2020 Asia–Pacific Statistics Week
A decade of action for the 2030 Agenda: Statistics that leaves no one and nowhere behind
15-19 JUNE 2020 | Bangkok, Thailand

Probabilistic Record Linkage: An Innovative Method to Improve the Quality of Data Integration (Case study in Iran)

Saeed Fayyaz[1]; Reza Hadizadeh[2]

[1] Statistician on Labour Force Statistics, Statistical Center of Iran
[2] Leader group of PPI, Statistical Center of Iran

Abstract:
with respect to the increasing use and availability of routinely collected 'Big data', changing the National Statistics Offices'(NSOs) approaches toward using alternative data sources and administrative registers instead of conventional surveys and censuses, it is becoming more useful to undertake researches that involves data integration and linking datasets from multiple sources mostly do not have enough convincing harmony and consistency. These sorts of problems; however, have resulted to ineffective record linkage and poor data matching which might be as a result of dissimilar alphabetic, wrong data entry methods and inconsistent standards in datasets. Frequently, many of big datasets and administrative registers have been created to meet the official needs instead of statistical purposes and emerging inconsistencies are inevitable. As a result, a significant number of them have no statistically acceptable standard and lack of either primary or secondary keys for inner joining and integrations process. Furthermore, in the lack of unique numeric key, text matching, especially in none English language (Persian) will be problematic as well. In such these cases, record linkage might be demanding and needs a great number of technical, expertise, time, energy and budget to deal with unpredictable and probable errors and mismatching difficulties. According to above mentioned, in this paper, after reviewing the record linkage methods and their efficiencies, an innovative text mining method based on asci code proposed to improve the linkage quality in these cases and sample datasets of Household Expenditure Survey of Iran is used to compare the outcomes. Obtained results showed that the proposed method could improve the linkage probability in corresponded case of Iran and increase the integration quality of datasets in comparison with conventional integration. This method can also be helpful not only for all of the National Statistical System (NSS) components but also for other countries and their NSOs in order to be utilized effectively and take better advantages from available administrative, big data and open access data sources.

Keywords: record linkage, primary and secondary key, probabilistic matching, alphabetic disorder, administrative registers, various data sets.

1. Introduction:

With the increasing use and availability of routinely collected 'big' data, it is becoming more useful to undertake research that involves linking data from multiple sources (Adrian Sayers et all, 2015). Record linkage is also referred to as data cleaning or object identification. It gives background on how record linkage has been applied in matching lists of businesses. It points out directions of research for improving the linkage methods (William E. Winkler, 2006). There are different methods for making links in data sets. enhancements to a record linkage methodology that employ string comparators for dealing with strings that do not agree character-by-character, an enhanced methodology for dealing with differing, simultaneous agreements and disagreements between matching variables associated with pairs of records, and a new assignment algorithm for forcing 1-1 matching (William E. Winkler, 2015). In other study, two main existing approaches for record linkage were compared: probabilistic and distance-based. The performance of both approaches are compared when data are categorical. To that end, a distance over ordinal and nominal scales are defined. The paper shows that, for categorical data, distance-based and probabilistic-based record linkage lead to similar results. (Josep Domingo-Ferrer et all, 2004). Also a study was done to assess the quality of your linkage algorithm, and how epidemiologists can maximize the value of their record-linked research using robust record linkage methods (Adrian Sayers et all,2016).

2. Methodology:

In other to make a record linkage it would be necessary to follow a series of steps. It is worth mentioning that there are different approaches to make record linkage in many countries but in this study following steps were considered with the help of software including SQL Server, SAS and Excel.

STEP 1 Standardization	•File formats •Data and Variable format •Variable definitions and their attributes •Para Data

1) **File formats:** As far as many data has been produced in administrative registers, surveys and all data providers based on their interest, so in it would be not varied to face with different data formats. Basically, it is important to harmonize the format of data in most common dataset formats including but not limited to SQL Server, Excel, Access.
*****In this study tow considered files for make linkage are in Excel format.

