LIRNEasia

Pro-poor. Pro-market.

Data driven policy

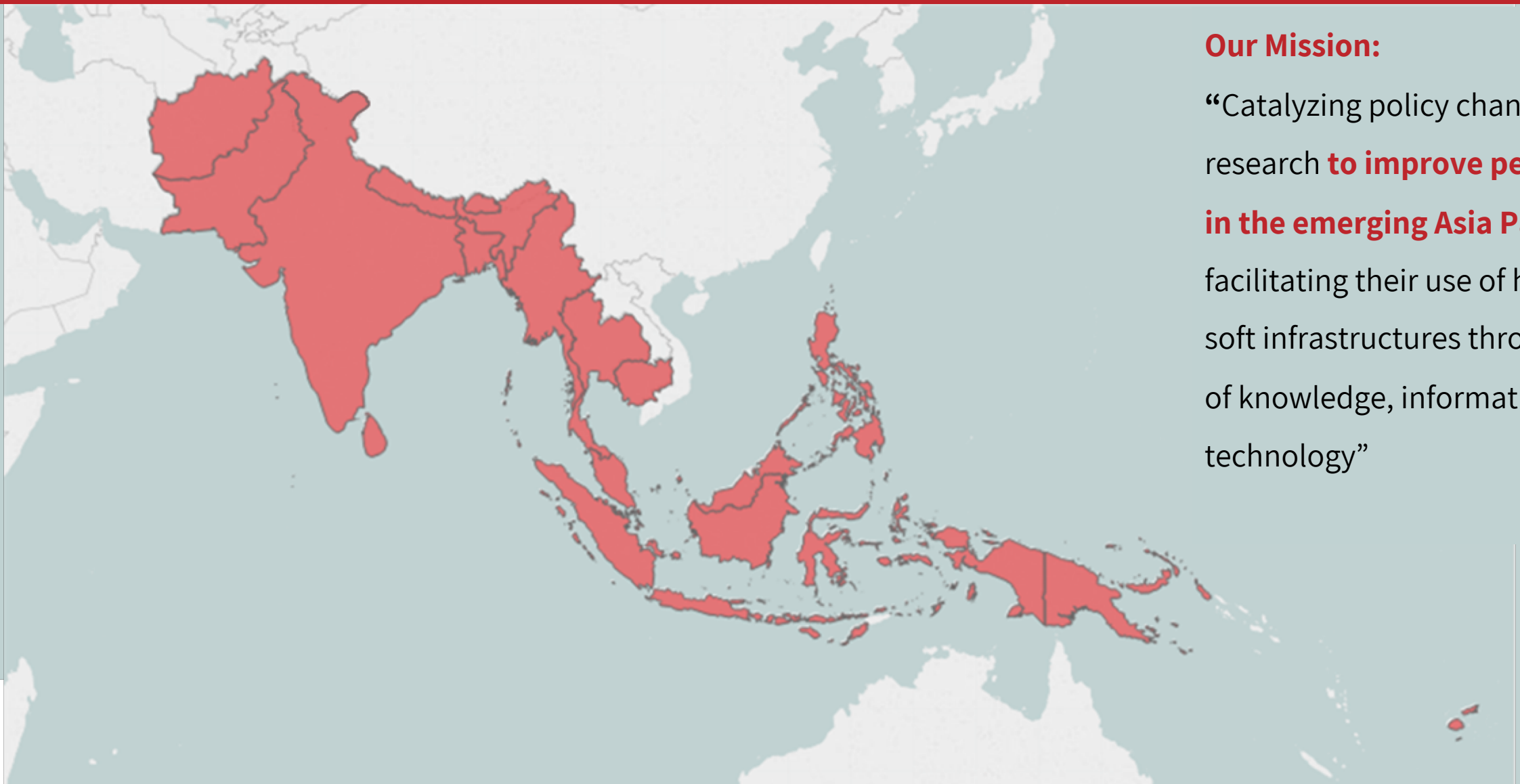
University of Moratuwa
1 February 2019

IDRC | CRDI

International Development Research Centre
Centre de recherches pour le développement international

Canada

LIRNEasia: a pro-poor, pro-market Asia Pacific think tank; since 2004



Our Mission:

"Catalyzing policy change through research **to improve people's lives in the emerging Asia Pacific** by facilitating their use of hard and soft infrastructures through the use of knowledge, information and technology"

Thing we do to achieve impact

- Research, building the evidence base
  - Often multi-country (regional comparisons), using mix of methods
  - Solutions that must work in our/local context

- Take this research --> policy
  - Communication to identified audiences, targeted messages, specified modes

- Capacity building
  - Of policy makers (improves policy horizons; gives them policy options; makes them more aware of our research, makes it easier for us reach out to them (later)
  - Of young/mid-career policy intellectuals (to enable them to do good research; to give them the tools to take their research to policy makers)
  - Of civil society and media (to enable them to identify good research from bad; to enable them to use research in their work and influence policy)

A FLAVOR OF WHAT RESEARCH WE DO (OR HAVE DONE)

“Old fashioned telecom sector stuff”: evidence based inputs leading to ICT sector reforms

- Reforms that increase access, provide differentiated price-quality bundles through market mechanisms (choice) etc.
- What kind of research have we done? What methods?
  - To quantify levels of access, gaps (e.g. urban vs rural; men vs women), understand the barriers, quantify the level of eCommerce use, etc.
    - Nationally representative sample surveys across 12 countries, since 2005
  - To understand the user experience, to understand why they face barriers; focus on marginalized
    - Focus :women, lower income persons, those with various disabilities. Qualitative, ethnographic methods
    - Broadband quality of service testing; Price and affordability benchmarking
  - To understand the policy, regulatory barriers that keeps access low
    - Sector performance reviews: expert interviews, surveys of key stakeholders

..contd.

- Some examples of “wins”
  - Changes to India Universal Service Policy, Myanmar Universal Policy, Bangladesh telecom license renewal, Reduction in Indonesia in-country back-haul pricing, overhaul of regressive SIM tax in Sri Lanka, push DoT IN to focus on bridging the gender divide in access.
- Partially won (ongoing) battles
  - “India has 500 million people online” narrative
  - But 19% of 15-65 aged Indians have used the Internet, in any form

Using digital devices & digital data to improve non-ICT sectors

- Some past examples
  - E.g. Agriculture: does more/better market price/other information (through mobiles) help agriculture markets work better, give farmers better livelihoods, include inclusion of small holders in global value chains?
  - E.g. Disaster Risk Reduction: models for ICT-based early warning systems for natural disasters in Maldives, Sri Lanka
- Today a big focus on:
  - How to improve other infrastructure (e.g. roads/transport, electricity, health) using digital traces we leave as mobile users (the IoT of developing countries)
- Along the way, attempting to answer questions such as:
  - How to do use our digital trace for development purposes in a non-privacy violating manner?
  - How to ensure decisions made using such data does not increase marginalization of certain groups
  - What regulatory and policy tools do we need to deal with issues of competition (to counter monopoly tendencies created due to networks effects; high switching costs, etc.)

Some development problems of interest to LIRNEasia...

- More people live in cities than in rural areas since 2008
  - How can we make cities more livable?
  - Is there a role for ICTs, not just more roads, transit, etc.?
- Infectious diseases are posing threats
  - Can we make better allocations of scarce resources?
- Governments are flying blind without timely data to better target expenditures, assess programs, and achieve the Sustainable Development Goals (SDGs)
  - Are there ways to remedy this?

