Predicting population-level socio-economic characteristics using Call Detail Records (CDRs) in Sri Lanka

Lasantha Fernando
LIRNE*asia*
Colombo 08, Sri Lanka
lasantha@lirneasia.net

Aparna Surendra
LIRNE*asia*
Colombo 08, Sri Lanka
aparnas1@alumni.stanford.edu

Sriganesh Lokanathan
LIRNE*asia*
Colombo 08, Sri Lanka
sriganesh@lirneasia.net

Thavisha Gomez
LIRNE*asia*
Colombo 08, Sri Lanka
thavisha@lirneasia.net

ABSTRACT

Prior work has shown that mobile network big data can be used as a high-frequency alternative data source to derive proxy measures that have strong predictive capacity to estimate census and poverty data in developing countries. Given that the observations from these studies can be dependent on local context and regional characteristics, we replicate this work targeting two regions in Sri Lanka. We focus on Northern Province, a post-conflict region with a highly vulnerable population and Western Province, an urban region that has been relatively untouched by the conflict. We analyze the relationship between aggregate features related to consumption, social and mobility behaviors derived from pseudonymized mobile phone CDRs and census data associated with population-level socio-economic characteristics. We show that Northern Province exhibits different social and mobility patterns when compared to populations with similar socio-economic characteristics in Western Province, which highlights the importance of replicating prior research studies under different local contexts. We go on to develop predictive models that estimate the census features using the derived CDR features. Our results confirm the applicability of this methodology in a Sri Lankan, post-conflict setting, and highlight potential areas that need to be addressed in order to improve the accuracy of our prediction models.

CCS CONCEPTS

• **Mathematics of computing → Regression analysis**; *Exploratory data analysis*; • **Information systems → Data analytics**; **Summarization**;

KEYWORDS

mobile network big data, regression analysis, large-scale data processing

DSMM'18, June 15, 2018, Houston, TX, USA

ACM ISBN 978-1-4503-5883-5/18/06…$15.00
https://doi.org/10.1145/3220547.3220549

ACM Reference Format:
Lasantha Fernando, Aparna Surendra, Sriganesh Lokanathan, and Thavisha Gomez. 2018. Predicting population-level socio-economic characteristics using Call Detail Records (CDRs) in Sri Lanka. In *DSMM'18: Data Science for Macro-Modeling with Financial and Economic Datasets, June 15, 2018, Houston, TX, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3220547.3220549

1 INTRODUCTION

The availability of accurate, timely, disaggregated, and comparable socio-economic data is crucial with regard to planning for economic development and resource allocation. Spatially-granular demographic data are often collected through the decennial national census, and population-level socio-economic characteristics are often captured more frequently, through representative surveys such as the Household Income and Expenditure Survey (HIES). In Sri Lanka, the HIES is conducted once every three years, and is representative only up to the district level, the second-level administrative unit. The census and surveys are expensive and time-consuming to conduct, and in the context of Sri Lanka, are not frequent enough to capture the changing dynamics of a fast-moving economy, especially one recovering from civil conflict. Similarly, other developing countries grapple with a lack of poverty data [12]. Our research seeks to determine the opportunity for mobile phone metadata to provide a reliable, cheap proxy for census data within Sri Lanka, focusing on a post-conflict region as well as a fast developing urban region, that have a greater need for frequent data collection.

Mobile phone metadata such as Call Detail Records (CDRs) can broadly describe three dimensions of human behavior: social networks, consumption activity, and mobility [11]. CDRs are passively collected by the mobile network whenever a subscriber uses the mobile phone to make or receive a phone call, send or receive a text, or when initiating a data session. A CDR that is generated by mobile phones yields new types of data - such as spatially disaggregated data at micro-regional levels (e.g. the household level). This, coupled with the near-ubiquitous adoption of mobile phones in developing countries, presents opportunities to leverage such data sources to complement traditional statistics in the intervals between official surveys.

Our paper seeks to identify relationships between features derived from Sri Lankan census and CDR data, replicating methods used by Frias-Martinez and Virseda [6] for an unidentified Latin American country. We use the 2011/12 Sri Lanka census data and CDR data for 3 months at the beginning of 2013 for approximately 2,100,000 mobile subscribers ((35-40% of the provincial population reported in 2011/12 census) within Western Province and 600,000 mobile phone subscribers (50-55% of the provincial population reported in 2011/12 census) from Sri Lanka's Northern Province. The analysis for Western Province, a densely populated region that includes the city of Colombo (the administrative and commercial center of the country), is intended to be used as a baseline to compare and contrast the behavioral relationships observed in Northern Province, which is a post-conflict region. Here, we assume the population of Western Province to be largely unaffected by the civil conflict that occurred prior to 2009 and to exhibit similar behavioral patterns as reported in prior work.

In our work, we primarily seek to answer two questions: (1) What relationships, if any, exist between Sri Lankan census and CDR data, and do these provide an opportunity for predictive models? (2) Are methods developed for census feature prediction in other countries applicable within a Sri Lankan context, especially in the regions severely affected by conflict?

2 RELATED WORK

Recent studies have highlighted the potential for mobile phone data to address a range of development issues, including: food security [3], disasters [9, 17], and disease propagation [1, 16]. In terms of understanding economic activity, numerous studies [2, 6, 13] have sought to map variables derived from CDRs to infer patterns of socioeconomic levels of populations, sometimes using them in conjunction with other mobile data such as airtime credit purchases [3, 7] or with other data sources such as satellite data [15]. Together, these studies highlight CDR's potential to complement traditional sources of data.

Frias-Martinez and Virseda [6] identified statistically significant correlates from CDRs and census based socio-economic features. Their research indicates strong linkages between populations with higher socioeconomic levels (SEL) and range of mobility. Using a somewhat related methodology, Smith-Clarke et al. [14] looked at mobile phone consumption features aggregated at a cell tower level and compared it with poverty data in Ivory Coast. However, the researchers were limited by the lack of ground truth data - the government data were older than the mobile data and of a lower resolution.

More recently, Blumenstock et al. [2] built a supervised model to predict wealth at an individual level. They leveraged CDR and phone survey data to develop a composite wealth index, which was used to predict the wealth of the out-of-sample population. The results were validated at a district level using national survey data, but the lack of micro-regional data prevented micro-region validation. Further, Steele et al. [15] showed that CDR data when combined with remote-sensing data was better positioned to model traditional measures of poverty at disaggregated geographic levels. Gutierrez et al. [7] leveraged CDRs and airtime credit purchase data to infer the relative income of individuals, based on the assumption that those who made larger airtime credit purchases were relatively more affluent than those mobile users who made multiple purchase of smaller airtime credit.

