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

Aparna Surendra, Sriganesh Lokanathan, Thavisha Perera-Gomez

1. INTRODUCTION

The availability of accurate, timely, disaggregated, and comparable socio-economic data is crucial for effective policymaking, especially with regard to 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 a poverty data (Serajuddin et al. 2015). Our research seeks to determine the opportunity for mobile phone data to provide a reliable, cheap proxy for census data within Sri Lanka, especially in post-conflict regions that have a greater need for frequent data collection.

Mobile phone meta-data such as Call Detail Records (CDRs) can broadly describe three dimensions of human behavior: social networks, consumption activity, and mobility (UN Global Pulse, 2013). 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) – which could provide novel opportunities for targeted policy design, implementation, and evaluation. 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. If CDRs can accurately predict Sri Lankan socio-economic characteristics, policymakers will have access to a wealth of reliable, timely data on which to base policy.

Our paper seeks to identify relationships between features derived from Sri Lankan census and CDR data, replicating methods used by Frias-Martinez and Virseda (2012) in relation to an unidentified Latin American country. We use the 2011/12 Sri Lanka census data and CDR data for approximately 600,000 mobile phone subscribers from Sri Lanka's Northern province, which is a post-conflict region. We 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 (Decuyper et al. 2014), disasters (Lu et al. 2012; Wilson et al. 2016), and disease propagation (Bengtson et al. 2015; Wesolowski, et al. 2015). In terms of understanding economic activity, numerous studies (Frias-Martinez et al. 2012; Smith et al., 2012; Blumenstock et al. 2015) 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 (Gutierrez et al. 2013; Decuyper et al. 2014) or with other data sources such as satellite data (Steele et al. 2017). Together, these studies highlight CDR's potential to complement traditional sources of data.

Frias-Martinez and Virseda (2012) identified statistically- significant correlates from CDRs and census socio-economic features. Their research indicates strong linkages between populations with higher socioeconomic levels (SEL) and range of mobility. Using somewhat related methodology, Smith-Clarke et al. (2012) 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. (2015) 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 micro-regional data prevented micro-region validation. Further, Steele et al. (2017) showed that CDR data when combined with remote-sensing data was better positioned to model traditional measures of poverty at disaggregated geographic levels. Guiterrez et al. (2013) 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 that 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" (Frias-Martinez and Virseda, 2012) 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 first 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 most sophisticated studies (e.g. Blumenstock et al., 2015) supplement big data with other sources, such as survey data.

Sri Lanka's Northern and Eastern Provinces were the most severely affected by the civil war, which ended in 2009. We chose to focus this research on the Northern Province, as 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).

3. BACKGROUND

The 2011/2012 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. It is Sri Lanka's least populous province, but population density varies widely. 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 Chart 1), and is one of the more densely-populated districts in the country.

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' ('Census of Population and Housing', 2012). 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 the northern Districts ('UNHCR Eligibility Guidelines', 2012). As such, the area under study 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 District Secretariat Division (DSD) level, calculated using the small-area estimation method developed by Elbers, Lanjouw, and Lanjouw (2003). However, this method still allows a three-year lag in population distribution reporting, and may not capture spatial heterogeneity.

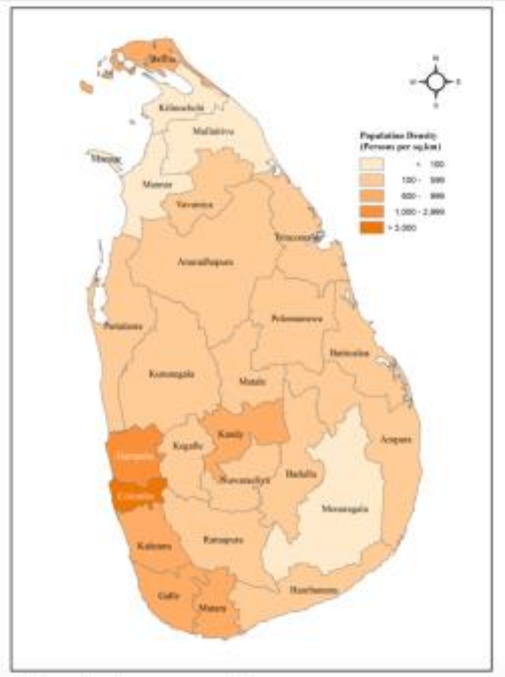


Chart 1, from 'Preliminary Report: Population of Sri Lanka By District'(Census and Statistics, 2012).

4. DESCRIPTION OF DATASETS

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 on the call; (3) The date and time 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 ID is mounted on a mobile phone tower (base trans-receiver station or BTS), which we can dereference to a physical (latitude, longitude) location. We only use CDRs associated with calls for the study.

We used the CDRs for approximately 600,000 subscribers identified as living in the Northern Province. The residence of each subscriber was determined as per the techniques in Lokanathan et al. (2016). We co-related BTS-level mobile subscriber population data with mapped census population data, which suggests that the dataset includes people with more than one mobile SIM (the true number of mobile phone subscriber 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 BTS level, as well as the approximated GN level).

The CDR features are loosely divided into three categories: those that describe user's phone call behavior (Consumption), their social network (Social), and their geographic movement (Mobility).

Consumption Features

- Total In - Total number of calls received (i.e. incoming calls)
- Total Out - Total number of calls made (i.e. outgoing calls)
- 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 Features

- 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 Features

- 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. (This can be considered the distance between home and work).
- Maximum Distance– The maximum distance between the BTSs typically visited by a person.

