STRATEGIES FOR RARE POPULATION DETECTION AND SAMPLING: A METHODOLOGICAL APPROACH IN LIGURIA

G. Lancia
Department of Mathematics
Università di Genova
Via Dodecaneso, 35. 16146, Genova.
giacomo.lancia@gmail.com

E. Riccomagno
Department of Mathematics
Università di Genova
Via Dodecaneso, 35. 16146, Genova.
riccomag@dima.unige.it

ABSTRACT

Economic policy sciences are constantly investigating the quality of well-being of broad sections of the population in order to describe the current interdependence between unequal living conditions, low levels of education and a lack of integration into society. Such studies are often carried out in the form of surveys, e.g. as part of the EU-SILC program. If the survey is designed at national or international level, the results of the study are often used as a reference by a broad range of public institutions. However, the sampling strategy per se may not capture enough information to provide an accurate representation of all population strata. Problems might arise from rare, or hard-to-sample, populations and the conclusion of the study may be compromised or unrealistic. We propose here a two-phase methodology to identify rare, poorly sampled populations and then resample the hard-to-sample strata. We focused our attention on 2019 EU-SILC section concerning the Italian region of Liguria. Methods based on dispersion indices or deep learning were used to detect rare populations. A multi-frame survey was proposed as the sampling design. The results showed that factors such as citizenship, material deprivation and large families are still fundamental characteristics that are difficult to capture.

1 Introduction

As our world becomes increasingly interconnected, the role of poverty and inequality in wealthy societies becomes ever more important. To capture and monitor the complex intricate evolution of social dynamics, many European countries participate in the European Union Statistics on Income and Living Conditions (EU-SILC) program (Eurostat, 2020). The task of this instrument is to examine income and living conditions across the European continent regularly. The most important aspects of EU-SILC focus on the links between poverty, well-being and inequality.

About three decades ago, poverty was described as widespread as in the past when prosperity and technology were much less diffused (Atkinson, 1998). Despite the proliferation of new technologies, poverty and inequality persist in European society, and the anticipated positive effects of technological advancement on these socioeconomic issues have not materialized as expected (Longford, 2014) In opposition to this, the role of education in contrasting poverty and inequality has been largely proven so far(Hofmarcher, 2021; Raffo et al., 2007). Among affluent countries, low-level education has correlated with a gap in income with a relevant impact on children's educational outcomes (Ferguson et al., 2007).

When considering the EU-SILC data to monitor the current living conditions in Europe, it has recently been shown that the EU-SILC microdata could substantially differ in describing the income-related variables among the European countries (Trindade and Goedemé, 2020). Further problems regarding the quality of the EU-SILC samples were already reported for the German section (Hauser, 2008). Major problems might also arise when such an open-access source of data is utilized to promote local policies. On the other hand, given the exponential growth of new emergent methods to analyse ever larger datasets, in the last decades the EU-SILC instrument has been completely revised to provide a more reliable and consistent source of data (Wirth and Pforr, 2022).

We propose in this work a novel methodology to investigate the EU-SILC microdata with special attention on rare populations. The problem of finding hard-to-sample populations is connected with finding those sections of the EU-SILC subjects that are poorly represented within the sampling population. Speaking more in general, we refer to rare populations as those sampling sub-populations that are widely dispersed among the reference sampling population (Lohr, 2021). Thus, we addressed the problem under two complementary aspects: proposing a methodology to detect and sample rare population strata.

To illustrate the methodology we utilized the 2019 EU-SILC dataset concerning the Liguria region. We are motivated to focus on this specific sub-population of the EU-SILC dataset because the government of the Liguria region has been developing a rich plan of assistance services for the strata of residents frail at most, i.e., those that are more exposed to inequality and poverty factors. This work was therefore aimed at devising an investigation parallel to EU-SILC in order to detect and newly sample the frail strata that are not well-represented in the EU-SILC data.

Thus, the detection phase, which is the first phase of this methodology, was accomplished through different univariate dispersion indices-based analyses and Artificial Neural Network (ANN)-based methods. The high potentiality revealed by ANN in deeply elaborating the EU-SILC data was already taken into account (Veljkovic et al., 2023; Pisati et al., 2010). When designing the sampling strategy, which is the second phase of our methodology, we opted for the Multi Frame Survey (MFS) Lohr and Rao (2006); Lohr (2007, 2011). Given its robustness and the consequent reduction of the under coverage that could occur if only one sampling frame was used Lohr and Rao (2006), such a sampling strategy turned out to be the most appropriate for the problem under the exam.

The rest of this work is thus organized as follows: In section 2 we reported a concise description of the 2019 EU-SILC data we utilized throughout the paper. Section 3 was dedicated to describing the three methods we implemented to detect rare populations within the 2019 EU-SILC dataset. In addition, we also described the methodology based on the MFS approach. The results of our two-phase methodology are then reported in section 5. In particular, we ran a simulation study to investigate the effective benefits of using an MFS-based approach for reducing survey costs. Finally, we let a summary discussion in section 6.

2 Data

In this section, we present the datasets we used to conduct our analysis. We are interested in the 2019 EU-SILC data for the Ligurian area. The analysis of these data was combined with the *2019 tax return data*; a complete ensemble of the annual declared income and economic situation of most of the households in Liguria. This tax dataset has the scope to provide a good measure of comparison for investigating the critical aspects of the EU-SILC data. This analysis was commissioned by the Bureaux of Financial programing and Statistics of the Liguria Region.

2.1 2019 EU-SILC data

The EU-SILC is a European program with the scope of collecting and comparing cross-sectional and longitudinal multidimensional microdata on income, poverty, social exclusion and living conditions (Eurostat, 2020). The collection of data is scheduled annually and involves households selected from all countries in Europe. The acquisition of data occurs after submitting a questionnaire to the selected households; this questionnaire helps explore the living conditions and social exclusion in Europe.

We considered the data available for the year 2019; in particular, we brought our attention to the Italian scenario with a special emphasis on the Ligurian region. The sub-dataset of the 2019 EU-SILC for Liguria amounts to 707 units. The EU-SILC database presents a huge variety of information of different kinds. We selected from the questionnaire a group of questions concerning the living conditions and social exclusion of the population in Liguria. These are 36 variables and all are categorical; a full description of these variables is in Appendix A. In addition, an in-depth investigation of the sampling distribution of these variables helped detect those strata of the Ligurian population which are rare to track.

All variables presented a rate of missing values lower than 10%; KNN multiple imputation (Troyanskaya et al., 2001) was utilized to impute missing values. In combination with the Ligurian dataset, we considered the 2019 Italian EU-SILC dataset as well. Likewise the Ligurian case, we extracted the same set of 36 categorical variables. We collected information for a total amount of 51981 subjects. Missing values, found with an incidence lower than 10% per each variable, were treated with the KNN multiple imputation.

2.2 2019 tax declaration dataset

The 2019 tax collection dataset is a broad collection of data available from the tax declaration in Liguria. The majority of the information in this collection concern the income and the economic situation of most the households in Liguria, i.e., a total amount of 1.216.747 declarations; information available for 731.537 households. The Ligurian population is therefore represented through the following economic and demographic features, namely

- Sex
- Age
- Employment
- Income
- Number of children
- Number of dependent disabled children.

In Section 6 we compare the reliability of the 2019 EU-SILC dataset for the Liguria reality with this source of administrative data.

3   Detecting Rare Populations

In this section we propose some tools to identify poorly represented or unrepresented sub-populations of EU-SILC data. The problem of determining whether a sub-population is under- or un-represented in a given dataset is related to the issue of *representation bias* (Shahbazi et al., 2022). Briefly, representation bias might arise from an intrinsic preexisting problematic nature of the data under exam (Mehrabi et al., 2021); in many other cases, the source of data from which the occurrences are taken might be lacking in a few of the sub-populations of interest (Shahbazi et al., 2022). Another potential source of representation bias is given by the sampling scheme (Shahbazi et al., 2022).

In the past years, *data coverage* for multicategorical attributes has broadly been considered to quantify representation bias (Asudeh et al., 2019, 2021; Jin et al., 2020). However, such a metric does not apply to the identification of relevant, unrepresented population strata. Furthermore, the problem of outlining a definition of rare or hard-to-reach population groups should go beyond the broad statement of "minority groups not involved in survey participation" (Brackertz, 2007). The hardest open issue remains the fact that the detection of unrepresented minority groups of data does not represent always a well-defined rare target population.

Based on these considerations, we propose three methods to detect rare and unrepresented subpopulations within the EU-SILC data. The central idea is to look for the less informative minority groups in EU-SILC data by means of anomaly detection techniques (Chandola et al., 2009) and entropy-based indices.

3.1   Entropy-based indices

For entropy-based measures, we expected that the variables with the lowest entropy were likely to have a high prevalence in a few classes; i.e., where we could identify a rare population of interest. Accordingly, we examine one by one the categorical variables that are far from showing a uniform occurrence in their classes. In other words, we focus our attention on the categorical variables that are not informative in terms of entropy.

We recall here a few properties of Shannon entropy. Given the generic probability mass function P, defined over the sample space Ω, we define the *Shannon Entropy* (or simply entropy) as

$$H[P] = -\sum_{x \in \Omega} P(x) \log P(x). \tag{1}$$

By construction, $H[P] \geq 0$ and is a concave functional (Bishop and Nasrabadi, 2006). Values of $H[P]$ close (or equal to zero) represent the case of a discrete distribution with most of (or all) the probability mass over one class. In contrast, the maximum entropy is reached when all classes are equally probable; in this case, the entropy reaches the value $\log N$ with N the total number of categories of P. The entropy $H[P]$ is often proposed as a measure of the overall *informativeness* of a source of random events following the distribution P.

This peculiarity of the entropy functional makes it suitable for identifying all those categories of the sampling population that are poorly represented or rare. In particular, due to the property of being a direct measure of the informativeness of a source of events, entropy was proposed in the last decades as an automated selection method of rare sample Daneshpazhouh and Sami (2014); Bereziński et al. (2015).

However, using Shannon entropy alone can only help to identify a type of subjects rarely interviewed; i.e., outliers and interesting rare populations, both of which present an intrinsic difficulty to be interviewed. Simply put, entropy can help uncover either some subjects who cannot be included in a particular category of the sampling population or the rare strata of the population that we want to detect.

To overcome the issue that arises from utilizing the sole entropy, we have opted for an additional entropy-based measure such as the Kullback-Leibler Divergence (KLD) (we will also refer to it as *relative entropy*). Given two discrete distributions P and Q defined over the same space Ω, the relative entropy (of P with respect to Q) $D[P||Q]$ is

$$D[P||Q] = \sum_{x \in \Omega} P(x) \log \frac{P(x)}{Q(x)}. \tag{2}$$

Relative entropy is proposed in many statistical learning contexts to quantify the *distance between two distributions* (Bishop and Nasrabadi, 2006). For example, it is often employed as a loss function in many classification and regression tasks. Similar to the entropy, the relative entropy is non-negative and concave (Bishop and Nasrabadi, 2006). Values close to zero express a small discrepancy between the analysed distributions while diverging positive values show that both distributions are far from being similar or equal. In the context of univariate variable selection, we utilized relative entropy in combination with the Shannon entropy. Specifically, we utilized relative entropy to compare the similarity between the categorical variables extracted from two different EU-SILC 2019 datasets, such as the dataset of the Ligurian region and the dataset of Italy. To avoid redundancy, we removed the data of the Liguria dataset in the one of Italy. A high relative entropy of a variable, combined with a low Shannon entropy, suggests that the variable may represent a category that is rare or difficult to sample in Liguria. This peculiarity diverges from the situation represented in the rest of the country. In fact, a low Shannon entropy denotes that one or more categories are undercovered, while a high relative entropy suggests a divergence between Liguria and Italy (except Liguria). From the analysis of the relative entropy values, we could therefore detect the populations poorly represented in the Ligurian EU-SILC dataset and in the Italian scenario.

The KLD has the drawback of being asymmetric, i.e., $D[P||Q] \neq D[Q||P]$. The Jensen-Shannon divergence is symmetric version of the relative entropyJensen-Shannon Divergence (JSD) (Nielsen, 2021) and is defined as

$$JS[P||Q] = \frac{D[P||M] + D[Q||M]}{2}. \tag{3}$$

with $M = \frac{P + Q}{2}$. The Jensen-Shannon divergence is bounded by both sides. As a lower bound it has the null value that represents the case where both P and Q are equal. The upper bound is the value $\log_2 N$, with N the number of categories of both P and Q.

