Using mobile network big data for land use classification

CPR*south* 2015

Kaushalya Madhawa, Sriganesh Lokanathan,
Danaja Maldeniya, Rohan Samarajiva

July 2015

LIRNE*asia*
info@lirneasia.net | www.lirneasia.net

LIRNE*asia* is a pro-poor, pro-market think tank whose mission is *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.*

Contact: 12 Balcombe Place, Colombo 00800, Sri Lanka. +94 11 267 1160.
info@lirneasia.net
www.lirneasia.net

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

Abstract

Understanding and monitoring land use characteristics is critical for urban planning. Unfortunately the traditional way of generating insights on land use involve surveys and censuses, which are both infrequent as well as costly. In developed economies that have greater levels of datafication, as well as use of remote sensors, the dependence on traditional methods is waning. However developing economies continue to depend on these traditional methods. Mobile phone use however is nearly ubiquitous even in developing economies. This enables the population to in-effect act as sensors of human activity. This paper explores the potential of leveraging massive amounts of human mobile phone usage data (i.e. mobile phone big data), to understand the spatiotemporal activity of the masses, and by extension provide a useful proxy for activity-based classification of land use. Using unsupervised clustering techniques on the mobile phone activity signatures of the population aggregated at the base station level, the paper shows the feasibility of inferring at the very least three distinct land use characteristics: commercial/ economic, residential, and mixed-use.

Keywords: Mobile Network Big Data, Call Detail Records, Urban Computing, Land Use Characteristics

1 Introduction

The emergent science of cities, in some ways evokes the tension of the great debates between the late Robert Moses, the 'master builder' from New York, and the late Jane Jacobs, the archetypical purveyor of the importance of complex socio-cultural inter-play between residents, space and other actors in the make-up and evolution of cities.[1] Greater opportunities for measurement to both describe as well as understand the growth of cities do not necessarily resolve the tension between the need for planning and organic development (Bettencourt, 2014). However by empirically appreciating the self-organization behavior of cities and utilizing technology to facilitate greater information flows for coordination, convergence may be possible with these divergent worldviews.

This complex interplay has relevance to post-conflict Sri Lanka. Colombo is the largest city in Sri Lanka by population and economic activity. With the post-war infrastructure boom, land use patterns in Colombo are changing quickly, with residential population decreasing, making way for commercial activities. The city population has decreased from 647,100 in 2001 to 561,314 in 2012 according to census figures, a loss of almost 14%.[2] Having an up-to-

[1] For a more thorough treatment of two world views as symbolized by the work of Jane Jacobs and Robert Moses refer to Flint (2011)

[2] Base census figures were obtained from the website of the Department of Census and Statistics Sri Lanka at http://www.statistics.gov.lk/page.asp?page=Population%20and%20Housing

date overview of land use characteristics is critical in such context. In seeking to understand the form and shape of the urban space, we still continue to rely on traditional land use and land cover classifications, measured through infrequent survey and census-based methods. Extant methods differentiate land use based on one or more dimensions. These dimensions can be based on inter alia the physical characteristics of the land, ownership characteristics, and quite importantly in the case of urban space, the type of activity undertaken on the land (Anderson, 1976).

The ubiquitous use of mobile phones in turn has enabled the population to in-effect act as sensors of human activity. This paper investigates the potential of massive amounts of human mobile phone usage data (i.e. mobile phone big data) to be leveraged to understand the spatiotemporal activity of the masses, and by extension provide a useful proxy for activity-based classification of land use.

2 Literature Review

Near ubiquitous mobile phone access and use in Sri Lanka affords possibilities to leverage aggregate calling data from mobile phones to understand broader human behavior. Currently Sri Lanka has a mobile penetration of over 100% and coverage of nearly 100% of the landmass (Telecom Regulatory Commission of Sri Lanka, 2014).

But assuming that every Sri Lankan owns at least one mobile device is an overestimation of the actual number of unique users and GSM Association (2014) lists the unique number of users as 50% of the total population. Lokanathan et al (2014) highlighted the non-homogeneous mobile penetration in the country by comparing the distribution of home locations of mobile users detected by their method against the resident population from census. They found a much higher correlation between the number of home locations and census population in Western Province than the correlation obtained for the entire country (R-squared values 0.8 to 0.59). Hence we can treat the behavior of observed SIMs as a closer approximation to the behavior of the entire population.

