BANCA D'ITALIA

EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

It's a match! Linking foreign counterparts in Italian customs data to their balance sheets

by Marta Crispino and Francesco Paolo Conteduca

December 2023

Number 823

BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

It's a match! Linking foreign counterparts in Italian customs data to their balance sheets

by Marta Crispino and Francesco Paolo Conteduca

Number 823 – December 2023

The series Occasional Papers *presents studies and documents on issues pertaining to the institutional tasks of the Bank of Italy and the Eurosystem. The* Occasional Papers *appear alongside the* Working Papers *series which are specifically aimed at providing original contributions to economic research.*

The Occasional Papers *include studies conducted within the Bank of Italy, sometimes in cooperation with the Eurosystem or other institutions. The views expressed in the studies are those of the authors and do not involve the responsibility of the institutions to which they belong.*

The series is available online at www.bancaditalia.it .

ISSN 1972-6643 (online)

Designed by the Printing and Publishing Division of the Bank of Italy

IT'S A MATCH! LINKING FOREIGN COUNTERPARTS IN ITALIAN CUSTOMS DATA TO THEIR BALANCE SHEETS

by Marta Crispino* and Francesco Paolo Conteduca*

Abstract

This paper describes the methodology underlying the matching between non-EU counterparts in the Italian Customs and Monopolies Agency data and firms in the Bureau van Dijk Orbis database. Through different validation exercises, we show that the matches stemming from our proposed procedure are largely correct regarding both records and transaction values. The resulting corresponding tables can serve as a useful tool to shed light on the features of the counterparts of Italian firms active in international trade.

JEL Classification: C81, F14, D22, C55, C88.
Keywords: record linkage, big data integration, customs data, balance sheet data, name harmonization, blocking, entity matching.
DOI: 10.32057/0.QEF.2023.0823

Contents

* Bank of Italy, Economics, Statistics and Research Department, Via Nazionale, 187 - 00184 Rome, Italy

1. Introduction[1]

Using firm-level datasets has become increasingly important in economics research. With the rise in the ability to process large data and the multiplication of available firm-level sources, combining different datasets allows the exploration of the activity of a firm under several complementary perspectives. However, the need for common identifiers across various sources may limit their integration and full exploitation. This limitation characterizes also the Italian customs data, specifically when it comes to transactions between Italian firms and entities located outside the European Union. In fact, these particular data lack any consistent identifier associated to the foreign counterparts of Italian firms.

This paper proposes a methodology to match the non-European Union members (extra-EU) counterparts of Italian firms' transactions reported in the Italian customs (*Agenzia delle Dogane e dei Monopoli*; henceforth, ADM) with the Bureau van Dijk (BvD) Orbis global database (henceforth, Orbis), collecting balance sheet and ownership data of companies worldwide. The primary goal of this study is to establish a link between records in these two datasets, specifically matching an extra-EU foreign counterpart in the ADM administrative data, as provided by compilers in custom declarations, with a corresponding firm in Orbis.[2]

The statistical task of matching, or record linkage, involves identifying whether pairs of records in one or more datasets refer to the same entity, such as a person or a company. It is a fundamental task in data integration, and it underlies various applications such as census data analysis, customer relationship management, and medical research. Record linkage aims at identifying all pairs of records that correspond to the same entity while minimizing the number of false matches. Usually, a record linkage procedure consists of two main parts: (*i*) the candidate selection step, when the choice of entities worth comparing takes place; (*ii*) the candidate matching step, when the comparison allows to determine whether the particular entities represent the same real-world object. Step (*ii*) involves pairwise time-consuming comparisons among all suitable entities shortlisted in step (*i*), which we compare through string similarity measures.

Because of the absence of a common identifier, we exploit other information in the databases, that is, the name and, if available, the address of the foreign counterpart to determine which records correspond to the same entity. This task is challenging since many entities with similar names and addresses may or may not correspond to the same firm. Moreover, the complexity of the exercise grows because of the large size of the two datasets involved, which typically consist of millions of records. This implies a significant number of records that need to be matched (ADM) and an extensive search space of entities (Orbis) to consider. For

[1] We are grateful to Riccardo Maria Nusca for his invaluable contributions to the advancement of the algorithmic procedure, encompassing the development of the harmonization pipeline and the preliminary studies. We also thank Elena San Martini for her assistance in creating a sample of Italian firms to calibrate the algorithm. Additionally, we would like to thank Michele Loberto, Alessandro Borin, Michele Mancini, Ludovic Panon, Andrea Linarello, Alberto Felettigh, and all the internal seminar participants for their helpful insights and comments. Finally, the Business Intelligence and Advanced Analytics division of the informatics department (SVI) also deserves credit for supporting us with the computing platform. The opinions expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy or the Eurosystem. All errors are our own.
Correspondence: marta.crispino@bancaditalia.it
[2] For transactions with intra-EU companies, foreign entities are identified by the VAT tax number.

example, if one dataset has 100,000 records and the other has 200,000, there could be up to 20 billion potential pairs to compare, which is computationally demanding or even infeasible.

As a consequence, reducing the matching problem's dimensionality and using efficient and effective techniques to achieve accurate results is crucial when dealing with large databases. Typically, workarounds include name harmonization and blocking, which both aim at decreasing the number of pairs to evaluate.[3] In our application, we resort to name harmonization routines developed in the context of patent data by the NBER patent data project.[4]

The goal of blocking is to narrow the search space by partitioning the datasets into smaller and more manageable chunks based on specific criteria (such as, the value of a third variable). This permits to compare only records within the same block and increases the chance of finding true matches while minimizing the computational burden of the procedure. In our application, we mostly use postal codes to define the blocks. This choice is sensible as postal codes serve as common location identifiers in many countries and are available in ADM and Orbis.

The proposed procedure consists in sequentially applying different matching techniques. At each step, a proportion of the firms appearing in the ADM data is matched with some company available in Orbis, relying on step-specific identifiers (the company's name or some function of it along with the postal code, if available). The initial steps focus on deterministic matches, in that they rely solely on finding exact correspondences between the two databases. In contrast, the subsequent steps employ probabilistic matching techniques, in that they apply record linkage algorithms that incorporate string similarities and specific blocking keys. Notably, we apply machine learning techniques to select the optimal acceptance rule for the candidate pairs. The outcome of the sequential matching procedure is a correspondence table that links records in ADM to at least one entity in Orbis.

Our record linkage procedure demonstrates good performance, as it can match a significant portion of the records. We consistently achieve an average matching rate of over 85 percent of the records when considering Italy's key trading partners. Notably, the contribution of deterministic and probabilistic matches is heterogeneous across partner countries, depending on country-specific characteristics of the underlying data sources.

Generally speaking, the validation of the algorithmic procedure poses significant challenges as we would need data labeled with the ground truth for testing purposes. Additionally, the sheer volume of data makes manual validation or expert evaluation impractical. Nonetheless, we test the performance of our matching algorithm using two approaches. First, we exploit the Brexit as a quasi-natural experiment. Before Brexit, the UK belonged to the EU Customs Union, and transactions with the UK had to comply with the intra-EU rules,

[3] For example, a company could be listed in one database as "ABC Corp." and in the other as "ABC Corporation", or a postal code could be listed with different hyphenation or formatting. Oftentimes the company's name may be spelled not consistently even within the same data source.
[4] Information on the project is available at https://sites.google.com/site/patentdataproject/Home.

which included the requirement to provide the VAT number of the British counterpart. Following the exit of the UK from the EU Customs Union, the requirement of filling in the VAT number no longer applies. Nonetheless, we observe a small sample of UK counterparts with both names and VAT identification numbers, a unique situation due to Brexit-induced changes in trade document compilation rules. Therefore, by comparing the VAT retrieved from the ADM data with those obtained from Orbis for the corresponding matched companies, we are able to determine whether our record linkage procedure correctly identifies links. In the second validation exercise, we focus on extra-EU companies exporting goods to Italy. In particular, we leverage the nexus between the commodities imported to Italy and the sector of activity of the foreign counterpart that produced those commodities, obtained from the match with Orbis. By examining the association between the exporter's sector and the code of the exported product, we assess the consistency and alignment of our matches, thereby validating indirectly the accuracy of our record linkage procedure.[5] In this second validation exercise, the accuracy (in terms of values matched) exceeds 95 percent for some partner countries (e.g., the USA, China, Switzerland, Mexico, and Japan), while it is lower than 90 percent for others (e.g., Egypt and Russia). A possible reason for the different performance may be the uneven country coverage of Orbis. Nonetheless, both these validation exercises indicate that the overall accuracy of the matching is satisfactory and that it crucially depends on the country of interest.

As mentioned, matching the two data sources is crucial to exploit the full potential of the rich information set represented by transaction data. By identifying the foreign counterpart of a transaction, one can delve into the microeconomic foundations of international flows of goods and international production. Several recent contributions in the international trade literature have tried to exploit the firm-to-firm dimension (e.g., Bernard et al., 2019; Huneeus, 2020; Carvalho et al., 2021; Dhyne et al., 2021; Adao et al., 2022; Alfaro-Ureña et al., 2022; Dhyne et al., 2022; Eaton et al. 2022; Alviarez et al., 2023; Amiti et al., 2023; Pustilnik, 2023). The available data usually provide information on domestic transactions or refer to developing countries. In contrast, our matching procedure yields unique domestic-firm to foreign-firm transaction data from a large developed economy. To our knowledge, we are the first to match the foreign counterparts of a dataset collecting customs transactions to a global dataset containing financial and ownership information.

