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 Mudiyanselage 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

LIST OF FIGURES

LIST OF TABLES

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

1. INTRODUCTION

1.1 Call Detail Record

A call detail record (CDR) is a data record produced by a telephone exchange or other telecommunication equipment that mentions the details of a telephone call or other telecommunications transaction (e.g., text message) that passes through that facility or device. A CDR contains many attributes about the call such as source number, destination number, time, duration, cell id and country code [3]. These data can be considered as structured data. According to the service provider's statements, thousands of records are generated in a few seconds. Therefore, CDR data stream is big data which show all its properties (big volume, big variety and big velocity).

Table 1.1: Example set of CDR Fields

Attribute	Example	Remarks
Source Number	9471XXXXXXX	Number of the caller
Destination Number	9471XXXXXXX	Number of the callee
Caller Cell Id	xxxxxx	
Callee Cell Id	xxxxxx	
Caller County Code	94	
Callee Country Code	94	
Caller IMEI Number	41301306104XXX	International Mobile Subscriber Identity
Callee IMIE Number	41301307205XXX	
Time Stamp of SSP	20161215094239	
Time Zone of SSP	32	
Call Duration	82	In seconds
Call Waiting Duration	1	In seconds
Prepaid or Postpaid	1	

Above table (Table 1.1) describes important attributes which can be found on a CDR. Format and attributes can be different with the practices of telco service providers.

A CDR reflects an accurate record of what has happened during a telephone call. It is used by telecommunication companies for many tasks. Calculating revenue and generating billing records are a few important tasks carried out. A CDR is an outcome of the call and it can be used for many things as it contains statistics of caller and callee, temporal and geographical information. With thousands of users, it will generate a data stream of CDR which can be considered as big data.

1.2 Telecommunication Systems

1.2.1 Data CDR

When someone is browsing the internet by navigating through the web pages, filling forms and clicking links in websites using mobile phones, a data session is started and telco service providers serve data needed for that customer. When a data session is finished, a record similar to CDR is generated in telco networking systems. Those CDR are called as data CDR in the telco domain.

A data CDR contains some valuable attributes such as customer identifier, from which tower or cell the data session was served, uploaded data flux amount, downloaded data flux amount, data session started time, data session ended time, application id if available, etc. Main attributes which can be found on a data CDR are mentioned in the following table (Table: 1.2). Those attributes can be used to analyze user consumption and his mobility. Using the application ids used in the data session, interesting user behavior patterns can be identified.

These data CDRs are used for billing purpose. They can be used for real-time campaigns as well. Following table (Table 1.2) has its attributes.

Table 1.2: Data CDR Fields

Attribute	Example	Remarks
Serving Number	9471XXXXXXX	Customer Mobile Number
Record opening time	2019-11-29 10:40:42	Session started time
Data type	1	2G/3G/4G data type
Cell id	xxxx	Geographic location
Duration	200	Data session time
Uploaded data volume	12244	Data upload in bytes
Downloaded data volume	4130	Downloaded data volume in bytes
Service id	232	Data consumed application id

1.2.2 Telecommunication data sources

Other that the CDR data, telco systems have many data sources. Some of them are related to the customers and some are related to the networking nodes such as cell towers. Following are few data sources which can be identified.

- Recharge data
  This data source includes data related to the prepaid card recharges, reloads, etc.
- Customer bills
  Customer bills are generated using CDR monthly. The bills contain summary of CDR and tax charges. Bills are posted to the customers at a scheduled time monthly. Most of the time, bills are for the postpaid customers.
- Customer payments

Customers make their payments when they are settling their bills or purchasing devices. Payments are done in retail outlets, banks, service provider shops, etc. Those data are stored to maintain customer account status.

- Device information
  These data contain the mobile device information of the customers. The device is mapped using the IMEI in CDR. That process can reveal which device customer is using.

1.2.3 Telecommunication data management

Since a massive amount of data is generated in telco systems, proper data management is required. Transactional data like bill payments are directly stored in a RDBMS. Reference data like cell information, device information are also stored in RDBMS. Data sources which should adhere to ACID properties are persisted in relational databases.

CDR data flow into the systems in high amount and in high velocity. Therefore, they should support big data capabilities. Most of the time, CDR data are stored in Hadoop big data warehouses. They are subjected to OLAP when required. RDBMS data also get replicated in big data storage in time to time. Data of Million customers are stored in a data lake for the analytical purposes and reporting requirements.

1.3 Research problem

Economy can be defined in many ways and it is a vast area which is hard to understand and interpret. The economy in an area depends on production, distribution, trade, consumption of goods and services in that area. Most of the time, analyzing economy is done at the macro level, which is defined in country level or provincial level. Getting a summarized view of the economy in such macro level requires huge amount of effort. Drilling down into the micro level of a geographic is even harder compared to the macro level.

When analyzing how the economy is behaving in a specific area, the population relevant to that economy should be identified first. There are a number of parameters and measures used in economic analysis. The data used for such analysis should be collected from a population who exactly living in the targeted area. Collecting data, focusing on a specific geographical area, is not straight forward. It needs more time and effort.

Another problem is how to analyze the evolution of the economy in a targeted geographic area. Analysis done in an exact time has to be repeated over a period of time to observe the behavior of the economy in that area. There are a number of events take place in the country or in a specific geographical area. Government takes decisions on the economy and they want to see how those decisions affect the country or an area. Examining the effect on the economy as a result of an event is hard because data has to be collected continuously and analyzed over a period of time.

As discussed above, the research problem is focused on how to analyze the economic behavior of a specific geographical area. The geographical area can be a small governance unit in the country. Data collected in the analysis should be relevant to the population who are living in that area. Furthermore, analyzing and interpreting the evolution of the economy in such area requires much more effort and more time.

1.4 Objective of the research

The objective of this research was to develop a methodology or a model using Data Science to analyze or interpret the economic behavior in a specific geographical area using telco domain data. Number of data sources which can be found in telco domain were used. Following are some of the data sources which were available for the research.

- CDR Data
- Click Stream Data
- Data CDR
- Device Information

There are thousands of attributes which can be used directly or derived using the above data sources. Finding the relevant features from those attributes which can be used to model the economic behavior was one of the objectives in this research. Developing a methodology or a model using Machine Learning to predict the socioeconomic level in Sri Lanka was the other objective which can save lot of cost and effort.

1.5 Expected outcomes of the research

Following are the expected outcomes of this research.

- Features extracted from data sources in telco domain data sources which have a correlation to the economic behavior
- Methodology or model which can be used to analyze the economic behavior of a specific geographical area using the features extracted
- Visualization or an interpretation which can show the evolution of the economy in a specific geographical area

2. LITERATURE REVIEW

This research is carried on CDR data with other telco data sources. It is about extracting features from data and modeling the economic behavior of a geographical area. Number of researches have been done on extracting knowledge from CDR in the past few years. The difference in this research is integrating CDR with other telco data sources in order to analyze the economic condition in a Sri Lankan geographical area. In the literature review, past work on mining patterns on CDR will be discussed as well as researches which have done on analyzing economic conditions.

Since analyzing economic condition may involve identifying customer behavior with various capabilities like expending capability and evolution of those behaviors, this literature review is mainly focused on feature extraction and Machine Learning techniques. Data sources play an important role in socioeconomic analysis because correct and relevant features should be identified. Literature review discusses how CDR were used to extract the features on subscriber wise or BTS wise.

Other than CDR attributes, various types of socioeconomic data sets have been used. Census data and survey data are the ones which have been used frequently. The literature discusses what types of data sources used to symbolize economic behavior. To carry out this research, the residence location of the mobile user has to be predicted. The literature review contains past work on how the home location of the user was derived.

Literature has revealed Machine Learning techniques such as Linear Regression, Support Vector Machine and Decision Trees which have been used to predict the economic level of a group of people in an area. Hypothesis testing is another approach used which has been discussed in the literature review.

2.1 Socioeconomic data

Socioeconomic level (SEL) changes over the time in country smoothly not suddenly with new government policies. Economic behavior in an area is linked with income, education, gross domestic product, population, infrastructure facilities and many more. In order to access those conditions and view in detail, government and non-governmental organizations carry out surveys on target population samples. Those surveys can help them to get quick insights in a short time with expensive effort.

Another way of accessing the ground truth data on the economic conditions is conducting a census covering domestic measures. Doing a census should incur huge capital and human effort. Research[1] done on predicting socioeconomic characteristics in Sri Lanka, used 2011/12 census data collected by Department of Census and Statistics in Sri Lanka.

Sri Lankan census data have been published at the district level (administration level 2), some data have made available at the Graama Niladhari Division (lowest administration level, 4th level). In the research, they have used 58 features from the census data. They have got biased towards household data like roof material, wall materials and infrastructure of housing in a specific area as shown in Table 2.1. GN-level data have not included data such as literacy and computer literacy rates.

Table 2.1: Feature list selected from census

Census Feature Category	Census Feature
Floor Materials	Cement Tile/Granite/Terrazo Mud Wood Sand Concrete
Roof Materials	Tile Asbestos Concrete Aluminium Sheet

	Metal Sheet Cadjan/Palmyrah/Straw
Wall Materials	Brick Cement Block Cabook Soil Bricks Mud Cadjan/Palmyrah Plank/Metal Sheet
Type of Structure	Single House - 1 Storey Single House - 2 Storey Single House - 2+ Storey Attached House/Annex Flat Twin House Row/Line Room Hut/Shanty
Housing Type	Permanent Semi-Permanent Improvised
Tenure	Encroached Rent Free Rent (Private Owned) Rent (Government Owned) Owned
Cooking Fuel	Fire wood Kerosene Gas Electricity Dust
Lighting	Electricity Kerosene Solar Power Bio Gas
Education	No schooling Primary Secondary O/L A/L Degree
Employment	Employed Unemployed Not Active
Gender	Male Female

Age	Young Middle-aged Senior

Poverty estimation research carried out in Ivory Coast [2], they have used poverty rate estimates provided by the International Monetary Fund, dating from 2008 for evaluation purposes. In the same work, for Region B, asset based index has been derived for the same purpose using Census data. The census data have contained 14 variables related to household ownership of assets such as laptops, mobile phones and various kinds of vehicles. Principal Component Analysis (PCA) was used to come up with the mentioned asset based index referred as the poverty level.

In a Latin American Country, census data collected by the local National Statistical Institute (NSI) have been used to identify relationship between socioeconomic factors and cell phone usage[3]. They have focused on three groups of variables extracted from census data. Those three groups were education variables, demographic variables and goods' ownership variables. Education variables contained attributes like primary, secondary school population. Demographic group included variables on male female population, age wise population. Household items and facilities like electricity and water were a few of the variables computed on goods ownership.

Similar research carried out on a Latin American country[4], they have defined three levels based on census data. National Statistical Institute of the country have defined three SELs where three categories were A, B and C. Category A was the highest SEL. The SEL value was created from the combination of 134 indicators. Few examples for those indicators are occupation of the members of the household, level of studies of the members in the house, combined income, the number of rooms in the house, the number of cell phones, land lines, or computers.

Poverty Prediction done in Rwanda[5], data gathered from small surveys done with individuals, were applied. They have used the merged data from those survey data which contained questions about the housing characteristics, asset ownership and several other basic welfare indicators of 856 phone survey respondents. Utilizing collected data, they have constructed a composite wealth index using the first principal component.

Empirical Application done to monitor poverty in Ivory Coast[6], Multidimensional Poverty Index(MPI) was calculated using data collected form 2011/2012 Demographic and Health Survey (DHS), held from December 2011 to May 2012. The attributes such as education, asset ownership and health were subjected in MPI calculation. But the consumption and income data were not collected in the survey. The MPI calculated is visualized geographically in Figure 2.1.

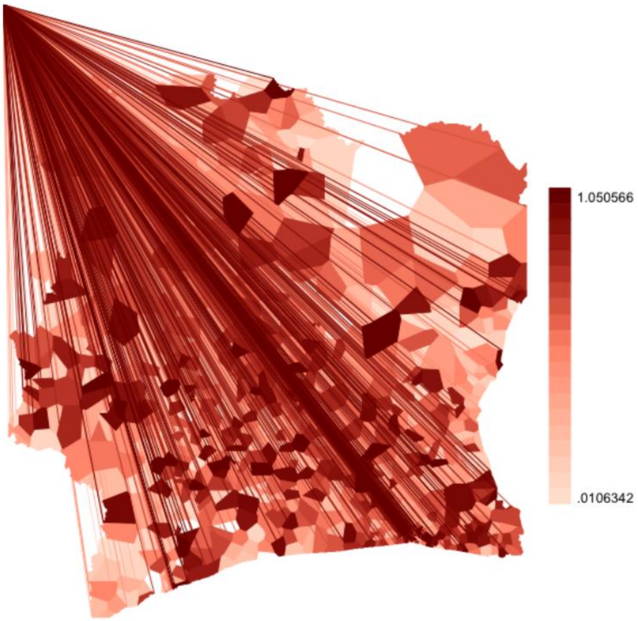


Figure 2.1: MPI calculated in Ivory Coast

As mentioned, in order to capture Socioeconomic Level (SEL), data tagged geographically should be incorporated to reveal the ground truth. Census data were frequently used. Other than that, the surveys conducted were used based on the requirement. Conducting a census is a very expensive task, but effective. Principle Component Analysis has been applied on data to create a poverty index for evaluation and model building purposes.

2.2 Feature extraction from CDR

When a mobile user makes a phone call, a record is generated in telco system called a CDR (Call Detail Record). There are many types of CDR as a result of a phone call, a SMS, a MMS or a data session. CDR has many attributes based on the configurations of the telco systems. Unique identifier to identify who is calling, unique identifier to identify who is receiving the call, time of the call, duration of the call, id of the BTS which caller connected, the id of the BTS which call receiver connected are some of the main attributes found in a CDR.

Base Transceiver Station (BTS) is the main component in a telecommunication network which is responsible for directing a call from source to destination. It is an antenna which has a longitude and a latitude of its geographical location. The id of the BTS is very important to identify the geographical location of the person who is making the call and location of the person who is receiving the call. Therefore, BTS represents a geographical area. Cell of each BTS tower can be calculated using Voronoi tessellation[4] which is a two dimensional region. Figure 2.2 represents BTS coverage and voronoi map of BTS after voronoi tessellation.

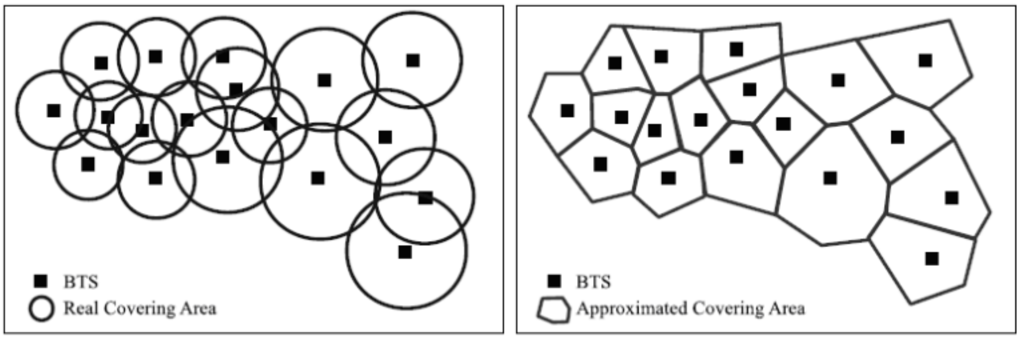


Figure 2.2: BTSs and their coverage (LEFT), approximated coverage after applying Voronoi Tessellation (Right)

Using CDR records, number of properties relevant to a mobile user can be derived. The properties can be loosely categorized into three categories [1][3]. These features were calculated subscriber wise or aggregated on BTS level. After deciding the subscriber's home location or BTS location, district or province wise data can be derived. Following are the three categories of features identified.

- Consumption - describe user's phone call behavior
- Social - social network status of the user
- Mobility - geographic movement of the user

Following are few features (Table 2.2) extracted[1] to build a model which was used to predict socioeconomic characteristics in Sri Lanka. All features were created in BTS level using CDR.

Table 2.2: Features derived from CDR data

Category	Feature	Description
Consumption	Total In	Total number of calls received
	Total Out	Total number of calls made
	Total	Total calls made and received
	Duration In	Average duration of the calls received
	Duration Out	Average duration of the calls made
	Duration	Average duration of the calls
Social	Contact Count	Number of unique contacts (phone calls made/received)
	Contact Rate	Average number of connections made by the user with his/her contacts
	Physical Distance	Average physical distance between user and his/her all contacts
Mobility	Unique Cell Counts	Number of different BTSs visited by a person
	Distance travelled	Distance between each pair of consecutively visited BTS
	Radius of gyration	Distance between home cell and each visited BTS, weighted by frequency of visits
	Maximum Distance	The maximum distance between the BTSs typically visited by a person

In another research[2], CDR were reduced to a graph where cell towers were denoted by nodes and weighted feature values were denoted by edges. The features were only the volume and duration of calls between cell towers because CDR were aggregated on cell wise for making user anonymous.

Since CDR have no information on users' data usage, SMS and MMS usage, it is hard to extract features related customer behavior other than voice usage. Therefore, some researches have extracted only few features like call duration and call volume [6]. Some of them extracted call duration and call volume on subscriber level and few[2] have calculated features on BTS level.

In the research[6], CDR from Ivory Coast and, Demographic and Health Survey (DHS) data were used. Voronoi cells were formed by dividing the map of Côte D'Ivoire into 1,214 cells, in a way that each cell contained an antenna. Then population counts for each cell were derived for each voronori cell and they were used to calculate call variables per capita. The variables were called volume per capita and call duration per capita. In the following figure (Figure 2.3), light circles show the antennas in the partitioned map of the Ivory Coast.