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, with over half of those in the Jaffna District. 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 2600 (Jaffna).

Census Data

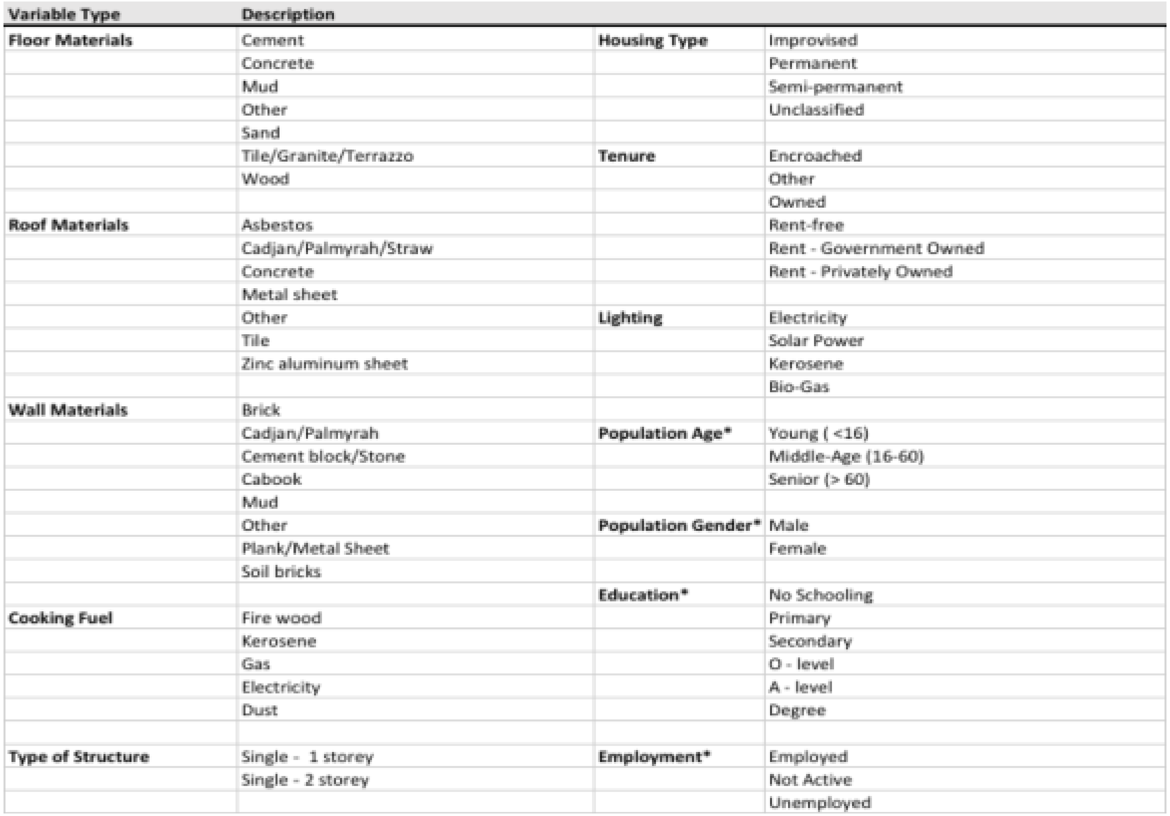
For high-resolution ‘ground truth’ population data, we decided to use census results. While census results are widely-reported at the district level (2nd level administrative unit), some data are made publicly-available at the Grama Niladhari (GN) division (4th and lowest level administrative unit).

The 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. We used 56 features from 12 categories for our analysis (Table 2). The Sri Lankan census focuses on population and housing, and does not collect or calculate figures to related to income, assets, or consumption. Ideally, we would compare CDR data with poverty-specific features. For instance, Frias-Martinez and Virseda (2012) 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 (Young, Middle-Age, and Senior – each calculated using the 5-year interval population age 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 Northern Province. 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.



Variable Type	Description		
Floor Materials	Cement	Housing Type	Improvised
	Concrete		Permanent
	Mud		Semi-permanent
	Other		Unclassified
	Sand		
	Tile/Granite/Terrazzo	Tenure	Encroached
	Wood		Other
			Owned
Roof Materials	Asbestos		Rent-free
	Cadjan/Palmyrah/Straw		Rent - Government Owned
	Concrete		Rent - Privately Owned
	Metal sheet		
	Other	Lighting	Electricity
	Tile		Solar Power
	Zinc aluminum sheet		Kerosene
			Bio-Gas
Wall Materials	Brick		
	Cadjan/Palmyrah	Population Age*	Young (<16)
	Cement block/Stone		Middle-Age (16-60)
	Cabook		Senior (> 60)
	Mud		
	Other	Population Gender*	Male
	Plank/Metal Sheet		Female
	Soil bricks		
		Education*	No Schooling
Cooking Fuel	Fire wood		Primary
	Kerosene		Secondary
	Gas		O - level
	Electricity		A - level
	Dust		Degree
Type of Structure	Single - 1 storey	Employment*	Employed
	Single - 2 storey		Not Active
			Unemployed

Chart 2 - List of Chosen Census Features.

**- Data collected at the individual level. (Features without asterisks are collected at the household level.)*

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.

