Statistical Papers
https://doi.org/10.1007/s00362-023-01468-3

REGULAR ARTICLE

Integrating rather than collecting: statistical matching in the data flood era

Riccardo D'Alberto[1] · Meri Raggi[1]

Received: 12 January 2022 / Revised: 28 June 2023

Abstract
Statistical matching is progressively emerging as a straightforward approach to data integration. This method of increasing importance and interest is useful to address the unsolved challenges posed by data shortage as well as the several opportunities occurring in the present data flood era. This paper offers an exhaustive review of the methodology from its early beginnings up to the most recent developments, considering also the most relevant applications. The links that statistical matching has with other integration methods are discussed, analysing how a 50-year-old method has been only recently proposed under a consistent but (yet) incomplete framework. Strengths and weaknesses of statistical matching are compared, considering different data features and sample representativeness frameworks, also, given future research ideas, always keeping an eye on uncertainty, the key problem to which statistical matching tries to answer.

Keywords Data integration · Data fusion · Imputation · Record linkage · Hot deck techniques

Mathematics Subject Classification 62D10 · 62G86 · 62P20 · 62P25

1 Introduction

Time constraints and budgetary restrictions are two relevant drivers of the ongoing process of rethinking the classic way of data collection operated nowadays by both the Official Statistics (OS) and researchers from different disciplines. Data collected

✉ Riccardo D'Alberto
riccardo.dalberto@unibo.it

Meri Raggi
meri.raggi@unibo.it

[1] Department of Statistical Sciences "P. Fortunati", Alma Mater Studiorum – University of Bologna, Via Delle Belle Arti, 41, Bologna 40126, BO, Italy

Published online: 19 September 2023

employing e.g., national censuses are being progressively cast off, while new scenarios for data integration emerge, offering new solutions for the OS, researchers, policymakers, and the general public.

The massive generation of Big Data experienced in the last two decades fosters the idea that information can be collected through countless approaches. Mobile devices, apps, social media, and the Internet of Things concur to offer the idea that there is no need to plan data collections anymore, rather, we only need modelling solutions to exploit the already available information (Iaccarino 2019). Nevertheless, raw data bring pending challenges in terms of quality, constraints due to privacy claims and security reasons, problems of data ownership as well as organizational, technological, and governance issues.

A gradual shift from data collection to data integration is already ongoing, primarily in the OS that is elaborating strategies to keep recursively up-to-date the available data sources, for example, by integrating both primary and secondary data, aggregating administrative registers, web data, project surveys, satellite, and geo-data. Thus, data integration represents the future of the incoming data production and sharing processes. The existing approaches lie in (1) traditional sources and administrative data, (2) traditional sources and Big Data and, (3) micro and macro-level data (UNECE 2017). There are different strategies with a common focus: to intensify the possibility to meet and properly answer the users' needs, assembling valuable information from multiple sources in a really broad research spectrum (Pentland 2019).

The integration of information originally collected in two (or more) data sources can be performed with different methods. Relevant ones are record linkage (RL), multiple imputation (MI), and statistical matching (SM).

RL consists of the exact and probabilistic approaches (Christen 2012). When two or more different data sources which refer to the same population must be integrated, exact RL allows us to merge on the basis of a common identifier for the units occurring in both data sets. If a record in one data set has exactly the same value in the common identifier as some records in the other data set, exact RL merges the records. This is the simplest case for the integration of different e.g., administrative sources. However, when (1) the sets of units collected by two or more data sources are (at least partially) overlapping, (2) no unique identifiers exist/can be used and, (3) the variables that the data sources have in common can serve as 'pseudo-identifiers' but they are misreported or change over time (Fellegi and Sunter 1969), probabilistic RL plays the role of the first actor in the integration process. Therefore, probabilistic RL detects the records of different data sets that refer to the same unit when exact identifiers cannot be used/are not available.

MI handles variables missing values. This is done, at the individual level, by a two steps approach. First, a small number of completed data sets are created and, from an imputation model, missing values are filled in. Second, estimates are computed in each completed data set and, finally, they are combined (Rubin 1987). MI can be used when a partially observed data set must be 'filled', for each record, by an estimated substitute of the variable's value that is randomly generated from the unknown conditional distribution of the missing variable given the observed one, using samples from an imputation model (Murray 2018). MI usually completes the records' missing entities by exploiting only one data set (Denk and Hackl 2003) and, "roughly speaking, the missing data are imputed more than once (...) being these imputations based on some distributional assumptions" (Rässler 2002, p. 5).

The present paper focuses on statistical matching and the methodology is depicted in detail in the following sections. However, to briefly frame SM in general terms, it allows us to integrate the information contained in two or more data sources when the operational context is characterized by the fact that (1) the different data sources collect information on (i) a set of common variables (**x**) and (ii) two sets of variables that are disjointly observed (**y** and **z**) and, (2) the units observed in the data sets are (potentially) disjoint sets of units (D'Orazio et al. 2006b). If RL deals with 'the same' units, SM deals with units that are as much as possible 'similar' (Judson 2005). The main difference among SM and RL/MI lies in the final integration goal. RL evaluates the coverage overlap between the data sets or the presence of duplicated records; it is applicable to add/remove records, potentially augmenting data in one source. Compared to MI, the integration focus of SM slightly differs from the one of MI which instead goes beyond the conventional two-databases situation (Judson 2005). Neither RL nor MI deals with the potentially widest goal of SM: building a synthetic (complete) data set from two (or more) data sources. SM creates a data set that is called 'synthetic' because it does not come from the direct observation/collection of information or, in other words, it is artificial. On the other hand, it is 'complete' in the sense that it ends up aggregating all the variables collected either in one or in the other data source. Furthermore, neither RL nor MI considers the amount of uncertainty behind the integration results, as, in contrast, SM allows us to do.[1] Moreover, SM serves the purpose(s) of data fusion more flexibly (Rässler 2002), being particularly useful when the missing data structure is such that there is the need to either acquire knowledge on the joint distribution function $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ or transferring from one source to the other the missing variable(s), only by exploiting the knowledge on **x**. In such context, the random variables (r.v.s) **y** are observed only in one data set, while the r.v.s **z** are observed only in another. The random variables **x** are observed in all the data sets at our disposal and, hereinafter, we call them 'matching' variables.

Nowadays, however, fruitful combinations of SM and RL/MI are emerging to deal with the challenges offered by Big Data integration, a field where non-probability samples must be considered (to date, Bethlehem 2016 identifies the main practical issues existing when matching different samples by dealing with mass imputation, while Rao 2021 offers an exhaustive review of the probability sampling methods, also, by focusing on models which bring valid inference from non-probability samples).

Two textbooks offer a cohesive dissertation of statistical matching: Rässler (2002) and D'Orazio et al. (2006b). The contribution of the present paper to the literature on SM and, more generally, to data integration is twofold. First, it reviews the latest SM developments and discusses the main findings on the identification, quantification, and treatment of the uncertainty behind data integration. Second, the paper considers these topics by covering the earliest SM developments up to the latest published articles. Both the methodological peculiarities and the SM strengths and weaknesses are discussed. The implications of the sampling frameworks in integrating data are investigated. The

[1] The key issue when using MI with non-overlapping data sets is that all the relevant model parameters cannot be estimated due to the fact that we measure sampling and imputation variance only. Hence, the model uncertainty is not considered by MI.

Table 1 Statistical matching: from the origins to the consecration, by schematically reading the state-of-the-art

Period	SM phases and main topics	Main contributions
The '70s	SM when it was just 'merging': first steps, trivariate normal, explanatory power of **x**	Okner (1972, 1974), Ruggles and Ruggles (1974) and Kadane (1978)
The '80s	The ascent of a method: categorical **x**, multivariate normal, alternative methods comparison, practical OS applications	Rodgers (1984), Walter (1984), Gavin (1985), Barry (1988), Singh et al. (1989) and Cohen (1991)
The '90s	A turning point: criticisms to CIA, comparisons among alternative approaches, sampling weights	Rubin (1986), Singh et al. (1993) and Renssen (1998)
The '00s	A look through cohesion: unequivocal notation, formalized approaches	Rässler (2002, 2004) and D'Orazio et al. (2006)
The '10s	Uncertainty and beyond: coherence, uncertainty definition/estimation, error measurement, independent samples	Conti et al. (2008), Vantaggi (2008), Conti et al. (2016), Conti et al. (2019) and Marella and Pfeffermann (2019)
The '20s	Big data and non-probability samples: the forthcoming integration	Chen et al. (2020), Kim et al. (2020) and Castro-Martín et al. (2022)

existing real data applications are not disregarded. The paper aims to provide casual readers as well as the interested ones with a useful map for the complete understanding of the method concerning its several shades of application.

To simplify the reader's journey across statistical matching, having a clearer look through the complex development and achievements of the method under a complete and shared theoretical framework, Table 1 shows the crucial contributions constituting the backbone of the SM state-of-the-art. These works and the others cited are listed in the References, but a schematic reading of the SM literature is proposed by grouping the most relevant contributions by the decades from the '70s up to nowadays, also, by following the macro-area subjects interested by the SM developments. Table 1 is meant to offer a clearer understanding of the SM evolution and to help the readers to efficiently focus on the main aspects characterising the method.

The paper is structured as follows. Section 2 briefly describes the method, highlighting its key features and presenting the two main SM goals (micro and macro). Section 3 reviews the 'merging' approach of the origins, discussing the need for formal cohesion that the preliminary SM proposals left aside. Section 4 investigates the non-parametric and Bayesian approaches, discussing the solutions offered to the problems of matching noise quantification and uncertainty definition/estimation. The uncertainty in SM is then analysed according to the most recent proposals in Sect. 5, with a specific focus on both the non-representative samples and the problems related to Big Data. Section 6 provides the concluding remarks and some considerations about further SM develop-

$$\begin{bmatrix} y^{\mathrm{A}}_{11} & \cdots & y^{\mathrm{A}}_{1m} & \cdots & y^{\mathrm{A}}_{1M} & x^{\mathrm{A}}_{11} & \cdots & x^{\mathrm{A}}_{1l} & \cdots & x^{\mathrm{A}}_{1L} & \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \\ y^{\mathrm{A}}_{i1} & \cdots & y^{\mathrm{A}}_{im} & \cdots & y^{\mathrm{A}}_{\mathrm{i}M} & x^{\mathrm{A}}_{i1} & \cdots & x^{\mathrm{A}}_{il} & \cdots & x^{\mathrm{A}}_{iL} & \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \\ y^{\mathrm{A}}_{n_{\mathrm{A}}1} & \cdots & y^{\mathrm{A}}_{n_{\mathrm{A}}m} & \cdots & y^{\mathrm{A}}_{n_{\mathrm{A}}M} & x^{\mathrm{A}}_{n_{\mathrm{A}}1} & \cdots & x^{\mathrm{A}}_{n_{\mathrm{A}}l} & \cdots & x^{\mathrm{A}}_{n_{\mathrm{A}}L} & \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \end{bmatrix}$$

$$\begin{bmatrix} \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & x^{\mathrm{B}}_{11} & \cdots & x^{\mathrm{B}}_{1l} & \cdots & x^{\mathrm{B}}_{1L} & z^{\mathrm{B}}_{11} & \cdots & z^{\mathrm{B}}_{1p} & \cdots & z^{\mathrm{B}}_{1P} \\ & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & x^{\mathrm{A}}_{j1} & \cdots & x^{\mathrm{A}}_{jl} & \cdots & x^{\mathrm{B}}_{jL} & z^{\mathrm{B}}_{j1} & \cdots & z^{\mathrm{B}}_{jp} & \cdots & z^{\mathrm{B}}_{jP} \\ & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & x^{\mathrm{A}}_{n_{\mathrm{B}}1} & \cdots & x^{\mathrm{B}}_{n_{\mathrm{B}}l} & \cdots & x^{\mathrm{B}}_{n_{\mathrm{B}}L} & z^{\mathrm{B}}_{n_{\mathrm{B}}1} & \cdots & z^{\mathrm{R}}_{n_{\mathrm{B}}p} & \cdots & z^{\mathrm{B}}_{n_{\mathrm{B}}P} \end{bmatrix}$$

Fig. 1 Data at hand in a statistical matching problem with two data sets (A and B)

ments. Appendix A offers an overview of the most considerable SM applications and software solutions.

