The Survey Statistician, 2024, Vol. 89, 47-58

Ask the Experts

What is the state of play on statistical matching with a focus on auxiliary information, complex survey designs and quality issues?

Marcello D'Orazio, Marco Di Zio and Mauro Scanu[1]

[1] Istituto Nazionale di Statistica – Istat, madorazi@istat.it, dizio@istat.it, scanu@istat.it

Abstract

The statistical matching problem consists in fusing information from two data sources that are representative of the same population but contain observations on disjoint sets of units. The lack of joint information on variables observed distinctly in the two data sources induces a source of uncertainty that usually statistics does not tackle directly, under the status of unidentifiability of the model given the data at hand. This paper gives an updated account of what has been proposed in order to deal with this problem.

Keywords: identifiability, uncertainty, imputation, data fusion.

1 Introduction

The widespread use of data integration for statistical purposes gave rise to new challenges: statistical matching aims at overcoming one of these challenges for which there is not an immediate answer under the statistical point of view. The statistical matching problem consists in making "estimates" on parameters of the joint distribution of Y and Z when Y and Z are observed in two distinct data sources (*A* and *B* respectively), and the sets of units on which Y and Z are observed, although representative of the same population, are disjoint. Hence, it is not possible to connect records through the use of identifiers or exact or probabilistic record linkage.

We avoid the description of well know methods, mostly already covered in D'Orazio et al. (2006a) and references therein. In this paper, we try to give the state of play on statistical matching on some specific issues. First of all, it should be clear once and for all that the use of the data in *A* and *B* is not enough for the two purposes for which statistical matching has been considered: i) a "fused" complete but synthetic data set on which whatever statistical analyses involving Y and Z could be performed (micro approach); ii) estimation of specific parameters on the joint Y and Z distribution (macro approach). Hence, a discussion on how to incorporate additional information in terms of data sources or constraints is given in Section 2. Secondly, most data sets are drawn according to complex survey designs, and Section 3 covers this issue. Section 4 illustrates different areas on which statistical matching applications have been conducted. Quality of the results obtained by statistical matching is an essential point that sometimes seems to be neglected in real applications: we describe the state of play on quality measures in Section 5. Some conclusions and hints on areas of research are finally discussed in Section 6.

2 Use of auxiliary information

The statistical matching (SM henceforth) problem is naturally affected by an identifiability problem in the sense that data in *A* and *B* are not enough to estimate the parameters of the joint distribution of Y and Z. To fill this gap, assumptions as the independence of Y and Z given the matching variables $\boldsymbol{X}$ (conditional independence, CI henceforth) is needed and it is explicitly or implicitly used in the base SM procedures. This assumption severely limits SM procedures applicability. One way to move away from this assumption and thus improve the conclusions is that of using auxiliary information, particularly on the variables Y, Z or $(Y, Z|\boldsymbol{X})$. Such information can take several forms: it can be a set of micro data, information on parameters or aggregate values of the variables under observation, or refer to logical and statistical constraints on the variables (partial information) see D'Orazio et al. (2006b).

2.1 Exploitation of additional data sources

An auxiliary data set *C* can be used for matching purposes if it is a representative sample of the target population, otherwise the inference will be affected by a bias. Several methods have been proposed in the literature: parametric, nonparametric and mixed, see D'Orazio et al. (2006a) for a review. In the presence of a micro data set, the approaches described in D'Orazio et al. (2006a) follow the idea of creating a data set by appending file *A*, *B* and the auxiliary file *C* and treating it as a statistical inference problem in the presence of missing data. This approach is studied and further explored in a Bayesian context by Fosdick et al. (2016). They use a data augmentation algorithm that intrinsically produces multiply imputed values. The simulations show that the size of the auxiliary file *C* essentially represents the degree of confidence with respect to the auxiliary information. Although this result was expected, it is important to remember that every time a file *C* is used, its observed size naturally becomes our degree of confidence about its quality.

The representativeness of *C* can be a limiting assumption as well. In fact, most of the times *C* is an outdated sample, or a file composed of proxy variables, i.e., it is composed of information related to the variables under investigation - and thus it is important to take *C* into account in order to avoid the CI assumption - but characterized by a sort of proxy information. How to deal with this issue is an important topic still under investigation. In Moretti and Shlomo (2023) and Fosdick et al. (2016) there are two ways of approaching the non-representativeness of the auxiliary file *C*. In the first, there is the research and use of further additional information, while the second is characterized by a further but weaker assumption about the representativeness of *C*.

Moretti and Shlomo (2023) propose calibrating the prediction regression model of a mixed approach (predictive mean matching) to known marginal totals of the variables $(\boldsymbol{X}, Y, Z)$ to make the estimation of parameters robust to misspecification of the model. They empirically show that this approach improves the results of the matched file.

