Predicting population-level socio-economic characteristics using mobile network data in Sri Lanka

Aparna Surendra, Thavisha Perera-Gomez, Sriganesh Lokanathan

CPR*south* 2017

Yangon, Myanmar

LIRNEasia

Pro-poor. Pro-market.

IDRC CRDI

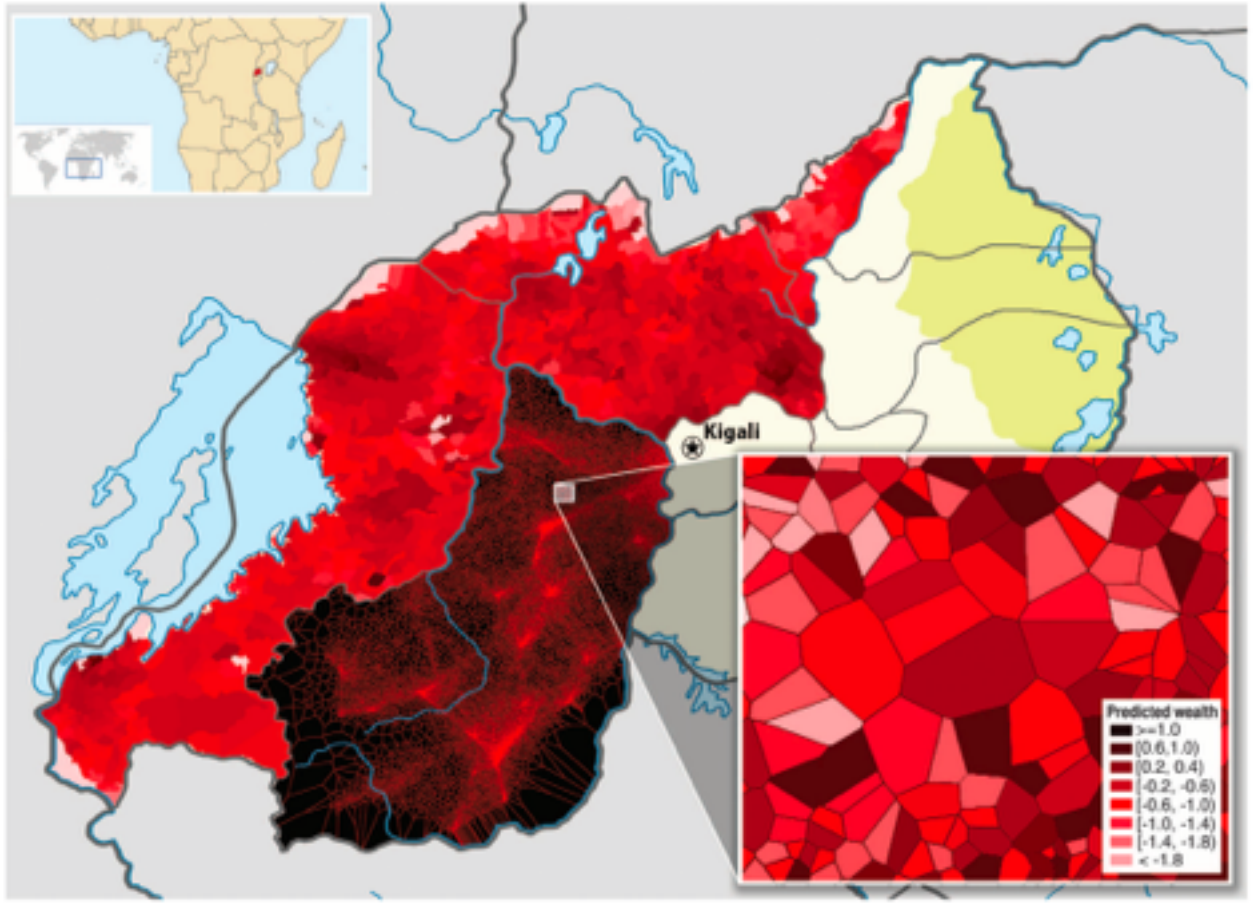
Canada

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..

UKaid

from the British people

What does success look like?



Predicting poverty and wealth from mobile phone metadata
Blumenstock et al.(2015)

Can mobile network data predict census socio-economic characteristics in Sri Lanka?

- More frequent collection
  - Sri Lankan census: Every 10 years
  - Household Income and Expenditure Survey (HIES): Every 3 years
- Data at a more granular spatial resolution
  - SL census: Publicly-available at Grama Niladhari (GN) level
  - HIES: District level
- Inexpensive

How do we get there?

Phase 1

- Investigate relationships between census and telecom data in Sri Lanka
- Build a preliminary predictive model

Phase 2

- Collect supplementary data (e.g. survey data) and use advanced machine learning methods to build a more accurate predictive model

Sophistication of tools used

How do we get there?

