



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



#apstatsweek2020



Virtual Event 15-18 June 2020

## 2020 Asia-Pacific Statistics Week

Leaving no one and nowhere behind

# 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 al, 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 al, 2016](#)).







Leaving no one and nowhere behind

# Introduction

**STEP 3**  
**Record Linkage**

- The common linkage tools
- The innovative approach

**\*Sort Tables\***

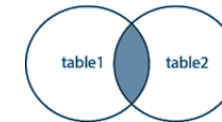
PROC SORT DATA=table 1 OUT=table1S; BY variable; RUN;  
 PROC SORT DATA=table2 OUT=table2S; BY variable 2; RUN;

**\* Merge Tables\***

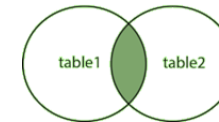
DATA Table 3;  
 MERGE Table1S Table 2S;  
 BY Variable (Primary key/ Secondary Key/ Name etc.);  
 RUN;

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

SQL Server Script for linkage



SAS Script for linkage



Using Fuzzy Lookup Tools in Excel

Default Match		Exact Match	
Month	Amount	Month	Amount
Jan	51	Jan	1,000.00
Feb	284	Feb	219
Mar	219	Mar	187
Apr	187	Apr	428
May	428	May	275
Jun	275	Jun	26
Jul	26	Jul	205
Aug	205	Aug	368
Sep	368	Sep	332
Oct	332	Oct	286
Nov	286	Nov	432
Dec	432	Dec	





