# Big Data to Improve Urban Planning

Rohan Samarajiva, Sriganesh Lokanathan, Kaushalya Madhawa, Gabriel Kreindler, Danaja Maldeniya

Data analytics is a frontier field where the tools and techniques are still being developed. Expertise, a critical input, is in short supply, the other being access to data. Even so, Colombo-based LIRNE*asia* has demonstrated the value of mobile network big data for urban planning in Sri Lanka's capital city. Pseudonymised, historical call detail records from multiple mobile operators have been analysed to understand and monitor land use, congregations of people, peak and off-peak travel patterns, communities, and traffic.

Rohan Samarajiva (*rohan@lirneasia.net*), Sriganesh Lokanathan, Kaushalya Madhawa, Gabriel Kreindler and Danaja Maldeniya are with LIRNE*asia*, Colombo, a regional information and communication technology policy and regulation think-tank.

"Smart city" is now a buzzword in South Asia, thanks to the Narendra Modi government's commitment to build 100 smart cities in India. In a "smart city," information and communication technologies (ICTs) are used to enhance feedback loops within the complex system of systems that constitutes the city. In the not-smart city, government and other decision-makers act without adequate and timely feedback. Surveys are the principal source of systematic data, but are expensive and cumbersome. They are thus rarely used. Experimentation is not a viable option. In a smart city, feedback is enormous. If processed properly, the data can improve the functioning of the city, both in terms of the design of its "hard" physical environment and in terms of the "soft" services it provides to citizens.

At one extreme of smart-city initiatives lies the vision of a centrally coordinated city resting on pervasive use of specialised sensors (for example, one under each parking space; multiple sensors at intersections), real-time or non-real-time analysis of the resultant big-data flows, and reliance on mathematical models. South Korea's Songdo is the exemplar. Reports of plans for green-field developments indicate that the Modi government is leaning towards this vision (*Times of India* 2015). At the other end of the continuum lies the "crowdsourced" smart city, where some aspects of the workings of the city are sought to be transformed by apps developed at "hackathons" by outside, and mostly volunteer, coders. Usually, the "problems" that are solved are defined by the data sets that are made available for the competitions.

Both approaches have weaknesses. Central coordination is very expensive and hard to get right. When mistakes are made, they are not easy to correct. Crowdsourcing is attractive because of its promise of unleashing decentralised innovation and because it is cheap. But, unless it is highly structured, both with regard to the data sets that are provided and in terms of implementing the solutions that are developed, transformational results are difficult to achieve (Townsend 2013).

There is a middle path that positions citizens, who are the ultimate beneficiaries of urban development, as primary sensors. Instead of seeing the city as a clockwork machine that can be perfectly controlled, this approach recognises the inherent complexity of the system and supports incremental changes, following the Deng Xiaoping dictum of "crossing the river by feeling the stones." Experimentation and learning are integral to the approach. This low-cost approach is especially appropriate for the organically developed, congested cities in developing countries, where the costs of installing and maintaining city-owned sensors would be quite high.

## Big Data

"Big data" is an all-encompassing term for any collection of data that is very large or complex, and therefore difficult to analyse using conventional data-processing applications. The challenges include analysis, capture, curation, search, sharing, storage, transfer, visualisation, and possible privacy issues. Examples are the 100 million or so call detail records (CDRs) generated in one day by the Sri Lankan mobile operators and the 10 terabytes of data generated by the Hubble Space Telescope each year.

Big data have always been there, but it is only recently that analysis has become tractable. Over the past few decades, more data have been "datafied." Mayer-Schonberger and Cukier (2013: 78–86) coined this neologism to describe very large sets that may include, but are not limited to, schema-less (unstructured, but processable) data. The data have to be in a digital format that allows tabulation and analysis. Until recently, constraints of computer memory, retrieval, and processing limited the use of these data to entities such as the United States National Security Agency and American Express, which could afford to use supercomputers (Samarajiva 1996). Hardware and memory have declined in price and improved in functionality, democratising the use of big data. In addition to proprietary and expensive software, open-source tools such as Hadoop have further broadened the range of those who can undertake big-data analytics.

