Virtual Event 15-18 June 2020
2020 Asia-Pacific Statistics Week
Leaving no one and nowhere behind

Statistical Center of Iran

Probabilistic Record Linkage: An Innovative Method to Improve the Quality of Data Integration

(Case study in Iran)

Action Area D. Modernizing statistical business processes (SD1)
Surveys, data management and uses

Presenter:
Saeed Fayyaz

Statistician on Labour Force Statistics
Iran's National Statistical Office

#apstatsweek2020

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

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

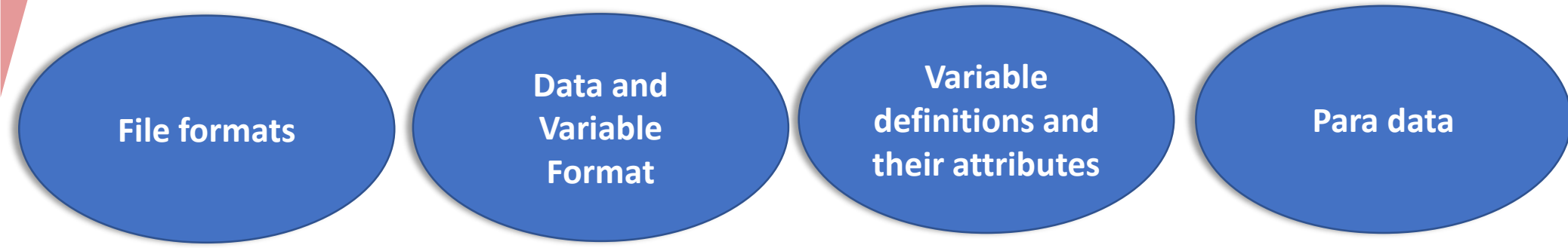
In other study, two main existing approaches for record linkage were compared: probabilistic and distance-based. The performance of both approaches are compared when data are categorical. To that end, a distance over ordinal and nominal scales are defined. The paper shows that, for categorical data, 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).

Methodology

STEP 1
Standardization

- File formats
- Data and Variable format
- Variable definitions and their attributes
- Para Data

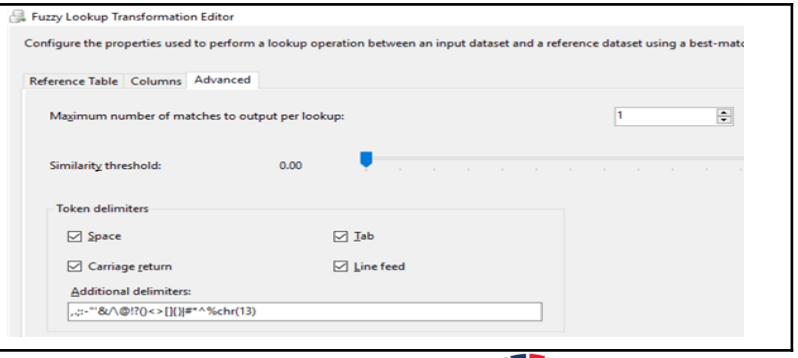


STEP 2
Purification

- Insure that no strange value/ charachter is on dataset

Function for Removing the Extra Character in SQL Server

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



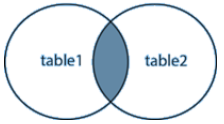
Introduction

STEP 3
Record Linkage

- The common linkaage tools
- The innovative approach

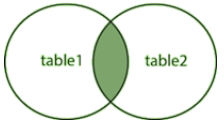
```
*Sort Tables*
PROC SORT DATA=table 1 OUT=table1S; BY variable; RUN;
PROC SORT DATA=table2 OUT=table2S; BY variable 2; RUN;
* Merge Tables*
DATA Table 3;
MERGE Table1S Table 2S;
BY Variable (Primary key/ Secondary Key/ Name etc.);
RUN;
```