Another entropy-based dispersion measure is defined as $h = \sum_{j \in \Omega} P^2(x)$ and it holds $-\log h = H_2(X)$ with H_2 the *Rènvi entropy* of order 2 (a.k.a *collision entropy*). The Rènvi entropy (Rényi, 1961) for a generic order $\alpha \in \mathrm{R}^+$ is given as

$$H_\alpha = \lim_{\gamma \to \alpha} \frac{1}{1 - \gamma} \log \left(\sum_{x \in \Omega} P^\gamma(x) \right). \tag{4}$$

Note also H_1 coincides with the Shannon Entropy defined in (1). The sum of squares of P is also bounded; lower and upper bounds are respectively $1/N$ and 1; with N the number of categories of P. This makes h suitable to be employed as an index over the informativeness and dispersion of P.

To summarise, we considered three entropy-based indices to describe the information content and redundancy of each variable. Shannon entropy and collision entropy have to do with the dispersion of occurrences in the categories that make up the variable. Conversely, the JSD helps show a contrast in the informativeness of some variables between the Ligurian region and the rest of Italy. All these indices are bounded, and we normalized them to reach values in the range $[0, 1]$. For simplicity, we have considered complementary quantities, such as the *complementary Shannon entropy* $1 - H$ and the *complementary collision entropy* $1 - H_2$.

For each of the 36 observed variables in the Ligurian region, say P, and its equivalent in Italy (excluding Liguria), say Q, we consider the point $ent = (1 - H[P], JS[P, Q], 1 - H_2)$ in $[0, 1]^3$. A point close to $(1, 1, 1)$ will be associated with a poorly represented variable, a point close to $(0, 0, 0)$ the opposite.

The evaluation of the entropy-based indices themselves does not provide any information on whether a single variable has some poorly sampled categories or not. Therefore, we considered an unsupervised classification problem to propose an automatic procedure to better discriminate those variables that might have some poorly sampled categories. We opted for a *K-means* clustering with two clusters. Avoiding complex analyses, we were interested in finding a net

discrimination to distinguish the poorly represented categories from those that do not exhibit this problem. We opted to proceed in this manner for pragmatic reasons. First, this method combines the distances between different variables, expressed as the Euclidean distances between the three entropy-based indices. Secondly, the interpretability of the results comes from the fact of using interpretable entropy-based indices. Simply put, the cluster of variables whose values of ent are closer to $(1, 1, 1)$ represents that cluster of variables poorly represented in some of their categories.

To evaluate the goodness of the K-means clustering we relied on the *Silhuhette score* metric (Rousseeuw, 1987). We readapted the silhouette score for the scenario of searching for two clusters Γ_0 (i.e., rare or poorly represented variables) and Γ_1 (i.e., variables that do not suffer from undercoverage). Let us consider the generic i-th variable among the 36 aforementioned. As introduced, the point $ent_i = (1 - H[P_i], JS[P_i, Q_i], 1 - H_2[P_i])$ describe the informativeness of the i-th variable, with P_i and Q_i the distributions of Liguria and Italy, respectively. To illustrate the method, we shall assume that e_i is associated with the cluster Γ_0. Thus, we define the *cohesion coefficient*, namely

$$\Xi_i = \frac{\sum_{ent_k \in \Gamma_0, ent_k \neq ent_i} d(ent_k, ent_i)}{|\Gamma_0| - 1},$$

with $d(ent_k, ent_i)$ denote the distance between the instances ent_k and ent_i (e.g., the Euclidean distance); $|\Gamma_0|$ denotes the size of Γ_0, i.e., the total number of items it comprises. Also, we take into account the *separability coefficient* of the i-th variable, namely

$$\Omega_i = \frac{\sum_{ent_k \in \Gamma_j, ent_k \neq ent_i} d(ent_k, ent_i)}{|\Gamma_1|}.$$

Combining the two coefficients, we get the Silhouette Coefficient for the i-th variable, namely

$$\Sigma_i = \frac{\Xi_i - \Omega_i}{\max(\Xi_i, \Omega_i)}. \tag{5}$$

In essence, one considers the discrepancy between the mean distance of the i-th instance concerning all other instances labelled in the same cluster (i.e., cohesion coefficient) and the mean distance between the i-th instance with respect to the elements of its closest cluster (i.e., the separability coefficient).

The values attained by Σ_i lay on the range [-1, 1]; values close to unity mean a net separation between the clusters; the opposite is represented by negative values. Values about 0 mean an overlapping of different clusters. To give an overall measure we utilized the jackknife cross-validation (Efron and Stein, 1981) to estimate the mean silhouette score and its 95% confidence interval.

3.2 Higher-order central moments-based indices

We propose here another univariate analysis to detect variables with some misrepresentation in the categories of the probability mass function. More specifically, we adapt the central moments of the distributions of these variables to provide some indices denoting the presence of an equal amount of occurrences among all the categories in question. Given the probability mass function P over the sample space Ω, the k-th central moment μ_k of P is defined as

$$\mu_k = \sum_{x \in \Omega} P(x)(x - \mathbb{E}(x))^k. \tag{6}$$

It holds $\mu_0 = 1$ and $\mu_1 = 0$ for all P; this implies that higher-order moments (say, $k \geq 2$) need to be taken into account. Moreover, the central moments of orders 2, 3, 4, 5, and 6 have a direct connection with some tendency of the distribution P, such as the deviation from the mean, the skewness about the mean value, the skewness and the fluctuation of outliers and extreme values about the mean value (Kamath and Liu, 2021). Using central moments for anomaly detection has been suggested recently (Renganathan et al., 2021), considering their property describes how data are spread out. The use of central moments has recently been proposed in the context of anomaly detection (Renganathan et al., 2021); the formulation of such a methodology was focused on time-discrete linear systems that undergo noise not necessarily Gaussian and possibly heavy-tailed.

Thus, starting from (6), we can construct the desired indices from the ratio

$$\frac{\tilde{\mu_k}}{\sum_{x \in \Omega} P^2(x) \sum_{x \in \Omega} (x - \mathbb{E}(x))^{2k}}; \tag{7}$$

with $\tilde{\mu_k}$ the unbiased estimator of the k-th order central moment. The heuristic of this approach is to consider (6) as the scalar product; then we consider the cosine similarity between P and $(x - \mathbb{E}(x))$ to produce the desired index. For odd orders, we took (7) as it is, since values close to unity indicate a skewness or an asymmetry of tails about the central

value. In contrast, for even orders, values close to unity tend to represent a well-spread mass of probability. Hence, we can define the index of k-th order as

$$\gamma_k = \begin{cases} 1 - \frac{\tilde{\mu}_k}{\sum_{x \in \Omega} P^2(x) \sum_{x \in \Omega}(x - \mathbb{E}(x))^{2k}} & \text{if } k = 2, 4, 6; \\ \frac{|\tilde{\mu}_k|}{\sum_{x \in \Omega} P^2(x) \sum_{x \in \Omega}(x - \mathbb{E}(x))^{2k}} & \text{if } k = 3, 5. \end{cases} \tag{8}$$

To identify which variables show some misrepresentation in their categories we reutilize the same approach proposed in section 3.1. That is, we consider an unsupervised classification problem to propose an automatic procedure based on the central moment-based indices to distinguish variables that suffer from undercoverage from the others. Hence, we again opted for a *K-means* clustering with 2 clusters; the quality of the obtained clustering was evaluated through the silhouette score.

3.3 Auto-Encoder Reconstruction Error-based index

So far we have considered univariate approaches. Both methods of sections 3.1 and 3.2 are based on the univariate extraction of features concerning the dispersion of a univariate discrete probability distribution. In contrast to this kind of approach, ANN have been proposed in recent years as a successful strategy to address anomaly detection in a multivariate setting (An and Cho, 2015; Zhou and Paffenroth, 2017). In particular, we focused on Variational Auto-Encoder (VAE); a type of Auto-Encoder (AE) broadly employed in anomaly detection problems. Generally speaking, the idea behind the AE is to learn simultaneously two distinct transformations, an *encoder* $\mathcal{E}$ and a *decoder* $\mathcal{D}$. The encoder has the scope to give the input data an efficient representation in a low-dimensional space; while the decoder attempts to reconstruct the input data from the encoding provided by the encoder. Usually, the decoder is placed just underneath the decoder and returns the output reconstruction. In other words, if we denote with $\mathcal{X}$ and $\mathcal{Z}$ the space of the input and encoded data, we can define the encoder as $\mathcal{X} : \mathcal{E} \to \mathcal{Z}$ and the decoder as $\mathcal{Z} : \mathcal{D} \to \mathcal{X}$. Accordingly, we can (informally) sketch it as $\mathcal{X} : \mathcal{E} \to \mathcal{Z} : \mathcal{D} \to \mathcal{X}$

In many applications for anomaly detection, AE are usually utilized to learn all the characteristics of one determined class of data. Unlike standard AE algorithms, VAE has the peculiarity of constraining the structure of the hidden units with a specific probabilistic structure. Usually, each output of the encoder is drawn from a Normal distribution whose parameters are learned from the input data; see figure 1. Such a structure leads to better generalization during the learning phase; the input data are not mapped into a single point but in a stochastic region of a reduced dimensionality space. As a result, small changes in the encoder output cannot affect largely the reconstruction through the decoder. Since this algorithm is devised to learn the interconnections among the features of training data, we expect that the propagation of a similar instance through this model will be encoded and then reconstructed with a very minimal loss. Conversely, when propagating an instance with different features from the ones utilized in the learning phase, the model will no longer be able to reconstruct the input data efficiently. When the last scenario occurs, one identifies that instance as an outlier or an anomalous instance; more in general, as a rare item that significantly differs from the majority of the training data.

Thus, we implemented a VAE to detect the set of categorical variables showing some misrepresentation in their categories. This model was implemented with the following encoder structure:

1. The input data are propagated through a single dense layer with 32 outputs *softplus* activation function (i.e., $f(x) = \log{(1 + \exp{x})}$.)

2. The 32-dimension description of input data is then propagated through two distinct dense layers with 16 output units Rational Activation Function (Boullé et al., 2020) (i.e., a parametric rational function whose parameters are learnt during the training phase). The first of the two dense layers provides the mean, while the second one the logarithm of the standard deviation, of the normally sampled outputs of the decoder.

3. Once each couple of parameters is evaluated, the output of each node of the encoder is drawn from the corresponding normal distribution.

For the decoder, we used this architecture:

1. The encoder's output is sent to the dense layers with 32 output nodes and a softplus activation function.

2. A further dense layer with a number of output nodes equal to the features of the input data is then placed underneath. The activation function is still a softplus function. Implementing such a number of outputs is necessary to ensure the input data reconstruction.
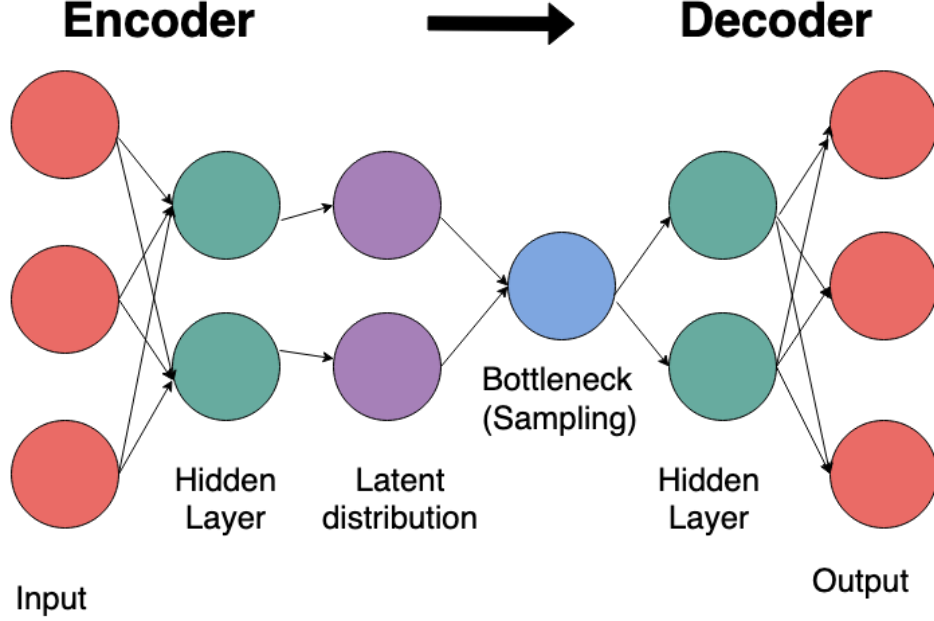
Figure 1: Schematic illustration of a VAE

The reconstruction was accomplished after minimizing the *mean square error* loss function; that is, denoted with $X_1, \ldots X_N$ and $\tilde{X}_1 \ldots \tilde{X}_N$ the input data and their reconstruction, respectively, we have

$$\Lambda = \frac{1}{N} \sum_{i=1}^{N} (X_i - \tilde{X}_i)^2. \tag{9}$$

To optimise the loss function use the *ADAM* algorithm (Kingma and Ba, 2014).