The statistical literature on record linkage is vast and continuously evolving, with novel procedures and tools to merge administrative databases frequently proposed.[6] Typically, the problem of matching textual data has been addressed in the economic literature with patent data (see the NBER patent data project, for example). These works often focus on specific tasks, are tailored to the requirements of a particular country or involve datasets smaller than the ones considered in this paper. In this regard, because of the aforementioned scalability problems, they are difficult to apply to our context, which is characterized by the large size of databases involved in the matching. Our methodology instead offers a versatile solution for handling large datasets across

[5] For example, a foreign firm listed in the manufacturing of footwear is more likely to export shoes.

[6] See for instance the Python package `name_matching` recently released by the Central Bank of the Netherlands (https://github.com/DeNederlandscheBank/name_matching), or the R package `fedmatch` released by the Fed (Cohen et al. 2021; https://cran.r-project.org/web/packages/fedmatch/fedmatch.pdf).

diverse countries simultaneously. In fact, while the record linkage literature presents specialized techniques, our distinctive approach accommodates extensive datasets, enabling integration and synchronization of data across international boundaries for comprehensive analyses.

The rest of the paper goes as follows. Section 2 describes the two datasets, namely the custom declaration data and Orbis. Section 3, after a brief introduction to some concepts of record linkage, is devoted to explaining the methodology adopted for data matching. In Section 4, we report some statistics describing the matching results, focusing on the validation of the procedure. Section 5 describes some open issues and concludes.

2. Data description

The database contains microdata on the foreign trade of goods by Italian firms, collected by the Customs and Monopolies Agency (ADM) and shared with the Bank of Italy based on a specific agreement for research purposes. The data collect information on transactions involving both extra-EU and intra-EU counterparts. Notably, the customs data for extra-EU trade and intra-EU trade are separate sources, and have different structures. For the former, the unit of observation is the single transaction, whereas for the latter Italian firms are required to fill in a monthly or quarterly report depending on established thresholds.

This paper focuses exclusively on the transactions with extra-EU countries between 2010-2021.[7] In the extra-EU customs data, a universal identifier of the foreign counterpart is absent, yet other potentially useful information is available. In particular, Italian firms have to fill in the Single Administrative Document (SAD) of the customs declarations. This document contains several details on the traded good (e.g., the 8-digit code according to the Combined Nomenclature, the value, the weight, the origin, the provenance, the destination) and the nature of the transaction (e.g., exports, imports, transit). Moreover, it collects information about the domestic firm and the foreign counterpart involved in the transaction. The main information of interest to us is the name and postcode of the foreign counterpart, which, when available, are typically located in the fields labeled "Consignor/Exporter" (for imports) or "Consignee" (for export). These fields provide relevant details regarding the foreign entities involved in the transactions.

However, identifying companies by their name, as reported in administrative forms, is not always straightforward, and it is also prone to errors because different entities may share the same denomination, reported denominations are not standard, and reporting errors may occur. In practice, several instances of the same entity may appear as different firms (due to typos or to misspelling, for instance). Hence, the country and postcode of the foreign counterpart represent useful information to sharpen our definition. This aspect is particularly relevant because preventing the comparison between irrelevant entities improves the matching's performance and accuracy. We detail these aspects in Sections 3 and 4.

[7] According to ISTAT data, Italian imports from extra-EU countries (respectively, exports to extra-EU countries) in 2021 accounted for 43% (47%) of the total value of imports (exports). Table A1 in the Appendix details the most important extra-EU trading partners.

Table 1 reports the number of records and their value in the raw ADM data by year and direction of the flow. Each year, records are in the order of million observations, with exports being generally higher than imports regarding both the number of records and values. It is worth emphasizing that the figures reported in Table 1 refer to the ADM raw data.[8] Before the matching with Orbis, we perform preparatory cleaning, which allows us to reduce substantially the number of records with a small effect on the value of the flows. Such cleaning procedures, detailed in Subsection 3.2.1, do not aim to increase the representativeness of the data vis-à-vis the official Italian trade statistics (for example, by removing outliers). Rather, we tailor our procedures to preserve as many observations as possible and to have the most comprehensive matching possible. Table 1 also reports the number of records and their value when the denomination of the foreign counterpart is present in the SAD. We observe that the name of the foreign counterpart is available most of the time in the export data, while it is frequently missing in the case of imports. However, since 2015, there has been a noticeable increase in coverage, and by 2021 it has reached full coverage, meaning that the name of the foreign counterpart appears to be now consistently provided for both imports and exports.

Table 1 – Extra-EU transactions and their value in the ADM data by year and direction of the flow

Year	Exports				Imports			
	Records (in mn.)		Value (in € bn.)		Records (in mn.)		Value (in € bn.)	
	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
2010	11.20	10.99	160.98	158.44	5.55	0.58	177.68	33.12
2011	12.39	12.33	179.96	179.75	6.45	0.68	198.77	41.29
2012	13.43	13.38	194.00	193.86	6.42	0.67	190.06	39.42
2013	14.98	14.92	192.69	192.41	6.85	0.82	175.37	45.43
2014	15.74	15.67	194.93	194.66	7.47	0.90	165.69	44.04
2015	17.10	17.05	201.51	201.34	7.59	2.16	168.79	51.02
2016	18.32	18.27	197.89	197.74	7.87	3.62	155.99	54.41
2017	20.51	20.51	216.85	216.85	6.93	3.26	144.88	50.57
2018	22.95	22.95	217.90	217.90	9.23	5.48	184.52	87.93
2019	24.44	24.44	226.54	226.54	9.42	5.76	184.41	95.91
2020	23.30	23.22	197.33	197.23	14.80	13.66	151.60	126.01
2021	35.97	35.22	260.25	260.06	19.01	19.01	211.83	211.83

Note: (a) indicates raw data, (b) indicates raw data with non-empty name.

BvD Orbis The Bureau van Dijk (BvD) Orbis global database (henceforth, Orbis) is the largest dataset containing historical information on the balance sheets and ownership structure of over 400 million companies worldwide from over 200 countries as of May 2023. Orbis is widely relied upon in the economics/empirical literature. The information contained in Orbis is sourced from over 100 local providers. The BvD harmonizes the information sourced from the data providers to limit the discrepancy across different financial entries at the country level. For the matching procedure, we identify a company through the BvD ID number, a consistent internal alphanumeric identifier. Orbis ranges in coverage (from detailed financial statements to limited or no recent information) depending on local accounting rules (on this point, see also Kalemli-Özcan et al., 2023; Bajgar et al., 2020).

[8] As the ADM data is raw, we may face several problems related to data quality, such as incorrect spelling of names, non-standardized postal codes, missing data, etc. We discuss how we address them in Section 3.

3. The methodology

In this section, we first briefly introduce some concepts of record linkage (Subsection 3.1). Then, we outline our proposed methodology (Subsection 3.2).

3.1 Probabilistic record linkage: a brief introduction and calibration

Record linkage (Fellegi and Sunter, 1969) consists of identifying and linking records that refer to the same real-world entity, such as a person or a company, across different data sources. Record linkage algorithms typically compare the values of certain fields or columns in each dataset, such as names, addresses, or other identifying information, to determine which records are likely to refer to the same entity. Typically, when the linkage is based on names, this comparison is made using a string distance, which assigns a score to each pair of names representing how similar they are. The choice of the distance function is crucial because each one is designed to measure similarity differently, allowing for the prioritization of specific types of dissimilarities based on the unique characteristics of the data. To illustrate this issue, we build a synthetic example with six pairs of fictitious company names and present it in Table 2.

Table 2 – A synthetic example on the comparison of string metrics

			Distance function			
Firm name	**Candidate match**	**GT**	**Levenshtein**	**Cosine**	**Jaro-Winkler**	**Jaro**
ABC CORPORATION	ABC CORP.		0.53	0.81	0.88	0.81
ABC CORPORATION	XYZ CORP.		0.33	0.58	0.56	0.56
ABC CORPORATION	JOY CORPORATION		0.80	0.89	0.73	0.73
ABC CORPORATION	ABC CORPèORATION		0.94	0.98	0.99	0.98
ABC CORPORATION	ACB CORPORATION		0.87	1.00	0.98	0.98
ABC CORPORATION	ABCCORP		0.47	0.83	0.88	0.82

Note: 'GT' indicates the ground truth, i.e., whether the entities in 'Firm name' and in 'Candidate match' are *a priori* the same (green cell) or not (red cell). 'Levenshtein', 'Cosine', 'Jaro-Winkler', and 'Jaro' report the similarity values between 'Firm name' and 'Candidate match' according to the corresponding distance. The cells are colored with a palette going from red (0) to green (1). The more the colors in each column reflect 'GT', the better the distance performs in the classification task.

We assume that pairs in rows 1, 4, 5, and 6 refer to the same entity (green in column GT), while pairs 2 and 3 do not (red in column GT). We then report similarity values normalized to range between 0 and 1. These values represent the similarity between pairs of strings as measured by different metrics.[9] Most distances assign high values (> 0.85) to pairs 4 and 5, where the dissimilarity is primarily due to a typo or minor variation. However, only a few distances, namely the cosine[10], Jaro, and Jaro-Winkler[11], can assign high values to pairs 1 and 6,

[9] Values close to 1 (0) indicate high (low) similarity.