2 Method in brief

For the sake of simplicity, let's consider two data sets: A and B. Let A be the 'recipient' data set, while B is the 'donor' data set. The number of observations in the two data sets is n_{A} and n_{B}, respectively. Let **x** be the random variables observed both in A and in B; **y** are the r.v.s observed only in A, while **z** are the r.v.s observed only in B. These r.v.s refer to the i-th and j-th observations collected in A and in B, with $i = 1, \ldots, n_{\mathrm{A}}$ and $j = 1, \ldots, n_{\mathrm{B}}$. Therefore, the observed r.v.s are

- $\mathbf{x} = \{X_1, \ldots, X_l, \ldots, X_L\}$, collected both in A and in B, (being X^{A}_l a vector of dimension n_{A}, while X^{B}_l is a vector of dimension n_{B}).
- $\underset{n_{\mathrm{A}}\times M}{\mathbf{y}} = \{Y^{\mathrm{A}}_1, \ldots, Y^{\mathrm{A}}_m, \ldots, Y^{\mathrm{A}}_M\}$, collected only in A (being $\mathrm{Y}^{\mathrm{A}}_m$ a vector of dimension n_{A}).
- $\underset{n_{\mathrm{B}}\times P}{\mathbf{z}} = \{Z^{\mathrm{B}}_1, \ldots, Z^{\mathrm{B}}_p, \ldots, Z^{\mathrm{B}}_P\}$, collected only in B (being Z^{B}_p a vector of dimension n_{B}).

Therefore, the data sets at hand are $\mathrm{A} = \left\{ \underset{n_{\mathrm{A}}\times L}{\mathbf{x}^{\mathrm{A}}}, \underset{n_{\mathrm{A}}\times M}{\mathbf{y}^{\mathrm{A}}} \right\}$ and $\mathrm{B} = \left\{ \underset{n_{\mathrm{B}}\times L}{\mathbf{x}^{\mathrm{B}}}, \underset{n_{\mathrm{B}}\times P}{\mathbf{z}^{\mathrm{B}}} \right\}$. The whole set of information that we have at hand is depicted in Fig. 1.

Statistical matching is applied to aggregate the information collected from different sources, by using two approaches: micro and macro (D'Orazio et al. 2006b). For the sake of simplicity, let's assume that $l = 1$, $m = 1$, and $p = 1$. In other words, let X, Y, and Z be univariate, continuous variables. Being $\mathcal{F}$ a family of distributions with each $f(X, Y, Z; \boldsymbol{\theta}) \in \mathcal{F}$ defined by a vector of parameters $\boldsymbol{\theta} \in \Theta$, macro SM aims at estimating the joint distribution function $f(X, Y, Z)$. On the other hand, micro SM aims at generating a synthetic (complete) data set from A and B. Whereas the former purpose should be clear, the latter deserves more explanation. Let S_d be a generic subset of d variables of interest (with $d = 1, \ldots, P$) chosen among the r.v.s $\mathbf{z}$. The goal of micro SM is imputing S_d from B to A and thus, generating the synthetic (complete) data set, named C, such that C = $\left\{ \underset{n_{\mathrm{A}} \times L}{\mathbf{x}^{\mathrm{A}}}, \underset{n_{\mathrm{A}} \times M}{\mathbf{y}^{\mathrm{A}}}, \underset{n_{\mathrm{A}} \times S_d}{\mathbf{z}^{\mathrm{A}}} \right\}$.

In the most general SM framework, let assume that

A.1. A and B collect information on two representative samples of the same target population (D'Orazio et al. 2006b).

A.2. The distinct samples A and B depicted in Fig. 1 can be considered as a unique sample A ∪ B of the $n_{\mathrm{A}} + n_{\mathrm{B}}$ i.i.d. observations from $f(X, Y, Z)$ (D'Orazio et al. 2006b).

A.3. From the overall sample given by A ∪ B, i.e., the sample of $n_{\mathrm{A}} + n_{\mathrm{B}}$ units from $f(X, Y, Z)$, a synthetic (complete) data set can be derived where the structure of missing information is missing completely at random (MCAR) or missing at random (MAR) (Rubin 1987; Rässler 2002, 2004).

The key estimation problem related to $f(X, Y, Z)$ has been often approached by resorting to the identifiable model derived from the conditional independence assumption (CIA). Briefly, the whole information set is defined by the A ∪ B sample and X, Y, and Z are independent and normal distributed r.v.s. Usually, CIA has been explicitly or implicitly adopted in SM for decomposing the aforementioned estimation challenge into smaller estimation problems by the factorization of the likelihood function (Anderson 1957). First, the solution was limited to the trivariate normal, while it was successively extended to multivariate distributions. Indeed, Rubin (1974) demonstrated that $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ is decomposable such that: $f(\mathbf{x}, \mathbf{y}, \mathbf{z}; \boldsymbol{\theta}) = f(\boldsymbol{x}; \boldsymbol{\theta}_{\mathbf{x}}) \cdot f(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta}_{\mathbf{y}|\mathbf{x}}) \cdot f(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}_{\mathbf{z}|\mathbf{x}})$. In other words, CIA allows computing the maximum likelihood estimator (MLE) for $\boldsymbol{\theta}_{\mathbf{x}}$ from the A ∪ B sample, while the MLEs for $\boldsymbol{\theta}_{\mathbf{y}|\mathbf{x}}$ and $\boldsymbol{\theta}_{\mathbf{z}|\mathbf{x}}$ are computed from A and B, respectively.

3 Where is the cohesion?

3.1 Statistical matching when it was just 'merging'

The key estimation problem in SM and a few solutions to overcome it were known since the '50s. However, only the availability of electronic computers gave spread to the first merging/matching applications: "the increased interest in social problems at the microeconomic level, as well as the chances offered by the developing technologies, fostered the demand for disaggregated socio-economic and demographic information" (Okner 1972, p. 325). The '1966 merge file' represents an early, rough approach in this direction. It was built from the 1967 Survey of Economic Opportunity and the 1966 Tax File referred to the U.S. families (Okner 1972) in response to the lack of consistent and comprehensive set of household data. The author answered to the need for official statistics about the distribution of the U.S. personal income or cross-classification by typical demographic characteristics of the population. He used the matching variables

x (wage, salary income, farm income...)[2] to set up a system of equivalence classes based on *major* and *minor* income sources and consistency scores. These, in turn, are used to assign 'points' to the units, thus to match them obtaining a new data file where the punctual pattern of income is recorded in addition to people's characteristics.

Sims (1972a, b) criticized such an approach since conditional independence was just implicitly assumed, while a more explicit theoretical framework was needed. Little is said about the outcome 'quality' and validity. Very limited considerations involved the adjustments required for evaluating the under-reporting or non-reporting that was eventually present in the original survey file. In this regard, a small improvement is offered by Alter (1974) who proposed to evaluate the concordance of the after-matching variables using cross-tabulation, integrating the 1970 Canadian Survey of Consumer Finances with the Family Expenditure 1970 Survey. However, the same author stressed that "the X–Y–Z problem remains unsolved (...) since a joint distribution of X, Y, and Z cannot be inferred from the known distributions of X with Y, and X with Z" (Alter 1974, p. 374).

Other pending challenges which were not taken into account by these contributions are related to the (implicit) assumption that the vector **x** is defined exactly in the same way in A and in B, although the matching variables may be affected by errors of different types, magnitude and frequency of occurrence.[3] Moreover, the peculiar, but not uncommon cases (e.g., in social sciences) of composite **x** or those of composite **z** and **y** were considered neither.[4]

An important residual issue is related to the rationale applied for the selection of the matching variables. Trivially, the choice of **x** was often data-driven, guided by the explanatory power of R^2. Using the coefficient of determination to assess the relationship strength between (X, Y) or (X, Z) (and, hence, the validity of the CIA) in an imputation by regression is straightforward. However, the cases in which we are not interested in assigning mean values but, rather, we want to reproduce the distributions of values in the original data and transfer complex sets of information do present further challenges. This was clear to Ruggles and Ruggles (1974) who made explicit that for matching purposes, no specific functional relationship must be determined in advance. Therefore, how to select the matching variables in the most efficient way? The authors proposed to match on the L-dimensional cross-tabulation using all the **x** variables between A and B. The matches will then be made stochastically with respect to the units which fall in the same cell. The assessment of the quality of the X variable

[2] Both originally observed and computed matching variables are used. Examples of computed matching variables are the total business income or property income, calculated from the sum of the absolute amounts of each of these components (taxable dividends, interests, savings, etc.).

[3] When such differences did occur, the adopted solutions consisted of mere clerical revision, as it is in Okner (1972).

[4] Composite variables are made up of two (or more) variables or measures that are highly related to one another, either conceptually or statistically. Scales, ratings, or categorical variables are usually used to make a composite variable (Grace 2006). The consequences related to the integration of composite variables can be linked to alterations of the relationship strength with potential outcome variables, modifications in statistical power, over(under)reduction of information, interpretation issues about the relationship of the composite variable with the outcome variable of interest. However, such variables offer potential pros; e.g., those linked to the possibility to exploit the information of proxies (observed in A) of the variable imputed from B.

intervals (i.e., the assurance that, within a specific interval of X, the distribution of Y and Z are invariant) is done with the Chi-square test on the Y and/or Z distributions. When significant differences are found, a correlation measure is computed to estimate *how much* the distributions differ. It is worth noticing that this discussion about the 'quality' of $\mathbf{x}$, also, embeds the considerations on the overall goodness and reliability of the final synthetic (complete) data set.

The aforementioned proposals have a common root: the (implicit) use of a pseudo-distance that is based on a hierarchically nested set of cross-tabulated cells, built on the variables $\mathbf{x}$ which are in common between A and B. Indeed, by successively partitioning these variables in narrower intervals, it is possible to tag and re-tag the units, then, by sorting the tags, the units can be selected for matching. This pseudo-distance is then similar to the weights attached to each matching variable X in a multivariate regression analysis that uses $\mathbf{x}$ (regressor) and $\mathbf{y}$ and $\mathbf{z}$ (dependent variables).

The works discussed so far proposed the SM approach under a methodological framework that was different from that of Record Linkage, even if the procedures implemented were often named equivocally like 'linkage', 'fusion', 'concatenation', etc. (Rässler 2002). The first, specific *matching* proposal that was presented within a framework characterised like the one described in Sect. 2 appeared only later, in Kadane (1978). Considering a triple of normal-distributed variables (X, Y, Z), the author concludes that the assumption of joint normality leads to the fact that all the regressions (X, Y), (X, Z) are linear, which is unlikely when real-world data are used. The solution proposed as a "way around the problem" (Kadane 1978, p. 424) is thus to adopt the aforementioned assumption limitedly, region-by-region, in the X space and, hence, to resort to separate estimates of the covariances (but for σ_{YZ} that is unobservable). Residual cases in which the information on σ_{YZ} can be retrieved consist of coarse samples which are yet perfectly matched, from which certain elements of σ_{YZ} can be known or, opting for the CIA. However, given that σ_{YZ} cannot be consistently estimated from the data at hand, the solution proposed by the author was to use a particular value for σ_{YZ} with the goal of getting results that would yield to a certain expected value (e.g., the expected amount of taxes willing to be raised by a specific tax schedule). Making assumptions about the distribution of σ_{YZ} and, hence, taking values for the latter from the distribution, finally bring results which could be weighted with the probability of the particular value of σ_{YZ}, such that σ_{YZ} is sampled. A drawback of this approach is that the more the assumption of normality loses reliability, the more methodological coherence diminishes. In addition, the proposed solution disregards any consideration about the validation of the final integration outcome.

