Research on mobile network big data: First results

Rohan Samarajiva & Sriganesh Lokanathan

Dhaka, 14 December 2014

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

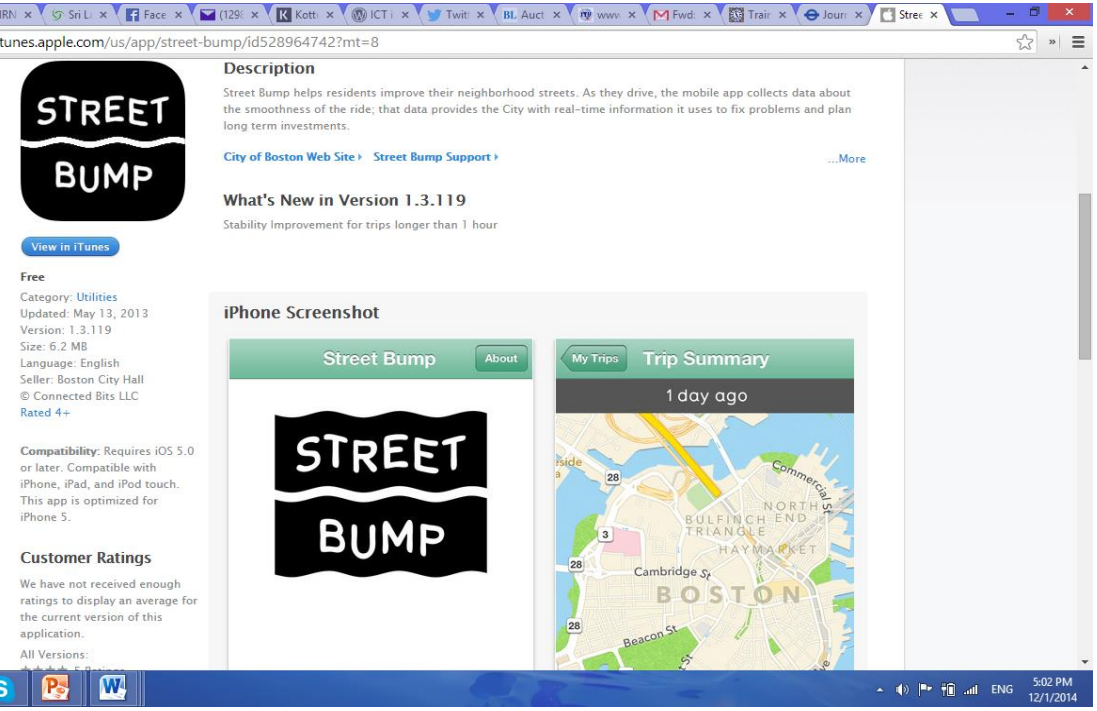
Our mission

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

LIRNEasia
Pro-poor. Pro-market.

Where we work

Bias in big data ➔ why mobile network big data in developing countries



- Streetbump is a Boston crowdsourcing + big data application that uses the natural movement of citizens to improve street maintenance
  - Data generated from an app downloaded to a smartphone "mounted" in a car

Can Streetbump be transplanted in 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

Mobile network big data are more inclusive, especially in our cities

	Mobile SIMs/100	Internet users/100	Facebook users/100
Myanmar	13	1	4
Bangladesh	67	7	6
Pakistan	70	11	8
India	71	15	9
Sri Lanka	96	22	12
Philippines	105	39	41
Indonesia	122	16	29
Thailand	138	29	46

Myanmar mobile SIMs/100 was 22.6 by September 2014

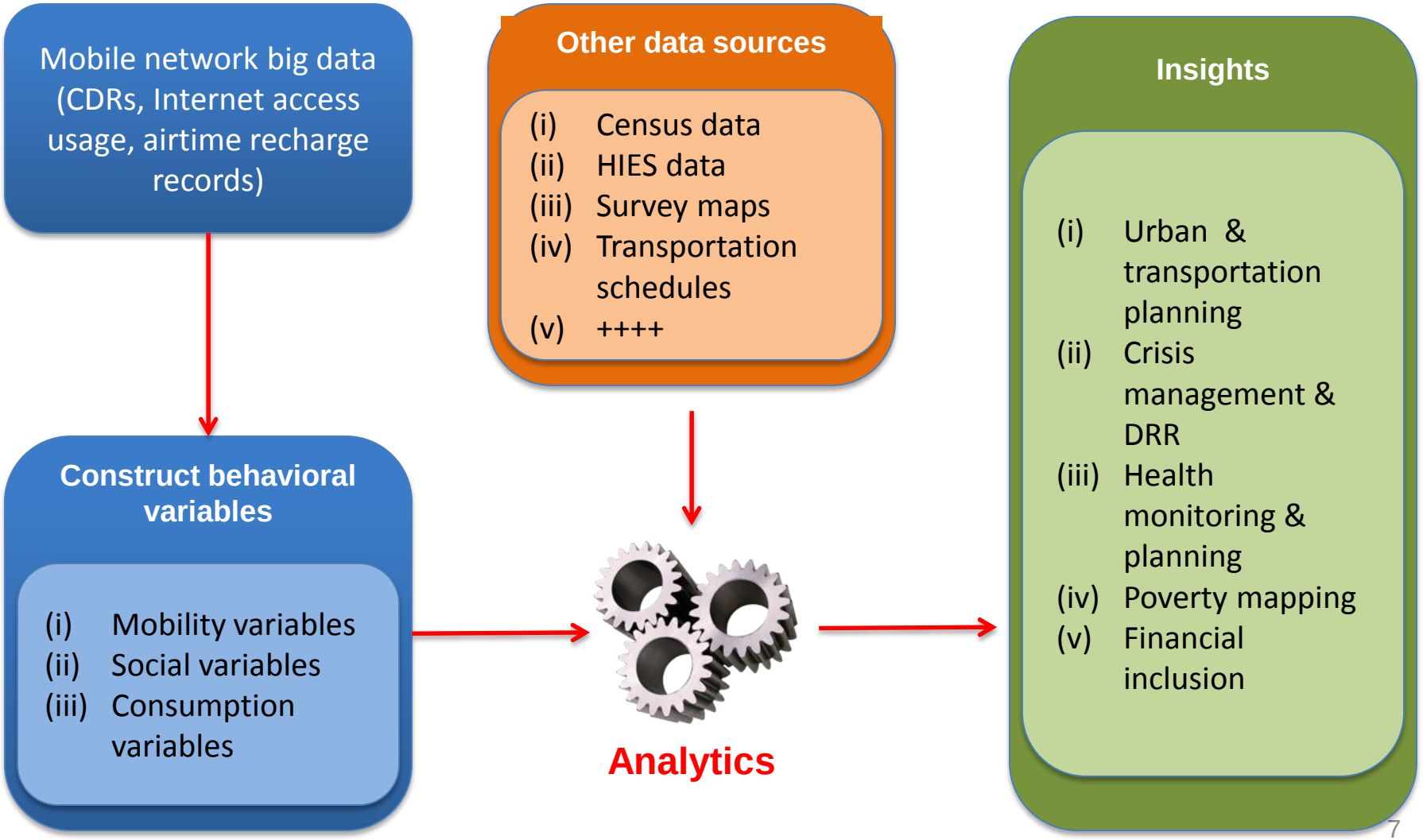
LIRNEasia
Pro-poor. Pro-market.

Source: ITU Measuring Information Society 2014; Facebook advtg portal

There is a role for other sources of big data

- But for smart cities, MNBD is the best
  - Low cost, compared to fitting all vehicles with GPS or electronic toll cards/toll infrastructure
  - But can/should be complemented with GPS and other sensor data
    - Proposed two-year study to LK Ministry of Urban Development, based on first results with MNBD, after which appropriate sensors can be installed
  - Global Pulse analysis of food-related Twitter content in Jakarta shows value in social media content, even if not as "representative"
- Visitor Location Register (VLR) data is best for physical mobility, but Call Detail Records (CDR) can serve as acceptable proxy
  - Something we plan to explore in relation to infectious diseases in 2015

Mobile network big data + other data ➔ rich, timely insights



Data used in the research

- Multiple mobile operators in Sri Lanka have provided four different types of meta-data
  - Call Detail Records (CDRs)
    - Records of calls
    - SMS
    - Internet access
  - Airtime recharge records
  - No Visitor Location Register (VLR) data
- Data sets do not include any Personally Identifiable Information
  - All phone numbers are pseudonymized
  - LIRNE*asia* does not maintain any mappings of identifiers to original phone numbers
- Cover 50-60% of users; very high coverage in Western (where Colombo the capital city in located) & Northern (most affected by civil conflict) Provinces, based on correlation with census data

