Estimating activity-based land-use through unsupervised learning from mobile phone event series in emerging economies

Gregor Engelmann^{*1}, James Goulding^{†1} and David Golightly^{‡2}

¹N-Lab, University of Nottingham ²Human Factors Research Group, University of Nottingham

January 13, 2017

Summary

Understanding and monitoring the makeup of a city is critical for effective urban planning. Traditional ways of describing land-use involve the use of surveys, which are costly and quickly outdated. Unparalleled mass urbanisation makes them particularly unsuited in emerging economies, where mobile phones are now nearly ubiquitous. This paper explores the potential of utilising automatically generated Call Detail Records to understand spatial-temporal activity and by extension provide a summary of activity-based land-use. Using non-negative matrix factorisation and unsupervised clustering, the paper shows the feasibility of describing activity-based land-use from CDR data in an exemplar East African city.

KEYWORDS: Activity-based land-use, Call Detail Records, Unsupervised learning, Non-negative matrix factorisation

1 Introduction

The understanding and monitoring of activity-based land-use is intrinsic to successful urban planning, and this is especially true in emerging economies. Land-use has traditionally been understood in terms of: land characteristics; ownership characteristics; and activity undertaken on the land itself (Anderson et al. (1976)). Traditional techniques to monitor land-use have focused on manual surveys or hand-crafted analyses of satellite imagery. These approaches are slow, expensive, infrequent and only provide a snap shot of a particular period in time and space. The ever-increasing ubiquity of mobile phones, its role as a proxy for human movement, and the vast amounts of data generated by such devices provide much potential to supplement or even replace these traditional methods. This is of particular relevance in East Africa where the data gaps being left due to the expense and logistics required to employ sensor technologies might be bridged via the data being

^{*}psxge@nottingham.ac.uk

 $^{^{\}dagger} james.goulding@nottingham.ac.uk$

 $^{^{\}ddagger} davig.golightly@nottingham.ac.uk$

generated by mobile phone usage. Closing such gaps is key in generating effective information to support infrastructure investments and urban planning. In this paper we introduce a framework for identifying *activity-based land-use* via call detail record (CDR) data. The approach applies non-negative matrix factorisation (NMF) to factorise underlying usage trends, prior to clustering regions into land-use categorisations via k-means clustering. Due to its dynamic nature, the use of CDR data to characterise actual behaviour rather than aggregate descriptors, is particular appropriate in emerging economies where mass urbanisation is resulting in rapidly changing environments.

2 Background

While CDR data has been widely used in the examination of epidemiology (e.g. Wesolowski et al. (2012); Brdar et al. (2016)) there have been far fewer studies utilising CDR data for land-use classification. Those that have been undertaken have tended to take an unsupervised approaches to identifying land-use patterns, clustering according to aggregate statistics related to network behaviour such as average number of connected mobile phone subscribers, or call volumes. Once clusters are formed, an archetype can be presented, and categorised spaces then compared to known locations of specific land-use categories for validation. One example is that of Kaushalya Madhawa (2015), who assessed broad land-use categories using CDR data from approximately 10 million mobile phone subscribers from multiple operators in Sri Lanka. A similar approach was taken by Reades *et al.* Reades et al. (2009) who extracted recurring patterns via k-means clustering, subsequently using these patterns to asses the extent of commercialisation in the study region.

In contrast, some researchers have used a semi-structured learning approach, seeding clustering algorithms with a small set of known points of interest to calibrate results. Pei et al. (2014) used semi-structured fuzzy c-means clustering to infer land-use areas in the Singapore metro area. This however, is a challenging approach, and only 58.03% accuracy was achieved, perhaps due to greater entropy across land-use within Singapore. As a consequence, Soto and Frias-Martinez (2011a,b) combined both an unsupervised k-means clustering and semi-structured fuzzy c-means clustering approach in their analysis of Madrid. Supervised approaches to land-use classification are hard to come across, due to the lack of ground truths for anything but broad brush categorisations. Toole et al. (2012) stand out here with their use of Zoning labels combined with CDR data to infer land-use areas for the Boston metropolitan area. Nevertheless, they found even then that outdated zoning data used for building models was the prime reason for the misclassification of the majority of areas.