What we do through our big data for development research practice

- Conduct analytics for development purposes using a combination of:
  - Big data (pseudonymized mobile network big data, electricity consumption data, CCTV, satellite imagery)
  - Administrative data + official statistics + other government and sectoral data
- Study the impacts of big data and AI on society
  - Privacy, marginalization/ bias, and competition
- Facilitate evidence to policy using new data sources
  - Catalyze the eco-system and mainstream the use of big data in various policy domains

https://lirneasia.net/big-data

LIRNEasia
Pro-poor. Pro-market.

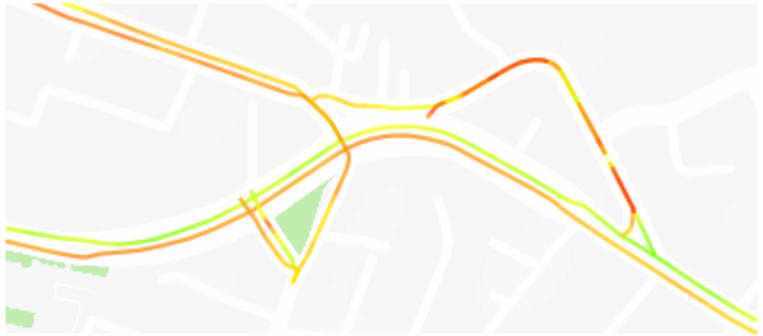
SOME EXAMPLES OF HOW WE USE BIG DATA FOR DEVELOPMENT

We consider representivity from start to finish, and throughout

- We know that reality is never perfectly represented through data; trick is to be aware of bias & strive for practical adequacy (from Critical Realism)
- Paucity of datafied data sets that included the poor and had enough variability caused us to choose mobile network big data (MNBD), rather than
  - Supermarket data (small slice, excluding poor)
  - Samurdhi/welfare data (mostly poor; but excluded some & lacked variability)
  - Smartphone data (major representivity issues)
- We always consider representivity of the data we use, in relation to phenomenon we're analyzing (e.g., speed of traffic)

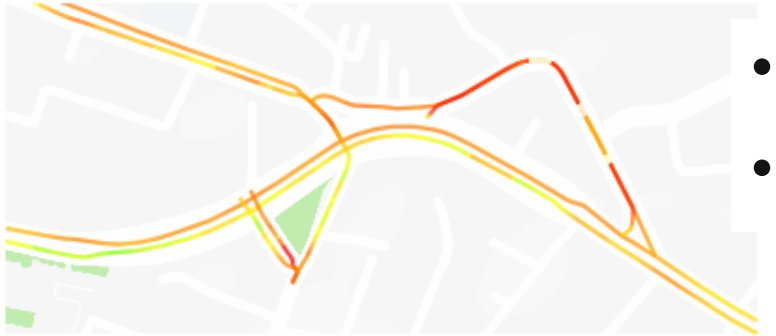
Traffic patterns at different times on the new Rajagiriya Flyover

Morning
8am - 9am



Afternoon
1pm - 2pm

Evening
6pm - 7pm



- 0 km/h
- 20 km/h
- 40 km/h

- Using only data from
- one ride-hailing firm

LIRNEasia
Pro-poor. Pro-market.

Using pseudonymized Mobile Network Big Data (MNBD) we can understand population density changes in Colombo region: weekday/ weekend
Pictures depict the change in population density at a particular time relative to midnight