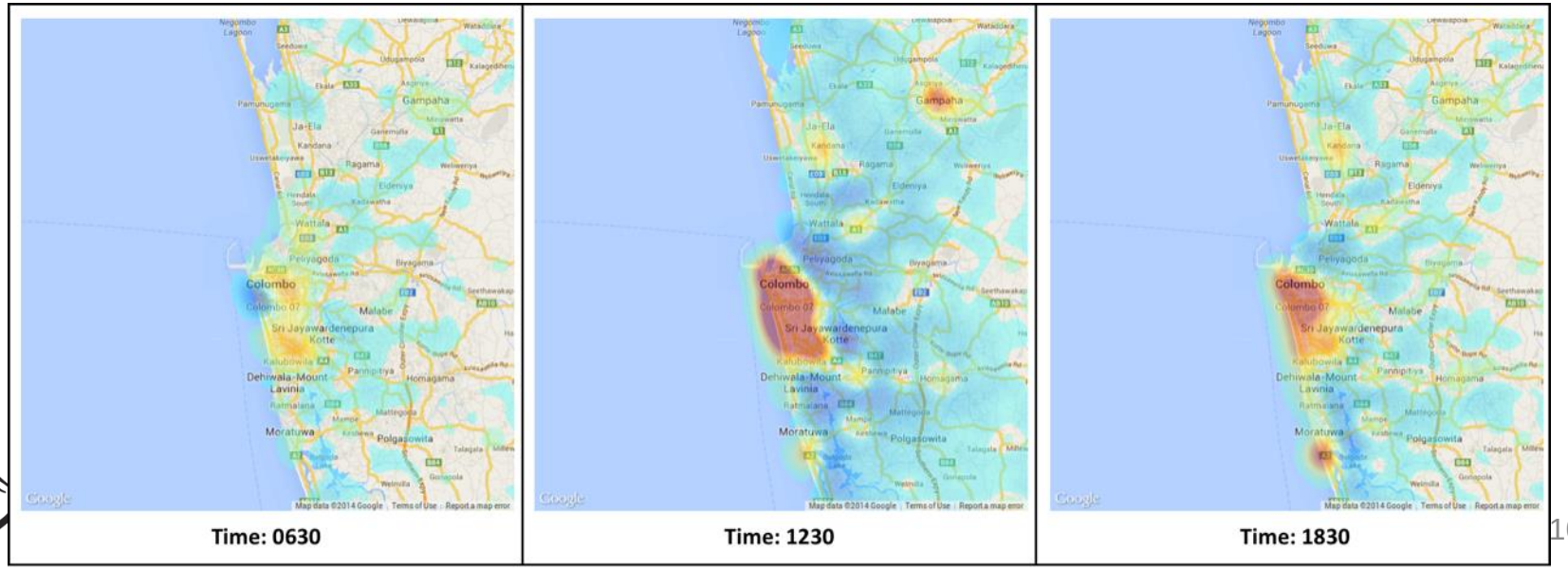
UNDERSTANDING CHANGES IN POPULATION DENSITY

Population density changes in Colombo region: weekday/ weekend

Pictures depict the change in population density at a particular time relative to midnight

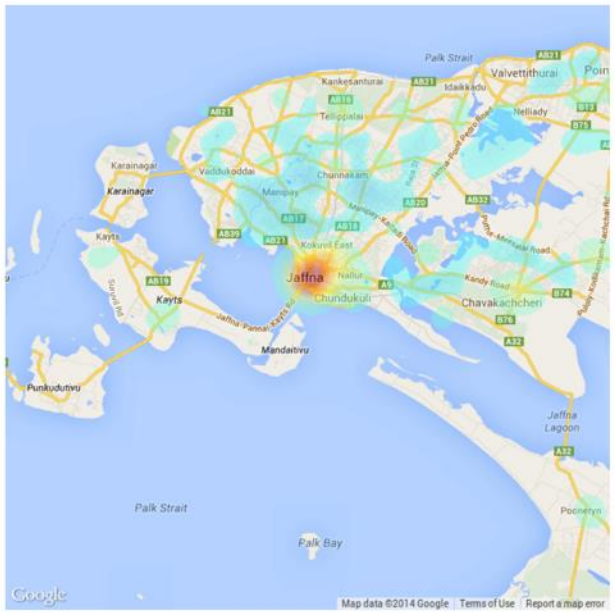
Weekday

Sunday

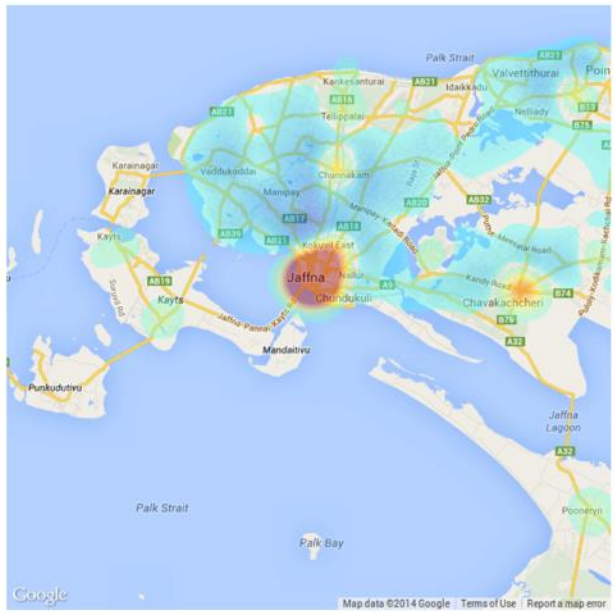


Population density changes in Jaffna region on a normal weekday

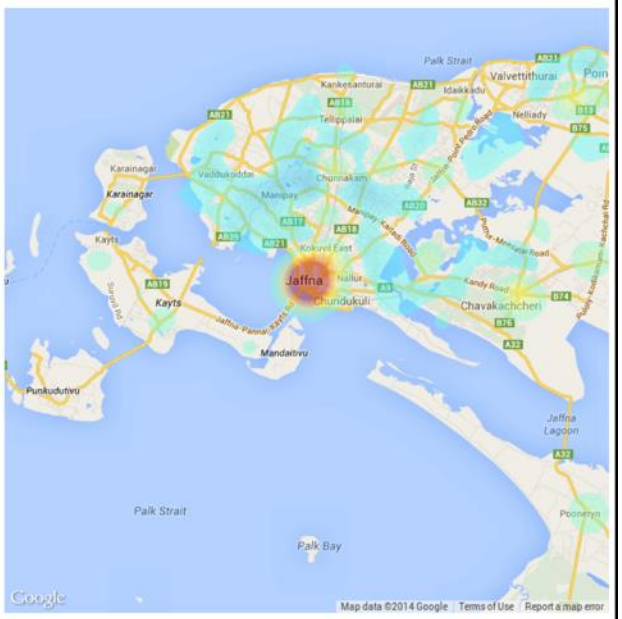
Pictures depict the change in population density at a particular time relative to midnight



Time: 0630

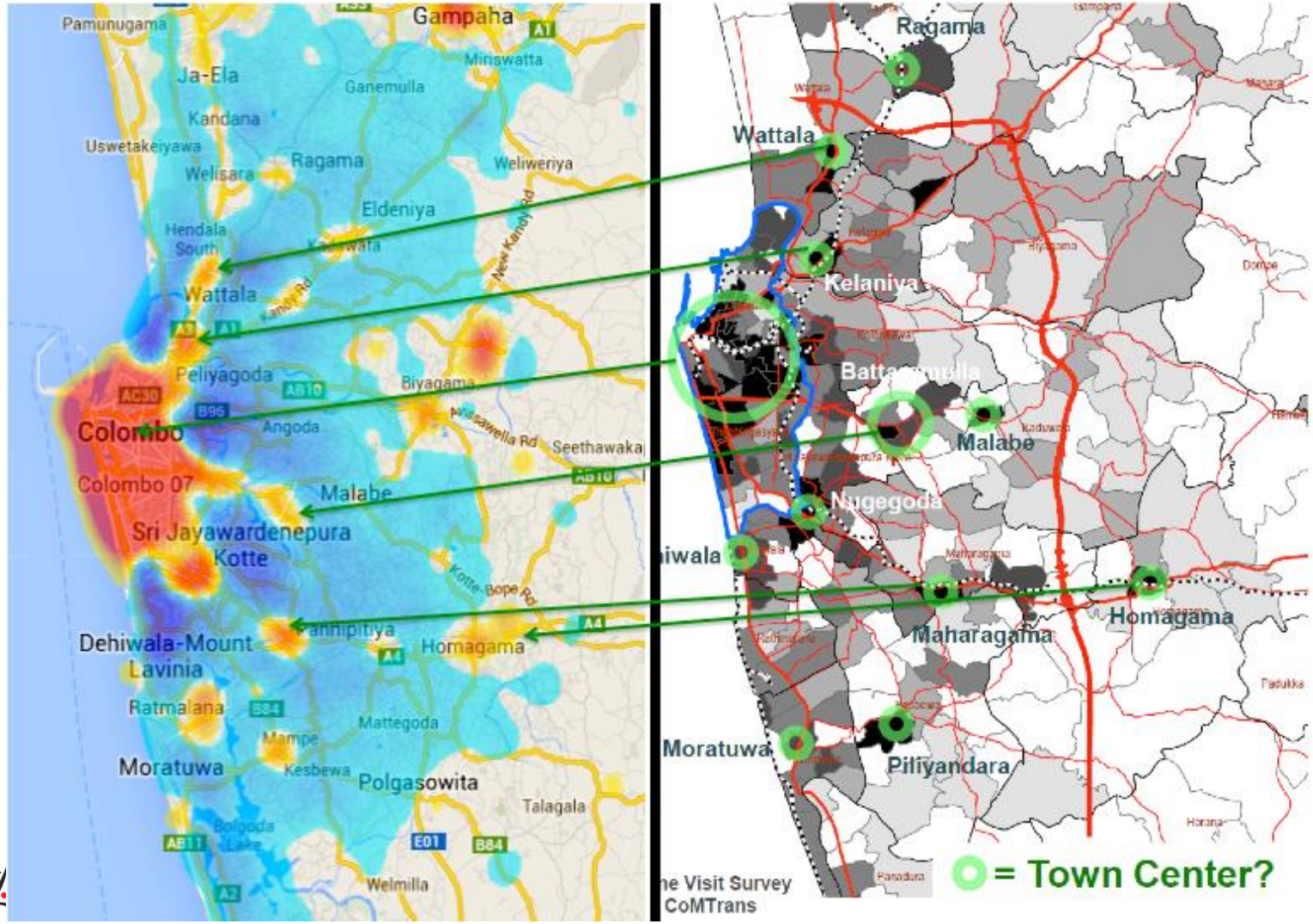


Time: 1230



Time: 1830

Our findings closely match results from expensive & infrequent transportation surveys

