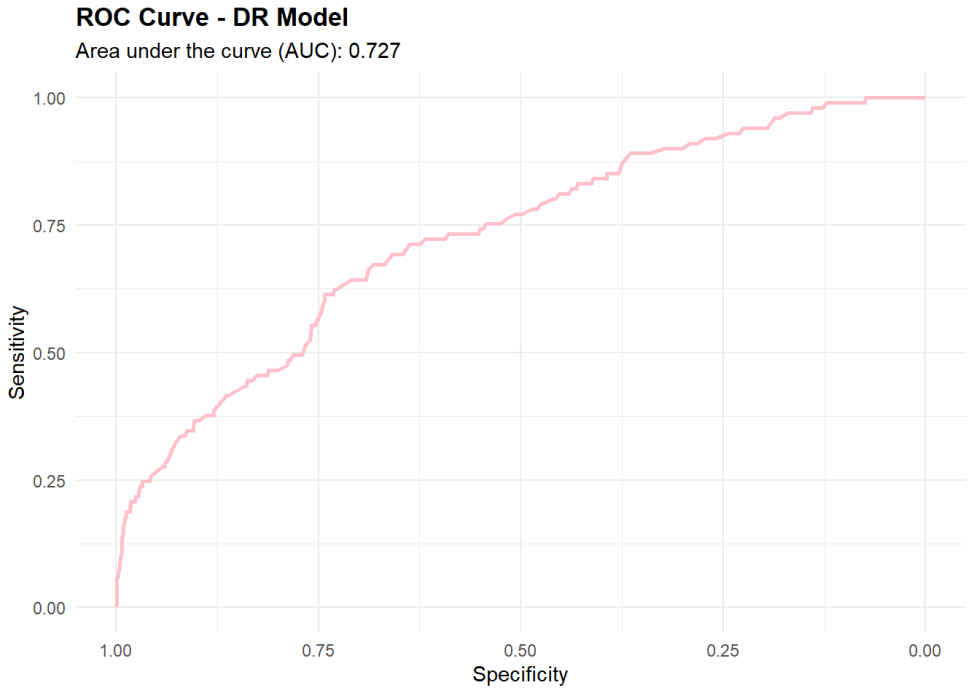
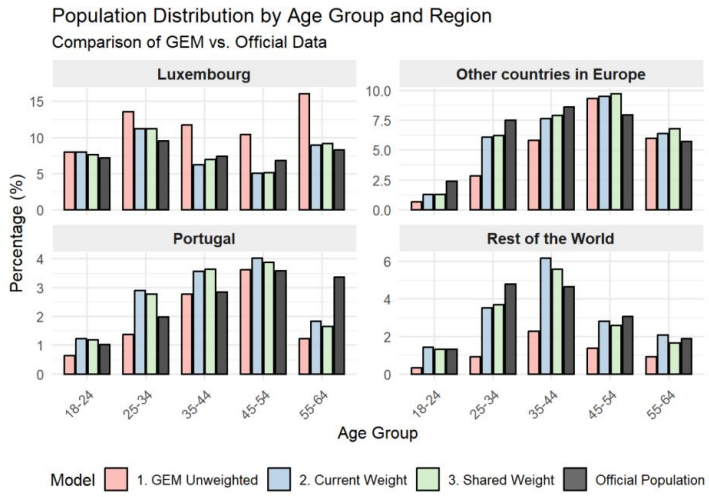


Table 14: Final comparison on the non-probabilistic side of the APS

Model	Proportion of TEA (%)	Standard error (%)	Maximum weight	Variance of weights	Deff	Deft
APS unweighted	8.3	0.8	-	-	-	-
APS current weighting	10.6	1.4	-	-	1.9419	1.3935
IPW Model 1	10	1.4	2497.5	64714.6	1.4886	1.2201
IPW Model 2	10.9	1.6	2275.6	64366.3	1.4860	1.2190
MI with GLM	10.5	1.2	-	-	-	-
MI with PMM	11.5	0.5	-	-	-	-
DR	10.9	1.5	2275.6	64366.3	1.4860	1.2190

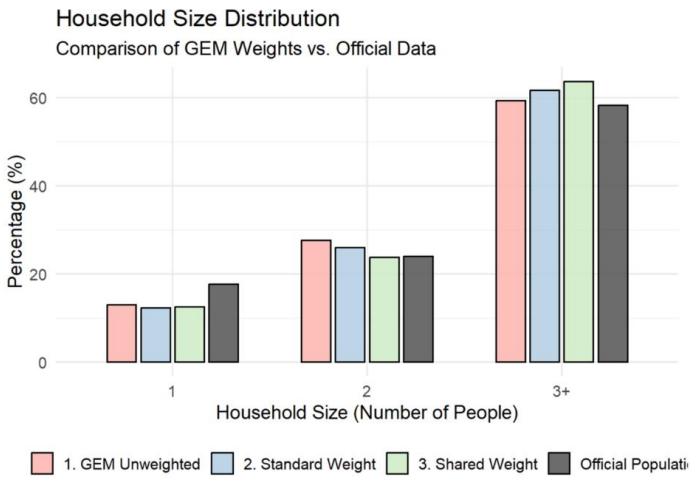
Figure 14: Country of Birth indicator analysis by Age Group



Whole GEM with	Sum of Absolute Error	Mean Absolute Error
Shared Weights	**2.31**	**0.82**
Current Weighting scheme	2.61	0.90
Unweighted	41.24	3.57