Phase 1

- Investigate relationships between census and telecom data in Sri Lanka
- Build a preliminary predictive model

Phase 2

- Collect supplementary data (e.g. survey data) and use advanced machine learning methods to build a more accurate predictive model

Sophistication of tools used

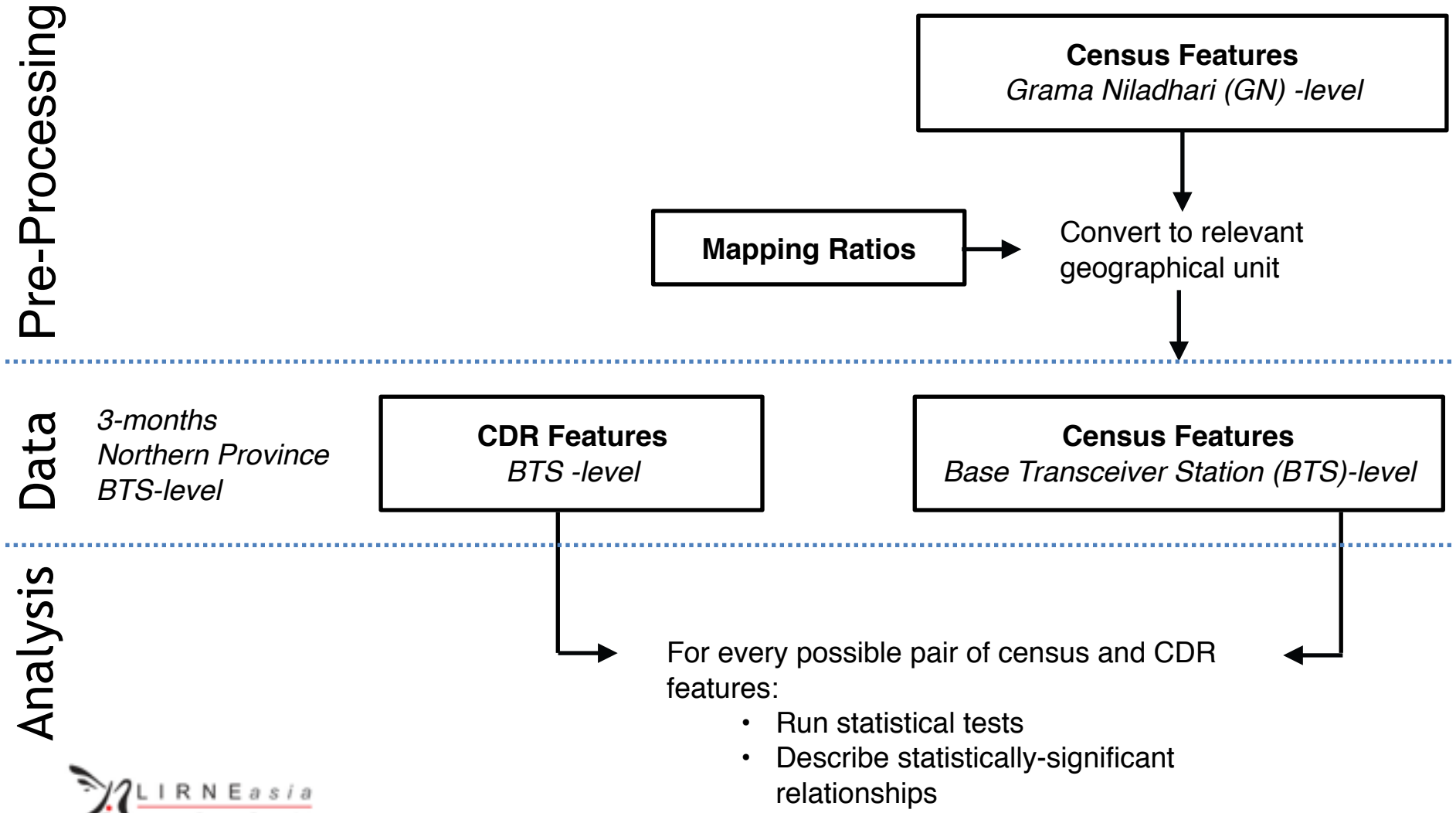
Call Detail Records Data (CDR)

- Records of all calls made and received by a person created mainly for the purposes of billing
- Similar records exist for all SMS-es sent and received as well as for all Internet sessions