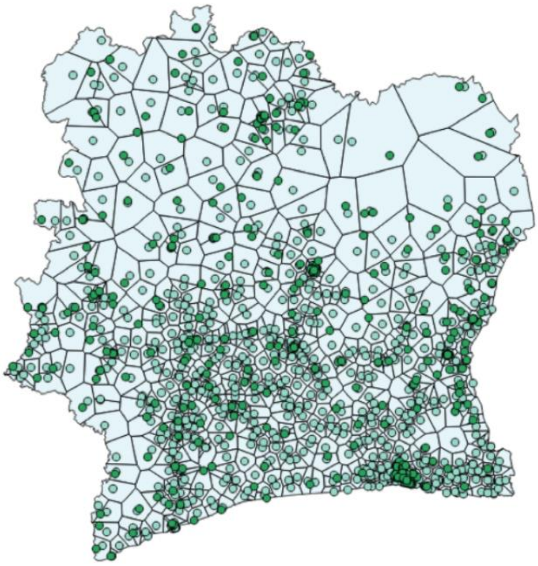


Figure 2.3: Location of Voronoi cells, cell phone antennas (light circles) and DHS clusters (dark circles) in Ivory Coast

2.3 Mobile user home location

When modeling the SEL of a geographical area, the residents of that area have to be identified first. The SEL mainly depends on the behavior of residents who live in a specific area. In order to identify the residents, resident's registry in Divisional Secretariat offices has to be queried. Since SEL is modeling with CDR, residential location and mobile number have to be linked which is an expensive operation.

Since CDR has the cell tower used by the subscriber, who is making the call and cell tower used by the subscriber, who is receiving the call, the location of each subscriber can be identified. Using this location, subscriber home location can be derived. It is much easier and requires low effort compared to finding home location from documents and linking with the mobile number.

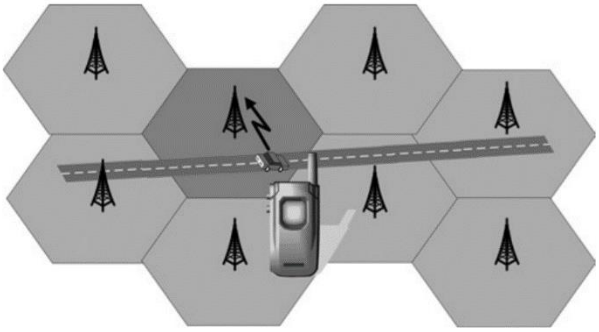


Figure 2.4: Schematic view of a car connected to a base station in a cellular network

Many researches have been carried out, in order to derive home location of the subscriber using CDR. Considering main mobility patterns of the customers, there are two major themes which can be found. They are work-home and home-work characteristics[7]. Most of the work was done to identify the time window to filter above two characteristics. CDR have been used to optimize transportation plans in Colombo, which resulted a time frame showing the commuting time between home and office for mobile users. If a time window is available, home location can be derived using CDR which were generated when the subscriber was at home.

For the accuracy of home location, CDR generated on weekends and holidays were filtered out[7]. The reason for that was people tend to go trips and visit relatives on weekends and holidays. Divisional Secretariat Division (DSD) of the home was calculated by taking the most frequent DSD of BTSs which related to the home DSD in the time window when the subscriber was at home. Home time frame was decided by analyzing the Euclidian distance between hour wise location throughout the hours in the day.

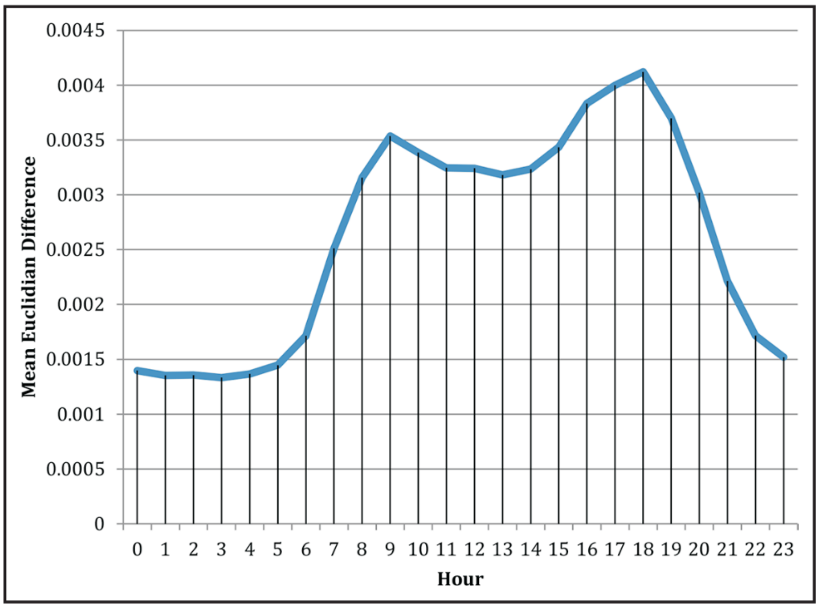


Figure 2.5: Average SIM mobility during hourly time brackets in a single day

After looking at the figure 2.5, it can be identified that people tend to move mostly between 6 AM and 6 PM in a day. People go from home to office or home to school at about 6AM. At about 6 PM, they return to home after work. According to the research done on Sri Lankan citizens, that is the best time window which can be used to distinguish the home and work behavior. Time restrained criteria has given promising results in detecting home location using CDR in France[8]. They have used five methods to define criteria based on activities, distance and time. Activities have been filtered out using time frame which is between 7 PM and 9AM which has given fine accuracy for the home location.

2.4 Modeling socioeconomic level

Feature extraction from CDR will output attributes such as call volume and call duration subscriber wise or geographical area wise. They may be in a summarized manner for each day, week, month or year. SEL data gathered for a given time period is the other data source which denotes the actual economic behavior of the people in the target area. Collecting SEL data requires huge amount of effort compared to CDR data mining.

Therefore, if a relationship can be found between attributes extracted from CDR and SEL data, modeling economic behavior of a specific geographic area is feasible. Various researches have been carried out to model economic behavior of an area using CDR. After identifying the economy related attributes, using Machine Learning (ML) techniques, economy level can be predicted. It will save a considerable amount of cost and will be very helpful for the policy makers.

The following section will summarize the analytical process and ML techniques which have been used in this research area.

2.4.1 Linear Regression

Linear Regression is a Machine Learning technique which is widely used to find the relationship between dependent variables and the explanatory variable. Linear Regression is implemented in predictive analysis by fitting the relationship of variables to a linear equation which is denoted in the following equation.

$$Y = XB + U$$

In this research, it is about predicting values of SEL data using CDR features derived. X is the explanatory variable and Y is the dependent variable.

In multivariate linear regression, Y is a matrix with a series of multivariate measurements and X is a matrix of observations on independent variables. Before attempting to fit a linear model to observed data, it should determine whether there is a relationship between variables of interest. Association of variables is measured by the coefficient of correlation. Coefficient of correlation value is between -1 and 1 indicating the strength of the association of the observed data for the two variables.

Analysis carried out on census data in Sri Lanka and CDR data[1], they have included 14 features in the final regression model. They have used multivariate linear regression. Mobile user population density and 13 other features were used to get the highest performing regression model. Performance was measured by highest adjusted R^2 values. That means the best fitting line for the observed data was calculated by minimizing the sum of the squares (R^2) of the vertical deviations from each data point to the line. Following graph (Figure 2.6) shows how well each CDR feature categories performed in predicting SEL characteristics.

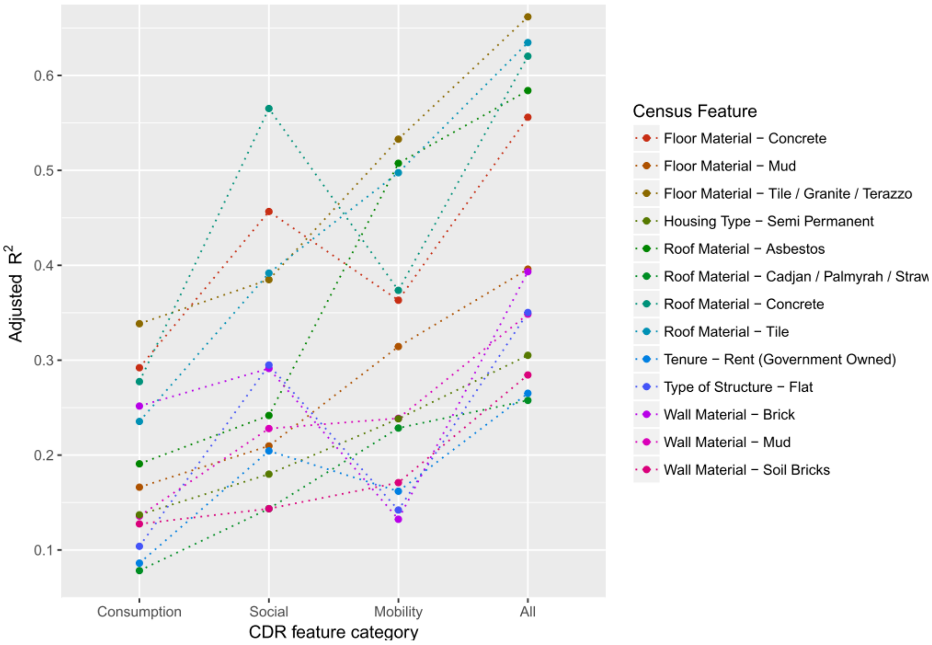
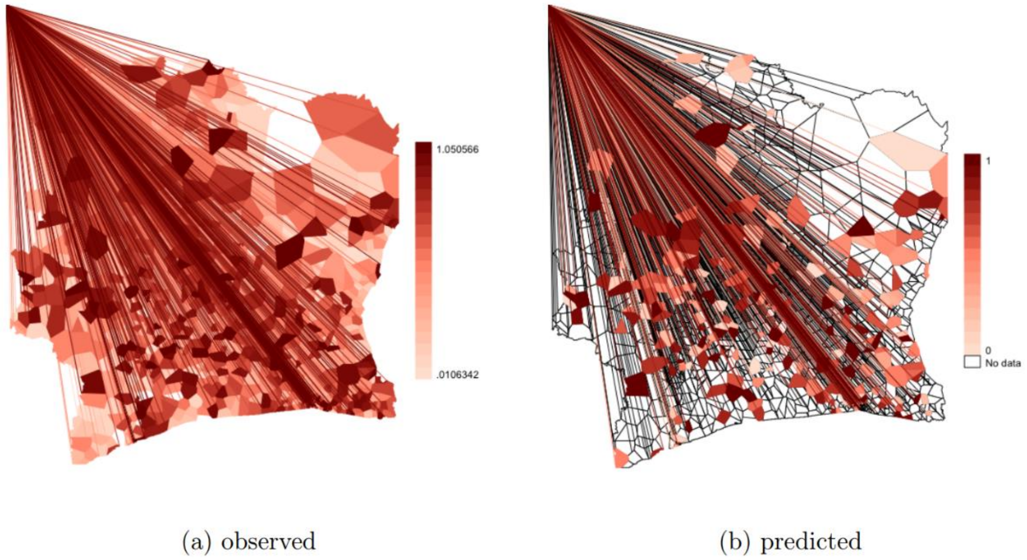


Figure 2.6: Performance of regression models against different variable types in Western Province

The results[1], in Northern Province, suggest that social features extracted from are a promising predictor of the number of households with permanent housing, but a poor predictor of households with semi-permanent structures. In Western Province, mobility can be considered as a very good predictor of the households with tile/granite/terrazzo floors, but a poor predictor for the type of the wall material.

A regression model has been used in predicting MPI (Multi-dimensional Poverty Index) using CDR of Ivory Coast[6]. They have extracted call volume and call duration using CDR and predicted MPI using a regression model. They have used least square method in building the model which predicts the MPI. Because of the noise, they used the model to estimate poverty in sub-prefecture level. There was a strong negative correlation between call volume and MPI score. Also a negative correlation was observed between call duration and MPI score. Using call volume and call duration MPI was predicted then. Following figure (Figure 2.7) shows the accuracy in predicted results.



(a) observed

(b) predicted

Figure 2.7: Observed and predicted poverty level in Ivory Coast

While the coefficients were strongly significant, the R-squared was around 0.3 for the model with headcount as the dependent variable and 0.15 for the models with intensity as the dependent model[6].

2.4.2 Support Vector Machine

Predicting SEL can be reduced to a classification problem. Support Vector Machine (SVM) is a promising Machine Learning technique used in classification problems. This is a supervised learning model which develops hyperplanes using training data. Based on the training labels data, it builds a model to assign the new data point to any of the pre-defined categories.

The SEL can be categorized into a few categories based on the economic level. In a research[4], SEL is categorized into A, B and C labels on BTS level. Then BTS has its feature vector derived using CDR. Following are the features included in the vector which is the input for the classification.

- Number of different BTS towers used (weekly)
- Diameter of the area of influence (weekly)
- Total number of weekly calls
- Closeness of incoming SMS contacts in relation to all communications
- Percentage of incoming SMSs with respect to all incoming communications
- Percentage of SMS contacts with degree of reciprocity 5
- Radius of gyration
- Total distance traveled(weekly)
- Median of total number of calls
- Percentage of voice contacts with degree of reciprocity 2

Using above features, the model was trained to classify new BTS data points. They have used 5-fold cross validation to improve the accuracy. According to the literature, the accuracy was not promising, because most of the A and C labeled test cased were misclassified as B. They have tried different feature selection methods to improve the accuracy. Three of them were maxrel, mRMR-MIQ and mRMR-MID. Maxrel methods have performed better compared to two other methods. Accuracy was lower than 80% and used more features to get more accuracy.

2.4.3 Decision Tree

The decision tree is a supervised machine learning technique which is a tree based mechanism developed using training data. It creates a model with nodes and edges which denotes the decision path. There are many variations in decision tree method. Random forest, Boosted Trees are few types of decision trees which used different approaches in classification.

As discussed in the previous section, random forest has been used in predicting SEL categories[4]. Socioeconomic categories used are A, B and C categories on BTS level. Using training data, they trained a model which can classify a new BTS based on its features extracted from CDR. Tree based regression model has been developed for the poverty prediction in Rwanda[5].

They have built Random Forest models with t trees where three feature selection methods were used. Three feature selection methods were mRMR-MIQ, mRMR-MID and Maxrel. It can be observed that maximum relevance feature selection (Maxrel) has produced results with better accuracy than mRMR-MIQ or mRMR-MID[4]. Following chart (Figure 2.8) shows the accuracy when number of features were changed on different feature selection method.

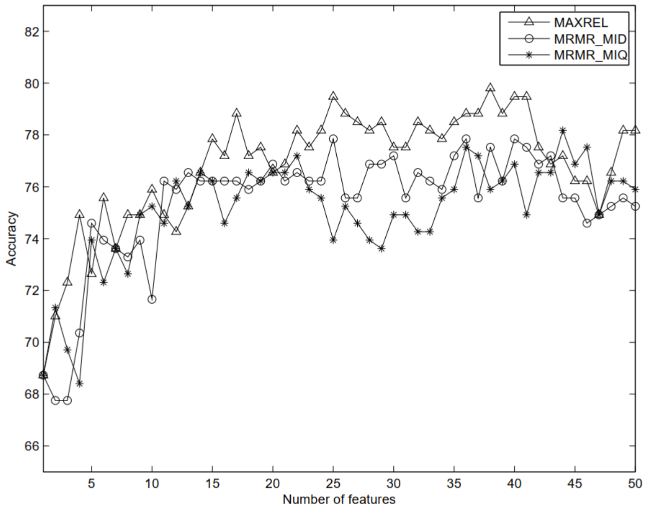


Figure 2.8: Correct classification rate for different feature selection method

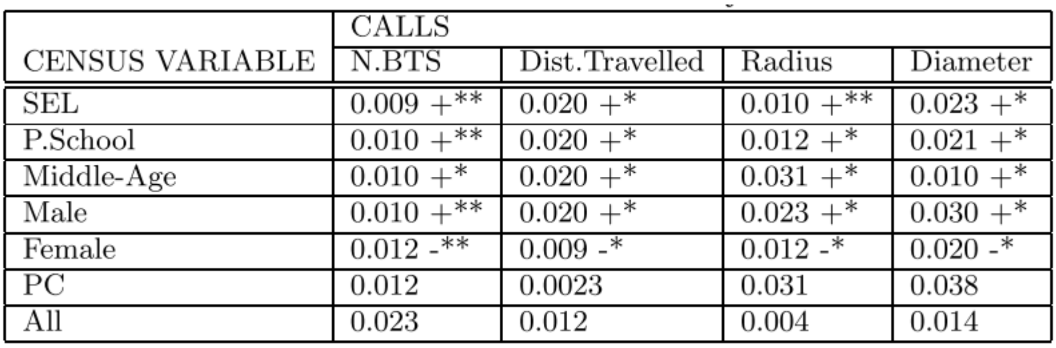
2.4.4 Hypothesis Testing

Hypothesis testing is a statistical inference method where two statistical data sets are compared to make a decision. A hypothesis is proposed for the statistical relationship between the two data sets, the null hypothesis is defined to propose no relationship between the two data sets. Hypothesis tests are used to determine what results in a study would lead to a rejection of the null hypothesis for a pre-specified level of significance.

Using this statistical approach, cell phone usage and census variables have been subjected to ANOVA test. ANOVA tests were used to understand whether it exists statistically significant differences in cell phone usage across different social groups[3]. Social groups were tagged by SEL that ranges from A/B (very high socioeconomic level) to E (very low socioeconomic level) with intermediate values C+, C, D+ and D. Cell phone usage variables were divided into three groups based on their characteristics. There groups were consumption, mobility and social. Call and SMS CDR were used to derive variable values. Mobility variables were only derived from Call CDR.

ANOVA tests have been executed on about three group variables. The relationship between mobility variables were highly correlated with SEL compared to the consumption and social variables as shown in Table 2.3.

