PREDICTING SOCIO-ECONOMIC CHARACTERISTICS OF THE SRI LANKAN POPULATION USING CALL DETAIL RECORDS (CDRS)

CPRSOUTH2017 **POLICY BRIEF**

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. However, there are several limitations of spatially-granular data on the Sri Lankan population: they are collected relatively infrequently (either through the decennial National Census or the triennial Household Income and Expenditure Survey -HIES), they have varying degrees of representation (the HIES is only representative to the 2nd administrative level), and there is often a 1-2 year lag between data collection and data publication. As a consequence, available data are not enough to capture the dynamics of a fast-moving economy like Sri Lanka, especially within its post-conflict zones.

Our research suggests that pseudonymized records of the population's use of their mobile phones can provide a cheap, reliable proxy for census data within Sri Lanka, especially in post-conflict regions that have the most need for more frequent, granular data.

FINDINGS & RECOMMENDATIONS

1. Our research indicates that CDR data are especially effective at capturing socio-economic characteristics of the displaced or recently resettled populations in Sri Lanka, indicating the data's relevance in post-conflict situations.
2. Use of CDR for socio-economic prediction is relevant in the Sri Lankan context, although patterns do vary from previous research. This demonstrates application of CDR data, but also reinforces need to do country-specific analysis.
3. CDR data can be used to develop a 'soft substitute' for census socio-economic characteristics but, for robust results, the census department will need to: (a) Collaborate with researchers to provide more granular census data; and (b) Complement CDR data with actively-collected data, such as survey data

THE RESEARCH

I LITERATURE REVIEW

Research has demonstrated the opportunity for call detail records data (CDR) to predict poverty, and provide a "soft substitute" (Frias-Martinez and Virseda, 2012) for data collected through the census. Frias-Martinez and Virseda identified statistically- significant relationships between CDR and census socio- economic features, and built a basic predictive model. 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 construct a composite wealth index, which they used to predict the wealth of the out-of-sample population. The results were validated at a district level using national survey data. To our knowledge, similar research has yet to be conducted in the immediate (< 5 years) aftermath of a conflict.

II THE DATA

We use the 2011/12 Sri Lanka census data and pseudonymized CDR data for approximately 600,000 mobile phone subscribers from Sri Lanka's Northern province, which is a post-conflict region.

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), from which we can dereference to a physical (latitude, longitude) location.

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

For high-resolution 'ground truth' population data, we used publicly-available census data at the Grama Niladhari (GN) division (4th and lowest level administrative unit).

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. In this study, we limit ourselves to existing datasets, and assume that at least some of the census features (e.g. Tile, Granite, or Terrazzo flooring, and University Degrees) correspond to high socio-economic levels (SEL), and others (e.g. Semi-permanent housing, cadjan or palmyrah walls) correspond to low SEL.

III METHODOLOGY

Statistical relationships

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.

Linear regression model

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 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. We call these the 'All' models, as they can include all 9 predictors.

To better understand the predictive capability of specific CDR feature types, we also 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.

IV RESULTS

Statistical relationships

Our set of CDR – census relationships correspond to findings in Frias-Martinez and Virseda (2012). These demonstrate that CDR data is sensitive to socio-economic characteristics in a Sri Lankan context.

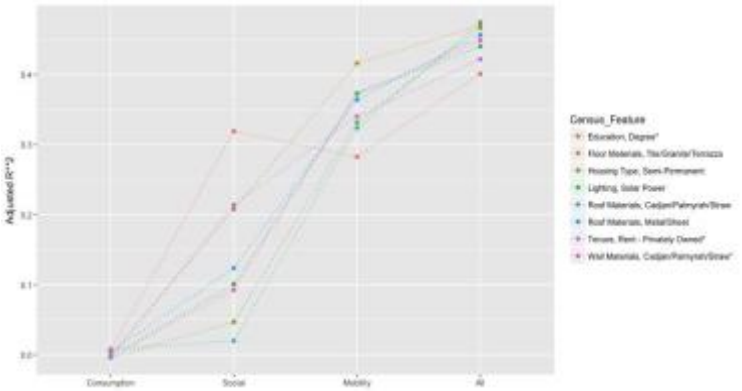
- Greater cell phone usage is positively correlated to census features associated with higher SEL, such as having tertiary education (degree) and roofs made of asbestos and concrete, and negatively correlated to census features associated with low SEL, such as living in improvised housing.
- A greater number of contacts is associated with high SEL.

A second set of interactions are inverse to those observed in previous research, especially around CDR mobility features. This discrepancy demonstrates sensitivity to the large percentage of displaced or recently resettled people in the area under study, and illustrates the opportunity for CDR-based prediction in post-conflict settings.

- Higher values for radius of gyration (distance traveled weighted by frequency), maximum travel distance, and physical distance between contacts is associated with low SEL.

Linear regression model

The linear regression did not produce robust results (the highest R^2 values were < 0.5). The chart below reports results for 'All' models with an adjusted $R^2 > 0.4$. All three sets of single feature type models have less predictive capability than the 'All' models, with mobility-only features producing the next-best results.



Our research seeks to understand the utility of CDR data in predicting Sri Lanka, for which we used publicly-available data. The statistical relationships demonstrate the opportunity area, but – for more robust predictive models – we need access to more granular census data, and to supplement our analysis with actively-collected research data (e.g. surveys).

V SOURCES

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

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.

ACKNOWLEDGEMENTS

This work was carried out with the aid of a grant from the International Development Research Centre, Canada and the Department for International Development UK

AUTHORS

Aparna Surendra | LIRNEasia | 12 Balcombe Place, Colombo 00800, Sri Lanka | aparna@lirneasia.net |

Thavisha Gomez | LIRNEasia | 12 Balcombe Place, Colombo 00800, Sri Lanka | thavisha@lirneasia.net |

Sriganesh Lokanathan | LIRNEasia | 12 Balcombe Place, Colombo 00800, Sri Lanka | sriganesh@lirneasia.net |