Existing research has utilized aggregate calling data from other countries to provide a greater understanding of urban environments. The real-time Rome project attempted using aggregate call volumes at base stations (in this case Erlang data[3]) to understand the land use characteristics in the city of Rome (Reades, Calabrese, & Ratti, 2009). Although the authors found a relationship between calling activity and the land use patterns at certain locations, they were not successful in segmenting the city into specific land use categories. This was due to the fact that Rome does not have clearly delineated zones for different types of urban activity making classification that much harder.

Automated identification of land use categories using calling data aggregated at the level of the Base Transceiver Station (BTS[4]) have been done in several countries. These approaches

[3] Erlang is a unit of traffic density and used to understand the load in a telecommunications system. One Erlang could be equivalent to one person talking for 60minutes or 2 people talking for 30minutes each, etc. Erlang data is used to understand the load on a base station at any given time

[4] A Base Transceiver Station (BTS) is the commonly utilized term for a base station. It consists of one ore more antennas and has a certain geographical coverage area.

can be divided into three potential categories depending on the type of techniques utilized: unsupervised, semi-supervised, or supervised.

Unsupervised techniques first cluster space according to various attributes related to calling behavior at the BTS level (average number of users connected over a certain time period, actual and normalized diurnal patterns of call volumes/ and or users connected to a base station, etc.). They then ex-poste qualitatively compare the clustered spaces with known locations of specific land-use categories. Depending on the specific technique the ways of qualitatively describing the derived geo-spatial cluster differs. For example Soto & Frías-Martínez (2011a) used k-means clustering to assign geo-spatial regions to specific existing land use categories in Madrid. Soto & Frías-Martínez (2011b) on the other hand utilized fuzzy c-means clustering techniques (again using data from Madrid) to assign each derived geo-spatial region probabilities of belonging to one or more existing land use categories. The latter approach seems more appropriate in regions where areas are not clearly delineated into different land use categories.

Supervised learning techniques a priori use existing land use information on known locations to train models that are then capable of predicting the land use categories of regions without additional information about those regions. Toole, Ulm, Bauer, & González (2012) utilized zoning data and mobile phone activity data to predict land use categories for the metro Boston region using such unsupervised techniques.

A third technique is semi-structured learning. Here a priori knowledge of land use categories of a small but known number of points of interests is used to calibrate the algorithms. Pei et al (2014) used a semi-supervised fuzzy c-means clustering approach to infer the land use types in Singapore.

The accuracy of supervised classification methods rely heavily on the quality of the external land use classification data that is utilized. Toole et al. (2012) for example found that their supervised learning approach failed to classify most of the areas accurately since the underlying zoning data used to train the model was often incorrect. Another issue is that greater heterogeneity in land use decreases the overall reliability of supervised and semi-supervised techniques. For example Pei et al. (2014) found that the detection rate of their approach was only 58.03%.

A common step found in all these methods is representing a location (i.e. a BTS cell) by the aggregated activity pattern of that region. The goal of the eigen-decomposition method proposed by Reades et al. (2009) is to identify and extract recurring patterns of mobile activity. Even though their method is capable of removing noisy fluctuations in the diurnal patterns at a location, it has not been used in other unsupervised clustering methods mentioned above. In this paper we have leveraged their method to extract the recurring patterns. We then used this insight to assign a land use category to each BTS using an unsupervised algorithm. We are restricted by the lack of good zoning data and/or other validation data. As such we are unable to attempt supervised and/or semi-supervised techniques. Our work goes further than previous unsupervised approaches, since we also quantify the extent of commercialization.

3 Data Source

The paper uses one month of Call Detail Records (CDRs) for nearly 10 million SIMs from multiple operators in Sri Lanka[5]. The data is completely pseudonymized by the operator i.e. the phone numbers have been replaced by a unique computer generated identifier. The researchers do not maintain any mapping information between the generated identifier and the original phone number.

