Statistical Center
of Iran

Innovations from Iran: Resolving quality issues in the integration of administrative and big data in official statistics

Saeed Fayyaz & Reza Hadizadeh

Challenges?!

1. Data is collected for an **administrative purpose**, not a statistical purpose.

2. Errors can be compounded when there are **multiple characters in a native language**, such as Farsi, with the same meaning.

3. Different **incomparable coding system** for example the Customs' code and Statistical Office coding system Iran's Export Price Index (XPI) and Import Price Index (MPI)

4. Difficulties in **categorizing the description of merchandises** as recorded in the Custom organizations registers into internationally classified groups

Innovative Solutions

Application of ASCII code for probablistic record linkage

Household income and Expenditure Survey

Application of Text Mining for categorizing and grouping

Import and Export Merchandise Price indices

Using machine coding algorithms

1

Using text mining to address coding challenges in administrative datasets

2

1

Differences arising from use of **non-English alphabetic** characters

Type of difference	Example in Farsi
Text with same alphabet but different characters	"كرمى", "كرمي"or "كرمئ"
Text with similarity but incorrect writing	"ک" with "گ" or "س" with "ش"
Text with more than one syllable missing some part	"محمدی پناه" with "محمدی"
Combination of above and more than one variable	"كرمى راد" with "كرمي"
Additional characters were written	ۀ ـٔ ـٕ – ۀ
Same person registered both in Farsi and English in different registers	"سعید" and "Saeed" they should change to the same ASCII code

American Standard Code for Information Interchange or ASCII

ASCII is a type of character encoding schemes. It has definitions for 128 characters which are represented by 7 bits and was originally developed from telegraphic codes.

Character encoding schemes work by converting text into a number. For example, in ASCII "A" is converted to 065 and "a" is converted to 097. The computer stores 065 and 097, not "A" or "a".
Specifying the character encoding scheme is very important as without it, a machine could interpret given bytes as a different character than intended.

Farsi keyboard

ASCII keyboard

Convert and Correct

Name	ASCII Code	Name	ASCII Code
ساویز	211199230237210	ساهره	211199229209229
ساویس	211199230237211	سایا	211199237199
ساوین	211199230237228	ساهري	211199229209

Example

Step	Action	Description
Standardization	Harmonize the File formats Harmonize Data and Variable format Control Variable definitions and their attributes Consider the related Para Data	These actions are necessary tasks before making linkages to remove many inconsistencies
Purification and data cleaning	Ensure that no strange value/ character is on dataset Remove additional characters (e.g. @)	This step removes all additional characters with functions
Record Linkage	Converting text/numeric variables to ASCII codes Design the linkage algorithms Make a linkage with/without primary key	Obtaining results and final controls to assure the quality of linkage is suitable