[10] The cosine distance between two strings is computed by taking the dot product of the bag-of-characters vector representations of the strings and dividing it by the product of their magnitudes or similarity.

[11] The Jaro distance is the number of matching characters divided by the total number of characters, including the number of transpositions required to match the characters in the two strings. The Jaro-Winkler distance is a modified version of Jaro distance that takes into account common prefixes between the two strings being compared. This metric gives higher weights to the prefix than to the rest of the string, thus favoring matches that start with the same characters. In addition, Jaro-Winkler distance includes a scaling factor based on the length of the common prefix, which can help to decrease the

which exhibit dissimilarity due to an abbreviation. Additionally, it is worth noting that certain distances, such as the Levenshtein[12] and cosine, assign high values to pair 3, which does not refer to the same entity. This example indicates that not all distances effectively distinguish between different entities in this particular case.

After choosing an appropriate string metric for the data being analyzed and calculating its value for each pair of records to compare, the next step is to evaluate when to accept that a given pair is a match. This decision involves selecting a cut-off threshold for match/non-match, that is, the value of the string distance above which the pair is accepted as a match. The choice of the threshold is linked to the algorithm's strictness or leniency. A lower threshold is more likely to identify true matches but increases the likelihood of false positives. Conversely, a higher threshold will reduce false positives but may miss valid matches. By looking again at Table 2, we see that a threshold larger than 0.73 and smaller than 0.88 would assign all six pairs correctly in the case of Jaro-Winkler, while a smaller one would result in false positives (incorrectly identifying at least pair in line 3 as a match), and a larger one would result in false negatives (failing to identify at least the pairs 1 and 6 as matches). It is common practice in the machine learning literature to determine the optimal threshold by relying on a sample where the ground truth is known, i.e., where the pairs of records are ex-ante known to be true matches or non-matches.[13] By varying the threshold and calculating the proportion of algorithmic matching errors,[14] one can control the trade-off between the risk of false positives and false negatives and calibrate the optimal threshold. Unfortunately, as already discussed, we do not have such a labeled sample of the data. However, we employ an alternative approach to calibrate the threshold. Specifically, we tested the algorithms on a sample of Italian companies for which we were able to retrieve the ground truth leveraging other databases.[15] Based on our tests the Jaro-Winkler distance proved to be the best performer and was therefore chosen for the task. Regarding the choice of the threshold value, we computed two performance measures – accuracy and precision.[16] Both measures were consistently high for threshold values below 0.85. However, as the threshold exceeded 0.85, the accuracy sharply declined. The threshold that maximized the combined accuracy and precision was determined to be 0.85, resulting in our final choice for the matching process (see Figure A3 in the Appendix). Equipped with the pair distance-threshold, we employed record linkage algorithms for our matching procedure.

An important consideration pertains to the external validity of this calibration study. Notably, this study was conducted specifically on Italian firms, which possess certain unique language characteristics that may differ

distance between two strings that are very similar except for a small difference in the beginning. This can be particularly useful when dealing with typos or misspellings in names or addresses.

[12] The Levenshtein distance, also known as the edit distance quantifies the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one string into another.

[13] See Tahamont et al. (2023) for a discussion on the importance of ground truth in linking administrative datasets.

[14] There are two types of errors the algorithm can make: false positives and false negatives. The choice of the threshold is related to the type of error the algorithm will make, and choosing the threshold involves a trade-off between the two types of errors.

[15] We used three different administrative sources available to the Bank of Italy.

[16] Accuracy is the proportion of correct matches (true positives and true negatives) out of the total number of pairs. Precision is the proportion of true positive matches out of all matches made, including false positives. Accuracy assesses overall correctness, while precision focuses specifically on the accuracy of positive predictions, providing insights into the model's ability to avoid false positives. The choice of which metric to prioritize depends on the specific requirements and goals of the problem at hand.

from those in other countries. This aspect is crucial as the distance measure involved in the pair comparisons may work better depending on the language or transliteration rules. In Section 4, we address these challenges by validating our algorithm using transactions from a diverse set of countries. This selection includes countries with variations in official language and alphabet, allowing us to account for such heterogeneity.

3.2 The steps of the matching procedure

3.2.1 Data cleaning and harmonization

Before proceeding with the actual matching, we perform some basic data cleaning and harmonization. Regarding cleaning the customs data, we try to safeguard as many transactions as possible. We exclude some commodity codes (e.g., gold, banknotes, works of art) because statistical compilation manuals dictate that they should be excluded from International Merchandise Trade Statistics, and eliminate transactions below one thousand euros[17] or with missing or null names of the foreign counterpart. It should be noted that a considerable number of transactions is lost during this preliminary cleaning procedure (see also Figures A1 and A2 in the Appendix). However, their volume represents a small share of the overall trade, because a large amount of these records involved less than € 1,000, and such a share remains constant over time. The cleaning in Orbis is also minimal. We only remove entities associated with natural persons.

We implement a procedure to clean and harmonize the names and postal codes of the companies in order to reduce the noise contained in the strings. More in detail, the name harmonization closely follows the routines explained in Thoma et al. (2010)[18] and available as STATA codes in the NBER patent data project page.[19] Concerning postcodes, we use open-source information[20] to obtain a list of accepted postcode formats across countries, which is the starting point of the harmonization procedure.[21]

3.2.2 The matching procedure

For each extra-EU country, we propose a fixed sequential procedure, consisting of seven steps, where at each stage, we match a proportion of the firms appearing in ADM data with some company available in Orbis relying on step-specific identifiers (the company's name or some function of it together with the postal code, if present). After a step is completed, we remove the matched entities from the pool of matchable ones, which instead go through the following step. The outcome of the matching procedure is a correspondence table linking companies in ADM with at least one entity in Orbis.

[17] The same approach ISTAT adopts in processing ADM data, for confidentiality reasons.

[18] The main procedures adopted within the harmonization are trimming of leading and trailing spaces, removing of punctuation, capitalizing names, converting foreign characters to their English-alphabet corresponding replacements, standardizing or removal of entity types, dropping of one-letter words, removal of spelling variations.

[19] The STATA codes where translated into Python by Riccardo Maria Nusca.

[20] https://en.wikipedia.org/wiki/List_of_postal_codes. While most countries use a postal code system, there are some exceptions, especially in Africa.

[21] In particular, for each country using a postal code system, we drop non-alphanumeric characters from postcodes and invalid entries (e.g., postal codes exclusively consisting of zeros). Moreover, we exclude those postcodes whose length was smaller than shortest accepted format for that countries. Finally, for each postcode, we only consider the first *n* digits (with *n* being the number of digits of the shortest accepted format for the corresponding country).