2) **Data and Variable Format:** The number of columns, rows and the structure of data for linkage is very important. Different data formats are developed based on local goals to pursued the data owners need and not for statistical purposes. In order to make a link between to different data sets in would be necessary to harmonize the whole data structures in them. In the other side, the variable format plays a vital role in linkage process. There are different types of formats for different variables including but not limited to numeric, text, integer number, percent, date, time etc. At this stage, the same variable should be the same formats in different data sets. It would be better that change the names of the same variable in data sets.
*****In this study all the variables harmonized in the same comparable and joinable formats. The data from household Cost and Expenditure for two different years and different family types has been considered as a case study. This might be an interesting area to focus on innovative methods to harmonize in this type of differences. The focus in this study was on innovative way of linking in the circumstance that assume data and variable formats had been harmonized before.

3) **Variable definitions and their attributes:** Tanking definitions and the coverage of each variable as well as its attributes in data producer/ provider as a key item, it would be necessary to control these matters before data record linkage. It is frequently seen that even the same variable in different data sets have different definitions or cover different target populations and attributes based on their owners' interest and usage.
* In this study, based on the data set selection criteria from a same surveys and control the variable definitions, coverage and other attributes, the number of inconsistencies was close to zero. But it would be very crucial step to control this items before making a record linkage, otherwise your join and linkage would be biased deliberately.

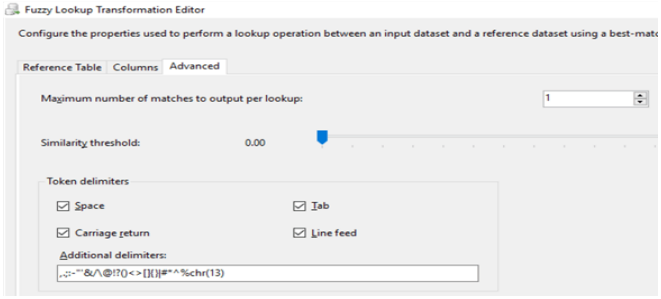
4) **Para data:** The Para data of a data set or survey are data about the process by which the data were collected. Para data of a survey are usually "administrative data about the survey. It is highly important to control the para data in order to find more about the variable, methods that data is collected and how many edit and computation are applied on data sets. This is helpful for efficient linkage between different data sets.

STEP 2 Purification	•Insure that no strange value/ charachter is on dataset

Before make any kind of record linkage it would be necessary to clean the data set to prepare a linkage. This might be removing extra characters including but not limited to below script and either replace or remove them. After removing this extra character, the linkage can be possible.

Table1. Function for Removing the Extra Character in SQL Server

```
BEGIN
DECLARE @ij NVARCHAR (50)
set @ij= REPLACE (@i, ' ', ''),',',''),'@',''),'-',''), '.',''),'-
',''),'*',''),'/',''),'\',''),'+',''),':',''),'=',''),
'!',''),'~',''),';',''),':',''),')',''),'(',''),'&',''),'^',''),'%','')
,'$',''),'#',''),'?',''),'>',''),'<',''),
'[',''),']',''),'{',''),'}',''),'|','')
RETURN @ij
END
```



This characters had been selected based on most common characters in different databases.

STEP 3
Record Linkage

- The common linkaage tools
- The innovative approach

In order to make linkage two circumstances were considered. In this study 3 ways were applied for inner- join of two data sets includes SQL Server (Table1), SAS (Table2) and Excel fuzzy lookup tools (Fig). The results in Excel method should be controlled manually by statisticians and its highly demanding effort and for each record statisticians must make a decision based on what they see in two files and similarity rate.

Table2. SQL Server Script for linkage

```
SELECT column name(s)
FROM table1
INNER JOIN table2
ON table1.column_name = table2. (Primary key/ Secondary Key/
Name etc.);
```

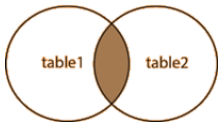
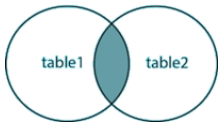


Table3. SAS Script for linkage

```
*Sort Tables*
PROC SORT DATA=table 1 OUT=table1S; BY variable; RUN;
PROC SORT DATA=table2 OUT=table2S; BY variable 2; RUN;
* Merge Tables*
DATA Table 3;
MERGE Table1S Table 2S;
BY Variable (Primary key/ Secondary Key/ Name etc.);
RUN;
```