Table 2.3: Census Variables and Mobility Variables [3]



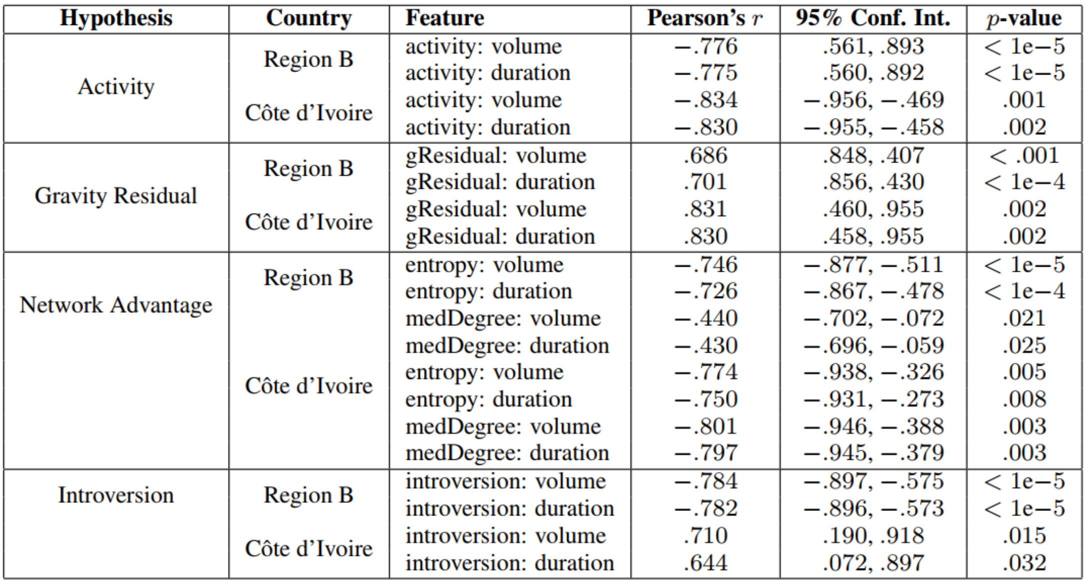
CENSUS VARIABLE	CALLS			
	N.BTS	Dist.Travelled	Radius	Diameter
SEL	0.009 +**	0.020 +*	0.010 +**	0.023 +*
P.School	0.010 +**	0.020 +*	0.012 +*	0.021 +*
Middle-Age	0.010 +*	0.020 +*	0.031 +*	0.010 +*
Male	0.010 +**	0.020 +*	0.023 +*	0.030 +*
Female	0.012 -**	0.009 -*	0.012 -*	0.020 -*
PC	0.012	0.0023	0.031	0.038
All	0.023	0.012	0.004	0.014

Poverty analysis done on CDR, used hypothesis testing on four main categories. Features were extracted on BTS level for Ivory Coast and Region B[2]. The four categories were

- Activity
- Gravity Residual
- Network Advantage
- Introversion

Activity was the level of mobile communication activity in specific areas. This reflected its social and economic activity. Gravity residual was the difference between observed and expected flows between areas. Network advantage was about advantageous position in the network. Then the level of introversion denoted the socioeconomic level. These hypotheses were tested and correlation between variables were measured (Table 2.4).

Table 2.4: Correlations between features derived from mobile phone data and poverty level [2]



Hypothesis	**Country**	**Feature**	**Pearson's r**	**95% Conf. Int.**	**p-value**
Activity	Region B	activity: volume	−.776	.561, .893	< 1e−5
		activity: duration	−.775	.560, .892	< 1e−5
	Côte d'Ivoire	activity: volume	−.834	−.956, −.469	.001
		activity: duration	−.830	−.955, −.458	.002
Gravity Residual	Region B	gResidual: volume	.686	.848, .407	< .001
		gResidual: duration	.701	.856, .430	< 1e−4
	Côte d'Ivoire	gResidual: volume	.831	.460, .955	.002
		gResidual: duration	.830	.458, .955	.002
Network Advantage	Region B	entropy: volume	−.746	−.877, −.511	< 1e−5
		entropy: duration	−.726	−.867, −.478	< 1e−4
		medDegree: volume	−.440	−.702, −.072	.021
		medDegree: duration	−.430	−.696, −.059	.025
	Côte d'Ivoire	entropy: volume	−.774	−.938, −.326	.005
		entropy: duration	−.750	−.931, −.273	.008
		medDegree: volume	−.801	−.946, −.388	.003
		medDegree: duration	−.797	−.945, −.379	.003
Introversion	Region B	introversion: volume	−.784	−.897, −.575	< 1e−5
		introversion: duration	−.782	−.896, −.573	< 1e−5
	Côte d'Ivoire	introversion: volume	.710	.190, .918	.015
		introversion: duration	.644	.072, .897	.032

3. METHODOLOGY

3.1 Introduction

As discussed in the literature review, CDR have been used to analyze poverty or socioeconomic level. Attributes of the subscriber and features of the area covered by a BTS can be extracted from CDR. The features extracted using CDR were evaluated or compared with census or survey data which symbolize the socioeconomic level. It reveals whether features extracted from CDR actually have a relationship with the poverty level.

If there is a relationship with the features extracted from CDR and census data, methodology can be developed to analyze poverty in an area using CDR without collecting data from citizens manually. Collecting socioeconomic data requires a huge amount of effort as well as cost. Identifying the feasibility of using CDR to view the SEL of people in fine grained level is very helpful for real-time decision making and policy making against the poverty.

This section will discuss the methodology to find out the feasibility of using CDR and other telecommunication data sources to analyze and predict the socioeconomic level in Sri Lanka. This will describe the top to bottom approach used in this research.

Types of data sources used in the research, details of socioeconomic data used to evaluate, steps carried out to identify the home location of Sri Lankan citizens, Machine Learning and statistical techniques used will be discussed in detail manner in the following section. This research was targeted in analyzing socioeconomic level covering the whole geographical area in Sri Lanka.

3.2 Socioeconomic Data in Sri Lanka

3.2.1 Official poverty line by district

Department of Census and Statistics (DCS) in Sri Lanka releases official poverty line data monthly. Other than the district level, it has the national level poverty line data. DCS calculates the poverty line using household income and expenditure survey (HIES). Official poverty line is fixed at a specific consumption level. The consumption level is that person who meets certain amount of nutrition intake which is 2030 kilocalories per day[9].

Since all of the population cannot be subjected to the HIES, sample of 20,000 houses is selected for the survey. Carrying out the survey for all the houses in Sri Lanka monthly is practically infeasible. Prices of the consumer goods are retrieved from the Colombo Consumer Price Index (CCPI). Price differences are captured from CCPI monthly and district wise price changes are tracked. The calculating poverty line incorporates spatial price indices. The following table (Table 3.1) is a sample of the poverty line in few districts of Sri Lanka. The calculation inputs the household food and nonfood consumption and expenditure data for a person.

Table 3.1: Official Poverty line from 2019 January to 2019 June for few districts in Sri Lanka

	National	Colombo	Gampaha	Kandy	Mathale	Galle	Mathara
2019 Jan Rs.	4752	5158	5001	4923	4835	4692	4541
2019 Feb Rs.	4730	5134	4978	4900	4812	4670	4520
2019 Mar Rs.	4722	5126	4970	4892	4804	4663	4513
2019 Apr Rs.	4737	5142	4986	4908	4820	4677	4527
2019 May Rs.	4812	5223	5064	4985	4895	4751	4598
2019 Jun Rs.	4856	5272	5111	5031	4941	4795	4641

Since poverty line is adjusted monthly, CCPI is monthly updated and spatial price indices are obtained monthly district wise. The simplest meaning of the poverty line is the minimum expenditure per person per month to fulfill the basic needs. The poverty head count can be defined as the number of people lives under the national poverty line. Poverty line data can be used by government to make policies to remove poverty in the society.

Since poverty data are released monthly, they can be stored and used to analyze how the poverty line changes with time in district level. Following diagram (Figure: 3.1) shows how the poverty line changes in last two years in Sri Lanka for each district.

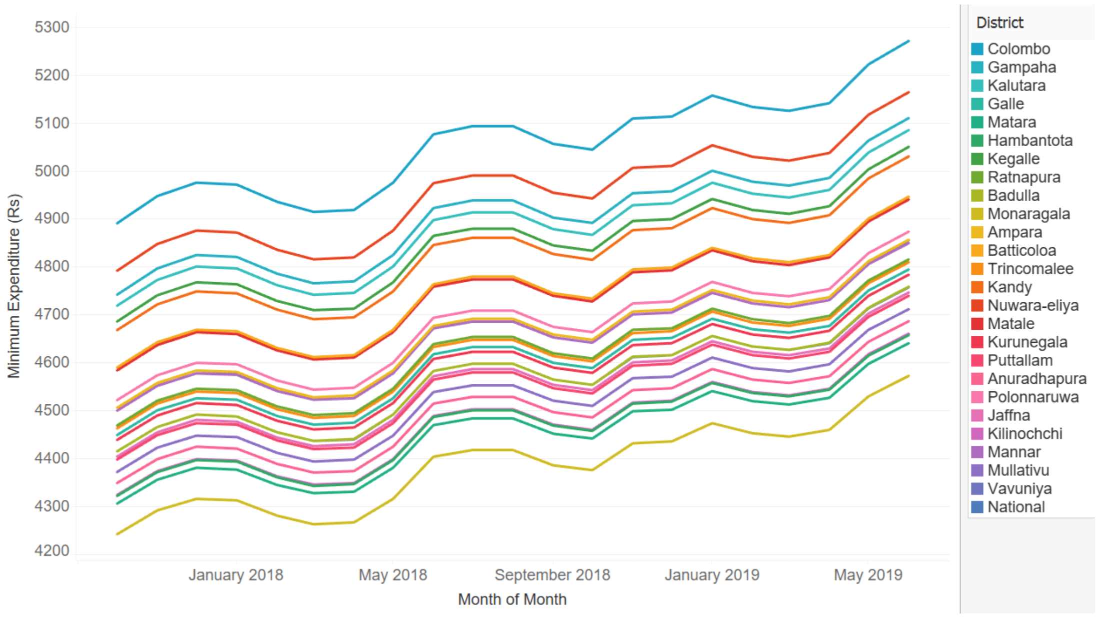


Figure 3.1: Official Poverty line by district for last two years in Sri Lanka

There are 25 districts in Sri Lanka. After looking at the poverty line over two years for each district, there is a similar behavior. There is an equivalent trend over the time for different districts. The reason for this behavior may be taking a sample of 20,000 households to calculate the poverty line. If a ratio between the poverty line of two districts calculated for the period stated in the figure 3.1, it has the same ratio over the time. The economic behavior of each district can be similar because Sri Lanka is a small country.

3.2.2 Socioeconomic Index data

Socioeconomic index is a numerical figure which can denote the economic status of a group of people or an area. Calculating socioeconomic index requires a huge amount of data which were gathered from a census or a survey. It requires a huge amount of effort and time. The socioeconomic index has been developed using the 2011/12 census data on Grama Niladhari Division (GND) wise [10].

Department of Census and Statistics has conducted the 14th population and housing census of Sri Lanka in 2012. It was held with the help of a large number of government officers covering all districts. They have collected information on demographic, social and economic details of households[11]. It is a gigantic process which consumes huge amount of effort and human hours.

The socioeconomic index was created using demographic and economic details of the people and household. Since 2011/12 census data are only available in GND level, index was calculated at GND level. Principal Component Analysis has been used to generate the numerical index. Variables used were categorized into two groups. They are population and household variables. Following Table (Table 3.2) shows the variable types used in index generation.

Table 3.2: Data Set and respective categories used for index

Data Set	**Category**
Household	Cooking Fuel
Household	Floor Material
Household	Housing
Household	Lighting
Household	Roof Material
Household	Structure
Household	Tenure
Household	Toilet Facilities
Household	Wall Material
Household	Waste Disposal
Household	Water Source

Population	Age
Population	Education
Population	Employment

Different categories have different variable values. Table 3.3 shows a few of the variables for sample categories. The original categorical variables surveyed at the household level have been converted to binary variables and aggregated for each GND.

Table 3.3: Few of the Category and values used for Socio Economic Index

Category	**Variable**
Education	Primary Secondary O Level A Level Degree & Above No Schooling
Employment	Employed Unemployed Economically Inactive
Gender	Male Female
Cooking Fuel	Firewood Kerosene Gas Electricity Sawdust / Paddy husk Other
Housing	Permanent Semi-permanent Improvised Unclassified

PCA can be used to reduce the number of variables in to a less number in a data set. It is a statistical method. In mathematical terms, from an initial set of n correlated variables, PCA creates uncorrelated indices or components, where each component is a linear weighted combination of the initial variables[12].

Following are the steps carried out to build the socioeconomic index for Sri Lanka GND level.

1. Curate the dataset to remove variables that are either redundant or non-indicative of socioeconomic status.
2. Normalize the dataset with respect to each category within each GND.
3. Standardize each variable.
4. Run PCA on the standardized data set.
5. Extract the weights given by the first principal component.
6. Multiply the standardized dataset by these weights.
7. Sum the above scores for each GND to get the socioeconomic index.

After generating the socioeconomic index, it has been visualized in GND level. The minimum index value is -14.73 and maximum index value is 15.43. Following Sri Lankan map (Figure 3.2) shows how the index has distributed geographically. Other Grama Niladhari divisions have the respective values given by the final step.

In order to generate the socioeconomic index for Sri Lanka, only few demographic details were captured from the census data. Income or salary details were not included in the index generated features. For the employment category, if the profession of the citizens (doctor, engineer, teacher, etc.) was included, the index will be more valuable and accurate.

When the distribution of the population is considered according to provincial level on census data, two third of this total population lives in the four provinces, Western (28.7 percent), Central (12.6 percent), South (12.2 percent) and North-western (11.7 percent)[11]. Attributes such as having electricity, having luxury household items were not captured. The main reason for the loss of the above data is the limited resource availability in the census. After looking at how economy index has distributed whole over Sri Lanka (Figure 3.2), it shows that people with higher socioeconomic level are the residents in urban areas.

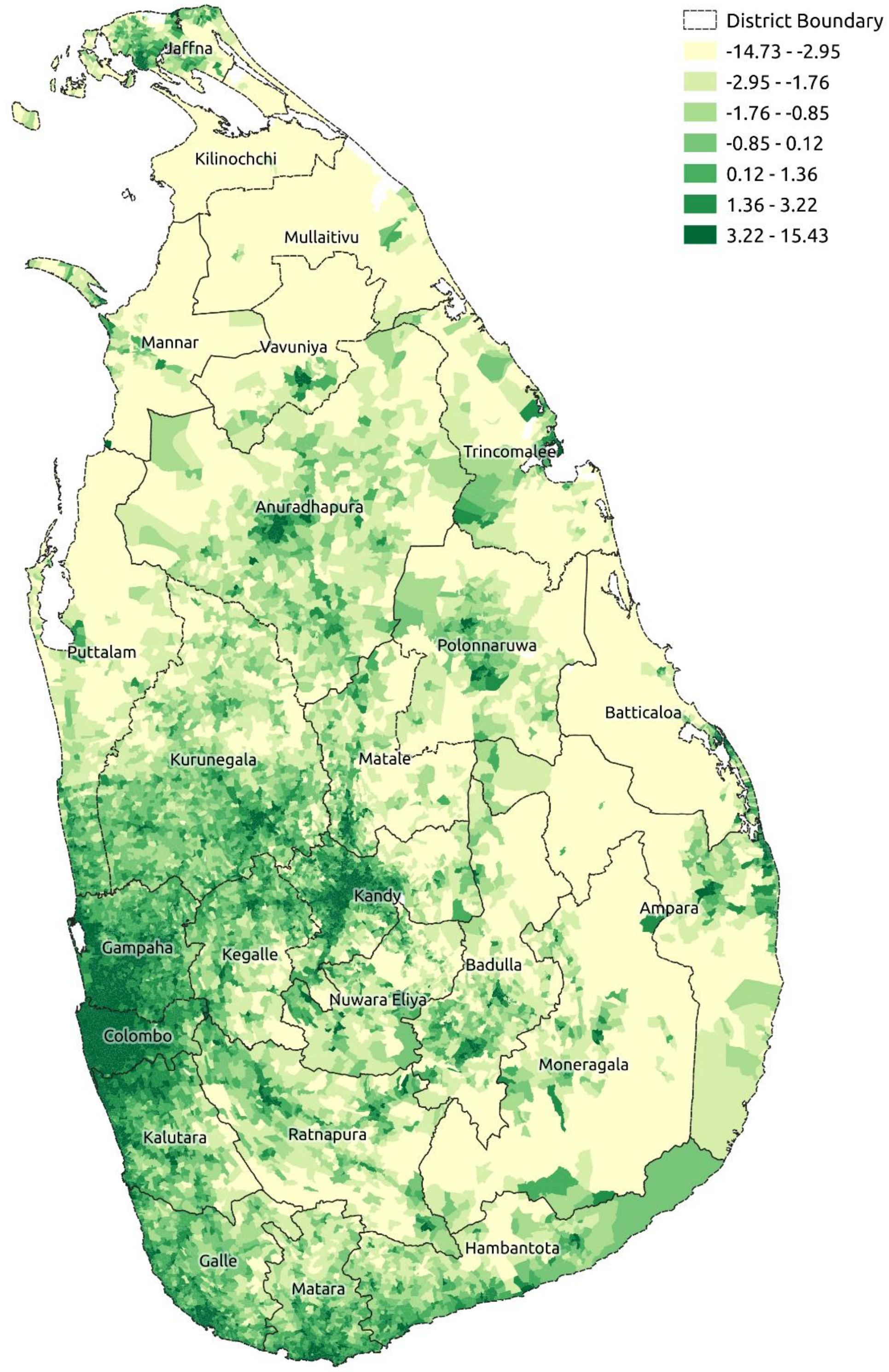


Figure 3.2: Socioeconomic Index distribution in Sri Lanka which was calculated using 2011/12 census data

3.3 Data sources for feature extraction

There are Peta bytes of data resides in telecommunication systems. The main data source is CDR. Other than CDR, there are transaction data like recharge information, bill data and device information. These data sources can be used for many tasks since they characterize customer behavior. Storing and processing huge amount of such data is the main challenge.

