Matching Techniques and Administrative Data Records Linkage

Haitham Zeidan*
Palestinian Central Bureau of Statistics (PCBS), Ramallah, Palestine – Haitham@pcbs.gov.ps

Abstract

The aim of this paper is to show 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 used in the experiment, We started our experiment by using The Fuzzy Lookup [2], it is an add-In for Excel was developed by Microsoft Research and performs fuzzy matching of textual data in Microsoft Excel, this tool use 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 used also Duke, see Lars M. (2013). [3] which is an existing and flexible deduplication (or entity resolution, or record linkage) engine written in Java. By using Duke engine we had written our matching algorithm and comparators to increase the matching results and matching accuracy, we had written also some data-cleaning functions for matching variables (Commercial Name, Owner Name, Telephone) in order to standardize each matching variable to get improved results. different matching algorithms used in the experiment such as Hamming Distance, e.g. Mohammad N. (2014) [4], Levenshtein distance, Mark P. (2014) [5], Jaccard Similarity, e.g. Suphakit N. et al. (2013) [6], exact match and multiple match.

The results showed that After cleaning the identification variables, the number of matches raises significantly, We also noted that the 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

Administrative records very important for official statistics instead of surveys to collect data for policy decisions also administrative records reduce the costs of data collection, increase the accuracy, for these reasons, administrative records are being used increasingly for statistical purposes, Stephen, P. (2007) [7], administrative records will help to build the business register that will contain different variables such as owner name, establishment name, telephones, address and other variables that will help to match different sources using matching techniques to build the final business register based on different sources from the agencies like ministries, municipalities, chambers and other sources, so this paper display PCBS Experience in matching different sources with census data.

1.1 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, William W. et al. (2003) [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 Hamming Distance

One of the earliest and most natural metrics is the hamming distance, e.g. Mohammad N. (2014) [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. In another way, 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.

1.3 Jaccard distance

A statistical way to measure similarities between sample sets. 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 $J(X, Y) = |X \cap Y |/|X \cup Y |$. 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 that the two sets have in common, the closer the Jaccard similarity will be to 1.0. , e.g. Suphakit N. et al. (2013) [6].

1.4 Edit (Levenshtein) Distance

Edit distance, Mark P. (2014) [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 find applications in natural language processing, where automatic spelling correction can determine candidate corrections for a misspelled word by selecting words from a dictionary that have 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. Objectives

The objectives of this study is to display the PCBS experiment in matching and administrative data records linkage, different data sources from different ministries and municipalities matched with PCBS Establishment Census 2012, Different matching algorithms, techniques and tools used in the experiment. PCBS intends to build an efficient statistical business register system that should serve its institutions' duties. So the objectives of matching process at the end: (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 serves all partners, (5) and Measuring the coverage of registered establishment comparing with establishment census 2012.

3. Methodology

3.1 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 look at reconciling the different data for the same establishments. The purpose of collecting files to match these files with establishment census 2012.

3.2 Matching Variables

In order to compare data and try to find out matching data, 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 M. (2013). Using it, it was easy to get useful results. The data, which we were working with, contain numerous columns, most of which were of no use whatever to find duplicates. For example, the internal identifier (called also primary key for each file) did not help as it was different in each file. only three columns had been used: **Telephone, Commercial Name and Owner Name**. We had

written using Duke for these columns some data-cleaning functions (phone numbers and Arabic text) in order to standardize each column so we can get improved results. We set the probability threshold (presently set to be 0,80) and define the two files to match and how we want to treat our two discriminating properties.

3.2 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, not all establishments had supplied one, phone numbers could correspond to different persons (local manager, owner). We provided a function to normalize the phone numbers, named Phone Cleaner, which allows cleaning up the registered data in each file. We had then specified our probabilities for the telephone: if one of the phone numbers present for an establishment registration in each file is the same, 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.

Commercial Name and Owner Name: These two columns were addressed with a single cleaner for several reasons after studying the files, sometimes data are not well filled-up; sometimes we have 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 suited 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 at the same time the commercial name and the owner name. Finally we have 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.

Telephone or Commercial Name and Owner Name: finally, to combine the phone analysis and the names analysis, we have two probabilities and we have 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 highest score reinforce the probability that we consider the two establishments as duplicates, unless we later find evidence indicating that the establishments are different.

3.2 Data Matching

We focused in matching process on duke, since duke can find duplicate records, also we can 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, geopositions, and more. Using a probabilistic model duke can handle noisy data with good accuracy. We made some matches exercises on some files from municipalities and ministries, and cleaning data provides in every cases good results; so cleaning data before matching 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. Whenever needed, we updated the cleaning specifications (if new cases appear) and we updated the cleaners based on new cases appeared.

4. 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, the files chosen from different municipalities

and ministries since each file different from others in the variables and specification, this will help us to test the algorithm accuracy.