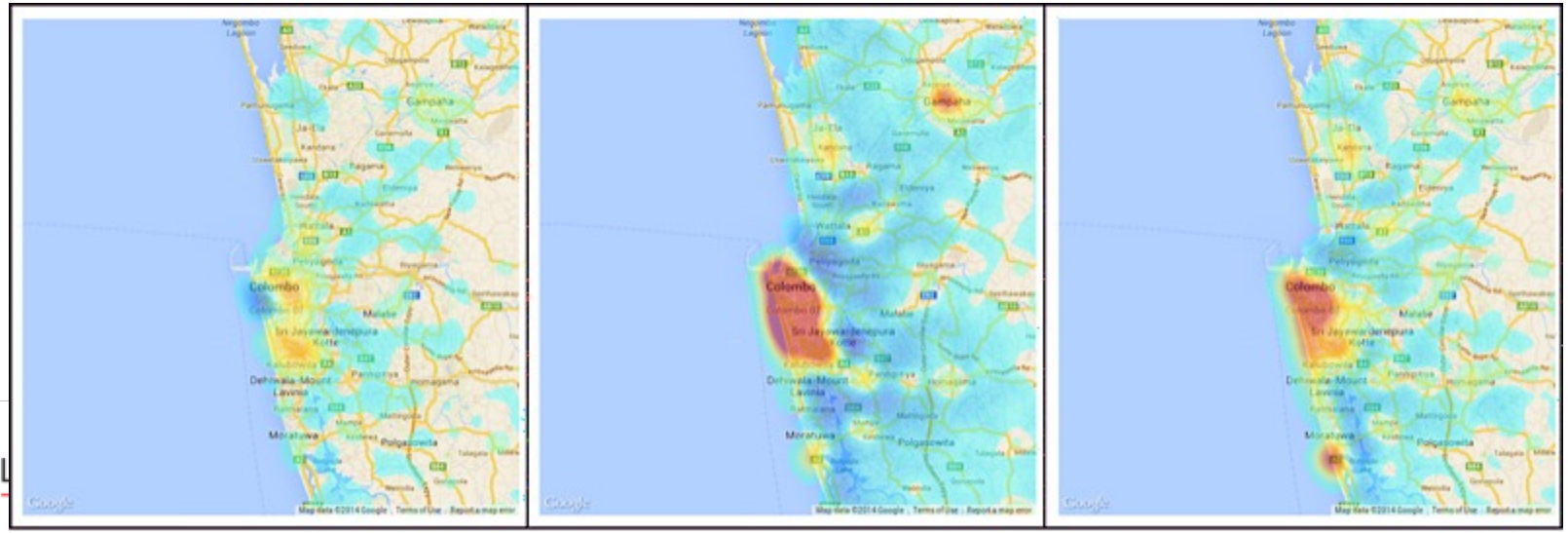
Weekday

Time 06:30 Time 12:30 Time 18:30

Decrease in Density

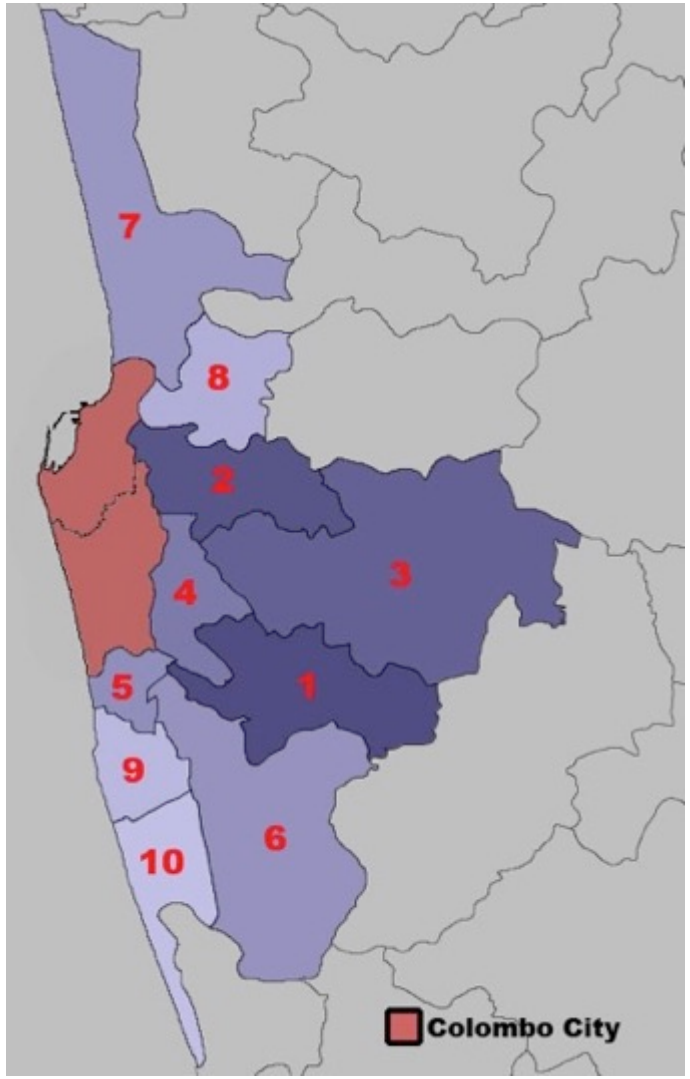
Increase in Density

Sunday



Census data from NSO + pseudonymized mobile phone big data ➔ mobility patterns in metro Colombo: nearly 47% Colombo's daytime population comes from outside

Colombo city is made up of Colombo and Thimbirigasyaya DSDs



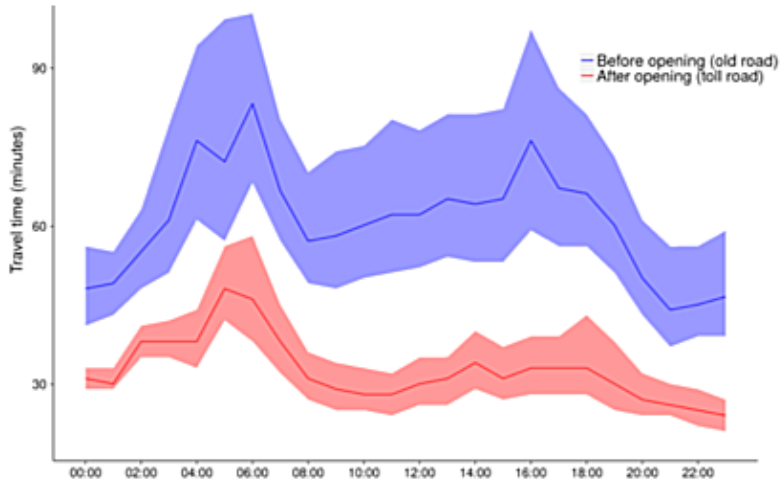
Home DSD	% age of Colombo's daytime population
Colombo city	53.1
1. Maharagama	3.7
2. Kolonnawa	3.5
3. Kaduwela	3.3
4. Sri Jayawardanapura Kotte	2.9
5. Dehiwala	2.6
6. Kesbewa	2.5
7. Wattala	2.5
8. Kelaniya	2.1
9. Ratmalana	2.0
10. Moratuwa	1.8

More info: Samarajiva, R., Lokanathan, S., Madhawa, K., Kriendler, G., & Maldeniya, D. (2015). Big data to improve urban planning. Economic and Political Weekly, Vol L. No. 22, May 30, 2015. Available at https://lirneasia.net/2015/05/journal-article-big-data-to-improve-urban-planning/

LIRNEasia
Pro-poor. Pro-market.

We can use MNBD to understand impact of new transportation infrastructure

New E03 highway connected the cities of Colombo and Negombo



Notes: Shaded areas denote 5^{th} and 95^{th} percent (bootstapped) confidence intervals

Impact on trip duration, by time of day

Findings using a Bayesian algorithm:

- People **changed their primary routes of travel, reducing overall congestion**,
- Overall **travel speeds increased**, and **reduced the amount of time spent in transit**.
- There was a **modest, but statistically significant, increase in the total amount of travel** in and around Colombo

- Next steps: understand the welfare impacts

– Blumenstock, JE, Maldeniya, D, and Lokanathan, S (2017). Understanding the Impact of Urban Infrastructure: New Insights from Population-Scale Data, *Proceedings of the 9th IEEE/ACM International Conference on Information and Communication Technologies and Development (ICTD 2017)*. Available at http://www.jblumenstock.com/files/papers/jblumenstock_2017_ictd_tollroad.pdf

LIRNEasia
Pro-poor. Pro-market.

But these analyses are of the mass

- They do not give us an understanding of mobility:
  - Disaggregated by gender, age, socio-economics
  - Disaggregated by type of mobility: car, (motor)bike, bus, train, walking
  - For those who do not use phones
- There is bias, but does it matter for the question we are addressing?
- As the use of data driven machine learning algorithms intensifies, should we worry?

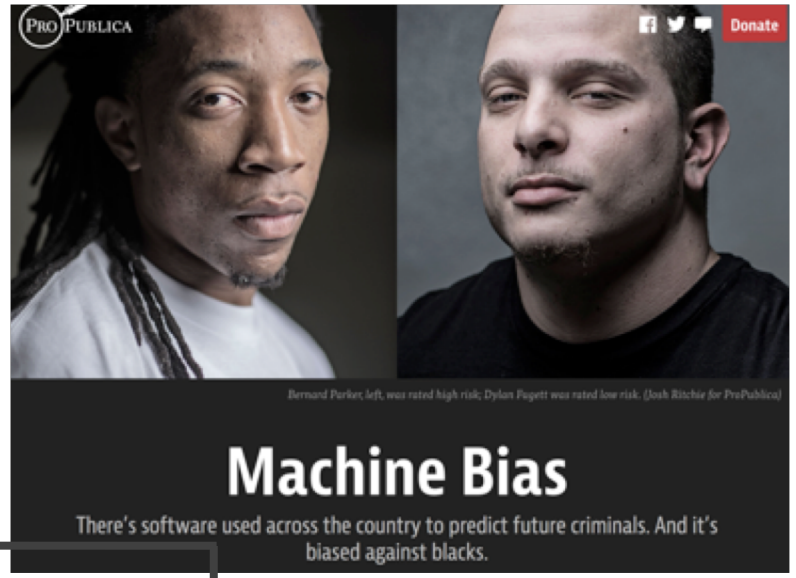
BUT FIRST LET US UNDERSTAND HOW THIS IS PLAYING OUT IN THE GLOBAL NORTH

Intelligent Machines

Microsoft is creating an oracle for catching biased AI algorithms

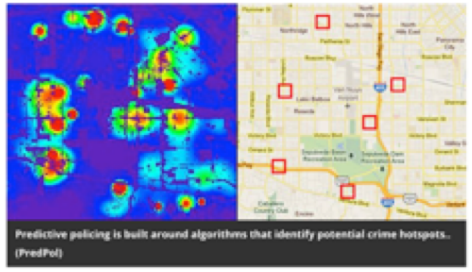
As more people use artificial intelligence, they will need tools that detect unfairness in the underlying algorithms.

by Will Knight May 25, 2018



Used to Predict Crime. But Is It Biased?

The software is supposed to make policing more fair and accountable. But critics say it still has a way to go.



By Randy Rieland
SMITHSONIAN.COM
MARCH 5, 2018

The Switch

Wanted: The 'perfect babysitter.' Must pass AI scan for respect and attitude.