$BTS_{census} = \Sigma\, GN_{i\ x}\, ratio_i$ $GN_i =$ GN-level census value

$$ratio_i = \frac{\text{geographic area of intersection between the BTS and GN}}{\text{total geographic area of the GN}}$$

5. STATISTICAL METHODS

For each census and CDR feature pair, we ran: 1) an ANOVA test, which indicates statistically significant differences features (within the census data) 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 Charts 4, 5,and 6, we record the results for pairs that had statistically significant ANOVA results (p-value < 0.001, indicated with ***) and their Spearman's ρ. Only results with p < 0.001 are considered in the analysis.

To interpret the tables, we can take the example of [Floor Materials – Tile/Granite/Terrazzo, Contact Count]. 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 BTSs within a quartile). Census features that can not be divided into quartiles are not included in this analysis. The p < 0.001 value (represented by ***) indicates a statistically significant difference in variances. The ρ value is represented by ++, indicating a moderate positive relationship (0.25<ρ <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.

6. FINDINGS

The mobility CDR features perform best (result in the highest number of statistically significant ANOVA tests), followed by social features and, finally, consumption features. This pattern corresponds to findings by Frias-Martinez and Virseda (2012).

Consumption Features

The consumption features yield few statistically significant ANOVA test results, and demonstrate weak monotonic relationships to census features. 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/terrazzo floor materials) exhibit strong or moderate positive correlations. This suggests that having a greater number of contacts corresponds to having a higher SEL – a trend that is true in the Frias-Martinez and Virseda (2012) findings, as well.

However, our results diverge from previous findings with respect to the two other social features: physical distance and contact rate. There are strong negative relationships between physical distance and the high SEL features (now including permanent housing). There are also strong or moderate positive relationships between physical distance and characteristics associated with low SEL: improvised housing, cadjan/palmyrah/straw roof materials, and metal sheet roof materials. Similar trends are seen with contact rate, which has moderate negative relationships between contact rate and higher socio-economic features, and moderate positive relationships with lower socio-economic features. This suggests that 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.