In East Africa accurate and fine grained land-use ground truths are extremely hard to find. As we are restricted by the lack of accurate zoning for validation improving unsupervised learning approaches is therefore our focus. Our work goes further than previous unsupervised approaches, however, in that we: 1. first convert event series to time series at different levels of temporal granularity to identify outliers with anomalous behaviour that would skew the observed behaviour patterns; and 2. we then also apply non-negative matrix factorisation (NMF) to extract latent features within population behaviour, reflecting daily patterns of life. This allows us not only to generate a vocabulary to describe underlying behaviours (such as nightlife, commuting, or industrial patterns),

but then also allows us to go on to produce interpretable clusters in this compressed behavioural space. The resulting land-use archetypes can be understood and characterised in terms of how much they express each of these building block behaviours.



Figure 1: Extent of Dar es Salaam, Tanzania, overlaid with our gridded areal representation of the city region

3 Methodology

The data used as part of this study covers a total of 433.6 million network events covering calls and SMS for 415k mobile phone subscribers taking place at a full set of cell towers (Base Transceiver Station, BTS) across the Dar es Salaam region of Tanzania (see Figure 1) over a period of 122 days

in the autumn of 2014^1 . A raw CDR record is automatically created for each network event and includes a range of attributes including: *timestamp*; an anonymised mobile phone *subscriber ID*; *call duration*; and a *Base Transceiver Station (BTS) identifier* (indicating the cell tower).

Our approach for estimating land-use patterns took this mass dataset and followed a multi-step process. First an event series of network events was extracted for each tower. These event series were then themselves split into week periods, and binned to produce time series of hourly event counts. This resulted in a dataset of approximately 9,500 weekly time series instances, each reflecting 168 hourly intervals (24×7) . Some BTS showed a uniform distribution of activity during day time, while others showed a higher network activity pattern either at evening times only, or at both morning and evening times. The total amount of network events differed quite significantly across the study area, making direct comparison and clustering difficult. In order to compensate for different network events counts, feature scaling was applied to each time series to standardise them.

NMF was then applied to these event series to identify latent features occurring in weekly usage behaviour. This factorisation process reduced both noise and reduced our feature space from 168 to 7 dimensions, eliminating outliers and addressing the curse of dimensionality. Moreover, manual analysis of the resulting latent features allowed us to interpret subsequent cluster archetypes. We note that other alternatives for factorisation are available such as Principal Component Analysis (PCA). However, due to allowance of negative loadings, the factors resulting from PCA are far harder to interpret. Similarly NMF was preferred to Latent Dirichlet Allocation (LDA) Lee and Seung (1999); Gautam and Shrestha (2010)due to its deterministic nature.

Following construction of our latent feature space, k-means clustering was applied in order to identify k different land-use areas based on our interpretable activity trends.

4 Results

The factors resulting from our analysis reflecting underlying activity patterns are illustrated in Figure 2. These are informative of population behaviours, with common weekly trends being revealed. Component 2, for example, reveals general underpinning network activity patterns (and is very similar to the average weekly time series for all towers) with a gradual increase from 7am until 10am, plateauing out before an early evening spike in network events. Component 3, however, reflects a predominant residential activity pattern, with population leaving an area in the middle of the day, and returning after work. In contrast, Component 5 reflects a workplace behaviour, with high daytime activity, zero nighttime events, and a significant weekend drop. Any particular region may be composed of a combination of land-uses (for example, half residential and half industrial), and so may express each of these building block behaviours to a different amount. Nonetheless we are provided with a vocabulary through which we can discuss those combinations, without need for imagery or demographic data.

 $^{^{1}}$ Due to both individual and commercial privacy, the anonymised data used as part of this study is unfortunately not publicly available, and was provided to us through a partnership with a major private sector network operator in Tanzania



Figure 2: Six trends extracted via NMF from the CDR data. Each describes a different underlying poulation behaviour, which form the building blocks for our activity based land-use clustering approach.