The validation of this model was mainly focused on the ability of the VAE to provide a good reconstruction of input data. To do so, we exploited the fact that the input data are categorical variables. Thus, for the i-th instance, we constructed the following metrics

$$\tau_0(X_i, \tilde{X}_i) = \begin{cases} 1 \text{ if } X_i - \lfloor \tilde{X}_i \rfloor = 0 \\ 0 \text{ if } X_i - \lfloor \tilde{X}_i \rfloor \neq 0 \end{cases} \tag{10}$$

$$\tau_1(X, \tilde{X}) = \frac{1}{N} \sum_{i=1}^{N} \tau_0(X_i, \tilde{X}_i). \tag{11}$$

where $\lfloor a \rfloor$ is the largest integer number smaller than a. We then used the *binomial test* to control whether at least 90% of the instances were perfectly reconstructed. Thus, we claimed the null hypothesis $H_0 : \tau_1 = 0.9$ with the alternative hypothesis $H_1 : \tau_1 < 0.9$. This approach was combined with a K-fold cross-validation; thus, we repeated the binomial test for each one of the K folds.

During the learning phase, we utilized $K - 1$ of the K folds, while the leftover was used to test the reconstruction ability of the model. To be sure that each of the validation folds does not significantly differ (in statistical terms) from the training ones, we evaluated the KLD between each couple of folds. Taking the total number of folds we reported the mean and the 95% confidence interval. The estimation of these quantities was accomplished through the jackknife resampling method.

To investigate which categorical variables were the most misrepresented through the VAE, we took inspiration from the approach described in Torabi et al. (2023). That is, we set a vectorised threshold based on the reconstruction error with the scope of pointing out anomalous instances.

More specifically, after training the model, the validation set was investigated to search for anomalous items. Thus, we propagated the training items through the model and for each one of them, we evaluated the *absolute reconstruction*

Method	Duration (sec.)
Entropy-based indexes	1.01
Central Moment-based indexes	1.04
VAE Reconstruction Error-based index	40.33

Table 1: Execution times of all methods when applied to the Iris Dataset

error per each feature. Given the generic i-th item, the absolute reconstruction error is

$$T_0(X_i, \tilde{X}_i) = \left(|X_i^{(1)} - \tilde{X}_i^{(1)}|, \ldots, |X_i^{(N)} - \tilde{X}_i^{(N)}| \right). \tag{12}$$

Note that T_0 is a vector with dimension equal to the dimension of item X_i. With regard to the data in our possession, the generic instance X_i is indeed an array of length 36 (since the aforementioned variables are 36 in total). By construction, $\tilde{X}_i$ is also an array with length 36. It follows that T_0 has length 36 as well. We denoted with $T_0^j(X_i, \tilde{X}_i)$ the absolute reconstruction error relative to the j-th feature of the i-th item. Next, we constructed the desired vectorized threshold; the generic j-th component was given by

$$\theta^{(j)} = \max_{x \in \{X_1, \ldots X_N\}} \{T_0^{(j)}(x, \tilde{x})\}. \tag{13}$$

Thus, when propagating one item of the validation set Y throughout the model, if the absolute reconstruction $T_0(Y, \tilde{Y})$ met $T_0^{(j)}(Y, \tilde{Y}) > \theta^{(j)}$ for any j-th component, the item Y was then flagged as anomalous. This procedure allowed us to detect which items needed more attention.

Finally, to identify which feature is more responsible for the anomalous feature of these outliers, we resorted to the *Feature Importance* algorithm (Breiman, 2001). The basic idea of this method is to break the correlations among the input variables to observe the resilience of the model in being able to make good predictions. During the validation phase, the variables of the validation data are shuffled, one at a time. The data with one shuffled variable are then propagated through the model; the importance of that shuffle feature, that is the prevalence of the anomalous items, is finally collected. The reshuffle of each variable was performed 30 times. The mean importance of each variable and the corresponding 95% Confidence interval was calculated employing the jackknife resampling method.

3.4 An illustrative example with the Iris Dataset

The indices above are tested on the Iris Dataset (Fisher, 1936), which consists of 150 instances of three types of irises (i.e., Setosa, Versicolour, and Virginica). For each instance, four features are reported: sepal length, petal length, sepal width, and petal width. The Setosa type differs from the others because of its short and tight petals. Based on this fact, we made a stratified sampling with a special focus on the Setosa type. Regarding the data sampling, we intentionally introduced an imbalance, particularly in sampling petal features. More specifically, we sampled a total of 80 samples; 40 samples from the Setosa population and 20 from each one of the remaining populations. A visualization of the pair plot of the sampled population is shown in Figure 2.

A summary with the results for all three methods is shown in Table 2. All methods showed that the sampled data revealed a shortage of information with the feature *Petal width*. In particular, the Entropy-based index and the VAE revealed a lack in both the petal characteristics, as expected. In opposition to this, the Moment-based index method indicated the *Sepal length* as a feature poorly represented together with the *Petal length* and *Petal width*. We observed that this group is characterized by high-scored indices γ_2, γ_4, γ_6, and low-scored indices γ_3, γ_5 (see Table 2), denoting the presence of an important sharp peak. This can be explained through a comparison with Figure 2; the method detected the shaped highly-tailed mass of Setosa items across all the three features under the exam. For completeness, we reported in Table 1 the execution times for all three methods. As expected, the VAE-based approach requires a larger amount of time compared to the others. This is due to the large number of parameters of VAE that need to be optimized.

Variable	Entropy Cluster	Central Moment Cluster	AE Rec Cluster	$1 - H_1$	JS	$1 - H_2$	γ_2	γ_3	γ_4	γ_5	γ_6	AE Rec Impor tance
Sepal length	Common	Rare	Common	0.11	0.02	0.21	0.97	0.01	0.99	0.01	0.99	0.28
Sepal width	Common	Rare	Common	0.07	0.02	0.12	0.97	-0.01	0.99	-0.01	0.99	0.19
Petal length	Rare	Common	Rare	0.29	0.04	0.43	0.66	0.05	0.84	0.06	0.91	0.48
Petal width	Rare	Rare	Rare	0.23	0.03	0.37	0.97	0.01	0.99	0.01	0.99	0.37

Table 2: Summary Table with the indices and results for the toy example. Results for all three methods were reported (Entropy-based indices, Central moment-based indices, and VAE reconstruction error-based indices). The first column shows the name of each categorical variable. Three columns show whether a variable was considered *rare* (i.e., there might exist some rare samples in it) or *common* (i.e., the opposite). The leftovers show the value of the indices of the three methods.

4   Sampling Rare Populations

We shall present here in detail the sampling strategy we developed. To illustrate our work, we shall first give a motivation for the benefits of using the MFS in sampling rare populations. After that, we will focus on applying this sampling strategy through a simulation study.

The Multi Frame Scheme (MFS) is a sampling technique pioneered by Hartley and Rao (1962). Its basic scope consists of aggregating, without duplicates, various incomplete sampling frames; each one containing a high prevalence of the rare population of interest, in order to obtain a larger sampling framework including all the rare populations we would like to investigate. When formulating an estimator of the expected value of the response variable Y, the following requirements should be met along the complete sampling framework (Lohr, 2011)

- The estimator must be unbiased.
- The estimator must be *internally consistent*. That is, if we consider only two incomplete frames A and B, the relation $Y = Y_a + Y_{ab} + Y_b$ must be met. We denoted with Y_a, Y_{ab}, Y_b represent the number of subjects, without considering intersections, in A, B, and $A \cap B$. By omitting duplicates, the variable of response Y must be conserved along all the incomplete frames.
- The estimator should be efficient, with low variance
- The estimator should be as robust as possible to control possible non-sampling errors from multiple frame surveys.

For the two-frame case, the method proposed by Hartley and Rao (1962) was based on the reweighting of the repose variable Skinner and Rao (1996). The optimal choice of the reweighted parameters was exploited to ask the efficiency of the estimator. Such an approach, denoted as *optimum* approach, is based on the optimization of the estimator variance and has been largely studied to obtain an estimator in the two-frame case Hartley and Rao (1962); Fuller and Burmeister (1972); Bankier (1986). A theoretical and general solution in the case involving an arbitrary number of frames (larger than 2) has already been proposed by Lohr and Rao (2006). Although the optimal approach gives the estimator some desired theoretical properties, it might often lead to an unstable solution when involving large sampling frameworks. In such a case, the method proposed by Mecatti et al. (2007) overcomes the issue brought by the optimal approach. As in the optimization approach, this method accomplishes a reweighing of the response variable; in this case, the weights are fixed and specific for each subject. Thus, the response variable of each subject is multiplied by the inverse of its multiplicity along the incomplete frames, i.e., the inverse of the number of frames that the subject belongs to. In the work of Mecatti et al. (2007) an unbiased and consistent estimator for the population mean is given. However, the efficiency is not guaranteed. In addition, such an approach does not provide any indication about the sampling size of the survey.

To bypass this issue, we have proposed here a method based on a constrained minimization problem. The idea consists of finding the number of subjects per frame minimizing the variance estimator (i.e., making the estimator the most efficient one) given the constraint on the total number of subjects to be included in the survey. Since the optimal solutions cannot be expressed in a close form, we searched for them through a computational approach. Briefly, we took under consideration the unbiased MFS variance estimator, namely

$$\hat{v}_{mfs} = \sum_{q \in Q} \left(\frac{N_q}{n_q} - 1 \right) \frac{N_q}{n_q - 1} \left[N_q \sum_{i \in \mathcal{S}_q} u_i^2 - \frac{N_q}{n_q} \left(\sum_{i \in \mathcal{S}_q} u_i \right)^2 \right], \tag{14}$$

with N_q the total population of the q-th incomplete frame, n_q the sample size of the q-th incomplete frame, and m_i the multiplicity of the i-th subject. Thus, we aimed to minimize the Lagrangian

$$\mathcal{L}[\pi_1, \ldots, \pi_n, \lambda] = \hat{v}_{mfs} + \lambda \left(\sum_{j \in \mathcal{Q}} \pi_j \nu_j - n \right); \tag{15}$$
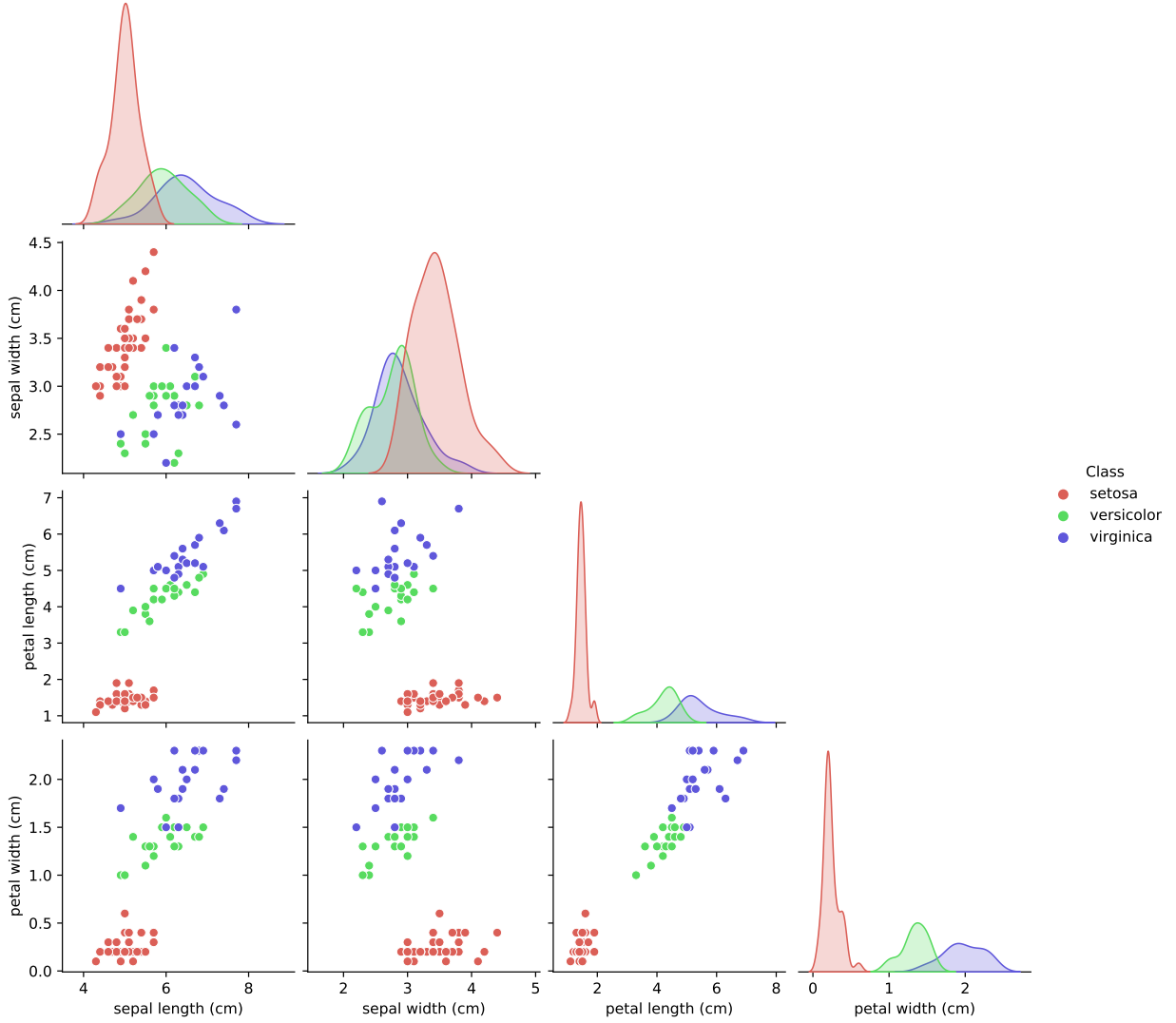
Figure 2: Pair plot for all the four features. Along the diagonal the Kernel Density Estimation, stratified per class, per each of the four features

with $\pi_1, \ldots, \pi_n$ the proportion of each population $N_1, \ldots, N_n$ to allocate in each corresponding incomplete frame of $\mathcal{Q} = \{\mathcal{Q}_1 \ldots \mathcal{Q}_n\}$; $\nu_j = \sum_{i=1}^{N_n} \frac{1}{m_i}$. λ is a the Lagrange multiplier.