3.2 The ascent of a method: applications from national agencies

The shortage of both theoretical foundations and empirical justification in Statistical Matching was made explicit for the first time by Rodgers (1984), pointing out that any finding that is drawn from the matched data sets is questionable as far its validity largely depends on the assumptions made, at first, on the matched variables. However, since these assumptions cannot be tested, it is compulsory to check for the consequences of the possible lack of validity. Then, the author proposed a first attempt for a cohesive

SM notation, introducing the fundamental concepts of distance function and donation classes. A distance function is defined as the absolute difference in the values of X (e.g., age) computed between two observations that come from different data sets. For example, a generic, basic distance is definable by $|x_i - x_j|$, for $i = 1, \ldots, n_{\text{A}}$ and $j = 1, \ldots, n_{\text{B}}$.[5] The donation classes are defined as homogeneous sub-groups of observations that help restrain the matched units' pairs (for example, by partitioning the units between male and female gender). Given that the data sets at hand collect information on the whole set of pairs made by the $n_{\text{B}}^{n_{\text{A}}}$ combination of donors and recipients, let $X^{\star}$ be a discretized variable whose categories $X^{\star}_{f}$, with $f = 1, \ldots, F$, identify the donation classes such that the size of the potential number of donor-recipient pairs can be restricted to $(n_{\text{B},X^{\star}_{f}})^{n_{\text{A},X^{\star}_{f}}}$.

Rodgers (1984) also hinted at a new integration perspective based on the usefulness of SM raising the topic of the validity of findings which result from analyses based on statistically matched data. Such a validity strongly depends on the accuracy of the underlying assumptions about the relationships between the variables. Given that A and B, separately considered, do not contain information about the relationships among variables $\mathbf{y}$ and $\mathbf{z}$, and SM only reflects the assumptions (implicit or explicit) made during the matching procedure, the matched data set we end up with is "a risky basis for analyses of such relationships" (Rodgers 1984, p. 96). The author considers SM simulations and empirical applications in different scenarios, testing the validity of CIA and discussing how much confidence to be placed in matching procedures and when, according to the set of variables at disposal. Namely, the topics of (1) unconstrained or constrained matching, (2) which matching variables to include in a distance function and, (3) the minimum size of the input data set that is required to carry out a matching process are investigated.

The integrated data set is valuable for the analyses which involve the relationships on $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ as far as the assumptions on such relationships made (or implied) by the analyst for the scopes of the integration procedure are robust. For example, let the case be that the following linear regression model has to be estimated: $\mathbf{z} = \mathbf{x}_{L-1} \cdot \boldsymbol{\beta} + \mathbf{y} \cdot \boldsymbol{\lambda} + \boldsymbol{\epsilon}$ (on the left side of the equation we have endogenous outcome variables which are explained by both endogenous and exogenous explanatory variables—on the right side of the equation—; there are vectors of non-zero parameters plus stochastic errors).[6] As Klevmarken (1982) points out, the possibility of estimating the parameters of such a linear model depends on the availability, for each Y included in the model, of at least one of the r.v.s $\mathbf{x}$ used in the argument of the distance function that is excluded from the set of $\mathbf{x}_{L-1}$ variables. Briefly, let the case be that $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ are all included in the system defined by the expression $\mathbf{x} \cdot \boldsymbol{B} + \mathbf{y} \cdot \boldsymbol{\Lambda} + \mathbf{z} \cdot \boldsymbol{\Gamma} = \mathbf{U}$, where we have the parameter matrices as well as the matrix of stochastic disturbances $\mathbf{U}$, with $E(\mathbf{U}) = \mathbf{0}$ ($\mathbf{y}$ and $\mathbf{z}$ are endogenous, $\mathbf{x}$ exogenous). This system clearly includes the previous equation, while another component is $\mathbf{y} = \mathbf{x} \cdot \boldsymbol{\pi} + \mathbf{V}$, a reduced form of the complete system where $\boldsymbol{\pi}$ and $\mathbf{V}$ are the corresponding sub-matrices of the parameters of interest. From the data set (sample) A, it is possible to estimate $\boldsymbol{\pi}$ and, hence, predict the values $\hat{\mathbf{y}}^{\text{B}}$

[5] For the sake of brevity, concerning the properties that must hold for defining a generic distance function, we refer to Mardia et al. (1980).

[6] Namely, the matching variables $X_1, \ldots, X_l \ldots, X_{L-1}$ are all included in the model but one.

based on the observed values of the matching variables. In addition, using sample B it is possible to estimate $\mathbf{z}^{\mathrm{B}} = \mathbf{x}^{\mathrm{B}}_{L-1} \cdot \boldsymbol{\beta} + \hat{\mathbf{y}}^{\mathrm{B}} \cdot \boldsymbol{\lambda} + \boldsymbol{\epsilon}$. By rewriting the latter equation as $\mathbf{z}^{\mathrm{B}} = \mathbf{M}^{\mathrm{B}} \cdot \boldsymbol{\delta} + \boldsymbol{\epsilon}$, where $\mathbf{M}^{\mathrm{B}} = (\mathbf{x}^{\mathrm{B}}_{L-1} | \hat{\mathbf{y}}^{\mathrm{B}})$ and $\boldsymbol{\delta} = (\boldsymbol{\beta} | \boldsymbol{\lambda})$, the parameters in $\boldsymbol{\delta}$ can be estimated by ordinary least square, i.e., $\hat{\boldsymbol{\delta}} = (\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{z}$, given that the inverse matrix $(\mathbf{M}'\mathbf{M})^{-1}$ exists. Therefore, the issue is that the rank of $(\mathbf{M}'\mathbf{M})^{-1}$ cannot exceed the number of variables $\mathbf{x}$, implying that at least as many of the matching variables be omitted from $\mathbf{z} = \mathbf{x}_{L-1} \cdot \boldsymbol{\beta} + \mathbf{y} \cdot \boldsymbol{\lambda} + \boldsymbol{\epsilon}$, as the number of $\mathbf{y}$ variables that are included.

Yet on the assessment of matching goodness, limited to categorical matching, Walter (1984) investigated the sampling effort required to obtain matches for all the units in a given sample. Using Markov chains, the author derived the first two moments of the exact distribution of the sample size required to complete the match quotas in all the categories of the chosen matching variable(s). Hence, the dependence of the matching difficulty in relation to samples size, the number of matching categories, as well as the distributions of category probabilities and quotas are considered. The author approached the problem assuming that the sample from the first population is given and by sampling the second population repeatedly (until all units of the first sample have been matched). Walter (1984) demonstrates that (1) for a fixed number of matching categories, larger samples are easier to match than small ones, (2) the mean sample size increases with the number of matching categories and, (3) matching gets easier if the category sampling probabilities are proportional to their quotas.

The latter situation occurs often when the matching variables are distributed similarly and there are only weak confounders. In addition, substantial oversampling is anticipated whenever the category probabilities and the quotas are far from being proportional. The problem is then the required, optimal degree of similarity that must exist between the matching variables in samples A and B in order to carry out an easier matching and not lose precision. In this regard, the case of continuous $\mathbf{x}$ emerges, while it was disregarded so far. In this direction, no further developments were proposed during the '80s: most of the contributions focussed on real data applications, with national departments and federal offices of the U.S. and Canada in the front line (see, for example, Radner et al. 1980; Rodgers and DeVol 1981; Gavin 1985; Armstrong 1989, and the references therein).

Despite these gaps, the practical contributions of the '80s still had a fundamental role in moving forward the whole SM framework which began to be thought, at that time, like a 'file-merging technique' distinct from Record Linkage. For example, Barry (1988) points out that RL is an 'exact matching' method, stressing that it is structured on pseudo-identifiers that allow linking entities from different data sources. In contrast, statistical matching deals with units that are similar but not (necessarily) the 'same'. The main goal of SM was stated such as "integrating data on an individual, from one source, with data on a different observation (from another source) if the two units are identified as the best matching or the most similar units" (Gavin 1985, p. 183). In other words, it became clear that statistical matching contrasts RL "because the set of units in the two files for statistical matching may be completely disjoint or have only a small unknown overlap" (Singh et al. 1988, p. 672). Such considerations spread light on the usefulness of SM for integrating data, particularly when privacy claims constraints do hold. Indeed, consequently to the data privacy concerns and the growing debate on this topic, a progressive shortage/lack of information (e.g., due to the removal of units' unique identifiers) made more complex the linkage among records from different data sources fostering, in turn, the diffusion of new data integration solutions.

To completion of the SM framework, Singh et al. (1988) and Armstrong (1989) analysed alternative approaches to SM, above all, through log-linear modelling imputation. The novelty of the proposal lies in the estimation of the conditional distribution that must be imputed in the categorical framework represented by $f(X^\star, Y^\star, Z^\star)$, where the star-variables are categorical covariates, and the related joint distribution $f(\cdot)$ is a probability mass function. The idea is to transform the classic SM problem related to $f(X, Y, Z)$, to one that involves the categorical variables. After having suitably selected the partitioning of the (X, Y, Z) space into categories for $f(X^\star, Y^\star, Z^\star)$, first, Z is imputed up to a $Z^\star$ category by exploiting $f(Z^\star|X^\star)$ within the imputation class $(X^\star, Z^\star)$, second, a value of Z within the $Z^\star$ category is chosen. The main advantage of such an approach is that CIA violations can be easily controlled in a categorical framework that is 'approximately the same' of $f(X, Y, Z)$. Hence, a subset of X as suitable predictors can be obtained, ending up with optimal imputation classes (as per an instability measure definable on the coarseness of the categorical partitions).

Such a solution is particularly relevant when the integration is oriented towards microsimulation models and there is a need for specific information that has a low probability to occur. In this sense, non-exhaustive examples can be high-income observations, frequent response errors, and/or poor information details. These problems can be addressed by feeding microsimulation modelling with SM imputation such that (1) the computational effort required by data integration can be reduced and, (2) the potential drawbacks from non-linear relationships among X, Y, Z (which could bias the results obtained from the analysis of the integrated data set) may be avoided or mitigated by means of a punctual control of the transformed categorical framework (Cohen 1991).

3.3 Finally, it came the methodology

"Micro-simulation databases which are frequently used by policy analysts and planners, are created by several datafiles that are combined by Statistical Matching" (Singh et al. 1993, p. 59), a method whose development was drastically speeded up by the discussion of the simulation results of three, alternative SM 'techniques': regression-based, distance-based, and log-linear ones. The empirical evidence offered by Singh et al. (1993) suggested that distance-based SM (i.e., the hot deck techniques that are discussed in detail in Sect. 4) performs better than regression-based SM. In contrast, log-linear methods should be preferred if auxiliary information is available and, hence, CIA can be relaxed by adopting categorical constraints. Similar conclusions are drawn by Schulte Nordholt (1998) who compared the results from simulations and real-world applications using Dutch data. Significantly, to date, this work can be considered, in addition to Renssen (1998), the first SM application with data referred to a European country (Netherlands), beyond the original German and French 'data fusion' attempts of the late '80s/first '90s (discussed in Appendix A).

At the dawn of the new millennium, considerations like "statistical matching has been widely used by practitioners without always adequate theoretical underpinnings" (Moriarity and Scheuren 2001, p. 407) and "throughout the world, today, we find synonyms used to describe the Statistical Matching process including 'data fusion', 'data merging' or 'data matching', 'mass imputation', 'microsimulation modelling' and 'file concatenation' (...) with a dragged discussion about a suitable and clarifying denotation" (Rässler 2002, p. 2) were suggesting the urge for more cohesion in SM formalization. The widespread idea behind SM was that it was largely used, mainly for practical purposes, by the OS (Moriarity and Scheuren 2003), as also Rässler (2002) states: "much of the literature describing traditional approaches and techniques are working papers, technical or internal reports" (p. 44).