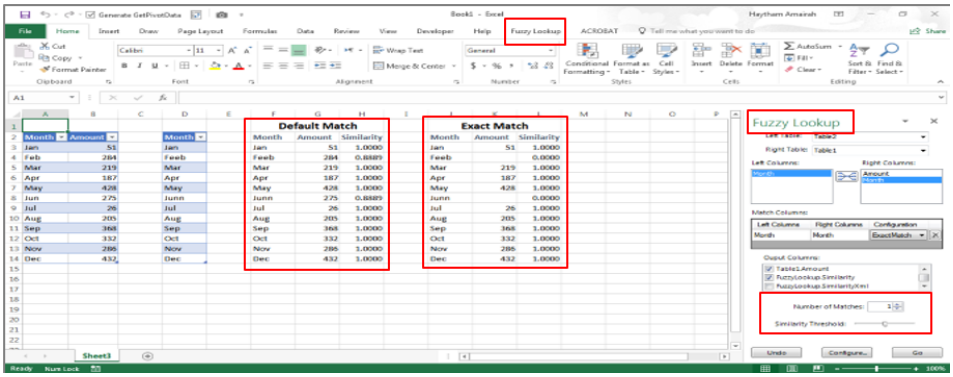


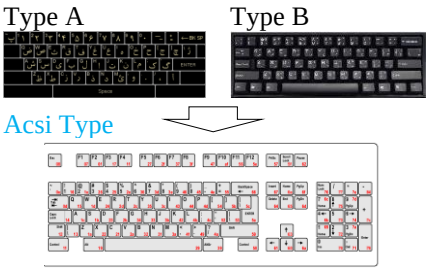
Fig1. Using Fuzzy Lookup Tools in Excel

Innovative approach

Most of computers keyboard are produced based on English alphabetic. In many languages there are some alphabetic that are same in meaning but different in writing and this may cause differences in text variable. In this innovative approaches all the characters in keyboard change to the ***ASCI*** *code*

Table4. SQL Server function for changing the characters asci codes

```
Create FUNCTION [dbo].[NameToString](@i NVARCHAR(50)) ; RETURNS
VARCHAR(max)
BEGIN
DECLARE @L int ; set @L=len (@i)
DECLARE @cnt INT = 1; DECLARE @asc VARCHAR(max); set @asc=' '
WHILE @cnt <= @L
BEGIN
  set @asc=replace(@asc+STR(asci(substring(@i,@cnt,1))) ,' ','')
  SET @cnt = @cnt + 1; END;RETURN @asc; END
```



Name	*Asci Code*	*Name*	*Asci Code*
ساویز	211199230237210	ساهره	211199229209229
ساویس	211199230237211	سایا	211199237199
ساوین	211199230237228	ساهر	211199229209

Probabilistic, deterministic, and clerical techniques may be combined in different ways depending upon the goal of the record linkage project. If a population parameter is being estimated for a purely statistical study, a completely probabilistic approach may be most efficient; for other applications, where the purpose is to make inferences about specific individuals based upon their data contained in two or more files, the need for a high positive predictive value would favour a deterministic method or a probabilistic method with careful clerical review. In order to make linkage, always primary and secondary keys are used that frequently exist but in many situations these numbers in not available. (DE Clark, 2004)

1) Primary and secondary keys were available

In this study, as mentioned before 1000 records of individuals' information had been selected as sample of households for 2 different implementations. In this data set *Personal Identification Number (PIN)*, which is a 10-digit unique number for each residence in Iran, and the *birth number* as were selected as primary and secondary number respectively. It is noticeable that the repetitive and incorrect INN had been removed, edited or control before the linkage.

**In both SQL Server and SAS script record linkage made by PIN as a Primary key (PIN= 10digit number) and with Fuzzy lookup similarity threshold 100% and number of matches was1.

If PIN number change to asci code, no significant difference occurred. The number of records that exactly match in the 3 methods are similar. It is shown that when correct unique Primary key in on offer in two data set there is no difference between the linkage method. It should be emphasized that the pervious requirement in pervious steps had been controlled before linkage.