The data integration process

SQL Server function for changing characters (Farsi or English) into 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
```

Different Steps for linkage with innovative approach

Step 1: Preparation before linkage includes standardization and purifications (remove extra characters)

Step 2: Convert alphabetic characters into ASCII codes for 12 linkage variables: (Variables: Name, Family Name, Province, City, Birth Certificate, Number, Birthday, Age, Gender, Marital status, Contact number, Postal code, Father Name)

Step 3: For each cell of ASCII code in dataset 1, make a fuzzy lookup search on corresponding column in dataset 2 and find 5 top similarities (there similarity threshold can be determined of 90% similarity) with Fuzzy Lookup function in SQL

Step 4: Sort the similarities from biggest to smallest

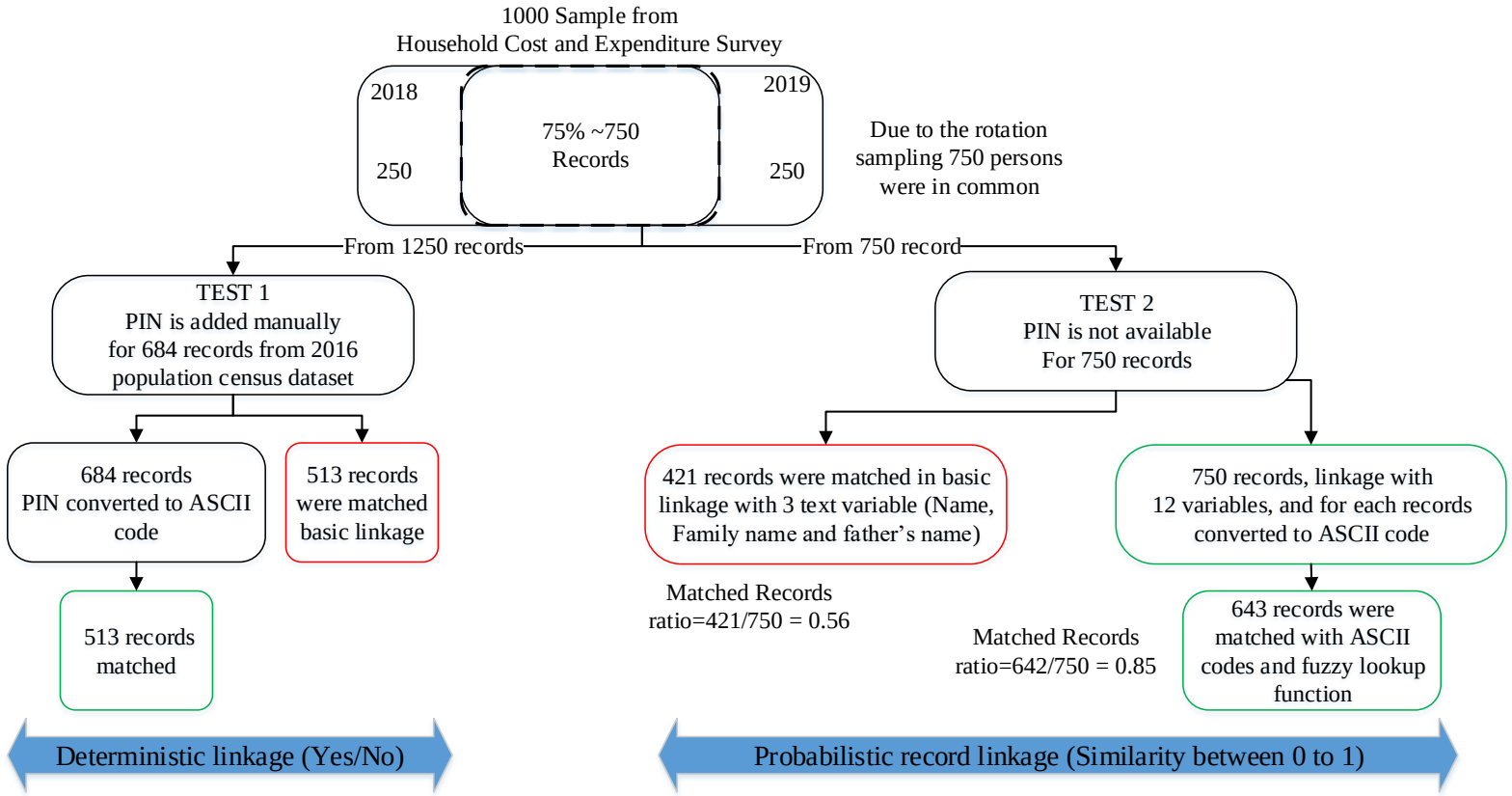
Step 5: Create the probability vector for each records with the maximum similarities for each record

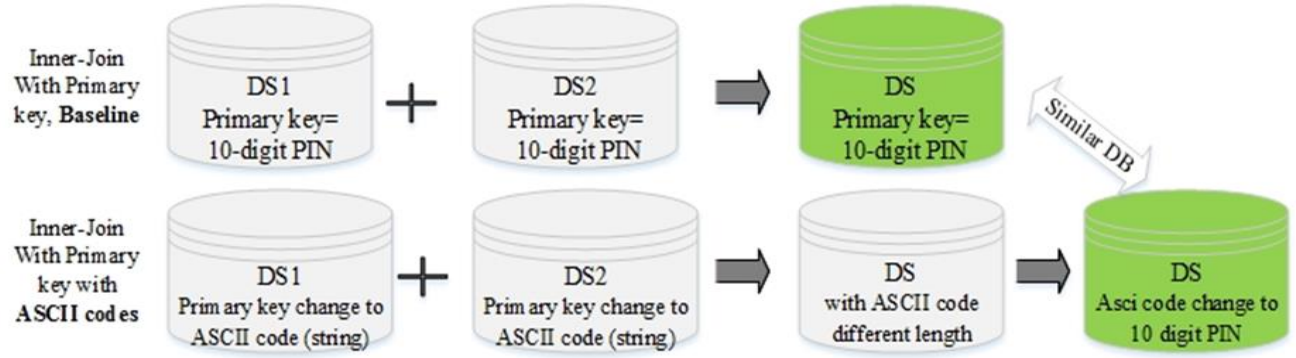
Step 6: compare the number of similarities with probability equal to 1 in the vector with threshold number (here the threshold was considered 7)

Step 7: if the number of similarities in vector greater than threshold 7 then the recorded were totally match