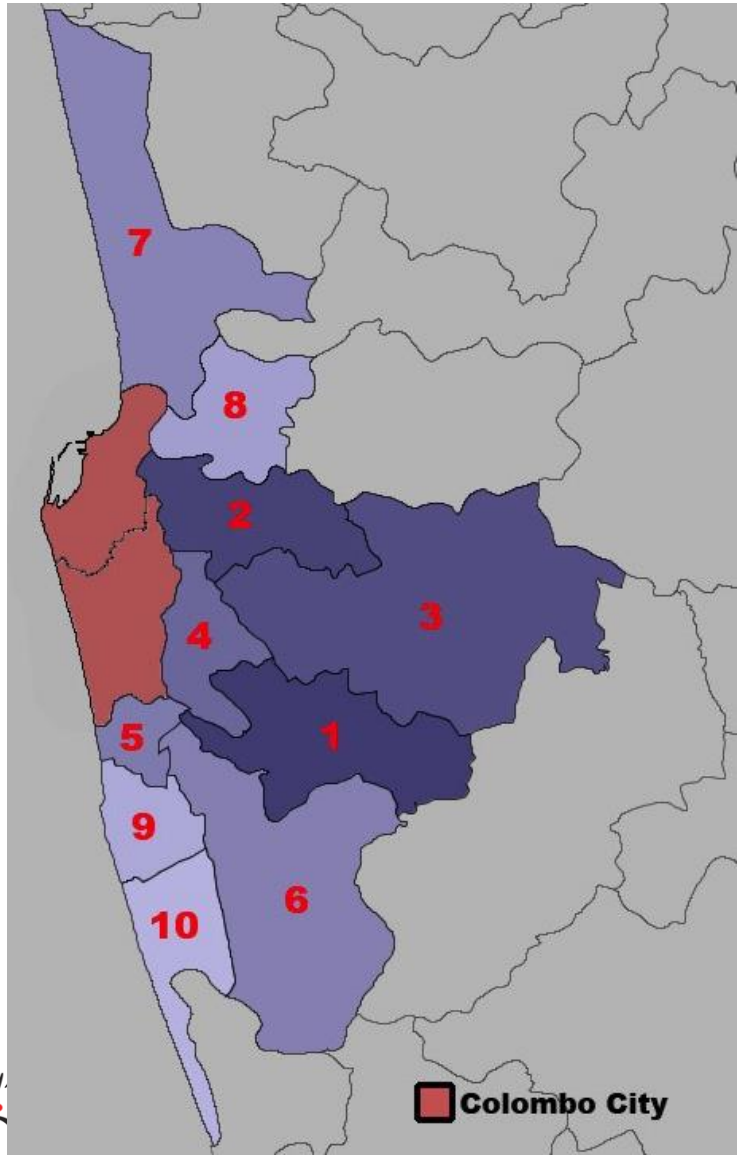


UNDERSTANDING WHERE PEOPLE LIVE AND WORK

Methodology

- Based on extracted average diurnal mobility pattern for population, choose time frames for home and work
  - Home time: 2100 to 0500
  - Work time: 1000 to 1500
- Calculate a home and work location for each SIM:
  - Match cell towers to Divisional Secretariat Division (DSD)
  - Count each DSD at most once per *day*.
  - Pick the DSD with the largest number of “hits”
  - For work consider only weekdays that are not public holidays

46.9% of **Colombo city's** daytime population comes from the surrounding regions



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

Implications for public policy

Urban Planning

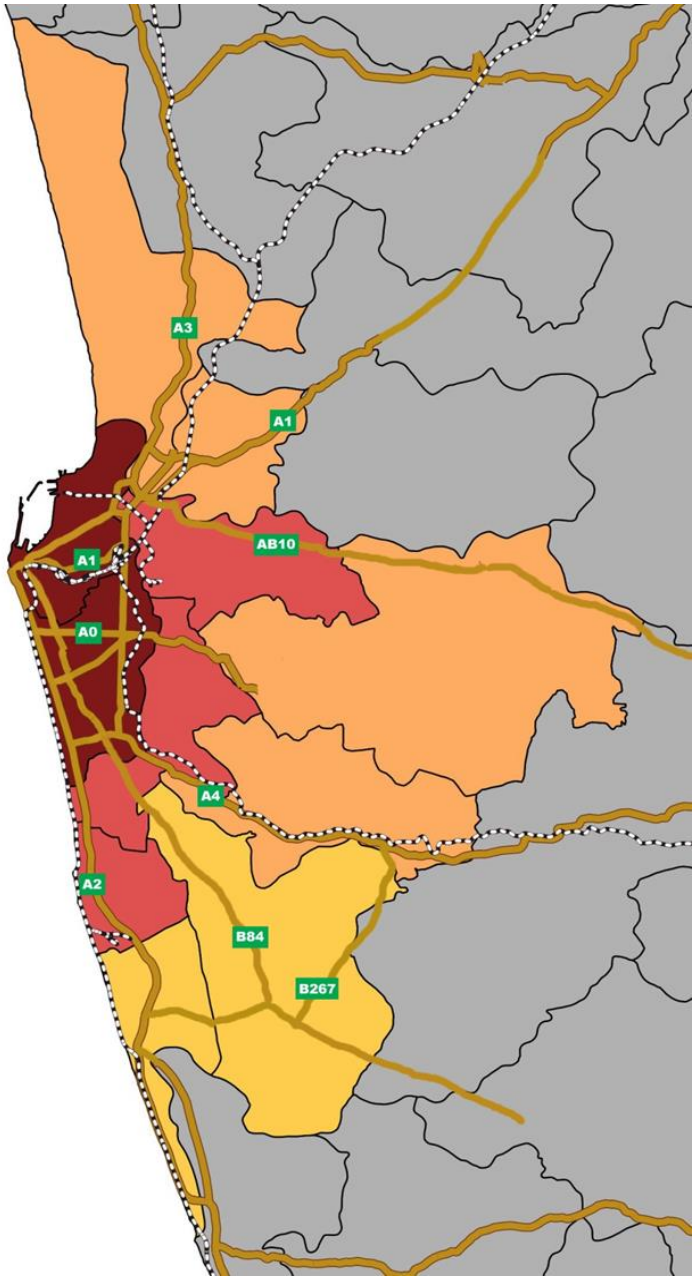
- Current municipal boundaries are obsolete; those from outside city limits cause costs but do not contribute adequate revenues; our data indicate logical boundaries of metro regions

Transportation Policy

- High volume transport corridors suitable for provision of mass transit
  - Kaduwela DSD (now served by AB 10 & A0) (3) already identified
  - High Level Road (now served by A4 & rail line) to Maharagama DSD (1)