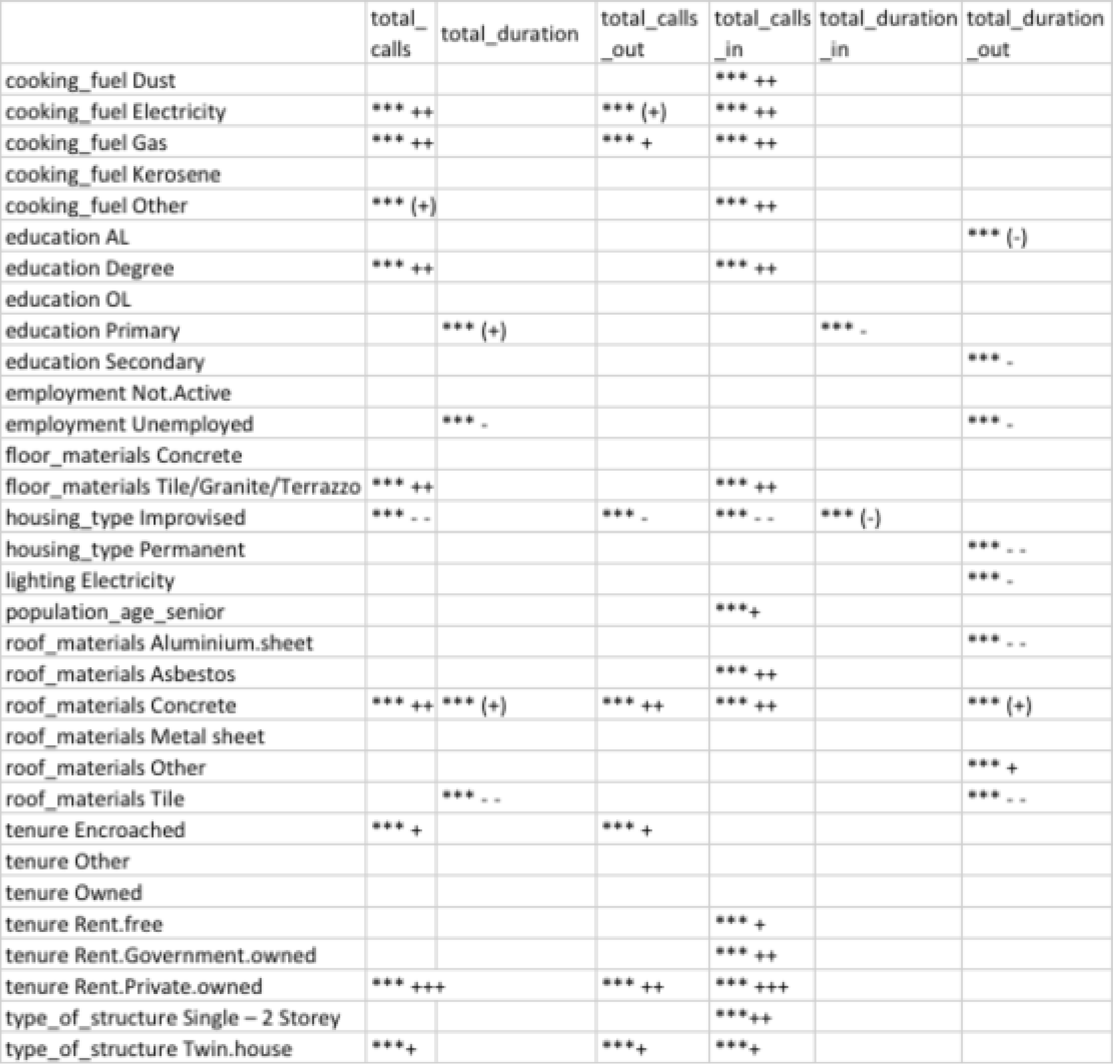
Mobility Features

Unique cell count and travel distance (both reflect averages at the BTS level) are two mobility features that produce expected relationships. In general, unique cell count has a strong or moderate positive relationship with some high SEL features (electricity, gas, permanent housing, tile/granite/terrazzo floors and degrees), and a strong or moderate negative relationship with low SEL features (improvised or semi-permanent housing, cadjan/palmyrah/straw walls or plank/metal walls). The trend suggests that greater mobility indicates higher SEL. While travel distance does not yield as many statistically-significant results, it generally moves in a similar direction. This corresponds to the Frias-Martinez and Virseda (2012) findings.

However, maximum travel distance and radius of gyration, the two mobility features that better capture extreme behavior, generally demonstrate strong or moderate negative correlations with high SEL features: degrees, tile/granite/terrazzo floors and electricity (only for radius of gyration; maximum distance does not have a statistically-significant ANOVA result for this feature). In addition, they generally demonstrate strong or moderate positive relationships with low SEL features: improvised or semi-permanent housing, kerosene lighting, mud floors, and cadjan/palmyrah/straw walls. 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 need to be further investigated. Nonetheless, the trend suggests that BTS regions with users who travel further, and/or travel long distances more frequently, have lower SEL features – this is the inverse relationship to that identified by Frias-Martinez and Virseda (2012).

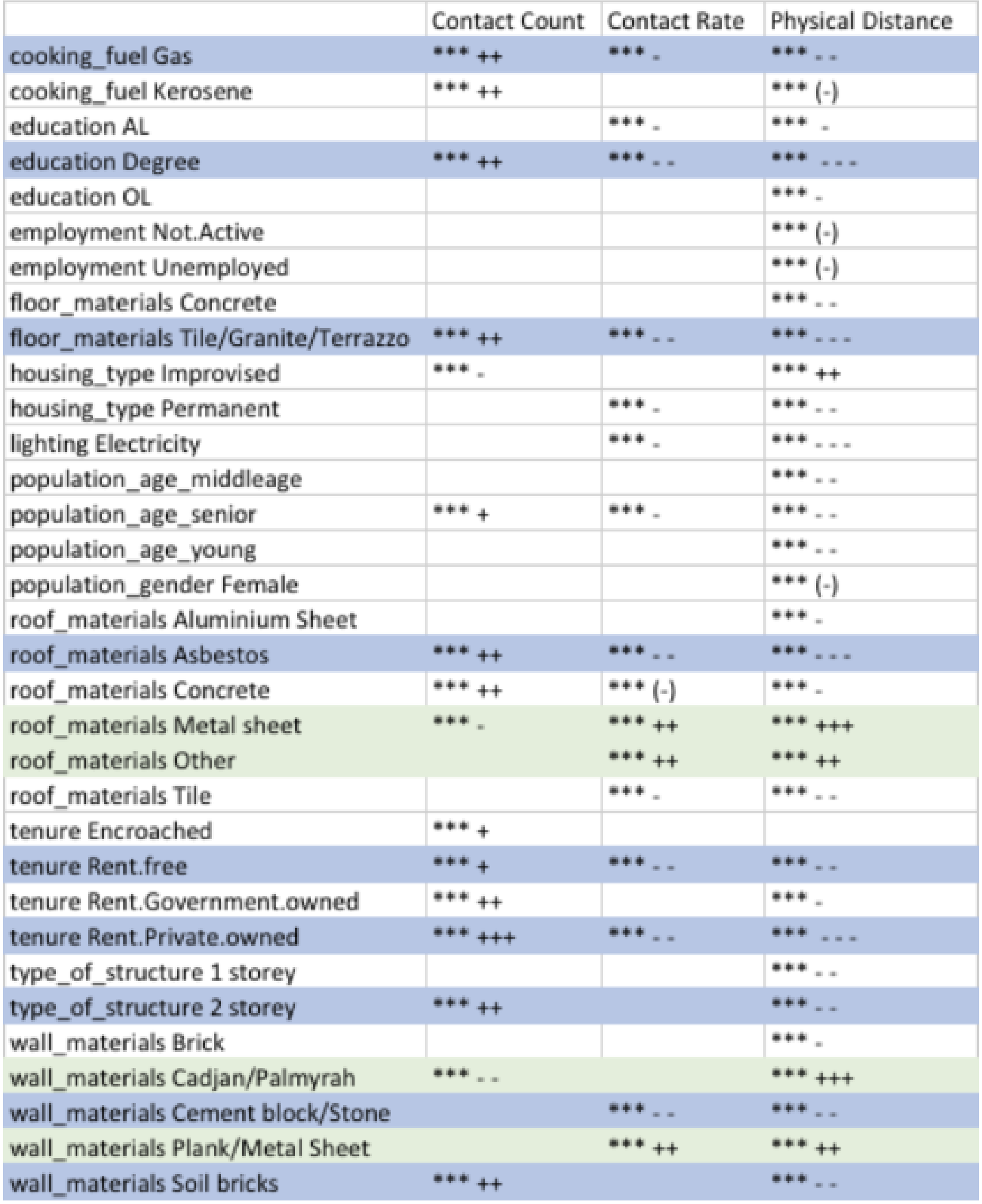
ANOVA test results
*** - $p < 0.001$

Spearman’s Rank Correlation
(+),(-) $\rho < 0.15$: very weak
+, - $0.15 < \rho < 0.25$: weak
++, - - $0.25 < \rho < 0.40$: moderate
+++, - - - $0.40 < \rho < 0.75$: strong