3.3.1 Voice CDR

As discussed in the first chapter, a Call Details Record (CDR) is the evidence generated by the telco systems when a telephone call was originated. Voice CDR contains the following key attributes of the call.

- Phone number of the subscriber who is making the call (a_number)
- Phone number of the subscriber who is receiving the call (b_number)
- BTS id connected to the a_number
- BTS id connected to the b_number
- Event started date and time
- Duration of the call in seconds
- Flag to indicate whether it is an incoming or a outgoing call
- Call charge
- Flag to indicate whether it is roaming or not
- Flag to indicate whether it is postpaid or prepaid
- IMEI or identifier for the device
- SIM number

Other than the above attributes, there are hundreds of columns in the CDR. Some of them are related to the billing criteria, network routing, service quality and system configurations. There are two types of CDR generated in the systems. One type is CDR which are generated before the billing. They are named in different terminologies. In this case it is called MSC CDR. The other type is one which is generated after billing.

MSC CDR can be accessed in real-time because they do not contain billing attributes. Therefore, the size of the CDR is low compared to the CDR after billing.

More than one billion voice CDR are originated in the system. In order to analyze them, CDR must be stored. Since the volume and velocity of the data are high, they were stored in a big data system based on Hadoop file system. Various types of indexing and compressing mechanisms were used to save the storage capacity in the Hadoop system.

Flume was used to inject CDR into the Hadoop file system. The process was scheduled to fetch CDR hourly from network systems. CDR can be considered as structured data. They have a specific schema and mapped into a table. Therefore, voice CDR is stored in a Hive table in Hadoop. As a result of that, Voice CDR can be processed using MapReduce jobs or using Spark. In this research, Spark was used to process voice CDR stored in Hadoop as Hive tables.

3.3.2 Data CDR

Data CDR is generated when subscribers use broadband data using their mobile devices. A CDR is created when a data session is started in order to persist the transaction. There are many important key features which can be found in the Data CDR produced in telecommunication network systems. Few of them are stated below with some details.

- Phone number of the subscriber who consumes data
- Data session started date and time
- Amount of data uploaded in bytes
- Amount of data downloaded in bytes
- Total amount of data served in bytes
- BTS id which served the data
- Flag to indicate the 2G/3G/4G signal type

- Data amount used as bundle in bytes
- Charge for the data session
- IMEI of the device
- SIM number used

Other than the above columns, Data CDR have hundreds of data attributes stating other information. Few of them are related to the network routing, service quality, billing configurations and service types such as roaming. More than one billion records were generated within a day as a result of customers' broadband data consumption.

Similar to voice CDR, there are two types of CDR generated in the subjected telco network system. One of them (MSC CDR) is generated before billing takes place and the other is generated after billing. MSC CDR can be accessed in real-time and they were used in this research.

Data CDR is considered as a structured data source which has well defined set of attributes. Data CDR has the three properties of Big Data which are huge in volume, velocity and variety. Storing and processing this amount of data is challenging. Data CDR were stored in a Hadoop data lake similar to voice CDR. It can manage Giga bytes of data flowed within a day.

Since Data CDR has a structured format, they were stored in a Hive table. Compressing and indexing techniques were used to store CDR. Parquet compression technique was used to optimize the storage capacity and partitioning was used to index the storage. Because of partitioning, CDR can be accessed without high latency. Flume was used to fetch the data CDR daily and put the files in Hadoop file system. Flume jobs were scheduled to get executed a few times a day. In order to process Data CDR, Map Reduce or Spark jobs can be used. Spark jobs were used to preprocess the CDR because of the high performance.

3.3.3 Recharge data

Majority of Sri Lankans use Prepaid connections as their main mobile connection. There are few reasons for the popularity of the Prepaid connection type. Simplicity, real-time charging and instant recharging facility can be considered as the main benefits of using Prepaid mobile connection. Prepaid subscribers use reloads, recharge cards, recharges from the web and money transfers to top up their account balance.

In this research, we have used recharge information from the subscribers. Recharge data are generated in telco systems similar to CDR. They contain important attributes which can characterize the behavior of the subscribers. Following are few attributes which can be found in recharge data.

- Mobile number of the subscriber who recharged the account
- Recharged amount in cents
- Recharged date and time
- Account balance before recharging
- Account balance after recharging
- Flag to identify recharge method

Other than the above fields, recharge data contain the account details of the transaction, transaction statuses, results such as bonus data and voice minute distribution and system configurations. There are hundreds of columns in recharge CDR which were used for billing purposes in telco systems.

Compared to voice and data CDR, recharge data are not in high volume. They can be stored in a transactional database (RDBMS) but recharge data are produced in telco systems similar to CDR. They are stored in the Hadoop based file system for analytical purposes. They are stored in hive tables which store structured data. Daily jobs were scheduled to fetch recharge data and put them in Hadoop. Flume was used to for the above process.

3.3.4 Postpaid bill data

Postpaid connections are not commonly used as Prepaid connections. But a considerable percentage of the population use postpaid connections. One of the main advantages of using Postpaid connection is the ability to consume an uninterrupted service. As discussed earlier, Prepaid connection gets charged on the fly. Postpaid customers are charged on the fly, but their bills are posted once a month. They have to pay the minimum in order to get an uninterrupted service.

Postpaid bill data denote the subscribers' usage pattern. There are important attributes which are included in the bill data. They are stored in the transactional database (RDBMS) in the telecommunication domain. Following are few of the important attributes available in Postpaid bill data.

- Unique identifier for the bill
- Customer account number
- Customer mobile number
- Total of the bill
- Bill cycle start date
- Bill cycle end date
- Payment due date of the bill
- Bill generated date
- Bill items
- Charges for the bill items

Since bill data are stored in an RDBMS, data have to be transferred to a data lake or data warehouse in order to execute analytical workloads. In this research, customers' bill data were used. They were stored in the Hadoop platform using Sqoop. Since bill data are structured data, they were stored using Hive tables. Sqoop jobs were scheduled daily to fetch the newly added bills in the Postpaid billing system. Since they were on the Hadoop file system, data can be processed using MapReduce or Spark jobs.

3.3.5 Postpaid payment data

Postpaid customers receive their bills on their consumption as discussed in the previous section. After receiving bills monthly, they have to pay bills for uninterrupted service. Payments can be done online, at service outlets, at banks or at retail shops. Payment details are stored in relational database as transactions. They are not Tera bytes of data in volume like CDR. Following are few attributes which can be found on payment data records.

- Unique identifier for each transaction
- Account number of the Postpaid connection related to the payment
- Payment amount
- Payment date and time
- Payment type
- Channel of the payment done
- Narration related to the payment

Payment data are structured data resides in RDBMS. They can be stored in a data warehouse for analytical purposes. In this research, they were stored in a Hadoop file system using Hive tables. Sqoop was used to fetch data daily to the data lake. Indexing and compression techniques were used to optimize the storage and data processing capability. Since data were stored in Hadoop, MapReduce or Spark can be used for data processing.

3.3.6 Customer details

Customer details can be considered as the most important in a telecommunication system. They are most sensitive data which can be used to link mobile connection to the person. They were stored in RDBMS as structured data. There are many important attributes which can be found in customer details. Customer details are entered into the system when someone is buying a new connection. Following are some key attributes observed on customer details.

- Unique identifier for each row
- Mobile Number of the subscriber
- Account number of the subscriber
- Connection started date and time
- Connection disconnected date and time
- Connection type of the subscriber (Prepaid or Postpaid)
- Billing address of the subscriber
- Permanent address of the subscriber
- Social id of the subscriber
- Birthday of the subscriber
- First name of the subscriber
- Last name of the subscriber
- Employment of the subscriber
- Employee name of the subscriber
- Employee address of the subscriber
- Email address of the subscriber
- Contact number of the subscriber

Other than the customer details, dealer details, SIM numbers can be found in the customer detail data. Using social id of the customer, customer demographic details can be extracted. For example, gender and birthday can be extracted if the subscriber has given the NIC. These details can be updated on the customer request or in events such as connection termination.

Since these data are in RDBMS, data have to be transferred to a data warehouse for analytical purposes. Sqoop was used to fetch data and stored in the Hadoop file system. Data were persisted in Hive tables. Spark or MapReduce is used to process data. In this research, Spark was used.

3.4 Usage of big data technologies

There are more than one billion records which were generated as a result of voice calls made by customers within a day. Because of the data consumption, more than one billion of data CDR records were generated as well. To fetch and store those data, Data warehouse or data lake with sufficient storage capacity is needed. Big data platform is the ideal solution for such a use case.

Big data has the capability to store high volume of structured or non-structured data. Other than storing, it has tools to process and summarize the huge volume of data coming in high velocity. Following section discusses the tools used in this research to handle CDR in big data perspective.

3.4.1 Hadoop as a CDR storage

CDR data cannot be stored in a RDBMS because of the huge volume and velocity of the data. Therefore, a big data platform is used to store CDR. Hadoop is used to store data. Hadoop is a framework which can support big data storage and processing. CDR is stored in Hadoop Distributed File System (HDFS). It uses a distributed storage mechanism. MapReduce is available in Hadoop for data processing. YARN is used to manage resources in Hadoop.

HDFS is a cluster of nodes where there are two types of nodes. Following are the types of nodes in HDFS and their responsibilities. HDFS cluster can be configured with high availability with more node addition with roles.

- Data Nodes – Store data with replications in each node
- Name Nodes – Keep track which node has which data

Following is an architecture diagram (Figure 3.3) showing how the nodes are associated. A name node can handle multiple data nodes. Other than CDR, data from

RDBMS like customer details and billing data were stored in the same HDFS layer for the ease of processing.

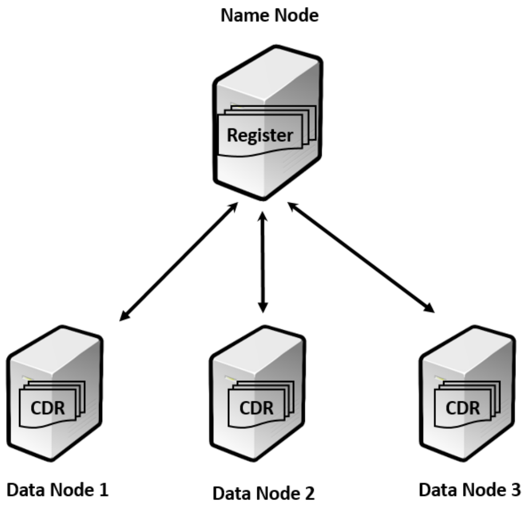


Figure 3.3: CDR stored in HDFS cluster

3.4.2 Data transferring

3.4.2.1 Flume

Apache Flume is used to move large amount of data in a reliable manner. In this research, Flume was used to inject voice and data CDR into Hadoop. Flume has a source, sink and channel. Source is the server location where CDR is first taken from. Sink is the HDFS location where data is finally stored. Flume takes data through the channel and stored in the target location.

3.4.2.2 Sqoop

Sqoop was used in this research to store transactional data in Hadoop as Hive table. Sqoop can fetch data from the RDBMS. It supports SQL and it fetches data from transactional databases and populates tables in Hive. Sqoop jobs were scheduled to import data daily. Capturing changed data was challenging in this task. Billing data, customer details and other required tables were transferred to HDFS using Sqoop. It is very efficient to use Sqoop in data transferring to Hadoop.

3.4.3 Data processing using Spark

After storing all the data sources including CDR in HDFS, data has to be processed to extract relevant features. If conventional data processing tools such as Java program is deployed, it will take a huge amount of time to get an output. In such scenarios, big data processing tools have to be used. There are two popular data processing techniques available. They are MapReduce and Spark. Following table (Table 3.4) shows a comparison between these two tools which support parallelize processing.

Table 3.4: MapReduce and Apache Spark comparison

MapReduce	Spark
For batch processing	For batch processing and real-time processing.
Slower because of disk I/O	Faster because of in memory processing
Low cost	High cost because of large memory
Highly scalable	Highly scalable
Have ML capabilities	Have rich ML capabilities
Compatible with all file types	Compatible with all file types
Bit complex to use	Simple to use because of available rich APIs

Because of the performance and ease of use, Spark was used to process CDR data over MapReduce. Spark is a cluster computing framework which supports parallel data processing.

Apache Spark is fault tolerant and much faster than MapReduce in data processing. It uses in memory data processing techniques which give higher performance. In the perspective of a developer, it exposes APIs in Java, Scala and Python. Spark SQL is also available for data querying. Streaming capabilities and Machine Learning capabilities are important factors for selecting Spark.

3.4.4 Overall architecture of CDR processing

Following high level architecture diagram (Figure 3.4) shows the technical aspects of CDR storage and processing.

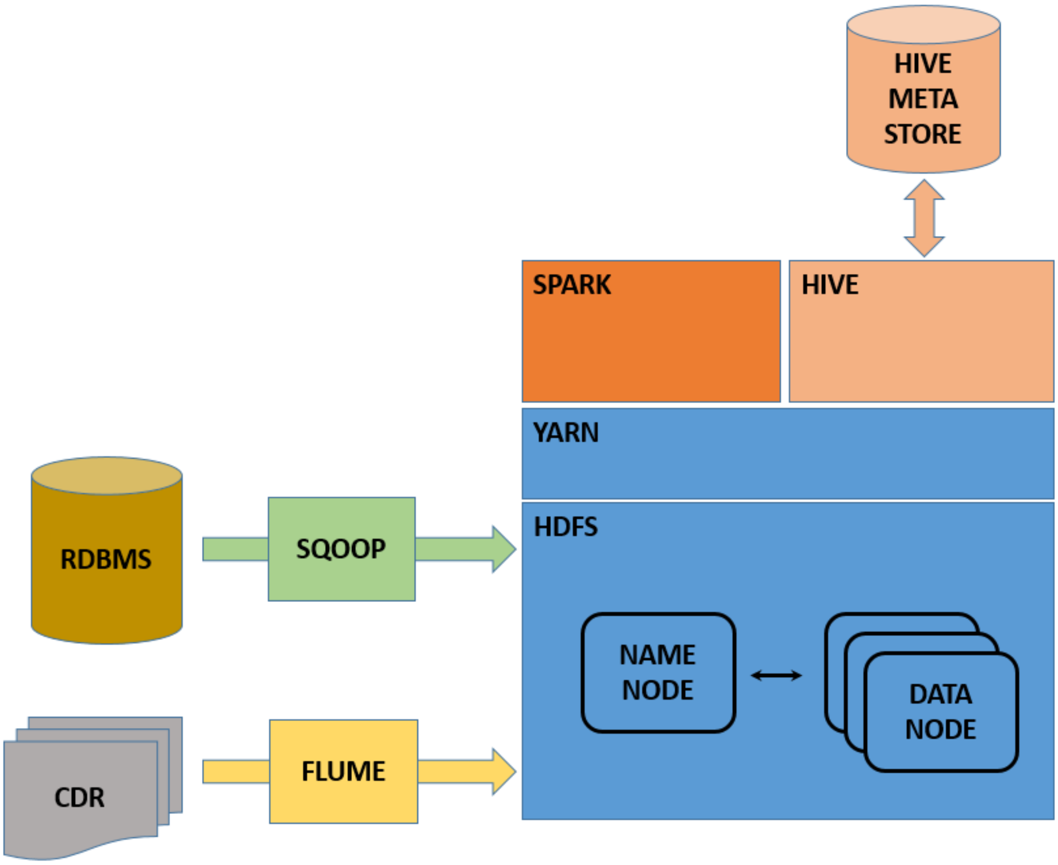


Figure 3.4: High level architectural diagram of the analytical platform

As discussed in the above sections, Flume and Sqoop were used to fetch data to HDFS layer daily. YARN serves as the resource negotiator in read and write operations in the big data platform. Spark is the technology used for processing huge amount of data in parallel. Hive was used to define schema on top of the HDFS layer.

3.5 Extracting the home location

This research was focused on analyzing economic conditions of a geographic area using CDR data. Geographical area can be defined in coarse grained level. This research has drilled down to Divisional Secretariat Division (DSD) level where the second smallest governing entity in Sri Lanka. In order to touch DSD level economic aspects, home location of the subscriber has to be derived.

In the literature review, research was carried out to find the home location of Sri Lankan citizens[7]. They have identified two main behaviors using CDR data. They were home to work and work to home occasions. There is a prominent difference of the call pattern which indicate when subscribers are heading to home or to work. This research has used that technique to derive the home location of the subscriber.

The home location was calculated using MSC Voice CDR. According to the previous researches, a time frame was used to capture the CDR of the subscriber when he was at home. Residents tend to stay at home after 6 PM of the day after work. This was decided after looking at the majority of the population. In Sri Lanka, government offices are closed at 4.15 PM on the day. The work normally starts at 8 AM on the day. Therefore, the majority of the residents start to move to their work place after 6 AM. Then this time window was used to filter out the CDR when residents were at home.

Time frame of 6 PM to 6AM was used to capture the CDR when subscribers were at home. The corporate connections were filtered out when calculating the home location. The cooperate connections were mainly used as the secondary mobile connection by most of the subscribers. That was the main reason for filtering out the cooperate connections.

In this research, home DSD and home district of the subscriber were used. CDR records have the id of the BTS only. Then separate mapping has to be used which has the BTS id and related DSD, district. Using the BTS ids derived using CDR and mapping, home DSD and home district can be found.