Health Policy

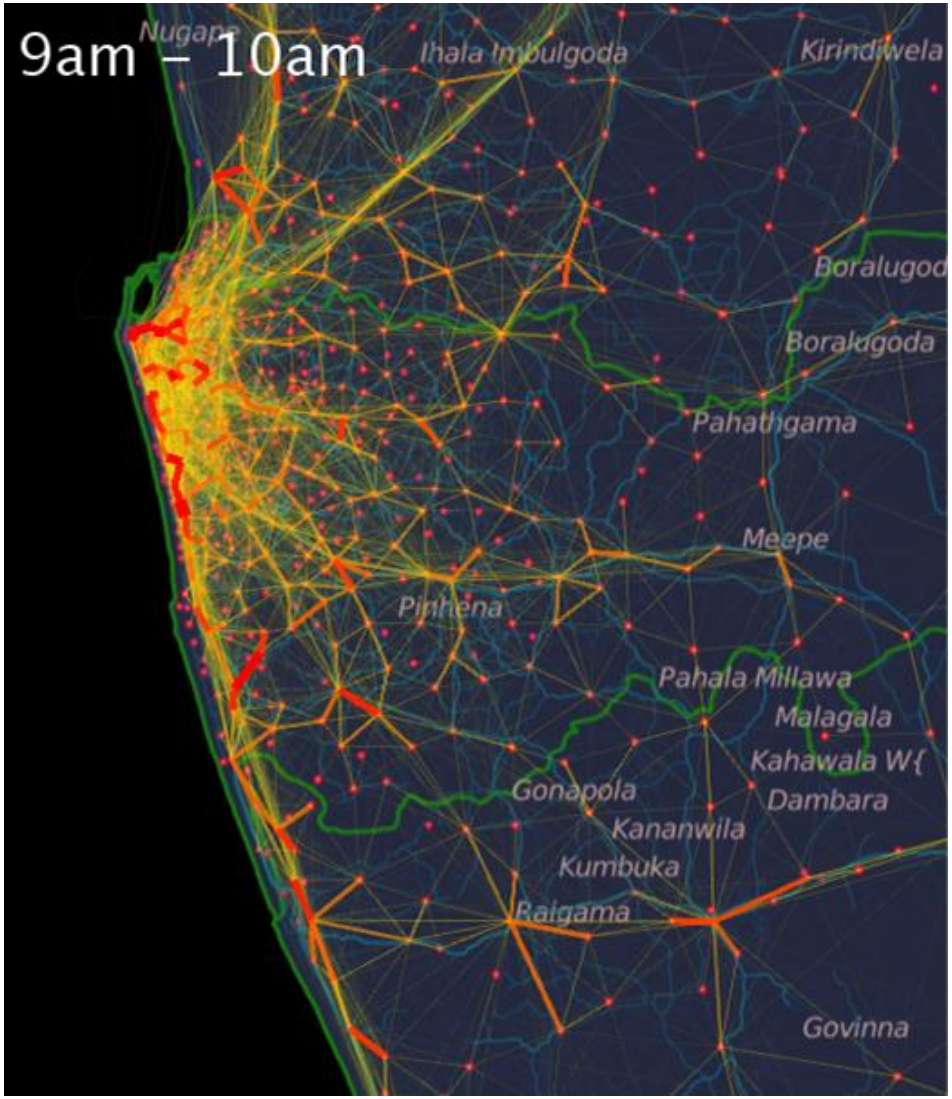
- Understanding people's regular mobility patterns can help model spread of infectious diseases (e.g. dengue)



LIRNEasia
Pro-poor. Pro-market.

MODELING TRAVEL

Understanding temporal variations in trips

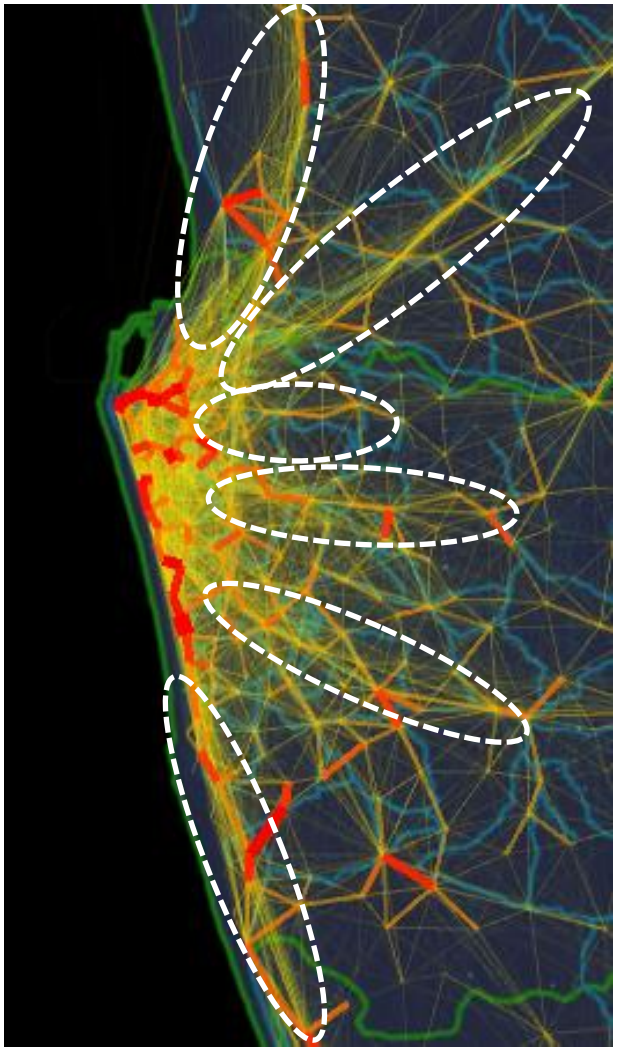


Low → High

Volume of people

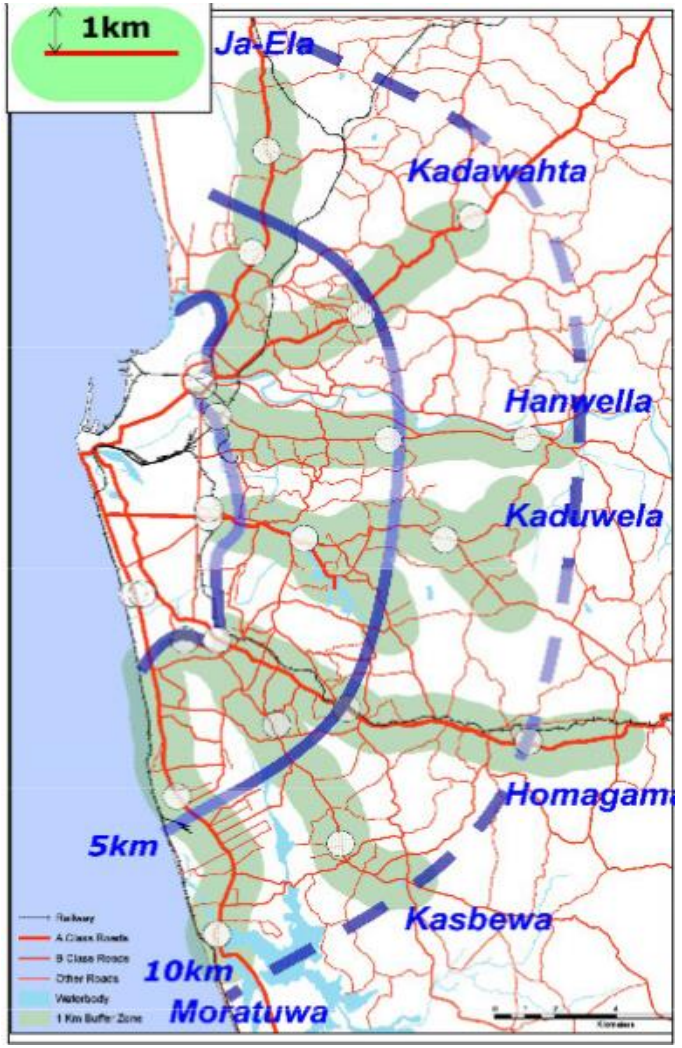
LIRNEasia
Pro-poor. Pro-market.

Mobility visualization for Colombo District identifies transport corridors



Low → High

Volume of People



Source: COMTRANS report,2013, Ministry of Transport

Implications for public policy

Transportation & Urban Policy

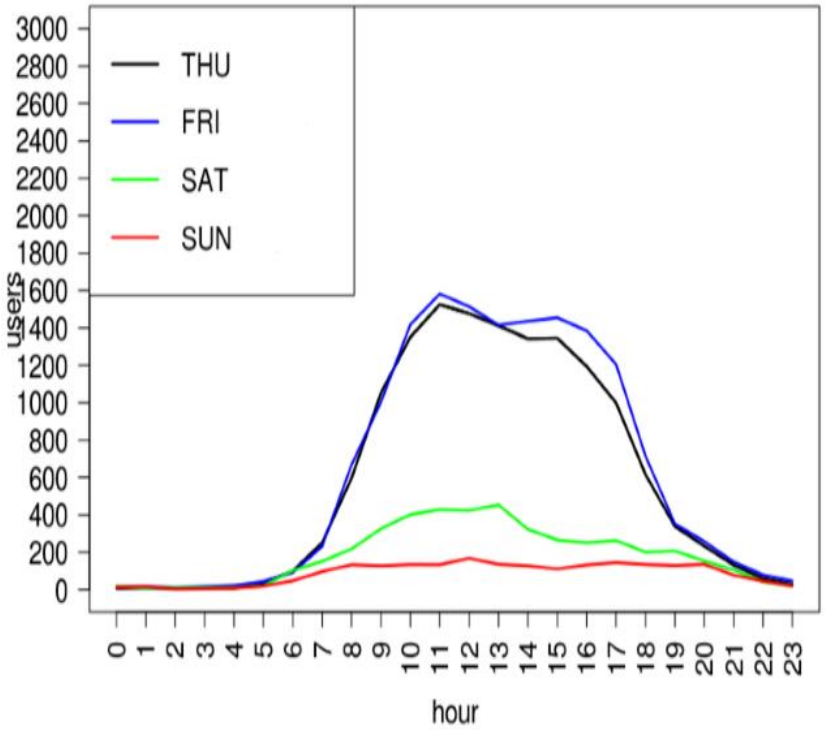
- CDR analysis can give us rough insights on principal transport corridors
- Then, with cooperation of mobile operators and additional computing power, we can zoom in on priority corridors to do detailed analysis using Visitor Location Register (VLR) data