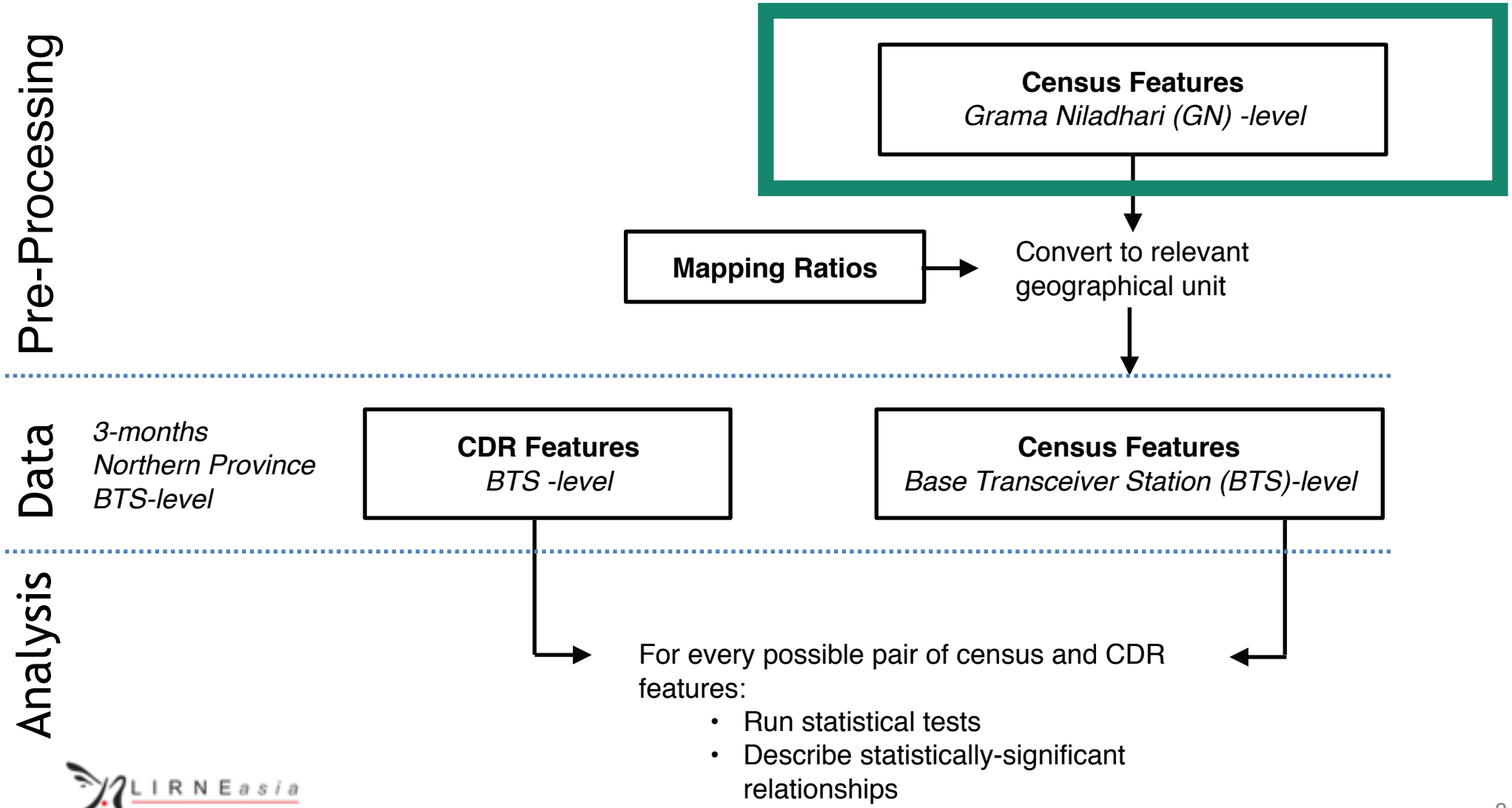
Calling Party Number	Called Party Number	Caller Cell ID	Call Time	Call Duration
A24BC1571X	B321SG141X	3134	13-04-2013 17:42:14	00:03:35

- The Cell ID in turn has a lat-long position associated with it
- All our data is pseudonymized

Methodology



Methodology



Census Features

We use 56 features from 12 categories

Individual

Education
- No Schooling
- Primary
- Secondary
- O - Level
- A - Level
- Degree

Employment
- Employed
- Not Active
- Unemployed

Population - Gender
- Male
- Female

Population - Age
- Young (< 16)
- Middle-Age (16-60)
- Senior (> 60)

Household

Floor Materials
- Cement
- Concrete
- Mud
- Other
- Sand
- Tile/Granite/ Terrazzo
- Wood

Roof Materials
- Asbestos
- Cadjan/Palmyrah Straw/Concrete
- Metal sheet
- Other
- Tile
- Zinc aluminium sheet

Wall Materials
- Brick
- Cadjan/Palmyrah
- Cement block/ Stone
- Mud
- Other
- Plank /Metal Sheet
- Soil Bricks

Housing Type
- Improvised
- Permanent
- Semi - permanent
- Unclassified

Type of Structure
- Single - 1 storey
- Single - 2 storey

Tenure
- Encroached
- Other
- Owned
- Rent-free
- Rent - Government Owned
- Rent - Privately Owned

Cooking Fuel
- Firewood
- Kerosene
- Gas
- Electricity
- Dust

Lighting
- Electricity
- Solar Power
- Kerosene
- Bio-Gas

LIRNEasia
Pro-poor. Pro-market.