	total_ calls	total_duration	total_calls _out	total_calls _in	total_duration _in	total_duration _out
cooking_fuel Dust				*** ++		
cooking_fuel Electricity	*** ++		*** (+)	*** ++		
cooking_fuel Gas	*** ++		*** +	*** ++		
cooking_fuel Kerosene						
cooking_fuel Other	*** (+)			*** ++		
education AL						*** (-)
education Degree	*** ++			*** ++		
education OL						
education Primary		*** (+)			*** -	
education Secondary						*** -
employment Not.Active						
employment Unemployed		*** -				*** -
floor_materials Concrete						
floor_materials Tile/Granite/Terrazzo	*** ++			*** ++		
housing_type Improvised	*** - -		*** -	*** - -	*** (-)	
housing_type Permanent						*** - -
lighting Electricity						*** -
population_age_senior				***+		
roof_materials Aluminium.sheet						*** - -
roof_materials Asbestos				*** ++		
roof_materials Concrete	*** ++	*** (+)	*** ++	*** ++		*** (+)
roof_materials Metal sheet						
roof_materials Other						*** +
roof_materials Tile		*** - -				*** - -
tenure Encroached	*** +		*** +			
tenure Other						
tenure Owned						
tenure Rent.free				*** +		
tenure Rent.Government.owned				*** ++		
tenure Rent.Private.owned	*** +++		*** ++	*** +++		
type_of_structure Single – 2 Storey				***++		
type_of_structure Twin.house	***+		***+	***+		

Chart 4 – Consumption CDR Features



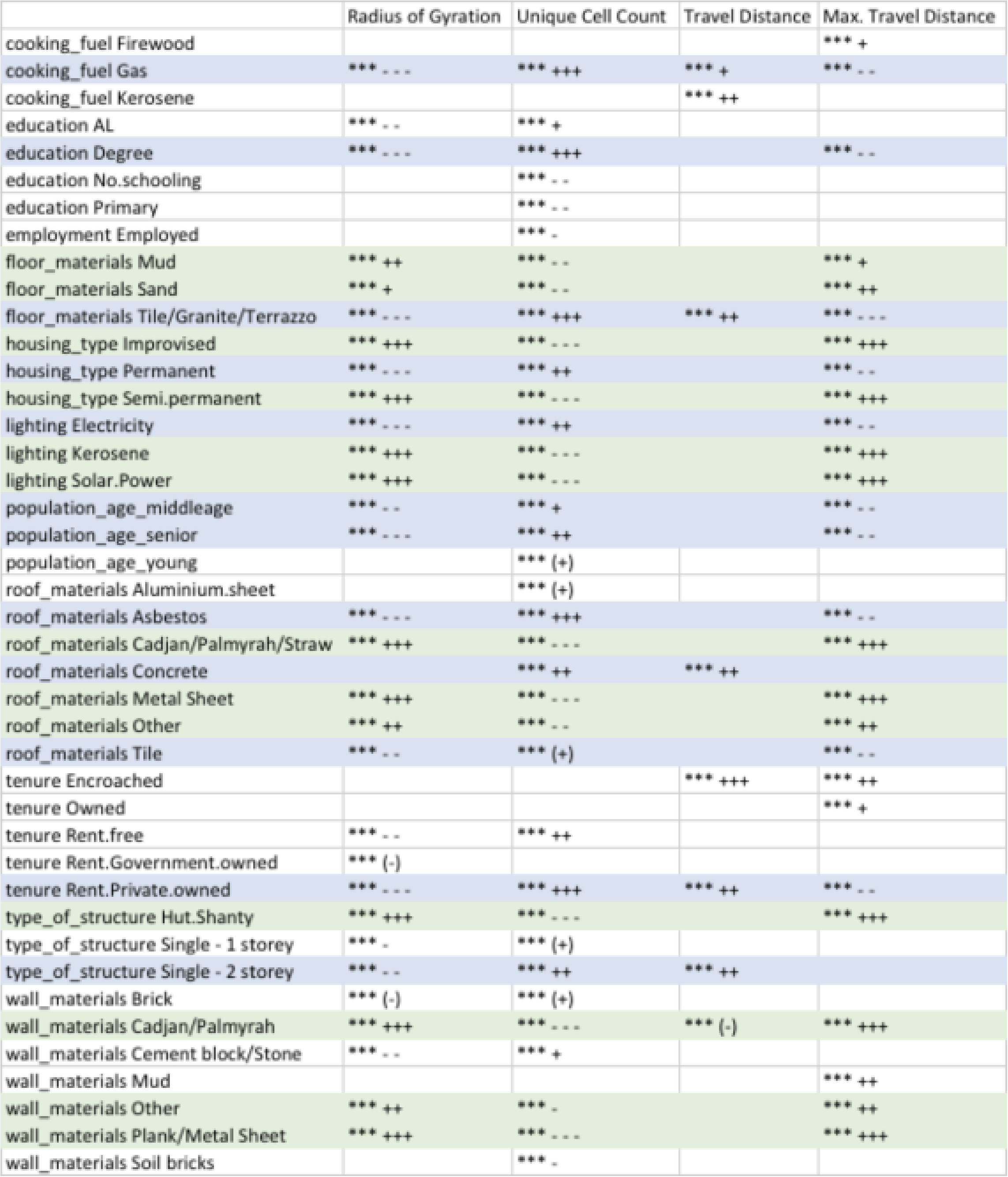
	Contact Count	Contact Rate	Physical Distance
cooking_fuel Gas	*** ++	*** -	*** - -
cooking_fuel Kerosene	*** ++		*** (-)
education AL		*** -	*** -
education Degree	*** ++	*** - -	*** - - -
education OL			*** -
employment Not.Active			*** (-)
employment Unemployed			*** (-)
floor_materials Concrete			*** - -
floor_materials Tile/Granite/Terrazzo	*** ++	*** - -	*** - - -
housing_type Improvised	*** -		*** ++
housing_type Permanent		*** -	*** - -
lighting Electricity		*** -	*** - - -
population_age_middleage			*** - -
population_age_senior	*** +	*** -	*** - -
population_age_young			*** - -
population_gender Female			*** (-)
roof_materials Aluminium Sheet			*** -
roof_materials Asbestos	*** ++	*** - -	*** - - -
roof_materials Concrete	*** ++	*** (-)	*** -
roof_materials Metal sheet	*** -	*** ++	*** +++
roof_materials Other		*** ++	*** ++
roof_materials Tile		*** -	*** - -
tenure Encroached	*** +		
tenure Rent.free	*** +	*** - -	*** - -
tenure Rent.Government.owned	*** ++		*** -
tenure Rent.Private.owned	*** +++	*** - -	*** - - -
type_of_structure 1 storey			*** - -
type_of_structure 2 storey	*** ++		*** - -
wall_materials Brick			*** -
wall_materials Cadjan/Palmyrah	*** - -		*** +++
wall_materials Cement block/Stone		*** - -	*** - -
wall_materials Plank/Metal Sheet		*** ++	*** ++
wall_materials Soil bricks	*** ++		*** - -