The principal constraints at this time are access to data and skilled data scientists. Most big-data sets are controlled by organisations. Organisations in competitive markets use them internally, but are reluctant to give access to outsiders for several reasons—they do not want to lose competitive advantage and forgo opportunities to monetise the data; they fear public-relations disasters if personally identifiable information leaks out; or they do not wish to incur the transaction costs of pseudonymising the data and managing non-disclosure agreements. Data in the possession of governments are being released as part of open-data initiatives in some countries.

With the explosion of interest in big data research, tremendous demand has arisen for computer scientists and statisticians who can analyse data. Universities are yet to gear up for the demand, resulting in scarcity. Good big-data research requires multidisciplinary teams made up of those willing and able to converse across disciplinary silos. Data scientists and domain experts who fit this profile are in short supply.

## What Kind of Big Data for Urban Planning?

Urban planning requires data such as those about land use, about where people live and congregate and when, about their mobility, their economic conditions, where they spend their money, and about their social networks. If one wishes to rely on citizens as sensors, there are two obvious sources. The ubiquitous mobile phones that are carried by almost all citizens, even those in poor cities, are the first. Cards that can be used in public transport are the second.

Smartphones can provide a lot more data than feature phones but cannot be assumed to be ubiquitous in developing economies. Unthinking reliance on smartphones, which would be concentrated among the rich at this stage, may yield biased findings that marginalise the poor. For example, reaching conclusions about the prevalence of potholes on streets based on sensors built into smartphones may result in the allocation of scarce resources to streets that are traversed by the rich and deny them to those used mostly by the poor. What works in Boston does not necessarily work fairly in a developing country.[1]

Mobile network big data (MNBD) are generated by all phones, smart and otherwise. MNBD include CDRs generated when calls and texts are sent/received, the internet is used, and prepaid value is loaded, and visitor-location registry (VLR) data are generated when handsets "tell" base transceiver stations (BTS) that they are in coverage areas. CDRs require some action by the user. VLRs do not. VLR data are generated independently of the user as long as the phone has power. CDRs, which include data elements such as calling-party number, called-party number, the BTS where the call originated, time of call, duration, and information about the device, are used for billing purposes. Therefore, the data is stored for some time. VLR data are larger in volume and tend to be written-over. Other than some data necessary for network management, such as those on load factors, VLR data are not routinely retained. CDR and VLR data can provide insights of value for urban planning.

If a city has a multipurpose stored-value card that is used in public transportation and/or to pay congestion charges, that too would serve as a valuable source of data for urban planning. London's Oyster Card and Hong Kong's Octopus Card are examples. However, such cards are usually not available in the public transport systems in most developing countries. Thus, MNBD are the best available sources.

The most accurate results of relevance to urban planning may be obtained if CDR or VLR data are obtained from all mobile operators. However, even if data from around 50% of mobile users can be obtained, insights of significant value can be generated.

LIRNE*asia*, a regional ICT policy and regulation think-tank, has demonstrated the value of MNBD in Sri Lanka. Pseudonymised, historical CDRs from multiple mobile operators have been analysed to understand and monitor land use, congregations of people, peak and off-peak travel patterns, communities, and traffic. Correlations have been validated using other data sets where available.

LIRNE*asia* has focused on MNBD because they have two characteristics of value in informing public policy in developing countries. First, it is the only "datafied" data set that comes close to comprehensive coverage of the populations of these countries, making it possible to treat citizens and the mobiles they tend to carry as the sensors that yield data as a by-product. Second, people carrying mobiles move across geographical space, producing data on their movement patterns. The findings, some of which are described below, have been shared with senior government officials in urban planning and those responsible for the collection of official statistics and also with professional associations.[2] For illustrative purposes, findings developed from three different kinds of MNBD—CDRs, the diurnal loading patterns of CDRs on BTS, and aggregate calling patterns among BTS, are discussed below.

## Illustrative Findings

### Changes in Population Density

Colombo is a small city of 5,50,000 citizens, according to the 2012 Census. It has lost population since the previous count. CDRs were analysed to measure diurnal changes in population density and gain insights into who commutes into the city and from where.

Using interpolation techniques to compensate for the fact that CDRs are only generated when owners take an action (for example, send/receive a call/text), the location of the phone can be plotted on an hourly basis. This allows diurnal variations in population density to be understood. The population "hot spots" identified using MNBD have been correlated with the findings of a conventional transport survey that cost over $3,00,000. The analyses suggested that certain areas, such as the southern and central parts of the city and a few industrial zones outside the city, serve as sinks, attracting large numbers of people from surrounding suburban sources. The northern part of the city, where the poor are concentrated, also functions as a source, showing lower density at midday on weekdays relative to midnight.

