# Land use classification using call detail records

Kaushalya Madhawa[1], Sriganesh Lokanathan[1], Danaja Maldeniya[1], Rohan Samarajiva[1]

[1]LIRNE*asia*, 12 Balcombe Place, Colombo 8, 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 continuously decreasing making way for commercial activities. Having an up-to-date overview of land use characteristics is critical in for urban planning, and traditional survey and census-based methods in addition to being expensive cannot give high-frequency insights. Mobile network big data could be utilized as a low-cost and high frequency alternative for understanding changes in land use. Our method potentially goes further than previous efforts, by allowing for a more fine-grained understanding of land use changes.**

Whenever a mobile phone is used to make or receive a call or SMS, or to access the Internet, the event is captured via meta-data stored in the operator's logs. This is called a Call Detail Record (CDR). The CDR also contains the location of the Base Transceiver Station (BTS) that services the subscriber, as well as timestamp for the event. Using these records it is possible to get a time-series of activity at each BTS. Building on prior work [1, 2, 3, 4], this usage pattern of BTS activity can be leveraged to understand the characteristics of the land beneath it.

For each BTS, we utilize a month of CDR data to construct a time-series measure of the number of users connected to the BTS at any given time. We then project this data into a 7-day week, excluding data from public holidays since human dynamics change dramatically during holidays. For each BTS we then end up with a time-series measure for each day of the week, where the diurnal activity pattern for that day is an average of all the same weekdays in that month.

Projecting the user data of a month to a week results in a unique 'signature' of mobile usage at each site as seen in Figure 1. A vector of 168 elements represents the time-series signature of a BTS. This signature provides a good measure of the variation in people's behaviors between days and even between the hours of each day.
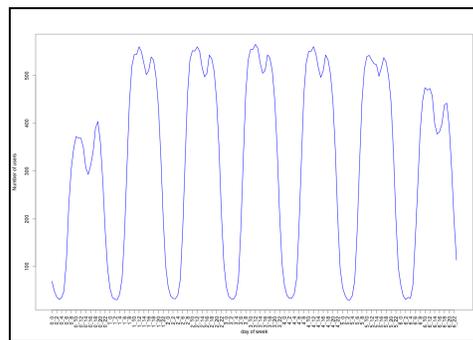


Figure 1: **Hourly user distribution for a week fro a sample BTS**

The actual number of users connected to a BTS at any given time depends on the area it covers and the place it is located at. To remove these biases we normalized the hourly user distribution by taking the z-score. The diurnal signature of each BTS also contains occasional noise that could affect the land use classification. Using methods articulated in [1] to deal with noise in Wi-Fi signals, we remove the random variations in our normalized diurnal data by decomposing the covariance matrix of user time series data into a combination of its principal components.

Each time series vector for a BTS can be represented by a linear combination of its principal eigenvectors as,

$$(1) \qquad T_i = C_{i1}V_1 + C_{i2}V_2 + \cdots + C_{in}V_n$$

Here $T_i$ represents the time series of a BTS *i*, and $C_{i1}$ represents the correlation of the first eigenvector $V_1$ and so on.

The first eigenvector accounts for the general diurnal pattern of the city. The first fifteen eigenvectors of the covariance matrix is capable of representing 95% of the variation in the time series data.

Then the time series distributions are redistributed using only the first fifteen eigenvectors.

## Land use classification

The time series of each BTS is then assigned to one of three clusters using a k-means algorithm considering each reconstructed time series as a 168-dimensional vector.

We use the signature of cluster centroids to infer the actual land use category that each cluster belongs to. Whilst prior work [4] had identified more than three clusters, we find that interpretation of the these additional categories is problematic for our data, especially when we validate using known land use characteristics in Colombo.

The centroid pattern of Cluster-3 is consistent with the dynamics of a commercial region, since more users are detected during weekdays than weekends and the daily peak occurs during the daytime. Similarly the centroid pattern of Cluster-2 is consistent with the dynamics of residential regions, with weekday/ weekend variations being minimal and the daily peak occurring during the late evening each day. Cluster-2 is classified as belonging to areas of mixed-use land characteristics, since there are no discernable variations in the time of the daily peaks or between weekdays and weekends.
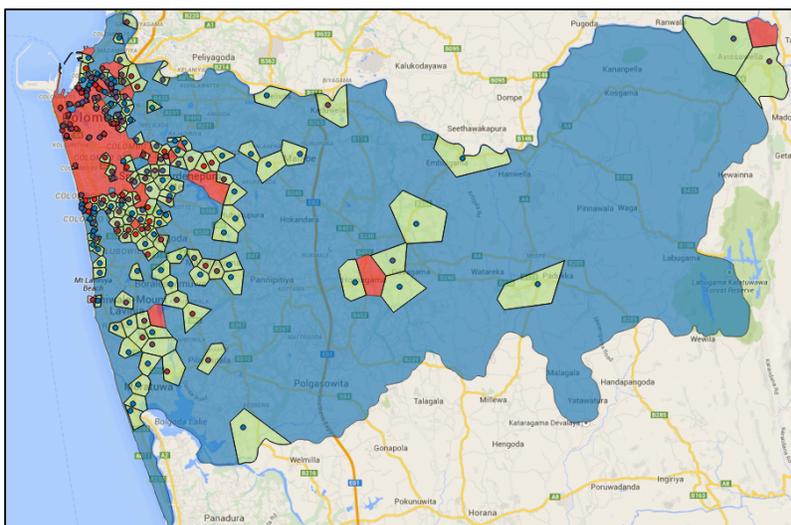


Figure 2: **Distribution of clusters in Colombo district.** Cluster-3 (Commercial) areas are in red, Cluster-2 (residential) areas are in blue, and Cluster-1 (mixed-use) areas are in green. Areas with blue dots signify a BTS signature closer to Cluster-2 than Cluster-3. Areas with red dots signify a BTS signature closer to Cluster-3 than Cluster-2

The spatial distribution of clusters is shown in Figure 2.

Cluster-1 also exhibits the least average silhouette value as compared to the other two clusters. We then calculate the distance from each point in Cluster-1 to the centroids of other two clusters. The ratio of the distance from the centroid of a Cluster-1 area to the centroids Cluster 2, and Cluster 3, is used to understand how the mixed-use regions compare with other two categories.

An analyses of these ratios for each Cluster-1 area (i.e. mixed use), suggests that mix-used regions neighboring a commercial area exhibit

a pattern more similar to the commercial pattern than a residential pattern. The inverse seems to hold true for mixed-use areas adjacent to mainly residential areas.

## Future Work

We plan to use this technique repetitively over a longer time period to investigate if the classifications of the silhouettes (blue and red dots in Figure 2) show changes. This might reveal a high-frequency measure of the changes in land use than is possible using just the three-cluster classification.

## References

[1]      F. Calabrese, J. Reades, and C. Ratti, "Eigenplaces : Segmenting Space Through Digital Eigenplaces : Segmenting Space through Digital Signatures," 2014.

[2]      J. Reades, F. Calabrese, and C. Ratti, "Eigenplaces: analysing cities using the space – time structure of the mobile phone network," *Environment and Planning B: Planning and Design, 36:824–836, 2009*.

[3]      G. Andrienko, N. Andrienko, M. Mladenov, M. Mock, and C. Politz, "Discovering bits of place histories from people's activity traces," in *2010 IEEE Symposium on Visual Analytics Science and Technology*, 2010, pp. 59–66.

[4]      V. Soto and E. Frias-Martinez, "Robust land use characterization of urban landscapes using cell phone data," *In Proceedings of the 1st Workshop on Pervasive Urban Applications (Purba'11)*, 2011.