Health Policy

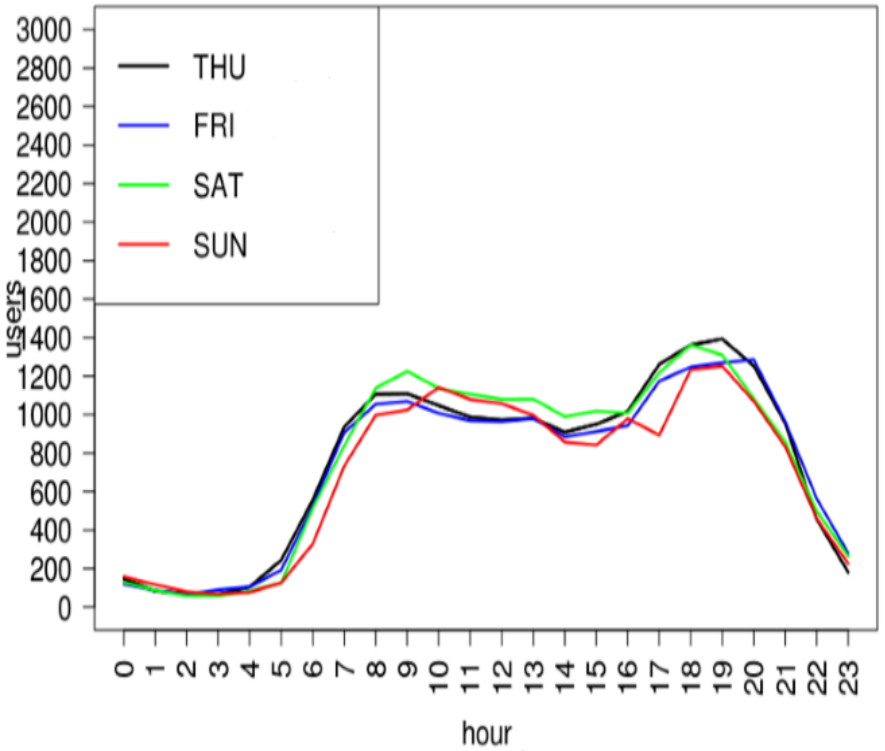
- Understanding people's regular mobility patterns can help model spread of infectious diseases (e.g. dengue)

UNDERSTANDING LAND USE CHARACTERISTICS

Hourly loading of base stations reveals distinct patterns



Type X: ?

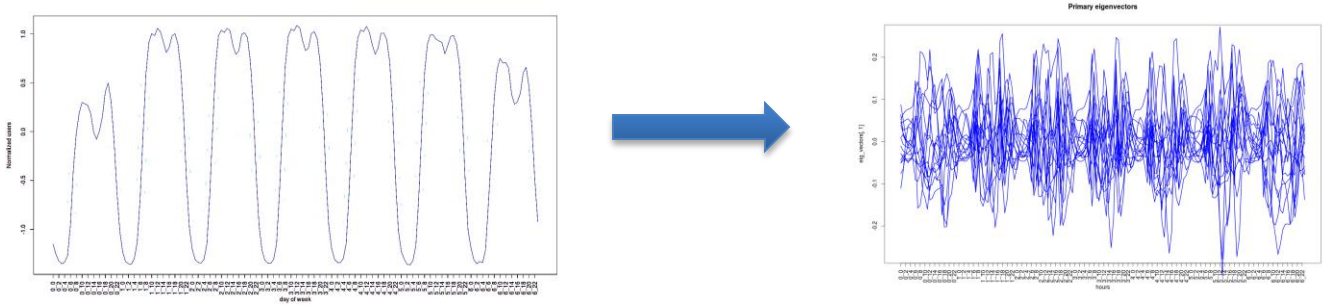


Type Y: ?

- We can use this insight to group base stations into different groups, using unsupervised machine learning techniques

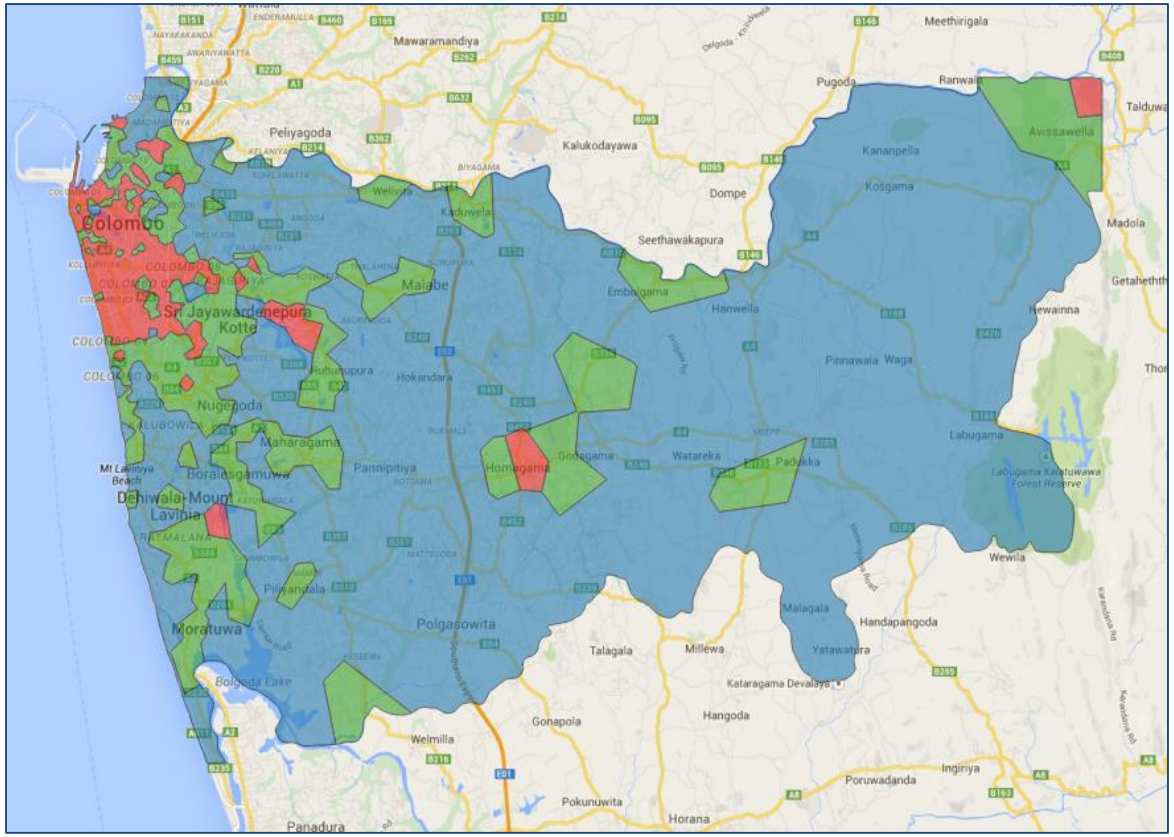
Understanding land use characteristics: methodology

- The time series of users connected at a base station contains variations, that can be grouped by similar characteristics
- A month of data is collapsed into an indicative week (Sunday to Saturday), with the time series normalized by the z-score
- Principal Component Analysis(PCA) is used to identify the discriminant patterns from noisy time series data



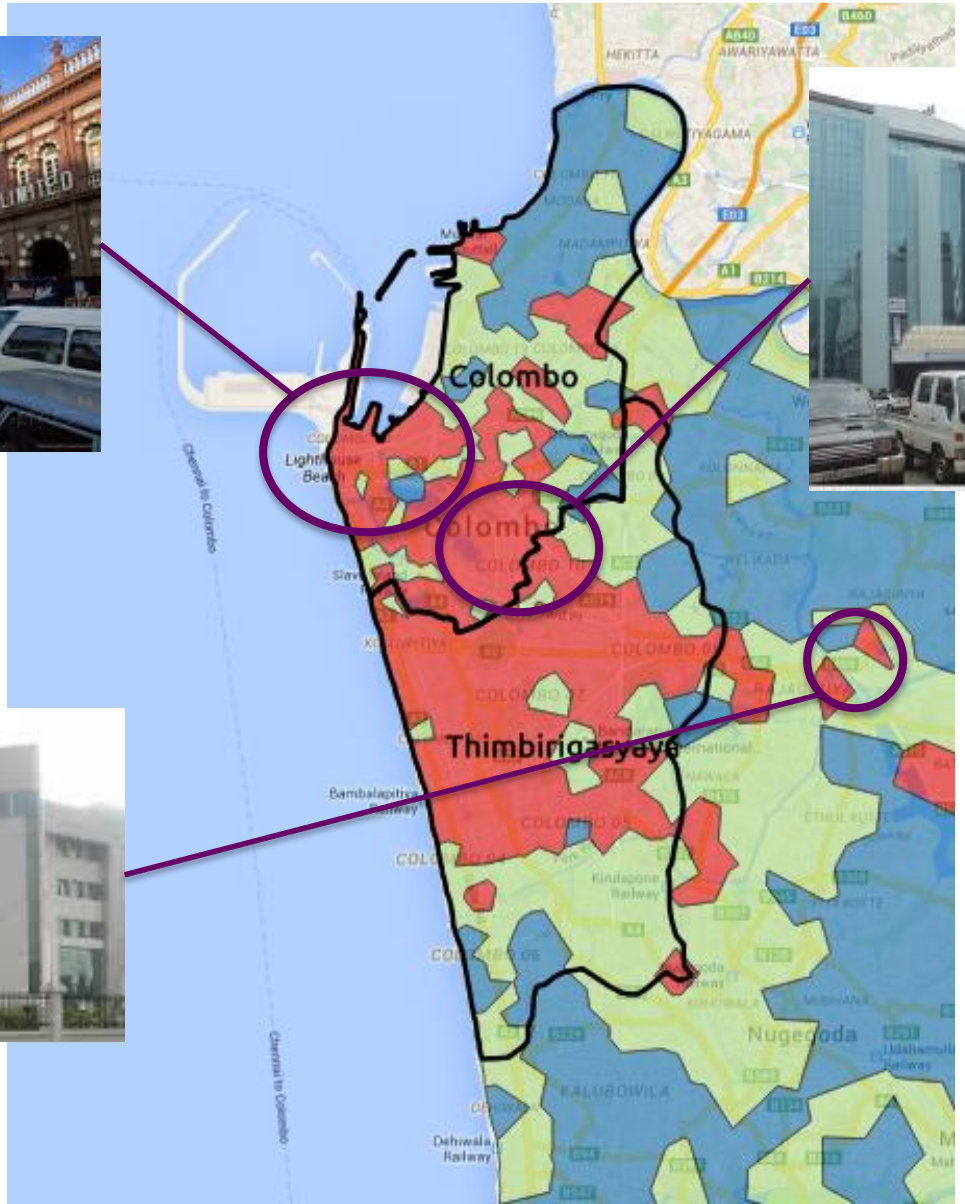
  - Each base station's pattern is filtered into 15 principal components (covering 95% of the data for that base station)
- Using the 15 principal components, we cluster all the base stations into 3 clusters in an unsupervised manner using k-means algorithm

