Statistical Journal of the IAOS 34 (2018) 599–603
DOI 10.3233/SJI-170423
IOS Press

Matching techniques and administrative data records linkage

Haitham Zeidan
Palestinian Central Bureau of Statistics, Ramallah, Palestine
E-mail: Haitham@pcbs.gov.ps

Abstract. The aim of this paper is to show a Palestinian Central Bureau of Statistics (PCBS) [1] experiment in administrative data records linkage. We focused in this paper on PCBS experiment in matching different data sources from different ministries, municipalities and other partners with PCBS Establishments Census 2012. Different matching algorithms and tools were used in the experiment. We started our experiment by using the Fuzzy Lookup [2]. It is an add-in for Excel developed by Microsoft Research. It performs fuzzy matching of textual data in Microsoft Excel. The tool uses the Jaccard Index of Similarity and Levenshtein distance; a statistical way to measure similarities between sample sets. In order to compare data and try to find out matching data, we also used Duke, see Lars [3] which is an existing and flexible deduplication (or entity resolution, or record linkage) engine written in Java. By using Duke Engine, we wrote our matching algorithm and comparators to increase the matching results and matching accuracy. We also wrote some data-cleaning functions for matching variables (Commercial Name, Owner Name and Telephone) in order to standardize each matching variable to get improved results. Different matching algorithms were used in the experiment such as Hamming Distance, e.g. Mohammad [4], Levenshtein distance, Mark [5], Jaccard Similarity, e.g. Suphakit et al. [6], exact match and multiple match.
The results showed that after cleaning the identification variables, the number of matches rises significantly. We also noted that there is an improvement in matching rates when going from the matching based only on phone numbers to the matching based on Telephone, Commercial Name and Owner Name.

Keywords: PCBS, Levenshtein distance, Jaccard similarity, Hamming distance

1. Introduction

Using administrative records is very important for official statistics instead of surveys to collect data for policy decisions. Using administrative record also reduce the costs of data collection and increase the accuracy. For these reasons, administrative records are being used increasingly for statistical purposes [7]. This paper displays PCBS experience in matching different sources with census data.

1.1. Objectives

The objective of this study is to demonstrate a PCBS experiment in matching and administrative data records linkage, where different data sources from different ministries and municipalities are matched with PCBS Establishment Census 2012. Different matching algorithms, techniques and tools were used in the experiment. PCBS intends to build an efficient statistical business register system that should serve the needs of the concerned institutions. Thus, the objectives of the matching process at the end are: (1) evaluating and analysing all registered establishments for all partners, (2) comparing administrative records with Establishment Census 2012, (3) developing a mechanism to improve the quality of administrative records, (4) getting a common definition for statistical business register that serves all partners, (5) and measuring the coverage of registered establishments compared with establishment census 2012.

1.2. String comparator metrics

When comparing values of string variables like names or addresses, it usually does not make sense to

just discern total agreement and disagreement. Typographical error may lead to many incorrect disagreements. Several methods for dealing with this problem have been developed. String comparators are mappings from a pair of strings to the interval [0, 1] measuring the degree of compliance of the compared strings [8]. String comparators may be used in combination with other exact matching methods, for instance, as input to probabilistic linkage, discriminate analysis or logistic regression. The simplest way of using string comparators for exact matching is to define compliance classes based on the values of the string comparator.

1.2.1. Hamming distance

One of the earliest and most natural metrics is the Hamming distance, e.g. Mohammad [4], where the distance between two strings is the number of mismatching characters. In information theory, the Hamming distance between two strings of equal length is the number of positions at which the corresponding symbols are different. Expressed differently, it measures the minimum number of substitutions required to change one string into the other, or the minimum number of errors that could have transformed one string into the other.

1.2.2. Jaccard distance

A statistical way to measure similarities between sample sets is Jaccard Distance. Jaccard similarity is defined as the size of the set intersection divided by the size of the set union for two sets of objects. For two sets X, Y, it is defined to be J (X, Y) = |X ∩ Y|/|X ∪ Y|. The Jaccard distance between the sets, defined as D (X, Y) = 1 − J (X, Y), is known to be a metric. for example, the sets {a, b, c} and {a, c, d} have a Jaccard similarity of 2/4 = 0.5 because the intersection is {a, c} and the union is {a, b, c, d}. The more the two sets have in common, the closer the Jaccard similarity will be to 1.0, e.g. Suphakit et al. [6].