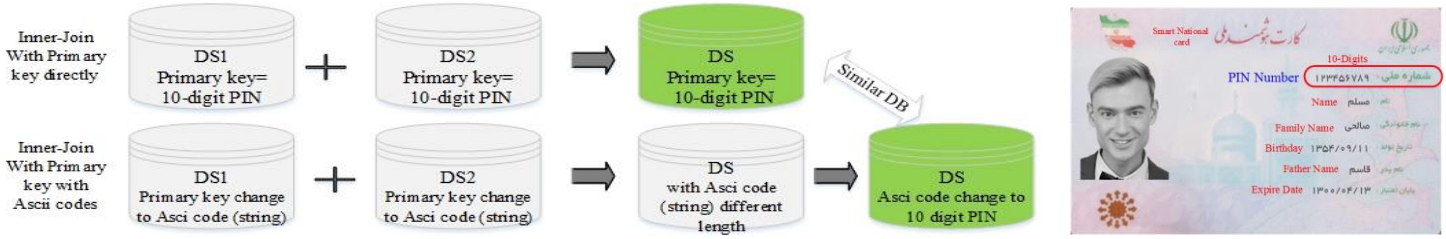


Fig2. Inner join of two databases when primary key is available

2) There were no Primary and secondary keys

In many cases, there are data sets with no numeric primary or secondary keys because of different causes confidentiality to different purposes of data producers. In Iran for example, although all the residences receive PIN at the birthday but there is no unique Business Identification Number (BIN) for enterprises or the Postal 10-digit code has been reminded under the coverage for many areas and Location Identification Number (LIN) is not available in many administrative or surveys data sets. In this case other variables on offer for making linkage between different data sets. These variables frequently are but not confined to Name, Family name,

Address, Age, Gender, Father's Name, Educational level, Occupation type etc. which are vary from one dataset to another. In this situation the number of linkage vary and it is highly related to the linkage method.