It was possible to identify how many people commuted to Colombo and from where using the following method. Based on the extracted average diurnal mobility pattern for the population, home time was defined as 9 pm to 5 am and work time as 10 am to 3 pm. A home and work location for each SIM was calculated. BTS from each location were matched to divisional secretariat divisions (DSDs; this is an administrative unit in a district). Each DSD was counted, at most, once per day to overcome biases due to calling activity. The DSD with the largest number of "hits" was used to determine home and work locations. It was found that 47% of Colombo's daytime population are commuters, with 10 adjacent DSDs contributing 27%, as shown in Table 1.

**Table 1: Colombo's Commuters and Where They Come From**

| Home DSD | Population | Percentage of Colombo City's Daytime Population | Percentage of Home DSD Population in Colombo City During Daytime |
|---|---|---|---|
| Colombo City (2 DSDs) | 5,55,031 | 53.1 | 86.6 |
| Sri Jayawardenepura Kotte | 1,07,508 | 2.9 | 24.3 |
| Kolonnawa | 1,90,817 | 3.5 | 23.6 |
| Dehiwala | 87,834 | 2.6 | 22.9 |
| Kelaniya | 1,34,693 | 2.1 | 19.7 |
| Ratmalana | 95,162 | 1.9 | 18.6 |
| Maharagama | 1,95,355 | 3.7 | 17.7 |
| Kaduwela | 2,52,057 | 3.3 | 17.0 |
| Wattala | 1,74,336 | 2.5 | 16.1 |
| Moratuwa | 1,67,160 | 1.8 | 14.7 |
| Kesbewa | 2,44,062 | 2.5 | 14.5 |

Source: Census 2011–12 and LIRNEasia research based on MNBD.

A metropolitan area may be defined, according to the Encyclopedia Britannica, as "a major city together with its suburbs and nearby cities, towns, and environs over which the major city exercises a commanding economic and social influence. ... Sometimes there may be two or more major cities, as in the Tokyo–Yokohama Metropolitan Area (Japan) or an agglomeration of metropolitan boroughs as in Greater London (England)."

One way to demonstrate the influence of a major city within a metropolitan region is by ascertaining the percentage of people who travel there daily. As may be seen from Table 1, almost one in four people in Sri Jayawardenepura Kotte, on the eastern border of Colombo, do so. The other DSDs around Colombo are also tightly integrated into the city, with more than one in eight commuting to it daily. This includes areas outside Colombo District, across the Kelani River, which demarcated the northern boundary of the city and the district from colonial times. The total population reaches 20,36,855 if all the DSDs in Table 1 are counted as part of the metropolitan region.

There is justification for coordination of transport, water supply, flood control, and the like at the level of a logically connected metropolitan region rather than at the level of smaller local government entities demarcated by features such as the location of wetlands that no longer exist, as in the case of the city's eastern boundary, or the now well-bridged Kelani River that defines the northern boundary. In 2011, a Cabinet Paper was submitted by the Ministry of Urban Development proposing the creation of a Colombo Metropolitan City Corporation, comprising the local government bodies for Colombo, Dehiwala-Galkissa, Kolonnawa, Kotikawatta, and Sri Jayawardenepura Kotte. These local government bodies cover the Dehiwala, Kolonnawa, Ratmalana, and Sri Jayawardenepura Kotte DSDs, with a total population of 10,36,352.

The rationale for the inclusion of the local bodies named in the 2011 Cabinet Paper is not known. The MNBD-based analysis shown in Table 1 suggests that a larger unit comprising all the DSDs has stronger justification. If the larger unit is selected, the total population of the metropolitan region would reach 20,36,855. This would still be small by Asian standards, but would be large enough to support the efficient provision of services to citizens and small enough for a system of municipal governance that does not distance the citizen too much.

In addition to contributing to an evidence-based demarcation of rational boundaries for urban governance, MNBD can also contribute to the identification of priorities for upgrading public transit. The data on where the largest numbers of commuters come from suggest two priority corridors, instead of the one that has already been identified.

