SPES

Sustainability
performances,
evidence & scenarios

D4.1
Report on mapping, harmonising and integrating novel data sources for research purposes

January 2024

Funded by European Union's Horizon Europe
Programme under Grant Agreement
No. 101094551

Authors

Veronica Ballerini – Researcher of the SPES Project, University of Florence
Davide Beraldo – Researcher of the SPES Project, University of Amsterdam
Chiara Bocci – Researcher of the SPES Project, University of Florence
Lisa Braito – Researcher of the SPES Project, University of Florence
Roberta Milana – Interning researcher of the SPES Project, University of Amsterdam
Emilia Rocco – Researcher of the SPES Project, University of Florence
Martin Trans – Researcher of the SPES Project, University of Amsterdam

Contributors and peer reviewers:
Mario Biggeri, University of Florence; Paolo Brunori, London School of Economics & Political Science; Jeroen de Vos, University of Amsterdam; Andrea Ferrannini, University of Florence; András Gábos, TARKI; Luca Lodi, University of Florence; Fabrizia Mealli, European University Institute; Stefania Milan, University of Amsterdam; Gregor Zens, IIASA.

Acknowledgements:
The authors would like to thank all SPES partners for their inputs provided in several SPES meetings and communications. We would like to thank also all interviewees involved in our data collection activities for having shared with us their opinion and insights on the topic.

Cite as:
Ballerini, V., Beraldo, D., Bocci, C., Braito, L., Milana, R., Trans, M. (2024). Report on mapping, harmonising and integrating novel data sources for research purposes. SPES Report no. 4.1, SPES project – Sustainability Performances, Evidence and Scenarios. Florence: University of Florence. Available at: https://www.sustainabilityperformances.eu/publications-deliverables/

Disclaimer

This Report D4.1 for the project SPES has been prepared by the University of Amsterdam and the University of Florence as part of Task 4.1 "Map complex and novel data sources and methods" of Work Package 4.

This task has allowed SPES research partners to identify and provide an evaluation of complex and novel data sources and associated methods that can be repurposed for the development of innovative measurement framework.

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Table of contents

Part II

Integrating traditional and innovative data sources

Lisa Braito, Veronica Ballerini[2], Chiara Bocci, Emilia Rocco

[2] Ballerini is supported by the European Union - Next GenerationEU, UNIFI Young Independent Researchers Call - BayesMeCOS Grant n°B008-P00634"

1.Introduction

Since the 1950s, a substantial portion of the statistical information generated by national statistical agencies has originated from sample surveys. Since they were established, sample surveys have been serving as the primary means to acquire reliable, accurate, and regularly updated information regarding the national population and businesses. Beyond describing the world, the data derived from these surveys have been serving various purposes, including informing policy decisions related to economics, social issues, and health. Additionally, they have been contributing to the assessment of the impacts, monitoring the overall health and economic conditions of the population, guiding decisionmaking for businesses and individuals, and fostering extensive research in economics, health, and social domains.

However, in recent years, surveys have encountered several challenges, including diminishing response rates, escalating costs, and a growing demand from users for more timely and granular data and statistics. The wealth of information is far from being threatened though. Nowadays, there has been a surge in data from alternative sources, such as administrative records generated by government agencies, satellite and sensor data, private-sector information like electronic health records and credit card transactions, and massive datasets accessible on the internet.

Novelties often come with questions: How can these emerging data sources be effectively utilised to complement or even substitute some of the information traditionally collected through surveys? How can they pave the way for new possibilities in generating information and statistics that contribute to the improvement of society?

The use of alternative data sources encounters some problems and at the same time affects classical survey programs. For instance, new data sources may lead to changes in measurement, and the existence of innovative data sources needs the development of approaches to link these alternative sources to universal frames to assess representativeness.

The past two decades have seen a profound “data revolution" in the field of social sciences, propelled by technological advancement. This era has provided researchers with access to diverse and extensive datasets in electronic formats. A significant portion of cutting-edge quantitative social science research stems from the creativity of researchers who skilfully integrate disparate datasets collected independently.

The emergence of new data sources, coupled with evolving perspectives, has broadened the scope of data integration and harmonisation, accentuating the need to explore, exploit, and develop statistical techniques in the realm of data integration. This expanding landscape necessitates ongoing research to address the complexities inherent in this dynamic field. Leveraging multiple data sources holds the potential to enhance the production of official statistics and advance research. The long tradition in the integration of survey data, such as EU-SILC, with national administrative data (income data registers, students registers; see, e.g., Jantti et al., 2013) is an example of how combining different data sources may allow the measurement and monitoring of social mobility, inter-generational inequality, and poverty. There are many other examples of opportunities given by the integration of different sources. For instance, if one is interested in analysing enterprises and their innovation and sustainable practices, one can integrate traditional sources, such as the national enterprises' census or ad hoc surveys, with information from web scraping data where text analysis is performed to the text on their website (see, e.g., Van der Doef et al., 2018; van den Brakel et al., 2019). When a probability survey is available for the target population of enterprises (e.g., a survey on start-ups), one can use methods to combine this information with the data arising from web scraping to make inferences about the determinants of innovation for this company. For a final example, we move to the dimension of environmental sustainability. The extensive availability of satellite imagery allows granularity and timeliness of data production on, e.g., land use (see, e.g., the CLMS project[3]). At the same time, the collection of daily pollution data via sensors is well established (see, e.g., the CAMS project[4]). Leveraging surveys at the municipal level, collecting - among other things - information on urban green areas, waste production, presence and activity of waste-disposal plants (among others, see the Italian urban environmental data survey), one could construct reliable and representative environmental indicators (UNOOSA, 2018). For further examples of the use of satellite imagery and remote sensing data, see Paganini et al. (2018); Anderson et al. (2017); Donaldson and Storeygard (2016). Yet, the integration of information from diverse sources requires meticulous attention, demanding a profound understanding of the distinctive properties inherent in each dataset and the statistical outcomes that arise from their harmonisation.

This part of the report aims to provide an overview for a broader audience of the statistical methods developed in the last decades in Official statistics and Survey statistics to deal with data integration issues. This part of the report is structured as follows. Section 1 concerns data integration as methods to create a synthetic dataset merging information from different sources. Section 2 deals with the issue of combining probability and nonprobability samples for inference. Section 3 is a collection of blue boxes that supply available statistical software. The reader can find other kinds of boxes throughout the text: the green ones propose toy/practical examples to exemplify both the issues at stake and the yellow ones provide statistical details of the suggested methods. Finally, the Appendix includes a flowchart supporting the users in the choice of the most suitable methods according to their needs.

1.Creation of a synthetic dataset

In this section, we present two data integration procedures, namely, record linkage and statistical matching. Such procedures aim to integrate two (or more) datasets that contain information on a set of common variables and variables that are not jointly observed. As output, the implementation of these integration procedures gives a set of pairs of records.

Although at first glance they could seem almost the same thing, record linkage and statistical matching differ in at least four aspects. The first two aspects concern the aim of the procedures and the nature

[3] Copernicus Land Monitoring Service

[4] Copernicus Atmosphere Monitoring Service

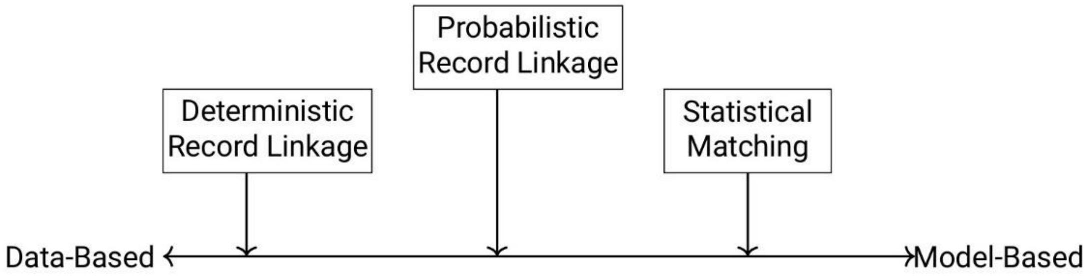


Figure 1: Ranking of creation of synthetic dataset from data-based to model-based methods (reproduced based on Asher et al. (2020))

of the information to integrate. Record linkage mainly aims to identify pairs of records corresponding to a single statistical unit that are present in different databases; statistical matching seeks to derive integrated statistical information by combining information from different datasets in which only some variables are observed twice and the overlapping of observed units is not necessary. Another aspect concerns the fact that, in a record linkage procedure, the common variables are sometimes misreported or subject to change, while statistical matching does not have to deal with the problem of the quality of collected data. Record linkage provides a solution to the problem of the quality of collected data collected introducing the linkage procedure parameters such as the probability of having observed the variable without noise. Finally, the procedures differ in the hypotheses at the base of the methodologies; while record linkage may not introduce any strong assumption on the conditional distribution of the variables of interest, statistical matching mainly works assuming conditional independence.

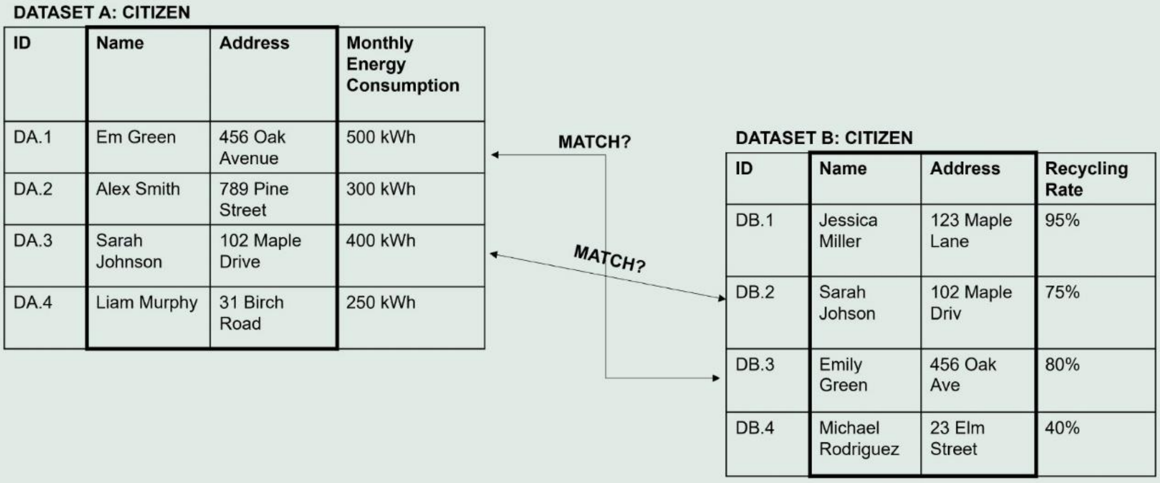
As portrayed in Figure 1, following the ranking outlined by Asher et al. (2020), first, we have deterministic record linkage which is the most traditional methods that rely on rules-based data integration. It involves simply combining records that share an identical key, such as a unique identifier, or when this is not available, a set of identifiers (referred to as a key) that uniquely identify an individual. Typically, this key is available for most records in one or more datasets. In probabilistic record linkage, a probability model is used to assess the probability that a pair of records represent the same entity. Statistical Matching, instead, is the most model-driven methods that match records with similar characteristics across multiple covariates, allowing for comparisons of values for a different set of variables.

1.1.Record Linkage

Record linkage, known by various terms such as data linkage, record linkage, data matching, entity resolution, deduplication, data matching, and instance matching, is the practice of identifying records in two or more data sources that pertain to the same entity. When the linkage process is precise and reliable, then the information from the sources can be merged, allowing researchers to study relationships among variables measured in the individual sources. The effectiveness of a record linkage method hinges on the extent to which the information available in the two sources can successfully distinguish and link records i.e., individual persons, households, or businesses. Many methods exist for linking records and assessing the quality of the linkages (for a comprehensive view see among others National Academies of Sciences, Engineering, and Medicine (2023), Binette and Steorts (2022), Asher et al. (2020), Reiter (2021).

Toy Example 1: Creation of a synthetic dataset via Record Linkage

Here is an example of a scenario where we find ourselves with the problem of creating a synthetic dataset. Let us suppose we have two distinct datasets collecting information about sustainability practices and the energy consumption of individuals in Ireland. In this application, we do not care about the design behind the datasets. Only dataset A collects information on individuals' energy consumption, whereas dataset B lists individuals' recycling rates - which are missing in A. Both datasets contain some common variables, namely citizen's name, and residence address.



DATASET A: CITIZEN

ID	Name	Address	Monthly Energy Consumption
DA.1	Em Green	456 Oak Avenue	500 kWh
DA.2	Alex Smith	789 Pine Street	300 kWh
DA.3	Sarah Johnson	102 Maple Drive	400 kWh
DA.4	Liam Murphy	31 Birch Road	250 kWh

DATASET B: CITIZEN

ID	Name	Address	Recycling Rate
DB.1	Jessica Miller	123 Maple Lane	95%
DB.2	Sarah Johson	102 Maple Driv	75%
DB.3	Emily Green	456 Oak Ave	80%
DB.4	Michael Rodriguez	23 Elm Street	40%

Our objective is to create a complete synthetic dataset that integrates information from both datasets linking individuals in dataset A with those in B according to the values of the common variables (often called *matching variables*) while accounting for variations in names and addresses. With a very small sample size and in the absence of errors, individuals' names could be considered as unique identifiers. However, on the one hand, it is common to work with large datasets, which implies that it is likely to encounter homonyms, and, on the other hand, measurement errors (e.g., typing errors) are frequent. Hence, when there are no unique identifiers, we are not able to simply match the individuals in the datasets. The challenge lies in accurately linking records with variations in the matching variables.

Record linkage techniques are essential in various fields, including epidemiology, social sciences, and business analytics, as they facilitate accurate identity resolution, leading to more informed decisionmaking and research outcomes. These techniques enable the incorporation of variables measured in additional data sources into the primary data sources. For example, survey records of individuals can be linked to attributes related to the geographic regions in which survey respondents reside or the affiliations and associations to which they belong. Furthermore, when merging two or more datasets, each representing a subset of the population, these techniques enable the expansion of the dataset by increasing the number of records. They also facilitate the construction of longitudinal datasets by establishing connections between records associated with the same individuals over time, such as merging high school records with data on college completion. In a broader context, they offer a means to validate the accuracy of data within a source by cross-referencing it with other sources (National Academies of Sciences, Engineering, and Medicine, 2023).

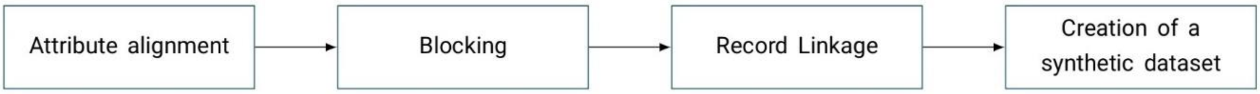


Figure 2: Reproduced based on (Binette and Steorts, 2022)

The process of linking records is a multi-stage operation that encompasses four distinct phases, with Record Linkage serving as just one of these pivotal stages (see Figure 2). Each of these stages plays a critical role in the overall process, from data preparation to record linkage, ultimately culminating in the successful integration of information from disparate sources. This sequential approach ensures a systematic and comprehensive handling of data, guaranteeing the accuracy and effectiveness of record linkage as an integral component of the broader procedure. In the initial stage, records are meticulously parsed to identify a common set of attributes or fields shared among the dataset. This step lays the foundation for subsequent comparisons. Attribute alignment includes also the standardisation of variables. Similar records are then thoughtfully grouped into ’blocks.’ Only records residing within the same block are subjected to direct comparison; records that do not share a block are promptly determined to be non-matches. This technique, often referred to as “blocking" is a crucial strategy aimed at reducing the number of potential candidate pairs that need to be examined. By grouping records based on specific criteria, blocking ensures that only records with shared characteristics are compared, thus effectively managing the scalability of record linkage in large datasets (Christen and Christen, 2012). Blocking, for instance, involves grouping records according to certain criteria, such as comparing records with identical zip or postcode values, which streamlines the linkage process (Asher et al., 2020). However, this approach comes with a caveat - the exact propagation of uncertainty from the blocking phase to the record linkage stage is not always achievable. Consequently, the record linkage task may inherit errors from the blocking stage, some of which may remain unresolved. Moving on to the record linkage stage, various methods are employed to identify matching records. This stage encompasses deterministic, probabilistic, or machine learning-based classification strategies, all of which are instrumental in determining the degree of similarity between records. In the final stage, entities that have been successfully resolved as matches in the preceding stage are consolidated to create a singular, representative record. This consolidation step ensures that the merged record accurately represents the linked information, thus concluding the record linkage process (Binette and Steorts, 2022).

Record Linkage tasks encounter a multifaceted trade-off, involving (i) the ability to handle large databases, (ii) ensuring the propagation of uncertainty throughout the entire data cleaning process, and (iii) developing methods that can effectively address the distortions and errors commonly found in databases.