Two main contributions answered the urge for more solid theoretical foundations: Rässler (2002) and D'Orazio et al. (2006b). They provided a cohesive theoretical framework for the SM methodology, discussing the main implications of the CIA and the use of auxiliary information (D'Orazio et al. 2006b), and comparing several alternatives to SM with a specific focus on Bayesian solutions (Rässler 2002). The latter was developed by Rässler (2003) who adapted and further implemented the framework of Rubin (1987), employing a non-iterative Bayesian alternative to his regression model.

The key challenge in SM became then finding a reliable alternative to the CIA. By approaching the SM problem as a non-response issue, the core idea embraced the fundamental 'identification problem'. Whereas the missing mechanism is ignorable, the association of the variables which are not *jointly* observed is not identifiable and, hence, it cannot be likelihood-estimated. Therefore, either there is additional information on $f(X, Y, Z)$, or the researcher must resort to several imputations, eventually based on informative priors. In such a context, Rässler (2004) proposed to frame the identification problem according to four levels of validity that SM may achieve. Namely, they are: 1st level—Preserving the individual values; 2nd level—Preserving the joint distributions; 3rd level—Preserving the correlation structures; 4th level—Preserving the marginal distributions. Usually, the latter level is the one that can be widely controlled in SM. If the conditional association (i.e., the one of the variables not jointly observed, given the variables in common between A and B) cannot be estimated from the data at hand, admissible values for the unconditional association of Y and Z can be estimated instead. How? Depending on the explanatory power of the matching variables, smaller/wider range of admissible values can be estimated (Rässler 2004).

4 The non-parametric and Bayesian approaches

During the last two decades, non-parametric SM gained relevant attention due to the fact that (1) it exploits, entirely, the 'live', observed information (D'Orazio et al. 2006b), (2) it reduces the possible model misspecification bias deriving from the assumption(s) made on the parameters of the joint family distribution $f(X, Y, Z)$ (Conti et al. 2017b) and, (3) it decreases the computational effort required by parametric SM (D'Orazio 2015). Even though non-parametric techniques require no ‘assumption’, it should be noticed that their application undergoes (1) the choice of the distance function to apply, (2) whether (and how) to build donation classes or not and, (3) the sampling mechanism for the selection of donors.

The methodological advances conveyed by the non-parametric SM are related to the concept of ‘matching noise’, its definition, quantification, and to the role that it plays in the integration procedure. The starting point is the joint distribution $f(X, Y, Z)$ obtained after statistical matching which may not coincide with the ‘true’ (unobservable) distribution. Hence, “the imputed data set is not a real data set and the statistical conclusions drawn from it are questionable” (Marella et al. 2008, p. 1593). Whenever the two distributions differ, there is matching noise and researchers must aim at minimizing it.

We have two data sets: $\mathrm{A} = \left\{ \underset{n_\mathrm{A}\times L}{\mathbf{x}^\mathrm{A}} \right\}$ and $\mathrm{B} = \left\{ \underset{n_\mathrm{B}\times L}{\mathbf{x}^\mathrm{B}}, \underset{n_\mathrm{B}\times P}{\mathbf{z}^\mathrm{B}} \right\}$. Let the case be that we want to build the synthetic (complete) data set.

$\mathrm{C} = \left\{ \underset{n_\mathrm{A}\times L}{\mathbf{x}^\mathrm{A}}, \underset{n_\mathrm{A}\times S_d}{\mathbf{z}^\mathrm{A}} \right\}$. Marella et al. (2008) pointed out that this generated data set will result from the distribution $f(\mathbf{x}^\mathrm{A}, \mathbf{z}^\mathrm{A}) = \int f(\boldsymbol{x}^\mathrm{B}_{\tilde{j}}|\boldsymbol{x}^\mathrm{A}) f(z^\mathrm{A}|\boldsymbol{x}^\mathrm{B}_{\tilde{j}})\mathrm{d}\boldsymbol{x}^\mathrm{B}_{\tilde{j}}$, where $\boldsymbol{x}^\mathrm{B}_{\tilde{j}}$ are the r.v.s observed for the donor units matched with the recipient ones (i.e., the $\tilde{j}$-th donor that has been matched with the i-th recipient), while z^A are the r.v.s imputed in the recipient data set based on the matched units’ pairs. It follows that the matching noise is a composite element of the donor distribution $f(\boldsymbol{x}^\mathrm{B}_{\tilde{j}}|\boldsymbol{x}^\mathrm{A})$ and the values of the imputed variables observed for the matched donor.

Conti et al. (2008) compared the performances (in terms of matching noise minimization) that are obtained from different non-parametric imputation strategies: hot deck techniques, k-Nearest Neighbour method (kNN), and local linear regression (when the assumption of linearity for the underlying population regression function is not mandatory). In addition, in previous works, the authors considered, specifically, the kNN method (Marella et al. 2008), evaluating the matching noise produced by imputation with both a fixed and variable number of donors. In the former case, the class of imputation procedures that includes distance-based and random hot deck techniques is defined by assuming that the k donors to a unit $i \in \mathrm{A}$ are given by the k nearest neighbours of $\mathbf{x}_i$ in B (with $i = 1, \ldots, n_\mathrm{A}$). Let d be the Euclidean distance such that $d(\mathbf{x}^\mathrm{A}_i, \mathbf{x}^\mathrm{B}_j) = \left[(\mathbf{x}^\mathrm{B}_j - \mathbf{x}^\mathrm{A}_i)'\mathbf{D}(\mathbf{x}^\mathrm{B}_j - \mathbf{x}^\mathrm{A}_i) \right]^{1/2}$, with $\mathbf{D}$ being a positive definite matrix. The k nearest neighbours of $\mathbf{x}^\mathrm{A}_i$ are the $k \geq 1$ observations in B which result to be the closest to $\mathbf{x}^\mathrm{A}_i$ according to d, i.e., the observations $\mathbf{x}^\mathrm{B}_{j(i)} = (\mathbf{x}^\mathrm{B}_{j_1(i)}, \ldots, \mathbf{x}^\mathrm{B}_{j_k(i)})$. With a number of fixed donors, it happens that some donors could be sparse and hence, the kNN method brings observations which are far from $\mathbf{x}^\mathrm{A}_i$ to be equally informative on $\mathbf{z}^\mathrm{A}_i$. The authors suggest that it is fruitful that the optimal value of k varies with $\mathbf{x}^\mathrm{A}_i$ to allow a different number of donors k to be matched with each $\mathbf{x}^\mathrm{A}_i$. This is done by fixing a threshold such that the observations which have a distance $d(\mathbf{x}^\mathrm{A}_i, \mathbf{x}^\mathrm{B}_j)$ smaller than the threshold are selected to be neighbours of $\mathbf{x}^\mathrm{A}_i$.

The simulation study results of Marella et al. (2008) hint at using (large) donor data sets with a variable number of k donors, possibly adjusting the mean imputation with residuals. In Conti et al. (2008), to evaluate the closeness between the data-generating model and the imputation-generating model, the authors propose a simulation study elaborating a Kolmogorov–Smirnov distance-based measure of divergence. Results show that kNN performs the worst when there are fixed k donors, underestimating variability since $f(z^{\mathrm{A}}|\boldsymbol{x}^{\mathrm{B}}_{\tilde{j}})\mathrm{d}\boldsymbol{x}^{\mathrm{B}}_{\tilde{j}}$ is condensed on the expectation of $\mathbf{z}|\mathbf{x}$. Moreover, the authors suggest preferring local linear regression estimators when a complex functional relationship holds between the variables.

As per the non-parametric SM, the approach proposed by Rässler (2002) represents a turning point for the suitable 'alternatives' to traditional SM. Indeed, the author innovated the SM framework by embedding it in Multiple Imputation, by proposing the transfer of information through Bayesian inference, while the results are validated in a frequentist way. Trivially, Rässler's starting point was the need for providing public use files for end-users by integrating two or more data sources. She stressed that the 'public use file' is characterized by the fact that the matched data are passed forward to others, usually outside the OS. Therefore, file users/data analysts often differ from the user who made the imputation. This problem poses a classical imputation challenge that, the authors says, cannot be solved by weighting, calibration, or the EM algorithm (Rässler 2002). Indeed, the SM problem cannot be handled by observed-data likelihood nor by the EM algorithm without making explicit assumption about the variables which are never jointly observed. Due to the inestimability of certain parameters (whenever the underlying model cannot be specified by the data at hand), SM poses a problem of identification, i.e., there are several feasible associations potentially describing the joint distribution of the variables not jointly observed.

At its core, the identification problem treated by Rässler (2002) can be framed as follows. Let n be a sample of individuals. To them, a question is asked and n_0 represents the number of people refusing to answer. We are interested in an outcome variable W taking values 0, 1. Let p be the proportion of the $n_1 = n - n_0$ individuals for whom we observe $W = 1$. Aiming to estimate $\mathrm{P}(W = 1)$, often we resort just to p. Consequently, the unobserved outcomes of W have the same distribution of the observed ones. But, if we consider p only as a good estimate of $\mathrm{P}(W = 1|R = 1)$, where $R = 1$ indicates that the outcome variable W has been observed, while $R = 0$ indicates that a person decided not to answer, we have that $\mathrm{P}(W = 1) = \mathrm{P}(W = 1, R = 1) + \mathrm{P}(W = 1, R = 0) = \mathrm{P}(W = 1|R = 1)\mathrm{P}(R = 1) + \mathrm{P}(W = 1|R = 0)\mathrm{P}(R = 0)$. Of course, $\mathrm{P}(W = 1|R = 0)$ is not known and, by using p as an estimate for $\mathrm{P}(W = 1)$, we are assuming that $\mathrm{P}(W = 1|R = 0)$ is also estimated by p. However, what is known is just that $\mathrm{P}(W = 1|R = 0)$ lies in [0, 1] and, thus, the lower and upper bounds of $\mathrm{P}(W = 1)$ can be estimated by $\mathrm{P}(W = 1) \leq \mathrm{P}(W = 1|R = 1)\mathrm{P}(R = 1) + \mathrm{P}(R = 0)$ and $\mathrm{P}(W = 1) \geq \mathrm{P}(W = 1|R = 1)\mathrm{P}(R = 1)$. $\mathrm{P}(R = 1)$ and $\mathrm{P}(R = 0)$ can be estimated by n_1/n and n_0/n, respectively, thus the bounds are estimated by $p\frac{n_1}{n} \leq \hat{\mathrm{P}}(W = 1) \leq p\frac{n_1}{n} + \frac{n_0}{n}$. This concept of 'identification' is further developed by Rässler (2002) who uses MI to estimates upper and lower bounds of the unconditional association.

The solution proposed is based on Bayesian inference and the data augmentation algorithm. A probability model for the observed data is specified given the vector of unknown parameters $\boldsymbol{\theta} \in \Theta$. Then, θ is treated as a random variable with a certain prior distribution, and inference about it is summarized by its posterior, given the data at hand. Rässler (2002) proved that the joint distribution likelihood receives contributions from both the observed data and the prior. Moreover, when data are unobserved, the prior (predictive) distribution does not condition on previous observations. The identification problem is then replaced using prior information and MI procedures. CIA is overcome by MI techniques using informative priors. Different prior settings on the conditional associations allow us to show the sensitivity of the unconditional association, as per the common variables occur in determining it.

The embryonic framework proposed by Rässler (2002) was further extended by Rässler (2003) through the Non-Iterative Bayesian Approach to Statistical Matching (NIBAS). No distributional assumptions are made on $\mathbf{x}$, while the only requirement is that the matching variables can serve as predictor matrices in a linear regression model. Due to the particular structure of missingness characterising SM, it is possible to define both a data model and a prior distribution and, consequently, derive the observed data posterior from them. By means of MI procedures, prior information is used for imputing missing data and, from the imputed data, lower and upper bounds can be estimated to achieve a range of values of the unconditional association parameters. Such a range serves as a quality measure for SM.