The above analysis was conducted on the basis of CDRs, which is arguably not the best source of data on where a person is at any given time. A CDR is generated when the person makes or receives a call, sends a text, or goes to the

**Figure 1: Base Transceiver Stations Serving Commercial and Residential Areas**

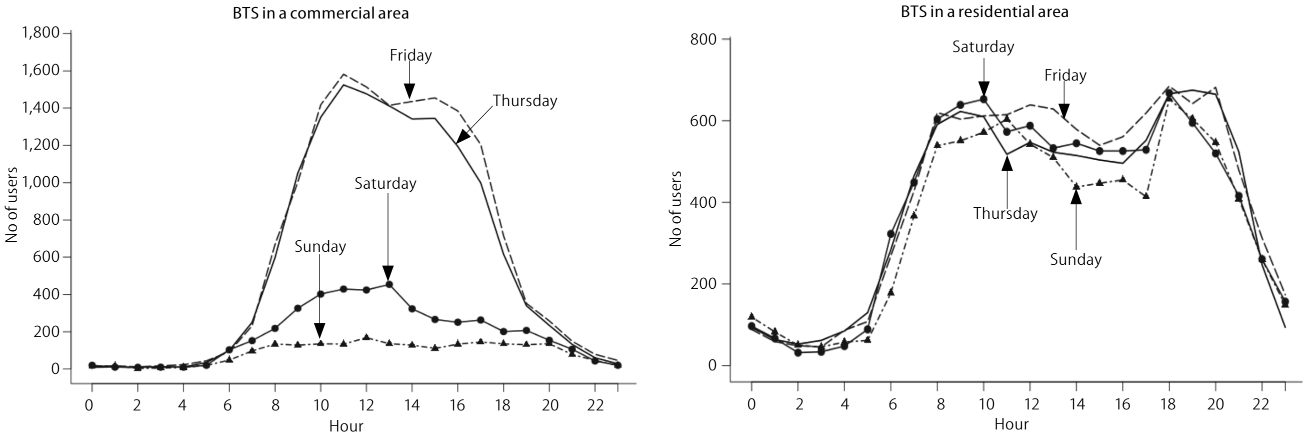
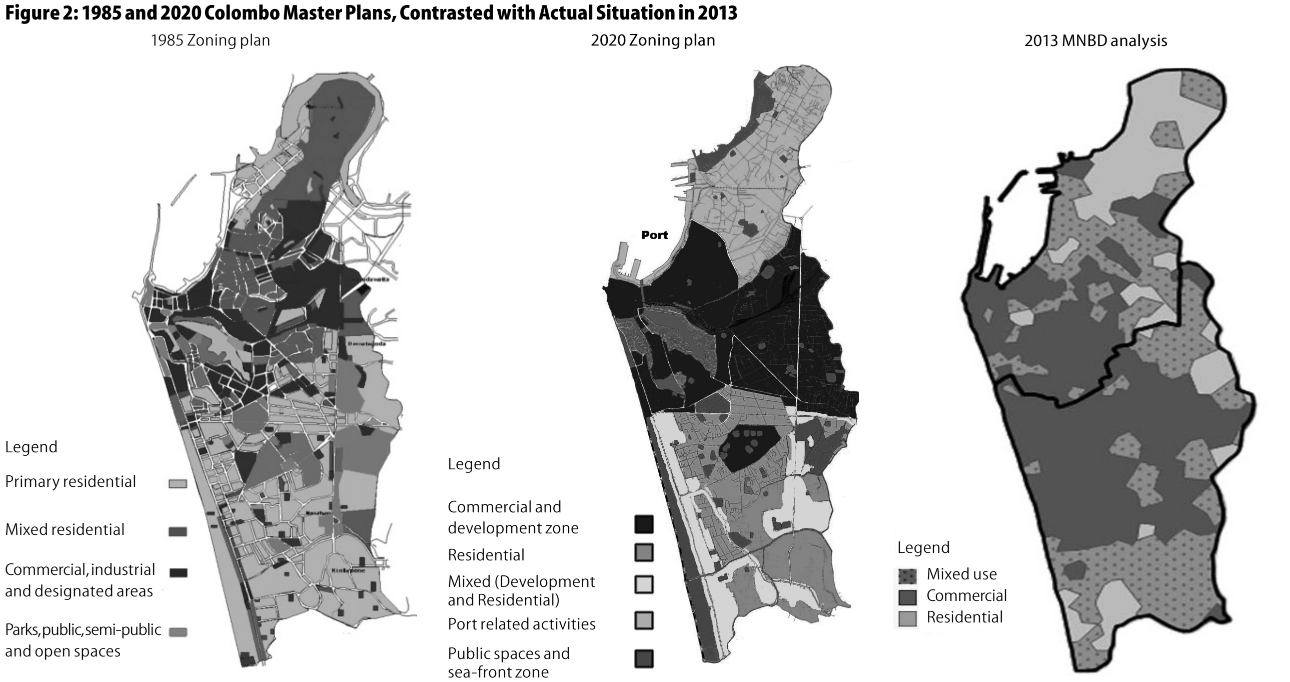