Three spatial clusters in Colombo District



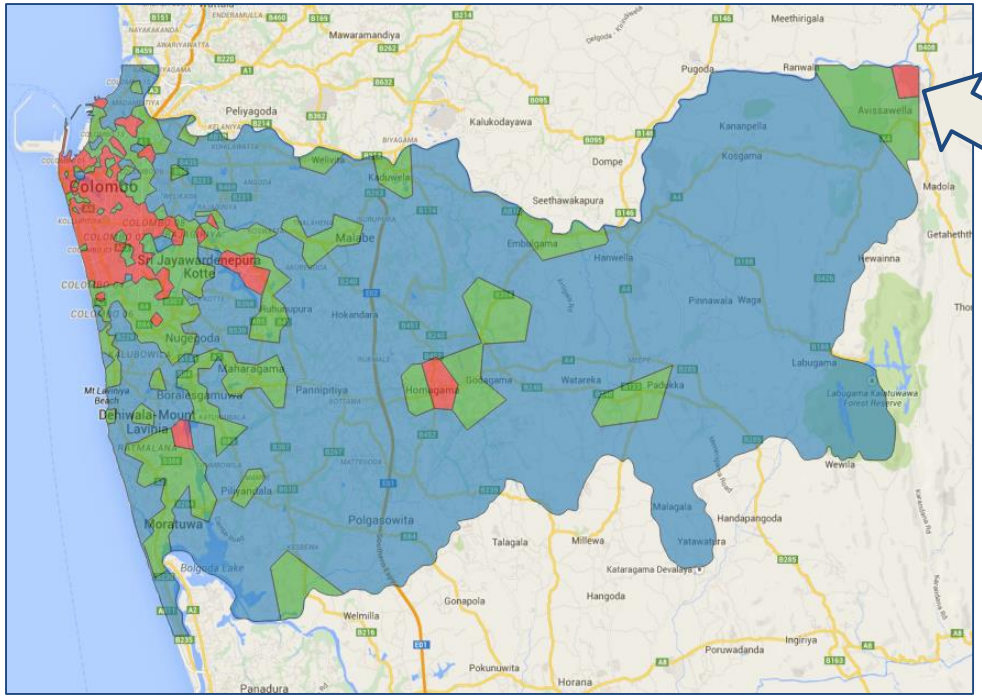
- **Cluster-1 exhibits patterns consistent with commercial area**
- **Cluster-3 exhibits patterns consistent with residential area**
- **Cluster-2 exhibits patterns more consistent with mixed-use**

Our results show Central Business District (CBD) in Colombo city has expanded



LIRNEasia
Pro-poor. Pro-market.

Small area in NE corner of Colombo District classified as belonging to Cluster 1?



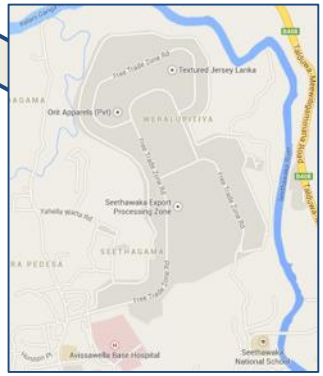
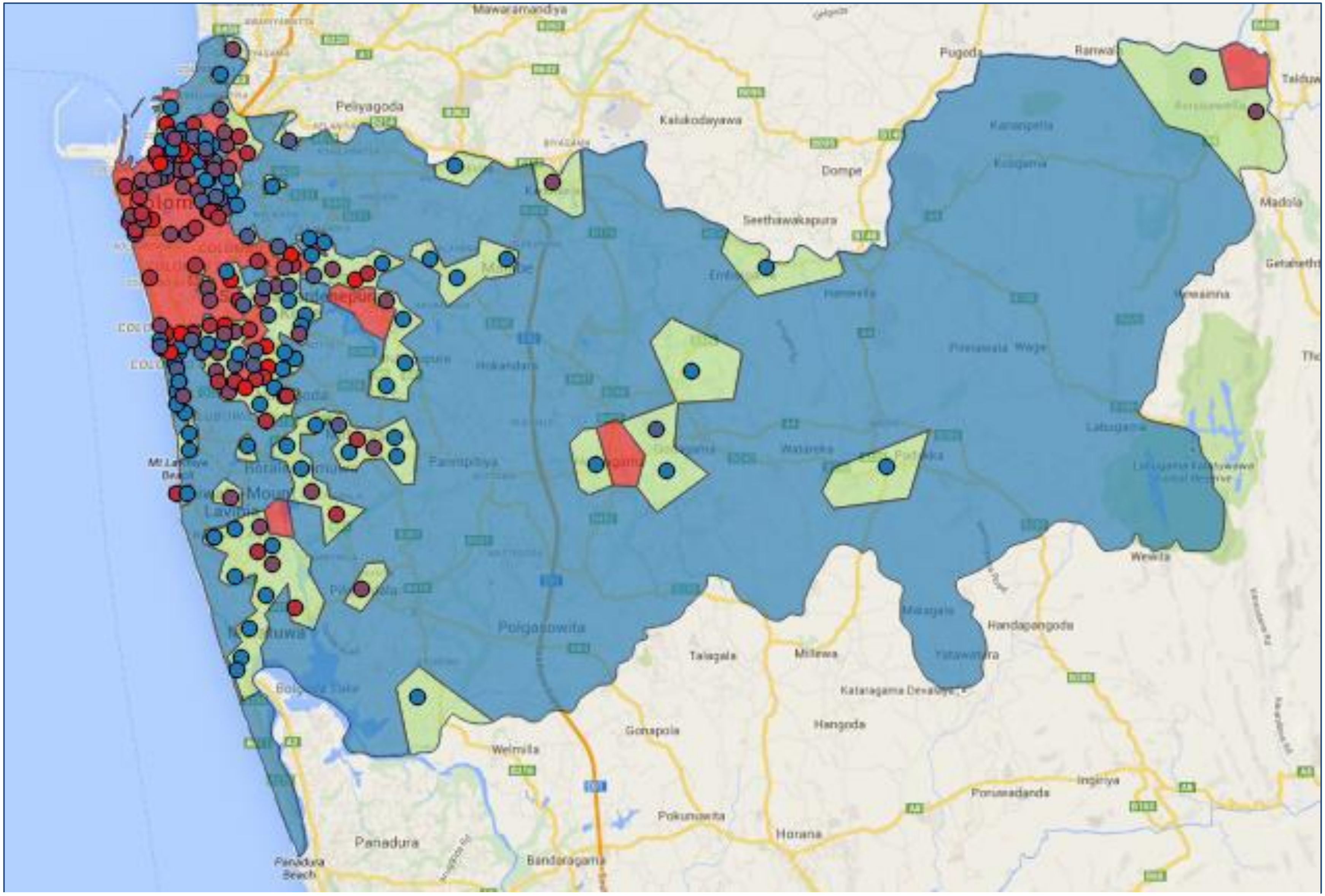


Photo ©Senanayaka Bandara - Panoramio

Seethawaka Export Processing Zone

Internal variations in mixed use regions: More commercial or more residential?