Jessie Battaglia holds her son, Bennett, in their home in Rancho Mirage, Calif. While screening for a new babysitter, Battaglia started using Predictim, an online service that claims to use "advanced artificial intelligence" to assess a sitter's risk of drug abuse, bullying or having a "bad attitude." (Kyle Grillot for The Washington Post)

By Drew Harwell
November 23, 2018

The reason why most of the images that show up when you search for "doctor" are white men

Algorithms are being introduced mainly with good intentions i.e. minimize subjectivity in decision making

- Initial algorithms were **deterministic**
- Some problems were too **complex** to be solved deterministically. A better approach was to construct algorithms that can learn from data (a.k.a. **machine learning**).
- With machine learning we no longer define the parameters of the algorithm, instead they are **derived from the data** fed into it and optimised by observing a set of **evaluation metrics**

LIRNEasia
Pro-poor. Pro-market.

However biases can, and do manifest

- Bias from the data
- Bias from the optimization used

Addressing algorithmic bias involves the singular goal of making the algorithm fair

- However, what is fair?
- The literature offers several definitions
  - The perspective affects which definition applies

Three fundamental notions of fairness (Kleinberg, et al., 2016)

- **Calibration within groups** - for each group and each bin the expected number of members with a positive outcome should be proportional to the score assigned to that bin.
- **Balance for the positive class** - the average score of members with a positive outcome should be the same for each group.
- **Balance for the negative class** - the average score of members with a negative outcome should be the same for each group.

LIRNEasia
Pro-poor. Pro-market.

Only a few special cases where all three notions can be satisfied

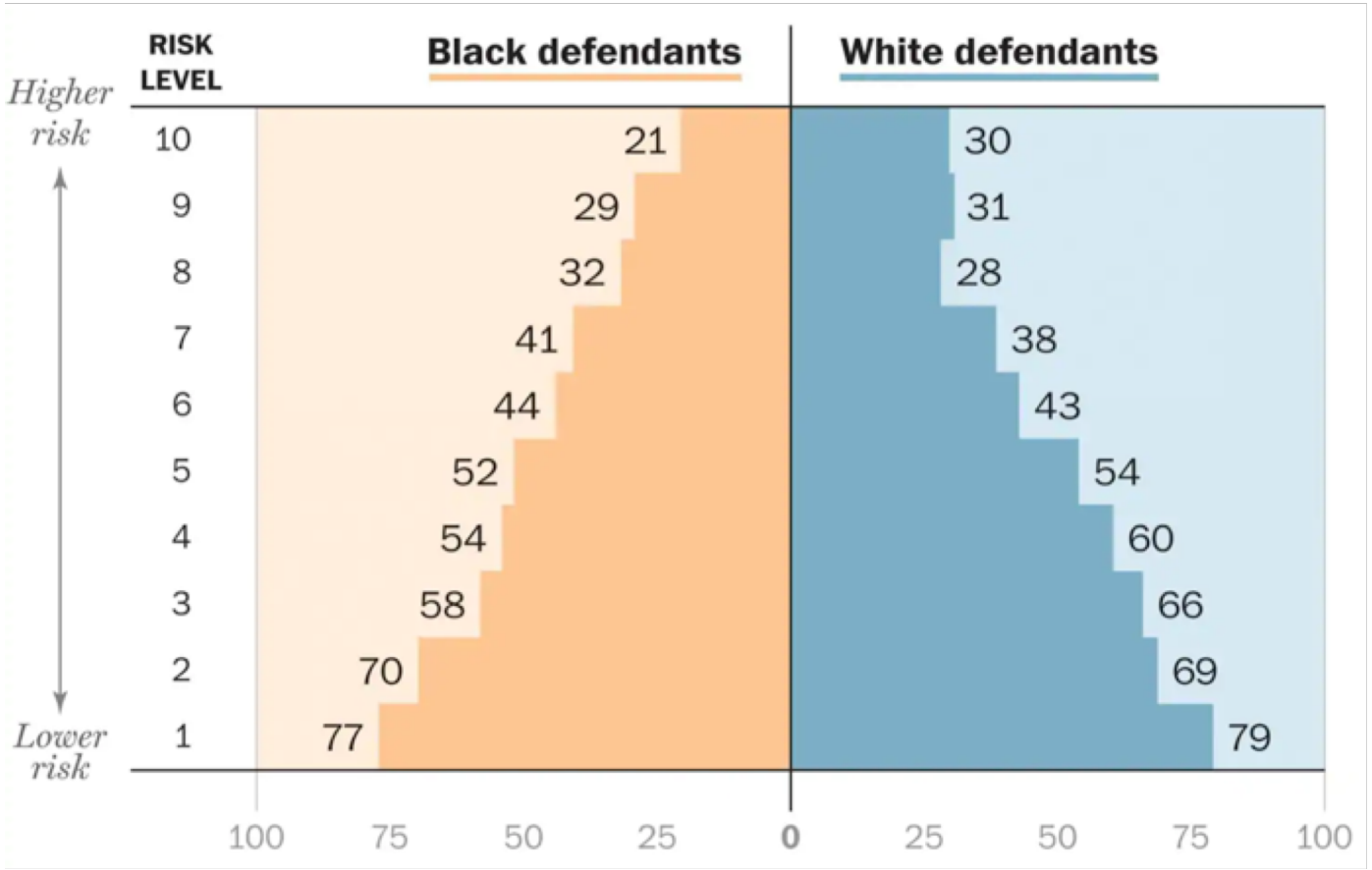
- Perfect prediction - for each feature vector, we know for certain what the outcome is.
- Equal base rates - the two groups have the same fraction of members that have a positive outcome.

Even satisfying all three notions approximately would require an approximate version of these special cases.

LIRNEasia
Pro-poor. Pro-market.

E.g. Looking a bit closely at the discourse around COMPAS (cont.)