In the context of databases comprising a cumulative total of N records, there exist $N(N-1)/2$ possible pairs of records that might be correlated, making it infeasible to evaluate each pair as the database size expands. The user should assess the trade-off involving (i), (ii), and (iii) within the context of each specific application to determine the most fitting method (Binette and Steorts, 2022).

1.1.1.Deterministic Record Linkage

Deterministic record linkage is a process that relies on a set of deterministic rules involving the comparison of attributes within records. It encompasses deterministic, rule-based, and similarity-based methods (for references on deterministic algorithm see Adena et al., 2015; Ansolabehere and Hersh, 2017; Berent et al., 2016; Bolsen et al., 2014; Cesarini et al., 2016; Figlio et al., 2014; Giraud-Carrier et al., 2015; Hill, 2017; Meredith and Morse, 2014). For instance, a straightforward example is exact matching, where two record pairs are linked only if they precisely match all common attributes. See Toy Example 1.1 for an illustration of the method.

Toy Example 1.1: Deterministic Record Linkage

Recall the Toy Example 1. We are interested in linking the records of two datasets on energy consumption and sustainability rates. We wonder if "Em Green" (DA.1) and "Emily Green" (DB.3) refer to the same individual; the address is spelled differently although it may appear the same if visually inspected. Similarly for "Sarah Johnson" (DA.3) and "Sarah Johson" (DB.2).

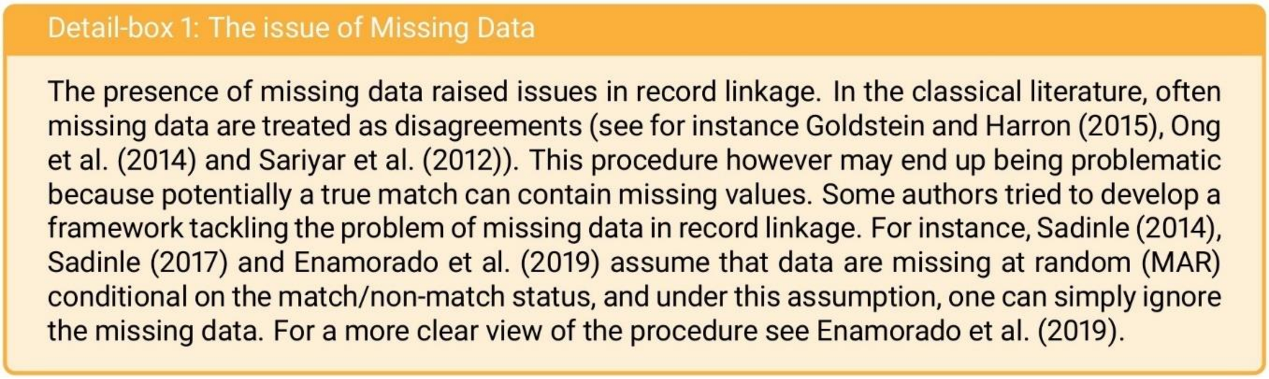
- Common Identifiers: We need to define which are the common identifiers. In this case, all common variables, namely "Name" and "Address" are chosen as matching variables.
- Linkage Rule: Apply a strict rule for exact matches. Spelling errors are not allowed, therefore only records with identical names, ages, and addresses will be linked.
- Outcome: the two pairs of records are not matches.

However, exact matching can be problematic, especially when dealing with data prone to enumeration and transcription errors, as it may exclude a significant number of true matches. While it effectively minimises the occurrence of false positives, it often results in a high rate of false negatives. Additionally, it tends to produce matched samples that are not representative of the overall population. To address these limitations, various strategies have been developed to relax the stringent matching criteria. This can involve allowing a certain number of attribute mismatches, employing disjunctions of exact matching rules, or utilising similarity functions to gauge the similarity of records. These methods introduce flexibility and accommodate errors in the data, typically implemented through an algorithm that specifies a set of decision rules to determine whether two records are sufficiently similar to be considered a match (see for instance ABE method in Abramitzky et al., 2012, 2014, 2019). However, it's important to note that deterministic approaches lack a built-in mechanism to account for uncertainty in the matching process. They do not employ probability models, nor do they provide a level of confidence in determining the matching status of record pairs. Despite this limitation, deterministic matching approaches can still be valuable, particularly when used as a blocking stage to prepare the data for subsequent probabilistic record linkage methods that incorporate uncertainty considerations. This two-stage approach allows for the scalability of more advanced techniques while benefiting from the efficiency of deterministic rules during the initial stage of record linkage.

1.1.2.Probabilistic Record Linkage

While many social scientists have traditionally relied on deterministic methods for record linkage, probabilistic modelling has emerged as a predominant approach within the statistics literature, especially since the seminal work of Fellegi and Sunter (1969) (see Detail-box 2 for a comparison between deterministic and probabilistic).

Probabilistic record linkage methods fundamentally involve calculating a match score for every possible pairing of records from two different datasets, based on their identifying variables. See Toy Example 1.2 for an illustration of the method. This match score is derived as the sum of the weights assigned to each identifying item used in the matching process. When two records agree for the specific item, a positive weight is assigned; in the case of disagreement of records falling outside a prescribed tolerance, a negative weight is assigned; and if an item is missing in either record, the weight is set to zero. For the treatment of missing variables see Detail-box 1.



Detail-box 1: The issue of Missing Data

The presence of missing data raised issues in record linkage. In the classical literature, often missing data are treated as disagreements (see for instance Goldstein and Harron (2015), Ong et al. (2014) and Sariyar et al. (2012)). This procedure however may end up being problematic because potentially a true match can contain missing values. Some authors tried to develop a framework tackling the problem of missing data in record linkage. For instance, Sadinle (2014), Sadinle (2017) and Enamorado et al. (2019) assume that data are missing at random (MAR) conditional on the match/non-match status, and under this assumption, one can simply ignore the missing data. For a more clear view of the procedure see Enamorado et al. (2019).

The weight allocated to an item is influenced, in part, by how effectively that item distinguishes one entity from another. For instance, when two records share an unusual first name like "Dexter", it suggests a higher likelihood of belonging to the same individual compared to two records with a common first name like "John". Consequently, the item weight for a match involving "Dexter" would be greater than that for a match involving "John". Following the scoring process, each pair of records can be categorised as a *match*, a *non-match*, or *indeterminate*. Pairs with scores exceeding a predefined cutoff value are identified as matches, indicating substantial agreement on numerous identification variables. Conversely, pairs with scores below a separate cutoff are categorised as non-matches, suggesting disagreements on enough identification variables to indicate distinct entities. If the highest score for a dataset record doesn't reach the non-match cutoff, it's considered to have no corresponding record in the other dataset. Pairs with scores falling between these two cutoff values may necessitate further review before a definitive determination is reached (National Academies of Sciences, Engineering, and Medicine, 2023).

These probabilistic methods are needed in scientific applications where there is a need to account for all sources of uncertainty that might influence the validity of results. The primary challenge in these scenarios lies in accurately quantifying this uncertainty and integrating it into subsequent analyses. These methods are primarily geared towards estimating the probability of a match between pairs of records based on their comparison vectors. This pairwise match score serves as a metric for assessing uncertainty regarding specific links, with false match and false non-match rates (or precision and recall) being metrics for evaluating performance. It is worth noting that, with the exception of Bayesian approaches, most methods treat record pairs as independent entities, often overlooking the implications of transitivity or other constraints within the linkage structure. This limitation can hinder their practicality when linking more than two databases and when dealing with applications featuring duplication across or within databases (Binette and Steorts, 2022).

In a simplified manner, the procedure followed in probabilistic record linkage can be reconstructed in various steps and can be portrayed in the workflow illustrated in Figure 3.

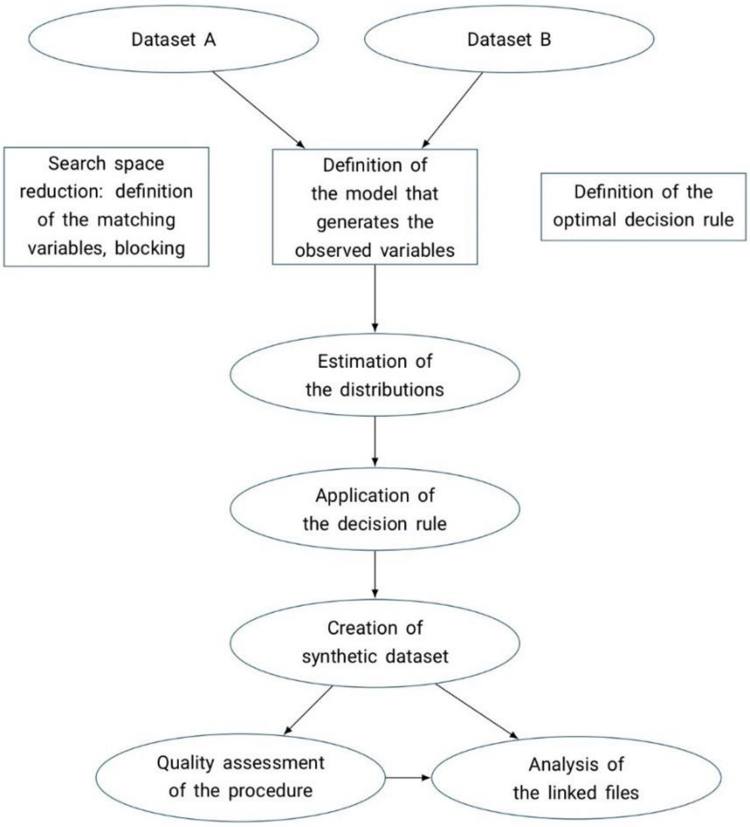

Dataset A
Dataset B
Search space reduction: definition of the matching variables, blocking
Definition of the model that generates the observed variables
Definition of the optimal decision rule
Estimation of the distributions
Application of the decision rule
Creation of synthetic dataset
Quality assessment of the procedure
Analysis of the linked files

Figure 3: reproduced based on ISTAT (2013)

Toy Example 1.2: Probabilistic Record Linkage

Recall Toy Example 1. We are interested in linking the records of datasets A and B. Probabilistic record linkage allows for a more flexible approach than deterministic approaches, considering the probability that records match based on various factors. It seeks to estimate, for pairs of records, the probability that they are a true match.

- Assumption: We assume conditional independence between the comparison variables given the match status of each pair.
- Common identifiers: As in the previous case, all common variables, namely "Name" and "Address" are chosen as matching variables.
- Linkage rule: A scoring method is used to assess the similarity for each attribute between the two records. For instance, among the many available options, one can choose the Jaro-Winkler distance for string variables, such as "Name" and "Address". Consider the pair DA.1 and DB.3.

Variable	Dataset A	Dataset B	Similarity
Name	Em Green	Emily Green	90%
Address	456 Oak Avenue	456 Oak Ave	95%

Using these similarity scores, a probabilistic record linkage algorithm might calculate the overall probability of a match. One needs to specify a certain threshold above which the records are considered to be a match.

- Outcome: The algorithm determines that there is a high probability of a match based on overall similarity, linking the records and providing a more nuanced understanding of individual sustainability transitions.
- Plus of probabilistic RL: The individuals' matching probability can be incorporated further into the analyses to keep accounting for the matching uncertainty in the inferential procedures.

Detail-box 2: Deterministic vs. probabilistic RL - A comparison

DETERMINISTIC Record Linkage

PROS	CONS
→ Rules-based approach, simply data join	→ These methods are in their basic application not robust to measurement error (e.g., misspelling) and missing data
→ Effectively minimizes the occurrence of false positives	→ Cannot quantify the uncertainty of the merging procedure and instead typically relies on arbitrary thresholds to determine the degree of similarity sufficient for matches
→ Used as a blocking stage to prepare the data for subsequent probabilistic record linkage methods	→ No probability model is used and no level of confidence is provided for the matching status of record pairs
→ Simplicity, interpretability, and computational scalability	→ High rate of type II errors (false negatives)
	→ Produces matched samples that are not representative of the overall population

PROBABILISTIC Record Linkage

PROS	CONS
→ One can estimate the false discovery rate (FDR) and the false negative rate (FNR). Researchers typically select, at their discretion, the value of the threshold such that the FDR is sufficiently small	→ Not always scalable (especially when entailing high dimensional data and Bayesian approaches)
→ Incorporation of auxiliary information in parameter estimation (see Enamorado et al. (2019))	
→ One can directly incorporate the uncertainty inherent to the merging process in the post-merge analysis	
→ More matches than exact matching/deterministic matching	
→ It has the property of requiring no training data for record linkage (it is entirely unsupervised)	

Classical Approach: The Fellegi-Sunter Model The Fellegi-Sunter framework (Fellegi and Sunter, 1969) formalises the approach of Newcombe et al. (1959) in a decision-theoretic framework. We will review the Fellegi-Sunter probability model, its interpretation, and its underlying assumptions (Binette and Steorts, 2022).

The Fellegi-Sunter framework operates on the principle of making independent decisions for each pair of records. Within this framework, three potential actions are considered for a given record pair: to establish a link, to indicate a possible link, or to abstain from linking altogether. The primary objective is to strike a delicate balance between minimising the number of potential link assignments and maintaining control over the rates of false matches (Type I error) and false non-matches (Type II error) (see Detail-box 5). The Fellegi-Sunter framework posits that an optimal linkage procedure should achieve the predefined error rates while concurrently minimising the count of potential link assignments. The framework introduces a significant concept known as the “fundamental theorem for record linkage" as established by its authors.

The key idea is that during the linkage process, two specific probabilities must be estimated for each pair of records: m(γ), the probability of observing the comparison vector (γ) used to represent the level of agreement/disagreement between two specified records, for two records that are an actual match and u(γ), probability of observing the comparison vector for two records that are not a match. It follows that, if γ is an observation generated from the distribution m(γ), then the two records are a match.

Instead, if it is generated from u(γ), then it can be asserted that the pair is made of two distinct units. This theorem demonstrated by the authors shows that the optimal linkage procedure corresponds to thresholding a likelihood ratio. More details on the model and the decision rule can be found in the Detail-box 3 and 4.

Detail-box 3: Pills of Fellegi and Sunter procedure - Notation (Scanu, 2003)

Let be A and B two partially overlapping files consisting of the same type of unit, for instance, individuals, households, firms, etc. Dataset A has size n_A and dataset B has size n_B.
The set comprising all potential record pairs originating from A and B is denoted as $\Omega = \{(a,b) : a \in A, b \in B\}$. The two files consist of vectors of quantitative and/or qualitative variables (X_A, Y_A) and (X_B, Z_B). Let us consider X_A and X_B. We are in a setting where these sub-vectors are of k-dimension and thus we have k key variables (common identifiers) such that each unit is uniquely identified by an observation x. Further, let γ_{ab} denote the vector of indicator variables pertaining to the pair of records (a, b), with $\gamma_{ab} = 1$ in the j-th position if $x^{\mathcal{A}}_{a,j} = x^{\mathcal{B}}_{b,j}$ and 0 otherwise $(j = 1, \ldots, k)$. These indicators are called *comparison variables*.
We can formally express record linkage as the task of assigning the pair $(a, b \in \Omega)$ to either one of the two subsets, $\mathcal{M}$ or $\mathcal{U}$. These subsets identify the matched and unmatched sets of pairs, respectively, based on the state of the vector γ_{ab}.
Assuming that the match status is independently and identically distributed, in Fellegi and Sunter (1969) there exists an initial bivariate random variable responsible for allocating each pair of records (a, b) to either the matched records (set $\mathcal{M}$) or the unmatched ones (set $\mathcal{U}$). This variable remains latent, or unobserved, serving as the primary objective in the record linkage process. Additionally, the comparison variables follow distinct distributions based on the pair’s status.
Let $m(\gamma_{ab})$ represent the distribution of comparison variables when the pair (a, b) is matched (i.e., $(a, b) \in \mathcal{M}$) namely $P(\gamma_{ab}|(a, b) \in \mathcal{M})$, and $u(\gamma_{ab})$ denote the distribution of comparison variables when the pair (a, b) is unmatched (i.e., $(a, b) \in \mathcal{U}$) namely $P(\gamma_{ab}|(a, b) \in \mathcal{U})$. These distributions play a pivotal role in determining the status of record pairs.

Detail-box 4: Pills of Fellegi and Sunter procedure - Model and decision rule (Scanu, 2003)

The primary objective of the entire record linkage process is to ascertain whether a pair of records pertains to the same entity. Consequently, the overall quality of the results obtained through the linkage procedure hinges on the effectiveness of the tool employed to make this assessment, namely, the decision rule. In this procedure, the space of actual outcomes consists of a real match or a real non-match for every pair of records belonging to Ω, and the space $\mathcal{D}$ of all possible decisions consists of assigning or not the pair as a link.