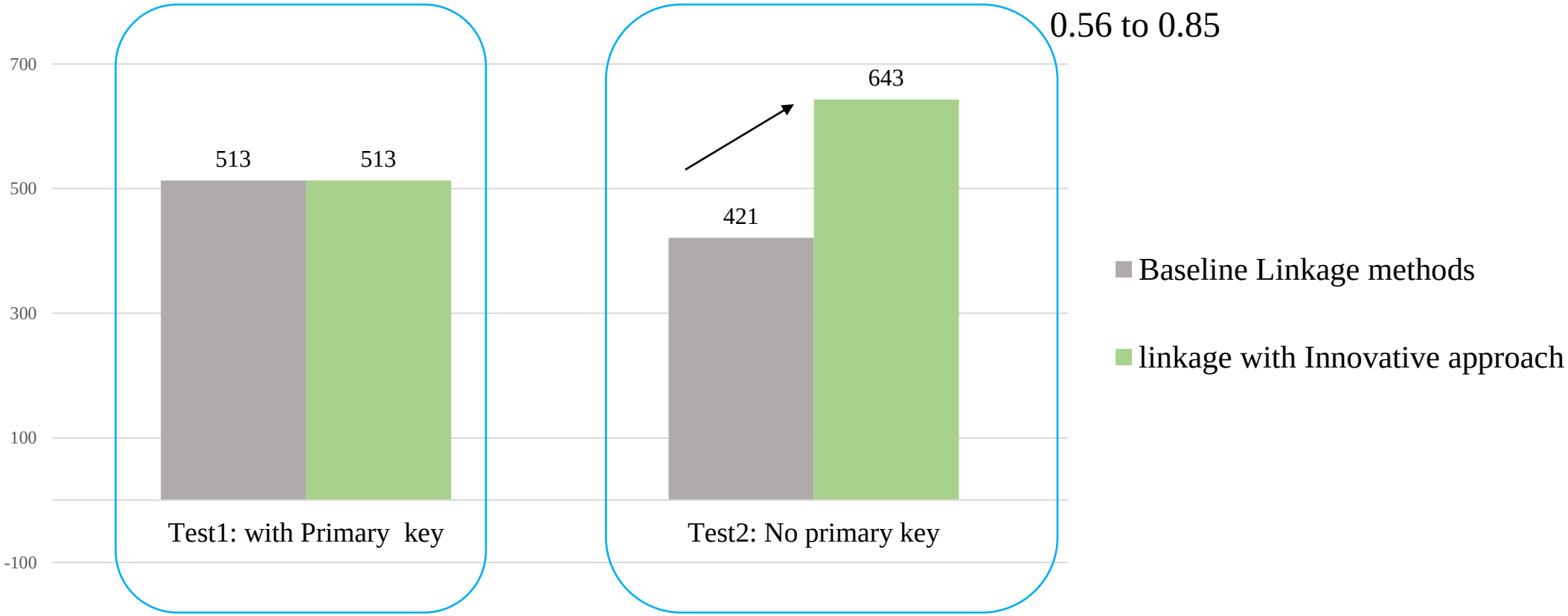
Step 8: In the new merged dataset put the variable with maximum length either from dataset 1 or dataset 2

Step 9: if the number of similarities less than threshold 7, then write both records in new merged dataset as separate records.




Inner-Join With Primary key, Baseline
DS1 Primary key= 10-digit PIN
DS2 Primary key= 10-digit PIN
DS Primary key= 10-digit PIN
Similar DB
Inner-Join With Primary key with ASCII codes
DS1 Primary key change to ASCII code (string)
DS2 Primary key change to ASCII code (string)
DS with ASCII code different length
DS Asci code change to 10 digit PIN

Results


0.56 to 0.85
700
500
300
100
-100
513
513
421
643
Test1: with Primary key
Test2: No primary key
Baseline Linkage methods
linkage with Innovative approach

2

Innovative approach 2: Using text mining to address coding challenges in administrative datasets

HS coding system

At the international level, the Harmonized System (HS) classifies merchandise trades into a six-digit code system. It comprises approximately 5,300 article/ product descriptions that appear as headings and subheadings, arranged in 99 chapters, grouped in 21 sections. The six digits can be broken down into three parts. The first two digits (HS-2) identify the chapter the merchandise are classified in, e.g. 