Northpointe's COMPAS Recidivism Algorithm

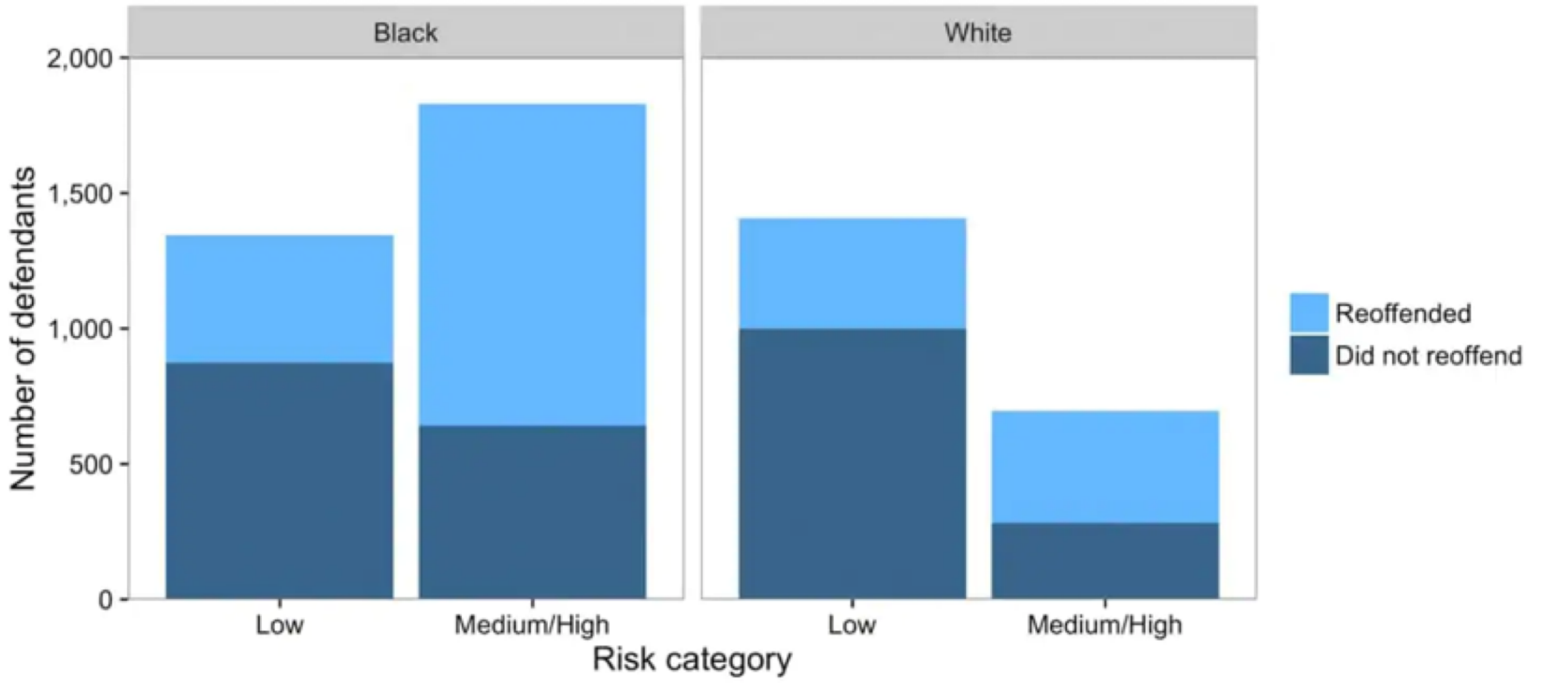


The percentage of black and white defendants in each decile bin that did not reoffend.

LIRNEasia
Pro-poor. Pro-market.

E.g. Looking a bit closely at the discourse around COMPAS

Northpointe's COMPAS Recidivism Algorithm



The number of black and white defendants in each risk category that did and did not reoffend

LIRNEasia
Pro-poor. Pro-market.

How do we tackle it?

- Simply removing any sensitive information from the training dataset **does not** help because the algorithm can **probabilistically infer** the **sensitive feature** using related information
- Algorithmic bias can be addressed by either **correcting the algorithm** or **correcting the training dataset**

Correcting the Algorithm

- Involves adjusting the **optimisation criteria** to include certain **fairness criteria**
- Commonly used optimisation criteria include:
  - **Maximum profit** - Uses a **single threshold** across all groups
  - **Demographic parity** - Uses a **different threshold** for each group such that the **fraction** of group members that are selected is the **same** across all groups

But is social engineering our objective?

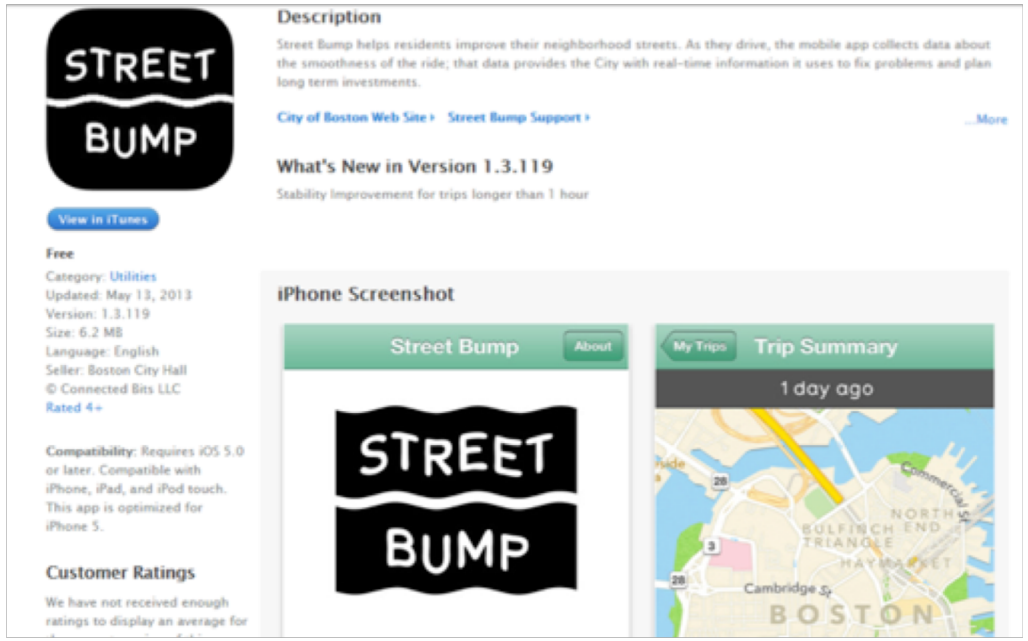
WHAT RESEARCH DO WE DO TO UNDERSTAND SUCH ISSUES IN THE GLOBAL SOUTH?

What research do we do to understand such issues in the Global South?

- Current global south context:
  - Low overall levels of ‘datafication’
  - Using algorithms for predictive policing, credit scoring, etc. are barely emergent
- Currently, limited scope for analyses of such issues because use cases are rare
- But value in understanding the issues so as to be able to engage when policy windows open

LIRNEasia
Pro-poor. Pro-market.

At the very least, appreciating issues of marginalization -> better public policy solutions



City of Boston has an app called Street Bump for smartphones. Any citizen can activate the app at the beginning of a journey. The accelerometer of the smartphone collects data proven effective in identifying pot holes and speed bumps. At the end, the collected data including the GPS coordinates of starting and ending points are sent to City Hall. Algorithms differentiate between bumps that should be there and those that should not be. Roads with an excess of the latter get routed into the work order system for repairs.