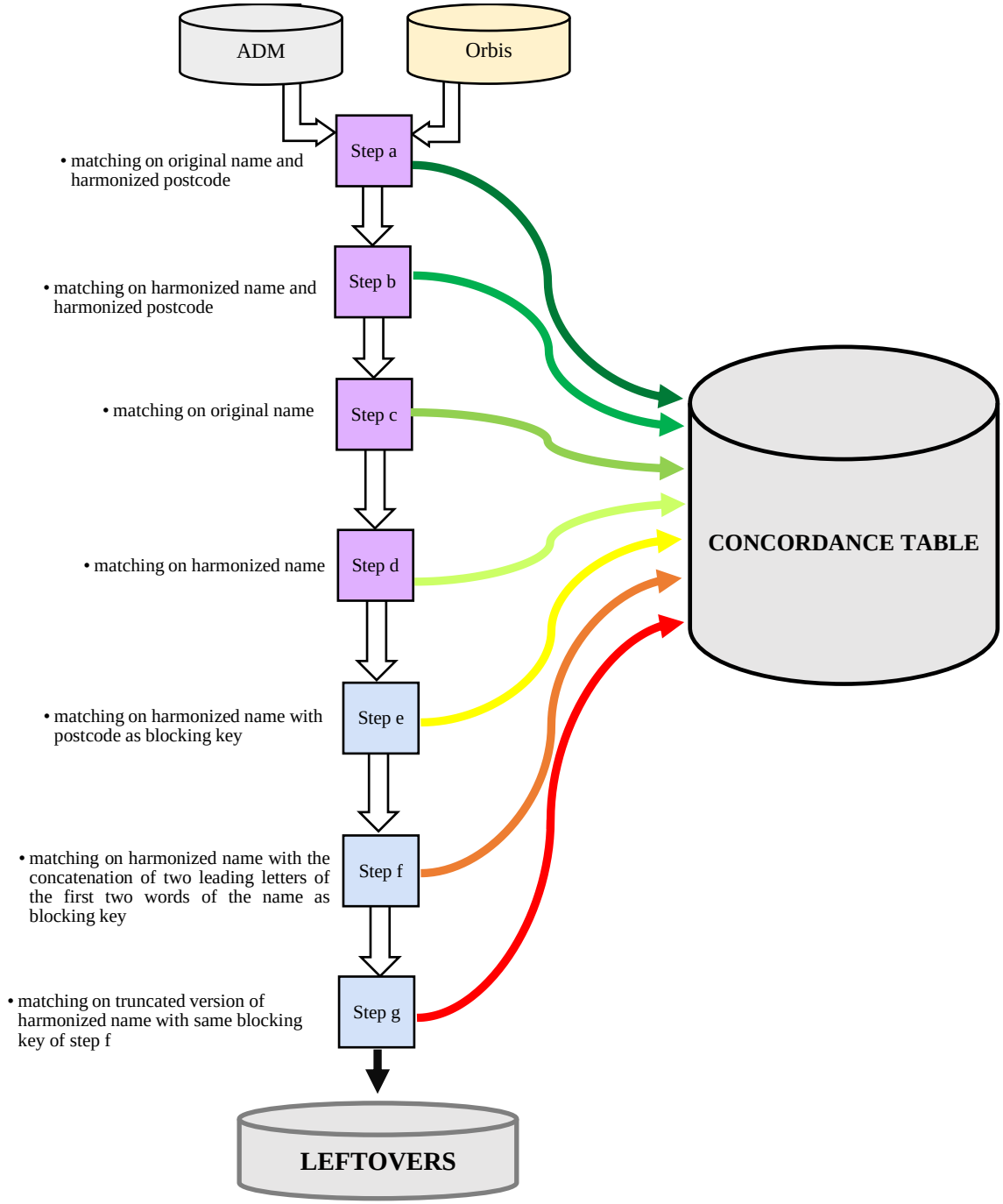
The proposed methodology goes as follows. After the basic data cleaning operations and harmonization routines on the name and postal codes of the records explained in Subsection 3.2.1, we proceed with the actual matching. We start with four *deterministic* steps. First, we identify all records with the same original name and harmonized postcode in ADM and Orbis (*step a*). Second, we isolate all records with the same harmonized name and postcode (*step b*). The other two deterministic steps do not rely on the postcode. In *step c*, we find all records sharing the same original name only, whereas we identify all records with the same harmonized name in *step d*. After completing the previous stages, we proceed with three additional probabilistic steps, which utilize record linkage algorithms, to find suitable matches for the remaining unmatched observations. In these additional steps, we leverage *ad hoc* blocking keys to effectively manage the matching process and reduce the complexity of the task.[22] The blocking keys enable us to subset the potential matches and focus on relevant pairs for further analysis. In *step e*, we perform record linkage on the harmonized name using the harmonized postcode as the blocking key. Hence, we first group the data based on the harmonized postcode, which helps to narrow down the potential matches. Then, within each postcode group, we apply the record linkage algorithm, with Jaro-Winkler distance and threshold set to 0.85, to compare and identify similar records based on the harmonized name. Moving to *step f*, we use a different blocking approach: we create a blocking key by concatenating the leading two letters of the first two words of the company name.[23] In *step g*, we continue using the blocking key approach from *step f*. However, we modify the record linkage process by working with a truncated version of the harmonized name in this step. Specifically, we cut the harmonized name to the average string length observed in the company names for that particular country. This average string length is a reference point to determine an appropriate length for the truncated version. By truncating the harmonized name, we aim to focus on relevant information while reducing noise and irrelevant details that may hinder the matching process. In these last two steps, we use the Jaro distance, instead of the Jaro-Winkler metric, as the latter is more conservative than the former. We represent the pipeline of the matching procedure in Figure 1. The deterministic steps (*a*-*d*) are expected to produce more accurate matches with fewer false positives. We also expect the extent of the accurate matches to decrease when we cannot include the postcode (steps *c* and *d*). Conversely, the probabilistic steps (*e*-*g*) introduce uncertainty and may yield more false positives. The accuracy of these steps largely depends on the signal-to-noise ratio of the matching key used to identify these matches. The probabilistic steps can yield satisfactory results if the matching key provides a strong signal and distinguishes true matches from non-matches. However, when the similarity between the compared strings is small or the signal weakens, the probabilistic steps may be less precise and prone to false positives. Within the deterministic steps, it is worth noting that false positives in steps *a* and *c* typically occur

[22] Specifically, given that the number of pairwise comparisons to assess in record linkage is the result of computing the Cartesian product of the entities between two datasets, and each comparison involves a time-consuming operation to compute a string similarity measure, the time complexity of these steps is very high. This complexity becomes infeasible, especially when dealing with a large number of potential candidates to compare, such as in the case of Orbis.

[23] For example, the record BRUNELLO CUCINELLI USA has as blocking key BRCU and will be compared with all records in Orbis that share the same block, such as BROWNS CUSTOMS, BRAD CUNNINGHAM, BROTHERS CUSTOM, BRAMBILLA CUTTING and so on.

due to name homonymy, i.e., where different entities have similar or identical names.[24] On the other hand, false positives in steps *b* and *d* may also be a by-product of the harmonization process, which can marginally introduce variation or errors in the data. The probabilistic record linkage steps instead are less precise, especially when the similarity between the two compared strings is relatively small.

Figure 1 – Pipeline of the steps of the matching procedure



Note: the color of the arrows represents the expected accuracy of the matching in the originating step (green for better accuracy, red for worse accuracy). Steps in mauve are deterministic. Steps in light-blue are probabilistic.

[24] Entities with identical or similar names often belong to the same corporate group, or represent different branches of the same company.

Table 3 – Example of the correspondence table for the US

year	*id ADM*	*id Orbis*	*matching type*	*similarity*
2010	18749820	US2301424	F	0.90
2010	2196693	US3007984	D	
2010	18096005	US3068210	D	
2011	7220558	US2509883	E	0.87
2013	3032944	US1204782	A	
2013	2058197	US2407984	G	0.96
2014	19429240	US1610312	E	0.94
2014	21698120	US1207314	D	
2014	11360954	US1408978	B	
2015	18582528	US1700441	G	0.92
2015	9992005	US2606933	A	
2016	19839359	US2602228	A	
2017	15528039	USFR07699	C	
2018	30740749	US2302391	F	0.85
2018	28321158	US2310945	C	
2019	19877879	US2775108	F	0.89
2020	28264246	US1210527	F	0.87
2021	13517488	US*104185	D	
2021	8628346	US1905524	A	

Note: *id Orbis* has been randomly changed to preserve the anonymity of the foreign counterparts.

Our procedure yields a correspondence table (see Table 3 for an example), which maps all extra-EU records in ADM to firms in Orbis. In particular, for each unique record in the customs data (given by the combination of *year* and *id ADM*[25]), we provide the BvD ID number of the matched company (*id Orbis*) and the matching step at which the record was achieved (column *matching type*), as well as the value of the string distance in case the matching is probabilistic (column *similarity*). The matching type and distance value can be valuable for empirical analysis as they help decision-making when facing the trade-off between a larger sample size (containing all matching types and/or all distance values), which has a larger probability of false matches, and a smaller sample size (selecting a subset of matching types or a threshold for distance), which likely presents a smaller share of false matches. This information allows researchers to make an informed choice based on their specific needs and requirements.

To conclude this section, it is worth mentioning that the algorithm may yield multiple matches for a given record. This occurrence can be attributed to various factors depending on the specific step of the procedure. One possible scenario, arising typically in steps *c* and *d*, is when multiple entries in Orbis share the same original/harmonized name but different postcodes. This situation may arise because these entries represent different branches or subsidiaries of the same company in the same country. Another scenario, happening in the probabilistic steps, is when different entries in the same block of Orbis share the same value of the string distance with a given record in ADM. In these cases, the algorithm is unable to discern among the multiple

[25] Note that *id ADM* is an internal identifier for the records (at the level of single transaction) that we built for technical reasons. It is not connected to any official firm identifier.

matches found, and therefore we decided to report in the correspondence table the ids corresponding to all potential firms of Orbis separated by a comma. Table 4 shows that the share of multiple matches varies by country, with a low share in India (8%), and a large share in Australia (75%).

Table 4 – Share of multiple matches by country

	Country														
	UK	**US**	**CH**	**CN**	**TR**	**JP**	**RU**	**CA**	**KR**	**BR**	**AU**	**IN**	**MX**	**EG**	**SA**
Multiple matches (%)	43%	56%	22%	11%	42%	29%	34%	45%	58%	51%	75%	8%	9%	29%	31%

Note: 'Multiple matches (%, companies)' indicates the percentage of records in ADM data for which there exist more than one matched companies in Orbis. Countries are reported according to their respective ISO-3166-1 alpha-2 code.

The presence of multiple matches indicates the need for further analysis and decision-making to determine the correct ones before using the microdata. One of the main factors contributing to the existence of multiple matches is the size and nature of the search space, namely Orbis. Several companies may share similar names, especially if they belong to the same corporate group. Moreover, Orbis includes inactive or dissolved firms, which may result in re-registration of businesses with the same name but with a different entity type. Addressing these cases becomes essential to mitigate the issue of multiplicity. However, we leave these additional steps to be handled by future users of the data.

In particular, future users can integrate the matching process with supplementary information or criteria specific to their research objectives. This approach can help reduce the problem effectively. In this regard, it is worth to emphasize that accurate, supplementary information plays a crucial role in this context. Generally, the frequency of multiple matches is lower for matching steps that utilize more data, such as the postcode (Table A4 in the Appendix). Therefore, leveraging such specific information during the matching process can significantly decrease ambiguity and improve results.

4. Results and validation

We report the results regarding matched shares of records and trade value in Subsection 4.1. The remaining subsections are devoted to validating the matches. As already mentioned, the common practice to evaluate the performance of a classification algorithm is to exploit ground-truth data, that is, data annotated with the correct answers. In our case, this would amount to a subset of the extra-EU firms for which we know the exact correspondence in Orbis. Unfortunately, this information is not available in our data by default, and we have to resort to other strategies to evaluate the performance of our algorithm.