Bayesian solutions to the identifiability problem of SM were fundamental for developing the method because they made explicit that whenever two variables are not (or, better, they are never) jointly observed, the related conditional association parameters cannot be estimated by likelihood inference. In contrast, the nearest neighbour solutions proposed and applied over the years are often undermined by the fact that conditional independence is produced *de facto*, even if it is not assumed. To overcome this drawback and its consequences, NIBAS assumes (at least) univariate normality for Y and Z, while $f(\mathbf{y}, \mathbf{z}|\mathbf{x})$ is assumed to be multivariate normal. Then, NIBAS assumes independence between the regression parameters of the general linear models for data sets A and B and the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{y},\mathbf{z}|\mathbf{x}}$ and, with a suitable non-informative prior, the observed-data posterior distribution and the conditional predictive distributions can be derived. From the latter, random draws for the parameters as well as the imputed $\mathbf{y}$ and $\mathbf{z}$ can be obtained.

Aiming to evaluate the predictive power of the matching variables $\mathbf{x}$, by employing simulations, Rässler (2004) demonstrated that the Bayesian approach offers a relevant advantage: whereas regression imputation ends up with estimates of the true population correlation that are not unbiased (not even asymptotically), Bayesian SM allows us to preserve the prior values of the conditional correlation, outperforming all the other approaches. This hints at the fact that, when auxiliary data are at disposal and prior information must be used, the Bayesian multiple imputation procedure proposed by the author is the best choice at hand.

5 Uncertainty: old issue, new challenges

5.1 A matter of constraints

The first half of the '00s saw the rise of the 'third way' to solve the identification problem[7] in SM. Usually, this had been addressed by means of specific modelling on (Y, Z) or, auxiliary information on $f(Y, Z)$. Different approaches to the problem were named 'uncertainty analysis', 'partial identification', or 'lower and upper probabilities study'. All these contributed to raise knowledge about the fact that the main goal of SM (at least from a macro point of view) had to be the estimation of the range of potential values identifying the unidentifiable parameters, consistently with the estimable ones (Di Zio and Vantaggi 2017). In other words, the focus must have been the reduction of the uncertainty about the association parameters of the variables never jointly observed, by means of the common variables.

D'Orazio et al. (2006a) state that, even if there could be complete knowledge of the distributions $f(X, Y)$ and $f(X, Z)$, it is not possible to conclude anything about $f(X, Y, Z)$ merely due to the fact that the joint distribution can be predicted only if there is a deterministic relationship between the two bivariate distributions. Considering the cross-tabulation approach in relation to the variables X, Y, and Z (a rather common practice of the '70s), let $(X^\star, Y^\star, Z^\star)$ be a triplet with number of categories F, G, H, respectively, such that the table's cells are definable as $\iota = \{(f, g, h) : f = 1, \ldots, F; g = 1, \ldots, G; h = 1, \ldots, H\}$. Therefore, the joint distribution $f(X^\star, Y^\star, Z^\star)$ is multinomial, unknown, and defined by $\theta_{fgh} = \mathrm{P}(X = f, Y = g, Z = h)$ for $f = 1, \ldots, F$, $g = 1, \ldots, G$, $h = 1, \ldots, H$. Thus, the true, unknown vector of parameters $\boldsymbol{\theta}^*_{fgh}$ define the distribution. D'Orazio et al. (2006a) state that this vector is *totally uncertain* but, by assuming complete knowledge on the marginal distributions of the pairs $(X^\star, Y^\star)$, $(X^\star, Z^\star)$, it can be restricted. The parameter θ^*_{fgh} lies in the interval defined by lower and upper limits such that all the plausible values for it determine a density function (made by the frequencies of all $\boldsymbol{\theta}^*_{fgh}$). By resorting to the MLEs for $\hat{\theta}_{fg.}$ and $\hat{\theta}_{f.h}$, it can be proved that suitable constraints help ruling out illogical values from Θ.

By approaching the problem from the point of view of the ecological inference, Conti et al. (2013) estimate the joint distribution of ordered categorical variables $f(X^\star, Y^\star, Z^\star)$ starting from a contingency table where the population counts provide the marginals. If the rows and columns counts arranging the table come from different samples (A and B), the problem is purely how to estimate the joint distribution function, i.e., a macro SM issue. The proposed solution is to estimate a class of possible distributions for $(X^\star, Y^\star, Z^\star)$, identifying a measure of uncertainty for the estimated model. The uncertainty is defined using the upper and lower bounds of the cells counts, with conditional and unconditional measures of uncertainty eventually restrained by means of structural zeros constraints. In SM, these are frequently used constraints for the parameters when the r.v.s of interest are categorical. Such constraints consist of defining $\theta_{fgh} = 0$ for some (f, g, h) (Agresti 2013). A structural zero occurs when (1) at least one pair of categories in (f, g, h) is not compatible or,

[7] This is framed in Sect. 4, as per the definition of Manski (1995) and Rässler (2002).

(2) each pair in (f, g, h) is plausible but the triplet is not compatible. Such a constraint is useful for integration purposes since its main effect is to potentially reduce the likelihood ridge to a unique distribution. How? When the goal is to restrict Θ to a subspace $\Omega \subset \Theta$ (closed and convex), we have to find the set of $\boldsymbol{\theta} \in \Omega$ such that the likelihood function $L(\boldsymbol{\theta}|\mathrm{A} \cup \mathrm{B})$ is maximized. Having $\mathcal{P}$ parameter subsets, when the case is that $\Omega \bigcap \mathcal{P}_{\hat{\boldsymbol{\theta}}_{\mathbf{x}} \hat{\boldsymbol{\theta}}_{\mathbf{y}|\mathbf{x}} \hat{\boldsymbol{\theta}}_{\mathbf{z}|\mathbf{x}}} \neq \emptyset$, i.e., the subspace has a non-empty intersection with the unconstrained likelihood ridge, structural zeros can be so informative that, e.g., $\Omega \bigcap \mathcal{P}_{\hat{\boldsymbol{\theta}}_{\mathbf{x}} \hat{\boldsymbol{\theta}}_{\mathbf{y}|\mathbf{x}} \hat{\boldsymbol{\theta}}_{\mathbf{z}|\mathbf{x}}} = \hat{\boldsymbol{\theta}}$. For example, by defining $(G-1)(H-1)$ independent structural zero constraints for each $X = f, f = 1, \ldots, F$ is sufficient for a unique ML estimate. In such a context, the simulation study results of Conti et al. (2013) show that the uncertainty reduction is directly proportional to the reduction of the support of the conditional distribution of $Y^\star$ and $Z^\star$ given $X^\star$. In addition, the uncertainty largely depends on the informativeness of the structural zero constraint.

The class of possible distributions for $(X^\star, Y^\star, Z^\star)$ (being these variables categorical, but Conti et al., 2017b considered continuous Z and Y, and discrete X, as discussed later here) is estimable by means of the so-called Fréchet bounds (or 'uncertainty class') (D'Orazio et al. 2017). Indeed, the latter allows us identifying the plausible lower and upper bounds for the parameters which must be estimated to define the marginal distributions $(Z|X)$ and $(Y|X)$. Namely, the cell frequencies θ_{yz} of the (Y, Z) contingency table, given the estimates $\hat{\theta}_{y|x}$ from A, $\hat{\theta}_{z|x}$ from B, and $\hat{\theta}_x$ from $A \cup B$ can be obtained by means of the class identified by

$$\max\{0; \hat{\theta}_{y|x} + \hat{\theta}_{z|x} - 1\} \leq \hat{\theta}_{yz|x} \leq \min\{\hat{\theta}_{y|x}; \hat{\theta}_{z|x}\}. \tag{1}$$

By means of this uncertainty class, we can evaluate the uncertainty in SM but we can, also, proceed in validating the whole integration. Indeed, Rässler (2002) proposed to evaluate the length of such class for the unidentifiable parameters in the normal multivariate case to finally define a measure of the reliability of the estimates under CIA. The author's results hint at the fact that, when short uncertainty classes do hold, the parameter estimates obtained by different models slightly differ from the ones obtained under CIA. In addition, a measure of uncertainty is defined by Rässler (2002) as $\frac{1}{K} \sum \hat{\theta}_k^{(\mathrm{U})} - \hat{\theta}_k^{(\mathrm{L})}$, where θ_k with $k = 1, \ldots, K$ are the unidentifiable parameters in a parametric model for (X, Y, Z), while $\hat{\theta}_k^{(\mathrm{U})}$, $\hat{\theta}_k^{(\mathrm{L})}$ are the estimated upper and lower bounds of the uncertainty class defined on these parameters.

The intuition was further developed by Conti et al. (2017b) who proposed a measure of uncertainty and studied its properties in the specific non-parametric context. From the parametric point of view, SM uncertainty is quantifiable in terms of the estimates range of the unidentifiable parameters. In contrast, non-parametrically speaking, such a measure relates to the 'intrinsic' association between the pair of variables (Y, Z). By using the Fréchet bounds as a starting point, a measure of uncertainty is given by the suitable functional that quantifies the length of the uncertainty class. Let $dF(x, y, z) = dQ(x)\, dS(y, z|x)$ be the joint distribution function of three r.v.s (X, Y, Z), where $Q(x)$ is the marginal distribution function of X, while $S(y, z|x)$ is the distribution function of (Y, Z) given X. The latter is a discrete matching variable, while Y and Z are continuous. Conditionally on X, the set of plausible models (or, in other words, the Fréchet class) of all distribution functions $S(y, z|x)$ can be obtained, in such a way that it is compatible with the univariate distribution functions $G(y|x)$ and $H(z|x)$. Let us consider L and U, the lower and upper bounds where $\mathrm{L}\,[G(y|x), H(z|x)] = \max\,[H(z|x) + G(y|x) - 1,\ 0]$ and $\mathrm{U}\,[G(y|x), H(z|x)] = \min\,[H(z|x), G(y|x)]$ (defined analogously in Eq. 1). Then, let us consider, for every (y, z), the inequalities $\mathrm{L}\,[G(y|x), H(z|x)] \leq S(y, z|x) \leq \mathrm{U}\,[G(y|x), H(z|x)]$. If this pair of inequalities does hold, where the bounds L and U are joint distributions functions with margins $G(y|x)$ and $H(z|x)$, the Fréchet class of these two distributions is defined as follows

$$\mathcal{S} = \{S(y, z|x) : \mathrm{L}\,[G(y|x), H(z|x)] \leq S(y, z|x) \leq \mathrm{U}\,[G(y|x), H(z|x)]\}\,. \quad (2)$$

Therefore, the set of distribution functions $\mathcal{S}$ defines the uncertainty class in the non-parametric SM framework. Taking the expectation with respect to the distribution of X, the unconditional Fréchet class can be defined as

$$\mathcal{S} = \{S(y, z) : E\,[\mathrm{L}\,(G(y|x), H(z|x))] \leq S(y, z) \leq E\,[\mathrm{U}\,(G(y|x), H(z|x))]\}\,. \quad (3)$$

Clearly, the uncertainty class in Eq. 3 does not take advantage of the common variables observed between A and B.