Census Features

QUESTION: If we're interested in socio-economic levels, why do we use census Population and Housing data, not poverty data?

ANSWER:

1) Accuracy. Poverty data is modeled through the HIES at the District level, and estimated at the DSD-level with small area estimates. The population and housing data is collected through census.

2) Granularity. Poverty estimates are not reported at the GN-level.

We use a multidimensional poverty perspective, and assume a relationship between publicly-available GN-level census features and socio-economic levels.

Census Features

QUESTION: Why do we focus on the Northern Province?

ANSWER:

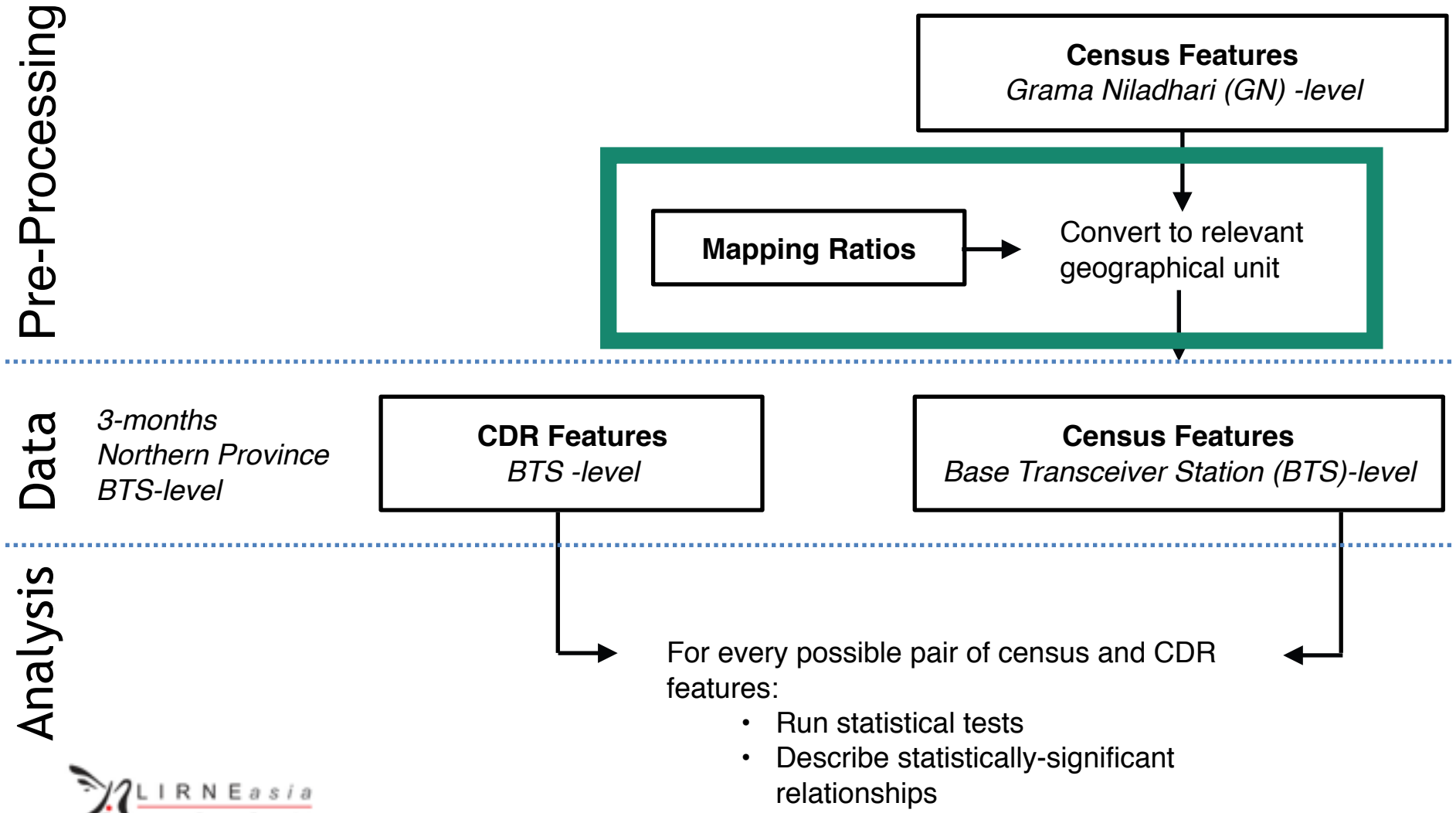
1)Need for data**.**

a. Greatest paucity of historical census data in the Northern Province (site of much of the conflict).

b.Population is changing rapidly in the aftermath of the conflict – this area has the greatest need for census proxies.