4.1 Results

We focus on Italy's most important extra-EU export partners, presenting country-level results demonstrating the variability in the share of matched records and values across different export destinations. This heterogeneity can be attributed to various factors, such as the initial data quality in customs declarations, different transliteration practices, and country coverage in the Orbis database.

Table 5 presents the share of value (second column) and of records (third column) of the cleaned ADM data[26] that the algorithms matches. The results indicate that a vast majority of records and values can be matched for Italy's most important trading partners, surpassing 85%. Russia, however, stands out as a notable exception, with a relatively low share, less than 70%, in terms of matched records and value. The proportion of matched value is generally higher than the share of matched records, suggesting that our procedure matches disproportionately high-value transactions, likely carried out with large counterparts. In Table A3 of the Appendix we report the matched records and value for some extra-EU country stratified by step of the algorithmic procedure.

It is crucial to emphasize that enhancing the algorithm and refining the matching procedure may not necessarily lead to a higher percentage of matches because the search space we operate within (Orbis) does not encompass a considerable portion of the ADM counterparts we endeavor to match. The latter indeed comprise various entities, including entities that do not have to present a balance sheet and may not be considered in Orbis (e.g., private individuals). As a result, the limited matches obtained may not solely be attributed to algorithmic or procedural limitations but rather to the inherent diversity and complexity of the data being processed. Part of these problems are addressed in the harmonization steps of the algorithm, but a deep, ad hoc cleaning may be necessary in specific situations (for example, by removing additional country-specific stop-words or customizing the company name harmonization).

Table 5 – Matched records and value for Italy's top-10 extra-EU exports partners

Country	Value of records (cleaned ADM data, %)	Records (cleaned ADM data, %)
United States	89.23%	92.08%
Switzerland	88.90%	82.15%
China	88.26%	86.27%
Turkey	79.01%	83.44%
Japan	86.67%	83.49%
Russia	67.41%	68.91%
Canada	85.88%	81.42%
South Korea	86.31%	84.03%
Brazil	89.27%	89.70%
Australia	90.50%	88.35%

Note: the reported countries represent Italy's top-10 extra-EU export partners with a postal code system.

4.2 Validation using Brexit

Before Brexit, the UK belonged to the European Union Customs Union, and transactions with UK counterparts had to comply with the intra-EU rules. This included the requirement to provide the VAT number of the British counterpart. As a consequence of Brexit, starting from January 2021, trade data with the UK is instead collected via the SAD, similar to any other extra-EU commercial partner. Consequently, the VAT number of the British counterpart is no longer requested, but there is a designated field for providing the name of the company (see

[26] That is, after the cleaning steps described in Subsection 3.2.1.

Section 2.1). However, there appears to be some behavioral stickiness, since very often the Italian exporters keep reporting the VAT number of their British counterparts after Brexit. This information, coupled with the name of the companies, is extremely useful as it makes validating our matching procedure possible: we first run the procedure considering only the name and the postal code fields, disregarding the information about the VAT number; then, we check the accuracy of the matches by comparing the VAT number in the SAD with the one retrieved from Orbis through our matching procedure.

Our procedure matches the UK foreign counterpart of 1.34 million transactions (about 90 percent of the total).[27] The evaluation of the accuracy of the matching algorithm requires a few steps. First, we keep the records with the VAT number field of the SAD filled in according to the standard format for the UK (11 characters long starting with 'GB'). Moreover, we exclude cases in which the VAT number of the UK counterpart coincides with that of the Italian exporter and those containing non-alphanumeric characters. This selection allows us to exclude from the validation the records containing firms whose identities cannot be verified by means of the VAT information. This cleaning procedure leaves a sample of 120 thousand matched records with a valid and verifiable VAT number. Second, we verify if the VAT number from the SAD corresponds to the VAT number from Orbis, according to our matching procedure.[28]

Table 6 - Results of the validation by number and value of Italian exports to the UK

	Number of records				Value of records (€ mn.)			
Matching step	tot	tot correct	% correct	% cumulative	tot	tot correct	% correct	% cumulative
A	21,409	20,656	96.48%	17.08%	241.31	231.90	96.10%	15.14%
B	44,469	42,181	94.85%	51.95%	617.58	516.59	83.65%	48.87%
C	11,770	11,435	97.15%	61.40%	166.22	141.23	84.96%	58.09%
D	23,513	19,921	84.72%	77.87%	257.59	199.60	77.49%	71.13%
E	15,486	11,813	76.28%	87.63%	199.05	127.10	63.85%	79.43%
F	3,714	2,393	64.43%	89.61%	44.26	27.44	62.01%	81.22%
G	604	410	67.88%	89.95%	5.49	3.29	59.96%	81.43%
	120,965	108,809		89.95%	1,531.52	1,247.16		81.43%

Note: Left panel reports the available number of records ('tot'), correctly matched records ('tot correct'), the percentage of correct records out of the matched ones for each step ('% correct'), the cumulative percentage of correct matches out of the number of matched ('% cumulative'). Right panel reports the total value (in millions of euros) of the SADs in millions of EUR ('tot'), the value of correctly matched records (second column), the percentage of the value of correctly matched records out of the matched value for each step ('% correct'), the cumulative percentage of the value of correctly matched records out of the matched value. All results are stratified by the step of the procedure described in Section 3, from A to G.

Table 6 reports the validation results regarding records (left, green heading) and their values (right, orange heading), differentiating across the various steps of the matching procedure outlined in Section 3. Overall, the validation exercise on UK imports shows that the algorithm effectively links entities within the Italian customs data to their counterparts in Orbis. Specifically, the counterparts of 90% of the transactions are correctly

[27] The procedure is unable to match 158 thousand records.
[28] For records associated with multiple entries in Orbis, we verify that at least one is correct.

assigned to an entity in Orbis, according to our VAT-based validation strategy. The algorithm's performance is slightly lower when considering the values associated with the matched records.

As anticipated, the algorithm's precision diminishes as we progress through the matching steps. This can be attributed to two key factors. Firstly, the remaining entities that have yet to be matched become increasingly challenging to link. Secondly, as we advance in the matching process, the quantity and quality of information available for matching purposes decrease. However, we note that the final steps only marginally contribute to the overall percentages of matched volumes, at least for the UK (see also Table A3 of the Appendix).

Concerning the precision of our matching, a careful inspection of the results allows us to reflag as correct a significant portion of the matches labeled as wrong by this VAT-based validation strategy. In fact, the administrative nature of the ADM data implies that there are issues with the compilation of the VAT number in the SAD. For instance, frequently the VAT number reported in the SAD happens to be invalid or outdated. Sometimes, the VAT number attributed through Orbis only differs by a few digits from that inserted in the customs data. In both cases the validation strategy automatically flags the matched pair as incorrect, but oftentimes having a visual look at such pairs, it is clear that they are indeed correct. Among the pairs flagged as wrongly matched, we only check, within each step, the most important one by aggregated value of the transactions. In step *a*, all matches, that by construction have the same name and postcode, can be classified as correct, despite not sharing the VAT number. In steps *b*, *c*, *d*, and *e*, this further inspection allows us to reclassify all the pairs checked as correct, resulting in a considerable improvement in accuracy, as reported in Table 6 (right panel, third column): in step *b*, it becomes 94%; in step *c*, 97.5%; in step *d*, 80%; and in step *e*, 78.2%. However, in step *f*, the checked pair was found to be genuinely incorrect.

The presented example serves as compelling evidence that the precision of the matching process is significantly underestimated by the figures in Table 6. This underestimation arises due to the inaccuracy of the non-harmonized VAT number information for the British counterpart in the two data sources. Unfortunately, this lack of harmonization increases the likelihood of mislabeling the matched pairs.

4.3 Correspondence between exporters' sector of activity and exported products

The second validation exercise focuses on extra-EU companies exporting goods to Italy.[29] For these transactions, the customs declarations do not contain any consistent identifier for the foreign counterpart, as discussed earlier in the text. Hence, we resort to an alternative strategy to assess the quality of our matching, which exploits additional variables contained in the ADM customs data, namely, the commodity code of a transaction, and the industrial sector of the matched foreign counterpart obtained through Orbis. The idea behind this validation exercise is that firms operating in a given sector are more likely to produce specific commodities closely associated with the nature of their industry as an output of their production process.[30]

[29] We focus on firms whose matched NACE code belongs to the first two NACE Rev. 2 Sections: (i) Agriculture, forestry and fishing, and (ii) Manufacturing, mining and quarrying and other industry. In particular, we do not include wholesalers and retailers.

[30] For example, a firm manufacturing motor vehicles is more likely to export cars.

Hence, checking the correlation between sectors of activity and exported products may constitute an indirect way to validate the matching. Clearly, there are some limitations and caveats to remember when interpreting this exercise's results. First, a firm may produce several products, sometimes even far from the main sector of activity indicated in Orbis. Second, the industry sector may not be reported in Orbis (as discussed in Subsection 2.2). Finally, due to name similarity, the matched firm from Orbis may be the holding or the financial conglomerate associated with the manufacturer. However, despite these limitations, we expect to find a strong correlation between exporters' sectors of activity and their exported commodities.