As illustrated, prior work has demonstrated the opportunity for CDRs to predict poverty, and provide a "soft substitute" [6] for data collected through the expensive and time-consuming census. However, to our knowledge, similar research has yet to be conducted in the immediate (< 5 years) aftermath of a conflict.

We consider this research the initial phase of a larger study. In this phase, we use readily-available data to understand the applications of this research in the Sri Lankan, post-conflict context. The data limitation, while self-imposed, better mimics the realities of working with big data in a policy setting. Most sources of big data are passively-collected for non-policy purposes. The more sophisticated studies (e.g. Blumenstock et al. [2]) supplement big data with other sources, such as survey data.

3 BACKGROUND

Sri Lanka's Northern and Eastern provinces were the most severely affected by the civil war, which ended in 2009. Northern Province was chosen as a region of interest for our study because this region suffers from the greatest paucity of historical census data, and its population is changing rapidly in the aftermath of the conflict (a combination of resettlement and continued displacement, as well as greater economic connectivity and restored infrastructure). Western Province was chosen assuming the behavior of its population to be similar to what has been reported in other studies and as such would be ideal to contrast any behavioral patterns specific to Northern Province.

The 2011/12 census marked the first country-wide census since 1981 (the census was not conducted in 1991, and the 2001/2 census only included areas of the Northern Province that were not under rebel control). The Northern Province accounts for 5.4% of Sri Lanka's population, and is made up five districts: Jaffna, Kilinochchi, Mannar, Mullaitivu, and Vavuniya. Out of its five districts, three districts - Killinochi, Mullaitivu, and Mannar - record less than 100 people per square kilometer. Yet Jaffna district has between 600 - 999 people per square kilometer (see Figure 1), and is one of the more densely-populated districts in the country. In contrast, Western Province accounts for 28.6% of the population and consists of the 3 most densely-populated districts of the country - Colombo, Kalutara, and Gampaha - with Colombo district having a population density of 3,438 according to the 2011/12 census.

The civil war significantly affected in-country migratory patterns. Many Northern Province residents migrated to Vavuniya during the conflict, and the district experienced the highest annual population growth rate within the country from 1981-2012. Correspondingly, Jaffna and Mannar districts reported a decrease in growth rates within the same period. Following the end of the war in 2009, there has been substantial resettlement activity. The 2011/2012 census records approximately 351,900 Northern Province residents who cite their reason for migration (to their current district of residence) as 'Displacement' or 'Resettlement after Displacement' [10]. The United Nations estimates that between April 2009 and the end of November 2012, the total number of people who returned to the Northern Province stood at 482,000 people, with several thousand additional internally displaced persons in transit situations within

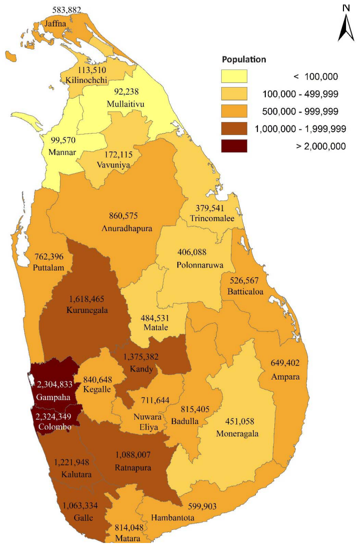

Figure 1: Population of Sri Lanka by district from the census of 2012[10].

the northern Districts [5]. As such, the area under study in Northern Province includes a highly vulnerable population.

The Sri Lankan Department of Census and Statistics measures poverty with: the official Poverty Line (LKR 3,264 per person per month in 2012/13), the poverty headcount index (share of population living below the poverty line), and the poverty gap index (the depth of poverty based on the aggregate poverty shortfall of the poor relative to the poverty line). The three measures are calculated at a district-level through the Household Income and Expenditure Survey (HIES), conducted once every three years with a sample population of 25,000 households. Recently, the Census department released more granular poverty data at the Divisional Secretariat Division (DSD) level, calculated using the small-area estimation method developed by Elbers, Lanjouw, and Lanjouw [4]. However, this method still allows a three-year lag in population distribution reporting, and may not capture spatial heterogeneity.

4 DESCRIPTION OF DATASETS

4.1 Call Detail Records (CDR) data

CDR data captures the following: (1) A unique identifier for the calling/sending party; (2) A unique identifier for the other party; (3) The timestamp at which the event was initiated; (4) The ID of the cellular antenna the subscriber was connected to at the time of the call. Each antenna is situated on a base transceiver station (BTS), which we can dereference to a physical (latitude, longitude) location. We only use CDRs associated with calls for the study.

The residence of each subscriber was determined as per the techniques in [8]. We correlated BTS-level mobile subscriber population data with mapped census population data, which suggested that the dataset includes people with more than one mobile SIM (the true number of mobile phone subscribers is lower than calculated).

The use of novel research methods and new datasets (including CDR data) have implications regarding privacy and representation, a full discussion of which is beyond the scope of this paper. While we have access to user-level CDR data, all data is pseudonymized. To further mitigate privacy concerns, we conduct and report analyses at the aggregate level.

The CDR features are loosely divided into three categories (Table 1): those that describe user's phone call behavior (Consumption), their social network (Social), and their geographic movement (Mobility). We calculated individual subscriber-level features, and then further aggregated them to the BTS level. There are approximately 250 BTS towers in the Northern Province, and approximately 650 BTS towers in Western Province. The estimated mean population within each BTS ranges from 3500 - 4500 people across districts, and the mean BTS mobile user population ranges from an estimated 1600 (Mannar) to 3500 (Colombo).

4.2 Census Data

For high-resolution 'ground truth' population data, we decided to use census results. While census results are widely-reported at the district level (2nd level administrative unit), some data are made publicly-available at the Grama Niladhari (GN) division (4th and lowest level administrative unit).

Northern Province has 921 GN divisions, with mean GN populations of over 1200 people in Jaffna, Kilinochchi, and Vavuniya districts, and less than 750 people in Mullaitivu and Mannar. Western Province has 2496 GN divisions, with mean GN population of over 1,600 for Gampaha and Kalutara districts, while Colombo district reporting a mean GN population of 4,000 people. We used 58 features from 12 categories of the national census for our analysis (Table 2). The Sri Lankan census focuses on population and housing, and does not collect or calculate figures related to income, assets, or consumption. Ideally, we would compare CDR data with poverty-specific features. For instance, Frias-Martinez and Virseda [6] used government-calculated socio-economic level (SEL) values, a weighted average of census features expressed from letters A (high SEL) to D (low SEL). A similar aggregate indicator does not exist in Sri Lanka. Therefore, we assume that at least some of the census features chosen (e.g. tile, granite, or terrazzo flooring, and university degrees) correspond to high SEL, and others (e.g. semi-permanent housing, cadjan or palmyrah walls) correspond to low SEL.