Being each category of X observed in A and B, the estimator of the Fréchet class can be obtained by re-writing Eq. 1 as follows

$$\left\{\max\left[\hat{H}_{n_\mathrm{B}}(z|x) + \hat{G}_{n_\mathrm{A}}(y|x) - 1, 0\right],\ \min\left[\hat{H}_{n_\mathrm{B}}(z|x), \hat{G}_{n_\mathrm{A}}(y|x)\right]\right\}. \quad (4)$$

In addition, the unconditional Fréchet bounds are estimated by

$$\left\{\sum_x \hat{p}(x)\,\max\left[\hat{H}_{n_\mathrm{B}}(z|x) + \hat{G}_{n_\mathrm{A}}(y|x) - 1,\ 0\right],\right. \\ \left.\sum_x \hat{p}(x)\,\min\left[\hat{H}_{n_\mathrm{B}}(z|x), \hat{G}_{n_\mathrm{A}}(y|x)\right]\right\}, \quad (5)$$

where $\hat{p}(x) = \left(\frac{n_{\mathrm{A},x}+n_{\mathrm{B},x}}{n_\mathrm{A}+n_\mathrm{B}}\right)$ is an estimate of $\mathrm{P}(X = x)$,[8]

Conti et al. (2017b) built a confidence region for the estimator of the Fréchet class depicted in Eq. 4 and, from it, by exploiting the Kolmogorov–Smirnov (KS) statistic, they set the confidence bands for $G(y|x)$ and $H(z|x)$, which are given by

$$\mathcal{G}_{n_\mathrm{A},x} = \left(\hat{G}_{n_\mathrm{A}}(y|x) - \frac{k_\alpha}{\sqrt{n_{\mathrm{A},x}}},\ \hat{G}_{n_\mathrm{A}}(y|x) + \frac{k_\alpha}{\sqrt{n_{\mathrm{A},x}}};\ y \in \mathbb{R}\right), \\ \mathcal{H}_{n_\mathrm{B},x} = \left(\hat{H}_{n_\mathrm{B}}(z|x) - \frac{k_\alpha}{\sqrt{n_{\mathrm{B},x}}},\ \hat{H}_{n_\mathrm{B}}(z|x) + \frac{k_\alpha}{\sqrt{n_{\mathrm{B},x}}};\ z \in \mathbb{R}\right), \quad (6)$$

[8] $n_{\mathrm{A},x}$ and $n_{\mathrm{B},x}$ are defined as $n_{\mathrm{A}x} = \sum_{i=1}^{n_\mathrm{A}} \mathrm{I}(X_i = x)$ and $n_{\mathrm{B},x} = \sum_{i=1}^{n_\mathrm{B}} \mathrm{I}(X_i = x)$, respectively, with $\mathrm{I}(x \in \mathrm{D})$ being an indicator function of the set D. It is equal to 1 if $x \in \mathrm{D}$, 0 otherwise.

respectively, with k_α being the $1-\alpha$ quantile of the KS distribution. By defining

$$\begin{aligned} \underline{\hat{\mathcal{S}}}(y,z|x) &= \max\left\{\hat{H}_{n_{\mathrm{B}}}(z|x) - \frac{k_\alpha}{\sqrt{n_{\mathrm{B},x}}} + \hat{G}_{n_{\mathrm{A}}}(y|x) - \frac{k_\alpha}{\sqrt{n_{\mathrm{A},x}}} - 1,\ 0\right\}, \\ \hat{\overline{\mathcal{S}}}(y,z|x) &= \min\left\{\hat{H}_{n_{\mathrm{B}}}(z|x) + \frac{k_\alpha}{\sqrt{n_{\mathrm{B},x}}},\ \hat{G}_{n_{\mathrm{A}}}(y|x) + \frac{k_\alpha}{\sqrt{n_{\mathrm{A},x}}}\right\}, \end{aligned} \tag{7}$$

we end up with a confidence region for the Fréchet class given by $\mathcal{S}_n^x = \left\{S(y,z|x) : \underline{\hat{\mathcal{S}}}(y,z|x) \le \mathcal{S}(y,z|x) \le \hat{\overline{\mathcal{S}}}(y,z|x)\right\}$. The measure of pointwise uncertainty is given by the interval $\{\mathrm{L}\left[G(y|x), H(z|x)\right],\ \mathrm{U}\left[G(y|x), H(z|x)\right]\}$ in terms of its length (i.e., $\mathrm{U}-\mathrm{L}$). Also, Conti et al. (2017b) summarized the pointwise measures of uncertainty (due to the fact that we have one measure for every triple x, y, z) into the unique measure of average length. Indeed, they define a weight function on $\mathbb{R}^3$, $T(x,y,z)$, and compute $\int_{\mathbb{R}^3}\{\mathrm{U}\left[G(y|x), H(z|x)\right] - \mathrm{L}\left[G(y|x), H(z|x)\right]\}\, dT(x,y,z)$. Therefore, taking $dT(x,y,z) = dQ(x)\; d\left[G(y|x), H(z|x)\right]$, the overall measure is given by

$$\int_{\mathbb{R}}\left\{\int_{\mathbb{R}^2}\{\mathrm{U}\left[G(y|x), H(z|x)\right] - \mathrm{L}\left[G(y|x), H(z|x)\right]\}\, d\left[G(y|x), H(z|x)\right]\right\} dQ(x).$$

The main finding related to such an uncertainty measure is that this intrinsic uncertainty (when no external auxiliary information is available) does not depend on neither the support nor the marginal distribution of $f(Y,Z)$, i.e., $G(y|x)$, $H(z|x)$. Indeed, independently of the sample data, Conti et al. (2017b) proved that the maximal uncertainty is 1/6. In contrast, such an uncertainty can be reduced when auxiliary information is available by imposing logical constraints.

Establishing boundaries for the illogical occurrences in the set of parameters finds practical relevance in real data applications. There are at least two reliable cases in which, either the *existence* of some information is doubtable, or it is marked by *inequality*. For example, the occurrence of an eight years old employee clearly belongs to the first case. For the second case, let's consider the probability of being a casual worker with a diploma to be higher than the probability of a manager without a degree. If "logical constraints naturally arise from applications" (Vantaggi 2008, p.710), constraints must be properly managed. Indeed, by re-adapting the probability theory of de Finetti (1974) and Vantaggi (2008) proposed to exploit coherent conditional probability for combining data from different sources without necessarily uptake strong assumptions on the relationships of (X, Y, Z). Furthermore, logical constraints can be considered but, when they are not present, the author proves that the conditional assessment can still be coherent even if we have to assume conditional independence. To date, in the case of same population and, same sample scheme, the proposal of Vantaggi (2008) is the only one that exploits the coherent conditional probability for integrating data (she considered both the case of two sources and multiple ones).

While Vantaggi (2008) proposed a setting for incoherences reduction based on MLEs, further developments towards the reduction of incoherences based on distance minimization are proposed by Brozzi et al. (2012). They suggested using specific adjustments which, by targeting weighted localization of parameters sub-domains from which the incoherences must be removed, prove to perform better than the originally coherent assessment of Vantaggi (2008).

A peculiar issue is tackled by Di Zio and Vantaggi (2017) in relation to the partial identification problem when the matching variables are misclassified. By disregarding the effect of the sampling variability of the estimates on the identification regions, the authors evaluate different scenarios of misclassification. By dealing with categorical variables, they describe the partially identifiable region (i.e., the class of probabilities which extend the conditional probabilities obtained by the information available in different sources) by means of lower and upper bounds on the consistent probabilities. When the common variable(s) is(are) misreported in only one of the two data sets that the researchers want to match, the authors demonstrate that the potential consistency of the distributions increases due to assumptions on the misclassification mechanism. In other words, it is possible to refine the identifiable region by means of such assumptions about the matching variables misclassification.

How much the integration uncertainty affects the quality of the synthetic (complete) data set is investigated by Conti et al. (2016, 2017a) by taking into account a stratified sampling design. In the first work, the authors propose a specific measure of the 'error' introduced by matching, estimating the distribution function for the variables not jointly observed as well as the corresponding measure of error (upper bounds of which are also introduced). If a class of plausible distributions for (X, Y, Z), conditional or unconditional on the matching variable, can be identified, the size of this class defines the measure of uncertainty. The authors prove that the difference between the admissible distributions in such class (that is a constrained Fréchet class) and the chosen matching distribution, basically gives the error of the matching procedure. The latter is estimable using iterative proportional fitting, offering the maximal error that can occur in choosing a distribution from such a class, i.e., by drawing a surrogate of the true but unknown $f(X, Y, Z)$.

5.2 Sampling frameworks and Big Data

Mainly, the different integration approaches discussed so far considered probability samples. Nevertheless, due to (1) the increasing rates of non-response, (2) the actual costs for data collection and, (3) the potentialities offered by Big Data, the trade-off between data quality and resources needed hinted at investigating other opportunities, e.g., non-probability samples which represent, to date, the most profitable solution for data integration (Lohr and Raghunathan 2017). Relevant examples are web surveys, social media data, mobile phone records, and web crawling software data. Rivers (2007) considered web surveys data proposing a nearest neighbour mass imputation approach that trains a predictive model of Y given X on the non-probability sample (e.g., a web panel) and uses it to predict the distribution $f(Y|X)$ for the probability sample, i.e., a conventional random sample from a population frame. This idea aims to tackle the non-response problems in probability-based surveys: individuals selected from the sampling frame (that covers the target population and contains some auxiliary variables) do not have to directly answer the questionnaire. Instead, they are allocated to a panel (that, also, contains the aforementioned set of auxiliary variables) that mimics the selected people who are then asked to complete the questionnaire. This sort of ‘sample matching’ is further investigated by Bethlehem (2016) who explores the conditions under which it works in the most efficient way. The author points out that such an imputation approach depends on the capacity of the auxiliary variables in explaining the participation behaviour completely: the non-response bias removal is higher as far as such capacity holds.

Alternatively, the ‘propensity to respond’ as a function of the covariates $\mathbf{x}$ for the non-probability sample is estimated and thus used to weight the non-probability data. By adapting the approach of Lee (2006) and Castro-Martín et al. (2022) estimate the individual propensity to participate in the non-probability sample by considering the hypothetical scenario of how would the sample have been if a probability sampling design was used to draw it. Selection bias reduction benefits from the training method proposed by the authors, offering “more importance in the prediction to the individuals who are more likely to appear in the population” (Castro-Martín et al. 2022, p. 17). Residual limitations lie in the fact that wider replicability of the results is envisaged (different data sets, more scenarios, etc.), additional prediction algorithms could be considered, and theoretical properties must be further developed.

A relevant challenge is represented by the fact that by combining Big Data from different sources (e.g., by incorporating large survey data) the promising matching-based imputation is essentially based on the MAR assumption. For example, this happens in Chen et al. (2020a) who propose a weighting adjustment based on parametric model assumptions on the selection mechanism of the non-probability sample, further extended by Chen et al. (2020b) to the non-parametric framework. Kim et al. (2021) go beyond MAR by proposing a sampling mechanism for Big Data that allows us to consider systematic differences among the samples even after having adjusted for the covariates. The probability sample is then used to estimate the missing data, correcting for the under-coverage bias of Big Data (that is considered an incomplete sampling frame for the finite population).

6 Conclusions (and the world beyond)

Recently, Statistical Matching has been (re)gaining attention within the OS, due to the role played by Big Data but, also, because of the increasing necessity of data providers for producing more detailed and punctual information, at the quickest time (de Waal 2015). If Multiple Imputation can be used when the missing information is partially present in a single data set, while probabilistic Record Linkage deals with the absence/misreporting of unique identifiers for the units observed in different files (which, in turn, must not be subjected to incompleteness), statistical matching (that is closely related to these methods) offers the possibility to deal with variables that are never jointly observed in two or more data sets. This feature is a strength for addressing many practical challenges in a world that is more and more characterised by several potential sources and tools for data collection and information sharing.

A relevant proposal aiming to shrink the gap between RL and SM is offered by Gessendorfer et al. (2018) who use SM as a supplement for RL when, dealing with non-consenter individuals observed in e.g., ad hoc surveys, the information collected on them in the administrative data cannot be aggregated with that of the surveys. In such a peculiar context of missing information, the proposal is to use SM to provide the values of the variables for the individuals who refused to give their consent for the linkage and, hence, for integrating the information that could not be integrated otherwise. However, the authors stressed that matching the non-consenter individuals previous to linking the observations does provide conflicting results hinting at problems which can be potentially worse than just ignoring the lack of consent.

Considering high-dimensional problems, Ahfock et al. (2016) deal with multivariate $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ aiming to identify the parameters characterizing the joint distribution function $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$. They propose to draw values from the identified set of parameters, such that the range of sampled values offers a measure of uncertainty of the partially identified parameters (i.e., the ones requiring a joint observation of $\mathbf{y}$ and $\mathbf{z}$). The solution proposed consists of a Gibbs sampler-based approach for estimating a set of positive-defined completions of a partially specified covariance matrix and it is a generalizable exit strategy for real-world data problems involving multivariate normal, skewed-normal, and normal mixture models. Comparing the results from both a simulation study and real data with those generated by a Bayesian approach, Ahfock et al. (2016) offer proofs that their frequentist sampling method largely outperforms the Bayesian one in providing correlation estimates in the neighbourhood of the true, observed one. In addition, the method shows flexibility and remarkable computational speed.

