TEXTUAL & DEMOGRAPHIC DE-DUPLICATION OF LPG BENEFICIARIES

The subsidized domestic LPG connections in India are almost 14.58 crores. The policy to avail LPG gas connection is "One Kitchen, One Connection" with Single bottle or double bottled cylinders for domestic purpose. Textual and Demographic Deduplication (TDD) technique assists in identifying a duplicate person using well defined algorithm and computing facility.

DR. NEERAJ MITTAL IAS
Joint Secretary (Marketing)
M/o Petroleum and Natural Gas
mittal1967@gmail.com

G. MAYIL MUTHU KUMARAN
Technical Director
muthu@nic.in

DEEPJOT KAUR
Scientific Officer
deepjot.kaur@nic.in

Edited by
MOHAN DAS VISWAM

There exist various deduplication techniques and technologies for identifying a duplicate person through DNA samples, biometrics like finger print, iris, facial photograph, hand geometry and other such methods. But, it is always difficult to identify the duplicates by the authentication or other modes of de-duplication, as a large chunk of data has accumulated with enormous duplicates with or without intentions over a period in almost all sectors. In governance, the duplicates pave the way for diversions and misuse of the subsidy meant for the intended beneficiaries.

Textual and Demographic Deduplication (TDD) can be achieved without the beneficiaries' presence for authentication with a well defined algorithm and computing facility. Though name and address matching may seem to be simple and straight-forward process, however, when the live data is analysed, immense complexities quickly emerge. It was experienced that originating a quality name and address matching solution requires focussed analysis into alchemy of data. These include detailed analysis and understanding of data to derive the causes where differences can originate.

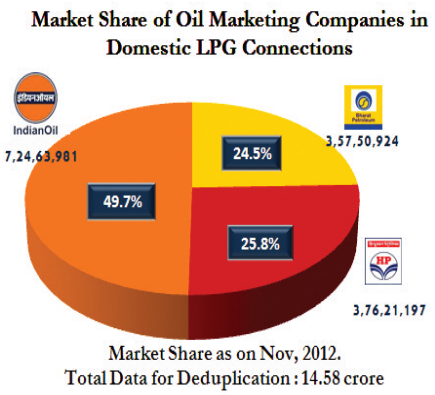
The algorithm was developed with an objective of name and address matching using the following strategies:

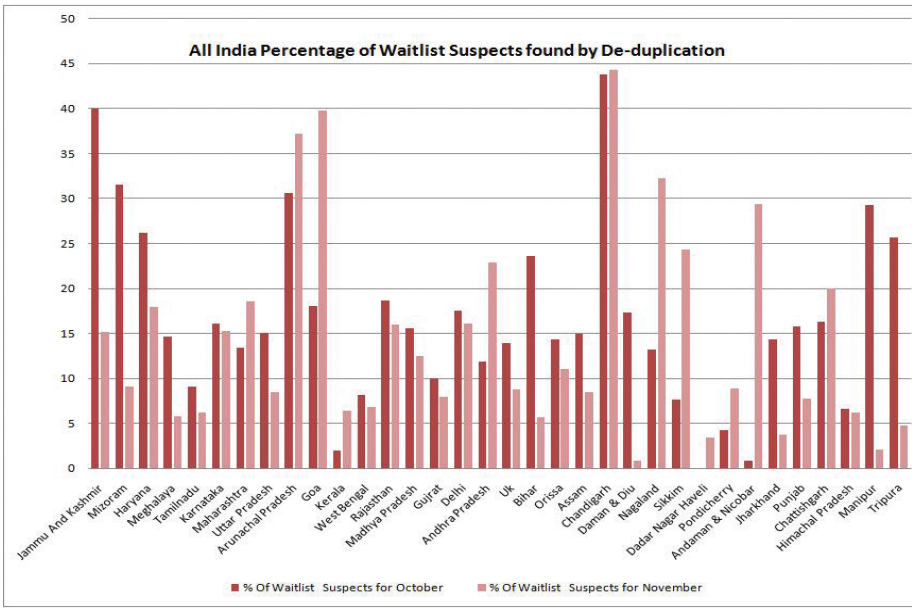
1. Spelling variations include interchanged or misplaced letters due to typing errors, substituted letters, and omissions
2. Phonetic variations due to mishearing and digitizing mistakes
3. When individual changes his/her name during the course of life
4. Issues due to various cultural adoptions
5. Capturing phonetic equivalents
6. Data Quality/Data Completeness

TEXTUAL AND DEMOGRAPHIC DEDUPLICATION

The subsidized domestic LPG connections in India are approximately 14.58 crores cumulatively for IOCL, HPCL and BPCL. In a set astir of curtailing the subsidised cylinders by tracking and blocking fake multiple connections, Textual and Demographic De-duplication activity is being performed on this data.