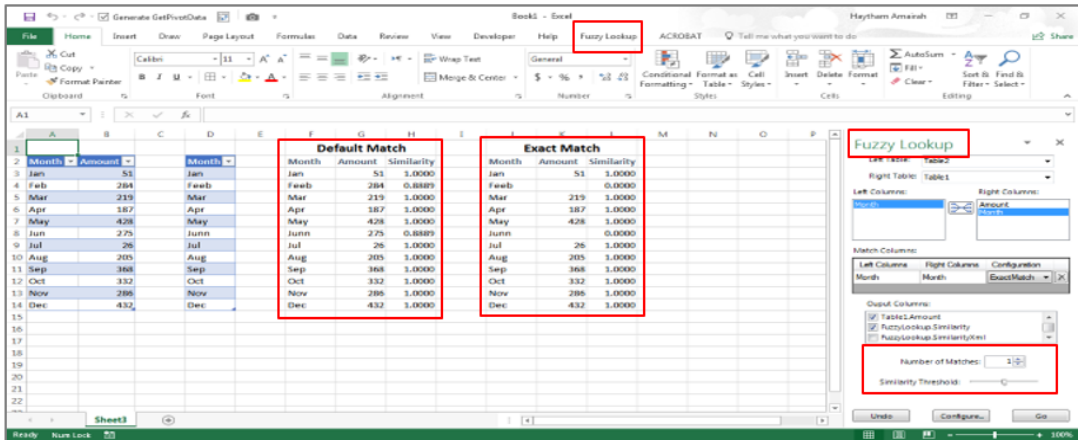
SQL Server Script for linkage




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

SAS Script for linkage





Using Fuzzy Lookup Tools in Excel

Introduction

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
```

Type A

Type B

Acsi Type

Name	Asci Code	Name	Asci Code
ساويز	211199230237210	ساهره	211199229209229