2)Most likely to reflect relevance of these methods in post-conflict settings.

Methodology



Mapping Ratios

a. Calculate census features as a percentage of the GN population

GN Name	Total	Brick	Cement block/ Stone	Cabook	Soil bricks	Mud	Cadjan/ Palmyrah	Plank/Metal Sheet	Other
Sammantranapura	**1687**	455	863	2	-	-	-	355	12
Mattakkuliya	**6143**	2534	3369	22	2	2	2	212	-

Raw Census Data - Housing Wall Materials

b. Map each BTS as fractions of the GNs that cover the BTS.

$BTS_i = s*GN_a+v*GN_b+\ldots+w*GN_d$

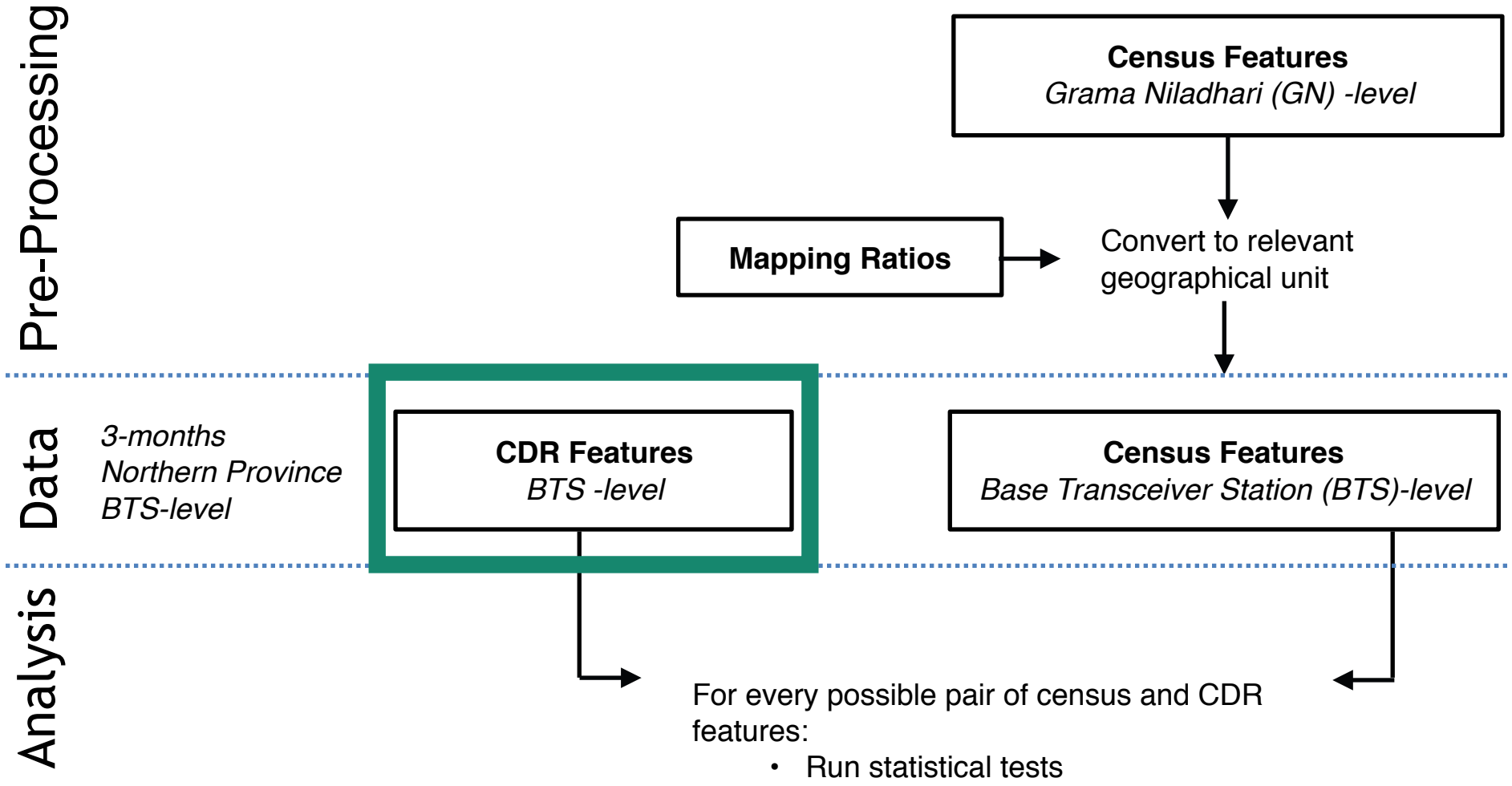
Where s, v, ...w represent the fractions of the GNs $GN_{a,\ldots,}GN_d$ that cover BTS_i.

c. Calculate BTS-level census value.