The census reports raw data in absolute figures, which we then converted into proportions. We also derived three additional features to capture population age from the census-provided data to identify whether larger age bands allowed clearer observation of statistical relationships.

Our census categories weigh heavily towards household-level data, especially observations of the housing unit (roof materials, wall materials, etc) and infrastructure. We were limited by publicly-available GN-level data, which does not include other collected data such as: literacy and computer literacy rates, more detailed (and gender-segregated) economic status data, and reasons for migration.

As a large number of census features involve infrastructure, it is crucial to understand the type of housing and building materials found in the regions of the study. The Department of Census classifies building materials into three types: 'ephemeral' (includes use of straw, cadjan leaf, sand), 'semi-durable' (includes clay walls, galvanized sheet roofs, stone floors), and 'permanent' (includes use of tile, asbestos, brick and cement). Mannar district has the highest percentage of temporary and/or shanty units (44.5%) in the country, followed by Kilinochchi (24.4%). A majority of the houses in the Mullaitivu and Kilinochchi districts are built with ephemeral and semi-durable materials. In Western Province, the percentage of permanent houses is at 91.6% (Colombo - 93.6%, Gampaha - 90.5% and Kalutara - 89.9%) with all 3 districts having less than 0.5% 'ephemeral' housing.

The CDR data was available only at a BTS-level, and our census information is reported at the Grama Niladhari (GN) administrative level. As there are far more GNs than BTS towers, mapping the BTS-level CDR data to the GN-level makes all the CDR values similarly small and strips them of interesting statistical properties. For this reason, we use the BTS as our unit of analysis.

To convert the GN-level census proportions to BTS-level census values, we employ a geographic coverage-based mapping method.

$$BTS_{census} = \sum_{i=0}^{n} \frac{val(GN_i, X)}{pop(GN_i)} * ratio_i$$

where $val(GN_i, X)$ = GN-level census value for feature X

$pop(GN_i)$ = Population of GN division

$$ratio_i = \frac{\text{Area of intersection between BTS \& GN}}{\text{Total geographic area of the BTS}}$$

5 STATISTICAL METHODS

For each census and CDR feature pair, we ran: 1) an ANOVA test, which indicates statistically significant different features based on CDR features; and 2) Spearman's rank correlation, to capture the strength and direction of the monotonic relationship. We chose to use Spearman's rank correlation for the analysis, to better capture monotonic (not only linear) relationships.In appendix A, we record the results for pairs that had statistically significant ANOVA results (p-value < 0.001) and their Spearman's ρ.

6 FINDINGS

In general, the mobility CDR features perform best (result in the highest number of statistically significant correlations), followed by social features and, finally, consumption features when both provinces are considered together, which corresponds to the findings by Frias-Martinez and Virseda [6]. However, our results deviate when considering consumption CDR features for Western Province, where the consumption CDR features showed the highest number of statistically significant correlations.

Consumption Features. The consumption features generally demonstrate weak to moderate monotonic relationships to census features in Northern Province. The features related to the number of calls (Total Calls, Total In, and Total Out) provide better results when compared to features related to call duration. The direction of relationships is as expected - in general, greater consumption is positively correlated to features associated with higher SEL (such as having tertiary education (degree) and roofs made of asbestos or concrete), and negatively correlated to census features associated with low SEL, such as living in improvised housing.

Social Features. Within the social features, the relationships between contact count and high SEL features (degree-holders, asbestos and concrete roof materials, tile/granite/terazzo floor materials) exhibit moderate to strong positive correlations. This suggests that having a greater number of contacts corresponds to having a higher SEL. In Western Province, we observe moderate to strong positive relationships between the contact rate, physical distance, and high SEL socio-economic characteristics as well - and vice versa. This conforms to the findings published by Frias-Martinez and Virseda[6]. However, in Northern Province, our results diverge from previous findings with respect to two social features: physical distance and contact rate. There are strong negative relationships between physical distance and the high SEL features. There are also moderate to strong positive relationships between physical distance and characteristics associated with low SEL: improvised housing, cadjan/palmyrah/straw roof materials, and metal sheet roof materials. A similar trend is observed with regard to contact rate as well. This suggests that in Northern Province, BTS regions with users who have greater geographic dispersion among their contacts, and/or who speak to their contacts more regularly, have lower socio-economic characteristics.

Mobility Features. Unique cell count (which reflect averages at the BTS level) is the only mobility feature that produce the expected relationship in Northern Province. Unique cell count shows a moderate to strong positive relationship with some high SEL features, and a moderate to strong negative relationship with low SEL features. However, travel distance, maximum travel distance and radius of gyration which are mobility features that better capture extreme behavior generally demonstrate moderate to strong negative correlations with high SEL features, and moderate to strong positive relationships with low SEL features. It should be noted that for each of these trends, there are outliers. For instance, there is a strong positive relationship with solar lighting (what we assume to be a high SEL feature) and a negative relationship with cabook walls (an assumed low SEL feature), which needs to be further investigated. Nonetheless, this trend suggests that within Northern Province, BTS regions with users who travel further, and/or travel long distances more frequently, have lower SEL features - an inverse relationship to that identified by Frias-Martinez and Virseda [6]. For Western Province, the results align with the findings of prior research[6].

7 REGRESSION MODEL

We built a multivariate linear regression model using the ordinary least squares method to predict the census feature values from the derived CDR features. Predictor and response features were log-transformed to better satisfy the conditions for linear regression models.

We included 14 predictors in the final regression. These included the 13 CDR features and one additional feature, mobile user population density, to control for variations in the number of active mobile users per unit of BTS geographic area of coverage. We reported the highest-performing model, as measured by highest adjusted R^2 values. Figure 2 & 3 reports results for the models with an adjusted R^2 > 0.25. As can be seen in these graphs, the consumption models have the poorest performance for both Northern and Western

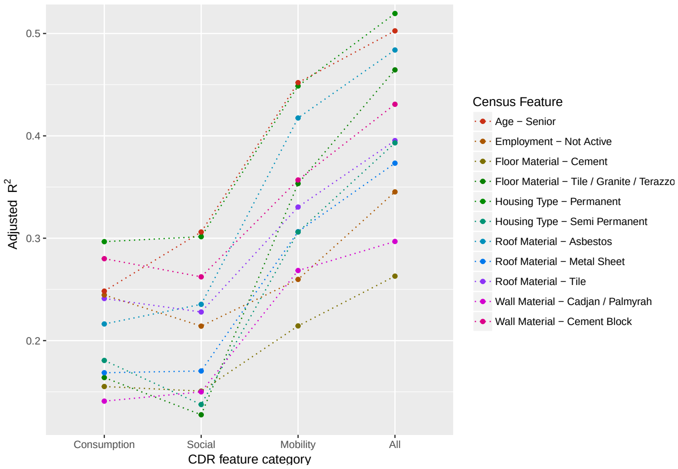

Figure 2: Performance of regression models against different variable types - Northern Province