The validation exercise includes three main ingredients. First, we observe the product sold by an extra-EU counterpart to an Italian importer. The SAD contains a field completed with the code identifying the type of commodity traded, that is, the 8-digit code from the Combined Nomenclature, (CN).[31] In this exercise, we select the leading two digits of the CN code (corresponding to the Harmonized System, HS, chapter). The second ingredient is the sector of the extra-EU company that sells the good to the Italian firm. We obtain this information for the matched firms through Orbis, which indicates the company's primary activity as a NACE 4-digit code. For simplicity, we focus on the leading 2-digit of the NACE 4-digit code (corresponding to the Division, see also Table A2 in the Appendix), and we keep only transactions associated with a unique identifier in Orbis. Finally, to link the NACE sector to the HS chapter, we use the Bilateral Trade in Goods by Industry and End-use (BTDIxE; OECD, 2021) conversion keys between different versions of the HS subheadings (HS 6-digit code) and the International Standard Industrial Classification (ISIC), which is the standard classification system underlying NACE. We consider a simplified version of such a conversion key, consisting of the leading two digits of each code. This choice allows us to obtain meaningful, easy-to-interpret patterns relating products to sectors.

In Table 7, we report the result of the validation for the UK in terms of the number of companies (left, green heading) and the total value of the transactions (right, orange heading), separately for the different steps of the matching procedure.

Overall, we observe similar patterns to those discussed in Subsection 4.2: initial matching steps produce more accurate matching than final ones. The matching between sectors and product categories is particularly good in values (almost 93%), whereas it is slightly worse regarding the number of companies, probably reflecting a relatively weak performance on small transactions. We repeat the same analysis for other extra-EU countries and report the aggregated results in Table 8. For the other extra-EU countries, we find results broadly in line with what was observed for the UK. We notice a low share of corresponding sectors and products if we look at the number of firms for Russia, Egypt, and Saudi Arabia. However, in values, the performance aligns with what we have for the other countries.

[31] The CN represents the good nomenclature valid with the European Union Customs Union. It integrates the standard Harmonized System (HS) with an additional 2-digit subheading.

Table 7 – Correspondence between sectors and exported products from the UK by matching step

	Number of companies				Value of records (€ mn.)			
Matching step	tot	matched sector-product	matched sector-product (%)	% cum	tot	matched sector-product	matched sector-product (%)	% cum
A	1,456	1,181	81.11%	12.83%	4,000	3,880	97.01%	36.23%
B	3,481	2,815	80.87%	43.41%	3,320	3,105	93.52%	65.23%
C	697	550	78.91%	49.38%	1,089	915	84.05%	73.77%
D	1,649	1,265	76.71%	63.12%	1,521	1,427	93.81%	87.10%
E	1,417	1,014	71.56%	74.14%	354	300	84.62%	89.90%
F	449	292	65.03%	77.31%	398	268	67.20%	92.40%
G	57	41	71.93%	77.75%	26	22	86.20%	92.61%
	9,206	7,158		77.75%	10,709	1,247		92.61%

Note: In the left panel, for each matching step, we report the number of unique companies matched in Orbis ('tot'), the number for which we find a correspondence among sector and product exported according to the OECD BTDIxE conversion key ('matched sector-product), its percentage ('matched sector-product (%)'), and the cumulative percentage of matched sector-product ('% cum'). In the right panel, for each matching step, we report the total value of the transactions of matched firms in € mn. ('tot), the total value of transactions for which we find a correspondence among sector and product exported ('matched sector-product), its percentage ('matched sector-product (%)'), and the cumulative percentage of corresponding sectors and products ('% cum').

However, as already mentioned, there are limitations to consider, such as firms producing multiple products and potential discrepancies in reported industry sectors. In order to assess the relevance of these limitations, we repeated the exercise for Italian exports for which we have the VAT number and are therefore able to retrieve the corresponding industry sector. In particular, we randomly select 2 million export transactions and check the correspondence between the products exported by the Italian firm and the latter sector of activity. By doing so, we can determine the maximum correlation we would expect to find. For this exercise the matching between sectors and product categories is 97%, whereas it is amounts to 87% regarding the number of companies. These figures help putting the results of this validation exercise into perspective.

Table 8 - Correspondence between sectors and products

	Country														
	UK	US	CH	CN	TR	JP	RU	CA	KR	BR	AU	IN	MX	EG	SA
Matched sector-product (%, companies)	78%	78%	81%	78%	77%	83%	66%	76%	80%	81%	71%	75%	71%	57%	60%
Matched sector-product (%, value)	93%	96%	99%	96%	93%	97%	83%	98%	95%	88%	88%	92%	97%	82%	99%

Note: 'Matched sector-product (%, companies)' indicates the percentage of companies for which we find a correspondence among sector and product exported according to the OECD BTDIxE conversion key; 'Matched sector-product (%, value)' indicates the percentage of the total value of transactions for which we find a correspondence among sector and product exported. Countries are reported according to their respective ISO-3166-1 alpha-2 code.

5. Discussion and concluding remarks

The algorithm proposed in this paper demonstrates good performance in matching the Italian customs data to the corresponding companies from Orbis, with 85% of overall matched foreign partners and satisfying accuracy. However, it is important to stress again some aspects that could affect the quality and limit the scope of the matching.

First, Orbis does not provide the universe of Italian firms' partners and varies in coverage across partner countries. Even within a given country, the representativeness of the data may differ across firm size and sectors. Furthermore, the information needed to identify the entities correctly, such as the company name or the location, may sometimes be inaccurate or incomplete. Similar concerns arise in the case of customs declarations, in which compilation errors, missing data, and inconsistencies may impair the ability of identifying the foreign counterparts. To address these issues, we rely on name harmonization routines, which, however, can be further customized and fine-tuned to accommodate the specific features of each country's language and business framework.

Second, our procedure often delivers multiple matches, which is largely a byproduct of the quality of the underlying sources. Because of that, the selection of the most plausible matches requires additional criteria. One possible direction is to resort on the correspondence between industrial sector and traded commodities. We leave this aspect for future research. Moreover, future research can leverage the sequential nature of new ADM intakes to improve the procedure by using previously matched entities as reference points.

We remark that our procedure primarily aims to match the most comprehensive part of ADM data to Orbis while preserving as many observations as possible to benefit the research community the most, providing a tool that serves as a starting point for their research. Users should critically evaluate the limitations outlined above and, possibly, tailor the matching procedure to their specific research context, making informed decisions and employing suitable validation techniques to ensure the reliability and accuracy of their results.

References

Adao, R., Carrillo, P., Costinot, A., Donaldson, D., & Pomeranz, D. (2022). Imports, exports, and earnings inequality: Measures of exposure and estimates of incidence. *The Quarterly Journal of Economics*, 137(3), 1553-1614.

Alfaro-Urena, A., Manelici, I., & Vasquez, J. P. (2022). The effects of joining multinational supply chains: New evidence from firm-to-firm linkages. *The Quarterly Journal of Economics*, 137(3), 1495-1552.

Alter, A., & Elekdag, S. (2020). Emerging market corporate leverage and global financial conditions. *Journal of Corporate Finance*, 62, 101590.

Alviarez, V.I., Fioretti, M., Kikkawa, K., & Fioretti, M. (2023). Two-Sided Market Power in Firm-to-Firm Trade (No. w31253). National Bureau of Economic Research.

Amiti, M., Duprez, C., Konings, J., & Van Reenen, J. (2023). FDI and Superstar Spillovers: Evidence from firm-to-firm transactions (No. w31128). National Bureau of Economic Research.

Bajgar, M., Berlingieri, G., Calligaris, S., Criscuolo, C., & Timmis, J. (2020). Coverage and representativeness of Orbis data (No. 2020/06). Organization for Economic Co-operation and Development.

Bernard, A. B., Moxnes, A., & Saito, Y. U. (2019). Production networks, geography, and firm performance. *Journal of Political Economy*, 127(2), 639-688.

Carvalho, V. M., Nirei, M., Saito, Y. U., & Tahbaz-Salehi, A. (2021). Supply chain disruptions: Evidence from the Great East Japan earthquake. *The Quarterly Journal of Economics*, 136(2), 1255-1321.

Cohen, G. J., Dice, J., Friedrichs, M., Gupta, K., Hayes, W., Kitschelt, I., ... & Webster, C. (2021). The US syndicated loan market: Matching data. *Journal of Financial Research*, 44(4), 695-723.

Dhyne, E., Kikkawa, A. K., Mogstad, M., & Tintelnot, F. (2021). Trade and domestic production networks. *The Review of Economic Studies*, 88(2), 643-668.

Dhyne, E., Kikkawa, A. K., & Magerman, G. (2022). Imperfect competition in firm-to-firm trade. *Journal of the European Economic Association*, 20(5), 1933-1970.

Eaton, J., Kortum, S. S., & Kramarz, F. (2022). Firm-to-Firm Trade: Imports, exports, and the labor market (No. w29685). National Bureau of Economic Research.

Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328), 1183-1210.

Huneeus, F. (2020). Production network dynamics and the propagation of shocks. Mimeo.

Kalemli-Ozcan, S., Sorensen, B., Villegas-Sanchez, C., Volosovych, V., & Yesiltas, S. (2023). How to Construct Nationally Representative Firm Level Data from the Orbis Global Database: New Facts on SMEs and Aggregate Implications for Industry Concentration. *American Economic Journal: Macroeconomics* (forthcoming).

OECD (2021), Bilateral Trade Database by Industry and End-use (BTDIxE), http://oe.cd/btd/.

Pustilnik, B. (2023). Trade policy on a buyer-seller network. Mimeo.

Tahamont, S., Jelveh, Z., McNeill, M., Yan, S., Chalfin, A., & Hansen, B. (2023). No Ground Truth? No Problem: Improving Administrative Data Linking Using Active Learning and a Little Bit of Guile (No. w31100). National Bureau of Economic Research.