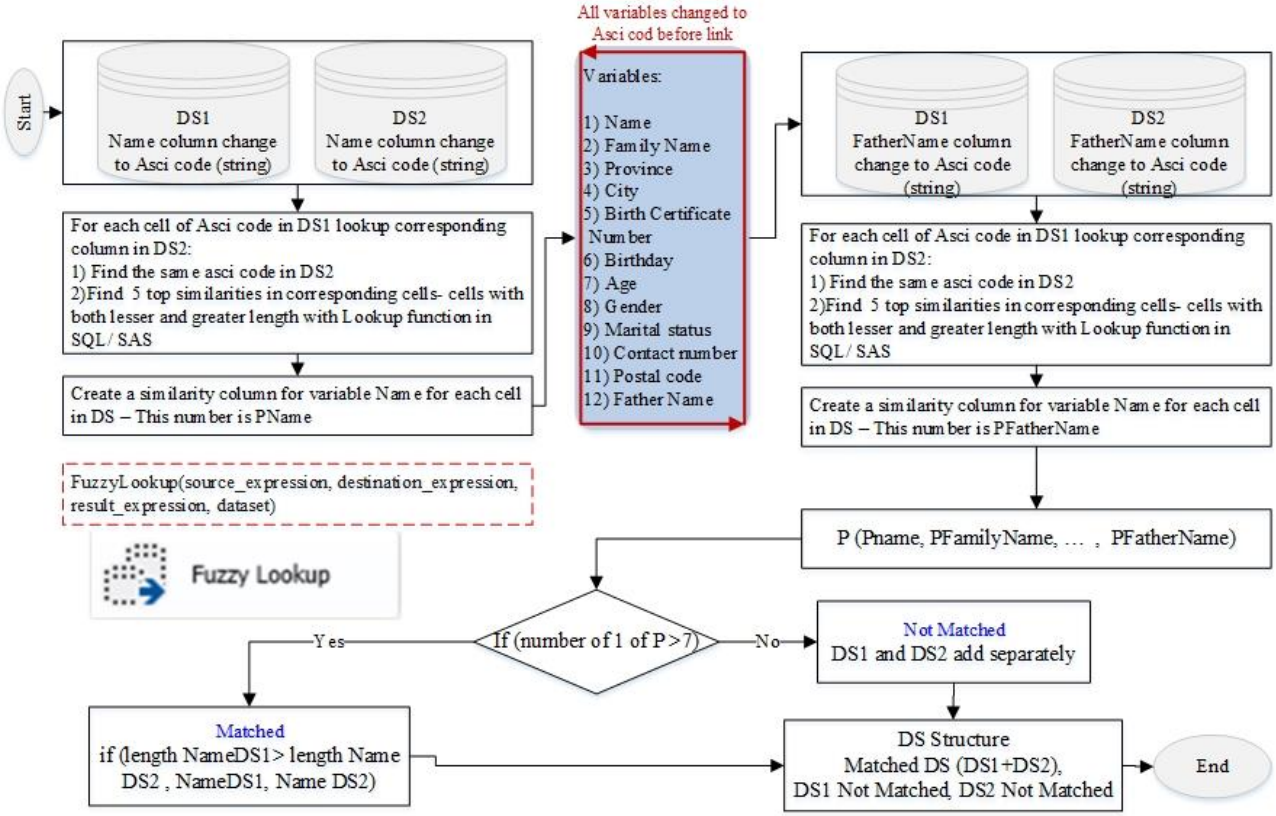


Fig3. Proposed algorithm for application of Asci code for record linkage

In proposed algorithm the number lookup of each record in database1 in dataset2 was considered to 5 records. This a trade-off between the number of similarities and the volume of the data set and how much time would be need to process the algorithm. Furthermore, 12 variables were considered in this case study that might be vary in other cases based on the type of data sets are going to be linked. Also, the threshold of 5 for each decision making of totally matched record in a compromising number. Taking this number, the bigger threshold is considered, the less likely to be matched the records.

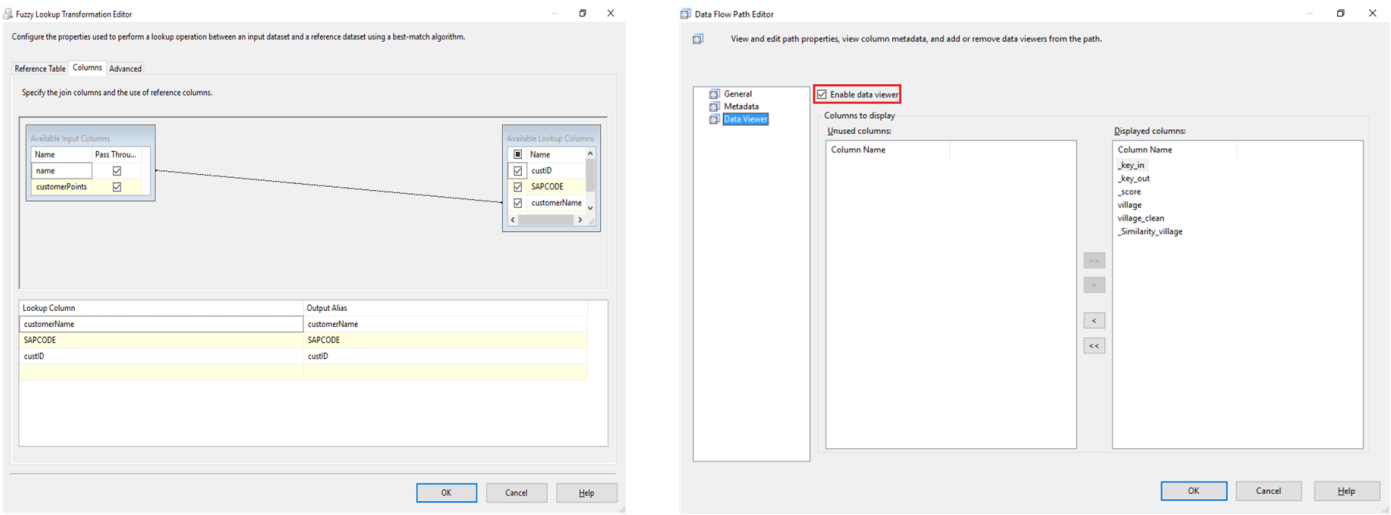
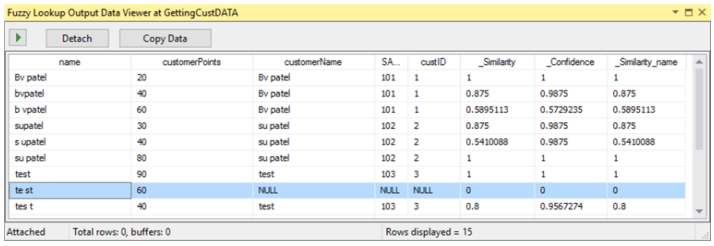