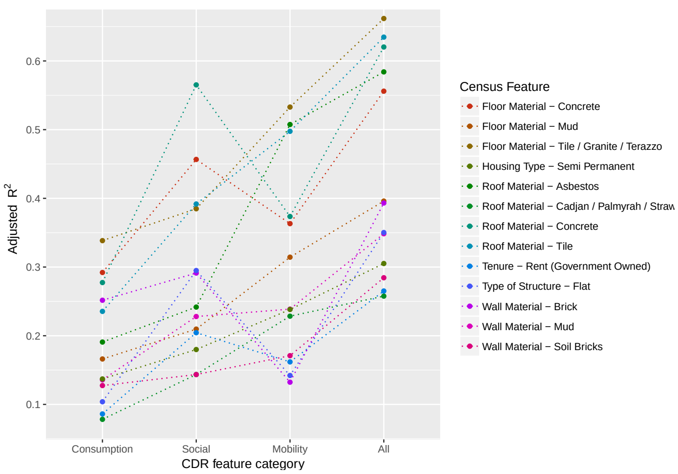

Figure 3: Performance of regression models against different variable types - Western Province

provinces. The predictive power of social and mobility models differ when considering the two provinces. For Northen Province, mobility models perform better than social models, whereas for Western Province, this observation seems to be reversed. We can also observe that different CDR features have different predictive capacity on a given census feature. For instance, in Northern Province, the results suggest that social CDR features are a good predictor of the number of households with permanent housing, but a poor predictor of households with semi-permanent structures. In Western Province, mobility is a very good predictor of the households with tile/granite/terrazo floors, but a poor predictor for the type of the wall material.

8 DISCUSSION

Our research shows some promising findings. First, it suggests that socio-economic levels can affect CDR data in a post-conflict, Sri Lankan setting. In particular, regions with high socio-economic levels may be observed through a greater number of contacts, greater number of unique cells being visited and greater consumption while having a lower values for contact rates, radius of gyration, maximum travel distance, travel distance, and physical distance between contacts. The inverse holds true for regions with lower socio-economic levels. The findings involving radius of gyration, maximum travel distance, travel distance, contact rate and physical distance are especially interesting, as they are the inverse trends to those observed in prior research [6], but only in Northern Province.

These results correspond to assumptions about the Northern Province population under study, which includes a high percentage of a vulnerable, highly mobile group that were displaced due to conflict. It suggests that our dataset, especially through features that capture human movement and geographic spread of networks, may have predictive capability for census features within a post-conflict context as well.

By considering the correlation results from both Northern and Western provinces, we can observe more nuanced patterns as well. For instance, in Northern Province, the number of households with cement floors has a weak positive correlation with unique cell count, and negatively correlated with the radius of gyration and maximum travel distance - a trend that is associated high SEL for Northern Province. However, in Western Province, the number of households with cement floors is negatively correlated with all of the mobility features - a trend that is associated with low SEL. While this discrepancy might be due to limitations in our data and/or methodology, it might also be due to the fact that high or low SEL populations in different regions exhibit different socio-economic characteristics . Additional analysis would be needed to verify whether such nuanced observations can be inferred accurately from CDR data at a regional level.

Our work emphasizes the need to replicate existing research methods within different contexts. CDR features are, ultimately, behavioral features, and these are not constant across time and space. For example, our work indicated a relationship between radius of gyration and low socio-economic features in Northern Province – the inverse of those identified by Frias-Martinez and Virseda [6]. Without understanding the local post-conflict context, our finding may have seemed incorrect. Through such replication in different local contexts, this papers moves the extant knowledge base forward.

However our R^2 values are less than ideal and well below the accuracy levels obtained in previous research [6]. In and of itself, this is not an indication of low utility of these techniques in the Sri Lankan post-conflict context. Rather, it is a reflection of underlying limitations in both our data and our methodology, some of which we can improve upon. These limitations are further articulated below.

8.1 Data Sources

We chose census features that could best illustrate aspects related to socioeconomic levels. However, we were limited by what GN-level census data was available in the public domain. It is possible that model could be more effective in predicting other potentially more appropriate census features, such as proportions of resettled or displaced persons and computer literacy, which are not publicly available. In an ideal scenario we would have had data at the census block level (which is a far smaller spatial unit than BTS coverage area or a GN), to better resemble the data used by Frias-Martinez and Virseda [6]. This would have reduced potential errors in mapping census features to BTS coverage areas.

In Sri Lanka, this data is not available to outside researchers due to the legally-mandated regulations governing the Department of Census and Statistics (DCS). To access such data requires formal collaboration with the DCS .

8.2 Mapping census features to BTS coverage areas

Our mapping method relies solely on geographic coverage. For simplicity, we assumed uniform population density across the overlap areas (i.e. the area overlaps between GNs and voronoi cells). This is problematic and is further compounded by the fact that we had access only to GN level census attributes, and not at a more granular census-block level. A better approach would have been to utilize a kernel density estimation technique when mapping census features to the voronoi cells.

8.3 Prediction Models

We built a basic multivariate linear regression model, controlling just for the number of mobile users per voronoi cell, and did not account for the varying population density that is seen in the Northern Province. In future work, we will control for this. In an ideal scenario, we would conduct a targeted survey to collect demographic and socio-economic data for a subset of the mobile subscribers in our study area. However, given that we have pseudonymized CDR data, matching survey results to CDR records may not be feasible with privacy safeguards.

In addition to the statistical methods employed in this study, we plan to explore machine learning techniques as well and develop models that can enable better feature selection and improve predictive capacity.

9 CONCLUSION

Our analysis demonstrates the potential for mobile network big data to predict census features, both in post-conflict regions as well as in a more urbanized region with a connected population. Our preliminary findings suggest that mobile network big data yields behavioral features that can help observe a specific vulnerable population, those of displaced and recently resettled persons. Applying this methodology to different regions yielded characteristics that are specific to that regional population. Our work also demonstrates the importance of replicating previous research studies in a local context. In this instance, the regional characteristics and the post-conflict situation of the region under study revealed interesting behavioral patterns that diverged from the anticipated result, providing new insight on behavioral norms associated with certain population-level socio-economic characteristics in a post-conflict setting.

ACKNOWLEDGMENTS

The authors would like to thank the Department of Census and Statistics, Sri Lanka for providing census data. This research was funded through a grant from the International Development Research Centre (IDRC) of Canada.