Example: $BTS_4 =(0.6*GN_3)+ (0.21*GN_1) + (0.19*GN_5)$

If the % of degree-holding population in GN_3 is 30%, GN_1 is 40%, and GN_5 is 30%: $BTS_4 =(0.6*0.3)+(0.21*0.4) + (0.19*0.3) = 0.32$

Methodology



CDR Features

12 CDR features from a 3-month period(Northern Province) from approximately pseudonymized data of 600,000 unique mobile subscribers.

Consumption

- Total - Total number of calls made or received per unique user
- Output - Total number of output calls per unique user
- Duration - Average duration of the calls per unique user
- Duration In - Average duration of the calls in per unique user
- Duration Out - Average duration of the calls out per unique user

Social

- Contact Count - Number of unique contacts
- 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 Count – Number of different BTSs visited by a person
- Travel Distance – Distance between each pair of consecutively visited BTS.
- Radius of Gyration – Distance between the home cell and each visited BTS, weighted by frequency of visit.
- Diameter – Maximum distance between the BTS' typically visited by a person.

CDR Features

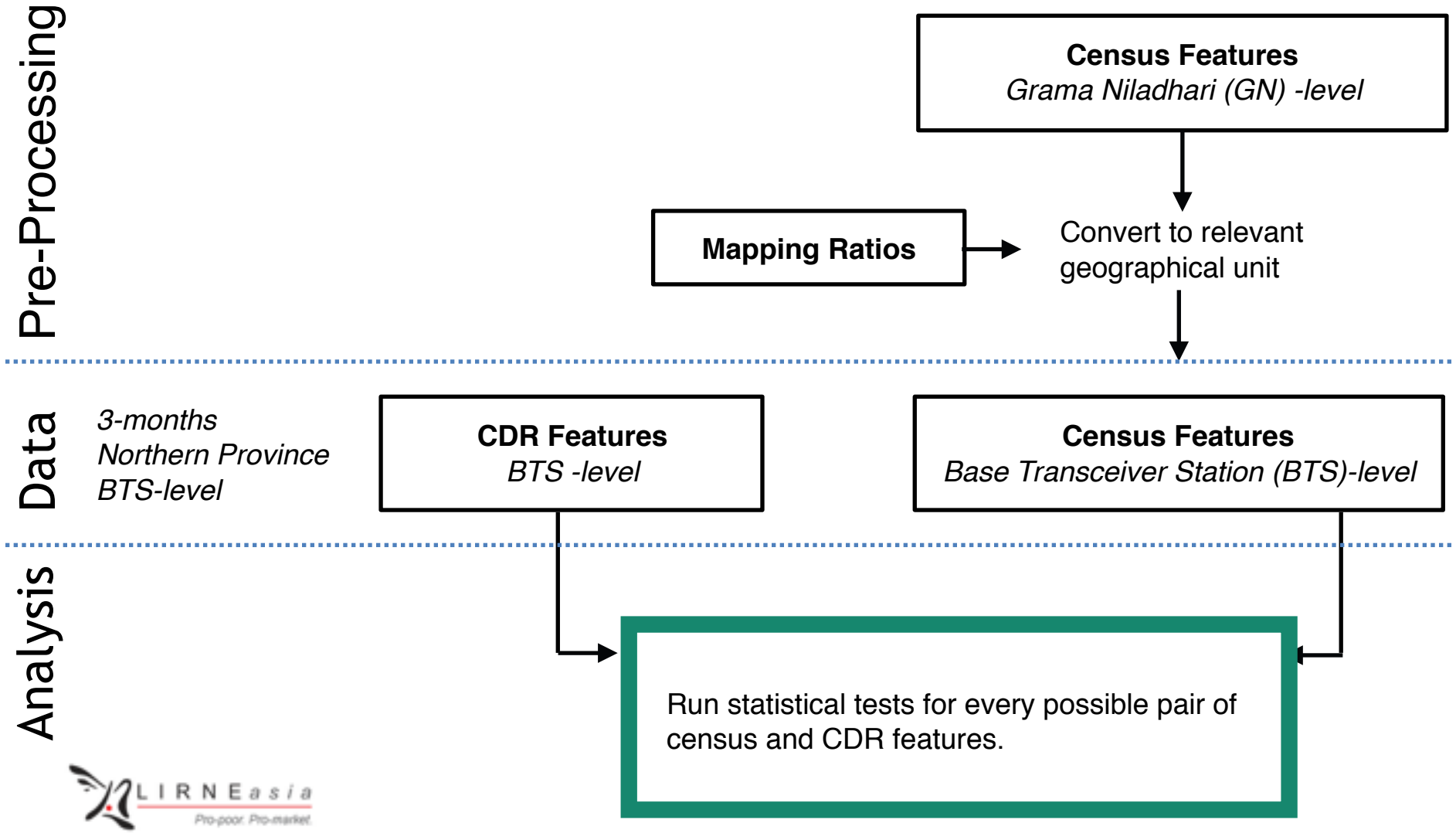
QUESTION: Why do we map census data to the BTS level, not CDR data to the GN level?