We are then mapping from Ω on Γ, thus mapping from the space of all the potential record pairs to the space of comparison. A function that yields a numerical comparison value for $\gamma_{j,ab}$ multiplied by a weight w_j provides a fundamental score indicating the degree of coincidence for the j-th key variable, determining the contribution of each common identifier. The approach introduced by Fellegi and Sunter (1969) consists of taking into consideration the amount of information provided by each key variable by using a log-likelihood ratio considering the agreement probabilities. Thus a weighted measure of agreement is set and it follows that each pair is then assigned the following weight: $w_{ab} = log(\frac{m(\gamma_{ab})}{u(\gamma_{ab})})$.

The following task is to map from Γ on a space of stated decisions which consists of three possible decisions: A_1 (that is, a link), A_3 (that is, a non-link), and A_2 (that is, a possible link). Each decision has its related probabilities which can be derived from the probability distribution $m(\gamma_{ab})$ and $u(\gamma_{ab})$ and the region of Γ associated to each decision. One can thus view record linkage as a common statistical hypothesis test with a critical and acceptance region. These regions are determined by the varied values of γ in Γ and their corresponding composite weight values, which are then compared against a set of fixed bounds. Additionally, a probability model based on $[m(\gamma_{ab}), u(\gamma_{ab})]$ is required to calibrate error rates, denoted as $\mu = P(A_1|(a,b) \in \mathcal{U})$ and $\lambda = P(A_3|(a,b) \in \mathcal{M})$.

The arrangement of γ_{ab} values should be such that the ratio $R_1(\gamma_{ab}) = m(\gamma_{ab})/u(\gamma_{ab})$ exhibits a monotonically decreasing pattern.

The intuitive meaning of the weight justifies the definition of two thresholds τ_μ and τ_λ, with $\tau_\mu > \tau_\lambda$, dependent on fixed values μ and λ, where $0 < \lambda < 1$ and $0 < \mu < 1$.

- $(a,b) \in A_1$ (link), when the ratio is bigger than or equal to τ_μ
- $(a,b) \in A_2$ (possible link) when the ratio is comprised in the region between τ_μ and τ_λ
- $(a,b) \in A_3$ (non-link) when the ratio is lower than or equal to with τ_λ

The thresholds are assigned solving equations that minimise both the size of the set of possible links and the false match rate and false non-match rate.

Figure 4 describes how the optimal rule works in a Fellegi-Sunter framework. Vertical lines in the diagram depict the thresholds. The left line denotes the lower threshold, while the right line denotes the upper threshold. The regions labelled FU and FM indicate the probabilities of false non-matches (FU) and false matches (FM). These regions correspond to the associated error rates for false nonmatches and false matches, respectively.

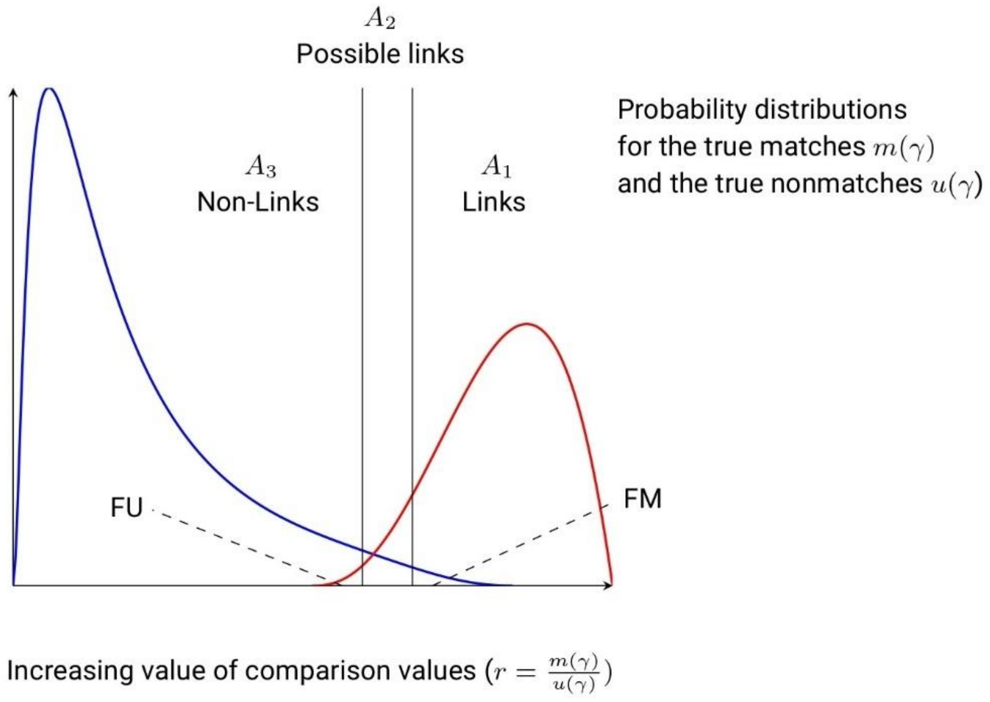

A_2
Possible links
A_3
Non-Links
A_1
Links
Probability distributions
for the true matches $m(\gamma)$
and the true nonmatches $u(\gamma)$
FU
FM
Increasing value of comparison values ($r = \frac{m(\gamma)}{u(\gamma)}$)

Figure 4: F-S optimal rule

The probabilistic record linkage framework operates based on two critical independence assumptions:

1. The latent (unobserved) variable that describes the match status is assumed to be independently and identically distributed.
2. The independence of comparison vectors among record pairs; this is the Conditional Independence Assumption (CIA) i.e., the assumption of independence between the comparison variables given the match status of each pair.

However, these assumptions often face challenges in practical applications. They often require transitivity closure within linkages, meaning that if a record links to another (a links to b) and that a second record links to a third (b links to c), it should imply that the first record links to the third (a links to c). Achieving transitivity closure can be complex and is not commonly realised in practice. Moreover, concerning the feasibility of estimating m and u distributions, the presence of latent variable risks makes the model parameter unidentified, and often methods rely on simplifying assumptions that may not be valid. Since these assumptions frequently do not align with the complexities encountered in real-world scenarios, numerous extensions and variations have been developed in the literature to address these limitations. We provide an overview of these extensions in the "Modern probabilistic record linkage" section, highlighting how researchers have expanded upon the Fellegi-Sunter framework to accommodate more nuanced and practical situations.

The other key aspect is related to the definition of the above-mentioned thresholds. It is important to note that the Fellegi-Sunter theory does not offer specific guidelines for establishing the thresholds to determine matches or non-matches. Instead, it operates on the principle that one can minimize Type I error (false positives) at the expense of Type II error (false negatives) or vice versa by setting these thresholds. Fellegi-Sunter proposes that a manual review of record pairs across a range of assigned weights can be conducted. This manual review aids in the identification of thresholds above which pairs are highly likely to be matches and below which pairs are highly likely to be non-matches. This empirical approach enables the fine-tuning of the thresholds based on the specific needs and characteristics of the dataset, ensuring that the linkage process aligns with the desired balance between minimising false positives and false negatives (Asher et al., 2020).

Detail-box 5: False matches and false non-matches

Let's consider:

- true positives: n_m
- false positives: n_{fp}
- true negatives: n_u
- false negatives: n_{fn}
- the total number of true matches: N_m
- the total number of true non-matches: N_u

We can then define the false match rate as the number of incorrectly linked record pairs divided by the total number of linked record pairs:

$$FMR = \frac{n_{fp}}{(n_m + n_{fp})}$$

It can be seen as the probability of the Type I error and thus corresponds to the significance αin a one-tail hypothesis test.
Analogously we can define the false non-match rate as the number of incorrectly unlinked record pairs divided by the total number of true match record pairs:

$$FNMR = \frac{n_{fn}}{N_m}$$

It corresponds to the probability of the Type II error, thus the probability β in a one-tail hypothesis test.

Modern Approaches to Record Linkage As mentioned in the previous sections, a crucial element in the application of probabilistic rules for record linkage revolves around the distributions of comparison variables corresponding to matches and non-matches, respectively. However, these distributions are typically unknown and necessitate estimation. Thus, the focus of the advancement in methods following the approaches by Fellegi and Sunter (1969) went in the direction of strategies for the estimation of these distributions. As described above, at the core of this framework is the idea that all pairs of records are assumed to be independently generated by a mixture of two distributions—one for matched pairs and another for unmatched ones. The assignment of matched and unmatched status to pairs is determined randomly by a latent (i.e., unobserved) dichotomous variable. In this section, we review modern probabilistic record linkage methods and their different approaches in estimation, which include extensions to the Fellegi-Sunter framework, Bayesian variants of Fellegi-Sunter, as well as machine learning semi-supervised and fully supervised classification approaches (Binette and Steorts, 2022)

1) Extensions to F-S The model described above facilitates the computation of a likelihood function to be maximised for estimating the unknown distributions of the comparison variables for matched and non-matched pairs. Since we are in the presence of a latent (unobserved) variable, the maximisation of the likelihood function typically involves iterative methods for handling it, commonly the EM algorithm or some of its generalisations

One of the primary directions for extending record linkage methods has focused on refining the criteria for setting thresholds (see Belin and Rubin, 1995). Part of the reason is that the error rates fixed in the Fellegi-Sunter framework, as well as the estimated false match rates, are not attained in practice. This discrepancy arises from several factors, including the simplifying assumptions and estimation errors inherent in the application of such probabilistic models. Moreover Winkler et al. (2000); Winkler (2002), and Larsen and Rubin (2001) instead considered fitting more complex models, allowing dependencies between field comparisons. Larsen and Rubin (2001) developed an iterative approach to lower as much as possible the number of records whose status was uncertain. Important to mention is the recent computational and methodological work of Enamorado et al. (2019) who scaled the F-S model to large databases, where they incorporated auxiliary information in the merge to inform parameter estimation and post-merge analyses which accounts for the uncertainty about the merge process.

2)BayesianF-S Bayesian techniques rely on probabilities of a match or non-match for specific agreement patterns that are either based on expert opinion or previous projects. For example, one may be interested in linking two datasets containing the English equivalents of Arabic names, such as for the Syrian data in Tancredi et al. (2020). Assume to have information on how the transliteration was obtained; such information may be formally included in the probabilities of match and non-match via appropriate prior elicitation. As is typical in a Bayesian approach, the prior information is then combined with the data, in with a new record linkage process involving new lists. At the end of the process, a posterior probability of match or non-match is determined for the record pairs, which allows the determination of links and non-links (Asher et al., 2020). Bayesian methods provide a way to quantify and propagate uncertainty for the joint linkage structure of a set of records (see for more detail Section 1.1.3 and Detail-box 6).

In the basic Bayesian Fellegi-Sunter framework the m and u probabilities were assumed to follow a probability distribution of some type; for example, a uniform distribution with all possible probability values being equally likely, or a Beta distribution, with the parameters of the distribution set based on expert knowledge. The optimal (mean) values from the posterior distributions of the m and u probabilities were then used to create the match weights and complete the linkage process.

For a simple example of a Bayesian model, have a look at Detail-box 6.

Several Bayesian extensions have been proposed in the literature and can be found in Fortini et al. (2001), Sadinle (2014), Sadinle (2017), Marchant et al. (2021), Tancredi and Liseo (2011), Ventura and Nugent (2014), McVeigh et al. (2019). In particular, McVeigh et al. (2019) addresses one of the main issues of the Bayesian approaches, which is their computational burden, which makes them difficult to implement with large datasets. They propose a blocking approach based on simpler probabilistic record linkage techniques. That is, the output of a simpler non-Bayesian probabilistic record linkage is used to perform "post hoc blocking", after which a Bayesian Fellegi-Sunter method is used for coherent modelling and uncertainty quantification. This allows the authors to scale their proposed method to voter registration and census datasets with millions of entries.

3)MachineLearningApproaches Computer science researchers have approached data linkage from a traditional binary supervised classification perspective (with the two classes being "matches" and "non-matches", but with no "potential matches"), or from a clustering perspective (where the aim is to group all records that refer to the same entity into one cluster) (Christen, 2019).

In machine learning, probabilistic record linkage can be replaced by one of several classification algorithms, whose goal is the creation of 3 clusters (links, non-links, possible links) which are formed to match the three regions of match weights in the Fellegi-Sunter algorithm (Asher et al., 2020). The ML approaches can be divided into:

- SUPERVISED: a training set of data is used to "teach" a classification algorithm, such as decision trees, support vector machines, ensemble methods (i.e., random forests), or conditional random fields.
- UNSUPERVISED: a method that does not rely on training data. An example is k-means clustering. In k-means clustering, for each pair, a measure of similarity for each field is calculated. Multiple similarity measures have been used within unsupervised record linkage Asher et al. (2020).

Once each of the pairs of records has an associated similarity measure vector, the distance between different pairs of records is measured. Different distance measures can be used, a common one is the Euclidean distance (i.e., summing the squared distance across all the fields being compared). Record pairs that are "close" to each other according to the distance measure are formed into clusters.

Fully supervised methods do not exploit the information provided by unlabelled examples; instead, they rely on larger numbers of labelled pairs. Given the significant class imbalance when considering record pairs (very few pairs match), vast amounts of reliable training data or carefully selected training data are required for the use of these methods. These training data may come from crowdsourcing or extensive manual record linkage efforts or they may be automatically generated using unsupervised methods to obtain an approximate training set. However, in practice, the amount of reliable training data necessary to train sophisticated learning algorithms such as deep neural networks are not always available for record linkage tasks. Some examples can be found in Kooli et al. (2018) and Kasai et al. (2019). Moreover, the use of deep learning techniques in entity resolution is especially promising in application to unstructured or textual problems, where, for instance, pre-trained language models can be used. For structured record linkage, simple classifiers (such as logistic regression, decision trees, random forests, Bayesian additive regression trees, and others) are often preferred.

Clustering-based approaches to record linkage can integrate multiple databases (Sadinle (2014), Sadinle (2017), Marchant et al. (2021), Tancredi and Liseo (2011), Ventura and Nugent (2014) etc.). Many clustering approaches to entity resolution are based on pairwise similarities, pairwise match probabilities, or determined links and non-links. Therefore, they can be seen as post-processing the result of other pairwise record linkage procedures. Clustering can be for instance used as a postprocessing step, namely as a second step to probabilistic record linkage to enforce transitivity of the output (for a review and concrete examples see Christophides et al., 2020; Monge, 1997; Ventura and Nugent, 2014).

1.1.3.The problem of assessing uncertainty

As emphasised multiple times throughout this discussion, an essential concern in the realm of record linkage and the creation of a synthetic dataset, particularly in the context of post-merge analysis, revolves around the evaluation of the uncertainty inherent in the linkage process. It is crucial to acknowledge that the linkage process is inherently imperfect, introducing a layer of error that further compounds the uncertainties present in statistical analyses. Remarkably, linkage errors are seldom, if at all, considered in the realms of inference or data dissemination. This absence of attention to linkage error means that its potential impact on the accuracy and reliability of data-driven conclusions remains largely unaddressed, highlighting the need for more comprehensive strategies to quantify and manage such uncertainties (for more detail see among other Reiter, 2021).