In Fosdick et al. (2016), an SM method is proposed for the case where the sample *C* is not representative for the joint distribution of $(\boldsymbol{X}, Y, Z)$ but is representative for its conditional distributions $Y|\boldsymbol{X}$ and $Z|\boldsymbol{X}$. The algorithm essentially consists of:

1. estimating the conditional distributions $Y|Z, \boldsymbol{X}$ and $Z|Y, \boldsymbol{X}$ from *C*,
2. obtaining a synthetic auxiliary file *C** by
    a) generating the observations from $Y, \boldsymbol{X}$ and $Z, \boldsymbol{X}$ observed in *A* and *B* respectively (e.g., by duplicating or sampling records with replacement from *A* and *B*),
    b) imputing the missing values of Y and Z in the respective subsets of *C** using the conditional distributions estimated in the first step given the observations generated by step 2a.

This algorithm produces an auxiliary file *C** that preserves the marginal distributions observed in *A* and *B* and the conditional distributions observed in *C*.

Even in this case, however, an assumption is made, namely that the conditional distributions observed in *C* are representative of those in the target population. To measure the validity of this assumption, the authors propose an empirical evaluation. They suggest comparing the marginal distributions of Y and Z obtained in the synthetic file *C** with those observed in *A* and *B*. A high discrepancy suggests that the conditional distributions of *C* are not representative and therefore this approach should be avoided.

Though developed in a Bayesian context, this approach can be an interesting proposal for dealing with non-representative auxiliary information in different inferential contexts.

2.2 Auxiliary information in terms of constraints

Another type of auxiliary information often available in official statistics concerns the use of logical or statistical constraints, known in the field of editing and imputation as edit rules (soft and hard edit rules). Hard constraints (hard edits) refer to relationships between variables that must be necessarily fulfilled by the values of each observation, for instance, babies cannot have an academic degree and the total costs of a company are greater than or equal to the amount spent on purchases. Soft constraints (soft edits), on the other hand, identify abnormal although possible behaviors, e.g., the ratio of purchases to sales is generally within an interval $[l, u]$, see De Waal et al. (2011). Auxiliary information in terms of constraints was firstly studied in D'Orazio et al. (2006a and 2006b), later with discretized continuous variables by Conti, Marella and Scanu (2016 and 2017). An interesting and extensive recent study with continuous variables is in Claramunt-González et al. (2023).

The introduction of hard constraints on the Y, Z variables naturally makes the conditional independence model unfit and impossible. However, the use of hard constraints is not immediate and needs further investigation. In D'Orazio et al. (2006a), constraints are introduced at the model estimation stage, i.e., the estimable parameters that determine the region of uncertainty is bounded by the introduced constraints. Claramunt-González et al. (2023) focuses on a mixed method, that is in fact a predictive mean matching, and the constraints are used not in the parameter estimation step, but in the stage of donor imputation, namely, donors are chosen taking into account the hard and soft constraints. As noted in D'Orazio et al. (2006a) and Claramunt-González et al. (2023), there is a general improvement in the matching results. However, some constraints may be more or less useful. In particular, the introduction of hard constraints in the imputation phase can lead to having empty imputation cells or cells with a low number of donors. So, the introduction of a constraint on $Y, \boldsymbol{X}$ or $Z, \boldsymbol{X}$ should be well thought because it does not make direct changes on the conditional independence model but may introduce problems for the imputation phase. It is also interesting to note that Claramunt-González et al. (2023) studied the use of constraints in the case of additional information on $(\boldsymbol{X}, Y, Z)$ that allows estimating a model that is not based on CI.

3 Approaches accounting for complex sample survey design

A large part of SM methods proposed in literature are designed to integrate random samples consisting of independent and identically distributed (iid) observations. This assumption is seldom valid in official statistics where the available data come from complex probabilistic surveys that commonly include stratification and clustering; these complex selection mechanisms consist of two or more stages that typically invalidate the independence assumption (units belonging to a cluster show a degree of homogeneity) and often result in unequal weighting of the final in-sample units (Base weights, which are the reciprocal of first order inclusion probabilities, are corrected to compensate for unit nonresponse and for coverage problems, so the final survey weights are often the outcome of calibration or post-stratification procedures). In this framework, the target of inference are finite population quantities and the approach to inference is typically *design-based* or *model-assisted design-based* (Särndal et al., 1992). Therefore, the application of SM methods has the

objective of estimating the correlation coefficient between Y or Z or the contingency table crossing Y and Z at finite population level, or to create a synthetic representative sample that can be used for the subsequent analyses (under the paradigm of design-based inference). In this context, when SM aims at estimating model parameters (e.g., the correlation between Y and Z), it should be considered that the model assumed for the data in the sample is often not the same as that in the population and the sampling design is said to be *nonignorable* or *informative* (cf. Opsomer, 2009).

The SM methods accounting explicitly for the sampling design (and survey weights) are quite limited. In the past, two main different approaches were suggested: (i) Rubin's file concatenation (Rubin, 1986) and, (ii) Renssen's matching by weight calibration (Renssen, 1998).