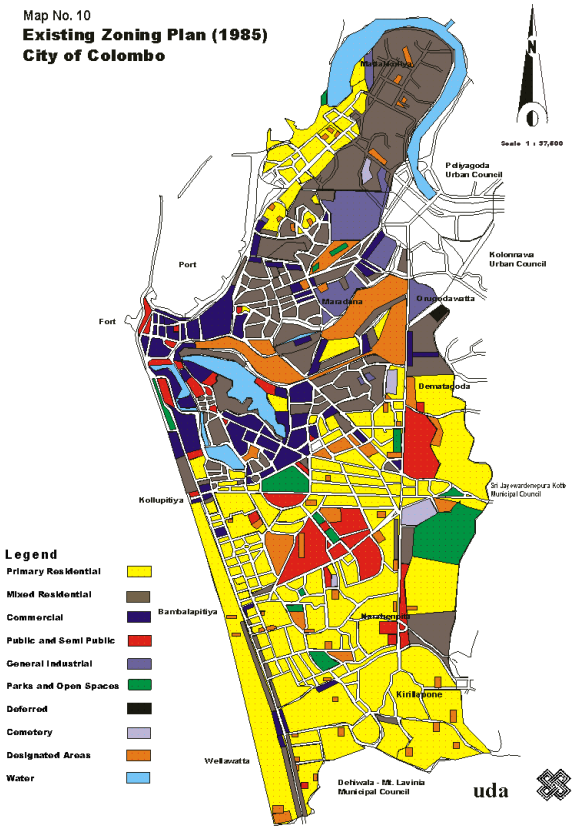


Blue dots: more residential than commercial

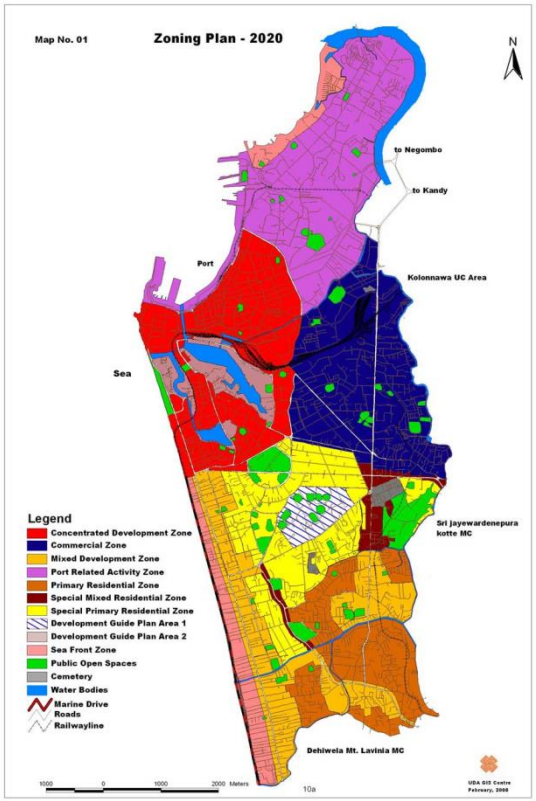
Red dots: more commercial than residential

Plans & reality

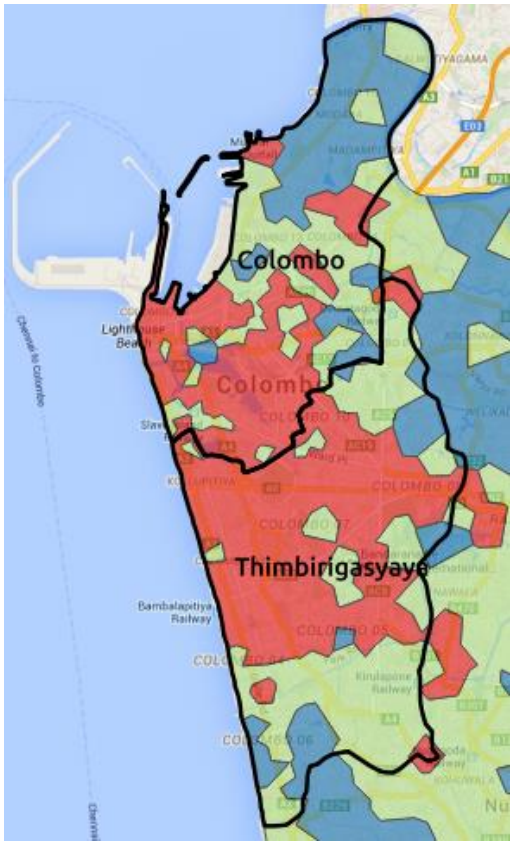
1985 Plan



2020 UDA Plan



2013 reality



Implications for urban policy

- Almost real-time monitoring of urban land use
  - We are currently working on understanding temporal variations in zone characteristics (especially the mixed-use areas)
- Can dispense with surveys & align master plan to reality
- LIRNEasia is working to unpack the identified categories further, e.g.,
  - Entertainment zones that show evening activity

UNDERSTANDING COMMUNITIES

Identifying communities: Methodology

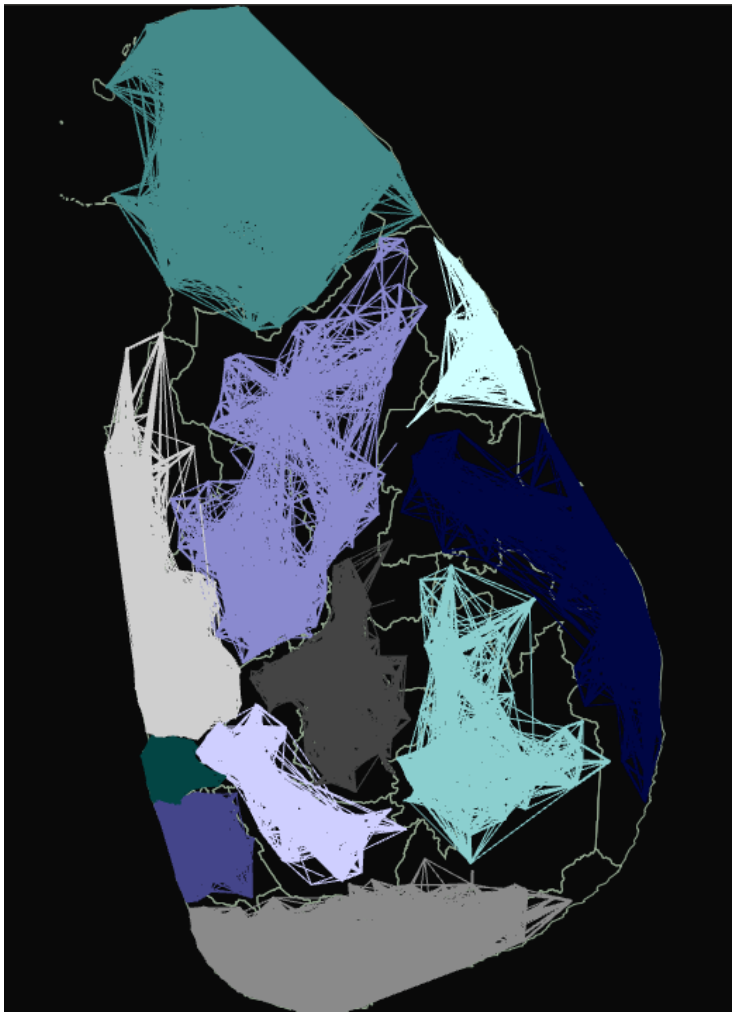
- Social networks segregated so overlapping connections between communities are minimized
- Strength of a community is determined by *modularity*
  - Modularity Q = (edges inside the community) – (expected number of edges inside the community)

$$Q = \frac{1}{2m} \sum_{a,b} (A_{a,b} - \frac{k_a k_b}{2m}) \delta(c_a, c_b)$$

M. E. J.-Newman, Michele-Girvan, "Finding and evaluating community structure in networks", Physical Review E, APS, Vol. 69, No. 2, p. 1-16, 204.

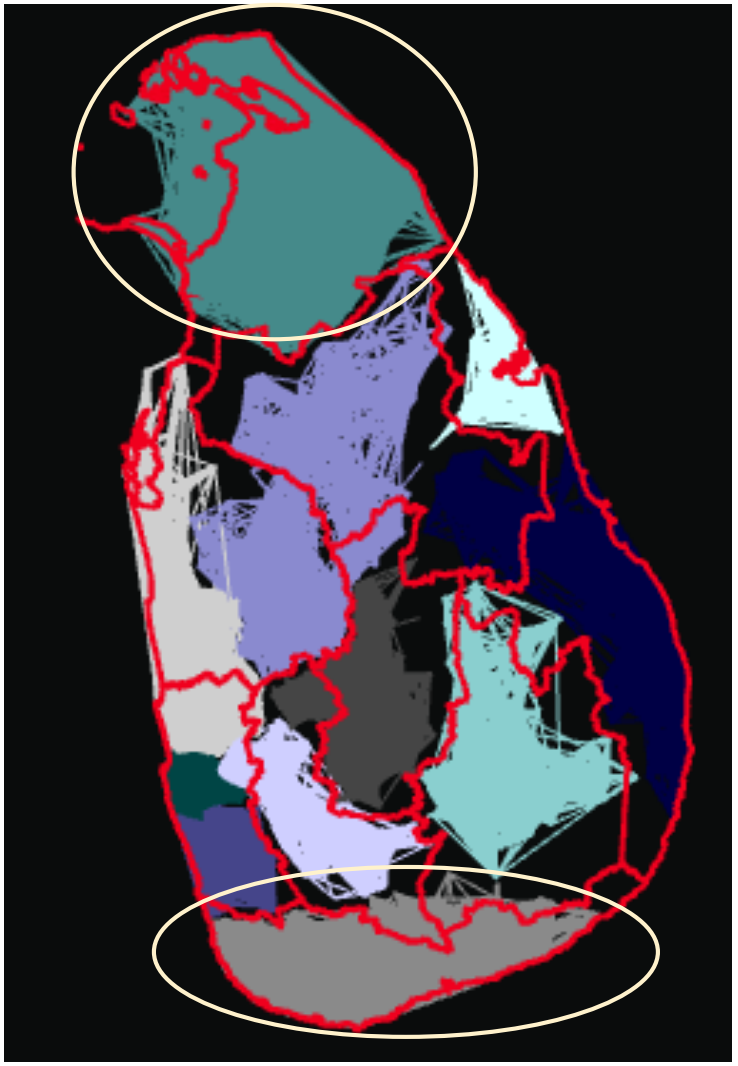
Resultant communities

- The optimal number of communities discovered by the algorithm was 11



How do these communities mesh with existing administrative boundaries?