Fig4. Fuzzy lookup transformation Editor



name	customerPoints	customerName	SA...	custID	_Similarity	_Confidence	_Similarity_name
Bv patel	20	Bv patel	101	1	1	1	1
bvpatel	40	Bv patel	101	1	0.875	0.9875	0.875
b vpatel	60	Bv patel	101	1	0.5895113	0.5729235	0.5895113
supatel	30	su patel	102	2	0.875	0.9875	0.875
s upatel	40	su patel	102	2	0.5410088	0.9875	0.5410088
su patel	80	su patel	102	2	1	1	1
test	90	test	103	3	1	1	1
te st	60	NULL	NULL	NULL	0	0	0
tes t	40	test	103	3	0.8	0.9567274	0.8

Fig5. Output for considered data

3. Discussion and Conclusion:

In connection with record linkage algorithm, a sample of two data set from Housing expenditure-income survey with 75% similarity was selected (From 1000 sample records, only 648 have PIN and 513 records are existing in two data sets. In these data set the PIN in not mandatory for family members and not all the records have PIN. Also, in order to assess the efficiency of proposed approach, the linkage was applied in two different circumstances: with and without PIN as primary key.

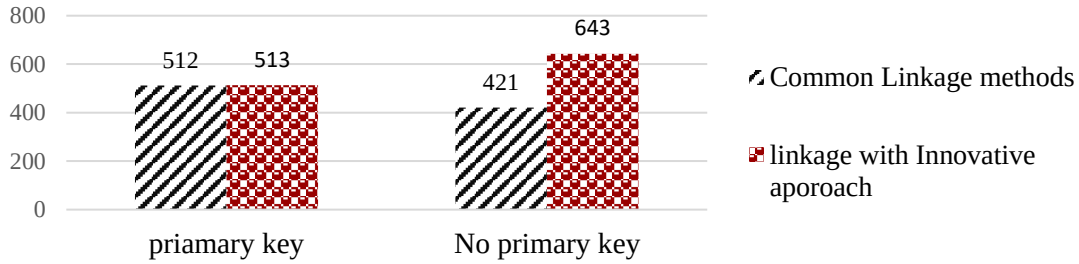


Fig6. the final result of two different methods of record linkage

In situation in which PIN number are available in two datasets, it is assumed that all PIN are correct and unique, the number of linked records when PIN applied directly as a primary key or when PIN changed to Asci codes are the same. It means that PIN and Asci code are similar effect on linkage process and this is because they have similar attributes and function here like numeracy, being unique and there is only one individual key for each number on keyboard. But also there is slight difference in the structure of final data set. In common linkage record, for example SQL/SAS programming, it is important that which data set mentioned first, because in matched record the information of first data set is considered as the basic and vacant cell will be filled by corresponding cell in other data set while in proposed algorithm the basic cell is determined by the max length of cells of both data sets resulting to more detailed information in final cell in final data set.

In the lack of PIN as primary key, the results significantly different in both common and innovative approaches. In common linkage record however; only 421 records from 684 (61.55%) and in innovative approach 643 from 684 (94.1%) was linked successfully that it emphasized on the superiority of innovative approach.

This study can be approached whenever the primary key for inter connection between two or more data sets will miss or has low quality.

References:

Adrian Sayers, Yoav Ben-Shlomo, AshleyW Blom, Fiona Steele (2016),' Probabilistic record linkage', doi: 10.1093/ije/dyv322, International Journal of Epidemiology, 2016, 954–964.

DE Clark 92004),' Practical introduction to record linkage for injury research', first published as 10.1136/ip.2003.004580 on 3 June 2004.

Josep Domingo-Ferrer, Vicenç Torra (2004),' Distance-based and probabilistic record linkage for re-identification of records with categorical variables', supported by the European Commission under project IST-2000-25069 "CASC".

Pradeep Ravikumar, William Cohen,' A Hierarchical Graphical Model for record Linkage', UAI. p 454-461.

William E. Winkler (2015), 'Advanced Methods for Record Linkage', Advanced Methods for Record Linkage.

William E. Winkler (2006),' Overview of Record Linkage and Current Research Directions', Statistical Research Division U.S. Census Bureau, Washington, DC 20233. Research Report Series.