As a response variable, we utilized the Equivalized Household Income (EHI). Such a variable is widely used in the EU-SILC context to describe the income situation of a household (Longford, 2014). The EHI can be evaluated by the formula

$$\text{EHI} = \frac{\omega}{1 + 0.5\psi + 0.3\chi}; \tag{16}$$

with ω the total income of the household, ψ the number of components of the household older than 18, and χ the number of children within the household older than 16 (and younger than 18). For more details see appendix B.

To test our sampling scheme, we utilized data randomly generated from the rare populations investigated through the methods shown in sections 3.1, 3.2, and 3.3. The data generation was accomplished through a Gaussian mixture model, with the optimal number of normal distributions sought after asking for the lowest Akaike information (Akaike, 1976).

We compared the results of our approach with a Simple Random Sampling (SRS)-based survey; that is we repeated the sampling procedure for the SRS approach and then we compared the results. We referenced the *Design Effect*, or simply Deff. Denoting with $\hat{v}_{mfs}$ and $\hat{v}_{srs}$, respectively, the estimator variance of both the MFS and SRS approaches;
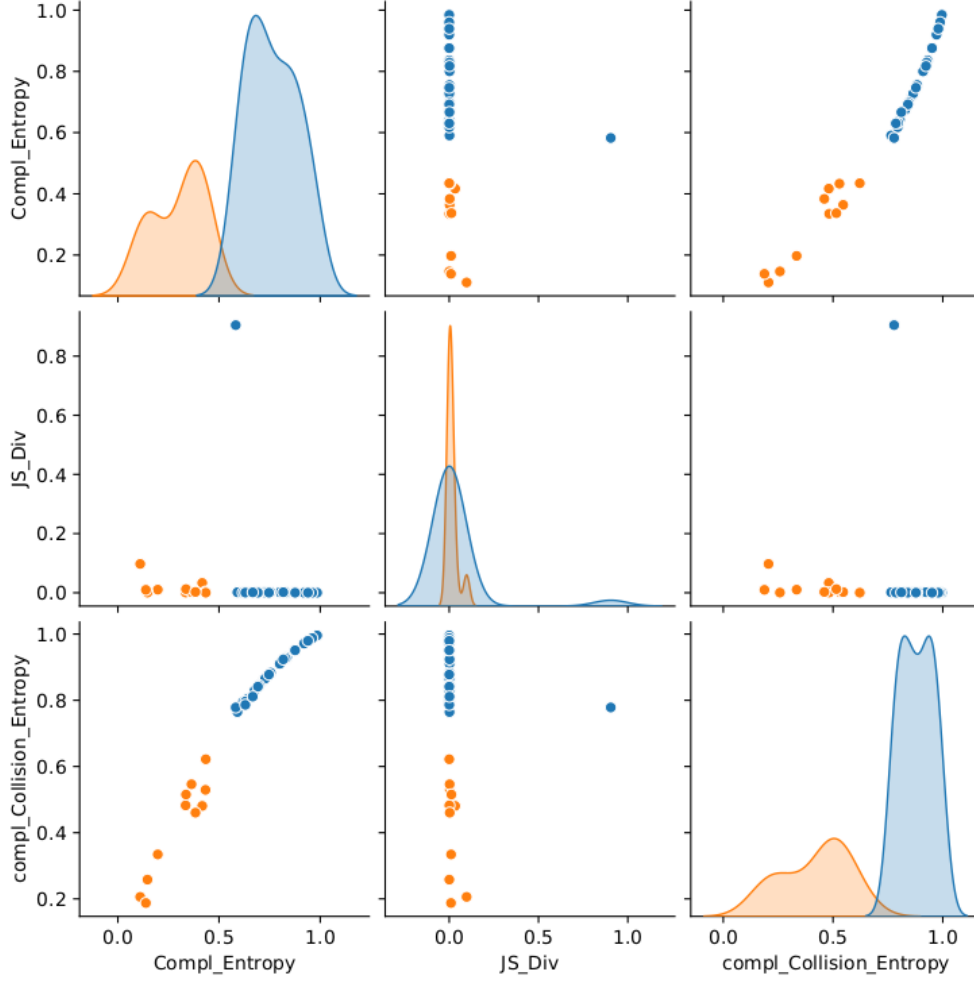
Figure 3: Pairwise plot of all entropy-based indices. Each plot shows also the two-cluster classification provided by the K-mean algorithm; in blue the *rare* items while in orange the others. The diagonal plots depict the Kernel Density Estimation for both classes.

the Deff is defined as

$$\Delta = \frac{\hat{v}_{mfs}}{\hat{v}_{srs}}.$$

The less the Deff, the more is the efficiency of MFS with respect to the SRS.

5 Results

In this section, we shall present all the results obtained in the two steps of the analysis. First, we shall show the rare populations that our methodology suggested. Next, we shall present a simulation of the MFS-based population sampling by using the EU-SILC 2019 data.

5.1 Phase one: detection of rare populations

A pairwise plot of all the entropy-based features is shown in figure 3. In that figure, the distinction between *rare* and *common* (i.e., well-represented) variables is also reported. For the sake of clarity, we meant here with the term *rare* a variable whose categories might not be well represented in the sampling population under the exam; the term *common* refers to the vice-versa. When validating the K-mean clustering model, we obtained an overall Silhouette coefficient of 0.66; the 95% confidence interval was equal to [0.62, 0.71].
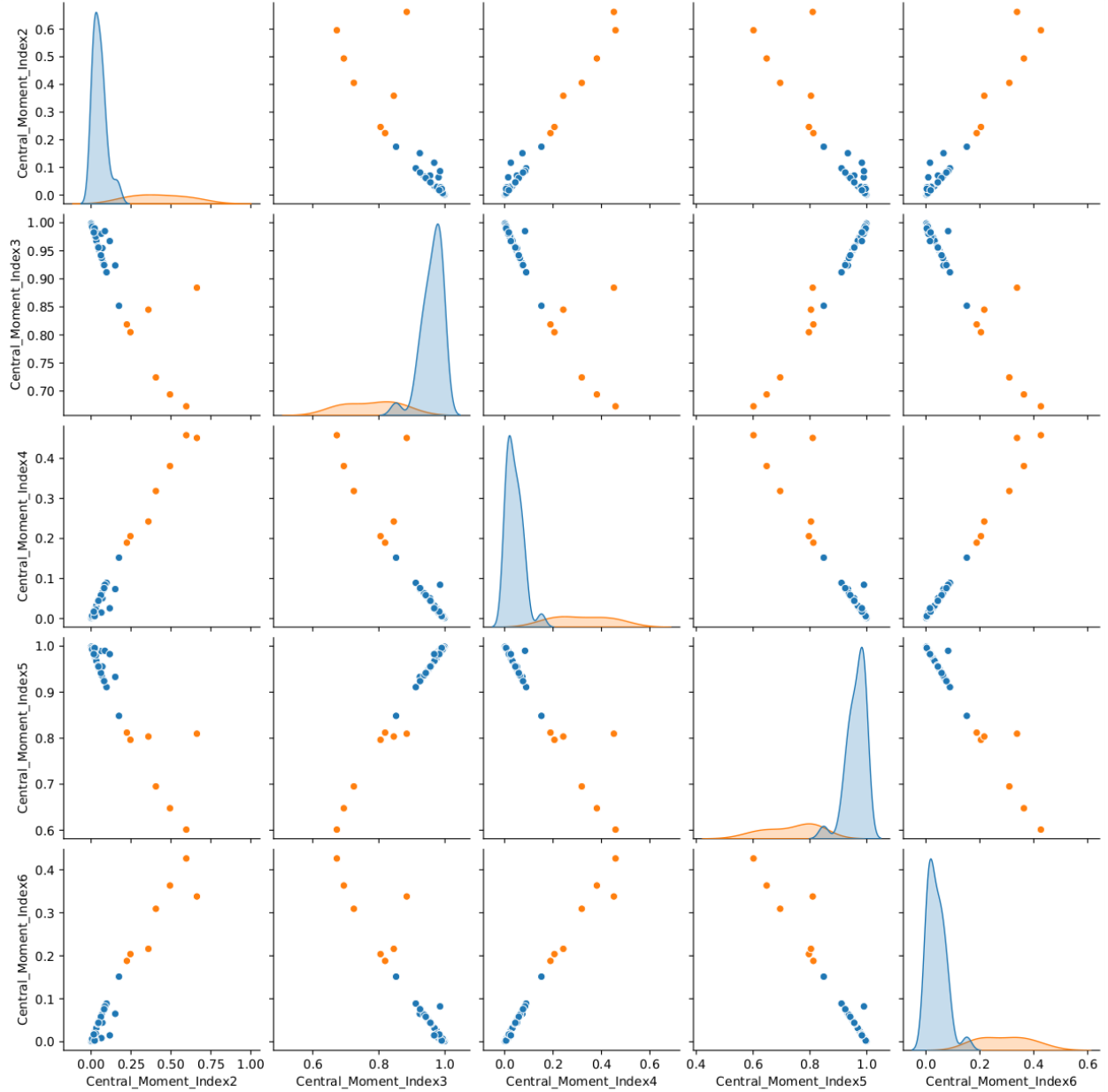
Figure 4: Pairwise plot of all central moment-based indices. Each plot shows also the two-cluster classification provided by the K-mean algorithm; in blue the *rare* items while in orange the others. The diagonal plots depict the Kernel Density Estimation for both classes.

When considering the central moment-based indices, we followed the same procedure as above. In figure 4, the pairwise plots of all indices are shown. The plots show also the clustering of variables in rare and common variables. The evaluation of the clustering guarantees an overall silhouette score of 0.52; the 95% confidence interval was equal to [0.41, 0.64].

The results for the VAE-based analysis are shown in figures 5 and 6. We examined plenty of different choices of folds, from two up to ten included, to validate the VAE. We observed that using a 10-fold validation reduced at best the risk of letting the model be trained and tested on different sets of data; see figure 5. The results from the binomial test, for each one of the ten training phases, are shown in table 3. The analysis of the variables' importance helped reveal the set of variables mostly misrepresented; we took all variables whose mean importance has an absolute error larger than 100% compared to the mean importance of the model when no shuffling is applied. We then identified variables

N. of fold	Statistic ($\hat{p}$)	p-value
1	0.96	0.99
2	0.97	1.0
3	0.98	0.99
4	0.98	0.99
5	0.98	0.99
6	0.97	0.99
7	0.97	0.99
8	0.98	0.99
9	0.97	0.99
10	0.97	0.99

Table 3: Results of the binomial test to validate the reconstruction ability of the VAE. The first column reports the number of folds which was used as a validation test. The statistic is the test statistic, i.e., the prevalence of correctly reconstructed items. The p-values are shown in the last column. We recall the null hypothesis is $\hat{p} = 0.9$, versus the alternative hypothesis $\hat{p} < 0.9$.

as *CITTADX* (whether the interviewed has Italian citizenship), *FERIE* (whether the interviewed can spend a week of vacation abroad or far from home), *ITALIA* (whether the intertwined was born in Italy), *RAGA015* and *RAGA117* (the number of children in the range of age, respectively, 0-15 and 1-17 who live in the household), *SEV_MAT_DEPRIV* (whether the interviewed is exposed to severe material deprivation); see figure 6.