Rubin's file concatenation consists in appending the two data sources and re-calculating the sampling weights in order to achieve representativeness of the target population. As noted by Ballin et al. (2008 and 2009) the recalculation of weights is not straightforward and requires knowledge of information on several aspects (sampling frame, design variables, etc.) typically available solely within the statistical agency that administers both the surveys. In addition, the approach does not consider unit nonresponse and corresponding weights' correction. In any case, the re-calculation of the weights does not solve the problem of lack of joint information regarding Y and Z. In practice, two imputations steps are still required (Y in subsample *B* and of Z in subsample *A*). For all these reasons the approach is seldom applied (see e.g., Ballin et al., 2008; Torelli et al., 2009).

The matching by weights' calibration proposed by Renssen (1998) is primarily intended to estimate the contingency table crossing Y and Z when Y and Z are categorical target variables. The method has the advantage of providing an estimated table whose marginal distributions are fully coherent with those estimated from the starting data sets via the Horvitz-Thompson (HT) estimator ($\widehat{N}_j = \sum_{k\in A} \widetilde{w}_A I(y_k = j);\ j = 1, \dots, J$). However, Renssen pinpoints that for coherence purposes, before estimating the Y and Z cross-table, it is necessary to align the marginal/joint distributions of the matching variables $\boldsymbol{X}$ in order to return the same known totals; this latter task requires two weights' calibration steps. Renssen's approach can exploit additional auxiliary information coming from a third sample *C* that observes both Y and Z (and possibly also $\boldsymbol{X}$). This method permits also the creation of a synthetic sample by means of a two-step procedure (resembling predictive mean matching). In fact, the estimation of the Y and Z cross-table requires the adoption of linear models (*linear probability models* in the case of categorical variables), whose predictions can be used as input of nearest neighbor hotdeck to impute in the recipient the target variable observed in the donor sample (see e.g., Donatiello et al., 2022). D'Orazio et al. (2010) extend Renssen's idea by replacing weights' calibration with the procedure suggested by Wu (2004).

Renssen's procedure is the most popular approach to handle survey weights in the two samples. It belongs to the larger class of SM methods that use survey weights in the matching step. A seminal proposal in this sense is that of Barr and Turner (1981) that consists in creating a synthetic sample using a constrained nearest neighbor hotdeck where the weights assigned to matched units in the synthetic sample are obtained by solving an optimization problem that guarantees reproducing the same estimated total amount of Y (Z) obtained by applying the traditional HT estimator in *A* (*B*) (this procedure assumes CI on Y and Z given $\boldsymbol{X}$ and that the starting weights in *A* and *B* return the same estimated population size). Renssen notes that this result can be achieved with much less computational effort by applying his procedure under CI (in absence of an additional auxiliary sample *C*), after the initial weights' calibration aimed at aligning the totals of the matching variables. D'Orazio (2015) suggests a slight modification of random hotdeck to use donors' survey weights in the random draw of a donor: this approach follows the ideas of weighted random hotdeck (cf. Andridge and Little, 2010).

Recently Kim et al. (2016) proposed a fractional imputation method aimed at creating a synthetic sample where the survey weights are used in the different steps of the imputation procedure.

Jauslin and Tillé (2023) follow Renssen's ideas to develop a nonparametric procedure; it at first applies weights' calibration to harmonize the totals of the matching variables, and then imputes the

recipient data set by using a nearest neighbor hot-deck approach that is constrained to use a donor only once and to return estimated totals of the imputed variable equal to those estimated from the donor sample. The optimization problem is solved by means of an algorithm commonly used in balanced sampling.

Schifeling et al. (2019) use SM to assess how measurement errors affect observation of a target variable (Z) in one survey when the same variable is observed free of errors (Y) in another survey. They suggest adoption of design-based inference to calculate the estimates of the cells of the contingency table crossing $\boldsymbol{X}$ and Y and the corresponding sampling variance. Then this information, coupled with assumed measurement error models (that avoid CI), becomes the input of a Bayesian approach that ends with an estimated posterior distribution of the true values and of model parameters.

Marella and Pfeffermann (2019) propose a unified framework for making inference on model parameters in the SM case when dealing with samples from complex sample surveys (informative samples). In particular, they define a sample likelihood that enables the estimation of the target population distributions and subsequently to impute the missing values. The authors give the conditions under which the sample models are identifiable and estimable from the starting data.

4 Some applications

SM methods have been and are applied in different domains: investigation of poverty and well-being, education statistics, travel and transportation statistics, agriculture statistics, etc. Although SM applications are mainly tailored to integrate data of probabilistic surveys, in some cases they are applied also to integrate probabilistic and non-probabilistic sources (administrative registers and more in general big data). Obviously, it is not possible to keep track of all the various applications of SM methods in the various domains and with different data settings. For this reason, in this section we limit our attention to some relevant applications with data stemming from probabilistic surveys referring to the same target population.