The home location was calculated in two steps. In the first step, home location on the weekend days was calculated per each connection. Then a home location per each person was calculated in the weekdays. The home DSD was marked as the weekend home location. If the weekend home location was not available, the weekday home location was used as the subscriber's home location. Home location was derived using voice calls in 18 months. Following diagram (Figure 3.5) gives a summary of the process used in deriving subscriber home DSD and home district.

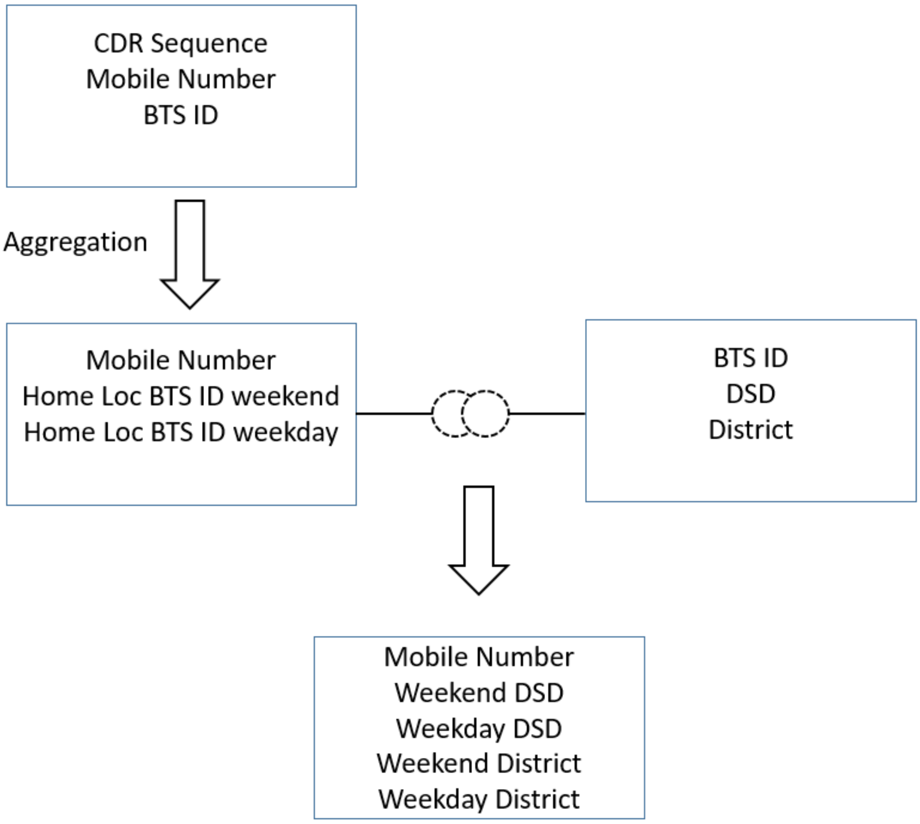


Figure 3.5: Data extracting process in deriving home location

In this research most frequent location was used to extract the home location of the subscribers. The weekend was considered first because residents tend to stay at home most of the weekend at night. On the other hand, holidays were disregarded because people go trips on holidays which will decrease the accuracy. Since 18 months' data were used, it took a huge amount of processing power and time. Following map and diagram (Figure 3.6 & Figure 3.7) shows an overview of the results in extracting home district of the subscribers.

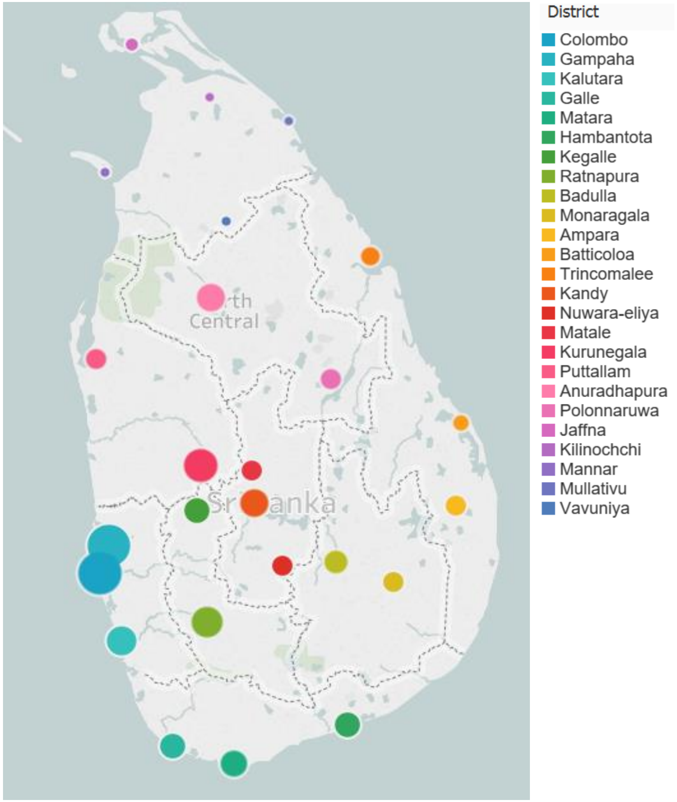


Figure 3.6: Geographical view of subscribers' home district distribution in Sri Lanka

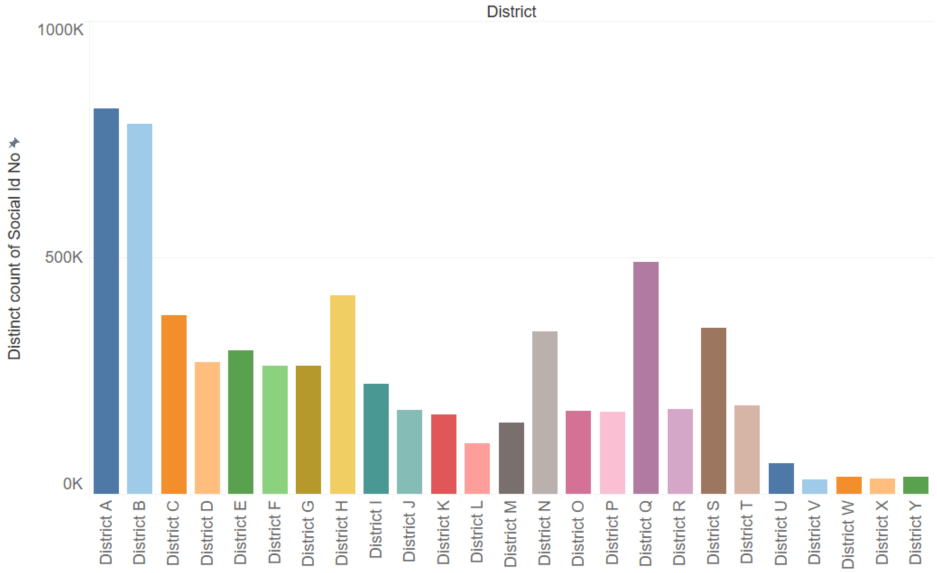


Figure 3.7: Subscribers' home district distribution in Sri Lanka

3.6 Feature extraction from data sources

3.6.1 Average outgoing call volume and average outgoing minute usage

Using MSC voice CDR, outgoing calls were filtered to calculate the features. Incoming call attributes were not calculated because, incoming calls were not charged in Sri Lanka. Therefore, incoming call usage would not reflect socioeconomic level of a subscriber. Outgoing call attributes were extracted in two approaches. They are

- Average outgoing call volume and average outgoing minute usage per subscriber, monthly district wise
- Average outgoing call volume and average outgoing minute usage for one month in DSD wise

For district wise outgoing call feature extraction, CDR of 2 years were used. District wise and DSD wise averages were calculated using the home DSD and home district of each subscriber. Following (Figure 3.8) was the process used in outgoing call feature extraction.

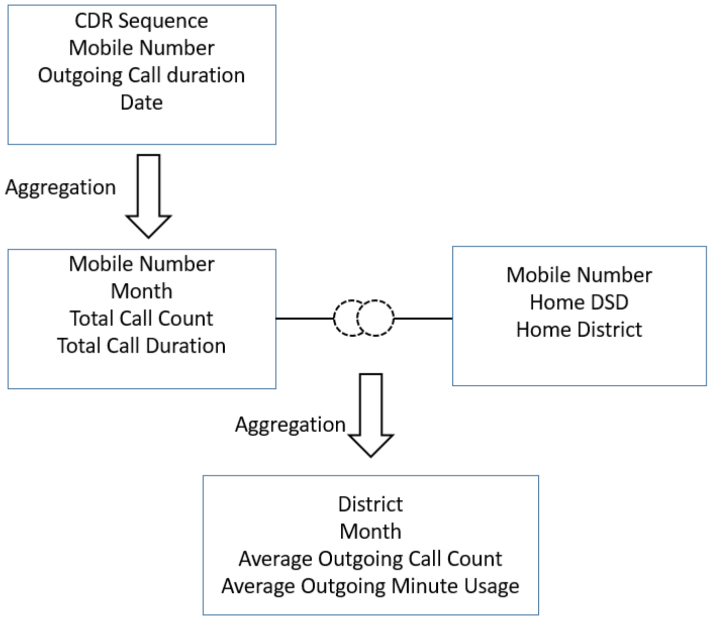


Figure 3.8: Outgoing Call feature extraction process

It was a huge amount of data processing to calculate average minute usage and call volume per person monthly. District wise average minute usage and DSD wise average minute usages (Figure 3.9 & Figure 3.10) are shown in the following figures.

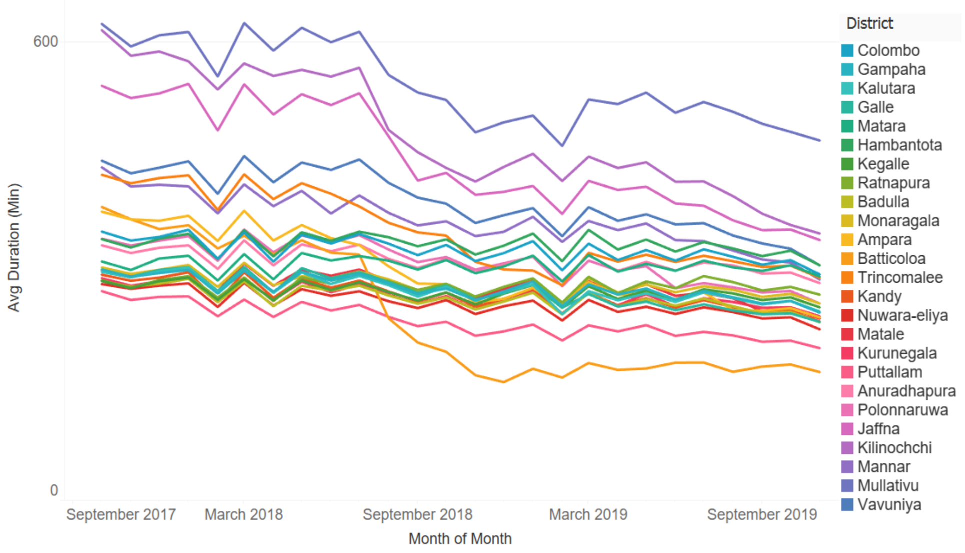


Figure 3.9: Average monthly outgoing minute usage district wise

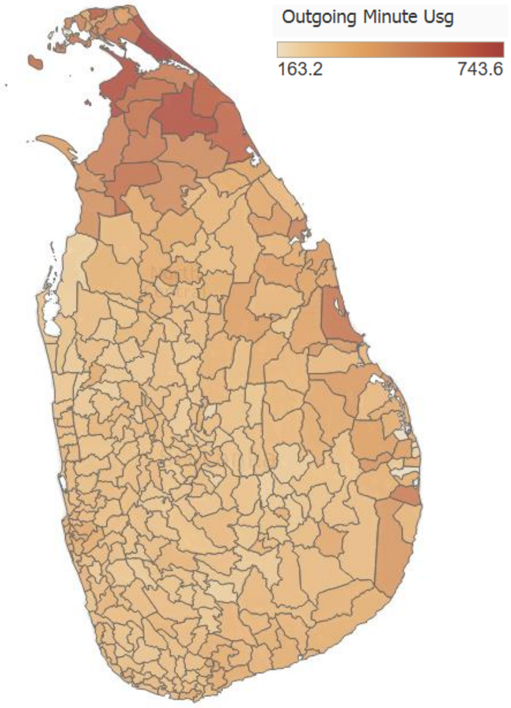


Figure 3.10: DSD wise average outgoing minute usage

3.6.2 Average data usage

Similar to extracting outgoing call features, data usage features were extracted using MSC data CDR. In Sri Lanka, broadband data were charged accordingly. People with less expenditure power limit their data usage as much as possible. Therefore, average data usage was calculated in two approaches. The approaches are mentioned in the following.

- Average data usage per subscriber, monthly district wise
- Average data usage for one month in DSD wise

The procedure was summarizing subscriber wise data usage monthly first. Then, using the home DSD and home district data, district wise and DSD wise average was generated using Spark. Following (Figure 3.11) is the resulted district wise average data usage after using data CDR in more two years. The data processing procedure was similar to extracting voice outgoing call features. But the results had different behavior compared to outgoing minute usage.

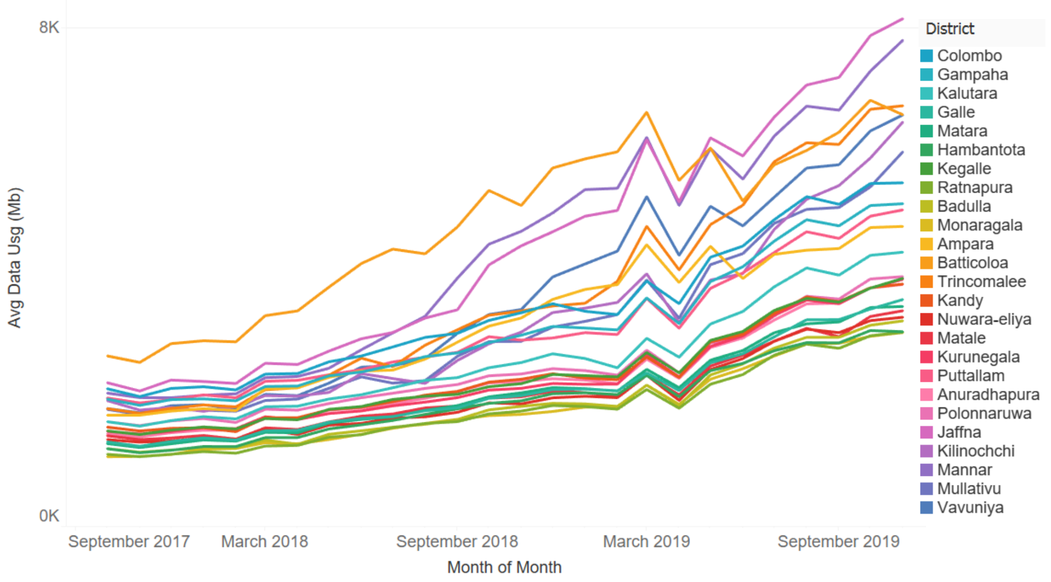


Figure 3.11: Average monthly data usage district wise

Following Map (Figure 3.12) shows the DSD wise average data usage results. Average data usage for each DSD was calculated using only data CDR of one month.

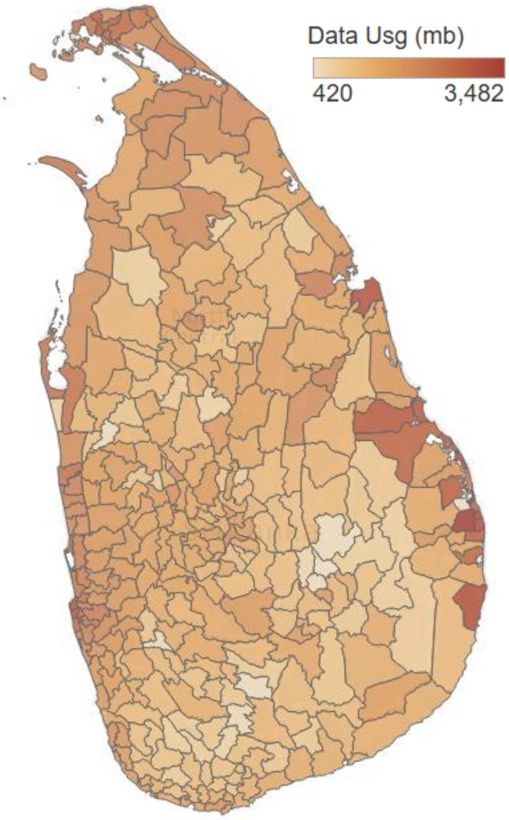


Figure 3.12: DSD wise average data usage

3.6.3 Average recharge amount

Prepaid users recharge their account using recharge cards, reloads, online top up and etc. Subscribers make a recharge when they have low credit or money is on their hands. Recharge amount can have a relationship with people's socioeconomic status. In this research, the average recharge amount was calculated monthly district wise as well as DSD wise. Average recharge value was calculated to get following attributes.

- Average recharge value per subscriber, monthly district wise
- Average recharge value for one month in DSD wise

Similar data preprocessing technique which was used in data usage feature extraction was used in this research. Recharge CDR stored in Hadoop for more than two years were subjected to the processing. Following map and diagram (Figure 3.13 & Figure 3.14) shows how the monthly recharge average behavior looked like.