REFERENCES

[1] Linus Bengtsson, Jean Gaudart, Xin Lu, Sandra Moore, Erik Wetter, Kankoe Sallah, Stanislas Rebaudet, and Renaud Piarroux. 2015. Using mobile phone data to predict the spatial spread of cholera. *Scientific reports* 5 (2015), 8923.

[2] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. 2015. Predicting poverty and wealth from mobile phone metadata. *Science* 350, 6264 (2015), 1073–1076.

[3] Adeline Decuyper, Alex Rutherford, Amit Wadhwa, Jean-Martin Bauer, Gautier Krings, Thoralf Gutierrez, Vincent D Blondel, and Miguel A Luengo-Oroz. 2014. Estimating food consumption and poverty indices with mobile phone data. *arXiv preprint arXiv:1412.2595* (2014).

[4] Chris Elbers, Jean O Lanjouw, and Peter Lanjouw. 2003. Micro–level estimation of poverty and inequality. *Econometrica* 71, 1 (2003), 355–364.

[5] United Nations High Commissioners for Refugees. 2012. UNHCR Eligibility Guidelines for Assessing the International Protection Needs of Asylum-Seekers from Sri Lanka. (December 2012). Retrieved March 13, 2018 from http://www.refworld.org/pdfid/50d1a08e2.pdf

[6] Vanessa Frias-Martinez and Jesus Virseda. 2012. On the relationship between socio-economic factors and cell phone usage. In *Proceedings of the fifth international conference on information and communication technologies and development*. ACM, 76–84.

[7] Thoralf Gutierrez, Gautier Krings, and Vincent D Blondel. 2013. Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. *arXiv preprint arXiv:1309.4496* (2013).

[8] Sriganesh Lokanathan, Gabriel E Kreindler, NH Nisana de Silva, Yuhei Miyauchi, Dedunu Dhananjaya, and Rohan Samarajiva. 2016. The Potential of Mobile Network Big Data as a Tool in Colombo's Transportation and Urban Planning. *Information Technologies & International Development* 12, 2 (2016), pp–63.

[9] Xin Lu, Linus Bengtsson, and Petter Holme. 2012. Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences* 109, 29 (2012), 11576–11581.

[10] Department of Census and Statistics. 2012. Census of Population and Housing. (2012). Retrieved March 13, 2018 from http://www.statistics.gov.lk/PopHouSat/CPH2011/Pages/Activities/Reports/FinalReport/FinalReportE.pdf

[11] UN Global Pulse. 2013. Mobile Phone Network Data for Development. (October 2013). Retrieved March 13, 2018 from http://www.unglobalpulse.org/sites/default/files/Mobile%20Data%20for%20Development%20Primer_Oct2013.pdf

[12] Umar Serajuddin, Hiroki Uematsu, Christina Wieser, Nobuo Yoshida, and Andrew Dabalen. 2015. Data deprivation: another deprivation to end. (2015).

[13] Christopher Smith, Afra Mashhadi, and Licia Capra. 2013. Ubiquitous sensing for mapping poverty in developing countries. *Paper submitted to the Orange D4D Challenge* (2013).

[14] Christopher Smith-Clarke, Afra Mashhadi, and Licia Capra. 2014. Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 511–520.

[15] Jessica E Steele, Pål Roe Sundsøy, Carla Pezzulo, Victor A Alegana, Tomas J Bird, Joshua Blumenstock, Johannes Bjelland, Kenth Engø-Monsen, Yves-Alexandre de Montjoye, Asif M Iqbal, et al. 2017. Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface* 14, 127 (2017), 20160690.

[16] Amy Wesolowski, Taimur Qureshi, Maciej F Boni, Pål Roe Sundsøy, Michael A Johansson, Syed Basit Rasheed, Kenth Engø-Monsen, and Caroline O Buckee. 2015. Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proceedings of the National Academy of Sciences* 112, 38 (2015), 11887–11892.

[17] Robin Wilson, Elisabeth zu Erbach-Schoenberg, Maximilian Albert, Daniel Power, Simon Tudge, Miguel Gonzalez, Sam Guthrie, Heather Chamberlain, Christopher Brooks, Christopher Hughes, et al. 2016. Rapid and near real-time assessments of population displacement using mobile phone data following disasters: the 2015 Nepal Earthquake. *PLoS currents* 8 (2016).

APPENDIX A

A.1 Features derived from CDR data

Table 1 shows the aggregate features that were derived from 3 months of CDR data recorded in the beginning of year 2013.

Table 1: 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

A.2 Census features used for analysis

Table 2 shows the census features that were chosen for the study. Feature categories marked with $*$ have data collected at an individual level while all others have data collected at a household level.

A.3 Spearman's rank correlation results

This section reports the Spearman's rank correlation, ρ, for different CDR features. Only values with a p-value below 0.001 are shown. ρ values are represented as below.

$(+), (-)$	$\rho < 0.15$:	very weak
$+, -$	$0.15 < \rho < 0.25$:	weak
$++, --$	$0.25 < \rho < 0.40$:	moderate
$+++, ---$	$0.40 < \rho < 0.75$:	strong

To interpret the tables, we can take the example of [Floor Materials - Tile/Granite/Terrazzo, Contact Count] for Northern province. We first assign each BTS to its respective quartile based on its average contact count. We then run an ANOVA test (also known as an F-test statistic) to compare the between-group variability with within-group variability for the Floor Materials - Tile/Granite/Terrazzo values of each group (where each 'group' is made up of all BTS's within a quartile). Census features that can not be divided into quartiles are not included in this analysis. The ρ value is represented by ++, indicating a moderate positive relationship ($0.25 < \rho < 0.39$). Together, the tests suggest that there is a relationship between Contact Count and Floor Materials - Tile/Granite/Terrazzo and, as the number of contacts increase, so does the proportion of households with tile, granite or terrazzo floors. In other words, it is an indication of higher socio-economic levels.

Tables 3 - 8 show Spearman's ρ between different CDR feature categories and the socio-economic characteristics. Features are color-coded if they have two or more moderate to strong correlations that move according to the general trends. Rows are shaded in blue if the census features are mostly associated with higher SEL characteristics, green if they are mostly associated with lower SEL characteristics.

Table 2: List of chosen census features

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

A.4 Regression analysis results

The results from the regression analysis where all CDR features were used is given in tables 9 & 10.