ساويس	211199230237211	سايا	211199237199
ساوين	211199230237228	ساهر	211199229209

main benefits of using the new method

TEXT variable

Convert to ASCI code
Remove doubles

ى
ي
ئ

371

Unique String / number variable

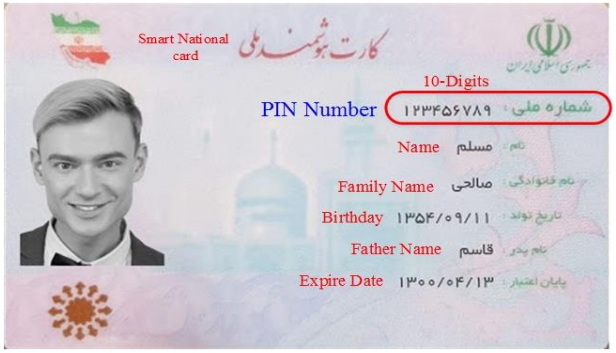
DS1 25%

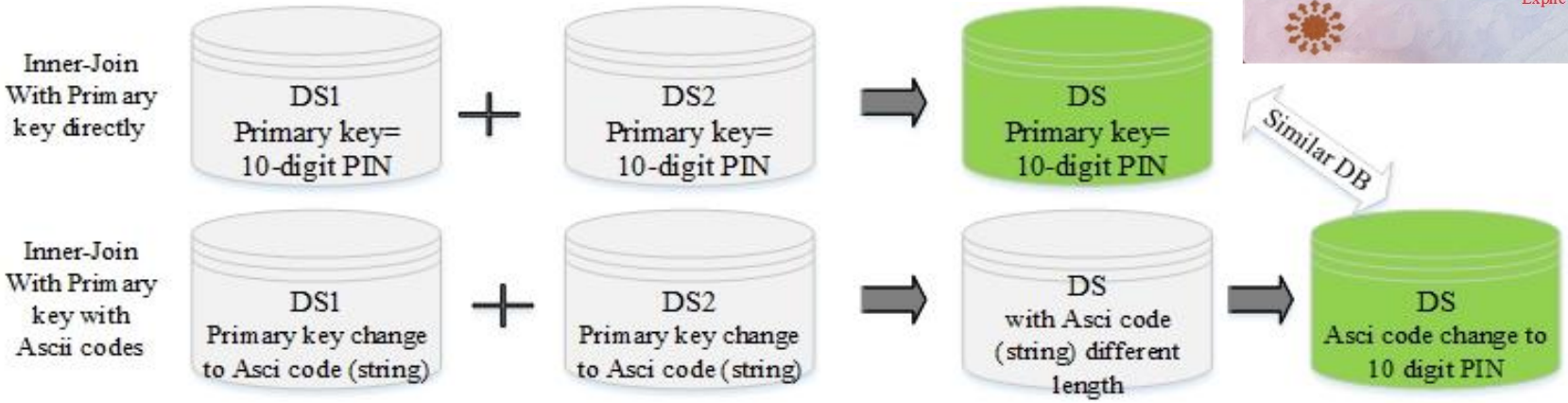
Household income and Expenditure Survey (HIES)
1000 Sample