Beyond the role potentially played by Big Data in the integration context, the near future of SM is linked to different theoretical challenges. While parametric SM has been extensively analysed, non-parametric SM left unsolved some challenges which are related, for example, to the optimality of the distance function to be used with distance-based hot deck techniques, or to the 'size' of the donation classes, and the discussion on how much these elements may affect the variance estimates in the synthetic (complete) data set.

Another pending issue is related to the use of survey weights and the sampling design used to build the different data sets at hand. Marella and Pfeffermann (2019) recently proposed a solution for combining the information when this is collected by different samples. Under informative sampling designs, the uncertainty of SM results is compared to the one generated by matching under a 'blind' CIA, or, in other words, by ignoring the informative sampling mechanisms. The simulation study proves that ignoring the sample selection process and its effects, the predictions on X, as well as those on Y and Z are negatively impacted. Hence, the synthetic (complete) data set generated differs from the underlying population distribution of $f(X, Y, Z)$, thus producing bias. This can happen even if the estimates generated by ignoring the sample selection effects may show a smaller variance.

Conti et al. (2019) stress that SM applications are not very common because data are obtained by means of different complex survey designs which, in turn, prevent the straightforward reconciliation of information. However, the authors suggest that, if ecological inference made effective the drawing of conclusions at the individual level starting from aggregated data, SM, also, could be re-directed for drawing inference by using the matching variables which can be thought of as a sort of 'grouping'.

Considering that no solution has been shared between these two fields, the authors hint at exploring this possibility. In addition, further developments could target the hidden incompleteness of the measure of uncertainty that the authors proposed. Indeed, if their proposal is very useful to capture the 'uncertainty of the data sets', it somehow lacks of measuring the 'quality of the matching'. An indicator that measures such a SM quality may be relevant for future research.

The inclusion of the sampling weights in the matching procedure and, more generally, the considerations related to the characteristics of the samples to be aggregated is of particular relevance since only two other works treated such challenges: Rubin (1986) and Renssen (1998). The former proposed to compute new sampling weights from the units produced by the A ∪ B *supersample*. This idea found scarce applicability, in practice, because the inclusion probabilities in the A sample, under the sampling design of B are not known. The latter proposed to calibrate the actual weights of the distinct A and B samples to the common information and, hence, obtain distributions that are compatible with the marginals (Y, X), (X, Z). However, D'Orazio (2009) demonstrated that the two proposals lead to very similar results.

Practically speaking, the main future improvement to take into account is related to auxiliary information. Which kind of auxiliary variables have to be used, in the most efficient way, for obtaining sufficiently accurate statistics from the integrated data? Which kind of information exploited from an additional source can be more proficiently used in integrating data? The knowledge about population totals or the knowledge about the relationship(s) of the variables at hand? Furthermore, would it be beneficial to recursively conduct ad hoc surveys to obtain information on a subset of the variables of interest from different data sources? How, then, would be possible to assess the quality of the inference drawn from the integrated data, when the latter is not available in complete form, in practice? In other words, how to assess this quality when we are the users carrying out the integration? These questions go with the need for additional simulation studies that investigate different parameters specification and dependence structures behind the imputation performances. Moreover, real-data applications should be addressed to set up straightforward data quality criteria for the matched data sets.

Author contributions The article is based primarily on the Ph.D. dissertation thesis of RD, supervised by MR. Conceptualization: RD and MR; Literature search: RD; Data analysis: RD; Writing—original draft: RD; Writing—review and editing: RD; Revision: RD.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement.

Appendix A Applications and software

This Appendix focuses on several SM contributions which share the following key features: (1) they are (mainly) practical applications of the method, (2) empirical examples, or (3) technical reports based on real-world data.

A.1 '70s–'90s applications

The decades from the early '70s up to the late '90s are characterized by two major 'data integration schools': the North-American one of which Okner (1972) is the first representative (see, in this regard, Sect. 3.1) and the European one. The latter, which is less known in the SM state-of-the-art, can be divided into at least three schools (French, German, and British) with different authors applying matching methods above all in the field of marketing research.

In the U.S. and Canada, Statistical Matching was applied mainly to integrate administrative registers and primary data from the OS and/or national departments and federal offices.

- Budd (1971) integrated micro-data files to be used for estimating the size distribution of income, resulting from the matching of various sources. The main aim is to correct and/or supplement the income estimates in the original U.S. Current Population Survey (context: Office of Business Economics, U.S. Department of Commerce).
- Okner (1972) integrated micro-data files from the 1967 U.S. Survey of Economic Opportunity and the 1966 U.S. Tax File (context: U.S. Office of Economic Opportunity).
- Alter (1974) integrated data set from the 1970 Canadian Survey of Consumer Finances with the 1970 Family Expenditure Survey (context: Expenditure Division, Statistics Canada).
- Ruggles and Ruggles (1974) offers an empirical example of integration between the 1970 U.S. Public Use Sample and the Social Security Longitudinal Employer-Employee Data File.
- Radner et al. (1980) discussed several empirical examples for evaluating the integration of different data sources produced by U.S. departments, research institutes, and the OS.
- Gavin (1985) integrated data from the Survey of Income and Education and the 1976 National Health Interview Survey (context: U.S. Department of Health and Human Services).
- Armstrong (1989) integrated data from the Survey of Consumer Finance and the Revenue Canada's Tax (context: Business Survey Methods Division, Statistics Canada).

In Europe, Statistical Matching was developed more or less independently among France, Germany, and the United Kingdom but under the same field of (media)marketing research. Indeed, SM was used by both public agencies and private institutes, on the one hand, to integrate television and other media data, while, on the other hand, media data and purchasing information.

- Bergonier et al. (1967) is, to the best of our knowledge, the first contribution that applied matching techniques to media marketing analysis.
- In Rässler (2002) (pp.46–47, and the references therein) there are several German empirical examples of the '70s from the German Media Analysis Association and the Bureau Wendt which carried out and integrated annual surveys on magazine readership and radio-television watching information.
- Wiegand (1986) integrated data of different characteristics and media schedule figures for transferring broadcast media information into press media surveys.
- Antoine and Santini (1987) integrated data from various media survey information, e.g., cinema audience survey and readership self-administered survey, in the context of the 'media-market programme'.
- Baker et al. (1989) and O'Brien (1991) integrated data sets from the British Target Group Index (TGI) data and the Broadcasters' Audience Research Board (BARB) data.
- Roberts (1994) integrated BARB data and the AGB Superpanel, a large market-tracking panel of the United Kingdom.
- Adamek (1994) offers empirical examples for analysing the integration techniques performances using TGI and BARB data.
- Darkow (1996) integrated data from *ad hoc* surveys on television viewing behaviour and other media data.
- Kamakura and Wedel (1997) integrated data from surveys on customer satisfaction related to multi-branch banks in Latin America with internal records.

A.2 '00s: today applications

The last two decades are characterised by SM applications carried out in several research fields but with a common, coherent, and cohesive, theoretical framework.

- Sutherland et al. (2002) integrated data for use in fiscal policy simulations from the British Family Expenditure Survey and the British Family Resources Survey.
- Denk and Hackl (2003) offers empirical examples of integration of different data sources on income and tax returns within the context of the EU-funded Project 'Development of a System of Indicators on Competitiveness and Fiscal Impact on Enterprises Performance' (DIECOFIS).
- Abello and Phillips (2004) integrated data from the National Health Survey and the Household Expenditure Survey (context: Methodology Advisory Committee, Australian Bureau of Statistics).
- Ballin et al. (2009) integrated data from the Farm Accountancy Data Network (FADN) and the Farm Structure Survey (FSS) (context: Italian National Institute of Statistics).
- Agafitei and Leulescu (2013) discussed two empirical examples of data integration related to (1) quality-of-life data, by matching European Union Statistics on Income and Living Conditions with the European Quality of Life Survey and, (2) labour and wages data, by matching European Union Statistics on Income and Living Conditions data with the Labour Force Survey (context: EUROSTAT, European Commission).

- Gutman et al. (2013) integrated micro-data files for use in estimating and forecasting health care costs and end-of-life expenditures, by matching data from the U.S. Centers for Medicare and Medicaid Services with Visual Statistics Mortality data.
- Roesch and Lips (2013) integrated data from FSS and FADN on Swiss agricultural holdings.
- D'Orazio and Catanese (2016) integrated data for estimating the energy production performances of the Italian agricultural holdings by combining the Economic Outcomes of Agricultural Holdings annual survey and the FADN data (context: Economic and Social Development Department, Food and Agriculture Organization of the United Nations).
- D'Alberto et al. (2018) integrated data for agricultural policy impact evaluation by combining Italian FADN data and an EU-funded FP7 Project survey.

A.3 Software

- Rässler (2003): non-iterative Bayesian based imputation (**NIBAS**) algorithm, S-PLUS 2000 (MathSoft, Inc.)
- Alpman (2016): **smpc** and **smmatch** commands for Statistical Matching, Stata 14 (StataCorp, LLC.)
- D'Orazio (2020): **StatMatch** package, R (R Foundation for Statistical Computing)

References

Abello R, Phillips B (2004) Statistical matching of the HES and NHS: an exploration of issues in the use of unconstrained and constrained approaches in creating a basefile for a microsimulation model of the pharmaceutical benefits scheme. ABS Technical Working Paper. Technical report. pp 1–44

Adamek JC (1994) Fusion: combining data from separate sources. Market Mag Manag Appl 6:48–50

Agafitei M, Leulescu A (2013) Statistical matching: a model based approach for data integration. Eurostat methodologies and working papers. Technical report. pp 1–100

Agresti A (2013) Categorical data analysis. Wiley, London

Ahfock D, Pyne S, Lee SX, McLachlan GJ (2016) Partial identification in the statistical matching problem. Comput Stat Data Anal 104:79–90. https://doi.org/10.1016/j.csda.2016.06.005

Alpman A (2016) Implementing Rubin's alternative multiple-imputation method for statistical matching in Stata. Stata J 16:717–739. https://doi.org/10.1177/1536867X1601600311

Alter HE (1974) Creation of a synthetic data set by linking records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey. Ann Econ Soc Meas 3:373–394

Anderson TW (1957) Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. J Am Stat Assoc 52:200–203. https://doi.org/10.2307/2280845

Antoine J, Santini G (1987) Fusion techniques: alternative to single-source methods. Eur Res 15:178–187

Armstrong J (1989) An evaluation of statistical matching methods. Business Survey Methods Division—statistics Canada. Technical report, 1–48

Baker K, Harris P, O'Brien J (1989) Data fusion: an appraisal and experimental evaluation. J Market Res Soc 31:153–212

Ballin M, D'Orazio M, Di Zio M, Scanu M, Torelli N (2009) Statistical matching of two surveys with a common subset. Università di Trieste Working papers. Technical report. pp 1–12

Barry JT (1988) An investigation of statistical matching. J Appl Stat 15:275–283. https://doi.org/10.1080/02664768800000038

Bergonier H, Boucharenc L, Irrmann P (1967) Une nouvelle methode d'analyse globale des resultats d'une enquete etablissement de typologies. Rev Française de Market 25:31–41

Bethlehem J (2016) Solving the nonresponse problem with sample matching? Soc Sci Comput Rev 34:59–77

Brozzi A, Capotorti A, Vantaggi B (2012) Incoherence correction strategies in statistical matching. Int J Approx Reason 53:1124–1136. https://doi.org/10.1016/j.ijar.2012.06.009

Budd EC (1971) The creation of a microdata file for estimating the size distribution of income. Rev Income Wealth 11:317–334. https://doi.org/10.1111/j.1475-4991.1971.tb00785.x