1.2.3. Edit (Levenshtein) distance

Edit distance [5] is a way of quantifying how dissimilar two strings (e.g., words) are to one another by counting the minimum number of operations required to transform one string into the other. Edit distances finds applications in natural language processing, where automatic spelling correction can determine candidate corrections for a misspelled word by selecting words from a dictionary that has a low distance to the word in question. In bioinformatics, it can be used to quantify the similarity of macromolecules such as DNA, which can be viewed as strings of the letters A, C, G and T.

2. Methodology

2.1. Data collection and specification

We started our experiment by collecting files from ministries and municipalities and writing the specification for each file, normalizing and analysing data supplied by various organizations (PCBS Census, Municipalities, Tax Administration), and we worked on reconciling the different data for the same establishments. The purpose of collecting the files was to match these files with Establishment Census 2012.

2.2. Matching variables

In order to compare data and try to find out matching cases, we used Duke which is an existing and flexible deduplication (or entity resolution, or record linkage) engine written in Java on top of Lucene, see Lars [3]. It was easy to get useful results by using it. The data, which we were working on, contain numerous columns, most of which were of no use whatever to find duplicates. For example, the internal identifier (also called primary key for each file) did not help as it was different in each file. Only three columns only were used: Telephone, Commercial Name and Owner Name. We wrote some data-cleaning functions using Duke for these columns (phone numbers and Arabic text) in order to standardize each column to get improved results. We set the probability threshold (presently set to be 0.80) and defined the two files to match and how we wanted to treat our two discriminating properties.

2.3. Data cleaning

Based on data specification, we put our cleaning rules for the matching variables; telephone is a phone number (or several phone numbers) linked to the establishment. The telephone numbers were formatted differently, where not all establishments supplied one, and the phone numbers could belong to different persons (local manager, owner, etc.). We provided a function to normalize the phone numbers, named Phone Cleaner, which allows cleaning up the registered data in each file. Then, we specified our probabilities for the telephone: if one of the phone numbers is the same for an establishment registration in each file, the probability that the establishments are the same is valued 90%. This is above our threshold of 0.80, so unless we later

find evidence indicating that the establishments are different, we will consider them as duplicates.

The columns Commercial Name and Owner Name were addressed with a single cleaner for several reasons after studying the files. Sometimes data were not well filled-up, sometimes we had the commercial name instead of the owner name and vice-versa. We provided a function to normalize the Arabic text, named "text-cleaner". It removes some key words and replaces some other words by more convenient ones in order to standardize data. We provided a function to clean up the data, and then build a comparator in order to be able to compare a variable containing both the commercial name and the owner name. Finally, we specified our probabilities: if the names are the same, the probability that the records themselves are the same is 95%. This is above our threshold of 0.80, so unless we later find evidence indicating that the establishments are different, we will consider them as duplicates.

The columns Telephone or Commercial Name and Owner Name were also used. To combine the phone analysis and the names analysis, we had two probabilities and we had to combine them in order to build a global probability. Let's assume that two organizations have the same commercial and owner name according to our comparator, and same phone numbers, using the formula used by Duke, which is inspired by *naive Bayes* inference. That gives us 0.95 and 0.90 probability, which combines to 0.97, higher than the score for each match using Telephone or Commercial Name and Owner Name. This high score reinforces the probability that we consider the two establishments as duplicates, unless we later find evidence indicating that the establishments are different.

2.4. Data matching

In matching process we focused on Duke, since Duke can find duplicate records. We can also use it to connect records in one data set with other records representing the same thing in another data set. Duke has sophisticated comparators like Levenshtein, Jaro-Winkler, and Dice coefficient that can handle spelling differences, numbers, geo-positions and more. Using a probabilistic model Duke can handle noisy data with good accuracy. We made some matching exercises on some files from municipalities and ministries. The cleaning of the data provided in every cases good results; thus, cleaning data before matching is very important to increase the accuracy of matching and to enhance the matching results. We used files provided by other municipalities and their description in order to run other matching exercises. Whenever needed, we updated the cleaning specifications (if new cases appear) and we updated the cleaners based on new cases appeared.