Thoma, G., Torrisi, S., Gambardella, A., Guellec, D., Hall, B. H., & Harhoff, D. (2010). Harmonizing and combining large datasets. An application to firm-level patent and accounting data (No. w15851). National Bureau of Economic Research.

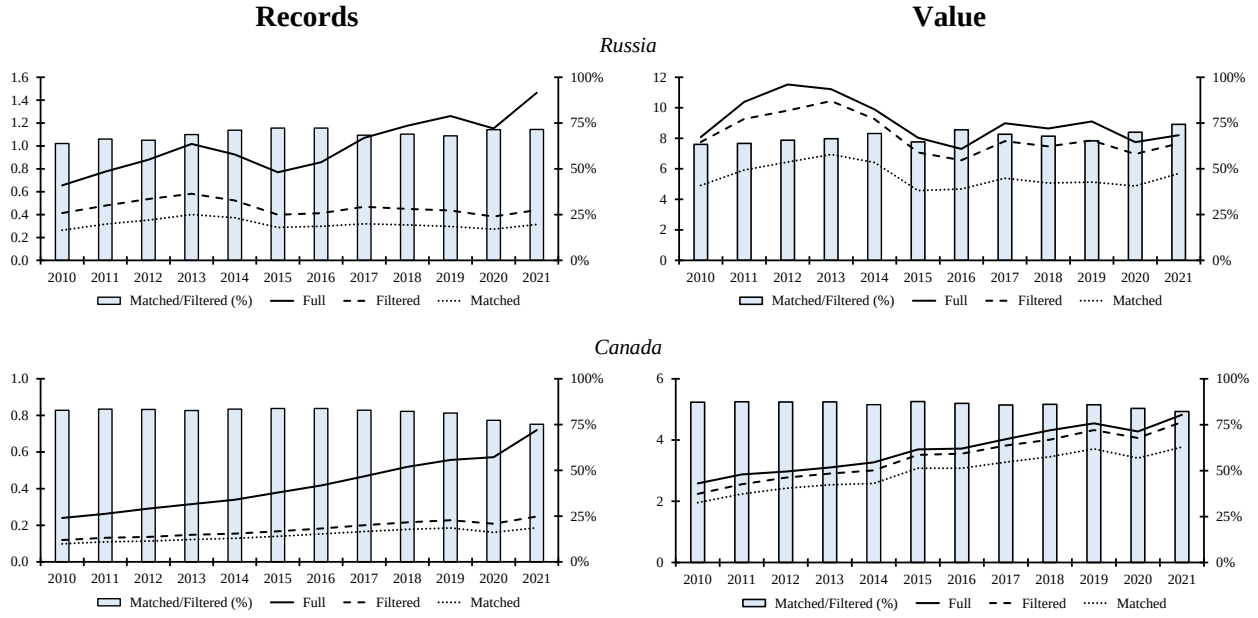
Appendix: additional tables and figures

Table A1 – Top-20 Italy's import and export extra-EU partners in 2021

Import		Export	
Country	*Share*	*Country*	*Share*
China	18.53%	USA	20.05%
Russia	8.95%	Switzerland	11.06%
USA	7.59%	United Kingdom	9.51%
Switzerland	5.38%	China	6.36%
Turkey	4.73%	Turkey	3.87%
Azerbaijan	4.43%	Russia	3.12%
United Kingdom	3.88%	Japan	3.07%
India	3.17%	South Korea	2.14%
Lybia	3.03%	Canada	1.96%
Algeria	2.97%	United Arab Emirates	1.96%
Saudi Arabia	2.33%	Hong Kong	1.95%
Japan	2.14%	Brazil	1.86%
Brazil	2.05%	Australia	1.75%
South Korea	2.03%	India	1.58%
Vietnam	1.69%	Mexico	1.57%
Iraq	1.61%	Egypt	1.54%
Ukraine	1.58%	Saudi Arabia	1.36%
Tunisia	1.26%	Israel	1.25%
Indonesia	1.16%	Tunisia	1.16%
Taiwan	1.13%	South Africa	0.91%

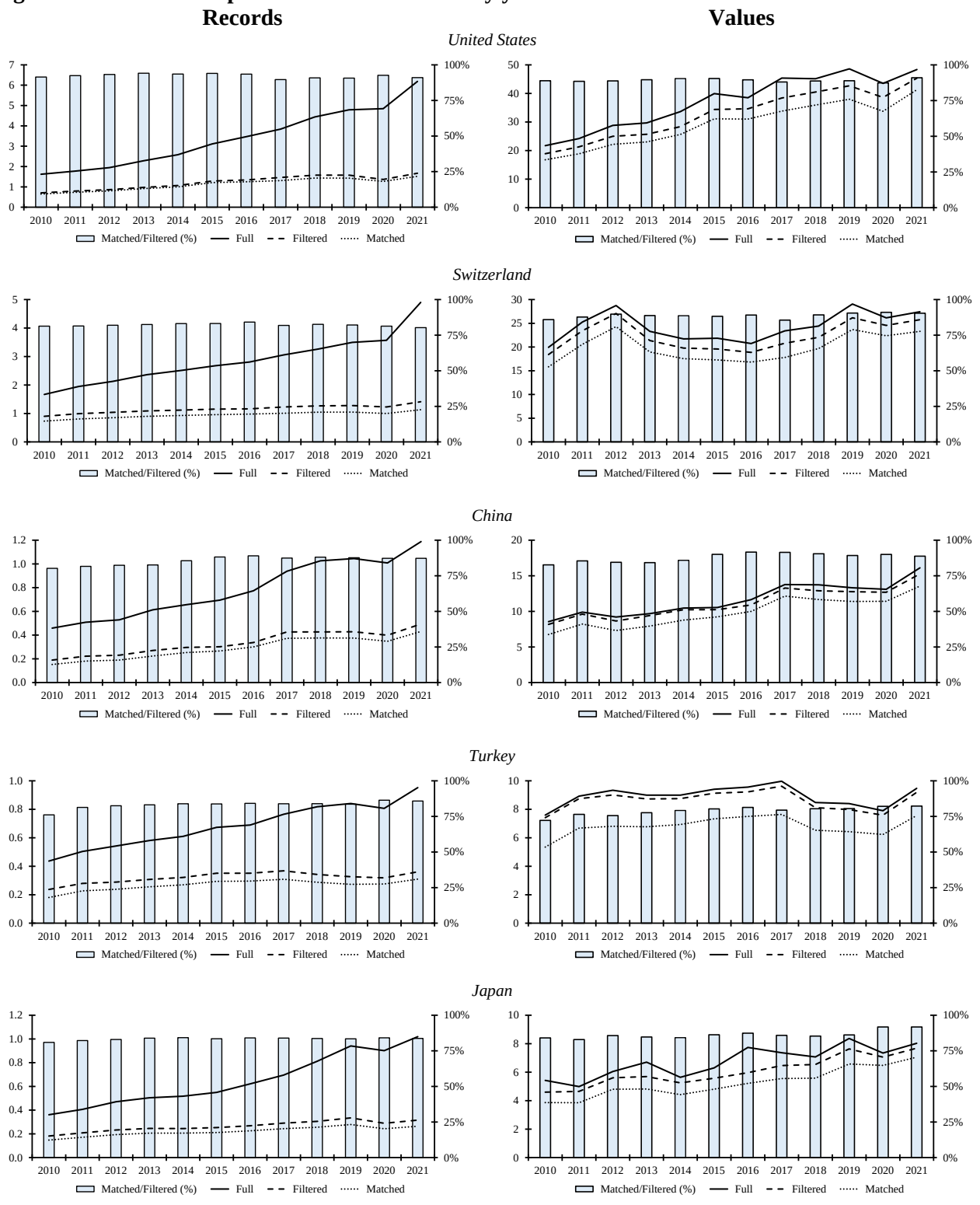
Note: authors' elaboration on ISTAT data.

Figure A1 - Matched exports records and values by year



Note: Plots on the left-hand side: the left axis is the number of records (in mn.), the right axis is the share of matched records over the filtered ones. Plots on the right-hand side: the left axis is the value of the records (in € bn.), the right axis is the share of value matched over the filtered value. For each country and year, 'Full' represents the total number of exports records (value) in the Italian Customs Data; 'Filtered' represents the residual number of exports records (value) after the cleaning procedure of Subsection 3.2; 'Matched' represents the total number of exports records (value), which is matched to an Orbis counterpart. The bars represent the share of matched filtered records (value).

Figure A2 – Matched exports records and values by year



Note: Plots on the left-hand-side: the left axis represents the number of records (in mn.), the right axis the share of matched records over the filtered ones. Plots on the right-hand-side: left axis represents the value of the records (in € bn.), the right axis represents the share of value matched over the filtered value. For each country and year, 'Full' represents the total number of exports records (value) in the ADM raw data; 'Filtered' represents the residual number of exports records (value) after the cleaning procedure of Subsection 3.2.1.; 'Matched' represents the total number of exports records (value), which is matched to an Orbis counterpart. The bars represent the share of matched filtered records (value).