**Figure 2: 1985 and 2020 Colombo Master Plans, Contrasted with Actual Situation in 2013**




internet. These actions do not occur at each and every location a person is at. On the other hand, no action is required for the generation of VLR data. That happens independently of the person's communication behaviour. Not having access to massive VLR data sets, the above reported research combined data over a period of several months, an acceptable second-best solution.

## Changes in Land Use

In the case of generating insights on urban land use, the CDR data was first-best—the diurnal loading patterns on BTS. It was observed that BTS in Colombo District (population 2.34 million; includes most of the Colombo metropolitan area and the city) could be classified into distinct categories by the application of unsupervised machine learning techniques to the diurnal loading data. The two polar cases are shown in Figure 1. The left-hand profile, where the peak use occurs at around midday, along with low loading on the weekends, is from a commercial area. The right-hand profile is from a BTS in a residential area. Here, the peak occurs at around 7 pm and there is no significant difference between weekdays and weekends.

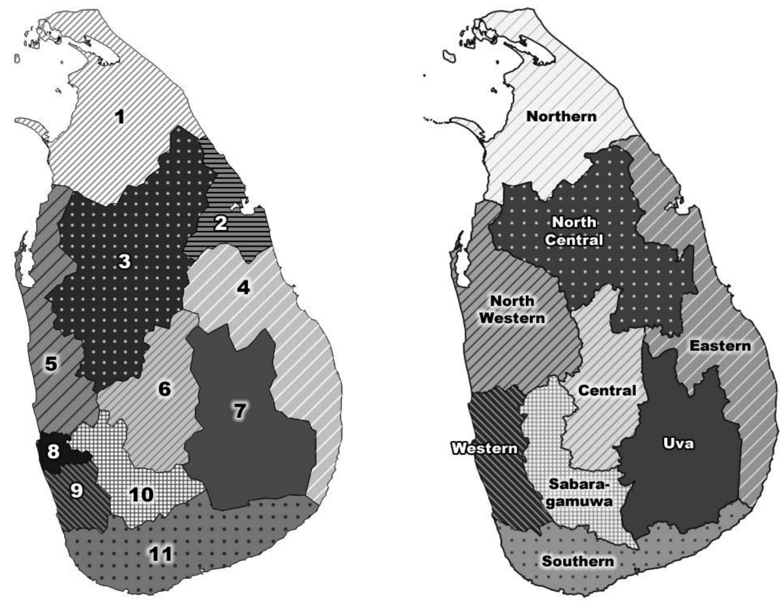
Principal Component Analysis (PCA) was used to identify discriminant patterns in each BTS's loading pattern to remove the effects of noise.[3] Fifteen principal components that collectively accounted for 95% of the variation were chosen. Using an unsupervised machine learning technique called K-means clustering, the 15 principal components of each BTS were used to effectively group all the BTSs into three categories (that is, K was set to 3). The three categories reflect three types of land use—predominately commercial, predominantly residential, and an intermediate category of mixed use. It is possible to further disaggregate the intermediate locales to show which way they "lean," that is, "leaning commercial" or "leaning residential."

The analysis costs very little and can be done at frequent intervals, unlike industry surveys that are conducted by national statistical organisations every three or four years.