Each CDR corresponds to a particular subscriber of the operator's network and is created every time a subscriber originates or receives a call. In the case of an in-network call (i.e. both parties on the call were subscribers in the same mobile network), two records are generated, one for each party. Each record contains the following attributes:

- Call direction: A code to denote if the record is an incoming or outgoing call
- Subscriber identifier: Anonymized identifier of subscriber in question
- Identifier of the other party: Anonymized identifier of the other party on the call
- Cell identifier: an ID of the cell (i.e. antenna) that the subscriber was connected to at the time of the call
- Date and time that the call was initiated
- Duration of the call

4 Research Methodology

Our approach to inferring land use categories from mobile network big data, followed a three-step process. We first constructed activity signatures for each BTS. We then broke down the activity signatures of each BTS into its principal components. Finally we classified each BTS into one of three potential categories by leveraging the principal components. As such we assign the BTS' derived classification to its coverage area. In order to determine the coverage area of the BTS we used Voronoi tessellation, whereby we partition the geospatial area of Sri Lanka into non-overlapping polygons, where each polygon's center of gravity is its associated BTS. Urban regions by virtue of having a greater density of BTS, have voronoi cells of much smaller area (on average about 0.25km^2), whilst in remote and rural areas, the voronoi cells may be a several km^2.

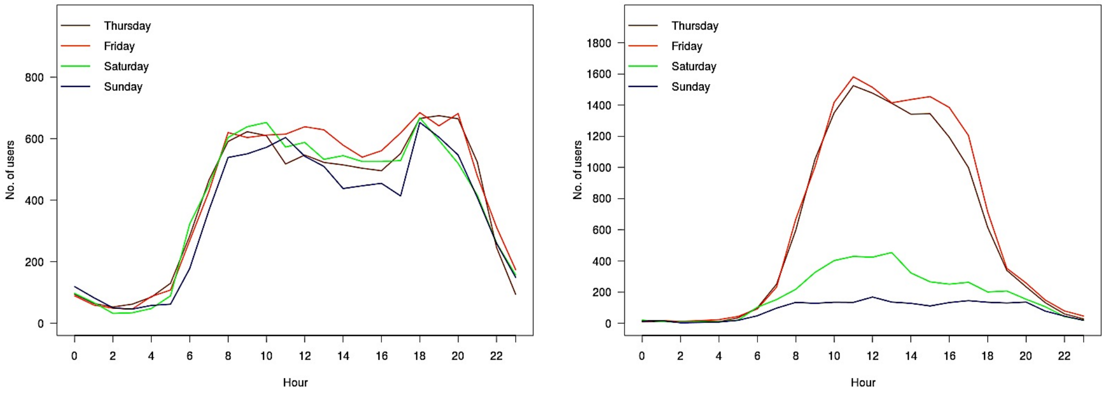
4.1 Construction of BTS activity signatures

We constructed a temporal variable for each BTS to capture the average diurnal pattern of human activity for a specific temporal frame. To do this we considered the number of users utilizing the BTS for making or receiving a call at any given moment. By aggregating the number of users at hourly intervals, we ended up with a temporal variable containing 720 points representing the aggregate number of users connected every hour for a period of 30 days (720 = 30 days x 24 hours). If a user had more than one activity during an hour, we counted her at most only once.

Whilst the distributions for each BTS varied, two prominent patterns emerged as symbolized by the two graphs show in Figure 1. BTS 1 in Figure 1, shows minimal variation in activity signatures between weekdays and weekends and the daily peak occurs in the evening. BTS

[5] Due to the agreements with the operator we are unable to name the operators and cannot give a precise figure for the number of SIMs that were analyzed.

2, on the other hand shows significant difference in the signature between weekday and weekends and the daily peak occurs in the morning. It was hypothesized that BTS 1 and BTS 2 exhibit behavior that is consistent with what would be expected of a residential and commercial area respectively. A visual inspection of the actual areas covered by these BTSs confirmed the hypothesis.



(a) BTS 1 **(b) BTS 2**

Figure 1: The hourly user volume at two selected BTSs

We found that public holidays distort the average activity pattern that we wished to construct. Therefore all data points pertaining to public holidays were removed. We then average the remaining values by hour and day of the week to remove the effect of variations associated with specific dates. This resulted in a temporal variable consisting of 168 (7 days x 24 hours) data points, which represented the average activity at a BTS on an hourly basis for every day of an average week.