Figure 5 – Social CDR Features

Features are color-coded if they have two or more strong or moderate correlations that move according to the general trends.

Blue *– Contact Count (positive), Contact Rate (negative), Physical Distance (negative). These are mostly features associated with high SEL characteristics.*

Green **-** *Contact Count (negative, Contact Rate (positive), Physical Distance (positive). These are mostly features associated with lower SEL characteristics.*



	Radius of Gyration	Unique Cell Count	Travel Distance	Max. Travel Distance
cooking_fuel Firewood				*** +
cooking_fuel Gas	*** - - -	*** +++	*** +	*** - -
cooking_fuel Kerosene			*** ++	
education AL	*** - -	*** +		
education Degree	*** - - -	*** +++		*** - -
education No.schooling		*** - -		
education Primary		*** - -		
employment Employed		*** -		
floor_materials Mud	*** ++	*** - -		*** +
floor_materials Sand	*** +	*** - -		*** ++
floor_materials Tile/Granite/Terrazzo	*** - - -	*** +++	*** ++	*** - - -
housing_type Improvised	*** +++	*** - - -		*** +++
housing_type Permanent	*** - - -	*** ++		*** - -
housing_type Semi.permanent	*** +++	*** - - -		*** +++
lighting Electricity	*** - - -	*** ++		*** - -
lighting Kerosene	*** +++	*** - - -		*** +++
lighting Solar.Power	*** +++	*** - - -		*** +++
population_age_middleage	*** - -	*** +		*** - -
population_age_senior	*** - - -	*** ++		*** - -
population_age_young		*** (+)		
roof_materials Aluminium.sheet		*** (+)		
roof_materials Asbestos	*** - - -	*** +++		*** - -
roof_materials Cadjan/Palmyrah/Straw	*** +++	*** - - -		*** +++
roof_materials Concrete		*** ++	*** ++	
roof_materials Metal Sheet	*** +++	*** - - -		*** +++
roof_materials Other	*** ++	*** - -		*** ++
roof_materials Tile	*** - -	*** (+)		*** - -
tenure Encroached			*** +++	*** ++
tenure Owned				*** +
tenure Rent.free	*** - -	*** ++		
tenure Rent.Government.owned	*** (-)			
tenure Rent.Private.owned	*** - - -	*** +++	*** ++	*** - -
type_of_structure Hut.Shanty	*** +++	*** - - -		*** +++
type_of_structure Single - 1 storey	*** -	*** (+)		
type_of_structure Single - 2 storey	*** - -	*** ++	*** ++	
wall_materials Brick	*** (-)	*** (+)		
wall_materials Cadjan/Palmyrah	*** +++	*** - - -	*** (-)	*** +++
wall_materials Cement block/Stone	*** - -	*** +		
wall_materials Mud				*** ++
wall_materials Other	*** ++	*** -		*** ++
wall_materials Plank/Metal Sheet	*** +++	*** - - -		*** +++
wall_materials Soil bricks		*** -		

Chart 6 – Mobility CDR Features

Features are color-coded if they have two or more strong or moderate correlations that correspond to the general trends.

Blue *– Radius of Gyration(negative), Unique Cell Count(positive), Travel Distance (positive), Max. Travel Distance (negative). These are mostly features associated with high SEL characteristics.*

Green - *Radius of Gyration(positive), Unique Cell Count(negative), Travel Distance (negative), Max. Travel Distance (positive). These are mostly features associated with high SEL characteristics. These are mostly features associated with lower SEL characteristics.*

7. REGRESSION MODEL

To better understand whether CDR features can be used to predict census values, we built a multivariate linear regression model using the ordinary least squares method. We log-transformed the predictor and response features to better satisfy the conditions for linear regression models, and removed CDR features that exhibited high multicollinearity. For each group of collinear features, we kept the feature that produced the model with the highest adjusted R^2 values.