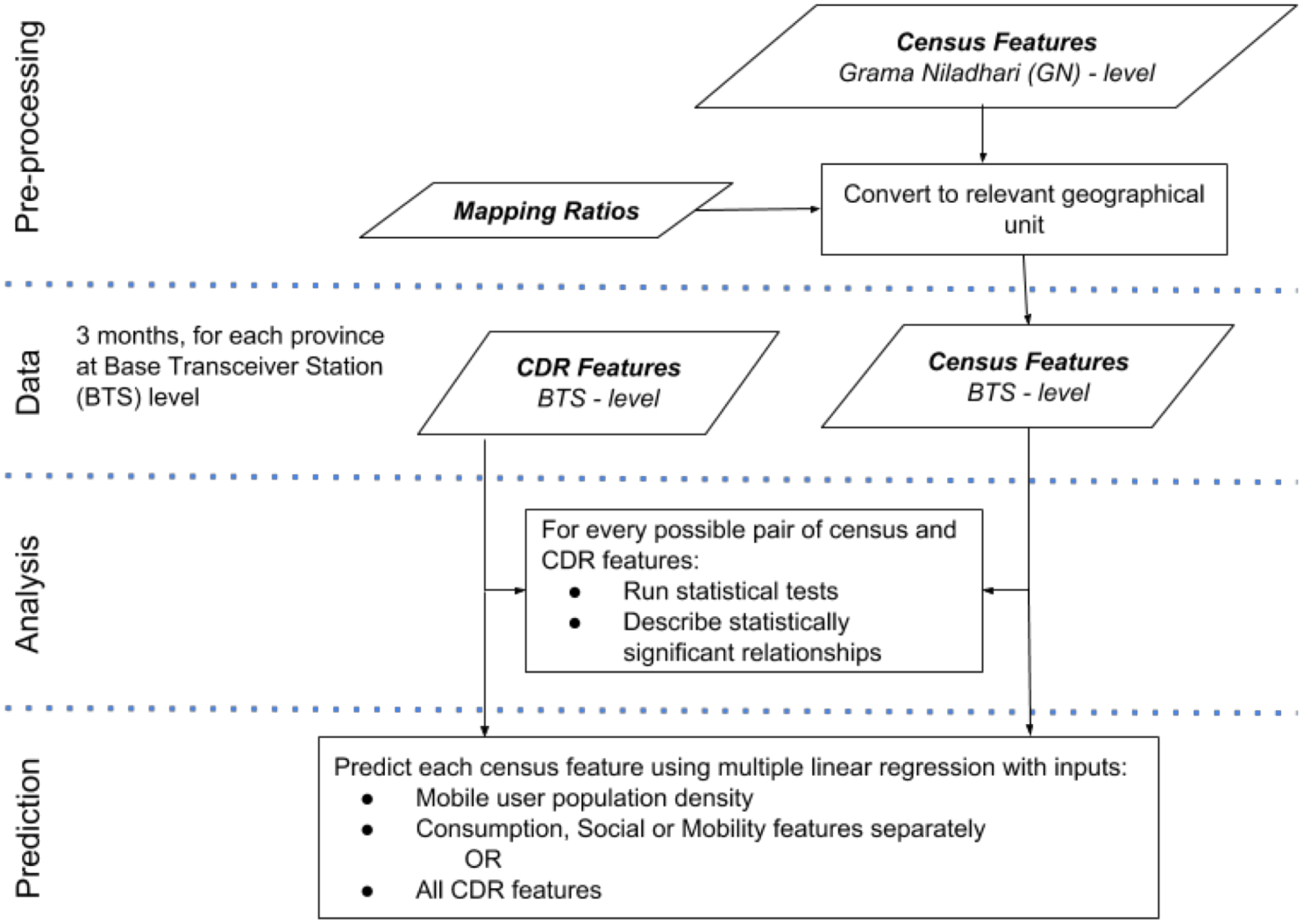
- Can Streetbump be transplanted to Colombo at this time?
  - Feature phones >> Smartphones
- "Something better than nothing" may not apply
  - Bias toward roads traversed by smartphone owners ➔ In conditions of limited resources, may skew resource allocation
- Is there a public policy solution that enables use of such apps even in developing countries with low smart phone penetration?
- Realizing benefits now will require governments to collaborate; cannot wait till they themselves build capacity

But other areas amenable to technical analyses exist

- Mapping socio-economic and demographic layers into our analyses of pseudonymized MNBD
  - Understanding mobility, social networks, consumption behavior disaggregated by gender, age, income level
  - Use case: Informing transportation policy

PROBING POVERTY: CAN MNBD YIELD INSIGHTS?

Methodological Overview



LIRNEasia
Pro-poor. Pro-market.

Summary of derived CDR features

Category	Features	Additional Description
Consumption	Total duration of incoming calls	Sum of duration of all incoming calls for a subscriber
	Total duration of outgoing calls	Sum of duration of all outgoing calls for a subscriber
	Total number of incoming calls	Count of incoming calls per subscriber
	Total number of outgoing calls	Count of outgoing calls per subscriber
	Total duration of all calls	Total duration incoming + Total duration outgoing
	Total number of calls	Total number of incoming + Total number of outgoing
Social	Contact count	Unique contacts per subscriber
	Contact rate	Total no. of calls / Total no. of contacts
	Physical distance of contacts	Avg (Distance to home location of each contact)
Mobility	Radius of gyration	Avg (Distance between home & visited BTS * # trips to that BTS)
	Unique cell counts	Unique cell visits
	Travel distance	Total distance travelled by a subscriber
	Maximum distance travelled	Maximum distance travelled

Census features used for our study

- We used 58 features from 12 categories from the 2011/12 National census
- Census feature categories were:
  - Floor Material
  - Roof Material
  - Wall Material
  - Type of Structure
  - Housing Type
  - Tenure
  - Cooking Fuel
  - Lighting
  - Education
  - Employment
  - Gender
  - Age
- Features related to education, employment, gender and age are reported at individual level
  - All other features are reported at household level
- Census features are reported at GN level
  - We scale it to BTS cell areas for comparison with CDR features

LIRNEasia
Pro-poor. Pro-market.

Why is mapping done from GN to BTS cell area?

In order to correlate aggregate CDR features to census features, we need to bring both values to the same spatial unit. We map GN to BTS areas.

Reason: Northern province has more GNs than BTS cells (Refer map)

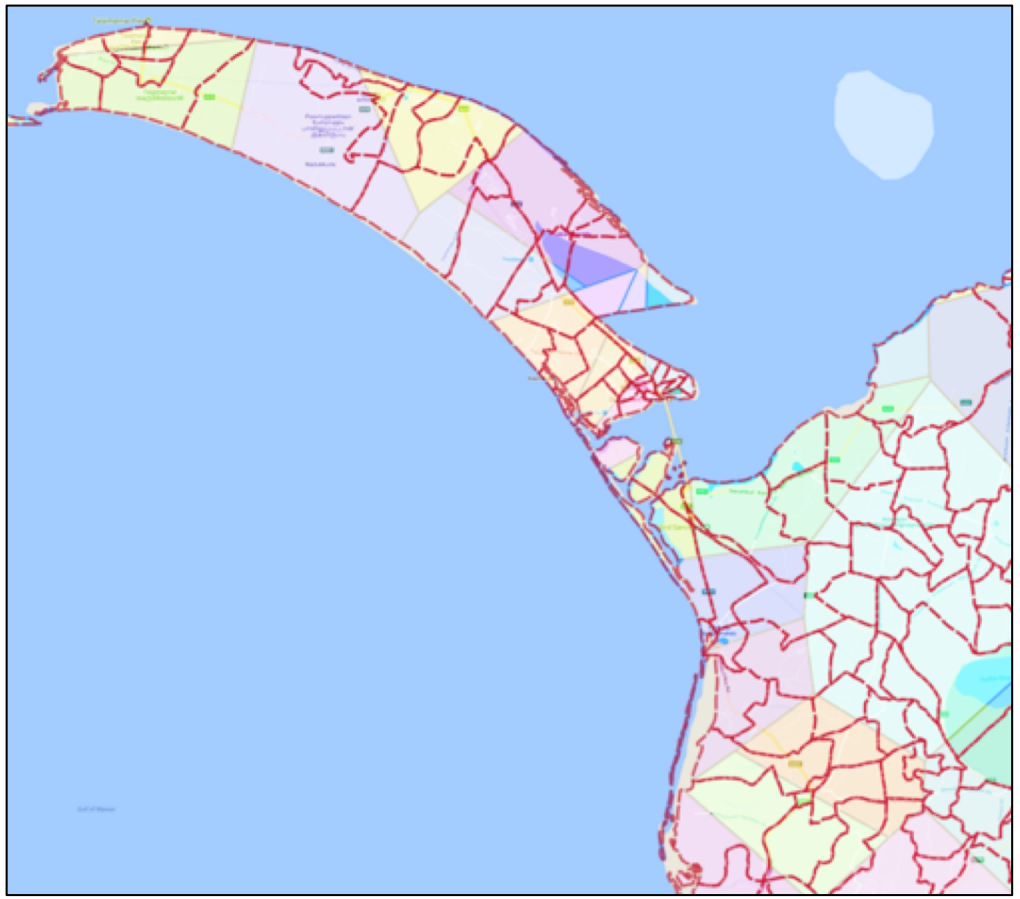
- 900+ GNs vs ~300 BTS cells
- GN boundaries are outlined in red while BTS cells are shaded in different colors