Table 3: Correlation of consumption CDR features for Northern province - Blue : All consumption features are positively correlated with the census feature - Green : All consumption features are negatively correlated with the census feature

Census Feature Category	Census Feature	Total Calls	Total Calls In	Total Calls Out	Total Duration	Total Duration In	Total Duration Out
Floor Material	Tile / Granite / Terazzo	++	+ + +	++			
	Cement				−−		−
	Concrete		++				
Roof Material	Asbestos	++	+ + +				
	Concrete	+ + +	+ + +	++			
	Tile				−−		−−
	Aluminium Sheet				−		−−
	Metal Sheet	−	−−				
	Cadjan / Palmyrah / Straw	−	−	−			
	Other				++		+
Wall Material	Brick	++	++				
	Cement Block		++		−		−
	Cabook		+		−		
	Soil Bricks		++				
	Cadjan / Palmyrah	−−	−−	−−			
Type of Structure	Single (1 Storey)		+				
	Single (2 Storey)	++	++	++			
	Attached House / Annex	+	+				
	Flat	++	+	+			
	Twin House	++	++	++			
	Row / Line Room	+	+				
Housing Type	Permanent	+	++		−		−
	Semi Permanent	−	−−				
	Improvised	−−	−−	−			
Tenure	Rent (Government Owned)	++	++	+			
	Rent (Privately Owned)	+ + +	+ + +	++			
	Rent Free		++				−
	Encroached	++	++	++			
	Other	++	++	+			
Cooking Fuel	Kerosene	++	++				
	Gas	++	+ + +	++			
	Electricity	++	++	++			
	Dust	++	+ + +	++			
	Other	++	++	++			
Lighting	Electricity	++	++				
	Solar Power		−				
	Other	++	++	+			
Education Level	Degree	++	+ + +	+			
	A/L	++	+ + +				
	O/L	++	++				
	Secondary		+				−
	Primary				−		−
Employment	Employed				−		
	Unemployed				−		−
	Not Active		++		−		−−
Gender	Male				−		
	Female				−−		−−
Age	Young			−	−−		−
	Middle Aged				−		−
	Senior		++		−		−−

Table 4: Correlation of consumption CDR features for Western province - Blue : All consumption features are positively correlated with the census feature - Green : All consumption features are negatively correlated with the census feature

Census Feature Category	Census Feature	Total Calls	Total Calls In	Total Calls Out	Total Duration	Total Duration In	Total Duration Out
Floor Material	Tile / Granite / Terazzo	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +
	Wood	++	++	+ + +	+ + +	++	+ + +
	Cement	− − −	− − −	− − −	− − −	− − −	− − −
	Concrete	− − −	− − −	− − −	− − −	− − −	− − −
	Mud	− − −	− − −	− − −	− − −	− − −	− − −
	Sand	−−	−	−−	−−	−−	−−
Roof Material	Asbestos	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +
	Concrete	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +
	Tile	− − −	− − −	− − −	− − −	− − −	− − −
	Metal Sheet	−−	−−	−−	−−	−−	−−
	Cadjan / Palmyrah / Straw	− − −	− − −	− − −	− − −	− − −	− − −
	Other					(−)	
Wall Material	Brick	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +
	Cement Block	−−	−−	−	−−	−−	−−
	Soil Bricks	−−	−−	−−	− − −	−−	− − −
	Mud	− − −	− − −	− − −	− − −	− − −	− − −
	Cadjan / Palmyrah	−−	−	−−	−−	−−	−−
	Other	−−	−−	−−	−−	−−	−−
Type of Structure	Single (1 Storey)	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +
	Single (2 Storey)	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +
	Single (2+ Storey)	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +
	Attached House / Annex	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +
	Flat	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +
	Twin House	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +
	Row / Line Room	++	++	++	++	++	++
	Hut / Shanty	++	++	++	++	++	++
Housing Type	Permanent	++	++	++	++	++	++
	Semi Permanent	− − −	− − −	− − −	− − −	− − −	− − −
	Improvised	−−	−−	−−	−−	−−	−−
	Unclassified	+	(+)	+	+	+	+
Tenure	Owned	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +
	Rent (Government Owned)	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +
	Rent (Privately Owned)	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +
	Rent Free	+ + +	++	+ + +	+ + +	+ + +	+ + +
	Encroached	+	+	++	++	+	++
	Other	+ + +	++	+ + +	+ + +	+ + +	+ + +
Cooking Fuel	Kerosene	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +
	Gas	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +
	Electricity	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +
	Dust	++	+	++	++	++	++
	Other	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +
Lighting	Electricity	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +
	Other	++	++	++	++	++	++
Education Level	Degree	+ + +	++	+ + +	+ + +	+ + +	+ + +
	A/L	+	+	+	++	++	++
	Secondary	−−	−−	−−	−−	−−	−−
	Primary	−−	−	−−	−−	−−	−−
	No Schooling	−	−	−	−	−−	−
Employment	Unemployed	−−	−−	−−	−−	−−	−−
Age	Young	−	−	−−	−−	−−	−−
	Middle Aged	+	+	+	++	++	++
	Senior				(+)	(+)	(+)

Table 5: Correlation of social CDR features for Northern province - Blue : Correlation for contact count is positive, contact rate is negative, and physical distance is negative - Green : Correlation for contact count is negative, contact rate is positive, and physical distance is positive

Census Feature Category	Census Feature	Contact Count	Contact Rate	Physical Distance
Floor Material	Tile / Granite / Terazzo	++	−−	−−−
	Cement		−−	−−
	Concrete	++		−−
Roof Material	Asbestos	++	−−	−−−
	Concrete	++		−−
	Tile		−−	−−−
	Aluminium Sheet		−	−−
	Metal Sheet	−	++	+++
	Cadjan / Palmyrah / Straw			+
	Other		+	++
Wall Material	Brick	++		−−
	Cement Block	+	−−−	−−−
	Cabook		−−	−−−
	Soil Bricks	++		−−
	Cadjan / Palmyrah	−−		+++
	Plank / Metal Sheet			++
	Other			++
Type of Structure	Single (1 Storey)	+	−−	−−
	Single (2 Storey)	++		−−
	Attached House / Annex	+		
	Flat	+		
	Twin House	++		−
	Row / Line Room	++		
	Hut / Shanty			++
Housing Type	Permanent	++	−−	−−−
	Semi Permanent		+	+++
	Improvised	−		++
Tenure	Rent (Government Owned)	++		−
	Rent (Privately Owned)	+++	−−	−−−
	Rent Free	+	−−	−−−
	Encroached	++		
	Other	++		−
Cooking Fuel	Gas	++		−−−
	Electricity	++		−−
	Dust	++	−	−−−
	Kerosene	++		
	Other	++		−−
Lighting	Electricity	++	−	−−−
	Kerosene			+
	Solar Power			++
	Other	++		
Education Level	Degree	+++	−−	−−−
	A/L	+++	−	−−−
	O/L	++	−	−−−
	Secondary	+		−−
	Primary			−−
	No Schooling			−−
Employment	Unemployed		−	−
	Not Active	++	−−	−−−
Gender	Male		−	
	Female		−−	−−−
Age	Middle Aged		−−	−−
	Senior	++	−−−	−−−