The general sources of uncertainty in Record Linkage are:

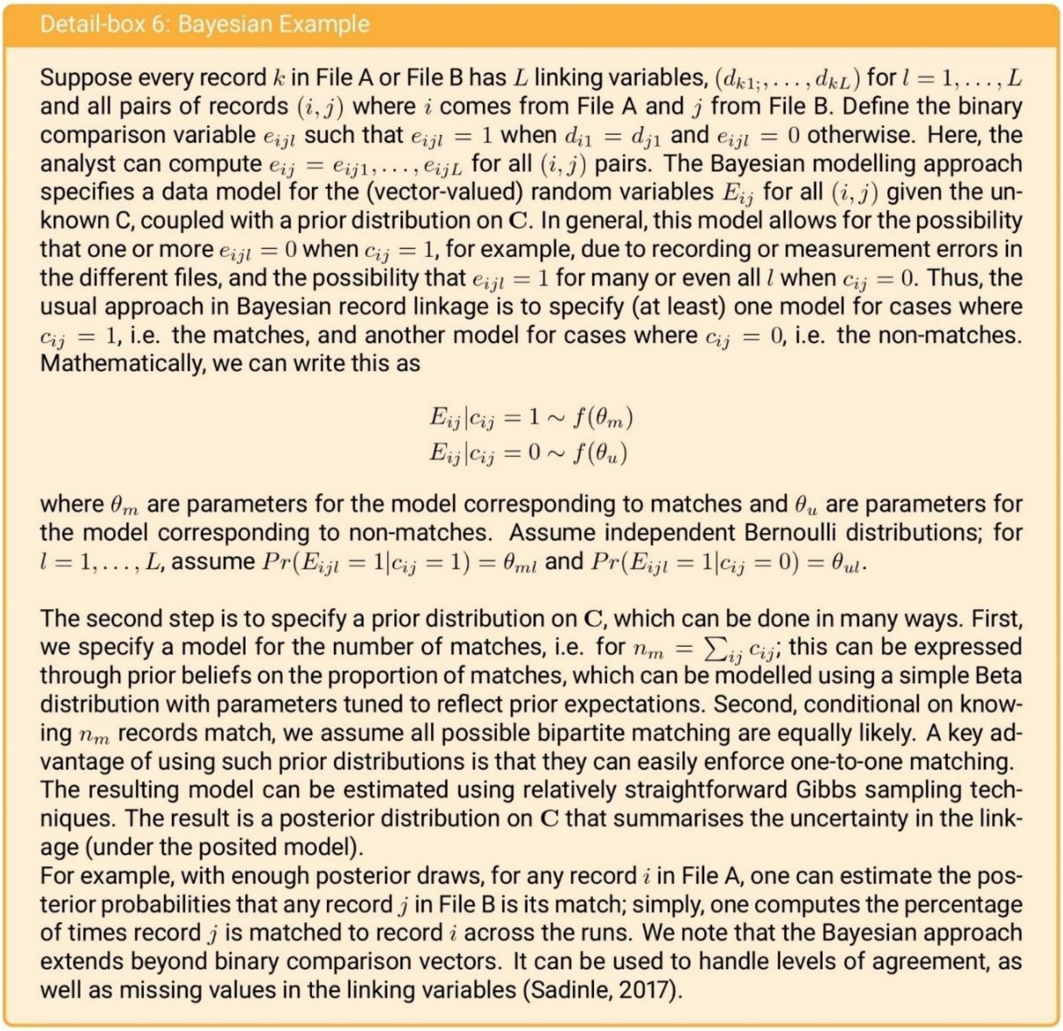
1. **Imperfect Matching**: Numerous record linkage methods result in a consolidated dataset containing a collection of record pairs that are considered the best matches based on specific criteria. In situations where perfect and unique identifiers are not available, this set may inadvertently include false matches. Furthermore, measurement errors can extend to the blocking variables used in the linkage process, further complicating the determination of true matches. The presence of these false matches can present challenges when drawing inferences from the linked data. When aggregated across many individuals, these errors can significantly distort estimates of the variables of interest, as well as their associated standard errors.
2. **Incomplete Matching**: Incomplete matching in the context of record linkage signifies that not all individuals from the primary dataset find corresponding counterparts in the secondary dataset. Incomplete matching can introduce a subtle yet impactful “selection mechanism” that has the potential to skew the outcomes of statistical inferences. Even in cases where no overt

selection mechanisms are at play, and the linkage process can be considered “ignorable” for the analysis (as defined by Rubin, 1976), it is important to recognise that solely utilising linked cases results in the partial utilisation of available information. This selective approach can lead to inflated variances in the inferences drawn.

Researchers have developed a variety of approaches to account for uncertainties in record linkage. The most common are: (i) modelling the matching matrix that indicates who matches whom as a parameter, (ii) embedding the linkage in a particular modelling task, and (iii) the imputation of incomplete links to address biases from incomplete matching.

Concerning the former, one effective approach is to leverage Bayesian methodologies, as exemplified in works like those by Fortini et al. (2001), Tancredi and Liseo (2011), Sadinle (2017). For more details on the Bayesian approach to account for uncertainty see Detail-box 6.

To account for the uncertainty in the linkage process, one can also embed the record linkage process directly in an analysis model, usually some regression involving (potentially multivariate) X and Y*,* to incorporate uncertainty in the inferences. Methods for doing so take two main forms: the adjustment approach (see Detail-box 7) and the use of hierarchical models. For the latter approach, the method implies the use of multiple imputations to fill in missing items for non-matched cases (for more details see Gutman et al., 2013).



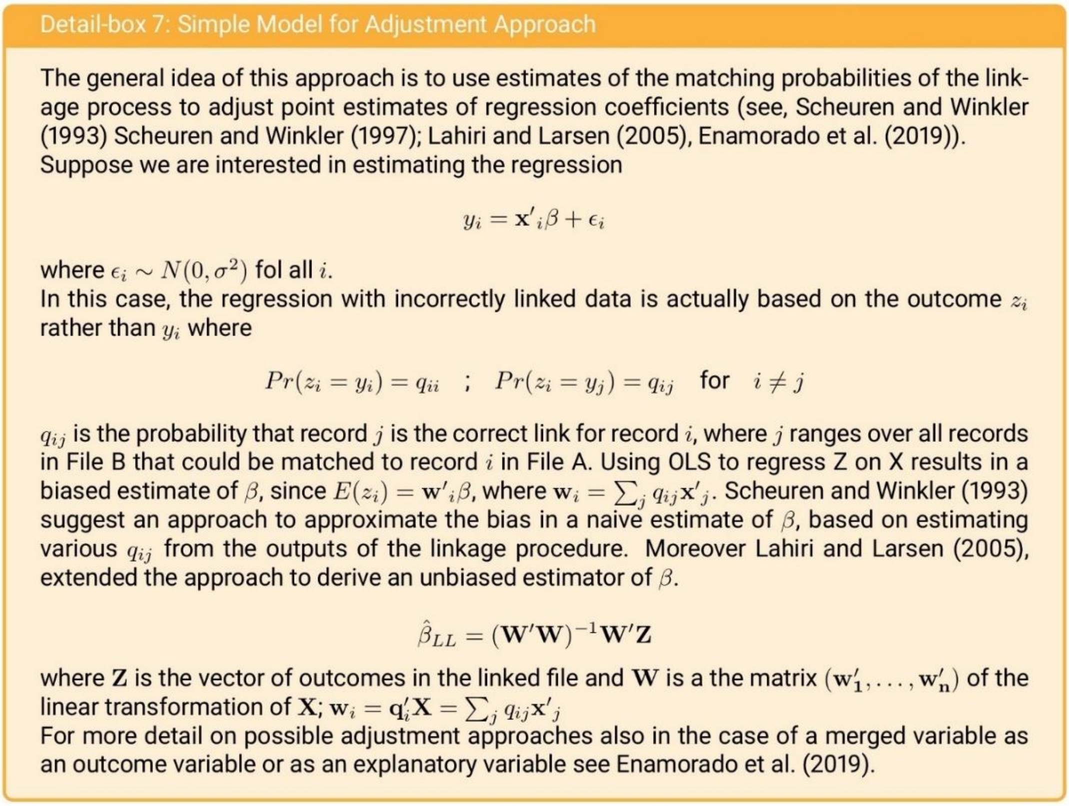
Detail-box 6: Bayesian Example

Suppose every record k in File A or File B has L linking variables, $(d_{k1;}, \dots, d_{kL})$ for $l = 1, \dots, L$ and all pairs of records (i, j) where i comes from File A and j from File B. Define the binary comparison variable e_{ijl} such that $e_{ijl} = 1$ when $d_{i1} = d_{j1}$ and $e_{ijl} = 0$ otherwise. Here, the analyst can compute $e_{ij} = e_{ij1}, \dots, e_{ijL}$ for all (i, j) pairs. The Bayesian modelling approach specifies a data model for the (vector-valued) random variables E_{ij} for all (i, j) given the unknown C, coupled with a prior distribution on $\mathbf{C}$. In general, this model allows for the possibility that one or more $e_{ijl} = 0$ when $c_{ij} = 1$, for example, due to recording or measurement errors in the different files, and the possibility that $e_{ijl} = 1$ for many or even all l when $c_{ij} = 0$. Thus, the usual approach in Bayesian record linkage is to specify (at least) one model for cases where $c_{ij} = 1$, i.e. the matches, and another model for cases where $c_{ij} = 0$, i.e. the non-matches. Mathematically, we can write this as

$$E_{ij}|c_{ij} = 1 \sim f(\theta_m)$$
$$E_{ij}|c_{ij} = 0 \sim f(\theta_u)$$

where θ_m are parameters for the model corresponding to matches and θ_u are parameters for the model corresponding to non-matches. Assume independent Bernoulli distributions; for $l = 1, \dots, L$, assume $Pr(E_{ijl} = 1|c_{ij} = 1) = \theta_{ml}$ and $Pr(E_{ijl} = 1|c_{ij} = 0) = \theta_{ul}$.

The second step is to specify a prior distribution on $\mathbf{C}$, which can be done in many ways. First, we specify a model for the number of matches, i.e. for $n_m = \sum_{ij} c_{ij}$; this can be expressed through prior beliefs on the proportion of matches, which can be modelled using a simple Beta distribution with parameters tuned to reflect prior expectations. Second, conditional on knowing n_m records match, we assume all possible bipartite matching are equally likely. A key advantage of using such prior distributions is that they can easily enforce one-to-one matching. The resulting model can be estimated using relatively straightforward Gibbs sampling techniques. The result is a posterior distribution on $\mathbf{C}$ that summarises the uncertainty in the linkage (under the posited model).
For example, with enough posterior draws, for any record i in File A, one can estimate the posterior probabilities that any record j in File B is its match; simply, one computes the percentage of times record j is matched to record i across the runs. We note that the Bayesian approach extends beyond binary comparison vectors. It can be used to handle levels of agreement, as well as missing values in the linking variables (Sadinle, 2017).



Detail-box 7: Simple Model for Adjustment Approach

The general idea of this approach is to use estimates of the matching probabilities of the linkage process to adjust point estimates of regression coefficients (see, Scheuren and Winkler (1993) Scheuren and Winkler (1997); Lahiri and Larsen (2005), Enamorado et al. (2019)). Suppose we are interested in estimating the regression

$$y_i = \mathbf{x}'_i\beta + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ fol all i.
In this case, the regression with incorrectly linked data is actually based on the outcome z_i rather than y_i where

$$Pr(z_i = y_i) = q_{ii} \quad ; \quad Pr(z_i = y_j) = q_{ij} \quad \text{for} \quad i \neq j$$

q_{ij} is the probability that record j is the correct link for record i, where j ranges over all records in File B that could be matched to record i in File A. Using OLS to regress Z on X results in a biased estimate of β, since $E(z_i) = \mathbf{w}'_i\beta$, where $\mathbf{w}_i = \sum_j q_{ij}\mathbf{x}'_j$. Scheuren and Winkler (1993) suggest an approach to approximate the bias in a naive estimate of β, based on estimating various q_{ij} from the outputs of the linkage procedure. Moreover Lahiri and Larsen (2005), extended the approach to derive an unbiased estimator of β.

$$\hat{\beta}_{LL} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Z}$$

where $\mathbf{Z}$ is the vector of outcomes in the linked file and $\mathbf{W}$ is a the matrix $(\mathbf{w}'_1, \ldots, \mathbf{w}'_n)$ of the linear transformation of $\mathbf{X}$; $\mathbf{w}_i = \mathbf{q}'_i\mathbf{X} = \sum_j q_{ij}\mathbf{x}'_j$
For more detail on possible adjustment approaches also in the case of a merged variable as an outcome variable or as an explanatory variable see Enamorado et al. (2019).

1.2.Statistical Matching

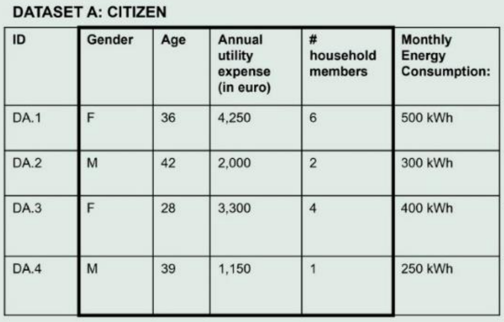
Statistical matching refers to a family of methods aiming at the integration of two or more datasets drawn from the same target population, that contain information on a set of common variables, X, and some not jointly observed variables, (Y,Z). Different from record linkage, statistical matching deals with data sources whose units are not necessarily overlapping.

There are two main approaches to statistical matching: a *macro approach* and a *micro approach*. On the one hand, the macro approach to statistical matching aims at directly estimating the joint model of the variables of interest that are not jointly observed. On the other hand, the micro approach is devoted to the generation of a synthetic dataset with complete information on the variables observed only in one data source and those observed in two data sources. The dataset produced as an output of the procedure is said to be *complete* because all the variables of interest are contained in it; it is said to be *synthetic* because it is not a product of direct observation of a set of units in the population of interest, but it is obtained by exploiting information in the source files in some appropriate ways. The boundary between the two approaches is not clearly defined. Although the main objective of the macro approach is not the creation of a complete dataset, a synthetic dataset may be obtained as a by-product of the estimation procedure; We will briefly discuss this approach in Section 1.2.3.

In practice, matching procedures devoted to the creation of a complete dataset can be regarded as an imputation problem of the target variables from a donor to a recipient survey (D'Orazio et al., 2006). The relation between the common variables with the target variables observed only in one of the datasets -the donor dataset- is explored and used to impute to the units of the other dataset -the recipient dataset- the variables that are not directly observed (see Detail-box 8) Thus, a synthetic dataset is generated with complete information on the target variables and the common ones.

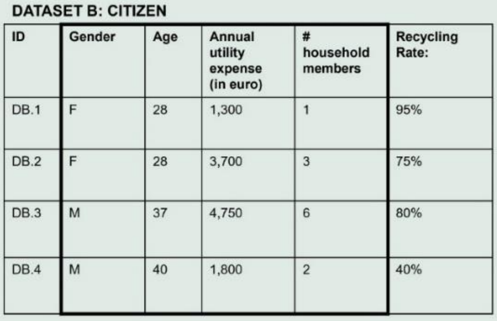
Toy Example 2: Creation of of a synthetic dataset via statistical matching

Here is another example of a complication one can encounter in dealing with data integration when the aim is to create a synthetic dataset. As in the first toy example, let us suppose we have two distinct datasets that are collecting information on sustainability practices and energy consumption of individuals in Ireland. Specifically, dataset A collects information on energy consumption, whereas dataset B includes data on recycling rates. In this application, we do not care about the design behind the datasets. The two datasets collect information on the same target population of individuals, but, in this case, we do not have quasi-identifier variables (i.e., name and address). The only common information we have are four common variables that describe the individuals and their households.



DATASET A: CITIZEN

ID	Gender	Age	Annual utility expense (in euro)	# household members	Monthly Energy Consumption:
DA.1	F	36	4,250	6	500 kWh
DA.2	M	42	2,000	2	300 kWh
DA.3	F	28	3,300	4	400 kWh
DA.4	M	39	1,150	1	250 kWh



DATASET B: CITIZEN

ID	Gender	Age	Annual utility expense (in euro)	# household members	Recycling Rate:
DB.1	F	28	1,300	1	95%
DB.2	F	28	3,700	3	75%
DB.3	M	37	4,750	6	80%
DB.4	M	40	1,800	2	40%

The aim is to integrate the two data sources leveraging the common variables to study the relationship between the two sets of variables not jointly observed, in this case "monthly energy consumption" and "recycling rate". Thus, the objective is to create a complete "synthetic" dataset, where for each row (a record) none of the variables is missing. Since we do not observe all the variables jointly, we need to leverage statistical methods to exploit the information contained in the distinct files.

Differently from record linkage, when setting a statistical matching procedure it is essential to explicit our assumptions. The most common situation is when the variables that are never jointly observed are assumed to be independent conditionally on the available covariates; this is the *Conditional Independence Assumption* (CIA; see Detail-box 8).

Detail-box 8: Conditional Independence Assumption

Consider two datasets A and B; we observe (Y, X) in A and (Z, X) in B. Denote with $f(Y, Z, X)$ the joint distribution of Y, Z, X. For the rules of probability, we can write

$$f(Y, Z, X) = f(Y \mid Z, X) f(Z \mid X) f(X) \,,$$

where $f(Y \mid Z, X)$ is unknown because we never observe Y and Z jointly. We say that Y and Z are conditionally independent given X, i.e., $Y \perp\!\!\!\perp Z \mid X$, if we can write

$$f(Y, Z, X) = f(Y \mid X) f(Z \mid X) f(X) \,.$$

Now, the elements on the right-hand side of the equation above are all known.

Under the CIA, statistical matching procedures ensure that the marginal and joint distribution of the variables in the source files is reflected in the statistically matched file (Rässler, 2012). Any discrepancy between the real data generation model and the underlying model of the synthetic complete dataset is called "matching noise".

Statistical matching can be performed in a parametric and a nonparametric framework.

1.2.1 Parametric micro approach

Consider two data sources *A* and *B*; we observe (Y, X) in A and (Z,X) in B. We are interested in learning something about the joint distribution of Y and Z (or Y, Z,X). Under the CIA, the joint relation between the common variables and the variables not jointly observed can be modelled (see Detail-box 8). In other words, we assume a parametric model for the joint distributions of (Y, X) and (Z,X). Once the parametric model is estimated, Z is imputed for each unit in A and Y is imputed for each unit in B. For the imputation, different techniques, such as conditional mean imputation and imputation from a distribution, can be implemented. See Toy Example 2.1 for an illustration of the method.

Toy Example 2.1: Parametric micro approach to statistical matching

Recall the Toy Example 2.