Castro-Martín L, Rueda MM, Ferri-García R (2022) Combining statistical matching and propensity score adjustment for inference from non-probability surveys. J Comput Appl Math 404:3414. https://doi.org/10.1016/j.cam.2021.113414

Chen Y, Li P, Wu C (2020a) Doubly robust inference with non-probability survey samples. J Am Stat Assoc 115:2011–2021. https://doi.org/10.1080/01621459.2019.1677241

Chen S, Yang S, Kim JW (2020b) Nonparametric mass imputation for data integration. J Surv Stat Methodol. https://doi.org/10.1093/jssam/smaa036

Christen P (2012) Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer, Berlin

Chung CK, Cheng PE (1995) Nonparametric regression estimation with missing data. J Stat Plan Inference 48:85–99. https://doi.org/10.1016/0378-3758(94)00151-K

Cohen ML (1991) Statistical matching and microsimulation models. In: Citro CF, Hanushek EA (eds) Improving information for social policy decisions—the uses of microsimulation modeling: Volume II, technical papers. The National Academies Press, Washington DC, pp 62–88

Conti PL, Marella D, Scanu M (2008) Evaluation of matching noise for imputation techniques based on nonparametric local linear regression estimators. Comput Stat Data Anal 53:354–365. https://doi.org/10.1016/j.csda.2008.07.041

Conti PL, Marella D, Scanu M (2013) Uncertainty analysis for statistical matching of ordered categorical variables. Commun Stat-Theor Methods 68:311–325. https://doi.org/10.1016/j.csda.2013.07.004

Conti PL, Marella D, Scanu M (2016) Statistical matching analysis for complex survey data with applications. J Am Stat Assoc 111:1715–1725. https://doi.org/10.1080/01621459.2015.1112803

Conti PL, Marella D, Neri A (2017a) Statistical matching and uncertainty analysis in combining household income and expenditure data. Stat Methods Appl 26:485–505. https://doi.org/10.1007/s10260-016-0374-7

Conti PL, Marella D, Scanu M (2017b) How far from identifiability? A systematic overview of the statistical matching problem in a non parametric framework. Commun Stat-Theor Methods 46:967–994. https://doi.org/10.1080/03610926.2015.1010005

Conti PL, Marella D, Scanu M (2019) An overview on uncertainty and estimation in statistical matching. In: Zhang L-C, Chambers RL (eds) Analysis of integrated data. CRC Press, Boca Raton, pp 73–96

D'Alberto R, Zavalloni M, Raggi M, Viaggi D (2018) AES impact evaluation with integrated farm data: combining statistical matching and propensity score matching. Sustainability 10:1–24. https://doi.org/10.3390/su10114320

D'Orazio M (2009) Uncertainty intervals for nonidentifiable parameters in statistical matching. In: Proceedings of the 57th Session of the International Statistical Institute, Durban (South Africa), August 2009

D'Orazio M (2015) Statistical matching and imputation of survey data with StatMatch. Italian National Institute of Statistics—ISTAT. Technical report. pp 1–35

D'Orazio M (2020) Statistical matching and imputation of survey data with StatMatch. R package version 1.4.0. https://cran.r-project.org/package=StatMatch

D'Orazio M, 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. In: Proceedings of the Seventh International Conference on Agricultural Statistics. pp 1–8

D'Orazio M, Di Zio M, Scanu M (2006a) Statistical matching for categorical data: displaying uncertainty and using logical constraints. J Off Stat 22:137–157

D'Orazio M, Di Zio M, Scanu M (2006b) Statistical matching: theory and practice. Wiley, Hoboken

D'Orazio M, Di Zio M, Scanu M (2017) The use of uncertainty to choose matching variables in statistical matching. Int J Approx Reason 90:433–440. https://doi.org/10.1016/j.ijar.2017.08.015

Darkow M (1996) Compatible or not? Results of a single source field experiment within a TV audience research panel. Market Res Today 24:150–161

de Finetti B (1974) Theory of probability. Wiley, London

de Waal T (2015) Statistical matching: experimental results and future research questions. Statistics Netherlands. Technical report. pp 1–33

Denk M, Hackl P (2003) Data integration and record matching: an Austrian contribution to research in official statistics. Austrian J Stat 32:305. https://doi.org/10.17713/ajs.v32i4.464

Di Zio M, Vantaggi B (2017) Partial identification in statistical matching with misclassification. Int J Approx Reason 82:227–241. https://doi.org/10.1016/j.ijar.2016.12.015

Dunn HL (1946) Record linkage. Am J Public Health 36:1412–1416

Fellegi IP, Sunter AB (1969) A theory for record linkage. J Am Stat Assoc 64:1183–1210

Gavin NI (1985) An application of statistical matching with the survey of income and education and the 1976 Health Interview Survey. Health Serv Res 20:183–198

Gessendorfer J, Beste J, Drechsler J, Sakshaug JW (2018) Statistical matching as a supplement to record linkage: a valuable method to tackle nonconsent bias? J Off Stat 34:909–933. https://doi.org/10.2478/JOS-2018-0045

Grace J (2006) Composite variables and their uses. In: Grace J (ed) Structural equation modeling and natural systems. Cambridge University Press, Cambridge, pp 143–180

Gutman R, Afendulis CC, Zaslavsky AM (2013) A Bayesian procedure for file linking to analyze end-of-life medical costs. J Am Stat Assoc 108:34–47. https://doi.org/10.1080/01621459.2012.726889

Harron K, Goldstein H, Dibben C (2016) Methodological developments in data linkage. Wiley, Chichester

Iaccarino G (2019) Metrics and methods for uncertainty quantification, presentation to the new techniques and technologies for statistics (NTTS – 2019), Brussels, 11–15 March 2019. https://ec.europa.eu/eurostat/cros/system/files/iaccarino_ntts2019.pdf

Judson DH (2005) Computerized record linkage and statistical matching. In: Kempf-Leonard K (ed) Encyclopedia of social measurement, vol 2. Elsevier, Amsterdam, pp 439–447

Kadane JB (1978) Some statistical problems in merging data files. Compendium of Tax Research—U.S. Department of the Treasury. Technical report. pp 159–171

Kamakura WA, Wedel M (1997) Statistical data fusion for cross-tabulation. J Market Res 34:485–498

Kim JW, Tam S-M (2021) Data integration by combining Big Data and survey sample data for finite population inference. Int Stat Rev 89:382–401. https://doi.org/10.1111/insr.12434

Klevmarken NA (1982) Missing variables and two-stage least squares estimation from more than one data set. In: Proceedings of the American Statistical Association—business and economic statistics section. pp 156–161

Lee S (2006) Propensity score adjustment as a weighting scheme for volunteer panel web surveys. J Off Stat 22:329–349

Lohr SL, Raghunathan TE (2017) Combining survey data with other data sources. Stat Sci 32:293–312

Manski CF (1995) Identification problems in the social sciences. Harvard University Press, Cambridge

Mardia KV, Kent JT, Bibby JM (1980) Multivariate analysis (probability and mathematical statistics). Academic Press, London

Marella D, Pfeffermann D (2019) Matching information from two independent informative samples. J Stat Plan Inference 203:70–81. https://doi.org/10.1016/j.jspi.2019.03.001

Marella D, Scanu M, Conti PL (2008) On the matching noise of some nonparametric imputation procedures. Stat Probab Lett 78:1593–1600. https://doi.org/10.1016/j.spl.2008.01.020

Moriarity C, Scheuren F (2001) Statistical matching: a paradigm for assessing the uncertainty in the procedure. J Off Stat 17:407–422

Moriarity C, Scheuren F (2003) A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputations. J Bus Econ Stat 21:65–73

Murray JS (2018) Multiple imputation: a review of practical and theoretical findings. Stat Sci 2:142–159. https://doi.org/10.1214/18-STS644

Newcombe HB, Kennedy J, Axford S, James A (1959) Automatic linkage of vital records. Science 130:954–959

Nielsen SF (2001) Nonparametric conditional mean imputation. J Stat Plan Inference 99:129–150. https://doi.org/10.1016/S0378-3758(01)00087-8

O'Brien S (1991) The role of data fusion in actionable media targeting in the 1990's. Market Res Today 19:15–22

Okner BA (1972) Constructing a new data base from existing microdata sets: the 1966 merge file. Ann Econ Soc Meas 1:325–342

Pentland S (2019) Better decisions with data, presentation to the new techniques and technologies for statistics (NTTS—2019), Brussels, 11–15 March 2019. https://ec.europa.eu/eurostat/cros/system/files/pentland_ntts_2019.pdf

Radner DB, Allen R, Gonzalez ME, Jabine TB, Muller HJ (1980) Report on exact and statistical matching techniques. Statistical policy paper 5—U.S. Department of Commerce. Technical report. pp 1–58

Rao JNK (2021) On making valid inferences by integrating data from surveys and other sources. Sankhya 83:242–272

Rässler S (2002) Statistical matching: a frequentist theory, practical applications, and alternative bayesian approaches. Springer, New York

Rässler S (2003) A non-iterative Bayesian approach to statistical matching. Stat Neerl 57:58–74. https://doi.org/10.20378/irbo-55154

Rässler S (2004) Data fusion: identification problems, validity, and multiple imputation. Austrian J Stat 33:1538

Renssen RH (1998) Use of statistical matching techniques in calibration estimation. Surv Methodol 24:171–183

Rivers D (2007) Sampling for web surveys. In: Proceedings of the American Statistical Association—Joint statistical meetings, Salt Lake City. pp 1–26

Roberts A (1994) Media exposure and consumer purchasing: an improved data fusion technique. Market Res Today 22:150–172

Rodgers WL (1984) An evaluation of statistical matching. J Bus Econ Stat 2:91–102. https://doi.org/10.2307/1391358

Rodgers WL, DeVol E (1981) An evaluation of statistical matching. In: Proceedings of the American Statistical Association—section on survey research methods. pp 128–132

Roesch A, Lips M (2013) Sampling design for two combined samples of the Farm Accountancy Data Network (FADN). J Agric Biol Environ Stat 18:178–203. https://doi.org/10.1007/s13253-013-0130-5

Rubin RD (1974) Characterizing the estimation of parameters in incomplete-data problems. J Am Stat Assoc 69:467–474. https://doi.org/10.2307/2285680

Rubin RD (1976) Inference and missing data. Biometrika 63:581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin RD (1986) Statistical matching using file concatenation with adjusted weights and multiple imputations. J Bus Econ Stat 4:87–94. https://doi.org/10.2307/1391390

Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley, New York

Ruggles R, Ruggles N (1974) A strategy for merging and matching microdata sets. Ann Econ Soc Meas 3:353–371

Schulte Nordholt E (1998) Imputation: methods, simulation experiments and practical examples. Int Stat Rev 66:157–180. https://doi.org/10.2307/1403488

Sims CA (1972a) Comments. Ann Econ Soc Meas 1:343–345

Sims CA (1972b) Rejoinder. Ann Econ Soc Meas 1:355–357

Singh AC, Armstrong JB, Lemaitre GE (1988) Statistical matching using log-linear imputation. In: Proceedings of the American Statistical Association—section on survey research methods. pp 672–677

Singh AC, Mantel HJ, Kinack MD, Rowe G (1993) Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption. Surv Methodol 19:59–79

Sutherland H, Taylor R, Gomulka J (2002) Combining household income and expenditure data in policy simulations. Rev Income Wealth 48:517–536. https://doi.org/10.1016/10.1111/1475-4991.00066

United Nations Economic Commission for Europe (UNECE) (2017) A guide to data integration for official statistics, technical report of the data integration project—version 1.0. High Level Group for the Modernisation of Official Statistics (HLG-MOS)

Vantaggi B (2008) Statistical matching of multiple sources: a look through coherence. Int J Approx Reason 49:701–711. https://doi.org/10.1016/j.ijar.2008.07.005

Walter SD (1984) Required sample size for categorical matching. J Am Stat Assoc 79:662–667

Wiegand J (1986) Combining different media surveys: the German partnership model and fusion experiments. J Market Res Soc 28:189–208