A huge number of papers apply SM methods to get insight on people's well-being (see e.g. Leulescu and Agafitei, 2013; Donatiello et al., 2016; Bernini et al., 2021). A large contribution to this objective is given by studies investigating the relationship between people's income and consumption, due to the well-known difficulties in collecting detailed information on both these items in the same survey. In the European Statistical System (ESS) this objective has been pursued by integrating the EU-Statistics on Income and Living Conditions (SILC) and the Household Budget Survey (HBS), in most of the cases by creating a synthetic sample that serves as basis for an in-depth investigation. Tonkin and Webber (2013) compare nonparametric methods and mixed SM methods. Donatiello et al. (2014, 2016 and 2022) investigate hotdeck imputation and warn about the consequences of assuming CI when matching the data of these two surveys. They show that CI approximately holds when considering a proxy of income or consumption in the matching process. As an alternative, they avoid CI by carrying out an assessment of uncertainty. Conti, Marella and Neri (2017) use Italian surveys to assess uncertainty due to the SM framework by including some constraints on the joint distribution of income and consumption. More recently, Donatiello et al. (2022) adopt Renssen's approach to derive a synthetic sample through a two-step procedure that uses a proxy of consumption in the matching process. The paper shows how crucial it is to think about SM when designing both the surveys, by harmonizing ex-ante the definitions and the classifications used for the common variables and by collecting the information (proxy of consumption in SILC) needed to make CI valid. In addition, the paper stresses that the nice feature of Renssen's approach of ensuring that the consumption imputed in the SILC survey maintains a marginal distribution aligned to that estimated in the HBS (donor) is crucial in official statistics, where the synthetic sample should provide estimates coherent with those obtained from the starting surveys. For the same reason, Rios-Avila (2015) suggests the use of a weight-splitting strategy to better comply with the constrained SM rationale and corresponding advantages. Ucar and Betti (2016) consider also the longitudinal dimension of the SILC survey and impute in it the consumption expenditure variable observed in the

cross-sectional HBS data; they also use a two-step Renssen procedure. Another application to Turkey data performs a constrained SM using propensity score ranking (Albayrak and Masterson, 2017). Decoster et al. (2020) compare different methods to impute consumption in a dataset on income in Belgium and create a data set for microsimulation purposes. Lopez-Laborda et al. (2020) suggest fitting Engel curves in a parametric SM approach with the objective of imputing consumption in SILC.

Recently attention moved from the joint distribution of income and consumption to a wider picture by including also wealth data. This is a very relevant domain of study, as it permits a thorough investigation of multidimensional poverty. A seminal paper on this topic is from Tedeschi and Pisano (2013) that investigate how to integrate the Survey on Household Income and Wealth (SHIW) carried out by the Bank of Italy with data on consumption available from the Italian HBS survey carried out by the Italian National Statistical Institute (Istat). Given the importance of the topic, recently Eurostat and OECD decided to join efforts and carry out an extensive SM exercise involving data from different countries within and outside EU (Balestra and Ohler, 2023), with the objective of measuring the joint distribution of household income, consumption and wealth at the micro level. In the same direction goes the work of Tram and Osier (2023) who want to explore multidimensional poverty in Luxembourg by applying a two-step approach that performs multiple imputation via Bayesian Bootstrap predictive mean matching.

Other survey data that are often involved in SM applications are those related to time use (TU), a very important topic, in particular when jointly investigated together with data collected in Labour Force Surveys (LFS). Gazzelloni et al. (2008) present a hot-deck application with Italian LFS and TU data; Ghahroodi (2023) considers Iranian data and suggests fitting tailored models for TU data (conditional predictive Dirichlet distribution or conditional predictive multinomial distribution) in the first step of an SM mixed approach. Zacharias et al. (2014) investigate the relationship between TU data and consumption expenditures by applying SM based on propensity scores. To investigate the measure of time and consumption poverty at microdata level Rios-Aviola (2016 and 2020) applies SM to integrate TU and living conditions data for different African countries.

Dalla Chiara et al. (2029) suggest an application of SM based on propensity scores for creating a synthetic dataset by integrating SILC, HBS, TU survey and data on household conditions and social capital; the fused file allows investigating households' living conditions in Italy.

Wiest et al. (2019) use SM to investigate effects of educational participation on well-being in later life. Bernini et al. (2021) aim at analyzing how happiness affects expenditure behaviour in different urbanized areas in Italy. Hossain et al. (2022) integrate data from the household travel survey with different specialized "satellite" surveys to assess the impact of COVID-19 on passenger travel demand in the Greater Toronto Area.

Torelli et al. (2009) and Ballin et al. (2009) explore the application of SM to the Farm Structure Survey and Farm Accountancy Data Network survey carried out on Italian farms, with the additional difficulty of managing dependent surveys. In the same framework, D'Orazio and Catanese (2016) use SM to assess the revenues and economic growth of farms producing renewable energies.

It is worth noting that a large part of the SM applications have a very challenging objective, namely the creation of a synthetic sample that serves as basis for in-depth analyses. Regardless of the SM method being applied and its complexity, a critical reading shows that several applications seem unaware of the assumptions underlying integration, in particular that of independence between the target variables conditional on the selected matching variables. Sometimes also the applications that are aware of CI and claim that it is valid, often ignore its consequences on extended analyses carried out on the synthetic sample. In fact, while CI permits to reliably explore the association/correlation between Y and the imputed Z, the same assumption may not lead to valid results when studying for instance the relationship between W, a different variable observed in the recipient file, and the imputed Z. In general, managing implications of the CI assumption can become quite difficult when SM is applied to integrate three or more data sources. For this reason, studies having very ambitious

objectives that require the integration of several samples should proceed very carefully and should dedicate much effort to understand whether the objectives can be reliably pursued given the available data and the underlying assumptions in the integration process.