- Assumption: We assume conditional independence between the monthly energy consumption (Y) and the recycling rate (Z) given the other variables.
- Common variables: For the sake of simplicity, in this toy example, one could consider just one common variable, e.g., annual utility expense (X). More expert readers may be familiar with the matrix notation $\boldsymbol{X} = (X_1, X_2, X_3, X_4)$, where the variable X is now a vector of variables; in other words, the set of common variables include gender (X_1), age (X_2), annual utility expense (X_3) and no. of household members (X_4).
- Approach: Parametric.
  1. We assume a parametric model on the relation between the monthly energy consumption and the annual utility expense, e.g.,

     $$\log Y_i = \alpha + \beta X_i + \varepsilon_i$$

     for all units i in dataset A. We estimate the model via OLS or maximum likelihood and obtain $\hat{\alpha}$ and $\hat{\beta}$. Using the estimates and the annual utility expense of all units j in dataset B, we can impute the monthly energy consumption $\hat{Y}_j$, e.g., drawing values from the estimated model.
  2. We assume a parametric model on the relation between the recycling rate and the annual utility expense, e.g.,

     $$\log Z_j = \gamma + \delta X_j + \eta_j$$

     for all units j in dataset B. We estimate the model via OLS or maximum likelihood and obtain $\hat{\gamma}$ and $\hat{\delta}$. Using the estimates and the annual utility expense of all units i in dataset A, we can impute the recycling rate $\hat{Z}_i$, e.g., drawing values from the estimated model.
- Outcome: A complete synthetic dataset as follows, where the bold entries are estimates from the models described above:

ID	**Gender**	**Age**	**Annual utility expense (in euro)**	**# of household members**	**Monthly energy consumption**	**Recycling rate**
DA.1	F	36	4,250	6	500 kWh	**73%**
DA.2	M	42	2,000	2	300 kWh	**67%**
DA.3	F	28	3,300	4	400 kWh	**65%**
DA.4	M	39	1,150	1	250 kWh	**72%**
DB.1	F	28	1,300	1	**489 kWh**	95%
DB.2	F	28	3,700	3	**308 kWh**	75%
DB.3	M	37	4,750	6	**413 kWh**	80%
DB.4	M	40	1,800	2	**240 kWh**	40%

1.2.2.Nonparametric micro approach

When we do not want to consider the possibility of assuming a distribution in advance, nonparametric frameworks are favored. One possibility is to use nonparametric estimation techniques to infer the relations between X and Y , and X and Z and then proceed as under the parametric framework just described. Another possibility is to avoid the estimation step and fill the missing values with existing ones; these kinds of procedures belong to the *hot deck* imputation family.

Especially when the matching variables are categorical, a popular technique is the “random hot deck”. It consists of randomly choosing a record in the donor file for each record in the recipient file. The pairing is often done within strata or donation classes; their function is similar to the “blocking variables” of record linkage procedures. Whereas in the case of quantitative variables, the most popular nonparametric techniques are the “Distance Hot Deck” ones (Okner, 1972; Rodgers, 1984; Ruggles and Ruggles, 1974); each record in the recipient file is matched with the closest record in the donor file, according to a distance measure computed using the matching variables. When two or more donor records are equally distant from the recipient record, one of them is chosen at random.

When each record in the donor file can be used as a donor only once, the distance hot deck procedure is said to be “constrained”. The main advantage of a constrained hot deck approach is that the marginal distribution of the imputed variable is maintained in the final synthetic file. On the other hand, the average distance of the donor and recipient values of the matching variables X is expected to be greater than that in an unconstrained case. D’Orazio et al. (2006) underlines the importance of the choice of the recipient file, which is usually the one to be used as the basis for further statistical analyses. For the sake of accuracy, as a rule of thumb, when the sizes of the two data sources are very different it is better to choose the smallest to be the recipient. In this way, the risk that the distribution of the imputed variable does not reflect the original one (estimated from the donor dataset) is low.

The R package StatMatch (D’Orazio, 2022) refers to the distance hot deck method as “nearest neighbor distance hot deck”, and implements it in the function NND.hotdeck (see Section 3 for references on the R packages to implement statistical matching). The function searches in the donor file the nearest neighbor of each unit in the recipient file according to a distance function computed on the matching variables. To reduce the effort in computing distances, D’Orazio (2011) suggests defining some donation classes, usually defined according to one or more categorical common variables.

For a simple example, see Toy Example 2.2.

Toy Example 2.2: Nonparametric micro approach to statistical matching

Recall the Toy Example 2.

- Assumption: We assume conditional independence between the monthly energy consumption (Y) and the recycling rate (Z) given the other variables.
- Common variables: For the sake of simplicity, in this toy example, one could consider just one common variable, e.g., annual utility expense (X). See the previous Detail-box for the generalisation to the multivariate case.
- Approach: Nonparametric - distance hot deck. It consists of two steps:
  1. Assign to either dataset A or B the role of the recipient: Let us give dataset A the role of the recipient.
  2. For each record in dataset A, measure the distance between its value of the matching variable X, namely the annual income, and the value of X of each record in B.

ID		DB.1	DB.2	DB.3	DB.4
	X	1.3	3.7	4.75	1.8
DA.1	4.25	|4.25-1.3|=2.95	|4.25-3.7|=0.55	**|4.25-4.75|=0.5**	|4.25-1.8|=2.45
DA.2	2	|2-1.3|=0.7	|2-3.7|=1.7	|2-4.75|=2.75	**|2-1.8|=0.2**
DA.3	3.3	|3.3-1.3|=2	**|3.3-0.37|=0.4**	|3.3-4.75|=1.45	|3.3-1.8|=0.15
DA.4	1.15	**|1.15-1.3|=0.15**	|0.115-3.7|=2.55	|1.15-4.75|=3.6	|1.15 -1.8|=0.65

- Outcome: The pairs of records with the minimum distance (above in bold) become matches.

ID	Gender	Age	Annual utility expense	# of Household	Monthly energy consumption	Matched ID	Recycling rate
DA.1	F	36	4.25	3	300 kWh	DB.3	**75%**
DA.2	M	42	2	1	250 kWh	DB.1	**40%**
DA.3	F	28	3.3	2	400 kWh	DB.2	**80%**
DA.4	M	39	1.15	0	350 kWh	DB.4	**95%**

1.2.3.Macro approach

Beyond the micro approach described in the previous sections, statistical matching provides another apparently different way to pursue data integration: the so-called *macro* approach (D'Orazio et al., 2006).

The macro approach uses the source files in order to have a direct estimation of the joint distribution function (or of some of its key characteristics, such as the correlation) of the variables of interest that have not been observed in common.

Most of the statistical matching methods are built on the aforementioned CIA. If the CIA holds, the joint density of the variables of interest in the two data sources and the matching variables can be factorised into the product of conditional densities (the pairwise relationship between the target variables and the matching ones) times the marginal density of the matching variables; this way, the model is identifiable and the macro approach consists of directly estimating it.

It is important to underline that the CIA cannot be tested; it could be a wrong and misleading assumption if it is introduced when it does not hold.

1.2.4 Assessing uncertainty

As mentioned before, the CIA cannot be tested since the variables of interest Y and Z are never jointly observed; when it is likely to be a misspecified assumption, more *uncertainty* must be yielded in the model.

In particular, as firstly suggested by Rubin (1986), one should consider a set of plausible parameters rather than a point estimate and generate a collection of synthetic datasets, rather than a single one. In the context of *multiple imputation*, fully Bayesian techniques are well-suited.

Detail-box 9: Statistical matching uncertainty

D'Orazio et al. (2006) formally define the statistical matching uncertainty. Assume to know exactly the partial joint distributions of (Y, X) and (Z, X), and denote with $\boldsymbol{\theta}_{YX}$ and $\boldsymbol{\theta}_{ZX}$ the parameters of such distributions. Denote with $\boldsymbol{\theta} \in \Theta$ the unknown parameter, defined in the parameter space Θ, defining the full joint distribution of (Y, X, Z). If the CIA does not hold, we still can say that the plausible values of $\boldsymbol{\theta}$ are those belonging to a subset of Θ, Θ^{SM} that satisfies the constraints given by the values of $\boldsymbol{\theta}_{YX}$ and $\boldsymbol{\theta}_{ZX}$.

2.Combining probability and nonprobability samples

Probability sampling continues to be the gold standard for acquiring a representative sample: in probability samples, the selection probability is known and therefore the estimations from a probability sample can rely on well-known statistical inference methods, often design-based. Nonetheless, the measurement of the study variable can also be derived from a non-probability sample or from *big data* sources.

The increasingly high costs for probability surveys, coupled with high non-response rates and the need to provide information at more granular levels have led to the necessity over time to integrate information from multiple sources.

Initially, methods were suggested for integrating multiple probability samples. Recently, in survey statistics, there has been a surge in the availability of nonprobability data for research purposes, offering unprecedented opportunities for new scientific discoveries.

However, they also pose additional challenges, including issues such as heterogeneity, selection bias, high dimensionality, and more. Specifically, this has emphasized the necessity to propose new suitable methods for combining probability and emerging nonprobability samples, as well as for merging probability samples with big nonprobability datasets as outlined in the review given by Yang and Kim (2020). Before describing these methods, let's briefly review the techniques for integrating probability samples.

Existing methods for probability data integration can be classified into two types, depending on the level of information to be combined: a macro approach combining the summary statistics from multiple surveys, and a micro approach creating synthetic imputations.

The macro approach was employed to combine data from two independent probability samples for estimating totals at the population and domain levels by Renssen and Nieuwenbroek (1997), Hidiroglou (2001), Merkouris (2004), Wu (2004), Ybarra and Lohr (2008), and (Merkouris, 2010). (Merkouris, 2004) and Merkouris (2010) provided a rigorous treatment of the survey integration through the generalized method of moments. Fuller and J. (1999) describe an application in the National Resource Inventory.

Very briefly, the proposals in this approach seek to devise a composite estimation method for aligning estimators of common characteristics in scenarios involving multiple surveys on the same population or multiple samples within the same survey. In addition, they allow to generate more efficient estimators for non-common variables, exploiting the strength of their correlation with the common ones.

In the micro approach, mass imputation (also called synthetic data imputation) is the commonly used technique for creating imputed values for items not observed in the current survey by incorporating information from other surveys. For simplicity, let's consider the case of two probability samples, a small sample A and a large sample B. In sample A, observe auxiliary information X and outcome Y, whereas in sample B, observe common auxiliary information X. The primary goal is to create proxy values of Y for the units in B, and then to use these values together with the associated sample weights in B to produce projection estimators. The proxy values are generated by first fitting a working model relating Y to X, based on the data from sample A used as a sort of training sample.

When we have more datasets and the missingness structure is not monotone, the mass imputation is still used but it becomes more complicated. A sample with partial information may contain additional information for parameter estimation. In such a situation a joint model of all variables needs to be considered, and the EM algorithm can be used to estimate the model parameters. Kim et al. (2016) used an instrumental variable assumption for model identification and developed fractional imputation methods for statistical matching. Park and Kim (2016) presented an application of the statistical matching technique using fractional imputation in the context of handling mixed-mode surveys. Park et al. (2017) applied the method to combine two surveys with measurement errors.

Toy Example 3: Combining probability and nonprobability samples

Suppose we are interested in estimating the average household's monthly energy consumption of our target population, but the only available information comes from energy suppliers; thus, in statistical terms, it is a nonprobability sample. This dataset is the synthetic dataset that was created by integrating Dataset A and Dataset B with one of the data integration techniques explained in Toy Example 1 or Toy Example 2. Let us suppose the dataset supplied by the energy suppliers (Dataset AB) was created by us using Record Linkage techniques.

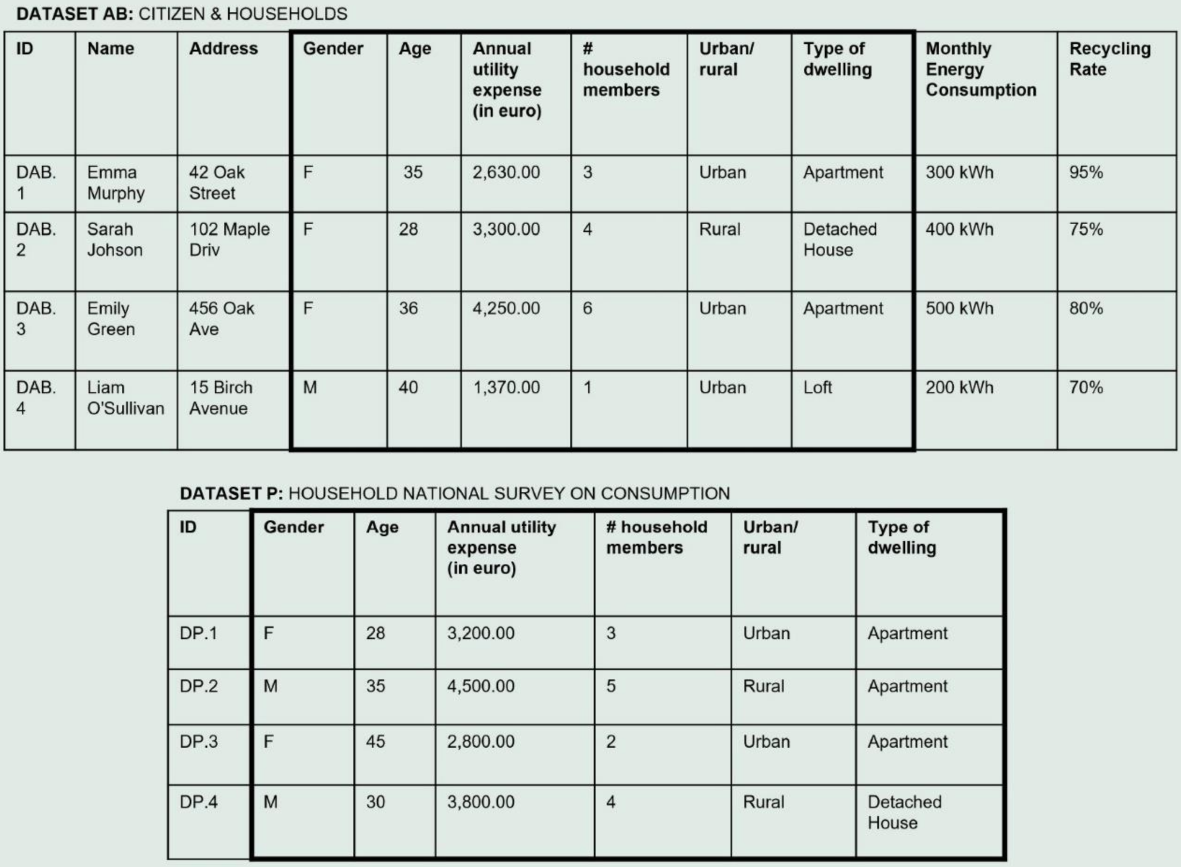
If we simply use this data source to make inferences, for instance, to investigate the relationship between "recycling rate" and "monthly energy consumption" or to compute the average of "monthly energy consumption" in the target population, our estimate will be biased. The bias arises from the nonprobability nature of Dataset AB. Indeed, we do not know if Dataset AB represents the target population since it is a nonprobability sample and we do not know the selection mechanism governing the selection of individuals into the sample.

However, luckily we know that the National Institute of Ireland annually surveys Irish households and their consumption behaviours (probability sample) therefore we know we have access to auxiliary variable information (i.e., a set of common covariates) from this probability survey sample (Dataset P). Multiple methods can be used to leverage this information coming from different data sources to make a reliable and representative inference and estimate the average household's monthly energy consumption.

- **Dataset AB (Nonprobability Sample):**
  Y = monthly energy consumption
  Z = recycling rate
  X = gender, age, annual utility expense, no. of household members, urban/rural, type of dwelling
- **Dataset P (Probability Survey Sample):**
  X = gender, age, annual utility expense, no. of household members, urban/rural, type of dwelling

The aim is to estimate the average household's monthly energy consumption μ_y. As mentioned above, because the sampling mechanism of a nonprobability sample is unknown, the target population quantity is not identifiable in general. Indeed, the sampling mechanism of dataset AB is unknown and, therefore, μ_y is not identifiable in general.