Table 6: Correlation of social CDR features for Western province - Blue : Correlation of census feature with all 3 CDR features is positive - Green : Correlation of census feature with all 3 features is negative

Census Feature Category	Census Feature	Contact Count	Contact Rate	Physical Distance
Floor Material	Tile / Granite / Terazzo	+++	(+)	++
	Wood	++		++
	Cement	−−−		(−)
	Concrete	−−−		−−−
	Mud	−−−		−−
	Sand	−		−−
	Other			−
Roof Material	Asbestos	++	++	+++
	Concrete	+++	+	+++
	Tile	−−−	−−	−−−
	Metal Sheet	−−		(−)
	Cadjan / Palmyrah / Straw	−−	−	−−
Wall Material	Brick	+++	−	
	Cement Block	−−−	++	
	Cabook		+	
	Soil Bricks	−−−		−−
	Mud	−−−		−−
	Cadjan / Palmyrah	−	−	−−
	Plank / Metal Sheet		(−)	
	Other	−−		−−
Type of Structure	Single (1 Storey)	++	+	++
	Single (2 Storey)	+++	+	+++
	Single (2+ Storey)	+++	+	+++
	Attached House / Annex	+++	++	+++
	Flat	+++	(+)	+++
	Twin House	++	+	+++
	Row / Line Room	+	(+)	+++
	Hut / Shanty	++		+
Housing Type	Permanent	++		+
	Semi Permanent	−−−		−−
	Improvised	−−		−−
Tenure	Owned	+++	+	++
	Rent (Government Owned)	+++	+	+++
	Rent (Privately Owned)	+++	+	+++
	Rent Free	++	(+)	++
	Encroached	+		++
	Other	++	+	++
Cooking Fuel	Fire Wood		+	
	Kerosene	+++	+	+++
	Gas	+++	+	+++
	Electricity	+++	+	+++
	Dust	+	+	+
	Other	+++	+	+++
Lighting	Electricity	+++	+	++
	Other	++	+	+
Education Level	Degree	++		+
	A/L	+		(+)
	Secondary	−	−	
	Primary	−	−	(−)
	No Schooling		−	
Employment	Employed			(+)
	Unemployed	−−		−
	Not Active		(−)	
Age	Young	−		−
	Middle Aged	(+)	+	+

Table 7: Correlation of mobility CDR features for Northern province - Blue : Correlation for radius of gyration is negative, unique cell count is positive, travel distance is negative, and maximum travel distance is negative - Green : Correlation for radius of gyration is positive, unique cell count is negative, travel distance is positive, and maximum travel distance is positive

Census Feature Category	Census Feature	Radius of Gyration	Unique Cell Count	Travel Distance	Max. Distance
Floor Material	Tile / Granite / Terazzo	− − −	+ + +		−−
	Cement	−−		−−	− − −
	Concrete	−−	+		−−
	Sand		−		
Roof Material	Asbestos	− − −	+ + +		− − −
	Concrete	−−	+ + +		
	Tile	−−		−−	− − −
	Aluminium Sheet	−−		−−	−−
	Metal Sheet	+ + +	− − −		++
	Cadjan / Palmyrah / Straw	+	− − −		
	Other	++			++
Wall Material	Brick	−−	++		−
	Cement Block	− − −	++	−−	− − −
	Cabook	− − −	++		− − −
	Soil Bricks	−			−
	Cadjan / Palmyrah	+ + +	− − −		++
	Plank / Metal Sheet	++	−−		++
	Other	++	−		
Type of Structure	Single (1 Storey)	−−	++		−−
	Single (2 Storey)	−−	++		
	Attached House / Annex		++		
	Flat		++		
	Twin House	−	++		
	Hut / Shanty	++	−−		++
Housing Type	Permanent	− − −	+ + +	−	− − −
	Semi Permanent	++	− − −		++
	Improvised	++	− − −		
Tenure	Rent (Government Owned)	−−	++		
	Rent (Privately Owned)	− − −	+ + +		−−
	Rent Free	− − −	++	−	− − −
Cooking Fuel	Gas	− − −	+ + +		−
	Electricity	−−	+ + +		
	Dust	− − −	+ + +		−−
	Other	−−	++		
Lighting	Electricity	− − −	+ + +		−−
	Kerosene	+	− − −		
	Solar Power	++	− − −		++
Education Level	Degree	− − −	+ + +		−−
	A/L	− − −	+ + +		−−
	O/L	−−	++		−−
	Secondary	−−	+	−	−−
	Primary	−−		−	−−
	No Schooling	−			−
Employment	Employed			−	
	Unemployed			−−	−−
	Not Active	− − −	++	−−	− − −
Gender	Male			−−	−−
	Female	−−	+	−−	− − −
Age	Young		−−	−−	
	Middle Aged	−−	+		−−
	Senior	− − −	+ + +	−−	− − −

Table 8: Correlation of mobility CDR features for Western province - Blue : Correlation of census feature with all 4 CDR features is positive - Green : Correlation of census feature with all 4 CDR features is negative

Census Feature Category	Census Feature	Radius of Gyration	Unique Cell Count	Travel Distance	Max. Distance
Floor Material	Tile / Granite / Terazzo	++	+ + +	+ + +	+ + +
	Wood	+	++	++	++
	Cement	−	− − −	− − −	−−
	Concrete	− − −	− − −	− − −	− − −
	Mud	−−	− − −	− − −	− − −
	Sand	−	−−	−−	−
	Other	−			
Roof Material	Asbestos	++	+ + +	+ + +	++
	Concrete	++	+ + +	+ + +	+ + +
	Aluminium Sheet	(−)		(−)	−
	Tile	−−	− − −	− − −	−−
	Metal Sheet	−	−−	− − −	−−
	Cadjan / Palmyrah / Straw	−−	− − −	− − −	−−
	Other		(−)	−	
Wall Material	Brick		++	++	++
	Cement Block		−	−	−
	Soil Bricks	−	−−	−−	−−
	Mud	−−	−−	−−	−−
	Cadjan / Palmyrah	−−	−−	−−	−−
	Plank / Metal Sheet	−		−	−
	Other	−	−−	−−	−−
Type of Structure	Single (1 Storey)	(+)	+ + +	++	+
	Single (2 Storey)	++	+ + +	+ + +	+ + +
	Single (2+ Storey)	++	+ + +	+ + +	+ + +
	Attached House / Annex	++	+ + +	+ + +	++
	Flat	++	+ + +	+ + +	+ + +
	Twin House	+	+ + +	++	++
	Row / Line Room	+	+	+	(+)
	Hut / Shanty		++	+	
Housing Type	Permanent	+	++	++	++
	Semi Permanent	−−	− − −	− − −	− − −
	Improvised	−−	−−	−−	−−
	Unclassified		+		
Tenure	Owned	+	+ + +	+ + +	++
	Rent (Government Owned)	++	+ + +	+ + +	++
	Rent (Privately Owned)	++	+ + +	+ + +	++
	Rent Free	+	+ + +	++	+
	Encroached	+	+	+	+
	Other	(+)	+ + +	++	+
Cooking Fuel	Fire Wood		+		
	Kerosene	++	+ + +	+ + +	++
	Gas	++	+ + +	+ + +	++
	Electricity	++	+ + +	+ + +	+ + +
	Dust	(+)	++	+	(+)
	Other	++	+ + +	+ + +	++
Lighting	Electricity	+	+ + +	+ + +	++
	Other		++	+	
Education Level	Degree	+	+ + +	+ + +	++
	A/L	+	+	++	++
	Secondary		− − −	−−	−
	Primary	(−)	− − −	−−	−
	No Schooling		−−	−−	
Employment	Unemployed	−	−−	−−	−−
Age	Young	−	−−	−−	−−
	Middle Aged	(+)	++	++	+
	Senior				(+)