5 Quality issues

As it is clear from the last considerations in Section 3, the big question in an SM problem is the assessment of the quality of the results. De Waal (2015) states that this is a primary field of research for SM, identifying two key issues: how to better extract information from the available data, and what kind of additional information could support SM?

As far as the first question is concerned, an update on the possible different estimators or imputation procedures has already been given in the previous sections. Hence, we focus here on the modelling issues that, frequently, imply the use of specific methods. As already said, a multivariate model that includes (Y, Z) is unidentifiable for the data at hand for SM. Since the publication of D'Orazio et al. (2006a), this additional source of uncertainty (i.e., uncertainty due to the lack of joint observations on Y and Z) has become more important in the evaluations of SM results than the uncertainty due to sampling (that can be always investigated by means of the usual tools, as for instance coefficients of variation). A thorough discussion on uncertainty in statistical matching is given in Conti, Marella and Scanu (2017). Up to now, this kind of uncertainty has been treated in the following ways.

1. It was resolved by assuming, possibly in an explicit way, specific models that are identifiable for the data. Much has been already said on the CI (mostly assumed subconsciously) and, as already remarked, we consider it important to be conscious of that assumption and to report it explicitly, if taken into account. This assumption seems appropriate just in those applications (see for instance Donatiello et al., 2022) that make use of at least one matching variable that is (very) highly correlated with either Y or Z, so that Y and Z become almost independent given the matching variables (something that can be imposed by construction when matching is planned while organizing the observation of the source files *A* and *B*). The CI cannot be tested by the data at hand.
2. A different identifiable model has been suggested by Kim et al. (2016). This model assumes that matching variables can be decomposed in two groups, say $\boldsymbol{X} = (V, W)$, and that V is an "instrumental variable" for Y, i.e., V is conditionally independent of Z given W and Y but V is correlated with Y given W. Under this model, the authors suggest the use of parametric fractional imputation (PFI, Kim et al., 2016). Also in this case, there is not a test that can validate the assumed model. Anyway, the authors state that "a sufficient condition for model identifiability is the existence of an instrumental variable in the model." Furthermore "The proposed methodology is applicable without the instrumental variable assumption, as long as the model is identified." Their estimation approach, based on the use of the EM algorithm, does not necessarily converge if the model is unidentifiable. They consequently claim: "In practice, one can treat the specified model as identified if the EM sequence converges." This seems the most interesting and intriguing aspect of this approach, that can justify the use of model assumptions even if untestable.
3. If no identifiable model can be constructed, a (possibly) large set of models are indistinguishable by the data at hand. D'Orazio et al. (2006a) use the notion of "likelihood ridge" as the set of all the equally likely maximum likelihood estimates of some parameters in order to represent the uncertainty on some parameters of the (Y, Z) distribution given the data at hand. For specific parameters, the width of the interval or space of all the equally plausible solutions quantifies how uncertain these parameters are given the available data on *A* and *B*. Results on Y and Z correlation coefficients are discussed in D'Orazio et al. (2006b), frequencies of contingency tables in D'Orazio et al. (2006a), for ordered categorical variables in Marella et al. (2013) and for generic empirical distributions on Y and Z in Conti et al. (2016).

The second question raised by De Waal (2015) was on what additional information can be considered in order to improve the quality of SM results. Much has already been described in Section 2 as far as additional data sources are considered. Here we focus on the effects of the use of constraints in terms of uncertainty. Conti et al. (2016) adopt as SM estimate for the distribution of (Y, Z) given $\boldsymbol{X}$ the central distribution among the ones in the estimated likelihood ridge given the data at hand. Even if the likelihood ridge is rather well known in general (consider for instance the Fréchet bounds for categorical variables and the parameter under the CI as a midpoint), this computation of the likelihood ridge's central distribution becomes cumbersome when constraints are imposed. They consequently define a very general estimator and derive its asymptotic properties as well as the width of the likelihood ridge in order to derive tests on the likelihood ridge's sparseness around the estimated distribution. This is just one of the papers that make use of the likelihood ridge width as a measure of the SM uncertainty due to the lack of joint observations on Y and Z.

A specific measure of the sparseness of the uncertainty set of distributions when the variables are categorical is given by the Fréchet bounds (D'Orazio et al., 2006a). Fosdick et al. (2016) compute the Fréchet bounds of Y and Z given $\boldsymbol{X}$ in order to verify, in a simulated context, the goodness of their estimator (described in Section 2.1) under the presence of different kinds of additional files *C*, and examine if these bounds are as tight as possible.