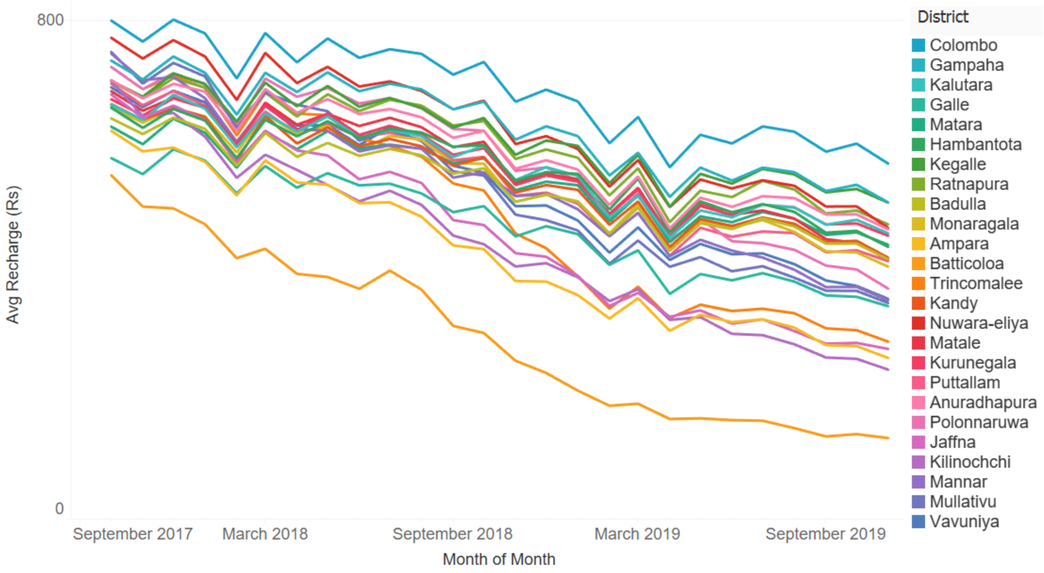


Figure 3.13: Average monthly recharge amount district wise

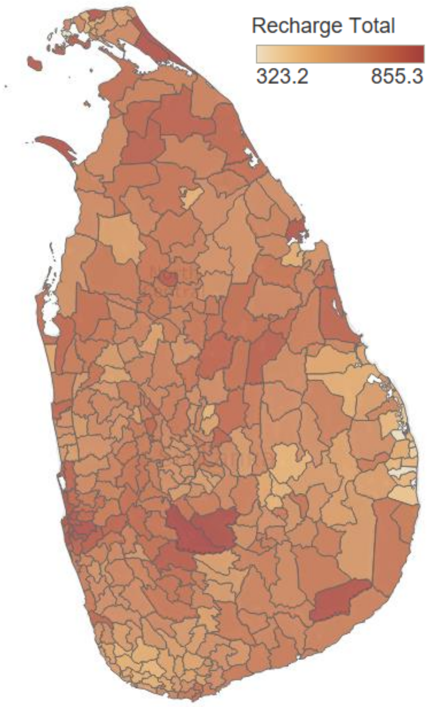


Figure 3.14: DSD wise average recharge amount

3.6.4 Average Postpaid bill amount

Postpaid customers receive their bill monthly stating the bill amount they should pay to continue uninterrupted service consumption. The Postpaid bill value may have a relationship with the customer socioeconomic status. In this research, average Postpaid bill value was calculated to analyze in two approaches. They are

- Average Postpaid bill amount per subscriber, monthly district wise
- Average Postpaid bill amount for one month in DSD wise

Postpaid data were stored in transactional database as discussed in previous sections. Therefore, bill data transferred to the Hadoop platform were summarized using Spark. Similar methodology compared to outgoing call processing, was followed to extract average Postpaid bill amount district wise and DSD wise. Since the data volume was lower compared to CDR data, processing power utilization was lower in this task. Following figure (Figure 3.15) shows how the district wise Postpaid bill average varies monthly.

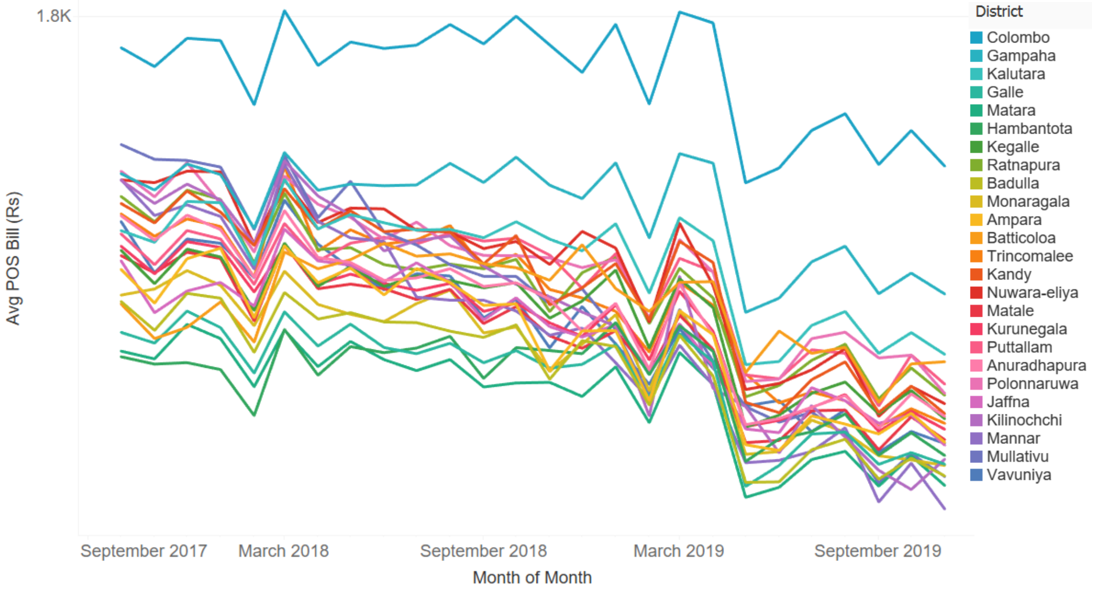


Figure 3.15: Average monthly Postpaid bill amount district wise

Similarly, following map (Figure 3.16) shows the DSD wise average bill value.

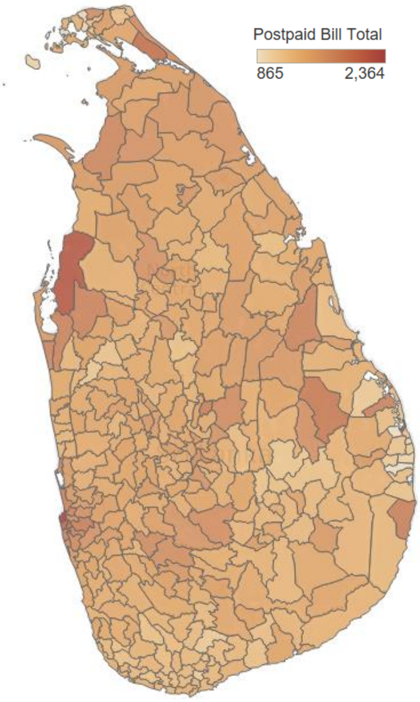


Figure 3.16: DSD wise average Postpaid bill amount

3.6.5 Average Prepaid loan amount

Prepaid loans are taken by the subscribers when their credit is over or balance is not sufficient to make a call. Prepaid loan details may also have a relationship with the socioeconomic level of the subscriber. Using the loan CDR of the Prepaid customers, analysis was done in two approaches. Following are the two variables extracted from the loan records.

- Average loan amount per subscriber, monthly district wise
- Average loan amount for one month in DSD wise

Same procedures similar to preprocessing voice CDR were followed to extract the average loan amount from Prepaid loan CDR. Loan CDR in 18 months' period were preprocessed using Spark to derive the parameters. District wise averages (Figure 3.17) and DSD wise averages were generated using the CDR. Since loan CDR had low volume, data processing was not much expensive as other CDR processing.

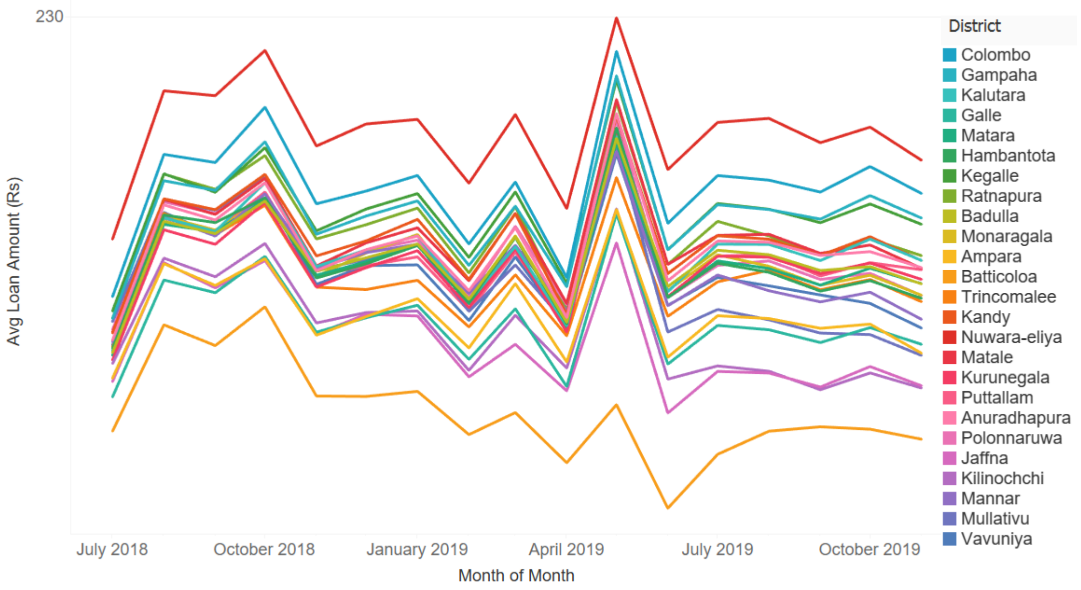


Figure 3.17: Average monthly Prepaid loan amount district wise

3.6.6 Average social media applications' data consumption

Mobile users consume most of the available broadband data to browse their social media networks. Social media networks like Facebook, Viber, WhatsApp are heavily used nowadays. The average social media usage parameters can symbolize the socioeconomic level of a person. Therefore, average data usage for some of the social media applications were extracted from Data CDR. It was also done in two methods, DSD wise and district wise.

- Average data usage of applications (Facebook, Viber, YouTube and Gmail), monthly district wise
- Average data usage of applications (Facebook, Viber, YouTube and Gmail) for one month in DSD wise

Similar data processing approach was followed similar to the above sections. CDR data in two years were used to calculate the average data consumption of each application. Spark was used for data preprocessing.

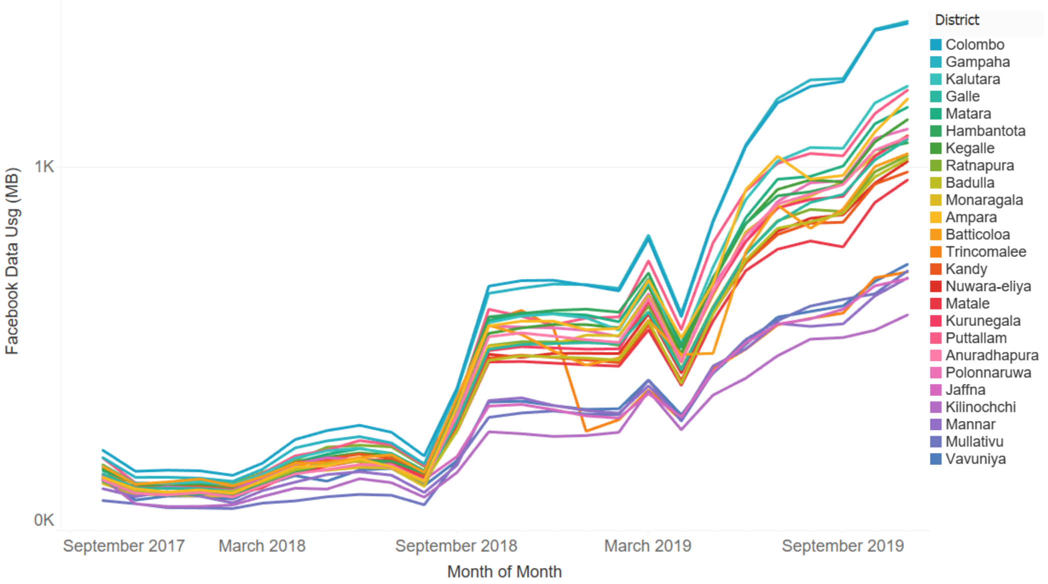


Figure 3.18: Average monthly Facebook data usage per connection district wise

Following map (Figure 3.19) shows the Facebook data consumption DSD wise for one month. Viber, YouTube and Gmail application data consumption were calculated in the same manner.

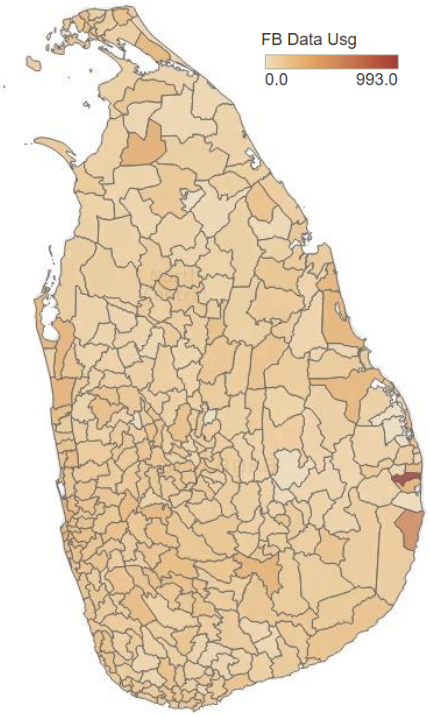


Figure 3.19: Average monthly Facebook data usage per connection DSD wise

3.6.7 On Net & Off Net average outgoing minute usage

The voice calls customers making within the network are called on net voice calls and other calls are named as off net voice calls. In Sri Lanka, off net outgoing calls are charged higher compared to the on net outgoing calls. Because of the difference in charging rates, mobile user calling pattern can have different characteristics according to his expending capabilities. Averages were calculated in two approaches. They are

- Average On Net and Off Net outgoing minute usage monthly district wise
- Average On Net and Off Net outgoing minute usage for one month DSD wise

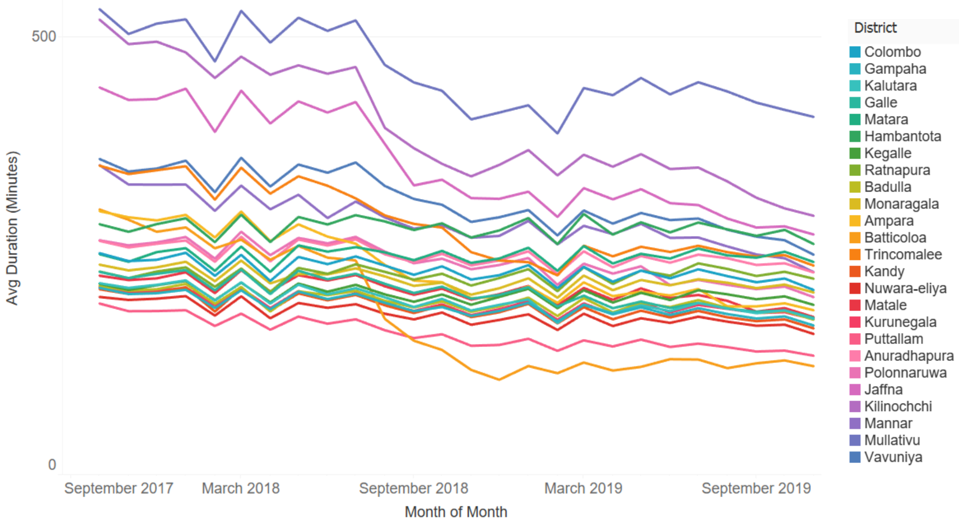


Figure 3.20: District wise monthly average on-net outgoing minute usage

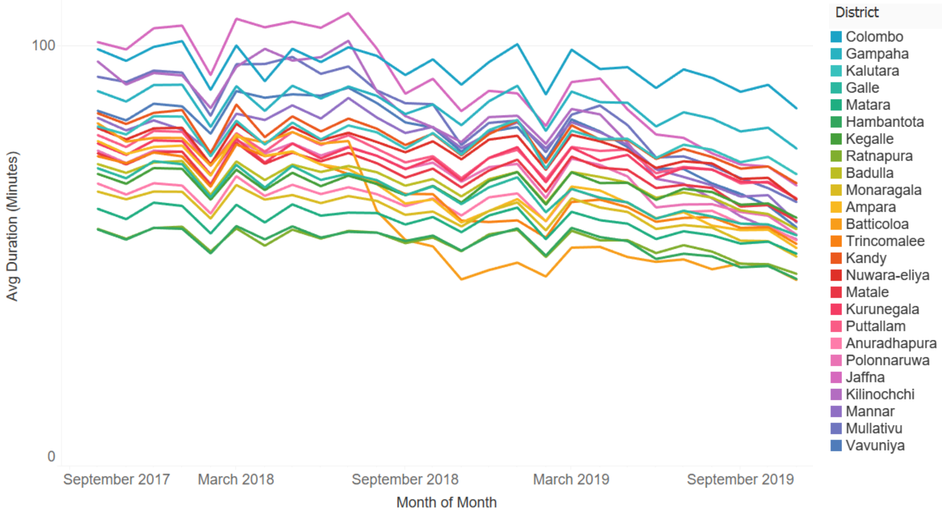


Figure 3.21: District wise monthly average off-net outgoing minute usage

3.7 Calculating feature relationship with SEL data

After extracting features from the CDR, their relationship with socioeconomic level of Sri Lanka has to be analyzed. There are many statistical methods available to calculate the relationship of two or more variables. Relationship between variables is called the correlation. Correlation index shows how much variables are associated with each other. In this research, Pearson Correlation Coefficient was used to determine the linearity and non-linearity between the features and SEL data in a specific geographical area.