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

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

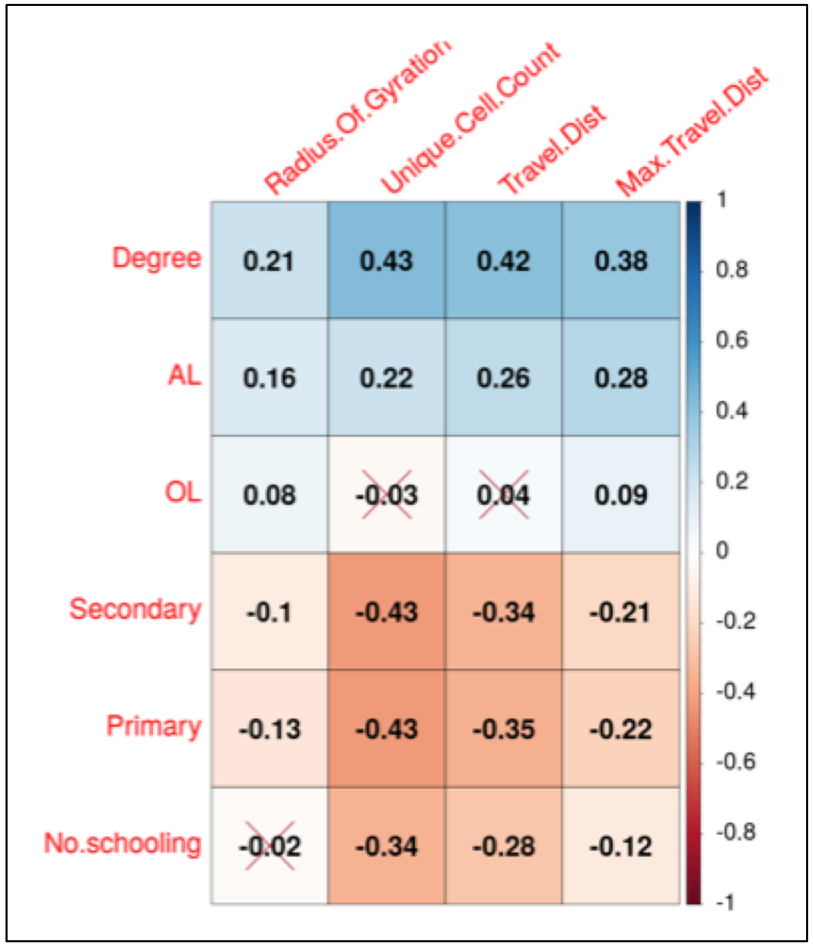
$pop(GN_i)$ = Population of GN division

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

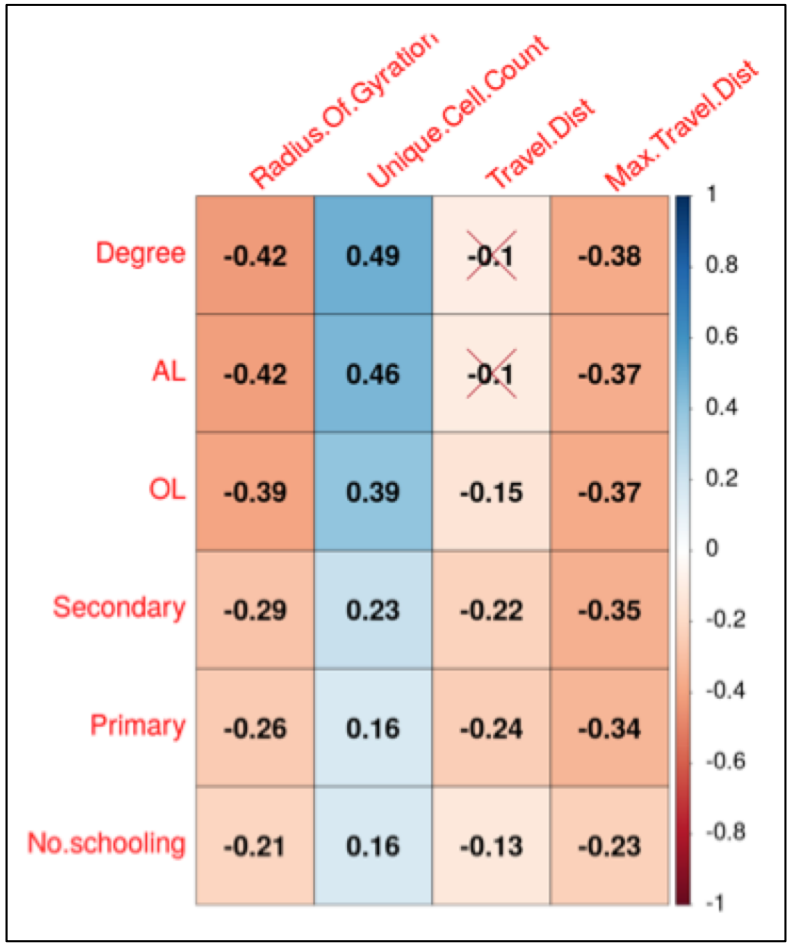


LIRNEasia
Pro-poor. Pro-market.

Correlation of education vs 'mobility'