The policy to avail LPG gas connection is "One Kitchen, One Connection" with single bottle or





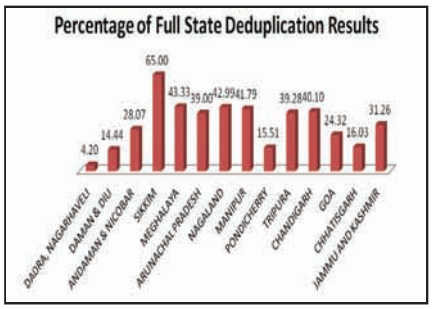
double bottle cylinders for domestic purpose. All India LPG domestic connections data as provided by Oil Marketing Companies (IOCL, HPCL & BPCL) have been collated and de-duplicated after which suspect report is generated state wise with a count of Inter and Intra OMC connections. The challenge in the activity is that the data is legacy and in different formats with names and addresses found incomplete or in varied transformations.

Results generated by NIC algorithm were compared with the Industry and the OMC's data and results found were superior and encouraging.

In the first phase, waitlist connection data till October 2012 of 34 states / UTs was taken up for deduplication. The details of this processing are as follows:

PHASE I	
Total no of consumers	12,35,67,627
Total no of wait list connections	14,97,538
Total Suspects	1,55,014
Clear List	13,42,524
Percentage of suspects	10.30%

In the second phase, waitlist connection data till December 2012 of 34 states was taken up for



deduplication. The details of this processing are as follows:

PHASE II	
Total no of consumers	14,58,36,1027
Total no of wait list connections	13,68,535
Total Suspects	1,90,045
Clear List	11,78,490
Percentage of suspects	13.80%

In the third phase, entire state data for fourteen states has been de-duplicated and other states are in progress till date.

PHASE III	
Total no of consumers	60,17,974
Total Suspects	17,29,348
Percentage of suspects	28.73%

OMC DATA DEDUPE AGAINST PDS BPL DATA

Oil Marketing Companies data was taken up for Deduplication against the Public Distribution Systems BPL (Below Poverty Line) data of Delhi and Chandigarh. The details of this processing are as follows:

OMC Data Dedupe against PDS BPL data - DELHI	
Total no of LPG consumers	52,34,152
Total no of PDS's BPL beneficiaries	3,74,358
Total Suspects	95,248

OMC Data Dedupe against PDS BPL data - CHANDIGARH	
Total no of LPG consumers	4,43,402
Total no of PDS's BPL beneficiaries	67,299
Total Suspects	42,611

The de-duplication activity is being carried out by NIC using the de-duplication algorithm developed by NIC and only computing facility is availed from CDAC's HPC PARAM YUVA Supercomputer with 24 nodes of 16 cores each which recently upgraded to 48 nodes of 16 cores each to reduce the LPG beneficiaries waiting period (who have applied for new connections) from months to hours.

PARAM YUVA	
Configuration	24 nodes (each nodes 16 CPU's)
Model name	Intel(R) Xeon(R) CPU: X7350 @ 2.93GHz
Cache size	4096 KB
Address sizes	40 bits physical, 48 bits virtual

FUTURE SCOPE AND WAY FORWARD

TDD has targeted to automate deduplication and identify duplicates in any variations to prevent large scale aberrant uses of the subsidies which are meant for the poor populace of this country. There is a lot of scope on the TDD for the egov projects, namely LPG beneficiaries, ration cards deduplication in PDS, other welfare schemes of Govt. of India.

FOR FURTHER INFORMATION:
G. Mayil Muthu Kumaran
Technical Director
NIC HQ
E-mail: muthu@nic.in
Ph: 01124305748