However, the MNBD-based analysis does not provide information on what exactly the commercial use is, though further analysis may show correlations between certain profiles and certain kinds of uses such as those characteristic of evening entertainment zones. But this will still not provide all the required information, especially in mixed-use areas. Therefore, the MNBD analysis complements conventional methods and also allows the identification of areas worthy of spot checks. It provides the ability to monitor change in land use almost in real time and therefore contributes to aligning master plans to reality and permits taking prompt action to prevent deviation from zoning rules. Figure 2 (p 45) shows how much actual use has deviated from the master plans for Colombo. The dark shades in each of the plans denote commercial areas. The darker zones in the map generated by MNBD analysis, also denoting commercial areas, shows that commercial use has expanded way beyond what was envisaged by the planners. The resulting conversion of residential premises to commercial establishments also explains the city's loss of population, as recorded in the 2011–12 Census.

**Figure 3: Sri Lanka's Communities in Relation to Administrative Boundaries**

(a) Eleven communities derived from MNBD analysis (b) Sri Lanka provincial map




## Identifying Communities

MNBD data provide measures of social networks, with geospatial features. It permits the use of community-detection algorithms to partition space (in this case, the whole of Sri Lanka) into distinct communities giving weight to intra-community communication as opposed to intercommunity communication. This can be done using different techniques, one being a modularity-optimisation technique. The modularity score Q for each potential community = (edges inside the community) – (expected number of edges inside the community) (Newman and Girvan 2004: 1–16, 204).

The purpose of the modularity optimisation technique is to identify sub-communities, such that there is a globally optimal solution that maximises each community's modularity score. It is possible to apply this technique at various scales, that is, a city, a region, or the country as a whole. Depending on the adopted scale, the identified communities may be different. In the research, it was applied to Sri Lanka as a whole and the communities that were revealed are shown in Figure 3.

The country-level analysis shows that the historically defined boundaries of three of the provinces most distant from the capital Colombo—the Northern, Southern, and Uva Provinces—hold true even according to community definitions emerging from MNBD analyses (communities 1, 11, and 7 respectively in Figure 3a). The Eastern Province, also remote (communities 2 and part of 4 in Figure 3a), comprising people of Tamil, Muslim, and Sinhala ethnic identities in roughly equal proportions, is actually two communities, a small one around the port city of Trincomalee (community 2 in Figure 3a) and a larger one (community 4 in Figure 3a) that includes the eastern part of the predominantly Sinhala North Central Province. This is an intriguing result, given the now-defeated Liberation Tigers of Tamil Eelam's (LTTE) claims of an ethnic "homeland" encompassing the Northern and Eastern Provinces and other proposals to merge the Eastern and Northern Provinces, even under a unified Sri Lanka. While two of the communities (communities 1 and 2 in Figure 3a) in the Northern and Eastern Provinces may be seen being constituted primarily, but not exclusively, by persons of Tamil ethnicity, the third (community 4) is comprised of a large number of Muslims and a significant number of Sinhala people as well. One hypothesis is that this reflects that these areas are among the most productive rice-growing areas in the country and that economic commonalties may be washing out ethnic factors.

Table 2 shows the ethnic composition of the nine provinces and the country as a whole. Here, the two distinct census categories of Sri Lanka Tamils and Indian Tamils (those who were brought to Sri Lanka during the colonial period) have been combined.

**Table 2: Percentage Distribution of Population by Ethnicity across Provinces**

| | Sinhala | Tamil | Muslim | Other |
|---|---|---|---|---|
| Western Province | 84.2 | 6.8 | 7.9 | 1.2 |
| Central Province | 66.0 | 23.8 | 9.9 | 0.3 |
| Southern Province | 95.0 | 1.7 | 2.9 | 0.4 |
| Northern Province | 3.0 | 93.8 | 3.1 | 0.1 |
| Eastern Province | 23.2 | 39.2 | 36.9 | 0.7 |
| NW Province | 85.7 | 3.0 | 11.0 | 0.3 |
| N Central Province | 90.9 | 1.0 | 8.0 | 0.1 |
| Uva Province | 80.8 | 14.7 | 4.3 | 0.3 |
| Sabaragamuwa Province | 86.4 | 9.2 | 4.3 | 0.1 |
| Sri Lanka | 74.9 | 15.3 | 9.2 | 0.5 |

It is possible to apply community detection techniques to a city or any sized unit. In the research, it was applied to Sri Lanka as a whole. Here, the data are first-best—

how one BTS is connected to another—based on aggregate calling patterns.

An analysis of communities within cities may provide useful insights for service provision, the demarcation of electoral districts, municipal wards, and so on. Public policy may use community insights to either reinforce communities, or perhaps build bridges between them. The commuting patterns described above may be usefully supplemented by insights from community analyses.