Leaving no one and nowhere behind

## SQL Server function for changing the characters ascii 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(ascii(substring(@i,@cnt,1))),',' ,',')
  SET @cnt = @cnt + 1; END;RETURN @asc; END
  
```

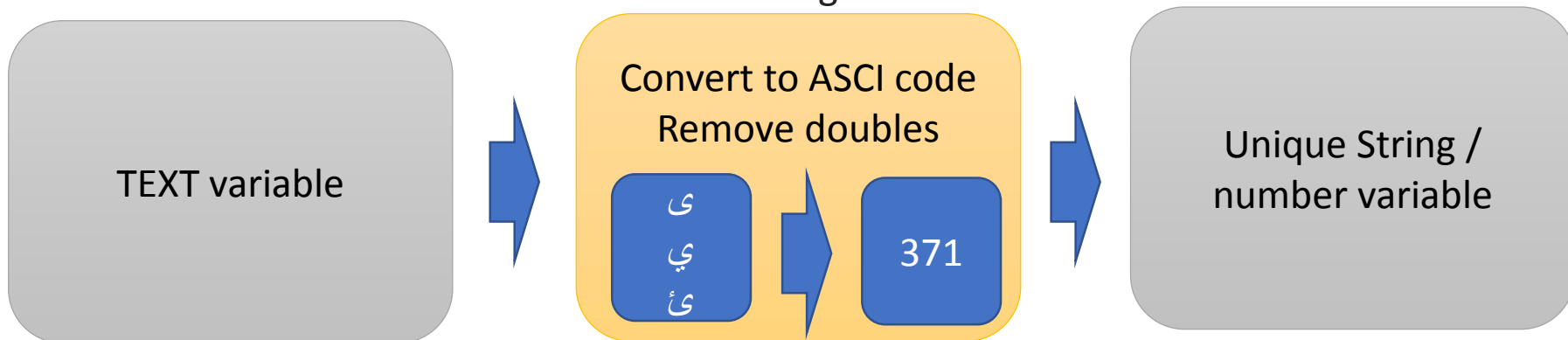
**Type A**

**Type B**

Aci Type

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

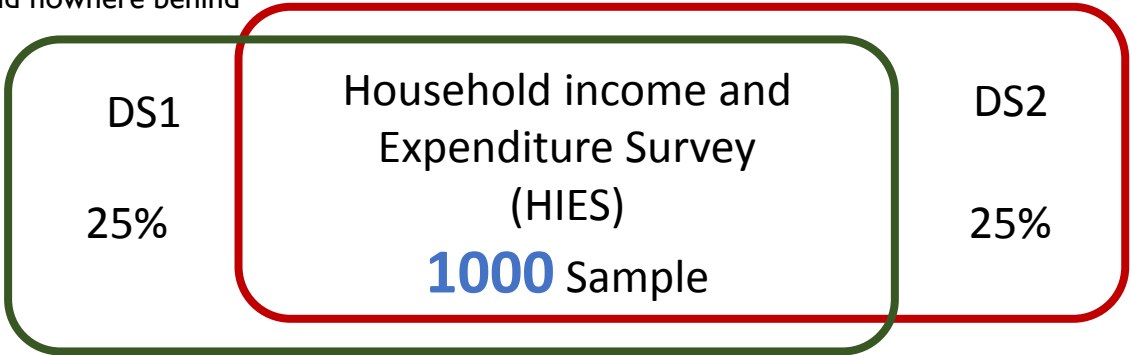
### main benefits of using the new method



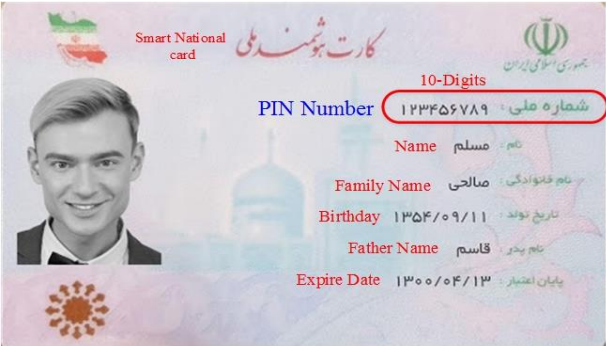
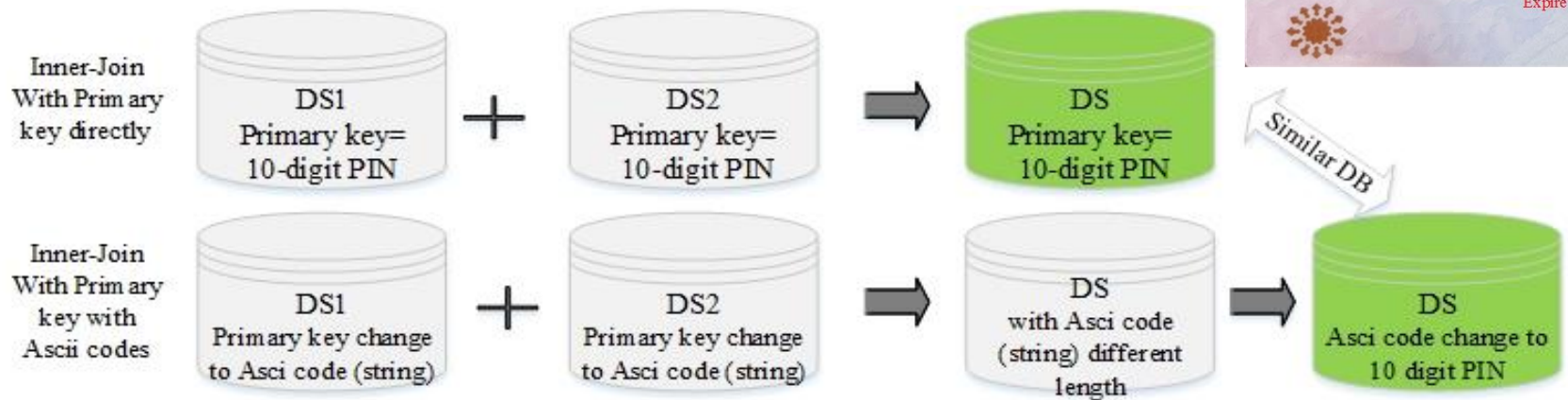


Virtual Event 15-18 June 2020  
**2020 Asia-Pacific  
 Statistics Week**

Leaving no one and nowhere behind



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



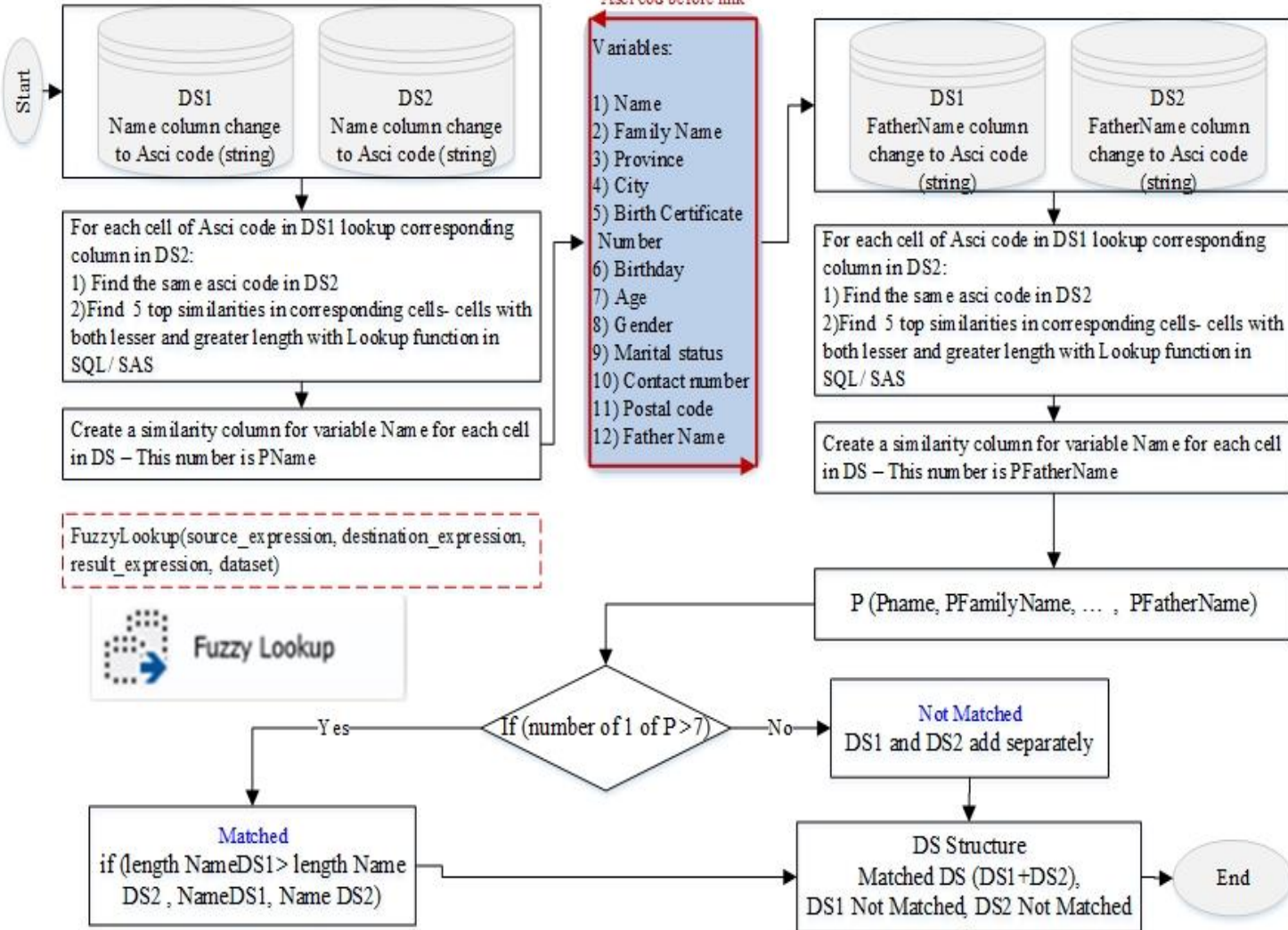


Virtual Event 15-18 June 2020  
**2020 Asia-Pacific Statistics Week**

Leaving no one and nowhere behind

All variables changed to Ascii cod before link

- Variables:
- 1) Name
  - 2) Family Name
  - 3) Province
  - 4) City
  - 5) Birth Certificate Number
  - 6) Birthday
  - 7) Age
  - 8) Gender
