

**SOCIOECONOMIC MAPPING USING MOBILE CALL  
DETAIL RECORDS FOR SRI LANKA**

Chandima Dileepa Rajaguru

(168258P)

Degree of Master of Science

Department of Computer Science and Engineering

University of Moratuwa  
Sri Lanka

February 2020

**SOCIOECONOMIC MAPPING USING MOBILE CALL  
DETAIL RECORDS FOR SRI LANKA**

Rajaguru Mudiyansele Chandima Dileepa Rajaguru

(168258P)

Dissertation submitted in partial fulfillment of the requirements for the  
degree Master of Science in Computer Science and Engineering

Department of Computer Science and Engineering

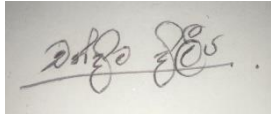
University of Moratuwa  
Sri Lanka

February 2020

## DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief, it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works.



2020-05-30

.....  
Chandima Dileepa Rajaguru

.....  
Date

The above candidate has carried out research for the Masters dissertation under my supervision.

.....  
Dr. Amal Shehan Perera

.....  
Date

## **Abstract**

CDR (Call Detail Record) is a data record that is generated by a telephone exchange or telecommunication equipment which contains details of that telephone call. These records are utilized by telecommunication service providers for their billing purposes. High volume of data generates in quick time which contains customer specific data with temporal and geographic information. Other than CDR data, telco systems have various data sources such as customer payment data and device information. Telco service providers collect CDR and store them for a limited period of time for various activities. It can be repurposed other than billing activities.

CDR data can denote various aspects of human behavior such as human relationships, expenditure power and mobility. Those aspects can help governance of the country regarding economic development and resource allocation in timely manner. In this research, CDR data records were integrated with other telco data sources in order to analyze and predict the economic behavior of a specific geographical area in Sri Lanka.

Big data and Machine Learning techniques were used to extract the customer behavior from CDR data. Big data processing techniques were applied on CDR data and telco data sources in order to identify properties of customers in a specific geographic area over a time period. Then those identified properties were evaluated to see whether they reflect the economic behavior in that area or not. After identifying dominant features related to the economy, Machine Learning techniques were applied on them to see the feasibility of predicting the economic behavior in the targeted area. The results were evaluated and interpreted as a part of this research. Such results will be very useful for the governance in order to understand the economic conditions in a specific geographical area and make the policies to address poverty over the time.

## **ACKNOWLEDGEMENT**

I would like to offer my sincere gratitude to my family for the continuous support and motivation given to me to make this dissertation a success. I also express my heartfelt gratitude to Dr. Amal Shehan Perera, my supervisor, for the supervision and advice given throughout this research for finishing it successfully. Then I pay my gratitude Mobitel (Pvt) Ltd for the guidance and facilities given to me in order to carry out this research. I also thank my friends who supported me in this whole effort.

# TABLE OF CONTENT

DECLARATION	i
Abstract	ii
ACKNOWLEDGEMENT	iii
TABLE OF CONTENT	iv
LIST OF FIGURES	vii
LIST OF TABLES	ix
LIST OF ABBREVIATIONS	x
1. INTRODUCTION	1
1.1 Call Detail Record	1
1.2 Telecommunication Systems	2
1.2.1 Data CDR	2
1.2.2 Telecommunication data sources	3
1.2.3 Telecommunication data management	4
1.3 Research problem	5
1.4 Objective of the research	6
1.5 Expected outcomes of the research	6
2. LITERATURE REVIEW	7
2.1 Socioeconomic data	8
2.2 Feature extraction from CDR	12
2.3 Mobile user home location	15
2.4 Modeling socioeconomic level	17
2.4.1 Linear Regression	17
2.4.2 Support Vector Machine	20
2.4.3 Decision Tree	21

2.4.4	Hypothesis Testing	22
3.	METHODOLOGY	24
3.1	Introduction	24
3.2	Socioeconomic Data in Sri Lanka	25
3.2.1	Official poverty line by district	25
3.2.2	Socioeconomic Index data	27
3.3	Data sources for feature extraction	31
3.3.1	Voice CDR	31
3.3.2	Data CDR	32
3.3.3	Recharge data	34
3.3.4	Postpaid bill data	35
3.3.5	Postpaid payment data	36
3.3.6	Customer details	36
3.4	Usage of big data technologies	38
3.4.1	Hadoop as a CDR storage	38
3.4.2	Data transferring	39
3.4.3	Data processing using Spark	40
3.4.4	Overall architecture of CDR processing	41
3.5	Extracting the home location	42
3.6	Feature extraction from data sources	45
3.6.1	Average outgoing call volume and average outgoing minute usage	45
3.6.2	Average data usage	47
3.6.3	Average recharge amount	48
3.6.4	Average Postpaid bill amount	50
3.6.5	Average Prepaid loan amount	51
3.6.6	Average social media applications' data consumption	52