A simulated set up is the context where quality measures can be defined and applied, exploiting all their potentiality, given that the actual parameters to be estimated are known in advance. For instance, this happens in the already cited paper by Claramunt-González et al. (2023) where a multivariate mixed method for SM for the estimation of the correlation between Y and Z is proposed: in that context, quality has been assessed by i) computing the estimator bias (which can be calculated due to knowledge on the actual parameters) in a multiple imputation case (allowing variance estimation) and ii) transforming estimates by means of Fisher *z*-transformation (in order to ensure that the resulting transformed estimates are generated by a normal distribution). Besides using the mean squared error as a measure of performance of an estimator, the same authors identify also a measure for the imputed data set, checking whether the individual imputations are "correct within a $100 \times \tau$ per cent". For instance, when the Z observations z_a (which are known in advance in file *A* in a simulation study) are imputed by an SM procedure, it is verified whether each imputation lies in the interval with extremes $(1-\tau)z_a$ and $(1+\tau)z_a$. The authors note that this is a way to derive the so-called "matching noise" (see Conti et al., 2010): given that the objective of the matching noise is to describe the distance between the actual data generation process and the imputation process, and that this computation can be cumbersome for some estimators as the mixed ones, the identification of such an empirical computation of the matching noise is a nice trick to take into account. The computation of the fraction of "correct within a $100 \times \tau$ per cent" imputed values should be considered as a quality measure for the statistically matched file. In fact, they say in the paper that: "we do not intend to release any statistically matched datasets." We agree with their approach. Anytime a statistically matched data set is created and released for whatever unplanned statistical data production, it could happen that the chosen but unplanned Y and Z are connected in ways that that are not taken into account in the SM method, e.g., even by hard edit rules (see Section 2.2 and discussion therein). The introduction of constraints, as already discussed in D'Orazio et al. (2006a), dramatically improves quality of results obtained by SM. In fact, the uncertainty set of equally plausible estimated distributions for the data set at hand changes significantly, and the use of hard constraints (as introduced for instance in Section 2.2) excludes the conditional independence model among the distributions that can contribute to the uncertainty space, moving the "conditional" uncertainty space towards the actual but unknown distribution. Claramunt-González et al. (2023), as noted in Section 2.2, suggest to include not only hard constraints in an SM problem, but also soft ones. As the authors note, these constraints generally need the help of a third complete data set that allows one to fix the characteristics of soft rules that allow to isolate those values that are unlikely, even if possible. The introduction of these rules should help an SM procedure in reducing the uncertainty space. However, it is yet not clear how effective they are, and this could be, in our opinion, an interesting area of research.

As far as the computation of uncertainty is concerned when dealing with samples drawn according to complex survey designs, D'Orazio (2015) includes survey weights in estimating the uncertainty in the case of categorical Y and Z for the cells in the contingency table $Y \times Z$ in a standard matching framework. In this case, before the assessment for coherence purposes, it is suggested to align the marginal/joint distribution of the chosen matching variables, e.g., by using the IPF algorithm (starting from v.1.3.0 of the R package StatMatch, D'Orazio 2022). In addition, robust estimation methods are introduced to handle the problem of statistical zeroes.

6 Conclusions

Although appealing, SM can be tricky and hides features that can be dangerous for the credibility of the results. As remarked in all the sections, a clear assessment and declaration of all the assumptions underlying the specific statistical matching application is absolutely necessary. In particular, the micro approach is the one that could be most harming, and could lead to the "dog food problem" (see Claramunt-González et al., 2023 and references therein). Even in case of additional information and/or specific assumptions, an evaluation of uncertainty should be given and the reasons which lead the analyst to choose just one of the equally plausible estimates for either the micro or the macro approach should be clearly stated.

Furthermore, there are some approaches that need more attention and additional research, also in an applied setting. We mention just two of them. The first considers the use of models that include instrumental variables and make use of PFI (Kim et al., 2016): this approach touches all the main statistical matching issues, such as the presence of additional data sources and the use of complex survey designs for *A* and *B*, while paying attention to model identifiability. The second consists of a micro approach in which uncertainty is taken into account in the imputation process (imprecise imputation, see Endres et al., 2019), and that could be worthwhile also outside the usual statistical matching framework.

References

Albayrak, O. and Masterson T. (2017). Quality of statistical match of Household Budget Survey and SILC for Turkey. Levy Economics Institute of Bard College, Working Paper No. **885**.

Andridge, R.R. and Little, R.J.A. (2010). A review of hot deck imputation for survey nonresponse. *International Statistical Review*, **78**, 40-64.

Balestra, C. and Ohler, F. (2023). Measuring the joint distribution of household income, consumption and wealth at the micro level. Methodological issues and experimental results. Edition 2023. European Union/OECD, Statistical working papers.

Ballin, M., Di Zio, M., D'Orazio, M., Scanu, M. and Torelli, N. (2008), File concatenation of survey data: a computer intensive approach to sampling weights estimation. *Rivista di Statistica Ufficiale*, **2-3**, 5-12.

Ballin, M., D'Orazio, M., Di Zio, M., Scanu, M. and Torelli, N. (2009), Statistical matching of two surveys with a common subset. Univ. di Trieste, Dip. Scienze Economiche e Statistiche, Working Paper, **124**.