3. Experimental results

To test and evaluate the accuracy of the matching process and matching algorithm using Duke in practice, we performed some experiments on many files chosen from different municipalities and ministries since each file was different from the others in the variables and specification, where it helped us to test the algorithm accuracy.

3.1. Matching results

We used two ways to match the files. The first way was First Exact Match where the "census" file was kept in memory. Then we navigated one record at a time in the Municipality file. Our goal was to match records that contain similar values for selected variables. For each record, the matching stopped at the first matched Census record (which doesn't mean that it is the right one; but it means that we found at least one establishment in the Census that matches the record in the Municipality file for the selected variables). The number of "first exact matches" is therefore equal to the maximum number of establishments in the municipality file for which we can find a corresponding establishment in the Census file for the selected variables.

The second way was Multiple Exact Match. In this process, all matched records are kept. Our goal using this way was to find out which one among all the establishments (of the Census file) that matched with a given establishment (in the Municipality file), is the right (or the best) one.

Table 1 below shows the results of matching on exact matching on phone number only. We matched Ramallah municipality with the census file without cleaner and with cleaner. Cleaning steps improved the result of matching (from 505 to 2462 records with First Exact Match or from 555 to 2828 records with Multiple Exact Match).

Table 2 below shows the results of exact matching on commercial name and owner name. We matched Ramallah municipality file with census file without cleaner and with cleaner. The total number of matches

Table 1
Results of matching: exact matching on phone number only

Variable	Without cleaner		With cleaner	
Number of records matched on phone number only	First exact match 505	Multiple exact match 555	First exact match 2462	Multiple exact match 2828

Table 2
Results of matching: exact matching on commercial name and owner name

Variable	Without cleaner		With cleaner			
			Replace only key words		Complete cleaner	
	First match	Multiple match	First match	Multiple match	First match	Multiple match
Commercial name and owner name	1401	2074	1526	2234	1448	1813

Table 3
Results of matching: exact matching on commercial name, owner name and phone number

Variable	Matched on all variables	Matched at least one of variables	Matched on all variables	Matched at least one of variables
	Display the first match in case of multiple matches for one record		Display multiple matches	
Commercial name, owner name and phone number	772	2798	801	3452

Table 4
Detailed results of matching for selected cities

	Ramallah	Al Bireh	Bethlehem	Hebron	Birzeit
Census (number of establishments)	14678	3566	9345	11151	370
Municipality (number of establishments)	7747	2921	6374	6522	279
Multiple matches using Duke	Match at least one of variables				
Number of matches using the telephone without cleaning	514	28	970	800	1
Matching rates and number of matches using the developed phone cleaner	2789 (36%)	614 (21%)	2050 (32%)	1074 (16%)	83 (30%)
Matching rates and number of matches using the phone cleaner and the owner name and commercial name	3057 (39%)	902 3(1%)	3015 (47%)	1352 (21%)	119 (43%)

was different. It was bigger using "replace only key words" than "complete cleaner".

Table 3 above shows the results of exact matching on commercial name, owner name and phone number. We matched Ramallah municipality with census file with cleaners functions. The total number of matchings were different, where they were bigger using multiple matching based on at least one of the variables than using all variables.

Table 4 above shows that matching rates, using simultaneously the three identification variables, were the best possible matching rates that we could obtain (before checking that all the establishments that the algorithm has considered as duplicates were really the same). They are rather different from one municipality to another (21% for Hebron to 47% for Bethlehem) which indicates that the quality of the files was also probably different according to each municipality.

We also noted that the improvement in matching rates when going from the matching based only on phone numbers to the matching based on all the variables was very different: +10/15% for Al Bireh, Birzeit and Bethlehem and only +3/5% for Ramallah and Hebron.