In table 4, a summary of all the results produced for each one of the three methods is reported. In that table, we highlighted the populations indicated as the ones to include in the sampling stage. Hence, the variables we selected are *CITTADX*, *ITALIA*, *RAGA015* and *RAGA117*, and *SEV_MAT_DEPRIV*. The rare populations are obtained after looking at the less frequent categories of these variables; the histograms of these variables are shown in figure 7. In the end, the rare populations we took under consideration are

- Households who do not have the Italian citizenship
- Households whose members were not born in Italy
- Households with 3 or more children in the age range either [1, 15] or [1, 17].
- Households subjected to severe material deprivation.

Variable	Entropy Cluster	Central Moment Cluster	AE Rec Cluster	$1 - H_1$	JS	$1 - H_2$	γ_2	γ_3	γ_4	γ_5	γ_6	AE Rec Impor tance
ACCORD	Rare	Rare	Common	0.876	0.001	0.951	0.018	0.983	0.017	0.983	0.017	0.019
ARRETR	Rare	Rare	Common	0.722	0.001	0.861	0.053	0.950	0.050	0.950	0.050	0.013
BAGNI2_B	Common	Common	Common	0.417	0.033	0.481	0.596	0.673	0.458	0.601	0.426	0.018
BORST	Rare	Rare	Common	0.985	0.000	0.996	0.001	0.999	0.001	0.999	0.001	0.016
CAMBCOND	Rare	Rare	Common	0.836	0.000	0.931	0.025	0.975	0.025	0.975	0.025	0.014
CAMB_CASA	Rare	Rare	Common	0.829	0.001	0.927	0.027	0.974	0.026	0.974	0.026	0.016
CASSAINT	Rare	Rare	Common	0.591	0.001	0.764	0.097	0.912	0.089	0.911	0.089	0.014
CIBO	Rare	Rare	Common	0.728	0.001	0.866	0.051	0.951	0.049	0.951	0.049	0.021
CITTADX	Rare	Rare	Rare	0.800	0.002	0.910	0.033	0.968	0.032	0.968	0.032	0.123
CONTRIB	Rare	Rare	Common	0.920	0.000	0.971	0.010	0.990	0.010	0.990	0.010	0.019
CONTRPU	Rare	Rare	Common	0.874	0.002	0.950	0.018	0.982	0.018	0.982	0.018	0.017
CORREG	Rare	Rare	Common	0.960	0.000	0.988	0.004	0.996	0.004	0.996	0.004	0.016
CRIME	Rare	Rare	Common	0.642	0.000	0.805	0.078	0.928	0.073	0.927	0.073	0.020
DIFCIB	Rare	Rare	Common	0.818	0.002	0.923	0.030	0.975	0.022	0.980	0.019	0.024
DIFFRED	Common	Rare	Common	0.335	0.001	0.483	0.151	0.924	0.074	0.933	0.065	0.024
DMEZZO	Rare	Rare	Common	0.755	0.001	0.883	0.044	0.958	0.042	0.958	0.042	0.019
DOM	Common	Common	Common	0.110	0.097	0.206	0.662	0.884	0.451	0.810	0.338	0.018
DON1555	Common	Rare	Common	0.433	0.003	0.529	0.064	0.980	0.015	0.989	0.009	0.021
FERIE	Common	Common	Rare	0.197	0.010	0.334	0.406	0.724	0.319	0.695	0.309	0.037
FTPT	Common	Common	Common	0.364	0.002	0.546	0.224	0.819	0.189	0.812	0.188	0.019
IMPREV	Common	Common	Common	0.146	0.000	0.258	0.494	0.694	0.381	0.648	0.363	0.026
ITALIA	Rare	Rare	Rare	0.698	0.002	0.845	0.060	0.943	0.057	0.943	0.057	0.110
LIQUID	Rare	Rare	Common	0.675	0.001	0.829	0.067	0.937	0.063	0.937	0.063	0.018
MALAT	Rare	Rare	Common	0.618	0.000	0.795	0.086	0.985	0.084	0.990	0.082	0.016
NOCERCO_I	Rare	Rare	Common	0.582	0.905	0.778	0.071	0.955	0.051	0.956	0.045	0.020
PAGAFF	Common	Common	Common	0.337	0.012	0.515	0.246	0.805	0.206	0.796	0.204	0.017
PAGAPP	Common	Common	Common	0.139	0.010	0.188	0.359	0.845	0.242	0.804	0.216	0.023
PAGBOL	Rare	Rare	Common	0.692	0.000	0.841	0.062	0.942	0.058	0.942	0.058	0.022
PASTO	Rare	Rare	Common	0.631	0.000	0.797	0.082	0.925	0.076	0.924	0.076	0.015
PESTERO	Rare	Rare	Common	0.939	0.000	0.980	0.007	0.993	0.007	0.993	0.007	0.016
RAGA017	Rare	Rare	Rare	0.630	0.000	0.786	0.028	0.987	0.008	0.995	0.004	0.203
RAGA115	Rare	Rare	Rare	0.667	0.001	0.811	0.023	0.990	0.006	0.996	0.003	0.203
RICMAN	Rare	Rare	Common	0.876	0.000	0.951	0.018	0.983	0.017	0.983	0.017	0.016
SEV_MAT_DEPRIV	Rare	Rare	Rare	0.747	0.000	0.878	0.046	0.956	0.044	0.956	0.044	0.031
SMOG	Common	Rare	Common	0.435	0.000	0.622	0.175	0.852	0.152	0.849	0.152	0.015

| SPESCUOL | Common | Rare | Common | 0.384 | 0.003 | 0.460 | 0.117 | 0.967 | 0.026 | 0.983 | 0.015 | 0.018 |

Table 4: Summary Table with the results for all three methods (Entropy-based indices, Central moment-based indices, and VAE reconstruction error-based indices). The first column shows the name of each categorical variable. Three columns report the detection of the *rare* populations (according to the three methods). The leftovers show the value of the indices used in the three methods. The highlighted rows denote a rare population for all three methods.

5.2 Phase two: detecting rare populations

The analysis of section 5.1 brought to light five different rare sub-populations within the 2019 EU-SILC dataset. We designed a simulation study to analyse the benefits of opting for a MFS to investigate these sub-populations. We generated the characteristics of 1500000 subjects through a Gaussian Mixture model. The Gaussian Mixture model was fitted on the 2019 EU-SILC data. According to the values of these characteristics, a single subject might be allocated to none, one or more than one rare sub-population. In case none of the rare sub-populations is represented, the subjects are said to belong to the *ground population*. Note that the choice of generating 1500000 items was determined by aiming to accurately represent the Ligurian population during the year 2019 (ISTAT, 2019). In figure 8 the Akaike Information Criterion (AIC) as a function of the number of normal distributions involved in the Gaussian Mixture model is shown; a single-Gaussian model proved most efficient for generating the synthetic population.

In figure 9 the values of the Design effect are shown for different orders of magnitude of the sampling population. The MFS approach became more accurate as larger populations were included. With a sampling size of 100 (or 1000), the MFS can be about 10^5 (or 10^7) times more efficient than a single (complete) frame-based survey with SRS sampling scheme.

When focusing on the sample size per each incomplete frame, we can observe a clear difference in how both sampling methods include subjects in each subframe. For large sample sizes ($>> 10^3$), the MFS tend to assign subjects keeping almost fixed some proportions between the sub-populations; see figure 10. Lower sample sizes are characterised by a similar proportion through all the incomplete frames, i.e. almost 10^{-2}. In opposition to this, the SRS turned out to be more focused on sampling the rare sub-population giving much less importance to the ground population; see figure 11.

5.3 Comparision with 2019 tax declaration data

To provide further validation of our methodology, we compared the rare subpopulations identified using the aforementioned methods with the subpopulations of the 2019 tax declaration dataset. (see 2.2). The goal was to illustrate that the undercoverage of hard-to-reach subpopulations revealed by our methodology was also present when inspecting a comprehensive administrative data source. Particularly, we put the accent on searching for statistical differences in the sub-populations concerning the number of children (age less than 17) in households. Note that such a subpopulation is the only one in which faithful data are available from both datasets (Ligurian section of 2019 EU-SILC dataset and the 2019 tax declaration dataset).

To reveal significant differences, we utilized the *2-sample Kolomogorov-Smirnov* test with a 5% significance level. Denoted with $\mathcal{F}_1$ and $\mathcal{F}_2$ the distribution of the number of children in the households for the two datasets, we tested the null hypothesis $H_0 : \mathcal{F}_1 = \mathcal{F}_2$ versus the alternative $H_1 : \mathcal{F}_1 \neq \mathcal{F}_2$. The test's result showed that the null hypothesis had to be rejected (p-value of $7 \cdot 10^{-4}$). The graphical analysis showed that the EU-SILC data have scarcely represented all the households with a very large number of children (say, > 5). As a result, the 2019 EU-SILC data of region Liguria appeared to be more focused on households with less than 3 children; see figure 12

6 Discussion and conclusion

Our work was committed to inquiring about the representativeness of EU-SILC data, for the Ligurian region in Italy, proposing a methodology to reveal hard-to-find sub-populations within the EU-SILC dataset together with a sampling scheme to acquire *ab novo* the information concerning the investigated sub-populations.

A comparison between the entropy and the central moment-based indexes revealed that the former was robust compared to the central moment-based one. Both methods identified similar hard-to-sample variables, but the Silhouette score comparison and graphical inspection favoured the entropy measure-based approach for detecting potential misrepresentations among EU-SILC variables. Compared to the VAE-based approach, most variables were labelled as not rare, and the multivariate-based approach aligned with others only for a limited set of variables.

About the sampling design analysis, the MFS approach turned out to be more valid than SRS-based unique frame strategy. The ability of MFS to be more efficient than a standard SRS approach is due to a wiser selection of rare
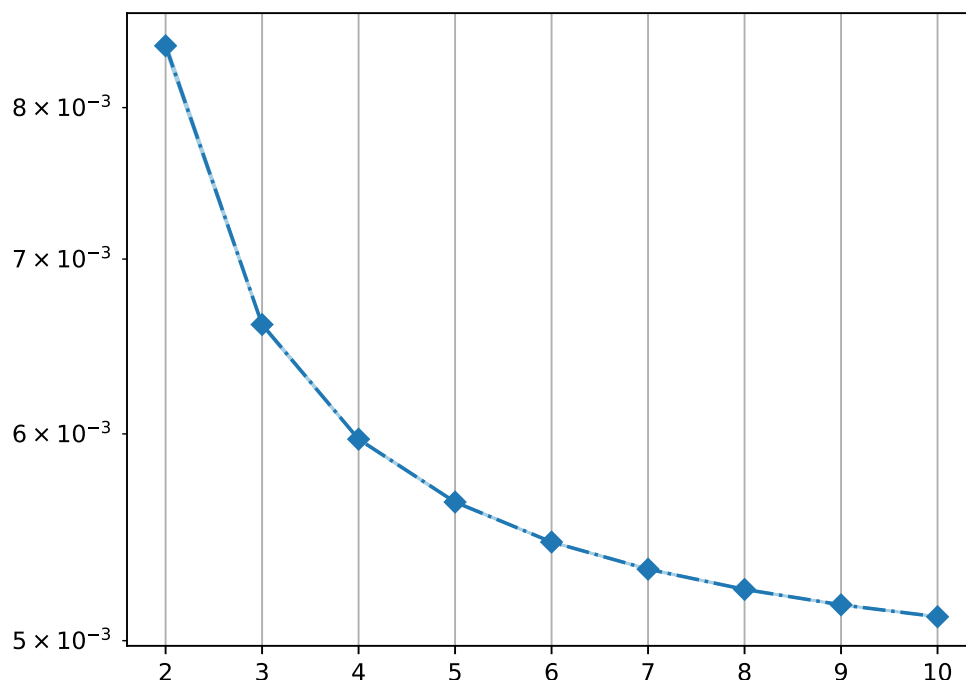
Figure 5: Mean value of KL divergence per each pair of data folds. Data were split using a K-fold approach. The number of folds is shown on the x-axis. When appreciable, the 95% confidence interval is represented by the light-coloured area.

subjects to interview. Consequently, the MFS suggests sampling a much lower number of rare subjects without a loss in the representability of the target population.