Barr, R.S. and Turner, J.S. (1981). Microdata file merging through large-scale network technology. *Mathematical Programming Study*, **15**, 1-22.

Bernini, C., Emili, S. and Galli, F. (2021). Does urbanization matter in the expenditure happiness nexus?. *Pap Reg Sci*., 1-26.

Claramunt-González, J., Van Delden, A. and De Waal, T. (2023). Assessment of the effect of constraints in a new multivariate mixed method for statistical matching. *Computational Statistics and Data Analysis*, **177.**

Conti, P.L., Marella, D. and Neri, A. (2017). Statistical matching and uncertainty analysis in combining household income and expenditure data. *Statistical Methods & Applications*, **26**, 485-505.

Conti, P.L., Marella, D. and Scanu, M. (2010). Evaluation of matching noise for imputation techniques based on the local linear regression estimator. *Computational Statistics and Data Analysis*, **53**, 354-365. DOI 10.1016/j.csda.2008.07.041.

Conti, P.L., Marella, D. and Scanu, M. (2016). Statistical matching analysis for complex survey data with applications. *Journal of the American Statistical Association*, **111**, 1715-1725. DOI 10.1080/01621459.2015.11128.

Conti, P.L., Marella, D. and Scanu, M. (2017). How far from identifiability? A systematic overview of the statistical matching problem in a nonparametric framework. *Communications in Statistics Theory and Methods*, **46**, 967-994.

Dalla Chiara, E., Menon, M. and Perali, F. (2019). An integrated database to measure living standards. *Journal of Official Statistics*, **35**, 531-576.

Decoster, A., De Rock, B., De Swerdt, K., Loughrey, J., O'Donoghue, C. and Verwerft, D. (2020). Comparative analysis of different techniques to impute expenditures into an income data set, *International Journal of Microsimulation*, **13**, 70-94.

De Waal, T. (2015). General approaches for consistent estimation based on administrative data and surveys, Discussion paper, Statistics Netherlands.

De Waal, T., Pannekoek, J., and Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons.

Donatiello, G., D'Orazio, M., Frattarola, D., Rizzi, A., Scanu, M. and Spaziani, M. (2014). Statistical matching of income and consumption expenditures. *International Journal of Economic Sc.*, **3**.

Donatiello, G., D'Orazio, M., Frattarola, D., Rizzi, A., Scanu, M. and Spaziani, M. (2016). The statistical matching of EU-SILC and HBS at ISTAT: where do we stand for the production of official statistics. DGINS - Conference of the Directors General of the National Statistical Institutes, 26-27 September 2016, Vienna.

Donatiello, G., D'Orazio, M., Frattarola, D. and Spaziani, M. (2022). The joint distribution of income and consumption in Italy: an in-depth analysis on statistical matching. *Rivista di Statistica Ufficiale - Review of Official Statistics*, **3**, 77-109.

D'Orazio, M. (2015). Integration and imputation of survey data in R: the StatMatch package. *Romanian Statistical Review*, **2**, 57-68.

D'Orazio, M. (2022) StatMatch: statistical matching or data fusion. R package version 1.4.1, https://CRAN.R-project.org/package=StatMatch.

D'Orazio, M. and Catanese, E. (2016). Evaluating revenues and economic growth for farms producing renewable energies: an investigation based on integration of FSS and EOAH 2013 survey data. Proceedings of the Seventh International Conference on Agricultural Statistics – ICAS VII, Rome 26-28 October 2016, 938-945 (DOI: 10.1481/icasVII.2016.e26c).

D'Orazio, M., Di Zio, M. and Scanu, M. (2006a). *Statistical Matching: Theory and Practice*. Wiley, Chichester.

D'Orazio, M., Di Zio, M., and Scanu, M. (2006b). Statistical matching for categorical data: displaying uncertainty and using logical constraints. *Journal of Official Statistics*, **22**, 137-157.

D'Orazio, M., Di Zio, M., and Scanu, M. (2010). Old and new approaches in statistical matching when samples are drawn with complex survey designs. *Proceedings of the* "XLV Riunione Scientifica" of the Italian Statistical Society (SIS), Padova, 16-18 June 2010.

Endres, E., Fink, P. and Augustin, T. (2019). Imprecise imputation: a nonparametric micro approach reflecting the natural uncertainty of statistical matching with categorical data. *Journal of Official Statistics*, **35**, 599-624.

Fosdick, B.K., De Yoreo, M. and Reiter, J.P. (2016). Categorical data fusion using auxiliary information. *The Annals of Applied Statistics*, **10**, 1907-1929.

Gazzelloni, S., Romano, M.C., Corsetti, G., Di Zio, M., D'Orazio, M., Pintaldi, F., Scanu, M. and Torelli, N. (2008). Time Use and Labour Force: a proposal to integrate the data through statistical matching. In: (Romano, M. C. ed.) I tempi della vita quotidiana: un approccio multidisciplinare all'analisi dell'uso del tempo, Argomenti N. **32**, Istat, 375-403.