Given that population density varies from place to place, and also given that our primary interest was in leveraging the broader pattern, we normalized the frequency distributions for each BTS. Thus we brought the activity signature of each BTS to a uniform scale by normalizing the frequency distribution for each BTS by its calculated z-score. Thus the effects of any deviations associated with specific dates were reduced.

4.2 Extraction of discriminant features

We then applied Principal Component Analysis (PCA)[6] on the covariance matrix of the weekly time series signatures of all BTSs that were constructed in the previous step. This results in 168 components (can also be considered as the discriminant features or eigenvectors) that can collectively be used to represent the entire data set. Each BTS activity signature is therefore a sum of these 168 components, with BTS-specific coefficients for each of the components. Thus the activity signature of any specific BTS i can be represented as a linear summation of its principal components in the following manner:

$$Activity\ signature(BTS_i) = \sum_{n=1}^{168} C_{ni} V_n$$

Where V_n represents the n^{th} eigenvector, and C_{ni} represents the coefficient associated with the n^{th} eigenvector for BTS i.

[6] Principal Component Analysis (PCA) is a mathematical technique used in many scientific fields to extract discriminant features from a large set of features.

The discriminant features for each BTS were then ordered based on how much of the variance of all the activity signatures they captured. They are ordered from highest to lowest. We then selected only the 15 principal components that can collectively represent 95% of the variation in the activity. This allowed us to remove random noise that may be associated with a BTS.

4.3 Grouping base stations based on time series signatures

Finally we utilized an unsupervised K-means clustering algorithm to classify BTSs into one of three clusters (i.e. K was set as 3) based solely on the average activity signature of each as represented by its 15 principal components. We chose the number of clusters to be three, as to represent three potential land use categories: commercial, residential, and mixed use. K-means algorithm was chosen since it doesn't require any a priori knowledge/assumptions on the relationship between usage characteristics and land use categories.

5 Results and discussion

The analysis was restricted to the greater Colombo region, for two reasons. Firstly the land demarcations were more ambiguous outside of the greater Colombo region. More importantly the density of base stations is much lower outside of the major urban areas, which means the coverage area for each BTS in a rural area is much larger than that of a BTS in an urban area with higher population density. This results in reducing the efficacy of the classification algorithm since it now relates to an average activity signature of a larger area.

The three clusters that were discovered using the k-means algorithm had average activity patterns as shown in Figure 2.

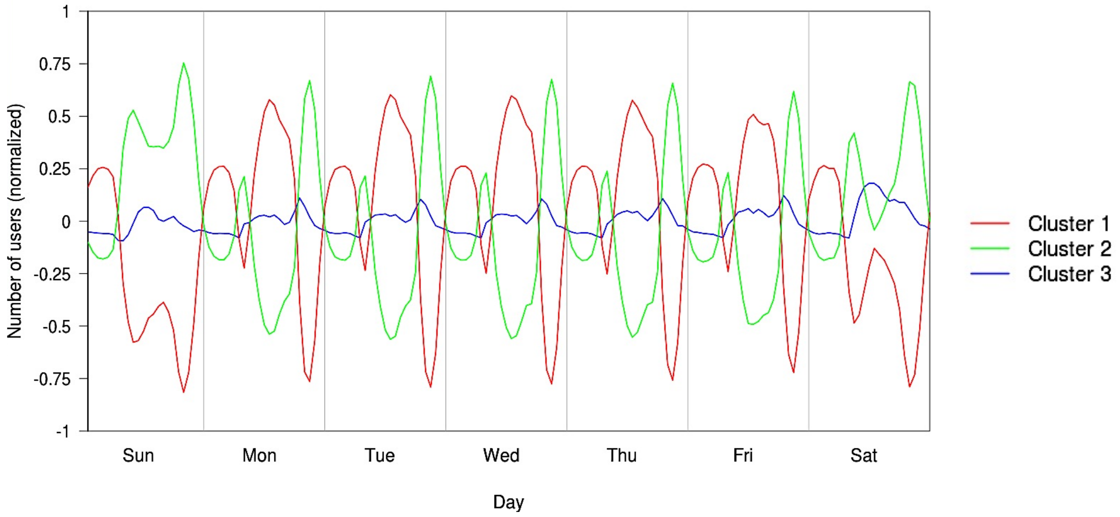


Figure 2:Average activity signatures for each of the 3 derived clusters

Cluster 1: The average activity signature of this cluster exhibits greater activity during weekdays, especially during the daytime. Furthermore the number of users during the weekend is very low as compared to weekdays. During weekdays, the peaks occur during the 1000 to 1200 time frame and the lowest number of users is detected during the late evening. These are all behavioral characteristics one would expect of working individuals and we hypothesized that this cluster contained base stations located in areas that are highly commercial and/or industrial.