ANSWER:
There are more GNs in the Northern Province than BTS towers (900+ to ~ 250).

When we map the BTS-level CDR data to the GN-level, all the CDR values are similarly small and lose their interesting statistical properties.

Image of BTS regions (grey) overlaid with GNs (red) in Mannar

Methodology



Statistical Tests

1. ANOVA Test

The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of three or more independent groups.

$$\text{F-Test Statistic} = \frac{\text{Variance between treatments}}{\text{Variance within treatments}}$$

2. Spearman's Rank Correlation

Spearman's correlation determines the strength and direction of the monotonic relationship between two variables.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Procedure 1 Process to compute statistical tests for all cell phone usage and census variables.

for each cell phone usage variable *pvar* **do**
  for each census var *cvar* **do**
    $q_1 = (min, min + \frac{max-min}{4})$
    $q_2 = (min + \frac{max-min}{4}, min + \frac{max-min}{2})$
    $q_3 = (min + \frac{max-min}{2}, max - \frac{max-min}{4})$
    $q_4 = (max - \frac{max-min}{4}, max)$
    for each q_i **do**
      $list[q_i]$ = select {BTS} with *cvar* value in q_i
      for each *BTS* in list$[q_i]$ **do**
        $D[q_i] = D[q_i] \cup pvar$
      end for
    end for
    ANOVA$(D[q_1], D[q_2], D[q_3], D[q_4])$
    $set(pvar) = pvar\ all\ BTS$
    $set(cvar) = cvar\ all\ BTS$
    correlation$(\{set(pvar)\}, \{set(cvar)\})$
  end for
end for

Frias-Martinez and Virseda (2012)

Multivariate Linear Regression Model

Data Preparation

- Log-transformed predictors and response features
- Removed features with high multi-collinearity.

Final Input Features

- 8 CDR predictors:
  - 1 consumption feature (Duration Out)
  - 3 social features (Contact Rate, Contact Count, Physical Distance)
  - 4 mobility features (Travel Distance, Gyration, Unique Cell Count, Maximum Travel Distance).
- 1 control (Mobile User Population Density)

Feature Selection

- [Best subset method] Build 2^9 models to predict each census feature. Report the model with the lowest adjusted R^2 value. (Note: some of these models use < 9 input features).
- Built three disaggregated linear regression models, each with one type of
- CDR feature ('Social', 'Consumption', 'Mobility').

Findings

Phase 1: Evaluate Opportunity

- Investigate relationships between census and telecom data in Sri Lanka
- Build a preliminary predictive model

Users in BTS regions with **higher SEL:**

- Have more contacts
- Travel to more unique BTS regions

Users in BTS regions with **lower SEL:**

- Speak more often to their contacts
- Have more geographic spread among their contacts
- Travel further and/or travel long distances more frequently

Findings

The **inverse** to previous findings (Frias-Martinez and Virseda, 2012), but **correspond to Sri Lankan context** : ~352,000 people in the Northern Province are displaced or resettled (Census 2011/12).

Demonstrates:

- Importance of replication
- Relevance of CDR data (especially social and mobility features) in predicting census socio-economic characteristics in Sri Lankan post-conflict settings

Users in BTS regions with **higher SEL:**

- Have more contacts
- Travel to more unique BTS regions

Users in BTS regions with **lower SEL:**

- Speak more often to their contacts
- Have more geographic spread among their contacts
- Travel further and/or travel long distances more frequently