- Southern & Northern provinces mesh the best
- Surprisingly, also Uva, hitherto thought to have a district aligned with Central plantations & other with Southern Province
- Eastern Province most intriguing
  - Trinco on its own
  - Polonnaruwa in NCP (predominantly Sinhala) tied to Batticaloa (Tamil/Muslim) and Ampara (Muslim/Sinhala) districts through rice economy
- Rest suggests administrative boundaries have been transcended



LIRNEasia
Pro-poor. Pro-market.

EXPLAINING CALLING BEHAVIOR: A GRAVITY MODEL APPROACH

Who Calls Whom in Sri Lanka?

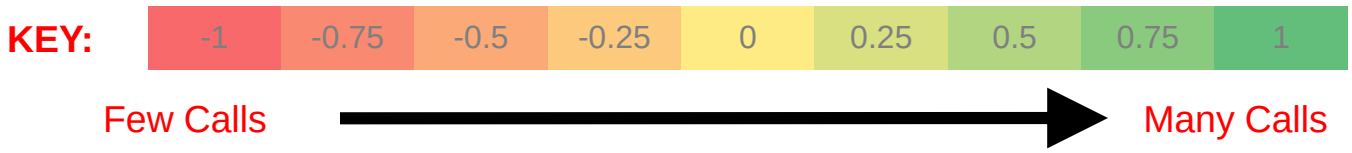
- We seek to understand the pattern of aggregate communication between areas
  - Depends on multiple factors: business and personal relations, ethnic, religious similarity, etc.

- We focus on three simple factors:
  - Population
  - Distance
  - IT literacy (broadly defined)

Total log, normalized volume of Calls between Provinces

- Within-province calls are generally high
- Western calls the most (both intra- and inter-province)

	N	NW	NC	E	Uva	Central	Saba	W	S
N	1.26	-2.13	-1.91	-2.01	-2.29	-1.83	-2.60	-0.77	-2.29
NW	-2.12	-0.53	-2.28	-3.42	-3.43	-2.36	-3.07	-0.76	-3.12
NC	-1.91	-2.28	-0.90	-2.89	-3.34	-2.11	-3.38	-1.19	-3.15
E	-2.01	-3.42	-2.89	-0.80	-2.85	-2.79	-3.68	-1.61	-3.47
Uva	-2.29	-3.43	-3.34	-2.85	-0.18	-2.32	-2.50	-0.80	-2.29
Central	-1.83	-2.36	-2.11	-2.79	-2.32	0.0	-2.48	-0.52	-2.83
Saba	-2.60	-3.07	-3.38	-3.68	-2.50	-2.48	-0.65	-0.32	-2.46
W	-0.77	-0.77	-1.19	-1.61	-0.80	-0.52	-0.32	2.33	-0.17
S	-2.30	-3.12	-3.15	-3.47	-2.29	-2.83	-2.46	-0.17	0.12



A Simple "Gravity" Model of Communication between pairs of DSDs

- Natural to assume:
  - Volume of calls increasing in population (at both ends)
  - Volume of calls decreasing in distance between DSDs
- Suggests a simple formula:

$$\textbf{Volume}\ (DSD_1, DSD_2) \propto \frac{\text{Pop}\ (DSD_1) \times \text{Pop}(DSD_2)}{\text{Distance}(DSD_1, DSD_2)}$$

- Similar formulas used extensively to understand transportation and trade flows.

Formula Validated by Regression Analysis

- The formula explains 42% of the variance in the data.
- Coefficients have the expected sign and *magnitude*.

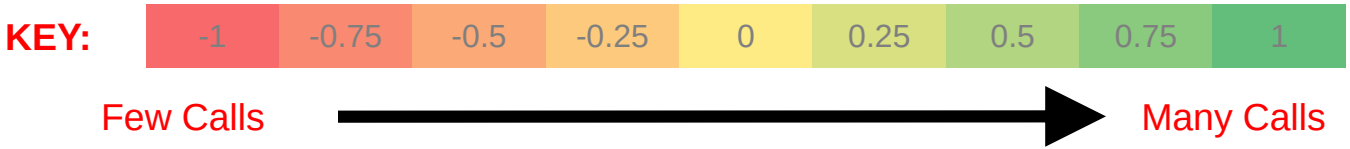
	(1)	(2)
	Log(Total Volume of Calls)	
Log(DSD population)	1.029***	0.962***
	(0.00760)	(0.00646)
Log(Distance between DSDs)	-0.945***	-1.169***
	(0.00852)	(0.00703)
District fraction with internet access		0.0538***
		(0.000436)
Observations	106,276	106,276
R-squared	0.423	0.535

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Total log, normalized volume of calls between Provinces, that remains unexplained after accounting for population and distance

- Once population and distance are accounted for, the North (and East) emerge as top callers.
- Sabaragamuwa and Uva are the main laggards.

	N	NW	NC	E	Uva	Central	Saba	W	S
N	4.39	1.92	1.84	2.89	2.42	2.19	1.57	2.48	2.54
NW	1.92	0.45	0.04	0.18	-0.61	-0.78	-1.26	0.03	-0.17
NC	1.84	0.04	0.60	0.40	-0.20	0.05	-0.72	0.66	0.37
E	2.88	0.18	0.40	1.95	0.22	0.07	-0.51	0.93	0.35
Uva	2.42	-0.61	-0.20	0.22	0.73	-0.69	-0.83	0.56	0.09
Central	2.19	-0.78	0.058	0.07	-0.69	0.0	-1.50	0.05	-0.54
Saba	1.58	-1.26	-0.72	-0.51	-0.83	-1.50	-0.64	-0.49	-1.03
W	2.48	0.03	0.66	0.93	0.56	0.05	-0.49	0.17	0.59
S	2.54	-0.17	0.37	0.35	0.09	-0.54	-1.03	0.59	0.77



ADDRESSING CHALLENGES

Addressing challenges

Challenge	Solution(s)
Negotiating access to data	• Win-win; insights/ techniques for public policy outputs can be leveraged for operator's business interests • Pro-active action by operator(s) rather than reactive to growing government interest in using such data
Minimizing harms from data sharing	• Development of self-regulatory guidelines for operators
Skills	• Assemble interdisciplinary teams that are superior to what consultants can offer
Research ➔ policy	• Policy enlightenment as step 1

Mobile-phone records would help combat the Ebola epidemic. But getting to look at them has proved hard (The Economist, Oct 25, 2014)

- "Releasing the data, though, is not just a matter for firms, since people's privacy is involved. It requires government action as well. **Regulators in each affected country would have to order operators to make their records accessible to selected researchers**, who would have to sign legal agreements specifying how the data may be used. Technically, this is fairly straightforward: **the standards are well established, as are examples of legal terms**. Orange, a big mobile operator, has made millions of CDRs from Senegal and Ivory Coast available for research use for years, under its Data for Development initiative. Rather, the political will to do this among regulators and operators in the region seems to be lacking."

Dissemination to mobile operators & governments

- Operators from Bangladesh, India, Pakistan, Sri Lanka, 8 August 2014
- Bhutan: operator & regulator, October 2014
- India: Story in *Hindu Businessline*, 3 October 2014
- Myanmar: Ministry of Communication & Information Technology & Ooredoo in October 2014
  - Also at seminar convened by UN Habitat
- UN Global Pulse & Govt of Indonesia, November 2014
- ITU Telecom World panel, 9 December 2014
- Bangladesh & Bhutan operators, December 2014
  - Also Bhutan Bureau of Statistics
- Government of Sri Lanka, Ministry of Urban Development & Department of Census & Statistics, also in January 2015
  - Based on presentations to Secretary Urban Development and DG Census & Statistics in Oct-Nov 2014
- Invited presentation at Sri Lanka Institution of Engineers, 16 January 2015

Work performed by collaborative inter-disciplinary teams

- LIRNE*asia*
  - Sriganesh Lokanathan
  - Kaushalya Madhawa
  - Danaja Maldeniya
  - Prof. Rohan Samarajiva
  - Dedunu Dhananjaya (lost to industry in Nov)
  - Nisansa de Silva (moved on to U of Oregon)
- LIRNE*asia*/ MIT
  - Gabriel Kreindler (Economics)
  - Yuhei Miyauchi (Economics)
- Technical partners:
  - WSO2 (Dr. Srinath Perara)
  - Auton Lab at Carnegie Mellon University

- University of Moratuwa
  - Prof. Amal Kumarage (Transport & Logistics Management)
    - Transport
  - Dr. Amal Shehan Perera (Computer Science & Engineering)
    - Data Mining
  - Undergraduates working on projects
- Other US Universities
  - Prof. Joshua Blumenstock (U Washington, School of Information)
    - Data Science
  - Saad Gulzar (NYU Poli Sci)
    - Political Science