Cluster 2: The signature of this cluster shows similar magnitude during weekends as well as weekdays. The number of users peaks during 1900 to 2100 in the late evening and drops during the daytime. This pattern implies more residential areas, with the signature characteristics of residential behavior where people leave to work in the morning and return in the evening. Further more the user count on Sunday is clearly higher than on any other day of the week further reinforcing the inference that Cluster 2 areas are highly residential.

Cluster 3: Unlike Cluster 1 and 2, the activity signature for Cluster 3 doesn't show as marked a difference between weekday and weekend or between morning and evening. There is an early evening peak during the weekdays but it is not as prominent as the daily peaks for Cluster 1 and 2. Similarly weekends and especially Sundays show a slight difference in the overall activity pattern. Based soley on the overall pattern it is difficult to make an inference, but given that this activity signature lies between those for Cluster 1 and 2, it was hypothesized that the base stations belonging to this cluster were in mixed-use areas.

The hypothesized inferences of the type of areas represented by the three clusters were verified through a visual inspection of the areas. Figure 3 shows the geographical distribution of the clusters found in the Colombo district. Neighboring voronoi cells are merged if they share the same land use category. The boundaries of the two constituent Divisional Secretariat Divisions (DSD) that make up the Colombo City municipality are marked in black.

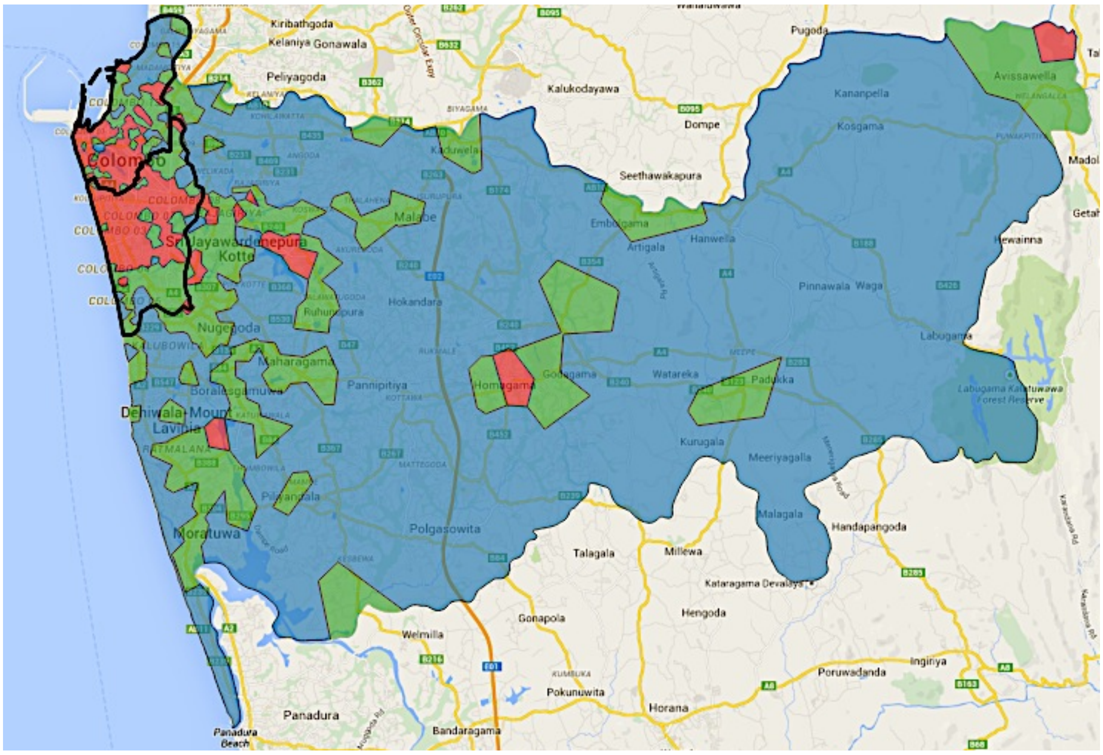


Figure 3: Spatial distribution of land use categories in Colombo district. The black border indicates the extent of Colombo city