The comparison with administrative data, such as the tax-declaration dataset, revealed that the EU-SILC dataset does not represent adequately rare strata of the sampled population, such as the numerous families. Such a result sheds some light on the weak faithfulness of the EU-SILC data when considering such a source of data to promote some policies for specific hard-to-reach strata of the population of interest.

It is important to stress, however, that the first phase of our methodology (i.e., the research of rare classes) was conducted on the sole EU-SILC data and in the absence of any other similar source of administrative data that could represent ground truth. This posed a challenge to developing our methodology; however, the comparison with the tax-declaration dataset showed a proof of concept regarding its practical potential.

Limitations in this study are certainly present; however, we remember we have just limited our goal to present a methodology that could delve into the application of advanced studies. We tried to combine and compare different approaches, such as deep learning and traditional dispersion indices, to ameliorate the precision of rare population detection. Additionally, longitudinal studies may provide a deeper understanding of the dynamics and evolution of rare populations over time. In conclusion, this paper lays the groundwork for a nuanced (and rigorous) approach to detecting misrepresented rare populations within population datasets. We investigated the issue of rare populations within the Ligurian area, offering a valuable contribution to the broader field of population research. As we continue to refine and expand our methodologies, we move closer to a more comprehensive and inclusive understanding of the diverse peculiarities of rare samples that are hard to include in large surveys.

A   Supplementary information on data

Name Variable	Description
ACCORD	Whether the subject has a written employment contract, oral employment agreement, or none of both
ARRETR	Whether the subject has received arrears from employment

BAGNI2_B	Whether the dwelling has two or more toilets
BORST	Whether the sbject has received a scholarship
CAMBCOND	Whether the subject changed profession during the year of reference (2019)
CAMB_CASA	Whether the subject changed dweliing during the year of reference (2019)
CASSAINT	Whether the subject has enrolled in a reduncy program
CIBO	Whether the subject could not afford to get the food nedded during the year of reference (2019)
CITTADX	Whether the subject jas Italain citizenship
CONTRIB	Whether the subject has received public contributions in cash to cover all or part of the costs of housing
CONTRPU	Whether the subject has received public contributions in cash to cover all or part of the rental costs
CORREG	Whether the subject is actually enrolled in a training course organized and/or recognized by the Region, of duration equal to or greater than 6 months (or 600 hours) and which grants a professional qualification
CRIME	Whether the area where the subject currently lives presents problems of crime, violence, and vandalism
DIFCIB	Whether the subject resorted to the help of someone who gave food, clothing or other essential goods to the family during the year of reference (2019)
DIFFRED	Whether the income of the subject in 2019 was increased, decreased, or unchanged compared to the year 2018
DMEZZO	Whether the subject's employer provided him/her a car, even for personal reasons
DOM	The population of the town hall where the subject lives: less than 5000, between 5000 and 10000, between 10000 and 20000, between 20000 and 50000, more than 50000
DON1555	Number of women with age in range 15-55 in the subject's household.
FERIE	Whther the subject could afforf a vacation far from his/her dwelling
FTPT	Whether the subject has a full-time or part-time contract
IMPREV	Whetehr the subbject would be able to cope, with own resources, with unforeseen expenditure of approximately EUR 800
ITALIA	Whether the subject was born in Italy
LIQUID	whether the subject has received one or more severance payments
MALAT	Whether the subject expirienced times or periods when he/she did not have the money to pay for the expenses for diseases
NOCERCO_I	The reason (when applicable) why the subject did not find a job in the last 4 weeks
PAGAFF	Whether the subject expirienced times or periods when he/she did not have the money to pay renting costs
PAGAPP	Whether the subject has utilized some apps to pay bills or transfer money
PAGBOL	Whether the subject expirienced times or periods when he/she could not pay the bills (e.g., gas, eletricity, water consumtion, and so on)
PASTO	Whther the subject could eat, meat, fish, or an equivalent vegetrain dish, at least once per two days
PESTERO	Whther the subject has recived one or more pensions from abroad
RAGA017	The number of children with age in the range 0-17
RAGA115	The number of children with age in the range 1-15
RICMAN	Whether the subject has received regular maintenance or maintenance payments from the ex-spouse/ex-partner or the separated/divorced parent who is not a co-owner
SEV_MAT_DEPRIV	Whther the subject has been expirencing a severe material deprivation
SMOG	Whether the subject lives in an area presenting pollution, dirt or other environmental problems caused by traffic or industrial activities
SPESCUOL	Whether the subject experienced times or periods when he/she could not pay school fees, such as school books, school or university taxes

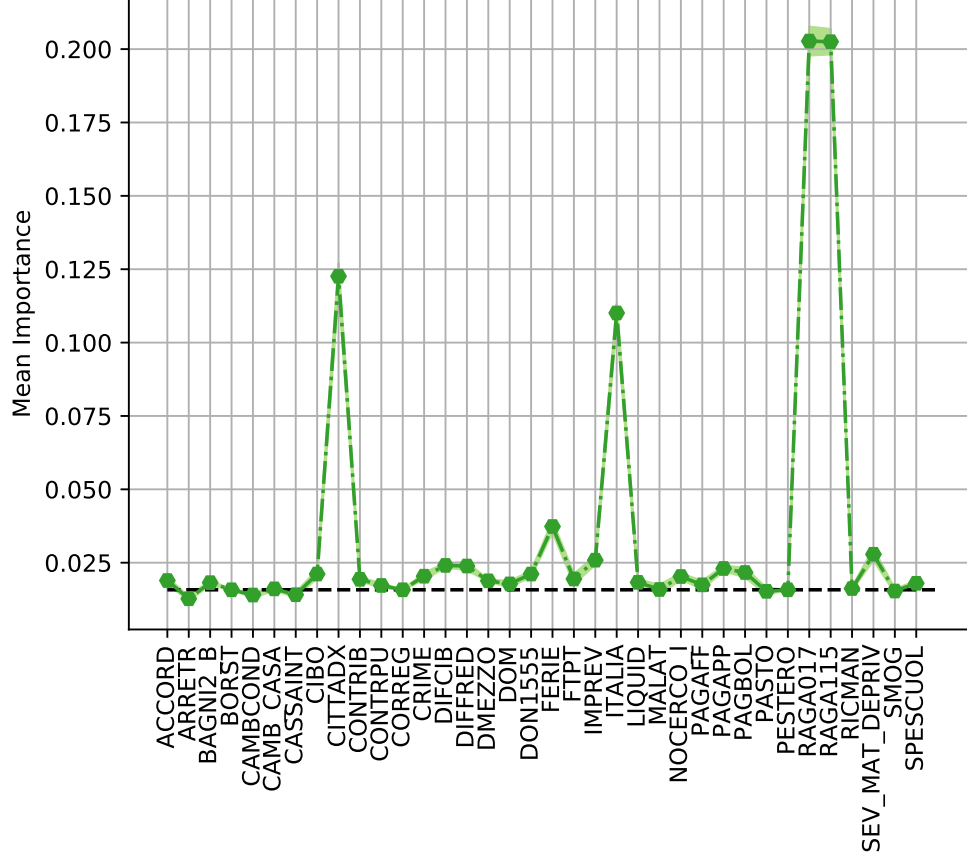Table 5: List with the description of the 36 categorical variables selected for the analysis.

Figure 6: Mean Importance per each variable. The light-coloured area denotes the 95% interval. The dashed black line denotes the mean importance when no shuffling is applied to any one of the variables.

B Mathematical details on the multi-frame approach

B.1 Population Evaluation

Let us consider a set of incomplete frames $\mathcal{R} = \{R_1 \dots R_n\}$ and a the population $\mathcal{P} = \bigcup_{k=1}^{n} R_k$. Since the frames are incomplete, each one of them contains subjects sharing the same unique characteristics. To evaluate the total population we can utilize the formula

$$t = \sum_{r \in \mathcal{R}} \sum_{k \in r} \frac{y_k}{m_k}; \tag{17}$$

with y_k a binary response variable and m_k its corresponding *multiplicity*. With the term multiplicity, we refer to the number of times that a sample is included in one or more of the incomplete frames $R_1 \dots R_n$. Note that the incomplete frames might be somehow overlapped. Consequently, the use of (17) to count the total items of $\mathcal{P}$ takes into account the eventuality that one subject might be included in one or more incomplete frames.

Dealing with the multiplicity can turn out to be unpractical. Thus, we can consider the *rescaled response variable* (i.e., $u_k = \frac{y_k}{m_k}$.). When evaluating the total population, we can still resort to (17). However, we can notice that we are now legitimised to consider the incomplete frames as independent of each other. For convenience, we shall denotes these independent sets with $Q_1 \dots Q_n$ with $\mathcal{Q} = \{Q_1, \dots, Q_n\}$. Thus, we have $\cup_{k=1}^{n} R_k = \cup_{k=1}^{n} Q_k$, with $Q_i \cap Q_j = \emptyset \; \forall i \neq j$. We rewrite the (17) as

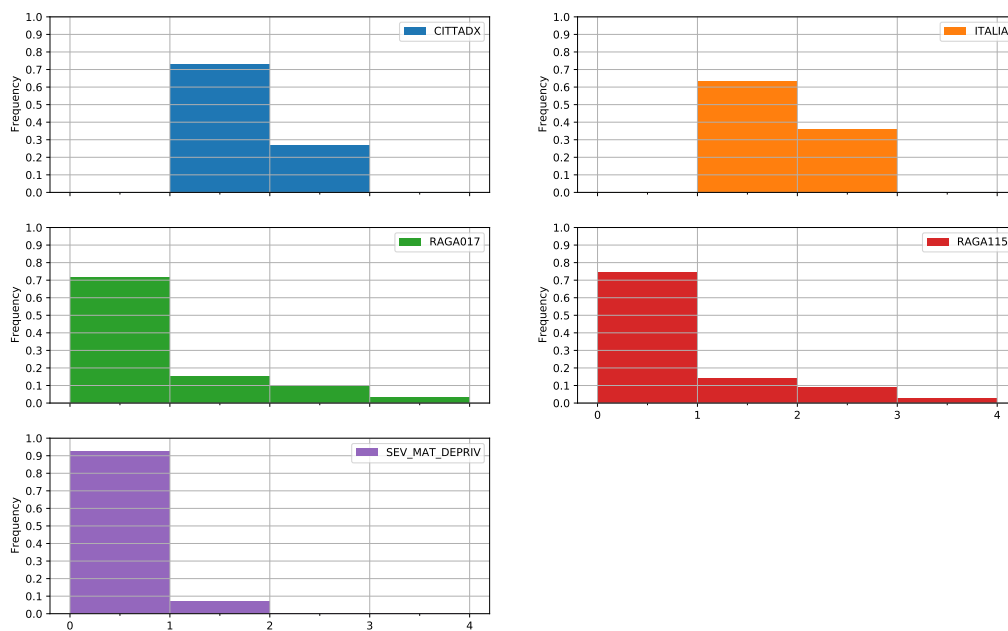$$t = \sum_{q \in \mathcal{Q}} \sum_{k \in q} u_k. \tag{18}$$

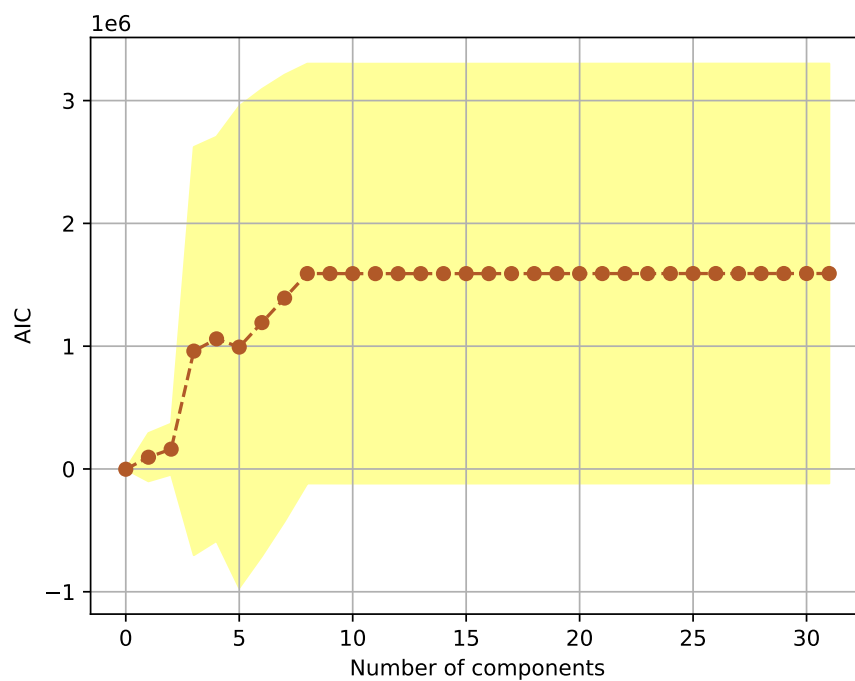Figure 7: Histograms for the rare categorical variables (as indicated in table 4)