4.1 Matching Results

We used to match the files two ways, the first way **First Exact match**: the "census" file is kept in memory, then we navigate one record at a time in the Municipality file. Our goal to match records that contain similar values for selected variables. For each record, the matching stops at the first matched Census record (which doesn't mean that it is the right one; but it means that we have 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 **Multiple Exact match**: same process but all matched records are kept. Our goal using this way to find among all the establishments (of the Census file) matched with a given establishment (in the Municipality file), which one is the right (or the best) one. Table (1) below shows the Results Matching (exact matching on phone number only), we matched ramallah municipality with census file without cleaner and with cleaner. Cleaning step improve the result matching (from 505 to 2462 records with First exact match or from 555 to 2828 records with Multiple exact match).

Table 1: Results Matching (exact matching on phone number only)

Variable	**Without cleaner**		**With cleaner**	
Number of records matched on Phone Number only	First Exact match	Multiple Exact match	First Exact match	Multiple Exact match
	505	555	2462	2828

Table (2) below shows the Results Matching (exact matching on commercial name and owner name), we matched Ramallah municipality file with census file without cleaner and with cleaner. the total matching are different, They are bigger using "replace only key words" than "complete cleaner".

Table 2: Results Matching (exact matching on commercial name and owner name)

	Without cleaner		**With cleaner**			
			Replace only key words		Complete cleaner	
	First match	Multiple match	First match	Multiple match	First match	Multiple match
Commercial Name & Owner Name	1401	2074	1526	2234	1448	1813

Table (3) below shows the Results Matching (exact matching on commercial name, owner name and phone number), we matched ramallah municipality with census file with cleaners functions. the total matching are different, They are bigger using multiple matching based on at least one of the variables than using all variables.

Table 3: Results Matching (exact matching on commercial name, owner name and phone number)

	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 & Phone Number	772	2798	801	3452

Table (4) below shows matching rates, using simultaneously the three identification variables, are the best possible matching rates that we could obtain (before checking that all the establishments that the algorithm has considered as duplicates are 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 is 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 is very different : +10/15% for Al Bireh, Birzeit and Bethlehem; only +3/5% for Ramallah and Hebron.

Table 4: Detailed results of the matching for several 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
Number of matches using the developed phone cleaner	2789 36%	614 21%	2050 32%	1074 16%	83 30%
Number of matches using the phone cleaner and the (owner name and commercial name)	3057 39%	902 31%	3015 47%	1352 21%	119 43%

The matching rate obtained with the cleaned phone number as an identification variable, table (5) below includes the establishments with no phone number registered. The number of records without phone number was as shown in table (5) below. those rates were extremely different between the different Municipalities (from 5% to 56 %), those 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: The establishments with no phone number registered

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

4.2 Results analysis

Some establishments are considered as matched whereas they shouldn't have matched as establishments are in reality different; Some 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 are

considered as unmatched although they should have matched as establishments are in reality the same; The following proposal for a "condition" in the comparator was made: if both commercial names contains 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 municipality files which match with only one record in the Census the results are 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 to extract from Duke successful matches all the multiple matched establishments has been developed. we studied of these multiple matched establishments and tried to improve the specifications in order to find specifications aiming at reduce the number of multiple establishments which need a manual check. So we tried to add activity variable in the matching in order to reduce multiple match cases

5. Conclusion and Future Work

This research aimed to show PCBS experiment in administrative data records linkage, without proper cleaning of the identification variables (improving the standardisation of the format in which they are registered in the file) only few establishments are going to match. For example, for Ramallah, using the phone number as it is in the files before cleaning, only 6% of the establishments matched. After cleaning the identification variables, the number of matches raises significantly. For example, for Ramallah, after the cleaning of phone numbers (standardising their format by introducing the area codes, deleting non numerical character) the number of matched establishments raises 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 establishments matched. For example, in Ramallah, the additional use of Commercial and Owner names (cleaned) allowed to reach almost 40% of the establishments matched and 47% in Bethlehem. The cleaning and the comparator can't solve all possible mistakes/discrepancies orthographic ones, variables wrongly registered (example: owner name instead of commercial name), missing values, data not up to date, format of registration of the same variable not standardized and so on. Improving the registration, using similar formats is key to improve significantly the matching. In the short run, as the identification data are not standardised, it is necessary that the Municipalities (and the other partners) provide the following data for as many as possible registered establishments: TELEPHONE NUMBER(S), COMMERCIAL NAME, OWNER NAME, ACTIVITY, LOCATION DETAILS. In a longer run, we would gain in setting a shared list of identification variables and in standardizing ways of capturing the information in the registers. It could give a good base for defining in common 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 J., & Ruslan S. (2014). Hamming Distance Metric Learning.

[5] Mark, P. (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 W., Pradeep R., & Stephen E. (2003). A Comparison of String Distance Metrics for Name-Matching Tasks.