As can be seen from Figure 3, the residential areas (i.e. Cluster 2) are the largest in terms of their geographic coverage. Commercial areas (i.e. Cluster 1) have the lowest geographic coverage. However when just consider the number of base stations that belong to each of the two clusters, they are roughly the same. This is because there is a higher concentration of base stations in the denser urban areas (thus resulting in a lower geographic footprint for each base station). Less commercial and sparsely populated regions on the other hand have a lower base station density.

The majority of the area within the city of Colombo (the area under the black border in Figure 3) is mainly commercial. It is also evident from the figure that the northern part of the city has been labeled as mostly residential (also see Figure 4). North Colombo is considered the inner city, and is a much poorer neighborhood with less development than the rest of the city. The population density in that area is much higher with the majority of the citizenry from that area employed in surrounding regions.

Looking more closely at Colombo city (Figure 4) we can see how the areas labeled as commercial reflects ground realities and in fact shows how the Central Business District (CBD) has expanded. Historically the Fort and Pettah region (see Figure 4) was the city's main CBD area. During the civil war it was also the target of much terrorism due to it being near the port area as well as it housing many government buildings. As a result the Navam Mawatha area (see Figure 4) near Beira lake, which has housed the Ceylon Chamber of Commerce for the past 150 years, became the location of choice for businesses from around the mid-1980s onwards. During the last decade the CBD has grown further south as is visible in Figure 4. This meshes with the data from the Department of Census and Statistics Sri Lanka that showed the city's residential population to have decreased by almost 14% from 2001 to 2012.

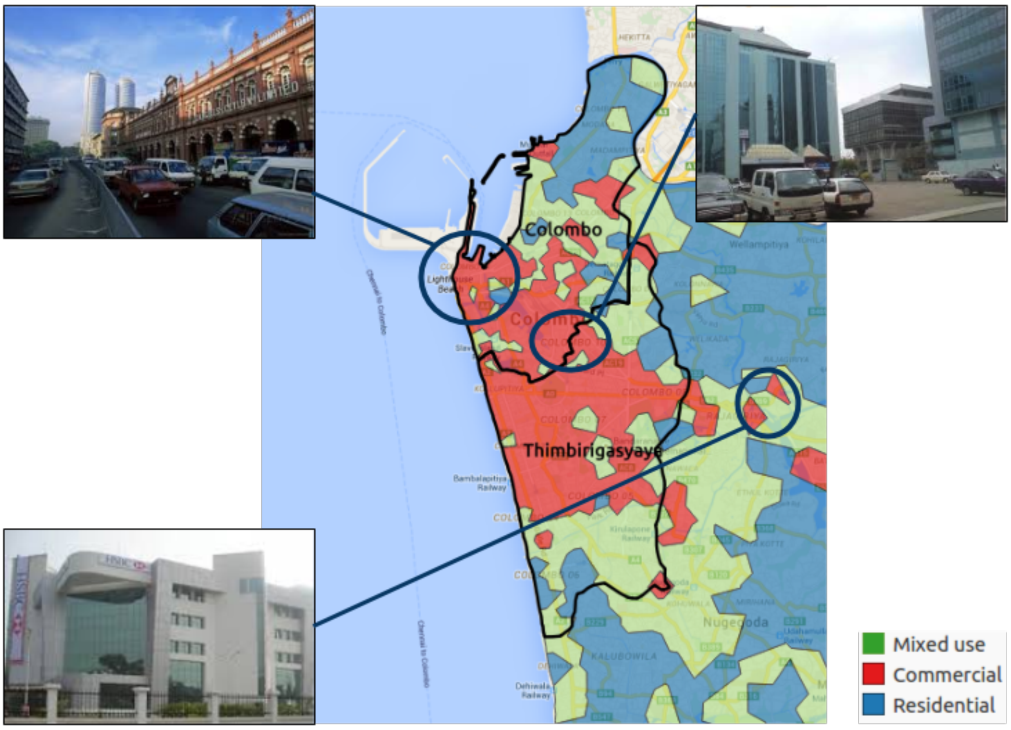


Figure 4: Expansion of Colombo city's Central Business District (CBD)