Western Province

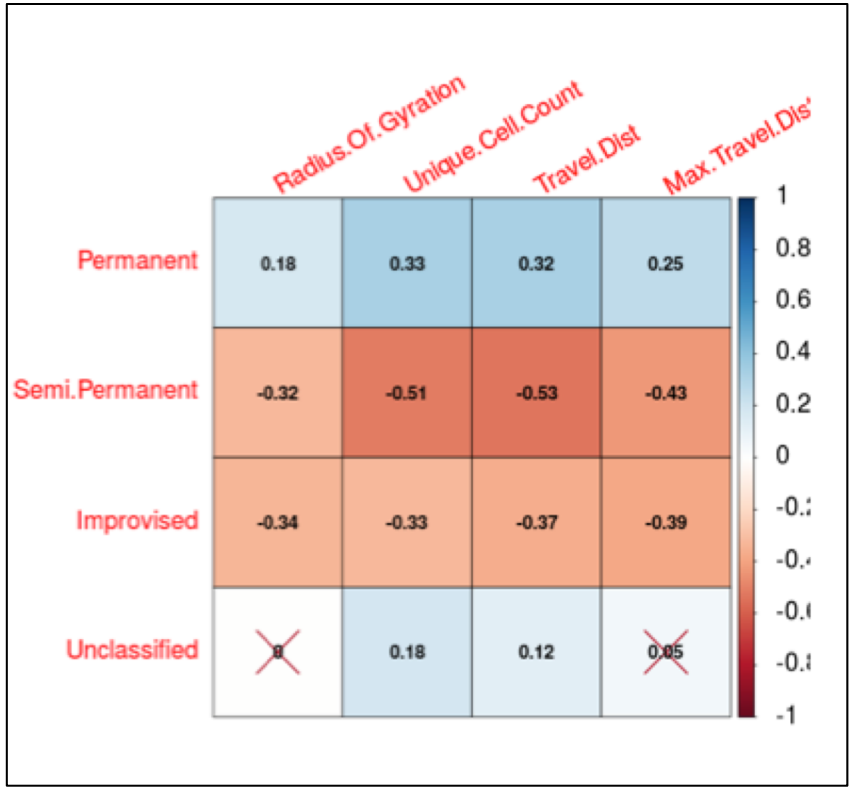


Northern Province

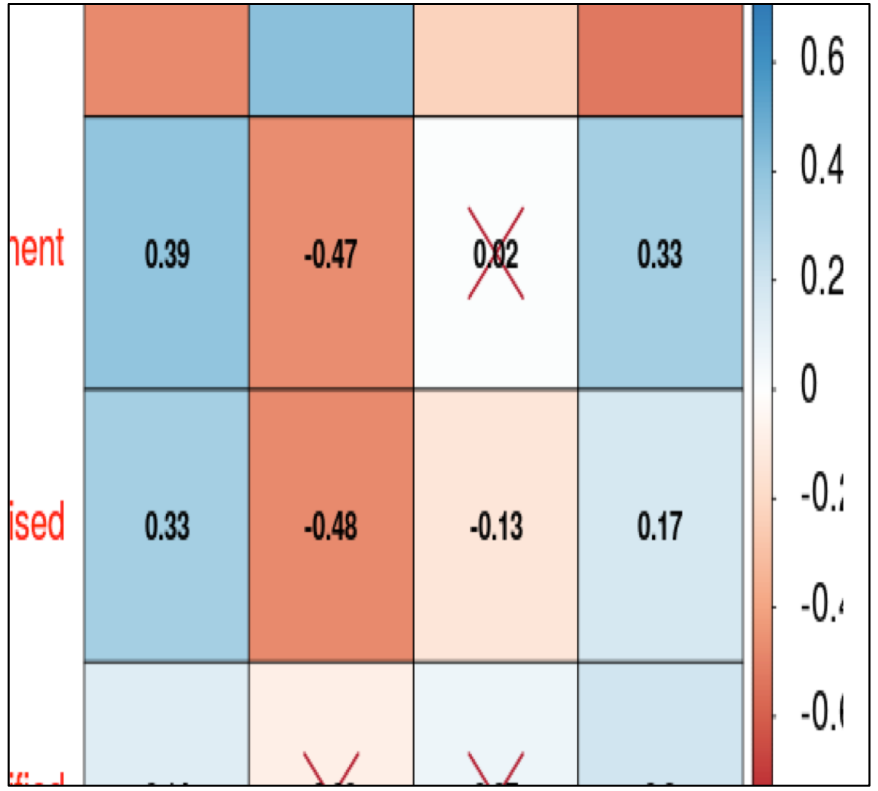
Correlation coefficients marked with a red cross indicate p-value > 0.001

LIRNEasia
Pro-poor. Pro-market.

Correlation of housing type vs 'mobility'



Western Province



Northern Province

Correlation coefficients marked with a red cross indicate p-value > 0.001

Correlation analysis: Observations from Western Province

- People in regions with higher SEL:
  - Have a higher number of unique contacts
  - Speak to their contacts more frequently (higher contact rate)
  - Have a greater geographic spread among their network (higher physical distance between contacts)
  - Travel to more unique BTS regions (higher unique cell counts)
- People in regions with lower SEL:
  - Have fewer unique contacts
  - Speak less frequently to their contacts (lower contact rate)
  - Have a smaller geographic spread among their contacts (lower physical distance between contacts)
  - Travel less frequently and for shorter distances (max. travel distance and radius of gyration is lower) - **This is in line with the findings of the study in Latin America**

LIRNEasia
Pro-poor. Pro-market.

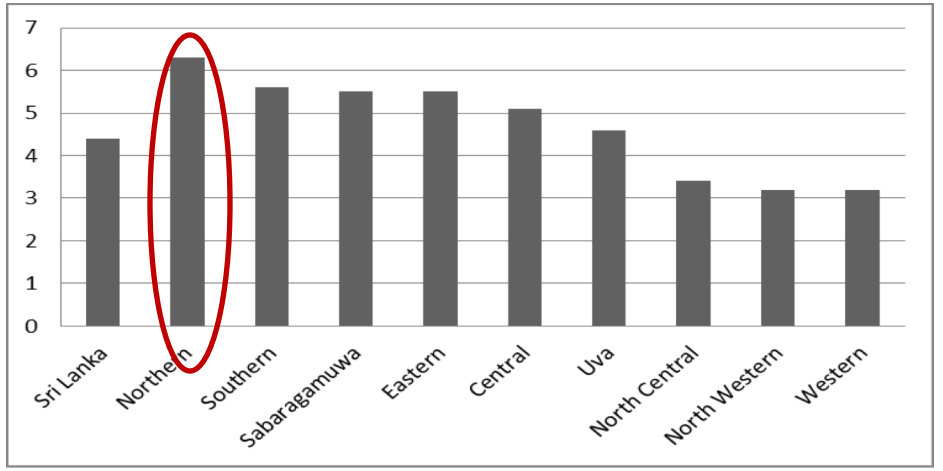
Correlation analysis: Observations from Northern Province

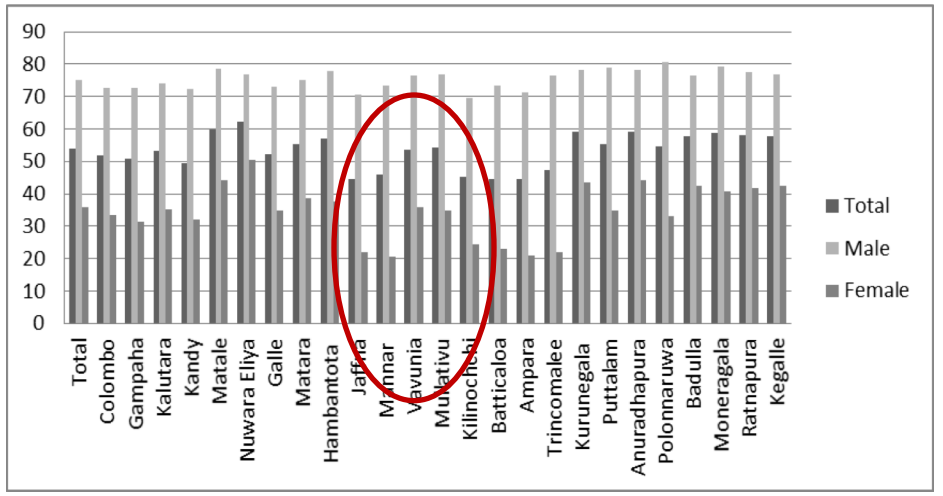
- People in regions with higher SEL:
  - Have a higher number of unique contacts
  - Speak to their contacts less frequently (lower contact rate)
  - Have a smaller geographic spread among their network (lower physical distance between contacts)
  - Travel to more unique BTS regions (higher unique cell counts)
- People in regions with lower SEL:
  - Have a fewer unique contacts
  - Speak more often to their contacts (higher contact rate)
  - Have a greater geographic spread among their contacts (higher physical distance between contacts)
  - Travel further and/or travel long distances more frequently (max travel distance and radius of gyration is higher) - **This is contrary to the findings of the study in Latin America**

LIRNEasia
Pro-poor. Pro-market.

“The outcome of any serious research can only be to make two questions grow where only one grew before”

- Usually, we use this Veblen quotation sardonically
- But in this instance, we are compelled to propose more research
  - Is Northern Province different because data is not representative?
  - Higher unemployment, esp among women?
  - Effects of war; more female-headed households?
  - Remittance-economy effects?





LIRNEasia
Pro-poor. Pro-market.

THANK YOU

www.lirneasia.net