Toy Example 3: Visualization of Nonprobability Synthetic Dataset & Probability Dataset



DATASET AB: CITIZEN & HOUSEHOLDS

ID	Name	Address	Gender	Age	Annual utility expense (in euro)	# household members	Urban/ rural	Type of dwelling	Monthly Energy Consumption	Recycling Rate
DAB. 1	Emma Murphy	42 Oak Street	F	35	2,630.00	3	Urban	Apartment	300 kWh	95%
DAB. 2	Sarah Johson	102 Maple Driv	F	28	3,300.00	4	Rural	Detached House	400 kWh	75%
DAB. 3	Emily Green	456 Oak Ave	F	36	4,250.00	6	Urban	Apartment	500 kWh	80%
DAB. 4	Liam O'Sullivan	15 Birch Avenue	M	40	1,370.00	1	Urban	Loft	200 kWh	70%

DATASET P: HOUSEHOLD NATIONAL SURVEY ON CONSUMPTION

ID	Gender	Age	Annual utility expense (in euro)	# household members	Urban/ rural	Type of dwelling
DP.1	F	28	3,200.00	3	Urban	Apartment
DP.2	M	35	4,500.00	5	Rural	Apartment
DP.3	F	45	2,800.00	2	Urban	Apartment
DP.4	M	30	3,800.00	4	Rural	Detached House

2.1.Dealing with the biased nonprobability samples

As outlined in the “Summary Report of the AAPOR Task Force on Nonprobability Sampling" by Baker et al. (2013), unlike probability sampling, there is no single framework that comprehensively encompasses all nonprobability sampling. However, all nonprobability samples share some common traits. First, they operate with unknown selection or inclusion mechanisms; second, they tend to be biased; third, they may not faithfully represent the target population. Nevertheless, nonprobability samples are readily available data sources that are more cost-effective and quicker to obtain. Consequently, a central focus for many survey sampling researchers in recent years has been addressing the challenge of developing methods to derive valid statistical inferences from these samples. A review is given in Wu (2022).

All these methods stem from the recognition that generating valid inferences from nonprobability samples requires additional information from the target population. The likelihood of having access to complete auxiliary information is typically implausible in most scenarios. A popular framework is to assume that auxiliary variable information on the same population is available from an existing probability survey sample called reference probability sample. This framework was first used by Rivers (2007) and followed by several other authors including Vavreck and Rivers (2008), Lee and

Valliant (2009), Valliant and Dever (2011), Elliott (2017), and Chen et al. (2020), among others. Existing methods for integrating data from a probability sample and a nonprobability sample can be classified into the four types described in the following subsections respectively. However, one needs first to outline which are the assumptions needed to pursue the following data integrating techniques (see Detailbox 10). The main assumptions are three:

A1 The auxiliary variables included in the nonprobability sample fully characterise the participation behaviour or the sample inclusion mechanism for units in the population. Statistically, this means that the sample inclusion indicator variable of the nonprobability sample and the outcome variable are independent given the set of auxiliary variables.

A2 Every unit in the target population has a non-zero probability of being included in the nonprobability sample.

A3 Units' participations in the nonprobability sample are independent of each other given the set of auxiliary variables. This means that, in a nonprobability sample, each person's decision to participate in the sample is not affected by what others decide given some auxiliary variables, e.g. some specific factors like age, gender, or other relevant variables.

Detail-box 10: Strong Ignorability Assumption (Rosenbaum and Rubin, 1983)

Let us define two samples: a probability sample $\mathcal{S}_A$ and a nonprobability sample $\mathcal{S}_B$. Let us define the sample inclusion and participation indicator I_B as an indicator that takes value 1 if the unit i belongs to the sample $\mathcal{S}_B$ and 0 otherwise. Then, denote with $\pi_B(\boldsymbol{X}) = \mathsf{P}(I_B = 1 \mid \boldsymbol{X})$ the sampling score, i.e., the probability of belonging to $\mathcal{S}_B$ given the common variables $\boldsymbol{X}$. Strong ignorability assumption implies:

(i) **Ignorability Assumption**(Rubin, 1976; Little and Rubin, 2002): the conditional independence of I_B and the study variable Y given the common variables $\boldsymbol{X}$.

$$\pi_B(\boldsymbol{X}) = \mathsf{P}(I_B = 1 \mid \boldsymbol{X}, Y) = \mathsf{P}(I_B = 1 \mid \boldsymbol{X})$$

(ii) **Positivity Assumption**: $\pi_B(\boldsymbol{X})$ is strictly positive

$$\pi_{B,i}(\boldsymbol{X}) > 0, \quad \forall i$$

This assumption is valid when the set of variables $\mathbf{X}$ includes all factors relevant for predicting the outcome Y influencing the probability of being chosen in sample $\mathcal{S}_B$.

2.1.1.Propensity score weighting

The first method for combining probability and nonprobability samples is the so-called propensity score weighting (or adjustment) (Rosenbaum and Rubin, 1983). In this approach, the (unknown) probability of a unit being selected into the nonprobability sample, which is referred to as the propensity or sampling score, is modelled and estimated for all units in the nonprobability sample. The propensity scores for the nonprobability survey sample are theoretically defined for all the units in the target population. Estimation of propensity scores for units in the nonprobability sample requires an assumed model on the propensity scores and auxiliary information at the population level. Usually, the complete auxiliary information is not available, and under the setting assumed above (in which auxiliary information is available from an existing probability survey sample) the population auxiliary information is supplied by the reference probability sample. This leads to some differences in the estimation process. For example, when a parametric model is assumed, the maximum likelihood estimation method is replaced by the maximum pseudo-likelihood one. Three common parametric models are the logit model, the probit model, and the complementary log-log one. See Toy Example 3.1 for an illustration of the method. Non-parametric methods without assuming an explicit functional form for propensity scores can be an attractive alternative. The non-parametric kernel regression estimator for the propensity score is given by Yuan et al. (2022). Chu and Beaumont (2019) considered regression-tree-based method for estimating the propensity scores. Their method involving the combined sample of the nonprobability sample and the reference probability sample, seeks to build a classification tree. The terminal nodes of the final tree are employed as homogeneous groups in terms of propensity scores in a manner akin to post-stratification.

After the estimation of the propensity scores for the units in the nonprobability sample, the subsequent propensity score weights can be directly used to calculate one of the two versions of the inverse probability-weighted estimator (the adaptation of the Horvitz-Thompson estimator and the adaptation of the Hàjek estimator for missing data problems and causal inference). Both these estimators may be sensitive to small values of estimated propensity scores. A robust alternative may be a post-stratified estimator in which the strata are formed based on homogeneous groups in terms of propensity scores.

In general, both the subsequent adjustments, propensity score weighting and stratification can be used to adjust for selection biases; see, e.g., Lee and Valliant (2009), Elliott (2017) and Chen et al. (2020). (Stuart et al., 2011, 2015) and Buchanan et al. (2018) used propensity score weighting to generalise results from randomised trials to a target population. O'Muircheartaigh and Hedges (2014) proposed propensity score stratification for analysing a non-randomised social experiment.

Finally, it is worth noting a weakness of the propensity score methods, is that they rely on a propensity score model (explicitly defined under the parametric approach) and are biased and highly variable if the model is misspecified (Kang and Schafer, 2007).

Toy Example 3.1: Propensity Score Weighting

The aim is to estimate μ_y, the average household's monthly energy consumption (Y).

- Assumption: Strong ignorability (Rosenbaum and Rubin, 1983)
- Approach:
  1. Propensity Score Model Definition: We need to define the *propensity score function*, namely a model for the probability of being in the nonprobability sample based on the auxiliary information. In this example, we want to specify the probability for an individual to belong to the dataset AB as a function of the available common variables: gender, age, annual utility expense, no. of household members, urban/rural classification, and type of dwelling. For the sake of simplicity, let us consider just one common variable, e.g., annual utility expense (X). For instance, we can use a parametric method, thus we can write

     $$\text{logit}\,(\pi_{AB}(x_i)) = \alpha_0 + \alpha_X x_i\ , \quad \forall i\ .$$

  2. Propensity Score Model Estimation:
     If we assume the logit model specified above, we can estimate $\hat{\alpha}_0, \hat{\alpha}_X$, maximising the following pseudo-log-likelihood function (Chen et al., 2020). Note that the method requires the use of the pseudo-log-likelihood function instead of the log-likelihood function because the auxiliary information derives from a probability sample and not from the whole population.
  3. Prediction: The unknown $\pi_{AB}(x_i)$ can be predicted for each unit i in the dataset AB.

  To make inference using the nonprobability sample, for each unit in AB assign weights equal to the inverse of the estimated propensity scores.

- Outcome: Compute the weighted average of monthly energy consumption in the nonprobability sample AB as follows (adaptation of the Horvitz-Thompson estimator) :

  $$\hat{\mu}_{IPW1} = N^{-1} \sum_{i \in AB} \frac{1}{\widehat{\pi}_{AB}(x_i)} y_i$$

  An alternative estimator (adaptation of the Hàjek estimator) is given by:

  $$\hat{\mu}_{IPW2} = \widehat{N}_{AB}^{-1} \sum_{i \in AB} \frac{1}{\widehat{\pi}_{AB}(x_i)} y_i$$

  where $\widehat{N}_{AB} = \sum_{i \in AB} (\widehat{\pi}_{AB})^{-1}$

2.1.2.Calibration weighting

A second approach for combining probability and nonprobability samples uses calibration weighting (Deville and Särndal, 1992; Kott, 2006). This technique calibrates auxiliary information in the nonprobability sample with that in the probability sample, so that after calibration the weighted distribution of the nonprobability sample is similar to that of the target population. Rather than estimating the propensity score model and inverting the propensity score to address the selection bias of the nonprobability sample, the calibration strategy directly estimates the weights. The justification of the calibration approach, by Deville and Särndal (1992), relies on the linearity of the model for the target variable or on the linearity of the inverse probability of sampling weight, but the linearity conditions are unlikely to hold for non-continuous variables. More in general, the robustness of the approach relies on the correct specification of the calibration approach. It should be noted that not everyone in the literature categorises this approach as a standalone method: at times, it is referred to as a weighting approach that encompasses both the propensity score and the calibration techniques. See Toy Example 3.2 for an illustration of the method.

Toy Example 3.2: Calibration weighting

The aim is to estimate μ_y, the average household's monthly energy consumption (Y).

- Calibration Model:
  Use the annual utility expense as an auxiliary variable shared between the nonprobability (dataset AB) and probability survey samples (dataset P). We aim to calibrate the annual utility expense in the nonprobability dataset AB with that of the probability sample P so that after calibration dataset AB is similar to the target population.

- Weight Adjustment:
  We directly estimate the weights. We assign a weight $\omega_{AB,i}$ to each unit i in the sample AB so that

  $$\sum_{i \in AB} \omega_{AB,i} x_i = \sum_{i \in P} d_{P,i} x_i$$

  $\sum_{i \in P} d_{P,i} x_i$ is a design-weighted estimate of the population total of X (annual utility expense) from the probability sample P ($d_{P,i}$ are inclusion probabilities). $\mathcal{Q}_{AB} = \{\omega_{AB,i} : i \in AB\}$ are the calibration weights we need to estimate. The balancing constraint calibrates the covariate distribution of the nonprobability sample (dataset AB) to the target population in terms of X. We are in an optimisation problem setting and the above is the covariate balancing constraint. We estimate $\mathcal{Q}_{AB}$ by solving a constrained optimisation problem.

  $$\min_{\mathcal{Q}_{\mathcal{AB}}} \{L(\mathcal{Q}_{\mathcal{AB}}) = \sum_{i \in AB} \omega_{AB,i} \log \omega_{AB,i}\}$$

  subject to $\omega_{AB,i} \geq 0$ for all $i \in AB$, $\sum_{i \in AB} \omega_{AB,i} = N$ and the balancing constraint.

- Outcome:
  Calculate the weighted average of the monthly energy consumption in the dataset AB (adaptation of the Horvitz-Thompson estimator).

  $$\hat{\mu}_{cal1} = N^{-1} \sum_{i \in AB} \omega_{AB,i} y_i$$

  An alternative estimator (adaptation of the Hàjek estimator) is given by:

  $$\hat{\mu}_{cal2} = \widehat{N}_{AB}^{-1} \sum_{i \in AB} \omega_{AB,i} y_i$$

  where where $\widehat{N}_{AB} = \sum_{i \in AB} \omega_{AB,i}$

2.1.3.Mass imputation

The third method is the mass imputation or prediction approach, which imputes the missing values for all units in the probability sample. In the usual imputation for missing data analysis, the respondents in the sample constitute a training dataset for developing an imputation model. In the mass imputation, an independent nonprobability sample is used as a training dataset for developing the imputation model, and imputation is applied to all units in the probability sample; see, e.g., Breidt et al. (1996); Rivers (2007); Kim and Rao (2012); Chipperfield et al. (2012); Bethlehem (2016); Yang and Kim (2018). Different imputation models, parametric and non-parametric, have been suggested in the literature, the choice depends on the nature of the response variable but also on the type and quantity of auxiliary variables available for both samples (probability and nonprobability). Brick (2015) discussed diagnostics and model checking.

After imputing the values of the response variable for the units in the probability sample, these values, along with the weights associated with the units in the probability sample (i.e., the reciprocal of the inclusion probabilities), are used to calculate a design-based estimator for the descriptive parameter of interest. See Toy Example 3.3 for an illustration of the method. The consistency of this estimator relies on a correct specification of the imputation model and of the estimate of the model parameters.

Toy Example 3.3: Mass Imputation

The aim is to estimate μ_y, the average household's monthly energy consumption (Y). For the sake of simplicity, one could consider just one common variable, e.g., annual utility expense (X).

- Assumptions: strong ignorability (Rosenbaum and Rubin, 1983).
- Approach: Mass imputation.
  1. Model estimation: Let us specify a model for the monthly energy consumption as a function of the annual utility expense.

$$\log(y_i) = \beta_0 + \beta_X x_i + \varepsilon_i \ , \quad \forall\, i \in AB \ .$$

  We use the observations i in dataset AB to obtain the estimates $\hat{\beta}_0, \hat{\beta}_X$.

- Prediction: We use the estimates $\hat{\beta}_0, \hat{\beta}_X$ to predict $\hat{y}_j$ for $j \in P$.
- Outcome: We define the following mass imputation estimator, which allows us to compute the average monthly energy consumption in dataset P:

$$\hat{\mu}_I = N^{-1} \sum_{j \in P} d_{P,j} \hat{y}_j \ ,$$

  where $d_{P,j}$ are the design-weights of sample P.

2.1.4.Doubly Robust Procedure

To improve the robustness against model misspecification, this approach combines the weighting and the imputation approaches, employing both the propensity score and the outcome model.

The Doubly Robust (DR) estimator of the mean is composed of two terms, one is the model-based prediction of the mean (i.e. corresponds to the estimator used adopting the mass imputation approach), the other is a propensity score-based adjustment using the errors given by the differences between the observed and model-predicted response for the units in nonprobability sample. See Toy Example 3.4 for an illustration of the method. The magnitude of the adjustment term is negatively correlated to the goodness of fit of the outcome model. Consequently, the DR estimator is unbiased if either the propensity score model or the outcome model is correctly specified, but not necessarily both. Moreover, it is important to note that its double robustness property does not require the knowledge of which of the two models is correctly specified. This estimator has been proposed by Chen et al. (2020) under the two-sample setting assumed at the beginning of the section, and assuming a logistic model for the propensity score and a parametric regression model for the outcome. Note that the DB mean estimator suggested by Chen et al. (2020), is analogous to the model-assisted generalised difference estimator discussed in Wu and Sitter (2001) under scenarios where the complete (i.e. at the population level) auxiliary information is available.

Recently, Chen et al. (2022) suggested an alternative approach to produce a double robust estimator by using the pseudo empirical likelihood method and considering both the normalisation constraint (on the probability measure over the units in the nonprobability sample) and the model calibration constraint (on an assumed outcome regression model).

Toy Example 3.4: Double Robust Procedure

The aim is to estimate μ_y, the average household's monthly energy consumption (Y). For the sake of simplicity, one could consider just one common variable, e.g., annual utility expense (X).

- Assumptions: strong ignorability (Rosenbaum and Rubin, 1983)
- Approach: Double Robust Inference
  It employs both the propensity score function (Section **??**) and the outcome models (Section **??**) to improve the robustness of the estimator.
- Outcome:
  Let's compute the average monthly energy consumption by using the Double Robust estimator:

$$\hat{\mu}_{dr} = \hat{N}_{AB}^{-1} \sum_{i\in AB} \frac{1}{\hat{\pi}_{AB,i}(x_i)}\{y_i - \hat{y}_i\} + \hat{N}_P^{-1} \sum_{j\in P} d_{P,j}\hat{y}_j$$

  where $\widehat{N}_{AB} = \sum_{i\in AB}(\widehat{\pi}_{AB})^{-1}$ and $\widehat{N}_P = \sum_{j\in P} d_{P,i}$

2.2.Combining probability sampling and big data

In the past decade, more and more data became available, including large administrative record datasets, remote-sensing data derived from satellite images (McRoberts et al., 2010), mobile sensor data (Palmer et al., 2013), and web survey panels (Tourangeau et al., 2013). While such data sources provide timely data for a large number of variables and population elements, they are nonprobability samples and often fail to represent the target population of interest because of inherent selection biases. Thus, it is essential to combine such data with probability samples. How to combine such information with survey data to provide better estimates for the population parameters is a new challenge that survey statisticians face today. Tam and Clarke (2015) presented an overview of some initiatives of big data applications in official statistics of the Australian Bureau of Statistics.