We included 9 predictors in the final regression. These included 8 CDR features: one consumption feature (Duration Out), three social features (Contact Rate, Contact Count, Physical Distance), and four mobility features (Travel Distance, Gyration, Unique Cell Count, Maximum Travel Distance). We included 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 fit the regression model with up to 9 predictors (max. $p = 9$). We used the best subset feature selection method to build 2^9 (when generalized, 2^p) models for each census feature, and reported the highest-performing model, as measured by lowest adjusted R^2 values. Chart 7 reports results for models with an adjusted $R^2 > 0.4$.

To better understand the predictive capability of specific CDR feature types, we built three additional sets of linear regression models: one with the single consumption feature, a second with mobility-only features, and the third with social-only features. All three sets of single feature type models have less predictive capability than the 'All' models. As can be seen in Chart 7, the mobility models have the best performance, followed by the social models. The consumption models have the poorest performance, with many adjusted R^2 values close to 0. When the consumption feature is included in the 'All' models, it adds – at most – 2 percentage points to adjusted R^2 values, and in three models (marked with an *) the consumption feature is not included in the highest-performing model.

Nonetheless, the disaggregated models do demonstrate interesting findings. For instance, the results suggest that social CDR features are better predictors of education (specifically, percentage of population with degrees) than housing characteristics, especially type of lighting and housing. It also indicates that mobility features are better good predictors of housing materials and lighting than of degrees.

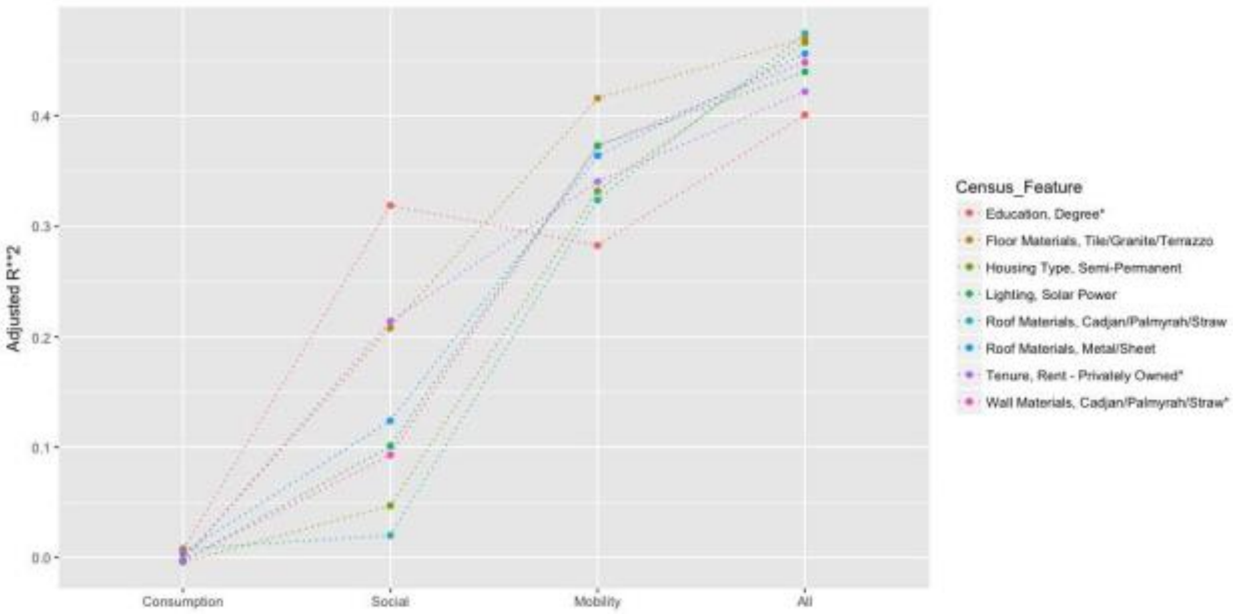


Chart 7

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, and lower values for contact rates, radius of gyration, maximum travel distance, and physical distance between contacts. The reverse trends hold for regions with lower socio-economic levels. The findings involving radius of gyration, maximum travel distance, and physical distance are especially interesting, as they are the inverse trends to those observed in the previous research.

Together, these results correspond to assumptions about the 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 movement and geographic spread of networks, may have predictive capability for census features within a post-conflict context.

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 -- the inverse of those identified by Frias-Martinez and Virseda (2012) . 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 are well below the R^2 values that Frias-Martinez and Virseda (2012) were able to achieve. 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.

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 (2012). 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 .

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, not more granular census-block level data. A better approach would have been to utilize a kernel density estimation technique when mapping census features to the voronoi cells. This will be done in subsequent extensions of this preliminary research.

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 future work, we will explore machine learning approaches (rather than just the statistical methods employed here), and develop a model that would better enable feature selection and predictive capacity.

8. CONCLUSION AND POLICY RECOMMENDATIONS

Our analysis demonstrates the potential for telecom data to predict census features, especially in post-conflict settings with a connected population. Our preliminary findings suggest that telecom data yields behavioral features that can help observe a specific vulnerable population, those of displaced and recently resettled persons.