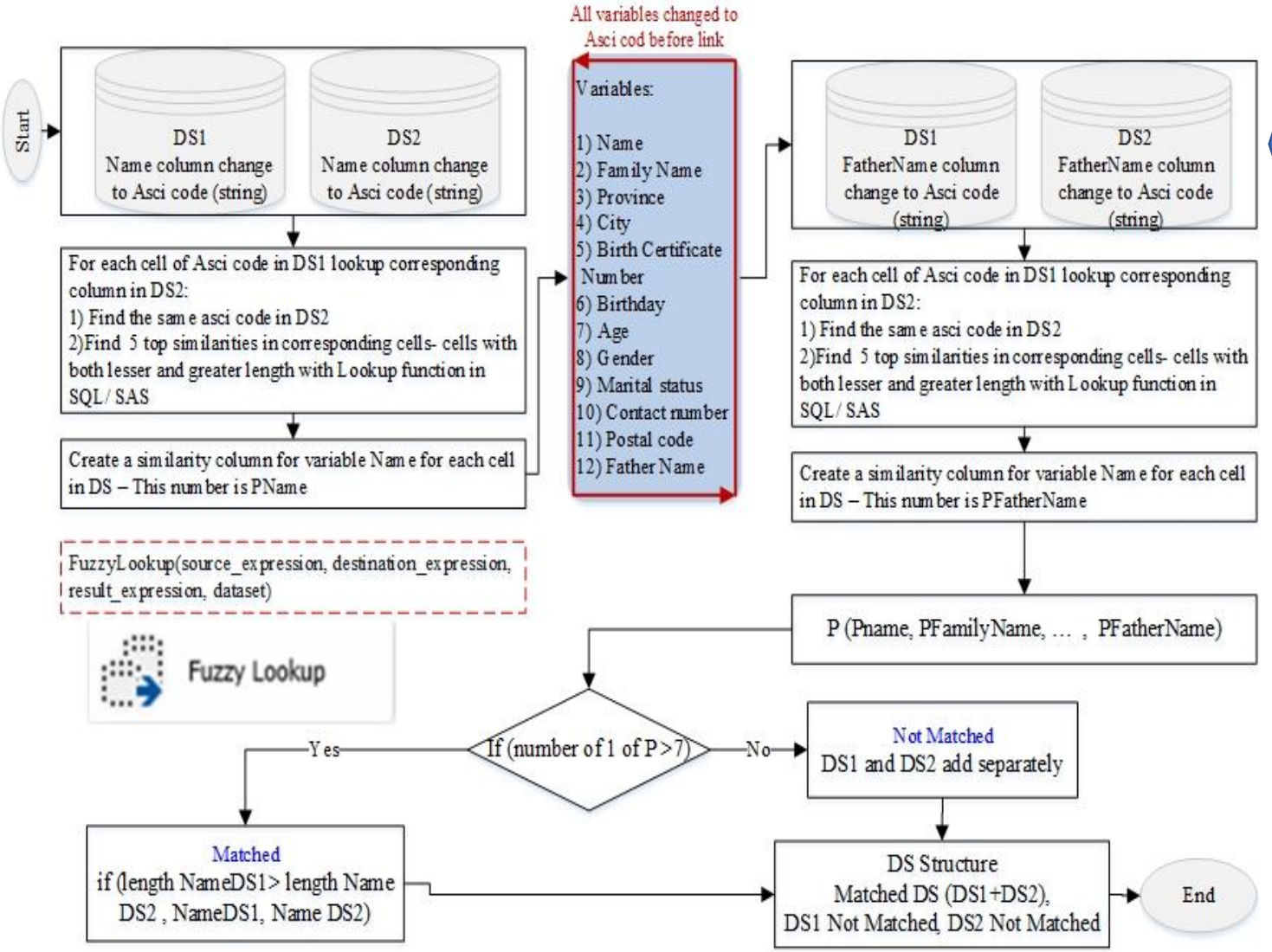
DS2 25%

Inner join of two databases when **primary key is available**

perfect match

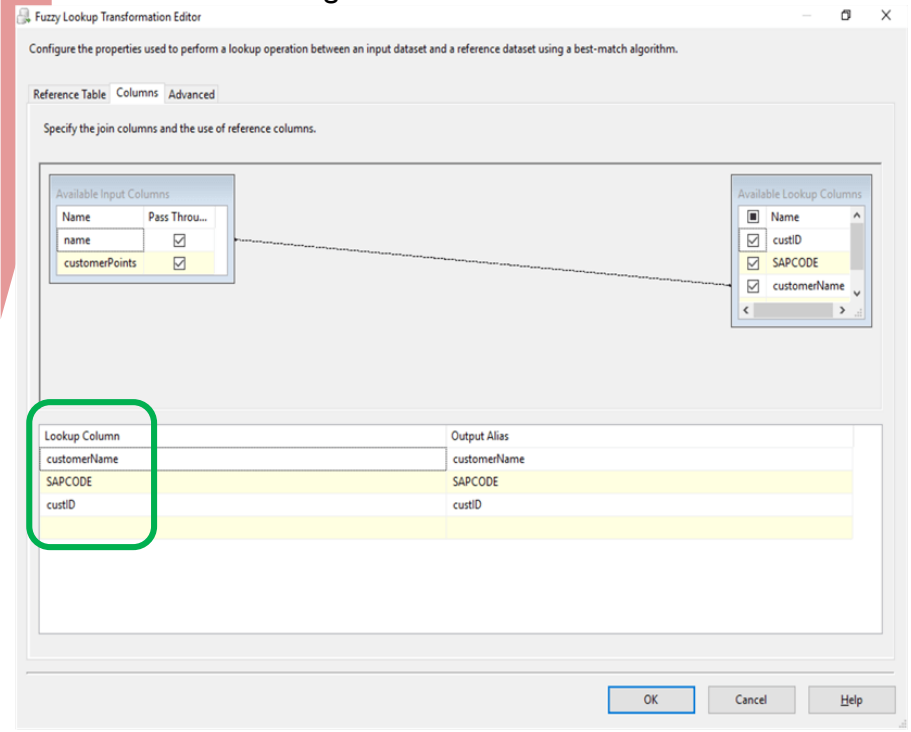


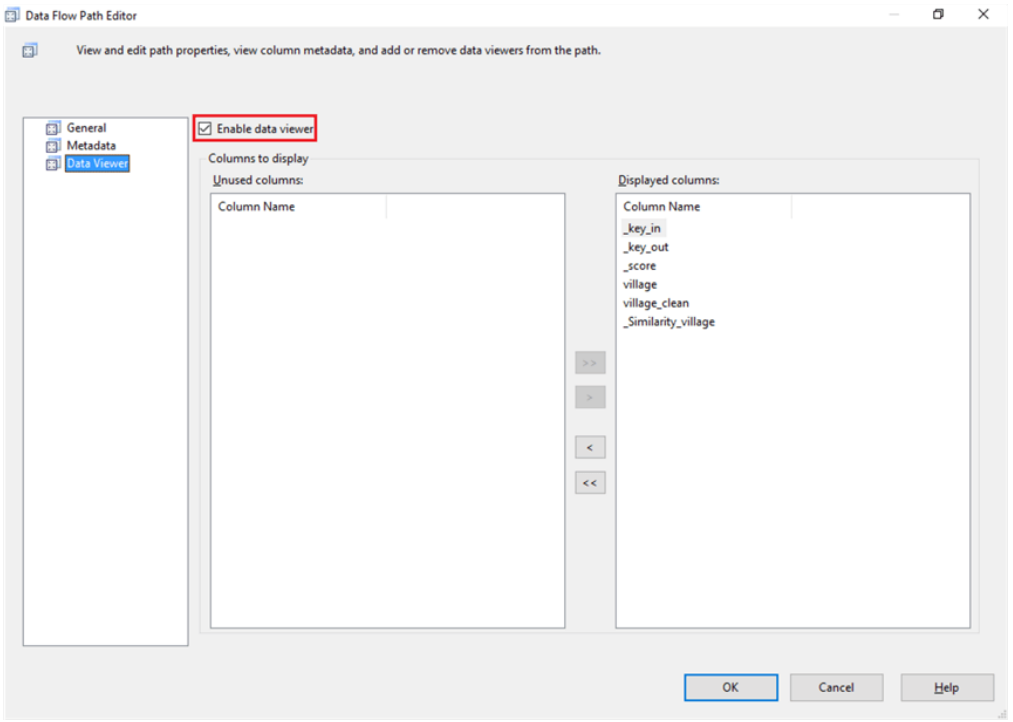


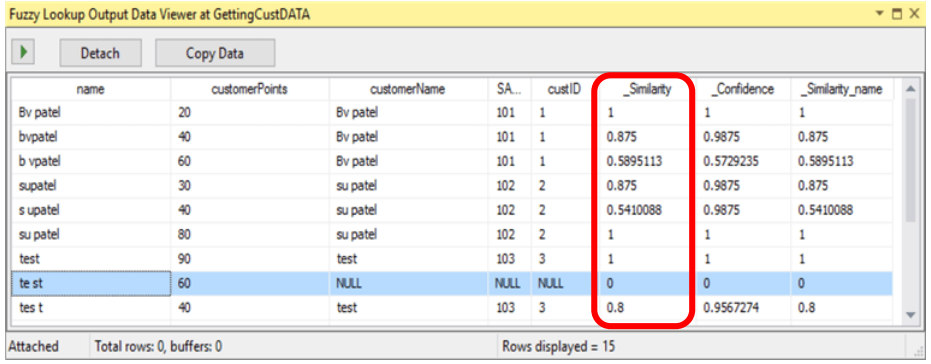


Proposed algorithm for application of Asci code for record linkage

Fuzzy lookup transformation Editor

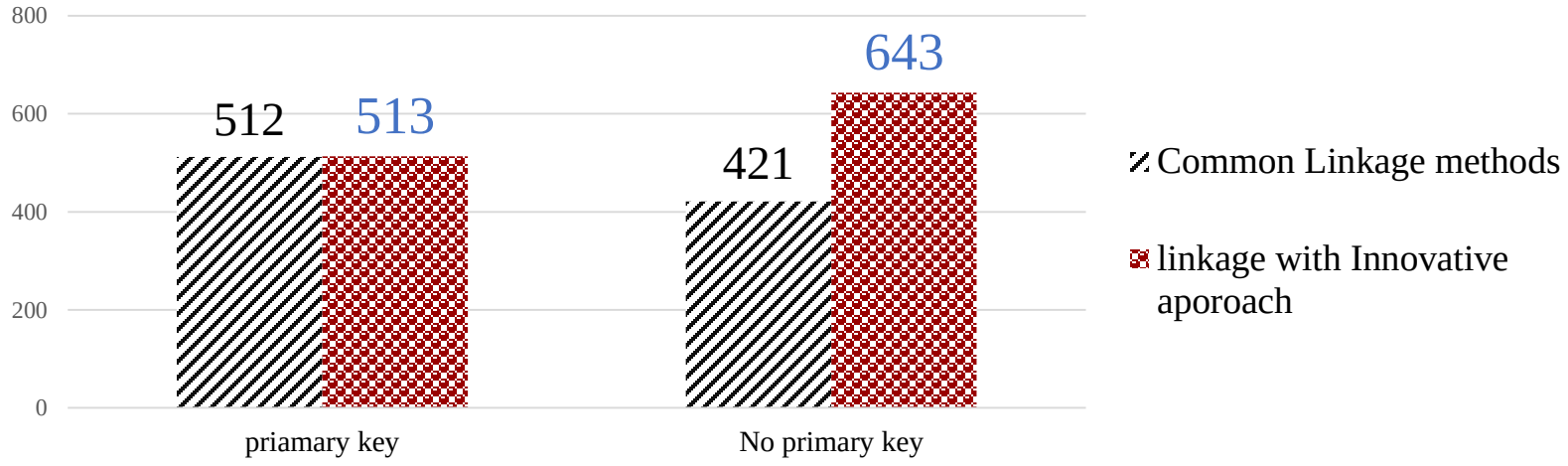






Output for considered data

Conclusion



the final result of two different methods of record linkage

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 method can applied to other cases in other languages as well

Introduction

Iran, DAMAVAND

Thank You For your attentions

✉ Saeed.Fayyaz@Gmail.com
☎ +98 (21) 85102299
+98 (21) 88970458
+98 (912) 6705085