Depending on the roles in statistical inference, big data can be classified into two types: one with large sample sizes (large n) and the other with a rich set of covariates (large p). In the first type, the nonprobability sample can be large in sample size. How to leverage the rich information in the big data to improve finite population inference is an important research topic that needs to be explored. In the second type, there are a large number of variables. There is a vast literature on variable selection methods for prediction, but very few contributions on variable selection for data integration that can successfully recognise the strengths and the limitations of each data source and utilise all captured information for finite population inference.

In cases where nonprobability data have large sample sizes (large n), it is crucial to differentiate between two scenarios based on whether the response variable is observed in the probability sample or not:

- When the probability sample includes the response variable, its mean, total, or other related parameters can be estimated by the commonly used estimator solely from the probability sample. In this situation, the auxiliary information in the big data can be used to improve this estimator. A common framework is to assume that the membership to the big data can be determined throughout the probability sample. Additionally, it is assumed the subsample of units in the probability sample constitutes a second-phase sample from the big data sample, which acts as a new population. Consequently, it is possible to calibrate the information in the second-phase sample to be the same as the new acting population. This idea has been explored by Yang and Ding (2020) and Kim and Tam (2021) among others. Kim and Tam (2021) have also implemented their calibration method, to incorporate big data auxiliary information, on the official statistics from the Australian Bureau of Statistics.
- When the response variable is present in the big dataset, but not in the probability sample, the situation is equivalent to the one described for combining probability and nonprobability samples. Hence, the same solution may be applied. Currently, the prevailing approach is mass imputation, wherein a predictive model is trained using big data and subsequently employed to impute the values of the response variable in the probability sample. Beyond parametric methods, nonparametric approaches such as nearest-neighbor imputation can also be considered (Yang and Kim, 2018). In the international forest inventory community, for combining ground-based observations with images from remote sensors, is popular to use a K-nearest-neighbour imputation strategy in which instead of using one nearest neighbour,

multiple nearest neighbours in the big data sample is identified, and the average response is used as the imputed value (McRoberts et al., 2010).

In the presence of big data of the second type, where there is a large number of auxiliary variables (large p), variable selection becomes crucial. This is essential for discerning the strengths and limitations of each data source and ensuring the utilisation of all and only the information relevant to inference about the finite population. With a large number of auxiliary variables existing integration methods may become unstable or even infeasible, since adding irrelevant auxiliary variables can introduce a large variability in the estimation. Gao and Carroll (2017) proposed a pseudo-likelihood approach for combining multiple non-survey data with high dimensionality; this approach requires all likelihoods to be correctly specified and therefore is sensitive to model mis-specification. Chen et al. (2018) proposed a model-based calibration approach using LASSO; this approach relies on a correctly specified outcome model. Yang (2020) proposed a doubly robust variable selection and estimation strategy that works in two steps. In the first step, it selects a set of variables that are important predictors of either the sampling score or the outcome model using penalised estimating equations. In the second step, it re-estimates the models' parameters by using the joint set of covariates selected from the first step and calculates the doubly robust estimator. This double robust estimator allows model misspecification of either the sampling score or the outcome model. Moreover, in the existing highdimensional causal inference literature, the doubly robust estimators have been shown to be robust to selection errors using penalisation (Farrell, 2015) or approximation errors using machine learning (Chernozhukov et al., 2018).

3.Implementation

This section provides useful information for the implementation of the methods described so far. The blue boxes below, one for each method introduced in this report, include the main packages for the most common software - R, Stata, and Python. For a general overview of the methods typically used in official and survey statistics that are implemented in R, see the CRAN Task View "CRAN Task View: Official Statistics & Survey Statistics"[5]. The task view is split into several parts:

- First part: "Producing Official Statistics". This first part is targeted at people working at national statistical institutes, national banks, international organizations, etc. who are involved in the production of official statistics and using methods from survey statistics.
- Second part: "Access to Official Statistics". This second part's target audience is everyone interested in using official statistics results directly from within R.
- Third part: "Related Methods" shows packages that are important in official and survey statistics, but do not directly fit into the production of official statistics.

[5] CRAN Task View:Official Statistics & Survey Statistics: https://cran.r-project.org/web/views/OfficialStatistics.html

Tools for Record Linkage

- **abeR** is an R package available on GitHub that makes allowance for errors in the data by not requiring matches to be exact. It uses an algorithm that specifies a set of decision rules to determine whether two records are sufficiently similar for them to be considered a match. A related Stata code is also available. The codes are based on methods described in Abramitzky et al. (2012, 2014, 2019).
- AMP is a met suggests a fully automated probabilistic method for linking historical datasets. The authors estimate these probabilities using the Expectation-Maximization (EM) algorithm. They suggest a number of decision rules that use these estimated probabilities to determine which records to use in the analysis. You can find the R code and Stata code on their website. The codes are based on methods described in Abramitzky et al. (2020).
- **blink** is an R package performing Bayesian entity resolution for categorical and text data, for any distance function defined by the user. In addition, the package allow the comparison to any other comparable method such as logistic regression, Bayesian additive regression trees (BART), or random forests. The package is based on the bipartite graph model by Steorts (2015).
- **Dedupe** is a Python library that uses machine learning to perform fuzzy matching, deduplication and entity resolution quickly on structured data. For more details on the methods offered by the package, see (Gregg and Eder, 2022).

Tools for Record Linkage

- **fastLink** is an R package that implements a Fellegi-Sunter probabilistic record linkage model that accommodates missing data and integrates auxiliary information. The software provides features for merging two datasets using the Fellegi-Sunter model, employing the Expectation-Maximization algorithm. The package is based on methods described in Enamorado et al. (2019)
- **fedmatch** Fast, Flexible, and User-Friendly Record Linkage Methods) is a R package which provides a flexible set of tools for matching two unlinked datasets. It allows for three ways to match data: exact matches, fuzzy matches, and multi-variable matches. It also allows an easy combination of these three matches. You can find detials on the package in the technical paper, Lee (2021).
- **Reclin2** is an R package implementing functions to assist in performing probabilistic record linkage and deduplication (i.e., generating pairs, comparing records, em-algorithm for estimating m- and u-probabilities (Fellegi and Sunter, 1969)). It allows also for pre- and post-processing for machine learning methods for record linkage.
- **RecordLinkage** is an R package implementing methods based on a stochastic approach and classification algorithms from the ML domain. For details, see Sariyar and Borg (2016).
- **Record Linkage** is a Python toolkit that provides most of the tools needed for record linkage and deduplication. The package contains indexing methods, functions to compare records and classifiers, and it is developed for research and the linking of small or medium-sized files. For more information on the methods implemented in the toolkit see de Bruin (2017).
- **RELAIS** (REcord Linkage At IStat) is an open source toolkit providing a set of techniques for dealing with record linkage projects elaborated by Istat (Italian national statistical institute).
- **Splink** is a Python package for probabilistic record linkage that allows you to deduplicate and link records from datasets that lack unique identifiers. For more details on the methods offered by the package, see Linacre et al. (2022).

Tools for Statistical Matching

- `MatchIt` is an `R` package performing matching, with a focus on the causal inference setting.
- `smpc` and `smmatch` commands in Stata perform Statistical Matching based on Alpman (2016), that implements the statistical matching procedure proposed by Rubin (1986).
- `StatMatch` is an `R` package for data integration through statistical matching and, as a by-product, the ability to impute missing values in a dataset.

Tools for estimation in nonprobability samples

- `NonProbEst` is an `R` package which implements a set of techniques for estimation in non-probability surveys, using various approaches. Functions in the package allow to obtain calibration weights, propensity scores and matching predictions for a reference sample. Machine learning classification algorithms can be used as alternatives for logistic regression as a technique to estimate propensities. For more details on the implementation, see Rueda et al. (2020)
- `nonprobsvy` is an R package available on GitHub; the package provides the implementation of modern methods for nonprobability samples when auxiliary information from the population or probability sample is available. The methods available are inverse probability weighting estimators with possible calibration constraints Chen et al. (2020), mass imputation estimators based on nearest neighbours (Yang et al., 2021), predictive mean matching and regression imputation (Kim et al., 2021), doubly robust estimators with bias minimization (Yang, 2020). Tutorials can be found at this link.

4 Conclusions

Throughout the last few decades, national statistical agencies and international organisations have relied heavily on sample surveys to produce crucial statistical information. Surveys have played an essential role in providing reliable, accurate, and regularly updated data. However, due to the high costs of probabilistic surveys and the decline in response rates it has become increasingly challenging, if not impossible, to obtain comprehensive and up-to-date information on the phenomena of interest through traditional investigation techniques based on probability samples. Moreover, many of these phenomena urge to be measured in a timely and granular manner, e.g., among others, climate change, pollution, social inclusion, poverty.

Data coming from a variety of non-probabilistic large-scale observations, such as digital administrative records, satellite data, and web data, have the potential to complement and enrich traditional data sources. These alternative data sources, that are becoming increasingly available and affordable, yet differ in their characteristics, levels of detail, and degrees of quality.

From these considerations arises the challenge of developing new statistical methods to appropriately combine data from multiple sources in order to make use of all the available information on a specific phenomenon under study.

In recent years, numerous researchers have made significant contributions in the domain of data integration methods in official statistics and survey methodology. However, the dissemination of these refined techniques is proving difficult among non-statisticians. This report provides an insight into the challenges and opportunities associated with various topics of data integration, such as record linkage, statistical matching, and probability and non-probability sample combination, providing a comprehensive overview for a broader audience interested in this evolving landscape of statistical methods. Thought for a multidisciplinary audience, each topic's presentation is intentionally kept non-technical and enriched with toy/practical examples to illustrate the various methods presented. Additionally, for each method, the primary packages for the most common software - R, Stata, and Python - are indicated.

The flowchart in the Appendix can be a useful tool to guide the reader in the identification of the most appropriate method according to the characteristics of the available data and the purposes for which the integration is being done.

5 Appendix

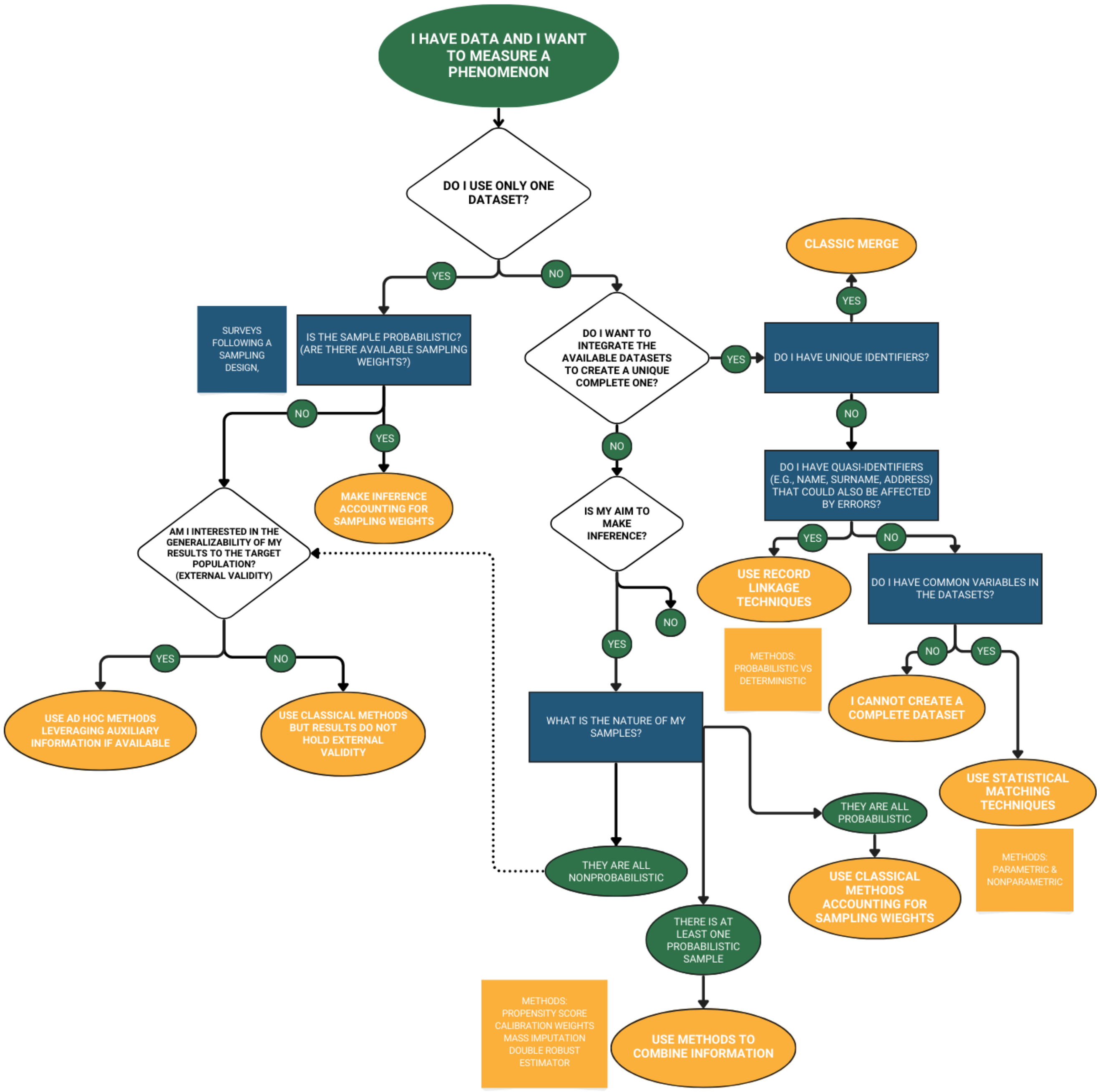

I HAVE DATA AND I WANT TO MEASURE A PHENOMENON
DO I USE ONLY ONE DATASET?
YES
NO
SURVEYS FOLLOWING A SAMPLING DESIGN,
IS THE SAMPLE PROBABILISTIC? (ARE THERE AVAILABLE SAMPLING WEIGHTS?)
DO I WANT TO INTEGRATE THE AVAILABLE DATASETS TO CREATE A UNIQUE COMPLETE ONE?
YES
DO I HAVE UNIQUE IDENTIFIERS?
YES
CLASSIC MERGE
NO
DO I HAVE QUASI-IDENTIFIERS (E.G., NAME, SURNAME, ADDRESS) THAT COULD ALSO BE AFFECTED BY ERRORS?
YES
USE RECORD LINKAGE TECHNIQUES
METHODS: PROBABILISTIC VS DETERMINISTIC
NO
DO I HAVE COMMON VARIABLES IN THE DATASETS?
NO
I CANNOT CREATE A COMPLETE DATASET
YES
USE STATISTICAL MATCHING TECHNIQUES
METHODS: PARAMETRIC & NONPARAMETRIC
NO
YES
MAKE INFERENCE ACCOUNTING FOR SAMPLING WEIGHTS
AM I INTERESTED IN THE GENERALIZABILITY OF MY RESULTS TO THE TARGET POPULATION? (EXTERNAL VALIDITY)
YES
USE AD HOC METHODS LEVERAGING AUXILIARY INFORMATION IF AVAILABLE
NO
USE CLASSICAL METHODS BUT RESULTS DO NOT HOLD EXTERNAL VALIDITY
NO
IS MY AIM TO MAKE INFERENCE?
NO
YES
WHAT IS THE NATURE OF MY SAMPLES?
THEY ARE ALL NONPROBABILISTIC
THEY ARE ALL PROBABILISTIC
USE CLASSICAL METHODS ACCOUNTING FOR SAMPLING WIEGHTS
THERE IS AT LEAST ONE PROBABILISTIC SAMPLE
USE METHODS TO COMBINE INFORMATION
METHODS: PROPENSITY SCORE CALIBRATION WEIGHTS MASS IMPUTATION DOUBLE ROBUST ESTIMATOR

Figure 5: Flowchart: Data Integration methods

References

Abramitzky, R., Boustan, L., and Eriksson, K. (2019). To the new world and back again: Return migrants in the age of mass migration. ILR Review, 72(2):300–322.

Abramitzky, R., Boustan, L. P., and Eriksson, K. (2012). Europe's tired, poor, huddled masses: Selfselection and economic outcomes in the age of mass migration. American Economic Review, 102(5):1832–1856.

Abramitzky, R., Boustan, L. P., and Eriksson, K. (2014). A nation of immigrants: Assimilation and economic outcomes in the age of mass migration. Journal of Political Economy, 122(3):467–506.

Abramitzky, R., Mill, R., and Pérez, S. (2020). Linking individuals across historical sources: A fully automated approach. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 53(2):94–111.

Adena, M., Enikolopov, R., Petrova, M., Santarosa, V., and Zhuravskaya, E. (2015). Radio and the rise of the nazis in prewar Germany. The Quarterly Journal of Economics, 130(4):1885–1939.

Alpman, A. (2016). Implementing Rubin's alternative multiple-imputation method for statistical matching in Stata. The Stata Journal, 16(3):717–739.