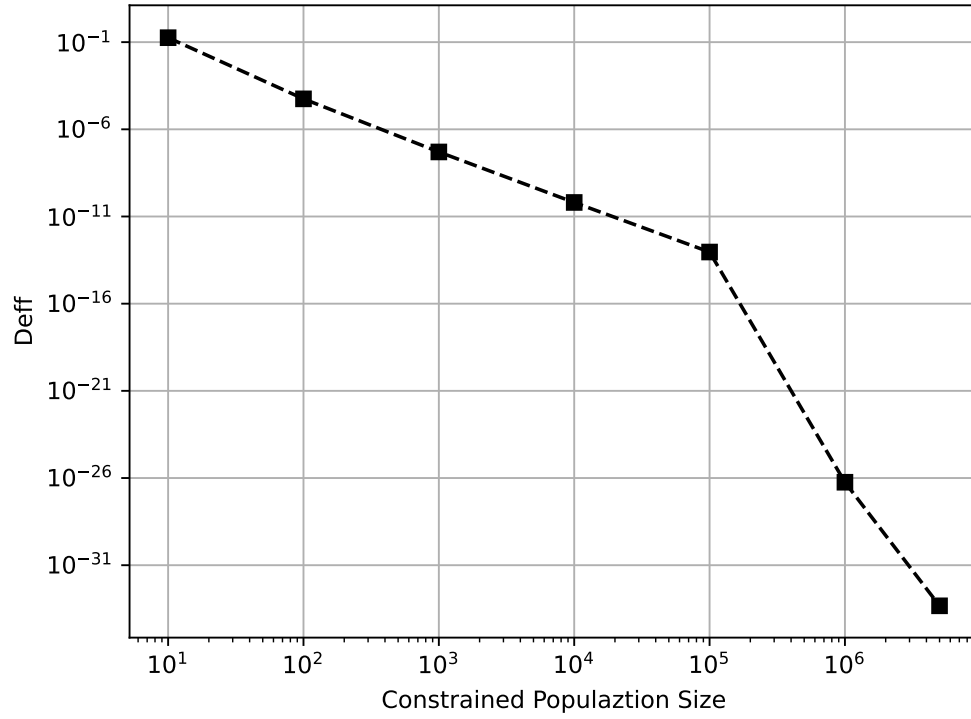Figure 8: AIC as a function of the number of clusters in the Gaussian Mixture density model.

Figure 9: Design effect as a function of the constrained population. Design effect between the schemes MFS and SRS. On the x-axis, the population size and on the y-axis the values of the design effect
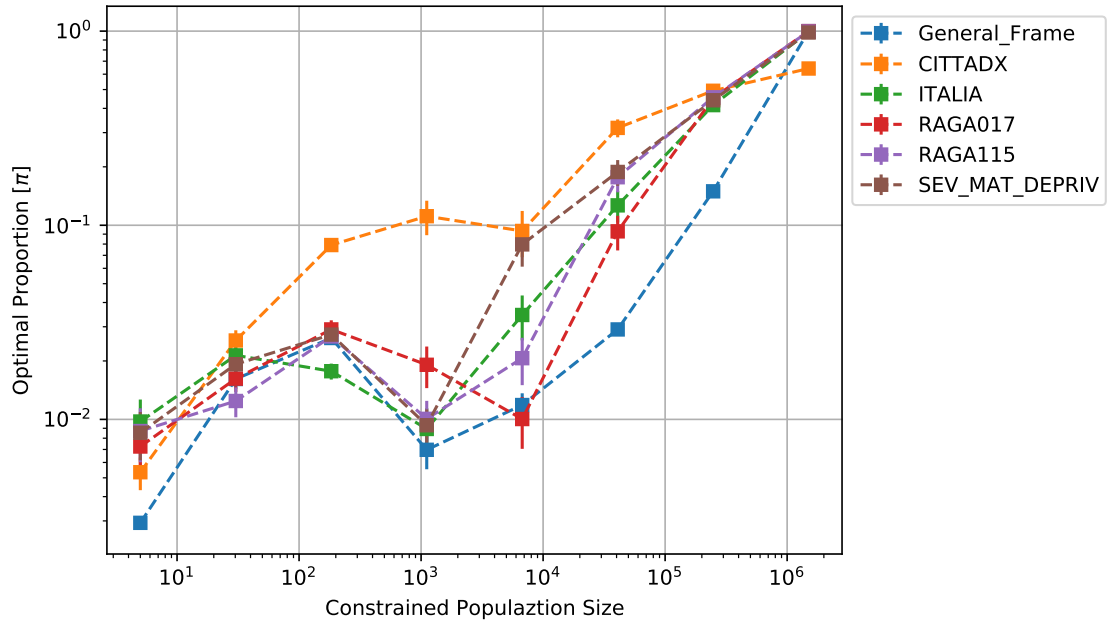

Figure 10: Optimal proportions for the MFS variance estimator as a function of constrained population size. Error bars represent the 95% Confidence interval.
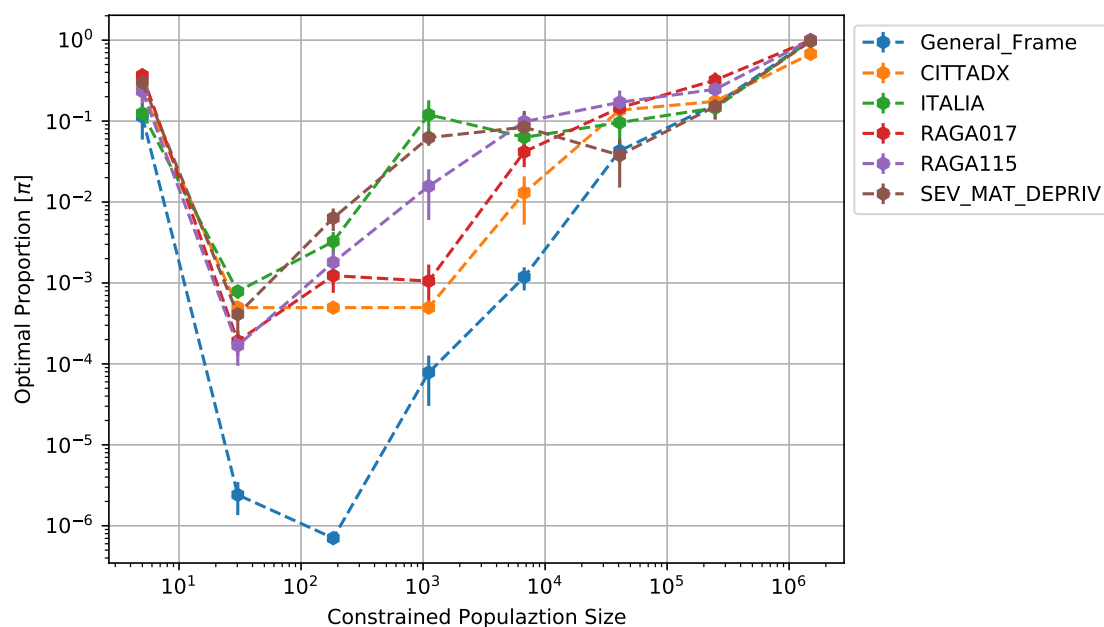
Figure 11: Optimal proportions for the SRS variance estimator as a function of constrained population size. Error bars represent the 95% Confidence interval.
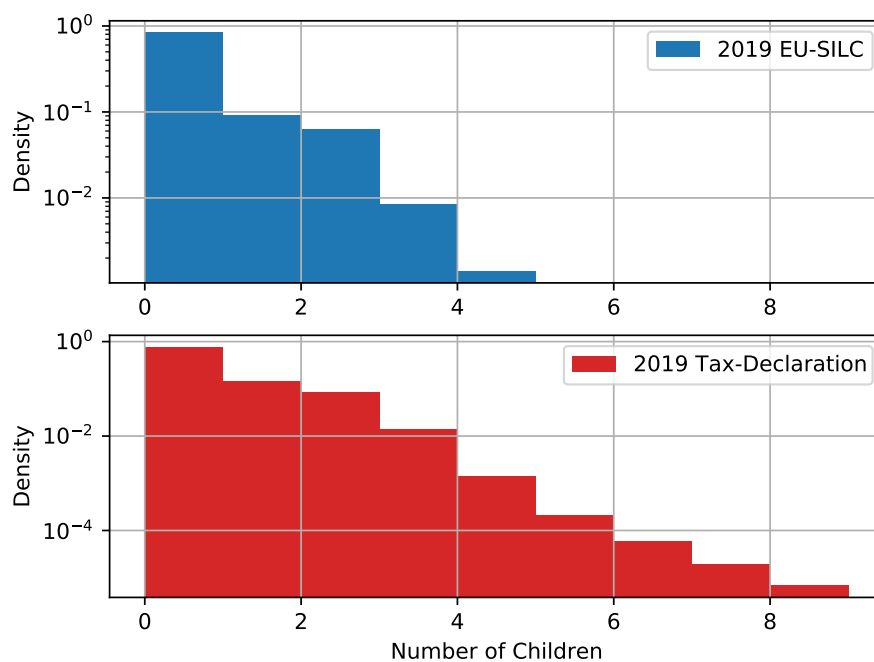

Figure 12: Historograms for the number of children in households in Liguria. Blue: 2019 EU-SILC data, Red: Tax-declaration

Hence, when considering a muti-frame framework we can simplify the approach to population estimation by resorting to this sort of equivalence due to a repositioning of the response variable. The mitigation of the overlapping over the incomplete frames allows us to study them separately. We shall discuss it more in detail in the next sections.

B.2 Population estimator

Let us consider the set of incomplete frames $\mathcal{R} = \{R_1 \ldots R_n\}$ and a the population $\mathcal{P} = \bigcup_{k=1}^{n} R_k$. Our goal is the estimate of the total population adopting a MFS scheme.

We can use the argumentation of section B.1 to consider the incomplete frames as independent. As a result, adopting the notation of section B.1, the MFS scheme over the frames of $\mathcal{R} = \{R_1 \ldots R_n\}$ can be regarded as a SRS over the frames of $\mathcal{Q} = \{Q_1 \ldots Q_n\}$.

To get the population estimation we thus start from a SRS; the population estimator is given as

$$\hat{t} = \sum_{q \in \mathcal{Q}} \sum_{k \in \mathcal{S}_q} u_k$$

with $\mathcal{S}_q$ a sample over the q-th frame. We evaluate the expectation of $\hat{t}$, so we have

$$\mathbf{E}\left(\sum_{q \in \mathcal{Q}} \sum_{k \in q} u_k \zeta_k^{(q)}\right) = \sum_{q \in \mathcal{Q}} \sum_{k \in q} u_k \mathbf{E}\left(\zeta_k^{(q)}\right) = \sum_{q \in \mathcal{Q}} \sum_{k \in q} u_k \pi_k^{(q)};$$

with $\zeta_k^{(q)}$ a random dichotomic variable (with fix probability $\pi_k^{(q)}$ which is the same per each sample of each frame). As a result, the unbiased estimator is obtained after reweighting u_k with the weight $\frac{1}{\pi_k^{(q)}}$, namely

$$\hat{t}_{unbiased} = \sum_{q \in \mathcal{Q}} \sum_{k \in \mathcal{S}_q} \frac{u_k}{\pi_k^{(q)}}. \tag{19}$$

When all subjects are equally likely to be acquired into $\mathcal{S}_q$, the weight $1/\pi_k^{(q)}$ is determined by the ratio $\frac{N_q}{n_q}$, with n_q the number of subjects sampled and N_q the total number of elements within the q-th frame.

B.3 Variance estimator

Let us determine the variance estimator of (19). Thus, we write down

$$\mathbf{V}(\hat{t}_{unbiased}) = \mathbf{V}\left(\sum_{q \in Q} \sum_{k \in \mathcal{S}_q} \frac{u_k}{\pi_k^{(q)}}\right).$$

Since the frames of $\mathcal{Q}$ are independent, we write down

$$\mathbf{V}(\hat{t}) = \sum_{q \in Q}\left[\sum_{k \in q} \mathbf{V}\left(\frac{u_k \zeta_k^{(q)}}{\pi_k^{(q)}}\right) + \sum_{i \in q; j \in q} (1 - \delta_{ij}) \mathbf{Cov}\left(\frac{u_i}{\pi_i^{(q)}} \zeta_i^{(q)}; \frac{u_j}{\pi_j^{(q)}} \zeta_j^{(q)}\right)\right],$$

with δ_{ij} denoting the Kronecker's delta.