However, to successfully use alternate data sources for policymaking (including CDR data) requires the confluence of several factors. First, data access must be negotiated from the private sector and shared with public sector officials in a manner that safeguards privacy and competitive concerns. Second, the public sector needs to develop capacity to process and analyze big data.

In the short-term, we recommend that Sri Lankan entities, especially DCS, partner with innovative external organizations who can help the government sector expedite the data negotiation and capacity-building processes. Such partnerships would allow government officials to develop proofs of concept quickly and mainstream them, help provide exposure for medium-term capacity-building, and help government officials and policy makers to become informed consumers of big data research.

REFERENCES

Bengtsson, L., Gaudart, J., Lu, X., Moore, S., Wetter, E., Sallah, K., ... & Piarroux, R. (2015). Using mobile phone data to predict the spatial spread of cholera. Scientific reports, 5.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. Science, 350(6264), 1073-1076

Decuyper, A., Rutherford, A., Wadhwa, A., Bauer, J. M., Krings, G., Gutierrez, T., ... & Luengo-Oroz, M. A. (2014). Estimating food consumption and poverty indices with mobile phone data. arXiv preprint arXiv:1412.2595

Department of Census and Statistics (2012). Census of Population and Housing. Retrieved May 24, 2017, from http://www.statistics.gov.lk/PopHouSat/CPH2011/Pages/Activities/Reports/FinalReport/FinalReportE.pdf

Department of Census and Statistics (2012). Population of Sri Lanka by District. Retrieved May 24, 2017, from http://www.statistics.gov.lk/PopHouSat/CPH2011/Pages/sm/CPH%202011_R1.pdf

Elbers, C, Lanjouw, JO & Lanjouw, P, 2003, 'Micro-Level Estimation of Poverty and Inequality', Econometrica, vol. 71, no. 1, pp. 355 – 364.

Frias-Martinez, V., & Virseda, J. (2012, March). 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 (pp. 76-84). ACM.

Frias-Martinez, V., Virseda-Jerez, J., & Frias-Martinez, E. (2012). On the relation between socio-economic status and physical mobility. Information Technology for Development, 18(2), 91-106.

Gutierrez, T., Krings, G., & Blondel, V. D. (2013). Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. arXiv preprint arXiv:1309.4496

Lokanathan, S., Kreindler, G., de Silva, N. D., Miyauchi, Y., Dhananjaya, D., & Samarajiva, R.

(2016). The Potential of Mobile Network Big Data as a Tool in Colombo's Transportation and Urban Planning. Special Issue of Information Technologies & International Development. Vol. 12, No. 2.

Lu, X., Bengtsson, L., & Holme, P. (2012). Predictability of population displacement after the 2010 Haiti earthquake. Proceedings of the National Academy of Sciences, 109(29), 11576-11581

Poverty Indicators Household Income and Expenditure Survey - 2012/13. (n.d.). Retrieved May 24, 2017, from www.statistics.gov.lk/poverty/PovertyIndicators2012_13.pdf
http://www.statistics.gov.lk/poverty/SpatialDistributionOfPoverty2012_13.pdf

Smith C, Mashadi A, Capra L (2013) Ubiquitous sensing for mapping poverty in developing countries, Proceedings of the Third Conference on the Analysis of Mobile Phone Datasets

Smith-Clarke C, Mashhadi A, Capra L (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. New York, NY, USA: ACM, CHI '14, pp. 511–520. doi:10.1145/2556288.2557358. URL
http://doi.acm.org/10.1145/2556288.2557358

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., ... & Hadiuzzaman, K. N. (2017). Mapping poverty using mobile phone and satellite data. Journal of The Royal Society Interface, 14(127), 20160690

The Spatial Distribution of Poverty in Sri Lanka. (n.d.). Retrieved May 24, 2017, from
http://www.statistics.gov.lk/poverty/SpatialDistributionOfPoverty2012_13.pdf

Tilakaratna, K. G., & Satharasinghe, A., Dr. (n.d.). Headcount Index and Population Below Poverty Line by DS Division – Sri Lanka: 2002 . Retrieved May 24, 2017, from
http://www.statistics.gov.lk/poverty/small%20area%20reportNEW.pdf

United Nations Global Pulse (October 2013) Mobile Phone Network Data for Development. Retrieved May 24, 2017, from
http://www.unglobalpulse.org/sites/default/files/Mobile%20Data%20for%20Development%20Primer_Oct2013.pdf

United Nations High Commissioners for Refugees (December 2012). UNHCR Eligibility Guidelines for Assessing the International Protection Needs of Asylum-Seekers from Sri Lanka. Retrieved May 24, 2017, from
http://www.refworld.org/pdfid/50d1a08e2.pdf

Wesolowski, A., Qureshi, T., Boni, M. F., Sundsøy, P. R., Johansson, M. A., Rasheed, S. B., ... & Buckee, C. O. (2015). Impact of human mobility on the emergence of dengue epidemics in Pakistan. Proceedings of the National Academy of Sciences, 112(38), 11887-11892

Wilson, R., zu Erbach-Schoenberg, E., Albert, M., Power, D., Tudge, S., Gonzalez, M., ... &

Pitonakova, L. (2016). Rapid and near real-time assessments of population displacement using mobile phone data following disasters: the 2015 Nepal earthquake. PLoS currents, 8.

World Pop. (n.d.). Retrieved from http://www.worldpop.org.uk/about_our_work/about_worldpop/