3.7.1 Pearson Correlation Coefficient

Pearson Correlation coefficient calculates the linear relationship between two variables. According to the Cauchy–Schwarz inequality, Pearson Correlation coefficient is a value between +1 and -1. 0 is considered as no correlation. +1 is a strong positive linear correlation and -1 is a strong negative linear correlation[13]. The method of calculating the coefficient of correlation is taking the covariance of the two variables divided by the product of their standard deviation.

Pearson Correlation coefficient is commonly used frequently when analyzing the relationship of the variables. In the literature review, there were instances where Pearson coefficient was used to measure the correlation of CDR features with SEL level. Following figure (Figure 3.22) shows how to interpret Pearson correlation index.

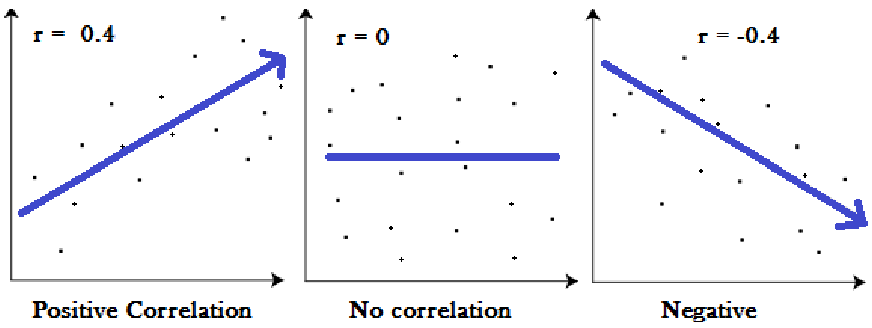


Figure 3.22: Pearson correlation coefficient values

3.7.2 Calculating correlation with poverty line

In Sri Lanka, Department of Census and Statistics releases poverty line data monthly. Poverty line data are available for each district in Sri Lanka. It denotes the minimum expenditure per person to fulfill basic needs. The percentage of the population which are below the poverty line is considered as the population suffered by poverty. It is very helpful to make policies to reduce that percentage in Sri Lanka.

In this research, we have extracted a number of features with monthly respected values for each district as explained in the above sections. CDR and a number of telco data sources were used to extract the following features.

- Average outgoing call volume
- Average outgoing minute usage, on-net & off-net minute usage
- Average data usage
- Average recharge amount
- Average Postpaid bill amount
- Average Prepaid loan amount
- Average data usage for applications
    - Facebook
    - Viber
    - YouTube
    - Gmail

In order to find the relationship between above features, Pearson Correlation Coefficient was calculated. More than 2 years' data were available for the correlation calculation which can give more accurate results. Using the correlation index, it can be decided which features in telco domain can be used to analyze the socioeconomic level of the people in Sri Lanka. It can expose the features to predict poverty in a macro geographical area. It can save a huge effort required to get a socioeconomic view in Sri Lanka covering all districts.

3.7.3 Calculating correlation with census data

Using 2011/12 census data in Sri Lank, a socioeconomic index has been created which characterize the housing and demographic details of the population. Principal Component Analysis (PCA) has been used to generate this index[1]. It was available for each Grama Niladhari Division (GND) in Sri Lanka which is the smallest governance unit in the country.

There are a number of features extracted from CDR in DSD level. Since socioeconomic index was related to one-time step, features were extracted for one month which is closer to the period census was held. Features from CDR were in Divisional Secretariat Division (DSD) level which is one step ahead of GND level. To calculate the correlation, the GND level socioeconomic index had to be converted to DSD level index. Following equation was used for the conversion.

$$IndexDSD = \frac{Index\ GND \times Population\ GND}{Population\ DSD}$$

After applying the above equation, poverty index for each DSD can be generated. Using the telco data sources, same feature set extracted for the district, were generated for each DSD. Then Pearson correlation coefficient was calculated for each feature to check the relationship with the poverty index. Following are the features generated for one month for each DSD.

- Average outgoing call volume
- Average outgoing minute usage, on-net & off-net minute usage
- Average data usage
- Average recharge amount
- Average Postpaid bill amount
- Average Prepaid loan amount
- Average data usage for applications (Facebook, YouTube, Viber, Gmail)

3.8 Predicting socioeconomic level

Using monthly poverty line data released by the Department of Census and Statistics, highly correlated features extracted from CDR and telco databases can be identified. Using these highly correlated features, district level future poverty indexes can be predicted. A Machine Learning technique will have to be applied in the future value prediction. Linear Regression was the model applied in this research.

Other than the district level, poverty index created on DSD level was used for the correlation analysis. Telco domain features extracted from CDR were subjected to statistical operations with those DSD level poverty indexes. Then Pearson correlation coefficient was calculated to see the relationship with the DSD level poverty index. Using the highly correlated features, ML model can be developed to predict the future poverty indexes. In order to achieve that target, Linear Regression was used as the ML technique.

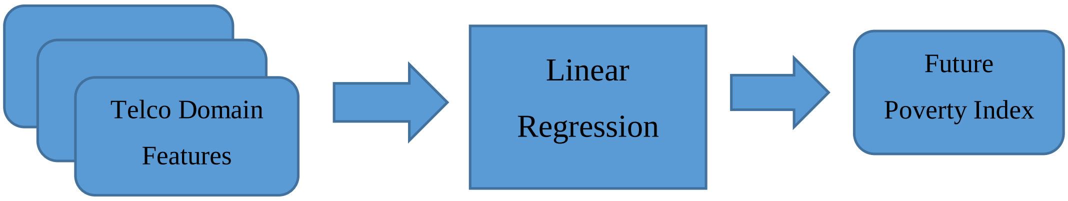


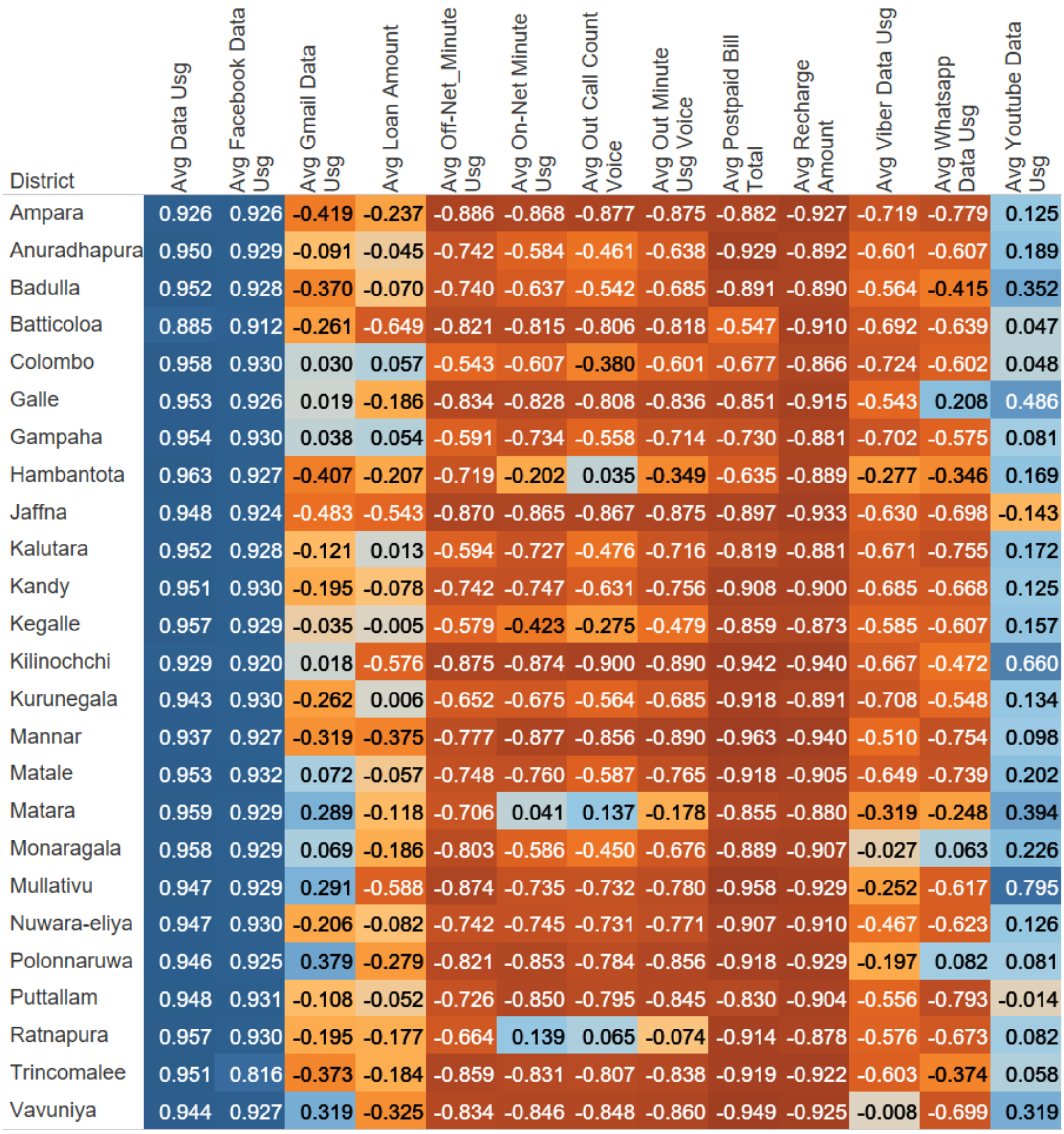
Figure 3.23: Poverty Index prediction process

4. RESULTS

4.1 Correlation analysis with district poverty line data

Pearson Coefficient of Correlation was calculated for each feature extracted from CDR and other telco data sources, against the district poverty line data. Features and poverty data were summarized monthly from October 2017 to December 2019. Following table (Table 4.1) shows the correlation with the poverty line.

Table 4.1: District wise Pearson Correlation Coefficient for each feature



District	Avg Data Usg	Avg Facebook Data Usg	Avg Gmail Data Usg	Avg Loan Amount	Avg Off-Net_Minute Usg	Avg On-Net Minute Usg	Avg Out Call Count Voice	Avg Out Minute Usg Voice	Avg Postpaid Bill Total	Avg Recharge Amount	Avg Viber Data Usg	Avg Whatsapp Data Usg	Avg Youtube Data Usg
Ampara	0.926	0.926	-0.419	-0.237	-0.886	-0.868	-0.877	-0.875	-0.882	-0.927	-0.719	-0.779	0.125
Anuradhapura	0.950	0.929	-0.091	-0.045	-0.742	-0.584	-0.461	-0.638	-0.929	-0.892	-0.601	-0.607	0.189
Badulla	0.952	0.928	-0.370	-0.070	-0.740	-0.637	-0.542	-0.685	-0.891	-0.890	-0.564	-0.415	0.352
Batticoloa	0.885	0.912	-0.261	-0.649	-0.821	-0.815	-0.806	-0.818	-0.547	-0.910	-0.692	-0.639	0.047
Colombo	0.958	0.930	0.030	0.057	-0.543	-0.607	-0.380	-0.601	-0.677	-0.866	-0.724	-0.602	0.048
Galle	0.953	0.926	0.019	-0.186	-0.834	-0.828	-0.808	-0.836	-0.851	-0.915	-0.543	0.208	0.486
Gampaha	0.954	0.930	0.038	0.054	-0.591	-0.734	-0.558	-0.714	-0.730	-0.881	-0.702	-0.575	0.081
Hambantota	0.963	0.927	-0.407	-0.207	-0.719	-0.202	0.035	-0.349	-0.635	-0.889	-0.277	-0.346	0.169
Jaffna	0.948	0.924	-0.483	-0.543	-0.870	-0.865	-0.867	-0.875	-0.897	-0.933	-0.630	-0.698	-0.143
Kalutara	0.952	0.928	-0.121	0.013	-0.594	-0.727	-0.476	-0.716	-0.819	-0.881	-0.671	-0.755	0.172
Kandy	0.951	0.930	-0.195	-0.078	-0.742	-0.747	-0.631	-0.756	-0.908	-0.900	-0.685	-0.668	0.125
Kegalle	0.957	0.929	-0.035	-0.005	-0.579	-0.423	-0.275	-0.479	-0.859	-0.873	-0.585	-0.607	0.157
Kilinochchi	0.929	0.920	0.018	-0.576	-0.875	-0.874	-0.900	-0.890	-0.942	-0.940	-0.667	-0.472	0.660
Kurunegala	0.943	0.930	-0.262	0.006	-0.652	-0.675	-0.564	-0.685	-0.918	-0.891	-0.708	-0.548	0.134
Mannar	0.937	0.927	-0.319	-0.375	-0.777	-0.877	-0.856	-0.890	-0.963	-0.940	-0.510	-0.754	0.098
Matale	0.953	0.932	0.072	-0.057	-0.748	-0.760	-0.587	-0.765	-0.918	-0.905	-0.649	-0.739	0.202
Matara	0.959	0.929	0.289	-0.118	-0.706	0.041	0.137	-0.178	-0.855	-0.880	-0.319	-0.248	0.394
Monaragala	0.958	0.929	0.069	-0.186	-0.803	-0.586	-0.450	-0.676	-0.889	-0.907	-0.027	0.063	0.226
Mullativu	0.947	0.929	0.291	-0.588	-0.874	-0.735	-0.732	-0.780	-0.958	-0.929	-0.252	-0.617	0.795
Nuwara-eliya	0.947	0.930	-0.206	-0.082	-0.742	-0.745	-0.731	-0.771	-0.907	-0.910	-0.467	-0.623	0.126
Polonnaruwa	0.946	0.925	0.379	-0.279	-0.821	-0.853	-0.784	-0.856	-0.918	-0.929	-0.197	0.082	0.081
Puttallam	0.948	0.931	-0.108	-0.052	-0.726	-0.850	-0.795	-0.845	-0.830	-0.904	-0.556	-0.793	-0.014
Ratnapura	0.957	0.930	-0.195	-0.177	-0.664	0.139	0.065	-0.074	-0.914	-0.878	-0.576	-0.673	0.082
Trincomalee	0.951	0.816	-0.373	-0.184	-0.859	-0.831	-0.807	-0.838	-0.919	-0.922	-0.603	-0.374	0.058
Vavuniya	0.944	0.927	0.319	-0.325	-0.834	-0.846	-0.848	-0.860	-0.949	-0.925	-0.008	-0.699	0.319

The Pearson correlation coefficient was calculated for every district in Sri Lanka using the poverty line data issued by the Department of Census and Statistics. Following table (Table 4.2) shows the variance and average coefficient value based on the district wise correlation values (Table 4.1).

Table 4.2: Pearson Coefficient average and variance

Feature	Variance of Pearson Correlation	Average Pearson Correlation
Avg Data Usg	0.00024	0.947
Avg Facebook Data Usg	0.00052	0.923
Avg Gmail Data Usg	0.06058	-0.093
Avg Loan Amount	0.04339	-0.195
Avg Off-Net Minute Usg	0.01026	-0.750
Avg On-Net Minute Usg	0.07572	-0.660
Avg Out Call Count	0.09158	-0.580
Avg Out Minute Usg	0.04697	-0.698
Avg Postpaid Bill Total	0.01107	-0.860
Avg Recharge Total	0.00048	-0.905
Avg Viber Data Usg	0.04534	-0.517
Avg Whatsapp Data Usg	0.07657	-0.515
Avg Youtube Data Usg	0.04297	0.199

After looking at the Pearson correlation coefficient, a person’s behavior characterized by the mobile phone usage has a significant correlation with the socioeconomic level. Data usage, Facebook data usage, average off net minute usage, average on net minute usage, average outgoing call count, average outgoing minute usage, average Postpaid bill total and average recharge total have a higher correlation with the poverty line.

The average outgoing call count has higher correlation, but the variance is too high. Viber and YouTube data usage have less correlation compared to the other features.

4.2 Predicting district poverty line

District wise poverty line prediction can save a lot of cost and effort. As discussed in the previous section, following features derived using CDR and other telco data sources can be used to predict the district poverty line. The average value was calculated on monthly basis and subscriber level.

- Average data usage
- Average Facebook data usage
- Average off net minute usage
- Average on net minute usage
- Average outgoing call count
- Average outgoing minute usage
- Average Postpaid bill total
- Average recharge total

Above features were used to build a Machine Learning model to predict poverty line. Following graph shows predicted vs actual values for several districts.

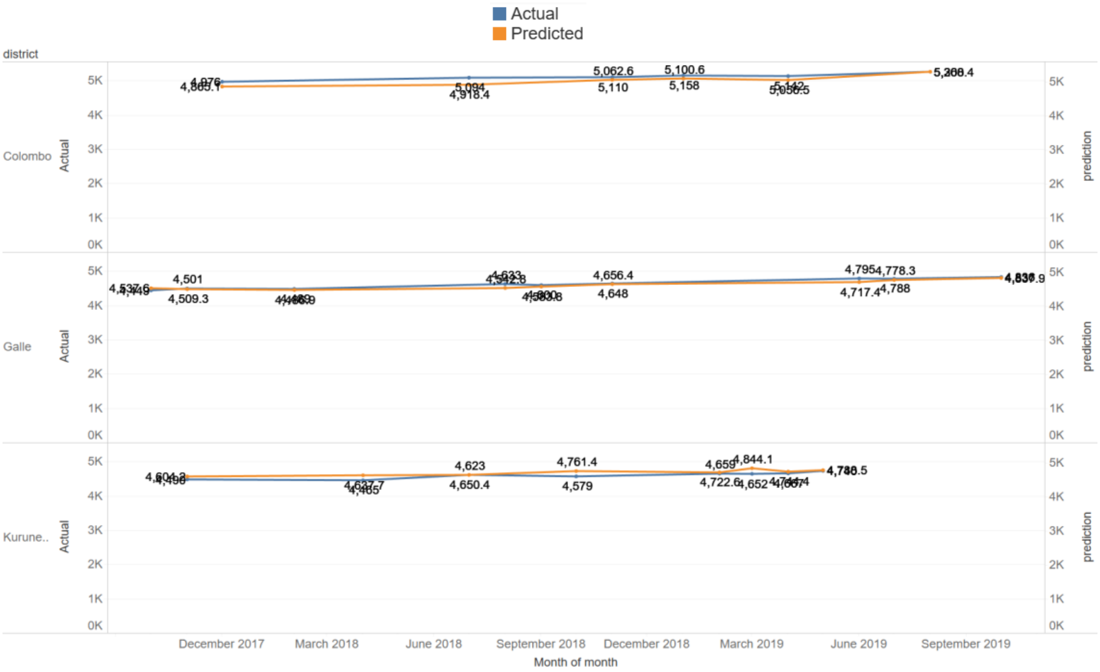


Figure 4.1: District wise prediction of poverty line

Linear Regression was used to do the prediction district wise. The r-squared value of the prediction was 0.72 when all the variables were used. It can be considered as a satisfactory accuracy level in Linear Regression.