What was surprising in Figure 3 was that the northeastern corner of Colombo district was labeled as commercial. Further investigation revealed that the area was home to the

Seethawaka Export Processing Zone (EPZ), a special purpose area developed by the government to attract industrial and manufacturing activities targeting the export sector. What is also interesting is that the area surrounding the Seetawaka EPZ has been classified as mixed-use. This characteristic is borne out in the rest of the district as well, with areas classified as commercial regions always having mixed-use regions surrounding it. This potentially indicates the spread of commercialization, but requires further investigation.

In order to understand the extent of commercialization in the mixed-use regions, we decided to leverage the silhouette coefficient for each of the BTSs after classification. A silhouette coefficient is a measure used to determine the quality of clustering achieved (Rousseeuw, 1987). The average silhouette coefficient BTSs in each of the 3 clusters is given in Table 1.

Table 1: Average silhouette coefficient for the BTSs in each of the 3 clusters

Cluster	Average Silhouette coefficient
1- Commercial	0.46
2- Residential	0.36
3- Mixed-use	0.22

The mixed-use cluster shows the least coherence, which is understandable since its temporal signature exhibits characteristics of both of the other two clusters. To quantify the closeness of a mixed-use region to the commercial temporal signature we constructed a measure C to capture the extent of commercialization. C is defined as follows:

$$C(BTS_x) = \frac{Distance\ to\ commercial\ signature\ (BTS_x)}{Distance\ to\ residential\ signature\ (BTS_x)}$$

Where BTS_x is the BTS that is being investigated.

Figure 5 maps each of the corresponding voronoi cells for those BTSs classified as mixed use. The C value for each cell is represented as a circle color from red (high values) to blue (low values). The redder the circle the greater its extent of commercialization.

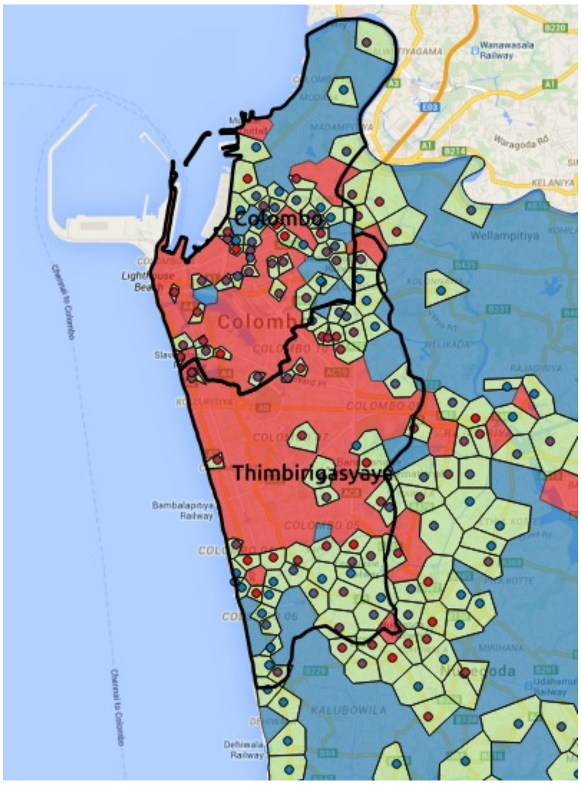


Figure 5: How close are the mixed-use regions to other two land use categories

6 Conclusion and future work

In this paper we explored how mobile network big data can be leveraged to understand land use characteristics in the city of Colombo. Using a month of CDR data we constructed activity signatures for each BTS. We grouped these activity signatures, resulting in three clusters of base stations. Since this grouping was done in an unsupervised manner, we labeled these clusters as residential, commercial and mixed used based on the average activity signature of each cluster. We could observe that the land use predicted by our method corresponded to the actual land use characteristics based on local knowledge on of known locations in Colombo city.

The proposed method is inexpensive as compared to the current survey/ census based methods, and furthermore can be done in near-real time, so as to have a high frequency understanding of the evolution of space. The latter is achievable through the measure we constructed to understand the extent of commercialization.