Table 9: Adjusted R^2 (Coefficient of determination) for final regression analysis - Northern province

Census Feature Group	Census Feature	Adjusted R^2	RSS
Floor Material	Tile / Granite / Terazzo	0.464	0.07
	Wood	0.053	0.00
	Cement	0.263	8.35
	Concrete	0.045	0.04
	Mud	0.138	1.72
	Sand	0.156	0.71
	Other	0.013	0.00
Roof Material	Tile	0.395	6.73
	Asbestos	0.484	3.00
	Concrete	0.163	0.05
	Aluminium Sheet	0.038	0.01
	Metal Sheet	0.373	5.42
	Cadjan / Palmyrah / Straw	0.171	1.29
	Other	0.197	0.09
Wall Material	Brick	0.125	1.57
	Cement Block	0.431	10.73
	Cabook	0.190	0.00
	Soil Bricks	0.098	0.01
	Mud	0.110	0.82
	Cadjan / Palmyrah	0.297	2.45
	Plank / Metal Sheet	0.223	0.83
	Other	0.200	0.21
Type of Structure	Single (1 Storey)	0.007	975.85
	Single (2 Storey)	0.079	1.43
	Single (2+ Storey)	0.043	0.14
	Attached House / Annex	0.057	0.06
	Flat	0.057	0.00
	Twin House	0.074	0.00
	Row / Line Room	0.023	2.61
	Hut / Shanty	0.116	21.23
Housing Type	Permanent	0.520	10.31
	Semi Permanent	0.393	5.28
	Improvised	0.168	0.65
	Unclassified	0.150	0.00
Tenure	Owned	0.003	3832.84
	Rent (Government Owned)	0.018	15.74
	Rent (Privately Owned)	0.025	133.71
	Rent Free	0.006	50.62
	Encroached	0.011	7.74
	Other	0.031	3.47
Cooking Fuel	Fire Wood	0.001	3639.04
	Kerosene	0.012	40.37
	Gas	0.023	418.72
	Electricity	0.037	0.03
	Dust	0.046	0.01
	Other	0.059	0.06
Lighting	Electricity	0.016	6176.80
	Kerosene	0.089	176.18
	Solar Power	0.092	4.95
	Bio Gas	0.021	0.00
	Other	0.033	0.00
Education Level	Degree	0.069	7.66
	A/L	0.062	57.13
	O/L	0.047	165.91
	Secondary	0.037	1059.40
	Primary	0.022	370.54
	No Schooling	0.013	3.00
Employment	Employed	0.132	1.74
	Unemployed	0.102	0.09
	Not Active	0.345	2.89
Gender	Male	0.188	2.09
	Female	0.307	2.41
Age	Young	0.195	1.32
	Middle Aged	0.219	2.46
	Senior	0.503	0.28

Table 10: Adjusted R^2 (Coefficient of determination) for final regression analysis - Western province

Census Feature Group	Census Feature	Adjusted R^2	RSS
Floor Material	Tile / Granite / Terazzo	0.662	3.99
	Wood	0.177	0.01
	Cement	0.204	12.00
	Concrete	0.556	0.36
	Mud	0.396	0.09
	Sand	0.086	0.00
	Other	0.029	0.00
Roof Material	Tile	0.635	13.19
	Asbestos	0.584	12.64
	Concrete	0.620	1.77
	Aluminium Sheet	0.040	0.04
	Metal Sheet	0.168	0.74
	Cadjan / Palmyrah / Straw	0.258	0.02
	Other	0.031	0.00
Wall Material	Brick	0.393	16.75
	Cement Block	0.224	16.30
	Cabook	0.100	1.77
	Soil Bricks	0.284	0.24
	Mud	0.348	0.35
	Cadjan / Palmyrah	0.200	0.00
	Plank / Metal Sheet	0.063	0.57
	Other	0.063	0.01
Type of Structure	Single (1 Storey)	0.071	18542.10
	Single (2 Storey)	0.153	1839.10
	Single (2+ Storey)	0.243	9.65
	Attached House / Annex	0.041	267.20
	Flat	0.350	16.57
	Twin House	0.078	1.87
	Row / Line Room	0.107	203.99
	Hut / Shanty	0.065	4.90
Housing Type	Permanent	0.120	16.93
	Semi Permanent	0.305	1.25
	Improvised	0.109	0.01
	Unclassified	0.037	0.00
Tenure	Owned	0.114	13729.26
	Rent (Government Owned)	0.265	24.39
	Rent (Privately Owned)	0.166	857.29
	Rent Free	0.122	12.52
	Encroached	0.030	27.67
	Other	0.052	3.84
Cooking Fuel	Fire Wood	0.030	4652.96
	Kerosene	0.243	181.46
	Gas	0.221	7107.79
	Electricity	0.166	0.59
	Dust	0.035	0.07
	Other	0.077	6.92
Lighting	Electricity	0.137	21633.82
	Kerosene	0.034	16.34
	Solar Power	0.130	0.01
	Bio Gas	0.023	0.00
	Other	0.075	0.03
Education Level	Degree	0.023	19808.99
	A/L	0.014	577717.12
	O/L	0.016	228707.40
	Secondary	0.013	390164.99
	Primary	0.015	72516.20
	No Schooling	0.020	1648.53
Employed	Employed	0.110	6.74
	Unemployed	0.121	0.04
	Not Active	0.077	4.63
Gender	Male	0.076	4.29
	Female	0.068	5.42
Age	Young	0.110	1.83
	Middle Aged	0.115	6.74
	Senior	0.152	0.62