In district poverty line prediction, p-values were calculated after applying the Linear Regression. P-values denotes whether relationships found were significant or not [14]. Features used in the model and their respected p-values are mentioned in the following table (Table: 4.3).

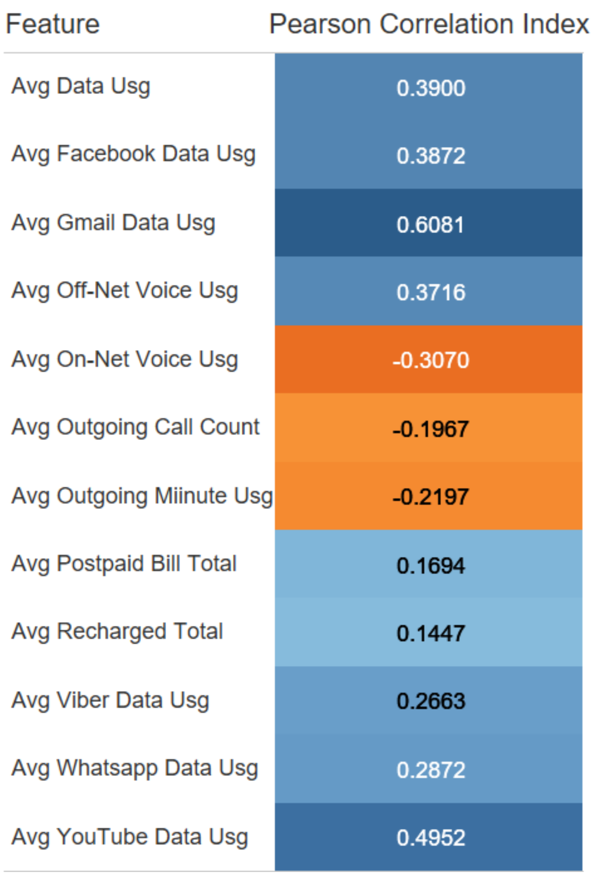
Table 4.3: P-values associated with district level features used in LR model

Feature	P-value
Average data usage	1.81×10^{-10}
Average recharge total	0.1267
Average Postpaid bill total	2.23×10^{-5}
Average outgoing minute usage	0.4569
Average outgoing call count	3.70×10^{-6}
Average Facebook data usage	0.00
Average on net minute usage	0.3018
Average off net minute usage	7.66×10^{-7}

4.3 Correlation analysis using DSD socioeconomic index

Using 2011/12 Census data in Sri Lanka, Grama Niladhari Division wise socioeconomic index has been derived [10]. As discussed in the methodology, Divisional Secretariat Division wise index was derived using those data. Then Pearson Coefficient of Correlation was calculated for the features derived in DSD level using CDR and telco data sources. Following table (Table 4.4) shows the Pearson coefficient calculated for each feature vs DSD socioeconomic index.

Table 4.4: DSD wise Pearson Correlation Coefficient

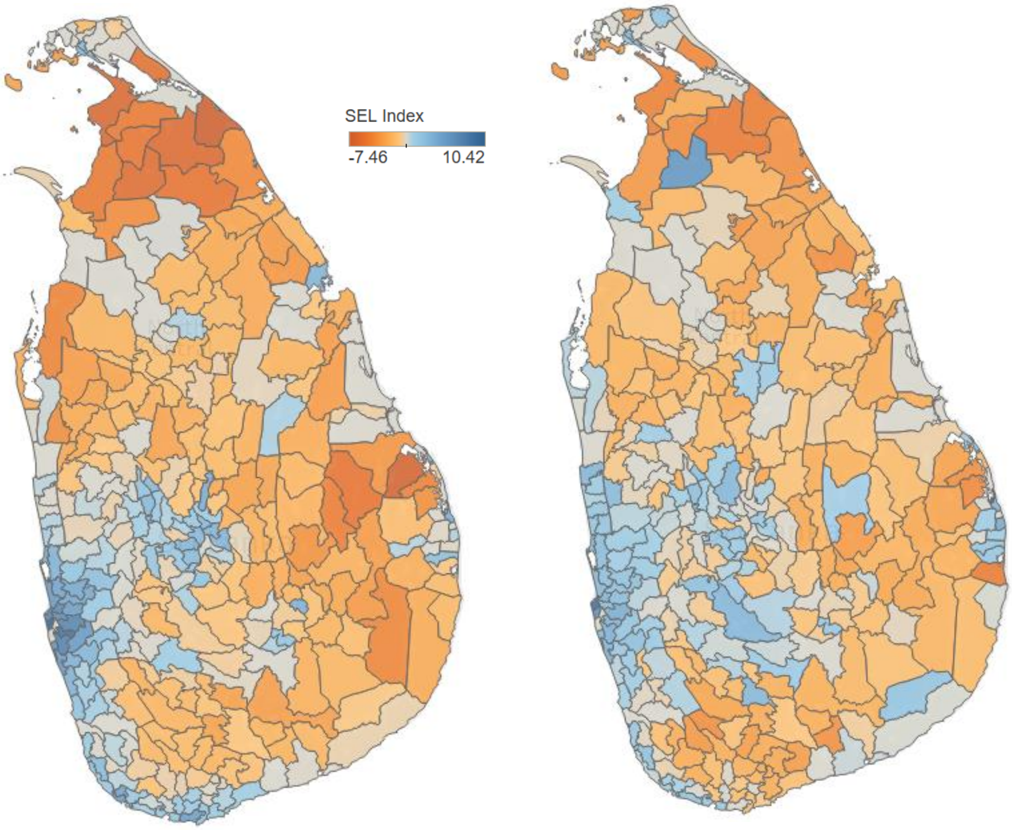


Feature	Pearson Correlation Index
Avg Data Usg	0.3900
Avg Facebook Data Usg	0.3872
Avg Gmail Data Usg	0.6081
Avg Off-Net Voice Usg	0.3716
Avg On-Net Voice Usg	-0.3070
Avg Outgoing Call Count	-0.1967
Avg Outgoing Miinute Usg	-0.2197
Avg Postpaid Bill Total	0.1694
Avg Recharged Total	0.1447
Avg Viber Data Usg	0.2663
Avg Whatsapp Data Usg	0.2872
Avg YouTube Data Usg	0.4952

4.4 Predicting DSD socioeconomic index

YouTube data usage, Gmail data usage and total data usage have the highest Pearson correlation index with the socioeconomic index at DSD level. Other features extracted from CDR, have a medium correlation. Using most of the features, ML model was developed to predict the socioeconomic index of any DSD.

Linear Regression was used to predict the socioeconomic index. The model has r-squared value of 0.52 which is over 0.5. That is a promising accuracy level with a low number of data points used for the prediction. The number of Divisional Secretariats used for the prediction is 285 because of the availability of SEL data. Following Sri Lankan map (Figure 4.2) shows the actual vs the predicted values.



Actual Index   Predicted Index

Figure 4.2: DSD wise actual SEL vs predicted SEL

After applying the Linear Regression, p-values were calculated for the features used in the model which gave r-squared of 0.52. P-value defines the statistical significance of the features in deciding the linear relationship [14]. Following Table (Table 4.5) has p-value associated with each feature used in building the Linear Regression model used to predict DSD level SEL values.

Average Gmail data usage and average off net minute usage have lowest p-values after applying the Linear Regression. That means those features have a significant relationship with the poverty index in DSD level. Average total data usage also has a low p-value.

Table 4.5: P-values associated with DSD level features used in LR model

Feature	P-value
Avg Data Usg	0.008
Avg Recharge Total	0.166
Avg Postpaid Bill Total	0.656
Avg Out Minute Usg	0.152
Avg Facebook Data Usg	0.349
Avg Gmail Data Usg	01.83×10^{-5}
Avg Youtube Data Usg	0.3252
Avg On-Net Minute Usg	0.0355
Avg Off-Net Minute Usg	1.805×10^{-4}

5. DISCUSSION

5.1 Computational complexity

CDR files were stored in HDFS in order to derive parameters related to the socioeconomic level. As discussed in the methodology, voice CDR, data CDR and recharge CDR were stored relevant to a period of two years. Following table (Table 5.1) has the approximate data amounts which were subjected in the research.

Table 5.1: CDR volume stored for parameter extraction

CDR Type	Daily Volume (GB)	Compressed Amount (GB)	Total Amount (GB)
Voice CDR	18	6	4320
Data CDR	40	15	10800
Voucher CDR	0.8	0.15	108

As stated in the table (Table 5.1), more than 15TB of data were subjected to the processing. CDR files were not stored as raw files because of the storage limitations. Parquet file compression technique was used which saved more than 30TB.

Processing 4 Terabytes of data to get the home location of subscribers was not feasible without partitioning. Data Records were partitioned using the date attribute in CDR. Then partitions were processed in separate manner to improve the processing performance. HDFS chunk size was 128MB which defines the block size of the CDR files stored. Lower chunk size was not recommended because it made the processing performance low.

Spark was used over the MapReduce because of the higher performance and usability. More than 500GB of memory in the Hadoop cluster was utilized by Spark in CDR processing.

5.2 SEL prediction in district and DSD wise

There is a large number of features which can be derived using CDR and telco data sources. This research was about finding the feasibility of analyzing and predicting socioeconomic level using those features. In order to analyze the socioeconomic level in a specific geographic area in Sri Lanka, this research used two approaches. The two approaches were

- Using district level poverty line data in Sri Lanka
- Using divisional secretariat level socioeconomic index derived using 2011/12 census data

Related to district level SEL analysis, average total data usage and average Facebook data usage had a strong positive correlation with the district poverty line. There were two other features which had a strong negative correlation. They were average Postpaid bill total and average recharge total. The averages were calculated in monthly and subscriber level basis.

Average outgoing minute usage, on net and off net average outgoing minute usages had a strong negative correlation in the district level analysis. Average Viber and WhatsApp data usages had a strong negative correlation. Average Prepaid loan amount and Average Gmail data usage had a low correlation with the SEL.

Therefore, except the average Prepaid loan amount and average Gmail data usage, other district wise features were used to predict district poverty line. The poverty line is the minimum cost needed to fulfill basic needs monthly per person. Linear Regression was used to train the model. Since 2017 October to 2019 December monthly data were used for 25 districts, model was built using 650 data points.

R-Squared is a statistical measure of fit that indicates how much variation of a dependent variable is explained by the independent variable(s) in a regression model [15]. District poverty line prediction model built had 0.72 r-squared value which was acceptable. The SEL prediction done in a Latin American country had a r-squared of 0.82 for a subset of the population [3]. Data set had only 5 months' data and SEL was given as a categorical value to a LR model. That was the maximum r-squared value which was found in the related work.

In the second approach, same variables used in district wise approach were calculated in divisional secretariat wise. Socioeconomic index was calculated in DSD wise with census data as discussed in the methodology. CDR and telco data sources which were closer to census executed time were used to derive the averages.

Average Gmail data usage and average YouTube data usage had a strong positive correlation with the socioeconomic index. Total average data usage, average off net outgoing voice minute usage and average Facebook data usage had marginally strong positive correlation with the SEL value in DSD wise. Average recharge value, average Postpaid bill total, average Viber data usage and average WhatsApp data usage had a normal positive correlation with SEL. Average On net outgoing voice minute usage had a marginal strong negative correlation, while average outgoing total minute usage and average outgoing call count had a normal negative correlation.

Linear Regression model built using the above features gave r-squared of 0.52 which can be considered as acceptable value. Number of data point used to develop the model was less compared to the district level. Only 285 DSDs have the socioeconomic index derived from 2011/12 census data. That can be considered as the main reason for the less accuracy than the district level model.

Previous research [1] carried out in Sri Lanka had not mentioned the r-squared values obtained. It has stated that r-squared value obtained was less than the ideal value. In predicting SEL in Ivory Coast [6], r-squared of 0.342 was recorded as the maximum value where only call volume and call count were used.

6. FUTURE WORK AND CHALLENGES

Handling large volume of CDR is challenging. Millions of records flow from the billing system daily. In order to process and summarize those data, big data skills and capabilities are required. A platform which is rich in processing power, memory and storage is necessary in such situations. In this research, Hadoop cluster was utilized for the storage and Spark was used for data processing. The cluster had a storage capacity of hundreds of Terabytes while memory was over hundreds of Gigabytes.

In order to improve the accuracy of the model, various features can be integrated after the correlation analysis. The features on mobility can be further added as discussed in the related work. In order to increase the number of data points used in ML model, district wise data can be gathered for coming years.

Considering DSD wise scope, census held in future days or public survey data can be used to improve the accuracy. In this research, time series analysis was not considered because of the low number of data points and time limitations. It will be a good option to try out in improving the model. The geographical level can be considered in more fine grained level, such as Grama Niladhari Division. It will be very helpful in addressing poverty problems in Sri Lanka.

7. CONCLUSION

Assessing socioeconomic level in a country can help the governance in many ways. Policies can be made to address the poverty while measuring the effectiveness of the current policies and newly established ones. The resources can be divided fairly where most of the people who suffered from poverty get maximum benefits. This research is about predicting socioeconomic level in Sri Lanka using CDR data.

Various features were extracted from CDR and telco data sources in district level and divisional secretariat level. Two approaches were followed in this research based on the availability of SEL data. District wise summarized features and district poverty line data issued by the Department of Census and Statistics were subjected to correlation analysis. Average data usage and social media usages had strong positive correlation while average Postpaid bill total, average recharge total, Average outgoing voice minute usage had a strong negative correlation. Predicting district wise poverty line was done using Linear Regression which gave an acceptable r-squared of 0.72.

Using DSD wise derived socioeconomic index using 2011/12 data was the other approach. Average Gmail data usage, average YouTube data usage and total data usage had a strong positive correlation with the DSD wise SEL index. Average on net outgoing voice usage had a marginal negative correlation with SEL. Linear Regression model built for DSD wise SEL prediction gave r-squared of 0.52 which was much promising than the previous research carried out in Sri Lanka. If the socioeconomic level can be predicted in more fine grained level geographically, it can save the high amount of effort and cost spent on holding census and surveys regularly all across the country to assess the economic level.

8. REFERENCES

[1] L. Fernando, S. Lokanathan, A. Surendra, and T. Gomez, “Predicting population-level socio-economic characteristics using Call Detail Records (CDRs) in Sri Lanka,” *Proc. 4th Int. Work. Data Sci. Macro-Modeling, DSMM 2018 - conjunction with ACM SIGMOD/PODS Conf.*, 2018.

[2] C. Smith-Clarke, A. Mashhadi, and L. Capra, “Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks,” *Conf. Hum. Factors Comput. Syst. - Proc.*, pp. 511–520, 2014.

[3] V. Frias-Martinez and J. Virseda, “On the relationship between socio-economic factors and cell phone usage,” *ACM Int. Conf. Proceeding Ser.*, pp. 76–84, 2012.

[4] V. Soto, V. Frias-Martinez, J. Virseda, and E. Frias-Martinez, “Prediction of socioeconomic levels using cell phone records,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6787 LNCS, no. 1, pp. 377–388, 2011.

[5] J. Blumenstock, G. Cadamuro, and R. On, “Predicting poverty and wealth from mobile phone metadata,” *Science (80-.).*, vol. 350, no. 6264, pp. 1073–1076, 2015.

[6] Emmanuel Francis Letouzé, “Applications and Implications of Big Data for Demo-Economic Analysis: The Case of Call-Detail Records,” *Foreign Aff.*, vol. 91, no. 5, pp. 1689–1699, 2016.

[7] S. Lanka *et al.*, “The Potential of Mobile Network Big Data as a Tool in Colombo’s Transportation and Urban Planning,” *Inf. Technol. Int. Dev.*, vol. 12, no. 2, pp. 63-73–73, 2016.

[8] M. Vanhoof, F. Reis, and T. Plötz, “Detecting Home Locations from CDR Data : Introducing Spatial Uncertainty to the State - of - the - Art.”

[9] Department of Census and Statistics, “Revision of District Official Poverty Lines.” Department of Census and Statistics, Sri Lanka, p. 1, 2017.

[10] LIRNEasia and V. Dias, “Socioeconomic Index,” 2019. [Online]. Available: https://github.com/LIRNEasia/socioeconomic-index. [Accessed: 08-Feb-

2020].

[11] Department of Census & Statistics, “Census of Population and Housing.” Ministry of Policy Planning and Economic Affairs, Sri Lanka, 2012.

[12] Vyas Seema and K. Lilani, “Constructing socio-economic status indices : how to use principal components analysis,” no. October, 2006.

[13] “Pearson correlation coefficient,” *Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient. [Accessed: 08-Feb-2020].

[14] J. Frost, “How to Interpret P-valuesand Coefficients in Regression Analysis,” 2018. [Online]. Available: https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/. [Accessed: 26-May-2020].

[15] A. Hayes, “R-Squared Definition,” 2019. [Online]. Available: https://www.investopedia.com/terms/r/r-squared.asp. [Accessed: 08-Feb-2020].