We recall that

$$\mathbf{V}(\zeta_i^{(q)}) = \frac{n_q}{N_q}\left(1 - \frac{n_q}{N_q}\right);$$

$$\mathbf{Cov}(\zeta_i^{(q)}; \zeta_j^{(q)}) = -\frac{1}{N_q - 1}\left(1 - \frac{n_q}{N_q}\right)\left(\frac{n_q}{N_q}\right)$$

and

$$\sum_{i \in q; j \in q} (1 - \delta_{ij}) u_i u_j = \left(\sum_{k \in q} u_k\right)^2 - \sum_{k \in q} u_k^2.$$

21

By putting them all together, under the assumption that $\pi_i^{(q)} = \frac{n_q}{N_q}, \forall i \in Q_q$, we obtain the variance estimator

$$\mathbf{V}(\hat{t}) = \sum_{q \in Q} \frac{1}{N_q - 1} \left(\frac{N_q}{n_q} - 1 \right) \left[N_q \sum_{k \in q} u_k^2 - \left(\sum_{k \in q} u_k \right)^2 \right]. \tag{20}$$

Alternatively, we can write the last equation as

$$\mathbf{V}(\hat{t}) = \sum_{q \in Q} N_q \left(\frac{N_q}{n_q} - 1 \right) S_q^2;$$

with

$$S_q^2 = \frac{1}{N_q - 1} \sum_{k \in q} \left(u_k - \frac{1}{N_q} \sum_{l \in q} u_l \right)^2 = \frac{1}{N_q(N_q - 1)} \left(N_q \sum_{k \in q} u_k^2 - \left(\sum_{k \in q} u_k \right)^2 \right).$$

To obtain the unbiased variance estimator, we can make first the following observation that in a SRS is valid $\mathbf{E}(s_q^2) = S_q^2$; with

$$s_q^2 = \frac{1}{n_q - 1} \left[N_q \sum_{i \in \mathcal{S}_q} u_i^2 - \frac{N_q}{n_q} \left(\sum_{i \in \mathcal{S}_q} u_i \right)^2 \right].$$

Thus, the unbiased variance estimator is

$$\hat{v} = \sum_{q \in Q} N_q \left(\frac{N_q}{n_q} \right) s_q^2;$$

or equivalently

$$\hat{v} = \sum_{q \in Q} \left(\frac{N_q}{n_q} - 1 \right) \frac{N_q}{n_q - 1} \left[N_q \sum_{i \in \mathcal{S}_q} u_i^2 - \frac{N_q}{n_q} \left(\sum_{i \in \mathcal{S}_q} u_i \right)^2 \right]. \tag{21}$$

To make evident the dependence on $\pi^{(q)}$, we can rearrange the last equation as

$$\hat{v} = \sum_{q \in Q} N_q \left(\frac{1 - \pi^{(q)}}{\pi^{(q)}} \right) \frac{1}{N_q \pi^{(q)} - 1} \left[N_q \sum_{i \in \mathcal{S}_q} u_i^2 - \frac{1}{\pi^{(q)}} \left(\sum_{i \in \mathcal{S}_q} u_i \right)^2 \right]. \tag{22}$$

B.4 Sample size estimation

To estimate the sample size we opted for solving an optimal problem. We required the estimator (22) to be minimal, given the constraint that the sample size must be equal to some fixed n. For convenience, unlike the previous sections, we shall change the notation $\pi^{(q)}$ with π_q.

The Lagrangian for such a problem is the following:

$$\mathcal{L}[\pi_1, \ldots, \pi_n, \lambda] = \hat{v}_{mfs} + \lambda \left(\sum_{j \in \mathcal{Q}} \pi_j \nu_j - n \right); \tag{23}$$

with $\nu_j = \sum_{i=1}^{N_j} \frac{1}{m_i}$; m_i the multiplicity of the i-th element with respect to the frames $\mathcal{Q}_1 \ldots \mathcal{Q}_n$. λ is a Lagrange multiplier. ν_j describe the *effective population* per frame, i.e., it holds $\sum_{j \in \mathcal{Q}} \nu_j = N_q$. Thus, we aim to find the optimal values of $\pi_1, \ldots, \pi_n$ leading to the minimal variance estimator, given n the sample size.

To find the stationary solution of the Lagrangian, we impose

$$\frac{\partial \mathcal{L}}{\partial \pi_q} = \frac{\partial \hat{v}_{mfs}}{\partial \pi_q} + \lambda \nu_q = 0 \tag{24}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \sum_{j \in \mathcal{Q}_j} \pi_j \nu_j - n = 0. \tag{25}$$

For convenience, we rewrite (22) as

$$\hat{v}_{mfs} = \sum_{q \in \mathcal{Q}} \Phi_q(\pi_q) \Psi_q(\pi_q).$$

With

$$\Phi(\pi_q) = \frac{N_q}{\pi_q} \frac{(1 - \pi_q)}{N_q \pi_q - 1}$$

and

$$\Psi(\pi_q) = \pi_q N_q \sum_{i \in \mathcal{S}_q} u_i^2 - \left(\sum_{i \in q} u_i \right)^2.$$

Accordingly, we formulate the gradient with respect to π as

$$\partial_{\pi_q} \mathcal{L} = \partial_{\pi_q} \Phi_q \left[\pi_q A_q - B_q \right] + \Phi_q A_q + \lambda \nu_q.$$

with

$$A_q = N_q \sum_{i \in \mathcal{S}_q} u_i^2$$

and

$$B_q = \left(\sum_{i \in q} u_i \right)^2.$$

We would like to solve (24); we try rewriting that equation as

$$\partial_{\pi_q}(A_q \Phi_q \pi_q - \Phi_q B_q) = -\lambda \nu_q;$$

we integer both sides and get

$$A_q \Phi_q \pi_q - \Phi_q B_q|_{\pi_q}^1 = -\lambda \nu_q (1 - \pi_q).$$

With $\pi_q \in [1/N_q, 1]$; we recall that $\Phi(1) = 0$. When putting it all together we have

$$\Phi(\pi_q) = \frac{\lambda \nu_q (\pi_q - 1)}{B_q - A_q \pi_q}. \tag{26}$$

We attempt to make the form of the stationary solution more explicit, so we write

$$\frac{N_q}{\pi_q} \frac{(1 - \pi_q)}{N_q \pi_q - 1} = \frac{\lambda \nu_q (\pi_q - 1)}{B_q - A_q \pi_q}$$

or equivalently,

$$\frac{N_q}{\pi_q} \frac{1}{N_q \pi_q - 1} = \frac{\lambda \nu_q}{A_q \pi_q - B_q}.$$

Thus, we obtain the optimal solution

$$\pi_q = \frac{\left(\frac{A_q N_q}{\lambda \nu_q} - 1 \right) + \sqrt{\left(\frac{A_q N_q}{\lambda \nu_q} - 1 \right)^2 + 4 \frac{B_q N_q}{\lambda \nu_q}}}{2 N_q} \tag{27}$$

However, the optimal solution (27) is still dependent on λ. To determine λ we need to solve

$$\sum_{j \in \mathcal{Q}_j} \frac{(A_j N_j - \nu_j) + \sqrt{(A_j N_j - \nu_j)^2 + 4 B_j N_j^2 \nu_j \lambda}}{2 N_j \nu_j \lambda} = n. \tag{28}$$

The solution of (28) together with (27) provide the desired optimal solution. As shown, it is not possible to find a closed-form solution.

Acknowledgements

Code Avaliability

Python code available at `https://github.com/glancia93/-Population-Detection-and-Sampling-in-Liguria/tree/main`

References

Akaike, H. (1976). Canonical correlation analysis of time series and the use of an information criterion. In *Mathematics in science and engineering*, volume 126, pages 27–96. Elsevier.

An, J. and Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. *Special lecture on IE*, 2(1):1–18.

Asudeh, A., Jin, Z., and Jagadish, H. (2019). Assessing and remedying coverage for a given dataset. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 554–565. IEEE.

Asudeh, A., Shahbazi, N., Jin, Z., and Jagadish, H. (2021). Identifying insufficient data coverage for ordinal continuous-valued attributes. In *Proceedings of the 2021 international conference on management of data*, pages 129–141.

Atkinson, A. B. (1998). *World of poverty*. Wiley-Blackwell.

Bankier, M. D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81(396):1074–1079.

Bereziński, P., Jasiul, B., and Szpyrka, M. (2015). An entropy-based network anomaly detection method. *Entropy*, 17(4):2367–2408.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Boullé, N., Nakatsukasa, Y., and Townsend, A. (2020). Rational neural networks. *Advances in neural information processing systems*, 33:14243–14253.

Brackertz, N. (2007). Who is hard to reach and why?

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58.

Daneshpazhouh, A. and Sami, A. (2014). Entropy-based outlier detection using semi-supervised approach with few positive examples. *Pattern Recognition Letters*, 49:77–84.

Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596.

Eurostat (2020). European union - statistics on income and living conditions.

Ferguson, H. B., Bovaird, S., and Mueller, M. P. (2007). The impact of poverty on educational outcomes for children. *Paediatrics & child health*, 12(8):701–706.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

Fuller, W. A. and Burmeister, L. F. (1972). Estimators for samples selected from two overlapping frames. In *Proceedings of the Social Statistics Section, American Statistical Association*, volume 245249.

Hartley, H. and Rao, J. (1962). Sampling with unequal probabilities and without replacement. *The Annals of Mathematical Statistics*, pages 350–374.

Hauser, R. (2008). Problems of the german contribution to eu-silc-a research perspective, comparing eu-silc, microcensus and soep.

Hofmarcher, T. (2021). The effect of education on poverty: A european perspective. *Economics of Education Review*, 83:102124.

ISTAT (2019). Il censimento permanente della popolazione in liguria.

Jin, Z., Xu, M., Sun, C., Asudeh, A., and Jagadish, H. (2020). Mithracoverage: a system for investigating population bias for intersectional fairness. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 2721–2724.

Kamath, U. and Liu, J. (2021). *Explainable artificial intelligence: An introduction to interpretable machine learning*. Springer.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lohr, S. (2007). Recent developments in multiple frame surveys. *cell*, 46(42.2):6.

Lohr, S. (2011). Alternative survey sample designs: Sampling with multiple overlapping frames. *Survey Methodology*, 37(2):197–213.

Lohr, S. and Rao, J. (2006). Multiple frame surveys: point estimation and inference. *Journal of American Statistical Association*, 101:1019–1030.

Lohr, S. L. (2021). *Sampling: design and analysis*. CRC press.

Longford, N. T. (2014). *Statistical Studies of Income, Poverty and Inequality in Europe: Computing and Graphics in R Using EU-SILC*. CRC Press.

Mecatti, F. et al. (2007). A single frame multiplicity estimator for multiple frame surveys. *Survey methodology*, 33(2):151–157.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Nielsen, F. (2021). On a variational definition for the jensen-shannon symmetrization of distances based on the information radius. *Entropy*, 23(4):464.

Pisati, M., Whelan, C. T., Lucchini, M., and Maître, B. (2010). Mapping patterns of multiple deprivation using self-organising maps: An application to eu-silc data for ireland. *Social Science Research*, 39(3):405–418.

Raffo, C., Dyson, D., Gunter, H. M., Hall, D., et al. (2007). Education and poverty: A critical review of theory, policy and practice. *Education and Poverty Programme*.

Renganathan, V., Hashemi, N., Ruths, J., and Summers, T. H. (2021). Higher-order moment-based anomaly detection. *IEEE Control Systems Letters*, 6:211–216.

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4, pages 547–562. University of California Press.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Shahbazi, N., Lin, Y., Asudeh, A., and Jagadish, H. (2022). A survey on techniques for identifying and resolving representation bias in data. *CoRR, abs/2203.11852*.

Skinner, C. J. and Rao, J. N. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91(433):349–356.

Torabi, H., Mirtaheri, S. L., and Greco, S. (2023). Practical autoencoder based anomaly detection by using vector reconstruction error. *Cybersecurity*, 6(1):1.

Trindade, L. Z. and Goedemé, T. (2020). *The comparability of the EU-SILC income variables: review and recommendations*.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525.

Veljkovic, A., Pohoryles, D. A., and Bournas, D. A. (2023). Heating energy demand estimation of the eu building stock: Combining building physics and artificial neural networks. *Energy and Buildings*, 298:113474.

Wirth, H. and Pforr, K. (2022). The european union statistics on income and living conditions after 15 years. *European Sociological Review*, 38(5):832–848.

Zhou, C. and Paffenroth, R. C. (2017). Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 665–674.