## Modalities of Conducting MNBD Research

### Gaining Access to Data

MNBD are of significant value to mobile network operators, for better running their operations, and for realising new value. For example, they may be used to develop location-based advertising. Access to one operator's big data by another could have serious implications for competitive advantage. Therefore, there is great reluctance to share data outside the firm, even if for a public purpose. Comprehensive coverage (N = all) is ideal for the kinds of analysis needed for urban planning. This would require data from multiple operators.

This poses a conundrum. There is less chance of competitively sensitive information leaking out or being inferred from published findings if data from multiple companies are aggregated, or, in the jargon of computer science, "mashed up." So it would seem that operators should be amenable to releasing data that will be mashed up before analysis. However, operators may also have concerns about what happens before the mashing up occurs—whether company-specific data will be leaked out from the research organisation before aggregation. The operators are wary about their data, even anonymised, getting into the hands of rivals.

LIRNE*asia* obtained the data on the basis of extensive non-disclosure agreements that included substantial financial penalties. Every researcher who worked with the data was bound by non-disclosure agreements that were explained and re-explained to them. In addition, data access was carefully controlled on a need-to-know basis. Data was pseudonymised within the operator's facilities and no keys were shared. Yet, there is no guarantee that the access provided to LIRNE*asia*, an organisation that has engaged with mobile operators for more than a decade, will be extended to other organisations. In light of the oft-made claims of confidential information provided by operators to regulatory authorities leaking to competitors, it is unlikely that government organisations will fare any better.

Of course, there is always the option of mandating data sharing, exemplified by the following recommendation in an editorial in the normally laissez-faire *Economist* in the context of the Ebola epidemic in West Africa.

> 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 (emphasis added; 25 October 2014).

It is not quite accurate to claim that the technical and legal standards are well established or that Orange provided complete data sets. And no data were released despite the publication's urgings. The guidelines developed by GSMA (the global lobbying group for mobile operators) during this period recommended in-house analysis rather than data sharing (GSMA 2014). But the issue is important. In the face of emergencies of various forms such as the Ebola epidemic, governments are likely to follow the advice of the *Economist*. This would not be the most productive outcome, since it is unlikely that government functionaries will be the best placed to engage in data analytics research, or that government organisations are likely to be very good at safeguarding personally identifiable information and competitively sensitive data. It appears that the *Economist* assumes qualified researchers will analyse the data, not the government itself. But this would involve the government picking research organisations to conduct the analysis, an activity that most governments are not necessarily good at.

In this context, there is merit in companies pre-empting government mandates by voluntarily sharing data on terms and conditions that are well thought-out and designed to both safeguard their interests, the customers who co-created the data, and the researchers who want to extract useful insights for urban planning. This would include the adoption of, and adherence to, self-regulatory codes of conduct.[4]

The above discussion focused on MNBD, the only data set that can at this time provide comprehensive insights for urban planning and other public purposes that do not marginalise the poor. This does not suggest that other datasets, such as those from Twitter, should not be used. They may be used for specific purposes, with care. In the future, it is likely that MNBD will be displaced by other data sets and the mobile operators will no longer enjoy the pivotal position they now do.

### Possible Organisational Forms

Data analytics is a frontier field where the tools and techniques are still being developed. Expertise is in short supply and is one critical input, the other being access to data.

Big-data research conducted for public purposes is unlikely to be able to offer compensation packages that can match the private sector. In any case, universities are not yet producing graduates with the full skill set. Therefore, this work is likely to be done by fresh graduates who will be attracted by the opportunity to learn on the job. The best positioned to attract bright young graduates who can learn on the job are likely to be well-organised research centres located in or outside universities.

While a university research centre may have the advantages of easy access to fresh graduates and brand recognition, the university setting is less than optimal in terms of nimble operation and ability to provide adequate compensation packages and a good working environment. Here, research centres with established brands (even if not in the big data space) such as LIRNE*asia* have an advantage. The challenge for either kind of research centre is ensuring adequate capitalisation. While the actual hardware and software costs are not high, the research centre has to be able to provide a good working environment and have adequate budgets for tasks such as marketing, negotiation of access to data, and ensuring that security is maintained at high levels. It will be difficult to have adequate capitalisation in a non-profit setting.