Ghahroodi, Z.H. (2023). Statistical matching of sample survey data: application to integrate Iranian time use and labour force surveys. *Statistical Methods & Applications*, **32**, 1023-1051.

Hossain, S., Loa, P., Wang, K., Mashrur, S.M., Dianat, A. and Habib, K.N. (2022). Comprehensive data fusion to evaluate the impacts of covid-19 on passenger travel demands: application of a core-satellite data collection paradigm. Available at SSRN: https://ssrn.com/abstract=4181189 or http://dx.doi.org/10.2139/ssrn.4181189.

Jausling, R. and Tillé, Y. (2023). An efficient approach for statistical matching of survey data trough calibration, optimal transport and balanced sampling. *Journal of Statistical Planning and Inference*, **225**, 121-131.

Kim, J.K., Berg, E. and Park, T. (2016). Statistical matching using fractional imputation. *Survey Methodology*, **42**, 19-40.

Lopez-Laborda, J., Marin, C. and Onrubia, J. (2020). Estimating Engel curves: A new way to improve the SILC-HBS matching process using GLM methods. *Journal of Applied Statistics*, **48**, 3233-3250.

Leulescu, A. and Agafitei, M. (2013). Statistical matching: a model based approach for data integration. Publications Office of the European Union, Methodologies & Working papers.

Marella, D., Conti, P.L. and Scanu, M. (2013). Uncertainty analysis for statistical matching of ordered categorical variables. *Computational Statistics and data Analysis*, **68**, 311-325. DOI 10.1016/j.csda.2013.07.004.

Marella, D. and Pfeffermann, D. (2019). Accounting for non-ignorable sampling and non-response in statistical matching. *International Statistical Review*, **91**, 269-293.

Moretti, A. and Shlomo, N. (2023). Improving statistical matching when auxiliary information is available. *Journal of Survey Statistics and Methodology*, **11**, 619-642.

Opsomer, J.D. (2009). Introduction to part 4. Alternative approaches to inference from survey data. In Pfeffermann D. and C.R. Rao (Eds) *Sample Surveys: Inference and Analysis*, **29B**. Elsevier, Amsterdam.

Renssen, R.H. (1998). Use of statistical matching techniques in calibration estimation. *Survey Methodology*, **24**, 171-183.

Rios-Avila, F. (2015). Quality of match for statistical matches using the consumer expenditure survey 2011 and annual social economic supplement 2011. Levy Economics Institute of Bard College, Working Paper No. **830**.

Rios-Avila, F. (2016). Quality of match for statistical matches used in the development of the Levy institute measure of time and consumption poverty (limtcp) for Ghana and Tanzania. Levy Economics Institute of Bard College, Working Paper No. **873.**

Rios-Avila, F. (2020). Quality of match for statistical matches used in the development of the Levy institute measure of time and consumption poverty (limtcp) for Ethiopia and South Africa. Levy Economics Institute of Bard College, Working Paper No. **970**.

Rubin, D.B. (1986). Statistical matching using sample concatenation with adjusted weights and multiple imputations. *J. Bus. Econ. Stat*., **4**, 87-94.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer, New York

Schifeling, T., Reiter, J.P. and De Yoreo, M. (2019). Data fusion for correcting measurement errors. *Journal of Survey Statistics and Methodology*, **7**, 175-200.

Tedeschi, S. and Pisano, E. (2013). Data fusion between Bank of Italy-SHIW and ISTAT-HBS, MPRA Paper No. **51253**.

Tonkin, R. and Webber, D. (2013). Statistical matching of EU-SILC and the Household Budget Survey to compare poverty estimates using income, expenditures and material deprivation. 2013 Edition. European Commission Statistical Working Papers.

Torelli N., Ballin, M., D'Orazio, M., Di Zio, M., Scanu, M. and Corsetti, G. (2009). Statistical matching of two surveys with a non-randomly selected common subset. In: Eurostat, Insights on Data Integration Methodologies. Office for Official Publications of the European Communities, Luxembourg, 68-79, isbn: 9789279123061

Tram, T.T.H. and Osier, G. (2023). Identifying the disadvantaged in Luxembourg – Measuring multidimensional poverty by statistical matching. Economie et Statistiques, Working papers du STATEC, N. **133.**

Ucar, B. and Betti, G. (2016). Longitudinal statistical matching: transferring consumption expenditure from HBS to SILC panel survey. Univ. di Siena, Quaderni del Dipartimento di Economia Politica e Statistica, N. **739**.

Wiest, M., Kutscher, T., Willeke, J., Merkel, J., Hoffmann, M., Kaufmann-Kuchta, K. and Widany, S. (2019). The potential of statistical matching for the analysis of wider benefits of learning in later life. *European Journal for Research on the Education and Learning of Adults*, **10**, 291-306.

Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method. *Can. J. Stat.*, **32**, 1-12.

Zacharias, A., Masterson, T. and Memis, E. (2014). Time deficits and poverty, the Levy Institute measure of time and consumption poverty for Turkey. UNDP & Levy Economics Institute of Bard College, Research Project Report.