The matching rate obtained with the cleaned phone number as an identification variable is shown in Table 5 below for the establishments with no phone number registered. The matching rates were extremely different for the different municipalities (ranging from 5% to 56%). The differences cannot be explained by phone ownership rates variability according to municipalities. The best matching rate (47% for Bethlehem) as shown in Table 4 was obtained in the municipality where the percentage of missing phone numbers is the best/lowest (only 5%) as shown in Table 5 whereas the worst matching rate (21% for Hebron) as shown in Table 4 was obtained where this percentage of missing phone numbers was the worst/highest (56%) as shown in Table 5.

Table 5
Establishments without registered phone numbers

	Ramallah	Al Bireh	Bethlehem	Hebron	Birzeit
Number of records without a phone number registered in municipalities files	1854	266	328	3620	36
Percent of phone numbers missing	24%	9%	5%	56%	13%

3.2. Analysis of results

Some establishments were considered as matched; whereas they shouldn't have matched because these establishments are in reality different. To help this problem, proposals have been made to improve the comparator. It is still a work in progress, which needs to be addressed by further work.

Other establishments were considered as unmatched although they should have matched as these establishments are in reality the same. In these situations, the following proposal for a "condition" in the comparator was made: If both commercial names contain at least three words and we have an exact match on commercial names, the two records are the same (even if the owner names are different). This condition was "too demanding" for finding possible matches. For the records from the municipality files which matched only with one record in the Census, the results were good, but not for the record which matched with more than one record in the Census. In order to make the study of the multiple matched establishments easier, a tool has been developed to extract the successful matches of all the multiple matched establishments from Duke. We studied these multiple matched establishments and tried to improve the specifications in order to find specifications aiming at reducing the number of multiple establishments that need a manual check. Thus, we tried to add an activity variable in the matching in order to reduce multiple match cases.

4. Conclusions and future work

This research aimed to show PCBS experiment in administrative data records linkage, where only a few establishments were going to match without having proper cleaning of the identification variables. For example, using the phone number as it is shown in Ramallah municipality files before cleaning, resulted in only 6% of the matched establishments. After cleaning the identification variables, the number of matches rises significantly. For example, after the cleaning of phone numbers in Ramallah municipality files (standardising their format by introducing the area codes, deleting non numerical character) the number of matched establishments rises to 36% even if 24% of the phone numbers are missing. It is crucial to get from the partners all the identification variables that are used to match establishments. Adding ID variables is conducive to raise the rate of matched establishments. For example, in Ramallah, the additional use of commercial and owner names (cleaned) allowed to reach almost 40% of the matched establishments and 47% in Bethlehem. The cleaning and the comparator cannot solve all possible orthographic mistakes/discrepancies, wrongly registered variables (example: owner name instead of commercial name), missing values, out of date data, registration format of the same variable is not standardized, etc. Improving the registration using similar formats is key to significantly improve the matching. In the short run and as the identification of data is not standardised, it is necessary that the municipalities (and the other partners) provide the following data of as many registered establishments as possible: TELEPHONE NUMBER(S), COMMERCIAL NAME, OWNER NAME, ACTIVITY, LOCATION DETAILS. In the long run, we would proceed in setting a shared list of identification variables and in standardizing ways of capturing the information in the registers, as it could ensure getting a good base for defining an Administrative Business Register ID.

References

[1] Palestinian Central Bureau of Statistics: http://www.pcbs.gov.ps.
[2] https://www.microsoft.com/en-us/download/details.aspx?id=15011.
[3] Lars M. (2013). Linking data without common identifiers, ISO 15926 and Semantics Conference, Sogndal.
[4] Mohammad N, David F, Ruslan S. (2014). Hamming Distance Metric Learning.
[5] Mark L. (2014). The stringdist Package for Approximate String Matching, The R Journal Vol. 6/1, June 2014, ISSN 2073-4859.
[6] Suphakit N, Jatsada S, Ekkachai N, Supachanun W. (2013). Using of Jaccard Coefficient for Keywords Similarity. Hong Kong, Proceedings of the International MultiConference of Engineers and Computer Scientists 2013, Vol I, IMECS 2013.
[7] Stephen P. (2007). Using Administrative Data for Statistical Purposes, the ICES-III, June 18–21, 2007, Montreal, Quebec, Canada.
[8] William C, Pradeep R, Stephen F. (2003). A Comparison of String Distance Metrics for Name-Matching Tasks.