However this work is still preliminary. A major shortcoming of this research is the lack of a systematic way to validate the findings. We had to rely on visual inspection of known locations with tacit knowledge of land use. Even this issue inhibits further improvements to the mathematical model as it’s harder to quantify the effect of such changes.

In future work, we intend to solve this problem by finding various data sets which could represent the actual land use distribution in Colombo. In the above section we witnessed how zoning plans could be unreliable for some regions as the intended land use diverts from how the land is being used. Currently the land use data constructed from the countrywide survey conducted in 2012 is only available at district level. As our method has done land use classification at the granularity of BTS cells, we need survey data at a comparable scale. With access to recent land use survey data at a much granular spatial level, we can better validate our current findings as well as improve the methodology to detect more fine grained land use categories such as nightlife and recreational parks.

In the meantime, we are exploring what alternative data sources can be leveraged to represent the actual land use patterns. Location sharing websites such as Foursquare has emerged as a new source to understand urban land use (Zhan, X., Ukkusuri, S. V., & Zhu, F., 2014). Foursquare possesses information on thousands of locations with their associated categories. Though it contains much more information compared to traditional data sources, these data can be biased by the user choices. As an example people tend to check-in more often from restaurants compared to a hospital or a school. We will investigate the feasibility of using Foursquare data to validate our land use classification in Colombo district.

7 References

Anderson, J. R. (1976). *A land use and land cover classification system for use with remote sensor data* (Vol. 964). US Government Printing Office.

Bettencourt, L. M. a. (2014). *The Uses of Big Data in Cities*. Big Data, 2(1), 12–22. doi:10.1089/big.2013.0042

Clawson, M., & Stewart, C. L. (1965). *Land use information. A critical survey of US statistics including possibilities for greater uniformity.*

Flint, A. (2011). *Wrestling with Moses: how Jane Jacobs took on New York's master builder and transformed the American city*. Random House Trade Paperbacks.

GSM Association (2014), *The Mobile Economy – Asia Pacific 2014*. Retrieved 8th July. Available at http://asiapacific.gsmamobileeconomy.com/GSMA_ME_APAC_2014.pdf

Lokanathan, S., Silva, N. de, Kreindler, G., Miyauchi, Y., & Dhananjaya, D. (2014). Using Mobile Network Big Data for Informing Transportation and Urban Planning in Colombo.

Pei, T., Sobolevsky, S., Ratti, C., Shaw, S.-L., Li, T., & Zhou, C. (2014). A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*, (August), 1–20. doi:10.1080/13658816.2014.913794

Reades, J., Calabrese, F., & Ratti, C. (2009). Eigenplaces: Analysing cities using the space–time structure of the mobile phone network. *Environment and Planning B: Planning and Design, 36*(5), 824-836. doi: 10.1068/b34133t

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53-65.

Soto, V., & Frías-Martínez, E. (2011a). Automated land use identification using cell-phone records. *Proceedings of the 3rd ACM International Workshop on MobiArch - HotPlanet '11*, 17. doi:10.1145/2000172.2000179

Soto, V., & Frias-Martinez, E. (2011b). Robust land use characterization of urban landscapes using cell phone data. *Proceedings of the 1st Workshop on Pervasive Urban Applications, in conjunction with 9th Int. Conf. Pervasive Computing.*

Telecom Regulatory Commission of Sri Lanka. (2014). Statistical Overview of Telecom Sector. Retrieved July 6th, 2015. Available at http://www.trc.gov.lk/2014-05-13-03-56-46/statistics.html

Toole, J. L., Ulm, M., González, M. C., & Bauer, D. (2012). Inferring land use from mobile phone activity. *Proceedings of the ACM SIGKDD International Workshop on Urban Computing - UrbComp '12*, 1. doi:10.1145/2346496.2346498

World Bank. (2012). *Turning Sri Lanka's Urban Vision into Policy and Action*. Colombo, Sri Lanka. Retrieved from https://openknowledge.worldbank.org/handle/10986/11929

Zhan, X., Ukkusuri, S. V., & Zhu, F. (2014). *Inferring Urban Land Use Using Large-Scale Social Media Check-in Data.Networks and Spatial Economics*, *14*(3-4), 647-667.