09=Coffee, Tea, Mate and Spices. The next two digits (HS-4) identify groupings within that chapter, e.g. 0902 = Tea, whether or not flavored. The next two digits (HS- 6) are even more specific, e.g. 090210 Green Tea (not fermented).

HS-2-Digit Codes for different sectors

HS code	Group's Name	HS code	Group's Name
01-05	Animal & Animal Products	50-63	Textiles
06-15	Vegetable Products	64-67	Footwear / Headgear
16-24	Foodstuffs	68-71	Stone / Glass
25-27	Mineral Products	72-83	Metals
28-38	Chemicals & Allied Industries	84-85	Machinery / Electrical
39-40	Plastics / Rubbers	86-89	Transportation
41-43	Raw Hides, Skins, Leather, & Furs	90-97	Miscellaneous
44-49	Wood & Wood Products		

Sub categories for specific ICCHS code 02041000

Iran Custom's Customized Harmonized Commodity Description and Coding System

ICCHS code	Descriptions
02041000	The carcass and lamb are left according to the value statement
02041000	The carcass of the remaining meat according to the declaration of value
02041000	The carcass of fresh mutton remains according to the value statement
02041000	Hot mutton according to the value statement
02041000	The remaining carcass of the sheep according to the declaration of value

Iran's customs register information has a **description of the goods** which include the specifications and attributes of the imported or exported goods. For example, ICCHS code 02041000 has been described. In order to calculate accurate price indices based on **the ICCHS 8-digit codes**, comparative prices should be calculated, and it would be necessary that each 8-digit code should have identical attributes between two different periods.

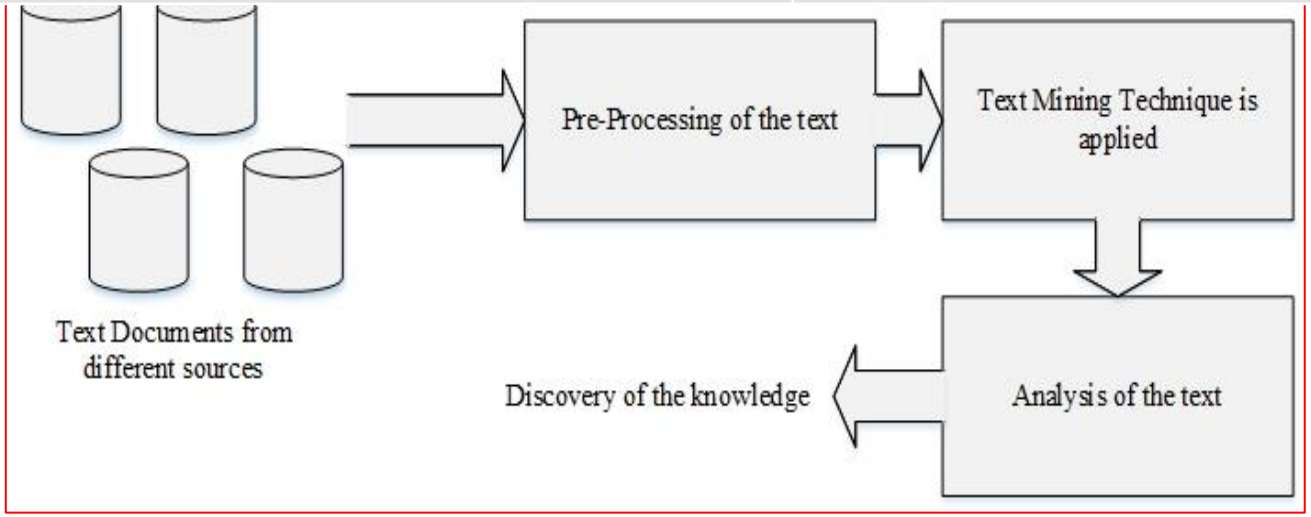
Challenges

Taking register data as a data source, there are some remarkable challenges of different type of goods as well as redundant characters resulting to low efficient linkage. So, if the linkage is applied based only on the ICCHS 8-digit, the calculated indices will be misleading and biased.

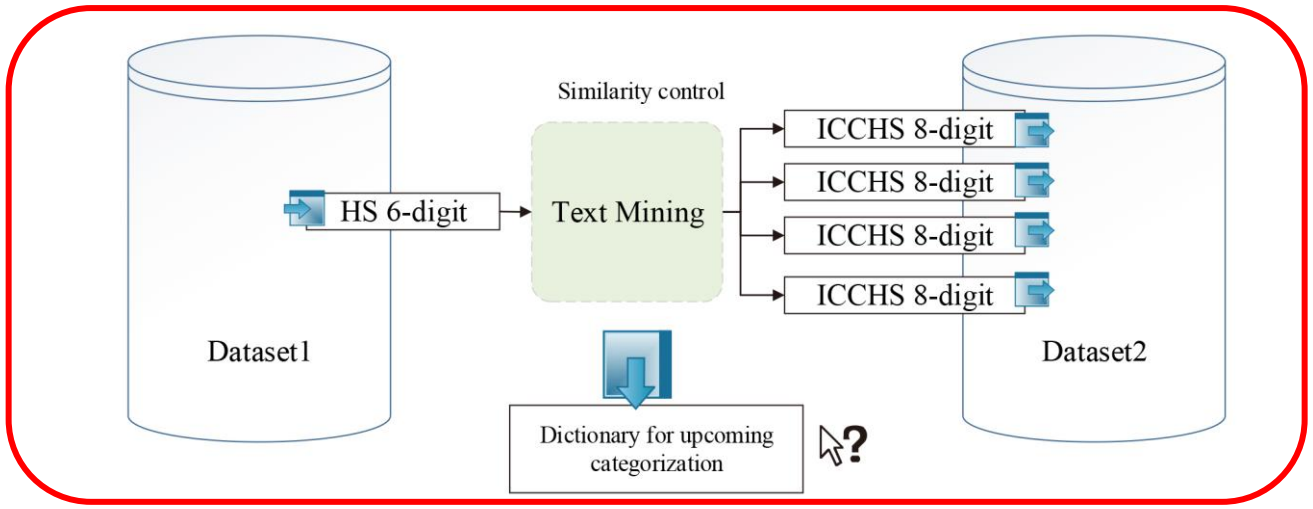
Iran's Customs dataset is a large dataset. More than 400,000 records are received on a daily basis. This makes it impossible to eye-control and review the data for inconsistencies between the Customs attribute set descriptions between one period and another.

The text mining process

Step	Action	Description
Pre-processing	Removing numbers Removing punctuation Removing stop words Removing strip whitespace Steaming	remove these characters (punctuation, numbers, stop words and whitespaces) @, “ “, !,%,(), …