Anderson, K., Ryan, B., Sonntag, W., Kavvada, A., and Friedl, L. (2017). Earth observation in service of the 2030 Agenda for Sustainable Development. Geo-spatial Information Science, 20(2):77–96.

Ansolabehere, S. and Hersh, E. D. (2017). Adgn: An algorithm for record linkage using address, date of birth, gender, and name. Statistics and Public Policy, 4(1):1–10.

Asher, J., Resnick, D., Brite, J., Brackbill, R., and Cone, J. (2020). An introduction to probabilistic record linkage with a focus on linkage processing for WTC registries. International Journal of Environmental Research and Public Health, 17(18):6937.

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

Belin, T. R. and Rubin, D. B. (1995). A method for calibrating false-match rates in record linkage. Journal of the American Statistical Association, 90(430):694–707.

Berent, M. K., Krosnick, J. A., and Lupia, A. (2016). Measuring voter registration and turnout in surveys: Do official government records yield more accurate assessments? Public Opinion Quarterly, 80(3):597–621.

Bethlehem, J. (2016). Solving the nonresponse problem with sample matching? Social Science Computer Review, 34(1):59–77.

Binette, O. and Steorts, R. C. (2022). (Almost) all of entity resolution. Science Advances,

8(12):eabi8021.

Bolsen, T., Ferraro, P. J., and Miranda, J. J. (2014). Are voters more likely to contribute to other public goods? Evidence from a large-scale randomized policy experiment. American Journal of Political Science, 58(1):17–30.

Bosco, C., Grubanov-Boskovic, S., Iacus, S., Minora, U., Sermi, F., and Spyratos, S. (2022). Data innovation in demography, migration and human mobility. Technical Report EUR 30907 EN, Publications Office of the European Union, Luxembourg.

Breidt, F. J., McVey, A., and Fuller, W. A. (1996). Two-phase estimation by imputation. Journal of the Indian Society of Agricultural Statistics, 49:79–90.

Brick, J. M. (2015). Compositional model inference. JSM Proceedings, Survey Research Methods Section, pages 299–307.

Buchanan, A. L., Hudgens, M. G., Cole, S. R., Mollan, K. R., Sax, P. E., Daar, E. S., Adimora, A. A., Eron, J. J., and Mugavero, M. J. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. Journal of the Royal Statistical Society, Series A, 181(4):1193–1209.

Cesarini, D., Lindqvist, E., Östling, R., and Wallace, B. (2016). Wealth, health, and child development: Evidence from administrative data on Swedish lottery players. The Quarterly Journal of Economics, 131(2):687–738.

Chen, J. K. T., Valliant, R., and Elliott, M. R. (2018). Model-assisted calibration of non-probability sample survey data using adaptive LASSO. Survey Methodology, 44:117–144.

Chen, Y., Li, P., Rao, J., and Wu, C. (2022). Pseudo empirical likelihood inference for nonprobability survey samples. The Canadian Journal of Statistics, 50(4):1166–1185.

Chen, Y., Li, P., and Wu, C. (2020). Doubly robust inference with nonprobability survey samples. Journal of the American Statistical Association, 115(532):2011–2021.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters, 21(1):C1–C68.

Chipperfield, J., Chessman, J., and Lim, R. (2012). Combining household surveys using mass imputation to estimate population totals. Australian New Zealand Journal of Statistics, 54:223–238.

Christen, P. (2019). Data linkage: The big picture. Harvard Data Science Review, 1(2).

Christen, P. and Christen, P. (2012). The data matching process. Springer.

Christophides, V., Efthymiou, V., Palpanas, T., Papadakis, G., and Stefanidis, K. (2020). An overview of end-to-end entity resolution for big data. ACM Computing Surveys (CSUR), 53(6):1–42.

Chu, K. C. K. and Beaumont, J.-F. (2019). The use of classification trees to reduce selection bias for a nonprobability sample with help from a probability sample. In Proceedings of the Survey Methods Section of SSC.

de Bruin, J. (2017). Record Linkage. Python library. version 0.8.1.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association, 87(418):376–382.

Donaldson, D. and Storeygard, A. (2016). The view from above: Applications of satellite data in economics. Journal of Economic Perspectives, 30(4):171–198.

D'Orazio, M. (2022). StatMatch: Statistical Matching or Data Fusion. R package version 1.4.1.

D'Orazio, M., Di Zio, M., and Scanu, M. (2006). Statistical matching: Theory and practice. John Wiley & Sons.

D'Orazio, M. (2011). Statistical matching and imputation of survey data with the package StatMatch for the R environment. R package vignette. http://www.cros-portal.eu/sites/default/files/ /Statistical_Matching_with_StatMatch.pdf.

Elliott, M. R., . V. R. (2017). Inference for nonprobability samples. Statistical Science, 32:249–264.

Enamorado, T., Fifield, B., and Imai, K. (2019). Using a probabilistic model to assist merging of largescale administrative records. American Political Science Review, 113(2):353–371.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. Journal of Econometrics, 189:1–23.

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. Journal of the American Statistical Association, 64(328):1183–1210.

Figlio, D., Guryan, J., Karbownik, K., and Roth, J. (2014). The effects of poor neonatal health on children's cognitive development. American Economic Review, 104(12):3921–3955.

Fortini, M., Liseo, B., Nuccitelli, A., and Scanu, M. (2001). On Bayesian record linkage. Research in official statistics, 4(1):185–198.

Fuller, W. A. and J., B. (1999). Estimation for supplemented panels. Sankhya:¯ The Indian Journal of Statistics, Series B, 61(1):58–70.

Gao, X. and Carroll, R. J. (2017). Data integration with high dimensionality. Biometrika, (104):251–272.

Giraud-Carrier, C., Goodliffe, J., Jones, B. M., and Cueva, S. (2015). Effective record linkage for mining campaign contribution data. Knowledge and Information Systems, 45:389–416.

Goldstein, H. and Harron, K. (2015). Record Linkage: A Missing Data Problem, chapter 6, pages 109– 124. John Wiley & Sons, Ltd.

Gregg, F. and Eder, D. (2022). Dedupe. [retrieved 31 October 2023].

Gutman, R., Afendulis, C. C., and Zaslavsky, A. M. (2013). A Bayesian procedure for file linking to analyze end-of-life medical costs. Journal of the American Statistical Association, 108(501):34–47.

Hidiroglou, M. (2001). Double sampling. Survey Methodology, (27):143–54.

Hill, S. J. (2017). Changing votes or changing voters? How candidates and election context swing voters and mobilize the base. Electoral Studies, 48:131–148.

ISTAT, CBS, G. I. S. S. E. (2013). Report on WP1: State of the art on statistical methodologies for data integration. Technical report, ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data.

Jantti, M., Törmälehto, V.-M., and Marlier, E. (2013). The use of registers in the context of EU? SILC: challenges and opportunities. Publications Office of the European Union.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science, (22):523–539.

Kasai, J., Qian, K., Gurajada, S., Li, Y., and Popa, L. (2019). Low-resource deep entity resolution with transfer and active learning. arXiv preprint arXiv:1906.08042.

Kim, J., Berg, E., and Park, T. (2016). Statistical matching using fractional imputation. Survey Methodology, 40:19–40.

Kim, J. K., Park, S., Chen, Y., and Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. Journal of the Royal Statistical Society Series A: Statistics in Society, 184(3):941–963.

Kim, J. K. and Rao, J. N. (2012). Combining data from two independent surveys: A model-assisted approach. Biometrika, 99(1):85–100.

Kim, J.-K. and Tam, S.-M. (2021). Data integration by combining big data and survey sample data for finite population inference. International Statistical Review, 89(2):382–401.

Kooli, N., Allesiardo, R., and Pigneul, E. (2018). Deep learning based approach for entity resolution in databases. In Asian conference on intelligent information and database systems, pages 3–12. Springer.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. Survey Methodology, 32(2):133.

Lahiri, P. and Larsen, M. D. (2005). Regression analysis with linked data. Journal of the American Statistical Association, 100(469):222–230.

Larsen, M. D. and Rubin, D. B. (2001). Iterative automated record linkage using mixture models. Journal of the American Statistical Association, 96(453):32–41.

Lee, M. F. . C. W. . B. M. . J. D. . S. (2021). fedmatch: Fast, flexible, and user-friendly record linkage methods. R package version 2.0.3.

Lee, S. and Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. Sociological Methods and Research, 37:319–343.

Linacre, R., Lindsay, S., Manassis, T., Slade, Z., Hepworth, T., Kennedy, R., and Bond, A. (2022). Splink: Free software for probabilistic record linkage at scale. International Journal of Population Data Science, 7(3).

Little, R. J. and Rubin, D. B. (2002). Statistical Analysis With Missing Data. Wiley, New York, 2nd edition.

Marchant, N. G., Kaplan, A., Elazar, D. N., Rubinstein, B. I., and Steorts, R. C. (2021). d-blink: Distributed end-to-end Bayesian entity resolution. Journal of Computational and Graphical Statistics, 30(2):406–421.

McRoberts, R. E., Tomppo, E., and Naesset, E. (2010). Advances and emerging issues in national forest inventories. Scandinavian Journal of Educational Research, 8(4):364–381.

McVeigh, B. S., Spahn, B. T., and Murray, J. S. (2019). Scaling Bayesian probabilistic record linkage with post-hoc blocking: An application to the California great registers. arXiv preprint arXiv:1905.05337.

Meredith, M. and Morse, M. (2014). Do voting rights notification laws increase ex-felon turnout? The ANNALS of the American Academy of Political and Social Science, 651(1):220–249.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. Journal of the American Statistical Association, 99(468):1131–1139.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. Journal of the Royal Statistical Society Series B: Statistical Methodology, 72(1):27–48.

Monge, A. E. (1997). An efficient domain-independent algorithm for detecting approximately duplicate database records. In Proc. of the ACM-SIGMOD Workshop on Research Issues on Knowledge Discovery and Data Mining (1997).

National Academies of Sciences, Engineering, and Medicine (2023). Toward a 21st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources. The National Academies Press.

Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959). Automatic linkage of vital records. Science, 130:954–959.

Okner, B. (1972). Constructing a new data base from existing microdata sets: The 1966 merge file. In Annals of Economic and Social Measurement, Volume 1, Number 3, pages 325–362. NBER.

O'Muircheartaigh, C. and Hedges, L. V. (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach. Journal of the Royal Statistical Society Series C: Applied Statistics, 63(2):195–210.

Ong, T. C., Mannino, M. V., Schilling, L. M., and Kahn, M. G. (2014). Improving record linkage performance in the presence of missing linkage data. Journal of Biomedical Informatics, 52:43–54.

Paganini, M., Petiteville, I., Ward, S., Dyke, G., Steventon, M., Harry, J., and Kerblat, F. (2018). Satellite earth observations in support of the Sustainable Development Goals. The CEOS Earth Observation Handbook.

Palmer, J. R., Espenshade, T. J., Bartumeus, F., Chung, C. Y., Ozgencil, N. E., and Li, K. (2013). New approaches to human mobility: Using mobile phones for demographic research. Demography, 50:1105–1128.

Park, S., Kim, J. K., and Stukel, D. (2017). A measurement error model for survey data integration: Combining information from two surveys. Metron, 75:345–357.

Park, S. and Kim, J. K.and Park, S. (2016). An imputation approach for handling mixed mode surveys. Annals of Applied Statistics, 10:1063–1085.

Rässler, S. (2012). Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches, volume 168. Springer Science & Business Media.

Reiter, J. P. (2021). Assessing uncertainty when using linked administrative records. Administrative Records for Survey Methodology, pages 139–153.

Renssen, R. H. and Nieuwenbroek, N. (1997). Aligning estimates for common variables in two or more sample surveys. The Journal of the American Statistical Association, 92:368–75.

Rivers, D. (2007). Sampling for web surveys, asa proceedings of the section on survey research methods. In Joint Statistical Meetings (Vol. 4) Alexandria. American Statistical Association.

Rodgers, W. L. (1984).An evaluation of statistical matching. Journal of Business & Economic

Statistics, 2(1):91–102.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3):581–592.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. Journal of Business & Economic Statistics, 4(1):87–94.

Rueda, M., Ferri-García, R., and Castro, L. (2020). The R package NonProbEst for estimation in nonprobability surveys. The R Journal, 12(1):406–418.

Ruggles, N. and Ruggles, R. (1974). A strategy for merging and matching microdata sets. In Annals of Economic and Social Measurement, Volume 3, number 2, pages 353–371. NBER.

Sadinle, M. (2014). Detecting duplicates in a homicide registry using a Bayesian partitioning approach. The Annals of Applied Statistics, 8(4):2404–2434.

Sadinle, M. (2017). Bayesian estimation of bipartite matchings for record linkage. Journal of the American Statistical Association, 112(518):600–612.

Sariyar, M. and Borg, A. (2016). Record linkage in R. R package. Version 0.4-10.

Sariyar, M., Borg, A., and Pommerening, K. (2012). Missing values in deduplication of electronic patient data. Journal of the American Medical Informatics Association, 19:e76–82.

Scanu, M. (2003). Metodi Statistici per il Record Linkage. Number 16 in Collana "Metodi e Norme". ISTAT, Roma.

Scheuren, F. and Winkler, W. E. (1993). Regression analysis of data files that are computer matched. Survey Methodology, 19:39–58.

Scheuren, F. and Winkler, W. E. (1997). Regression analysis of data files that are computer matched – Part II. Survey Methodology, 23:157–165.

Steorts, R. C. (2015). Entity resolution with empirically motivated priors. Bayesian Analysis, 10(4):849– 875.

Stuart, E. A., Bradshaw, C. P., and Leaf, P. J. (2015). Asessing the generalizability of randomized trial results to target populations. Prevention Science, 16:475–485.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., and Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. Journal of the Royal Statistical Society, Series A, 174:369–386.

Tancredi, A. and Liseo, B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. Annals of Applied Statistics, 5:1553–1585.

Tancredi, A., Steorts, R., and Liseo, B. (2020). A unified framework for de-duplication and population size estimation (with discussion). Bayesian Analysis, 15(2):633–682.

Tourangeau, R., Conrad, F. G., and Couper, M. P. (2013). The science of web surveys. New York: Oxford University Press.

UNOOSA (2018). European Global Navigation Satellite System and Copernicus: Supporting the Sustainable Development Goals - BUILDING BLOCKS TOWARDS THE 2030 AGENDA. Technical report, OFFICE FOR OUTER SPACE AFFAIRS - UNITED NATIONS OFFICE AT VIENNA.

Valliant, R. and Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. Sociological Methods Research, 40(1):105–137.

van den Brakel, J., Buelens, B., Curier, R., Daas, P., Gootzen, Y., de Jong, T., Puts, M., and Tennekes, M. (2019). Aspects of existing databases, traditional and non-traditional data sources and collection of good practices (The MAKSWELL Project). https://www.makswell.eu/attached_documents/ output_deliverables/deliverable_2.1.pdf.

Van der Doef, S., Daas, P., and Windmeijer, D. (2018). Identifying innovative companies from their website. In Abstract for BigSurv18 conference (ingediend).

Vavreck, L. and Rivers, D. (2008). The 2006 cooperative congressional election study. Journal of Elections, Public Opinion and Parties, 18(4):355–366.

Ventura, S. L. and Nugent, R. (2014). Hierarchical linkage clustering with distributions of distances for large-scale record linkage. In Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2014, Ibiza, Spain, September 17-19, 2014. Proceedings, pages 283– 298. Springer.

Winkler, W. E. (2002). Methods for record linkage and Bayesian networks. Technical report, Technical report, Statistical Research Division, US Census Bureau.

Winkler, W. E. et al. (2000). Machine learning, information retrieval and record linkage. In Proc Section on Survey Research Methods, American Statistical Association, pages 20–29.

Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method. The Canadian Journal of Statistics, 32(4):15–26.

Wu, C. (2022). Statistical inference with non-probability survey samples. Survey Methodology, 48(2):283–311.

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

Yang, S., K. J. K. . S. R. (2020). Doubly robust inference when combining probability and nonprobability samples with high-dimensional data. Journal of the Royal Statistical Society, Series B, 82:445–465.

Yang, S. and Ding, P. (2020). Combining multiple observational data sources to estimate causal effects. Journal of the American Statistical Association, 115(531):1540–1554.

Yang, S. and Kim, J. (2020). Statistical data integration in survey sampling: A review. Japanese Journal of Statistics and Data Science, 3:625–650.

Yang, S. and Kim, J. K. (2018). Integration of survey data and big observational data for finite population inference using mass imputation. arXiv: Methodology.

Yang, S., Kim, J. K., and Hwang, Y. (2021). Integration of data from probability surveys and big found data for finite population inference using mass imputation. Survey Methodology, 47(1):29–58.

Ybarra, L. M. R. and Lohr, S. L. (2008). Small area estimation when auxiliary information is measured with error. Biometrika, 95:919–931.

Yuan, M., Li, P., and Wu, C. (2022). Nonparametric estimation of propensity scores for non-survey samples. Working paper.