  - 9) Marital status
  - 10) Contact number
  - 11) Postal code
  - 12) Father Name



Proposed algorithm for application of Ascii code for record linkage



# Virtual Event 15-18 June 2020 2020 Asia-Pacific Statistics Week

Leaving no one and nowhere behind

## Fuzzy lookup transformation Editor

Fuzzy Lookup Transformation Editor

Configure the properties used to perform a lookup operation between an input dataset and a reference dataset using a best-match algorithm.

Reference Table: Columns Advanced

Specify the join columns and the use of reference columns.

Available Input Columns	Available Lookup Columns
Name	Pass Thru...
name	<input checked="" type="checkbox"/>
customerPoints	<input checked="" type="checkbox"/>

Available Lookup Columns
Name
<input checked="" type="checkbox"/> custID
<input checked="" type="checkbox"/> SAPCODE
<input checked="" type="checkbox"/> customerName

Lookup Column	Output Alias
customerName	customerName
SAPCODE	SAPCODE
custID	custID

OK Cancel Help

Data Flow Path Editor

View and edit path properties, view column metadata, and add or remove data viewers from the path.

General Metadata Data Viewer

Enable data viewer

Columns to display

Unused columns:

Column Name

Displayed columns:

Column Name
_key_in
_key_out
_score
village
village_clean
_Similarity_village

OK Cancel Help

Fuzzy Lookup Output Data Viewer at GettingCustDATA

Detach Copy Data

name	customerPoints	customerName	SA...	custID	<b>_Similarity</b>	_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

Attached Total rows: 0, buffers: 0 Rows displayed = 15

Output for considered data

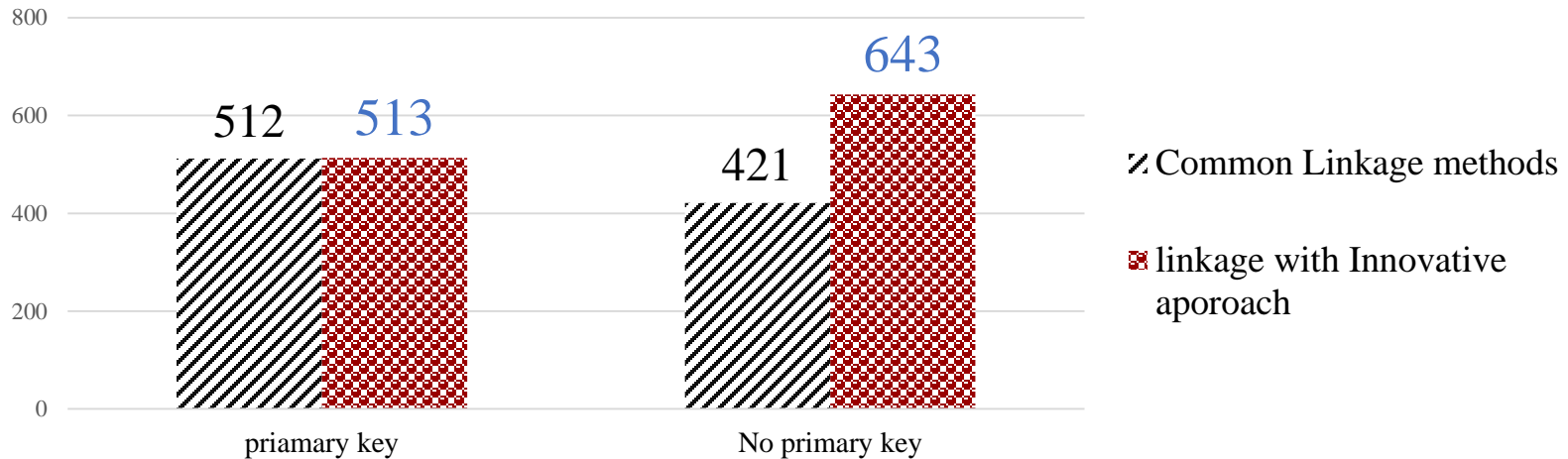






# Conclusion

Leaving no one and nowhere behind



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



Virtual Event 15-18 June 2020  
**2020 Asia-Pacific  
Statistics Week**

Leaving no one and nowhere behind

# Introduction



Iran, DAMAVAND

## Thank You For your attentions

✉ [Saeed.Fayyaz@Gmail.com](mailto:Saeed.Fayyaz@Gmail.com)

☎ +98 (21) 85102299

☎ +98 (21) 88970458

+98 (912) 6705085



#apstatsweek2020