Reference: *STATEC, demographic statistics.*

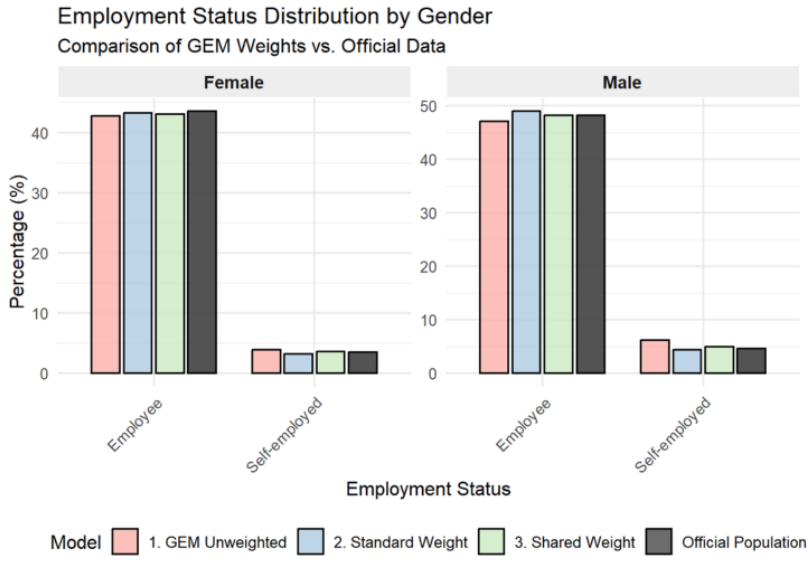
Figure 15: Household size indicator analysis



Whole GEM with	Sum of Absolute Error	Mean Absolute Error
Unweighted	**9.54**	**3.18**
Shared weights	10.71	3.57
Current Weighting scheme	10.74	3.58

Reference: *STATEC, EU-SILC.*

Figure 16: Employment status indicator analysis by Gender



Whole GEM with	Sum of Absolute Error
Current Weighting scheme	**0.94**
Shared Weights	1.10
Unweighted	4.06

Whole GEM with	Mean Absolute Error
Shared Weights	**0.62**
Current Weighting scheme	0.65
Unweighted	1.51

Reference: *STATEC, Labour Force Survey (LFS).*

Table 15: Bootstrap results (%)

Original TEA mean estimate	9.99
Bootstrap Mean	9.80
Standard error	0.92
CI in percentile	]8.08;11.59[
CI normal distribution	]7.99;11.61[
Coefficient of Variation (CV)	9.41
Monte Carlo's Stability	-1.94

Figure 17: Distribution of TEA Bootstrap

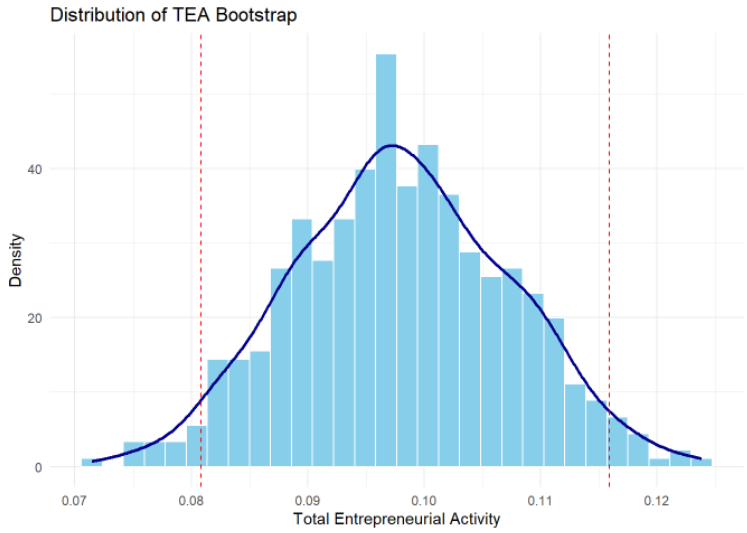
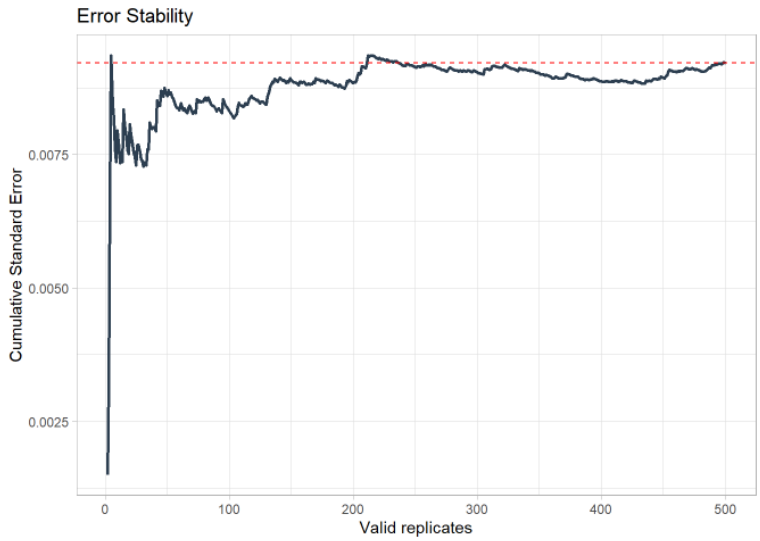


Figure 18: Error stability in the Bootstrap



FACULDADE DE **ECONOMIA**