Will a for-profit research firm focus solely on findings with private benefits and neglect research with a public purpose? Not necessarily, especially if it has its beginnings in a non-profit context and has leaders who are committed to public-interest research. But it may face difficulties in obtaining grant funding if it is not a non-profit.

If capital is required, what should be the source? As stated, data from all mobile operators is ideal for urban research. Data is a critical input. Will it be easier to obtain access to MNBD if all mobile operators have shares in the company? Decision-making in joint ventures can be difficult, but having more than one operator on board may have the benefit of reducing the hesitation of competitors to contribute data. Sri Lanka has shown that it is possible for multiple operators to participate in a mobile-money enterprise.[5] Here too, sensitive customer data moves through the joint venture. This leaves room for hope. However, the Sri Lanka collaboration is a unique one. All over the world, mobile operators use mobile money and similar new businesses as instruments to hold their customers or gain new ones.

However, if the operators do not agree to a joint venture, investment by one operator becomes an option to consider. In this, a number of questions need to be considered. If only one operator invests, will it be possible to obtain data from the other operators? Will the benefits of simple decision-making be washed out by perceptions that the firm is a creature of one mobile operator? Could a trusted third party (for example, a credible public-interest research organisation) play a useful role in blunting such negative perceptions?

Another possibility is an investor outside the telecom industry, partnering with a research firm or centre. This may assuage concerns about one operator gaining access to competitively sensitive information. But the value of MNBD extends beyond the industry. For example, the data would be of significant interest to any firm with even a potential interest in entering e commerce or to a bank. But it is an option worth exploring like all of the above.

Given the significant public purposes that would be served by big data analytics research conducted in the public interest, there may be a good case to be made for impact investing in such a firm. Operational costs and a modest return on equity could be covered by revenues from commercial and public-interest work, which would be undertaken. The possibility of the investor-partner gaining undue access to sensitive data is less likely to influence data release decisions in this scenario.

NOTES

1 See "Street Bump: Help Improve Your Streets" on Boston's mobile app to collect data on road conditions, http://www.cityofboston.gov/DoIT/apps/streetbump.asp

2 See, for example, this report on a public lecture organised by the Institution of Engineers Sri Lanka, http://lirneasia.net/2015/01/lirneasias-big-data-research-presented-at-a-public-lecture-in-sri-lanka/

3 PCA is the basis of multivariate data analyses and was first formulated at the turn of the 20th century (Pearson 1901: 559–72). It has relevance for multiple disciplines and domains. This cross-applicability is a unique characteristic of many of the techniques now deployed in the emerging science of cities that often have their origins in other domains. For example, the well-known gravity model has its origins in physics.

4 An example of the early-stage development of a self-regulatory code may be found at http://lirneasia.net/2014/08/what-does-big-data-say-about-sri-lanka/

5 See "Sri Lanka's Mobile Money Collaboration Recognized at MWC 2015," http://lirneasia.net/2015/03/sri-lankas-mobile-money-collaboration-recognized-at-mwc-2015/

REFERENCES

GSMA (2014): "GSMA Guidelines on Use of Mobile Data for Responding to Ebola," October, http://www.gsma.com/mobilefordevelopment/wp-content/uploads/2014/11/GSMA-Guidelines-on-protecting-privacy-in-the-use-of-mobile-phone-data-for-responding-to-the-Ebola-outbreak-_October-2014.pdf

Mayer-Schonberger, V and K Cukier (2013): *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, London: John Murray.

Newman, M E J and M Girvan (2004): "Finding and Evaluating Community Structure in Networks," *Physical Review E, American Physical Society*, Vol 69, No 2.

Pearson, K (1901): "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine*, (6) 2, pp 559–72.

Samarajiva, R (1996): "Surveillance by Design: Public Networks and the Control of Consumption" in R Mansell and R Silverstone (eds), *Communication by Design: The Politics of Information and Communication Technologies*, Oxford: Oxford University Press, pp 129–56.

*Times of India* (2015): "Chandigarh May Become India's First Smart City," 12 January, http://timesofindia.indiatimes.com/india/Chandigarh-may-become-Indias-first-smart-city/articleshow/45857738.cms

Townsend, Anthony M (2013): *Smart Cities: Big Data, Civic Hackers and the Quest for a New Utopia*, New York: WW Norton.