Text mining	determine Jaccard similarity and Cosine similarity	information retrieval, text classification, document clustering, topic detection, topic tracking, questions generation, question answering, essay scoring, short answer scoring, machine translation, text summarization
Analysis	Ensure that no strange value/ character is on dataset Match the new texts with proper code Library making	Assign a label to each ICCHS code based on the keyboard’s description and the number of repetition in the database. These labels however can be a 4-digit number which can be attached to previous 8-digit ICCHS codes. Categorizing after the matching the proper codes



Linkage process with **Text mining** in price indices



In order to create a dictionary of frequent attributes, a 4-digit number was added to ICCHS code. In Table, the results of this conversion of ICCHS 8-digit to 12-digit code is presented. This dictionary will be advantageous for further text mining applications and similarity identifications.

Sub-categories for specific ICCHS code

ICCHS code	Descriptions	Identify Code
02041000	carcass lamb left accord value statement	**0111**
02041000	carcass remain meat accord declaration value	**0112**
02041000	carcass fresh mutton remain accord value statement	**0113**
02041000	Hot mutton accord value statement	**0114**
02041000	remain carcass sheep accord declaration value	**0115**

Pre-processing

Commonly each Custom's attribute set's description contains a series of characters including but not limited to numbers, symbols, low importance signs and redundant spaces. Thus, in order to prepare high quality analysis on these descriptions, it is necessary to remove these characters (punctuation, numbers, stop words and whitespaces). The last is 'stemming' There is package for Stemming function in R programming (Package tm, function stem-document).

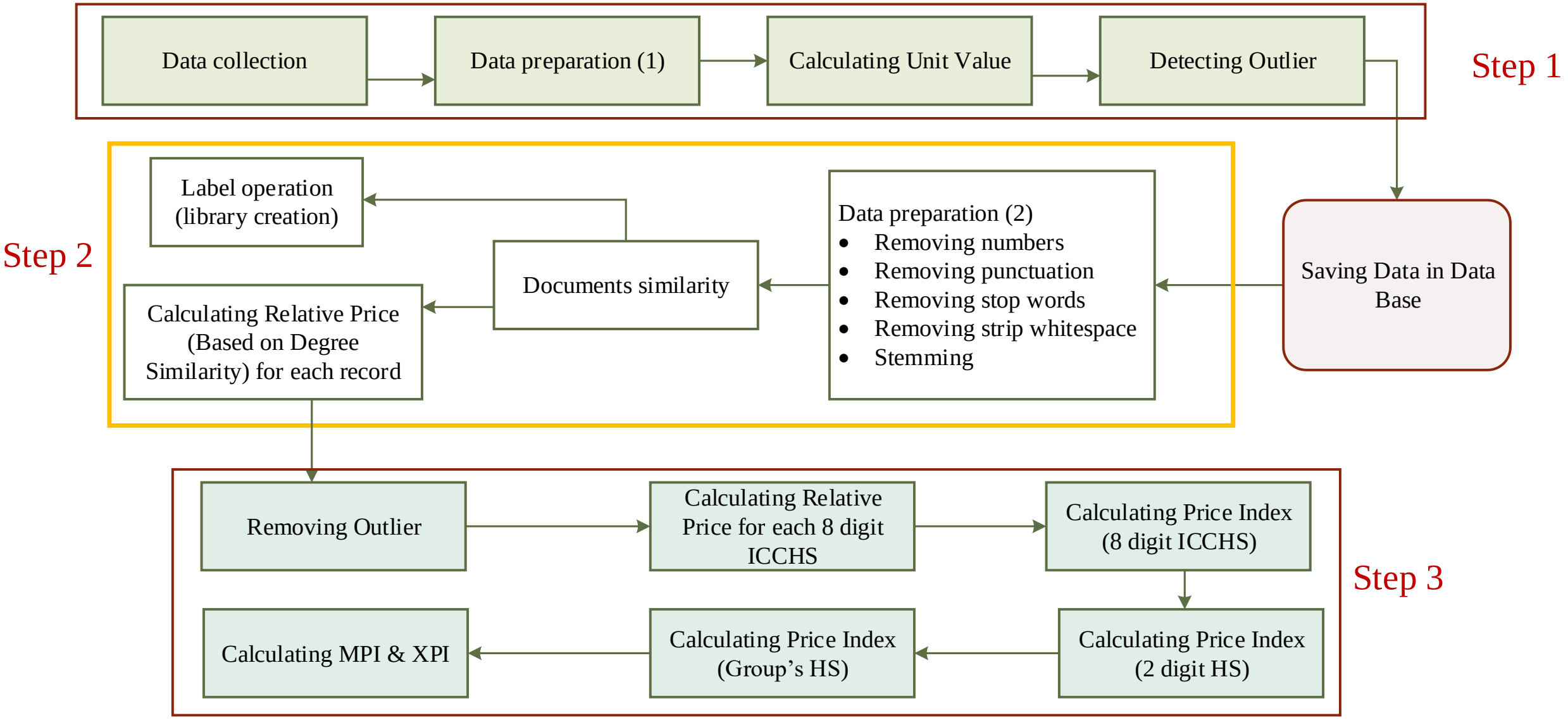
Text mining

One of the famous methods is the metric distance that exists between two texts. The Jaccard similarity and Cosine similarity methods to find similarities between two texts are used in the R programme

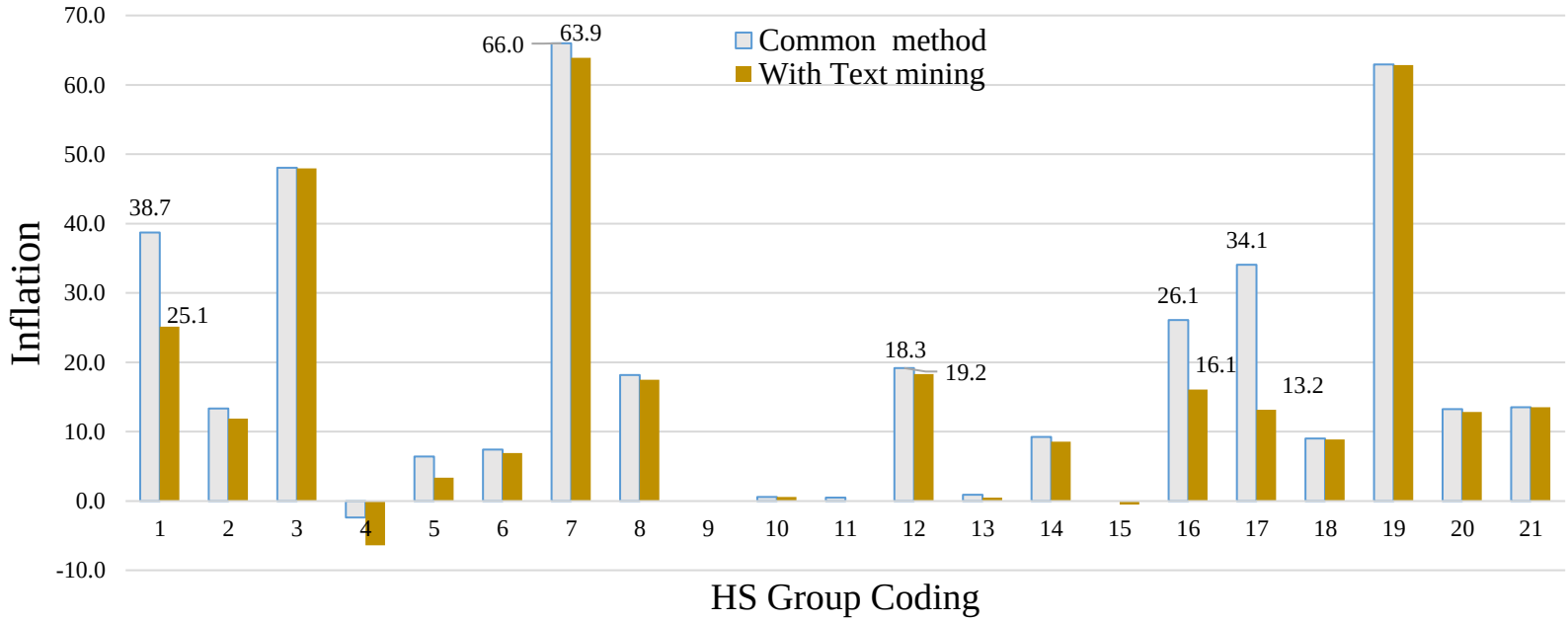
Analysis of the text

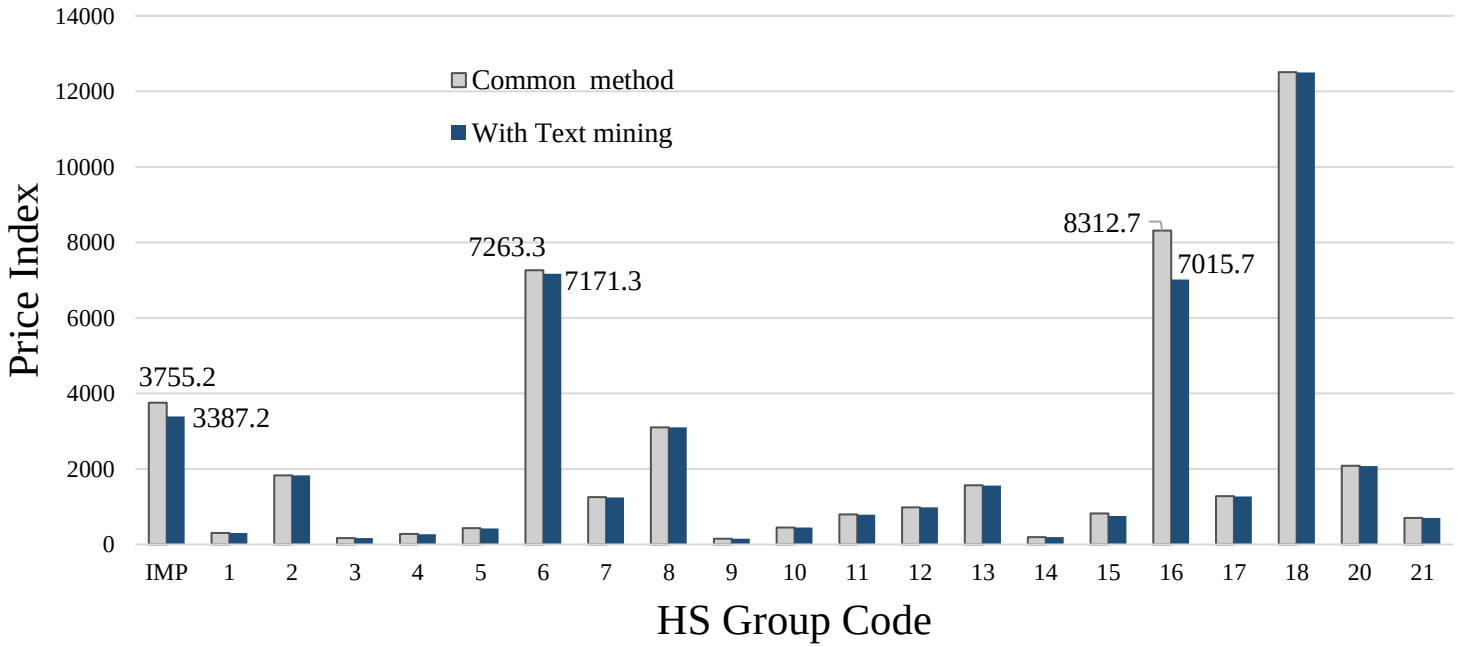
Following the similarity determination, a label was given to each merchandise's ICCHS code based on the key words mentioned in description and the number of repetitions in the database. These labels however can be a 4-digit number which can be attached to previous 8-digit ICCHs codes. The new codes had 12-digits that would be more beneficial for next linkage. The new 12-digit codes were combined from 4-digits and ICCHS 8 digit-code.

Different steps of MPI and XPI price indices with Innovative approach



Inflation rate for imported merchandise in Quarter 3 2o19 in Iran.





Price Index for imported merchandise in Quarter 3 2019 in Iran.

Conclusion

The first innovative approach replaced text variables with ASCII codes to address language traditions where many characters are used interchangeably. The results show improved linkage rates. This approach has wide applicability in other situations where alphabetic traditions may impede the quality of record linkage based on text variables. This method can be applied by statisticians and data scientists in other non-English countries and situations

1

The second innovative approach applied text mining techniques to enhancing the administrative classifications used in the Customs dataset for the purpose of compiling internationally comparable price indexes. The results showed lower prices. This approach has wide applicability when administrative datasets with classification systems optimized for administrative, not statistical purposes, are used. This approach can be extended to all price indices like the Consumer Price Index (CPI) to also improve the quality of inflation rate. In Iran, the two innovative approaches are still in their research phase

2

Quality of integration

Thank you for attention

Acknowledgments