3.6.7	On Net & Off Net average outgoing minute usage	54
3.7	Calculating feature relationship with SEL data	55
3.7.1	Pearson Correlation Coefficient	55
3.7.2	Calculating correlation with poverty line	56
3.7.3	Calculating correlation with census data	57
3.8	Predicting socioeconomic level	58
4.	RESULTS	59
4.1	Correlation analysis with district poverty line data	59
4.2	Predicting district poverty line	61
4.3	Correlation analysis using DSD socioeconomic index	62
4.4	Predicting DSD socioeconomic index	63
5.	DISCUSSION	66
5.1	Computational complexity	66
5.2	SEL prediction in district and DSD wise	67
6.	FUTURE WORK AND CHALLENGES	69
7.	CONCLUSION	70
8.	REFERENCES	71



## LIST OF FIGURES

	Page
Figure 2.1: MPI calculated in Ivory Coast	11
Figure 2.2: BTSs and their coverage (LEFT), approximated coverage after applying Voronoi Tessellation (Right)	12
Figure 2.3: Location of Voronoi cells, cell phone antennas (light circles) and DHS clusters (dark circles) in Ivory Coast	14
Figure 2.4: Schematic view of a car connected to a base station in a cellular network	15
Figure 2.5: Average SIM mobility during hourly time brackets in a single day	16
Figure 2.6: Performance of regression models against different variable types in Western Province	18
Figure 2.7: Observed and predicted poverty level in Ivory Coast	19
Figure 2.8: Correct classification rate for different feature selection method	21
Figure 3.1: Official Poverty line by district for last two years in Sri Lanka	26
Figure 3.2: Socioeconomic Index distribution in Sri Lanka which was calculated using 2011/12 census data	30
Figure 3.3: CDR stored in HDFS cluster	39
Figure 3.4: High level architectural diagram of the analytical platform	41
Figure 3.5: Data extracting process in deriving home location	43
Figure 3.6: Geographical view of subscribers' home district distribution in Sri Lanka	44
Figure 3.7: Subscribers' home district distribution in Sri Lanka	44
Figure 3.8: Outgoing Call feature extraction process	45
Figure 3.9: Average monthly outgoing minute usage district wise	46
Figure 3.10: DSD wise average outgoing minute usage	46
Figure 3.11: Average monthly data usage district wise	47
Figure 3.12: DSD wise average data usage	48
Figure 3.13: Average monthly recharge amount district wise	49
Figure 3.14: DSD wise average recharge amount	49
Figure 3.15: Average monthly Postpaid bill amount district wise	50

Figure 3.16: DSD wise average Postpaid bill amount	51
Figure 3.17: Average monthly Prepaid loan amount district wise	52
Figure 3.18: Average monthly Facebook data usage per connection district wise	53
Figure 3.19: Average monthly Facebook data usage per connection DSD wise	53
Figure 3.20: District wise monthly average on-net outgoing minute usage	54
Figure 3.21: District wise monthly average off-net outgoing minute usage	54
Figure 3.22: Pearson correlation coefficient values	55
Figure 3.23: Poverty Index prediction process	58
Figure 4.1: District wise prediction of poverty line	61
Figure 4.2: DSD wise actual SEL vs predicted SEL	64

## LIST OF TABLES

	Page
Table 1.1: Example set of CDR Fields	1
Table 1.2: Data CDR Fields	3
Table 2.1: Feature list selected from census	8
Table 2.2: Features derived from CDR data	13
Table 2.3: Census Variables and Mobility Variables [3]	22
Table 2.4: Correlations between features derived from mobile phone data and poverty level [2]	23
Table 3.1: Official Poverty line from 2019 January to 2019 June for few districts in Sri Lanka	25
Table 3.2: Data Set and respective categories used for index	27
Table 3.3: Few of the Category and values used for Socio Economic Index	28
Table 3.4: MapReduce and Apache Spark comparison	40
Table 4.1: District wise Pearson Correlation Coefficient for each feature	59
Table 4.2: Pearson Coefficient average and variance	60
Table 4.3: P-values associated with district level features used in LR model	62
Table 4.4: DSD wise Pearson Correlation Coefficient	63
Table 4.5: P-values associated with DSD level features used in LR model	65
Table 5.1: CDR volume stored for parameter extraction	66

## LIST OF ABBREVIATIONS

CDR	Call Detail Record
ML	Machine Learning
DSD	Division Secretariat Division
GND	Grama Niladhari Division
OLAP	Online Analytical Processing
RDBMS	Relational Database Manage System
SEL	Socio Economic Level
MPI	Multidimensional Poverty Index
DHS	Demographic and Health Survey
BTS	Base Transceiver Station
PCA	Principal Component Analysis
SVM	Support Vector Machine
DCS	Department Census and Statistics