Table A2 – Sectors description in Orbis

Sect.	Description	Sect.	Description	Sect.	Description	Sect.	Description
01	Crop and animal production, hunting and related service activities	24	Manufacture of basic metals	50	Water transport	75	Veterinary activities
02	Forestry and logging	25	Manufacture of fabricated metal products, except machinery and equipment	51	Air transport	77	Rental and leasing activities
03	Fishing and aquaculture	26	Manufacture of computer, electronic and optical products	52	Warehousing and support activities for transportation	78	Employment activities
05	Mining of coal and lignite	27	Manufacture of electrical equipment	53	Postal and courier activities	79	Travel agency, tour operator, reservation service and related activities
06	Extraction of crude petroleum and natural gas	28	Manufacture of machinery and equipment n.e.c.	55	Accommodation	80	Security and investigation activities
07	Mining of metal ores	29	Manufacture of motor vehicles, trailers and semi-trailers	56	Food and beverage service activities	81	Services to buildings and landscape activities
08	Other mining and quarrying	30	Manufacture of other transport equipment	58	Publishing activities	82	Office administrative, office support and other business support activities
09	Mining support service activities	31	Manufacture of furniture	59	Motion picture, video and television programme production, sound recording and music publishing activities	84	Public administration and defence; compulsory social security
10	Manufacture of food products	32	Other manufacturing	60	Programming and broadcasting activities	85	Education
11	Manufacture of beverages	33	Repair and installation of machinery and equipment	61	Telecommunications	86	Human health activities
12	Manufacture of tobacco products	35	Electricity, gas, steam and air conditioning supply	62	Computer programming, consultancy and related activities	87	Residential care activities
13	Manufacture of textiles	36	Water collection, treatment and supply	63	Information service activities	88	Social work activities without accommodation
14	Manufacture of wearing apparel	37	Sewerage	64	Financial service activities, except insurance and pension funding	90	Creative, arts and entertainment activities
15	Manufacture of leather and related products	38	Waste collection, treatment and disposal activities; materials recovery	65	Insurance, reinsurance and pension funding, except compulsory social security	91	Libraries, archives, museums and other cultural activities
16	Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials	39	Remediation activities and other waste management services	66	Activities auxiliary to financial services and insurance activities	92	Gambling and betting activities
17	Manufacture of paper and paper products	41	Construction of buildings	68	Real estate activities	93	Sports activities and amusement and recreation activities
18	Printing and reproduction of recorded media	42	Civil engineering	69	Legal and accounting activities	94	Activities of membership organizations
19	Manufacture of coke and refined petroleum products	43	Specialized construction activities	70	Activities of head offices; management consultancy activities	95	Repair of computers and personal and household goods
20	Manufacture of chemicals and chemical products	45	Wholesale and retail trade and repair of motor vehicles and motorcycles	71	Architectural and engineering activities; technical testing and analysis	96	Other personal service activities
21	Manufacture of basic pharmaceutical products and pharmaceutical preparations	46	Wholesale trade, except of motor vehicles and motorcycles	72	Scientific research and development	97	Activities of households as employers of domestic personnel
22	Manufacture of rubber and plastic products	47	Retail trade, except of motor vehicles and motorcycles	73	Advertising and market research	98	Undifferentiated goods- and services-producing activities of private households for own use
23	Manufacture of other non-metallic mineral products	49	Land transport and transport via pipelines	74	Other professional, scientific and technical activities	99	Activities of extraterritorial organizations and bodies

Note: 'Sect.' corresponds to a NACE Rev. 2 Division. Cells in light green are manufacturing sectors.

Figure A3 – Accuracy and precision as a function of the threshold

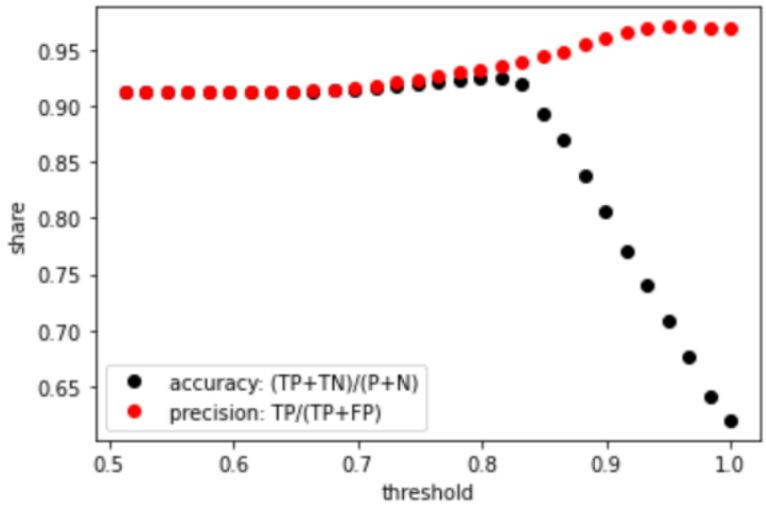


Table A3 – Matched records and value stratified by step of the procedure for some extra-EU countries

	Country									
	United States		Switzerland		China		Turkey		Japan	
Matching step	value (%)	records (%)	value (%)	records (%)	value (%)	records (%)	value (%)	records (%)	value (%)	records (%)
A	10.96%	11.03%	22.04%	18.31%	0.25%	0.22%	0.07%	0.05%	1.25%	1.75%
B	19.58%	18.23%	9.05%	7.66%	4.26%	3.51%	0.33%	0.27%	4.40%	5.42%
C	27.60%	27.10%	27.86%	26.26%	2.85%	2.71%	1.95%	1.54%	23.72%	16.78%
D	20.24%	21.14%	19.58%	15.08%	41.49%	39.40%	4.76%	7.14%	46.82%	48.05%
E	9.44%	9.41%	9.47%	13.73%	5.93%	6.14%	13.56%	13.29%	2.01%	1.99%
F	10.56%	11.31%	9.94%	15.23%	40.07%	42.10%	44.94%	44.04%	17.57%	17.95%
G	1.61%	1.78%	2.07%	3.74%	5.15%	5.92%	34.38%	33.66%	4.22%	8.05%

	Russia		Brazil		Mexico		Saudi Arabia		United Kingdom	
Matching step	value (%)	records (%)	value (%)	records (%)	value (%)	records (%)	value (%)	records (%)	value (%)	records (%)
A	1.87%	1.73%	5.66%	3.86%	4.48%	4.35%	3.37%	1.53%	9.42%	10.24%
B	5.92%	4.48%	5.99%	8.18%	7.91%	7.74%	8.54%	10.15%	20.16%	26.26%
C	18.23%	22.36%	15.21%	13.52%	14.36%	15.95%	9.95%	7.60%	11.60%	13.94%
D	20.20%	18.28%	15.30%	16.10%	23.08%	24.55%	17.46%	22.83%	15.07%	11.63%
E	14.63%	10.98%	13.76%	12.07%	12.43%	11.26%	16.30%	17.55%	32.13%	23.23%
F	33.64%	35.63%	37.42%	40.20%	32.84%	31.40%	27.13%	30.80%	9.97%	11.25%
G	5.52%	6.53%	6.67%	6.07%	4.89%	4.75%	17.24%	9.54%	1.64%	3.46%

	Australia		Egypt		Canada		India	
Matching step	value (%)	records (%)	value (%)	records (%)	value (%)	records (%)	value (%)	records (%)
A	9.87%	8.61%	0.16%	0.12%	4.25%	4.15%	2.57%	2.72%
B	6.45%	7.24%	0.46%	0.30%	7.29%	7.20%	2.74%	2.14%
C	21.39%	22.62%	14.20%	13.69%	33.06%	30.97%	24.09%	15.07%
D	16.94%	19.44%	23.89%	19.88%	32.60%	31.84%	21.66%	16.20%
E	9.28%	11.08%	0.59%	0.57%	3.23%	4.09%	8.21%	9.20%
F	17.66%	21.69%	49.94%	54.57%	16.71%	18.37%	41.88%	46.28%
G	3.94%	2.96%	10.76%	10.88%	2.86%	3.38%	3.76%	4.16%

Table A4 – Multiple matches by country and matching step

	records (%)				
Matching step	**United States**	**Switzerland**	**China**	**Turkey**	**Japan**
A	38.43%	5.53%	1.13%	0.00%	0.41%
B	40.26%	8.35%	3.28%	0.00%	2.19%
C	68.01%	21.38%	9.27%	16.14%	36.29%
D	81.45%	31.88%	10.70%	14.07%	35.82%
E	27.91%	2.62%	2.44%	2.57%	1.52%
F	51.66%	60.73%	9.55%	41.58%	27.39%
G	45.09%	21.59%	34.41%	65.62%	12.79%

Matching step	**Russia**	**Brazil**	**Mexico**	**Saudi Arabia**	**United Kingdom**
A	10.81%	31.65%	0.20%	7.51%	23.00%
B	14.72%	70.06%	1.65%	21.97%	33.34%
C	50.73%	56.64%	4.18%	21.94%	26.38%
D	49.35%	88.44%	12.07%	50.97%	51.13%
E	5.05%	32.15%	2.47%	14.76%	62.20%
F	28.35%	35.69%	9.17%	32.60%	44.57%
G	31.12%	66.10%	40.50%	24.94%	66.87%

Matching step	**Australia**	**Egypt**	**Canada**	**India**
A	51.57%	0.00%	12.44%	1.99%
B	58.96%	2.45%	16.14%	1.81%
C	85.21%	18.26%	52.50%	9.88%
D	93.14%	42.26%	56.26%	15.84%
E	42.51%	0.78%	11.74%	0.96%
F	80.65%	27.30%	37.92%	6.80%
G	74.23%	31.72%	43.23%	16.51%