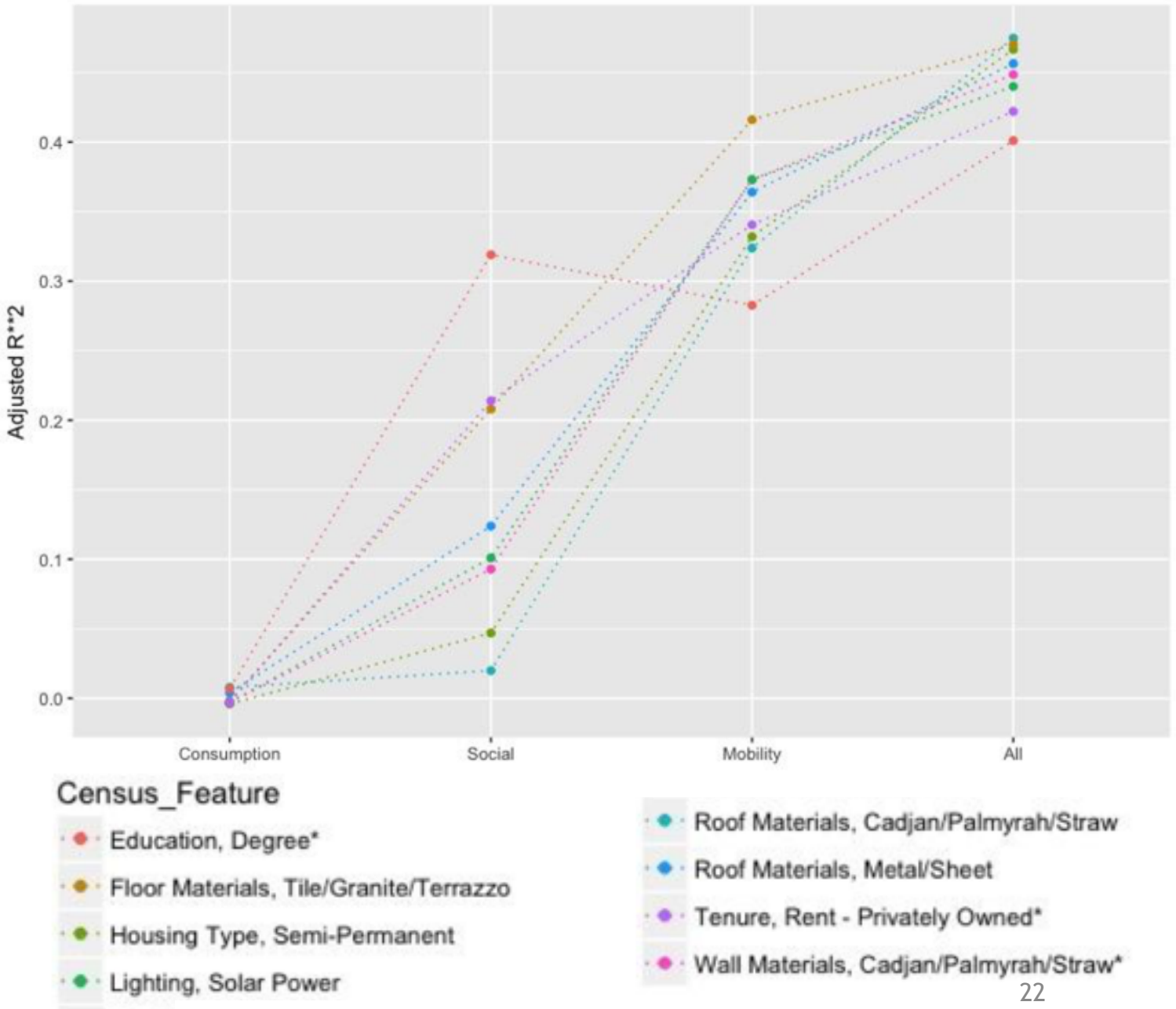
Findings

Phase 1: Evaluate Opportunity

- Investigate relationships between census and telecom data in Sri Lanka
- Build a preliminary predictive model

- Best performance: 'All' models (but still not robust result, max. adj. R^2 ~0.45)
- Poorest performance: 'Consumption' models (many adj. R^2 values close to 0.)
- Social CDR features may be better predictors of education (e.g. degree-holders) than housing characteristics.
- Mobility CDR features may be better predictors of housing materials and lighting than of education.

Only includes census features where 'All' model has an adj. R^2 > 0.4



* Indicates that consumption feature was not included in highest-performing' All' model

Can mobile network data predict census socio-economic characteristics in Sri Lanka?

Phase 1

- Investigate relationships between census and telecom data in Sri Lanka
- Build a preliminary predictive model

Phase 2

- Collect supplementary data (e.g. survey data) and use advanced machine learning methods to build a more accurate predictive model

Recommendations

<u>For policy partners (Department of Census and Statistics)</u>

- **Provide additional data:** that is not available at GN-level, incl.: numbers of resettled or displaced persons, computer literacy.
- **Provide more granular data:** Census-block level data may reduce errors in mapping method

<u>For research team</u>

- **Improve mapping method:** E.g. account for population density
- **Use different predictive models:** E.g. classification models

References

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

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

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.