Figure 3: Observed behaviour clusters

With these building blocks in hand and each time series projected into the lower dimensional space they represent, k-means clustering technique was applied. The choice of k remains an arbitrary one dependent on the task in hand. As such the number of clusters n was varied from 2 to 15 clusters. The most interpretable results were obtained at n = 5. The centroids for each of the resulting clusters are illustrated in Figure 3². Annotation of each class was performed by examining the extent it expressed each building block in combination with with external knowledge:



Figure 4: Spatial distribution of activity-based land-use areas in the central Dar Es Salaam region. Yellow is 'Affluent-Commercial', Blue is 'Slum', Green is 'Formal-Night-Active', pink is 'Industrial' and red is 'Residential-Commuting'

- **Cluster 1 Affluent-Commercial:** consistent activity throughout the day (zones that bring people in due to tourism, job opportunities and amenities).
- **Cluster 2 Slum:** characteristic of a poor demographic with lower daytime activity, low morning activity and significant peak in the early evening (perhaps due to lack of mobility).
- **Cluster 3 Residential-Commuting:** this profile expresses a far higher expression of component 3 (the residential activity pattern) than other behaviours, suggesting a commuting pattern.

 $^{^{2}\}mathrm{A}$ lower number of clusters mixed different behaviours, while a larger number of clusters produced some with insufficiently distinct behaviours.

- **Cluster 4 Industrial:** high expression of component 5 (non-residential). Some commuting, but a highly significant lack of mobility activity in the mornings, evenings and weekend.
- **Cluster 5 Formal-Night-Active:** average activity over the course of the day, but with significant spikes in the evening and night.

The map in Figure 4 shows a plot of the spatial distribution of these clusters for the centre region of Dar es Salaam. In order to protect commercial interests of the network operator that provided us with the data, we interpolate BTS catchment areas into a 500m x 500m grid-cell representation.

5 Conclusion

In this paper we have shown how NMF and k-means clustering might be used to identify interpretable activity-based land-use clusters from CDR data. The biggest difficulty lies in identifying outliers that see low levels of network activity over the study period. While not always feasible, as data may have been pre-aggregated, hourly aggregates seem to be the best temporal scale for identifying anomalous behaviour.

6 Biography

Gregor Engelmann is a PhD student in Intelligent Mobility at the Horizon Centre for Doctoral Training at the University of Nottingham carrying out his research with the N-Lab. His work focuses on how CDR data can be used to understand wider urban dynamics interrelated with transport and mobility trends.

Dave Golightly is a senior research fellow at the Human Factors Research Group at the University of Nottingham. His work focuses on the affect of domain knowledge and expertise on how individuals use technology with a specific interest in transport applications.

James Goulding is assistant professor and deputy director of N-LAB, a new centre for International Analytics at Nottingham University Business School, specializing in novel forms of data science time series analysis, summarisation and visualization of mass datasets. His work focuses on how closed source, commercial data sets can be harnessed to promote international development and social good.

References

- Anderson, J. R., Hardy, E. E., Roach, J. T., and Witmer, R. E. (1976). A land use and land cover classification system for use with remote sensor data. Technical Report 964, USGS.
- Brdar, S., Gavric, K., Culibrk, D., and Crnojevic, V. (2016). Unveiling spatial epidemiology of hiv with mobile phone data. *Scientific Reports*.

- Gautam, B. P. and Shrestha, D. (2010). Document clustering through non-negative matrix factorization: a case of hadoop for computational time reduction of large scale documents. In *International MultiConference of Engineers and Computer Scientists IMECS*, volume 1, pages 1–10.
- Kaushalya Madhawa, Sriganesh Lokanathan, D. M. R. S. (2015). Using mobile network big data for land use classification. Technical report, LIRNEasia.
- Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401:788–791.
- Pei, T., Sobolevsky, S., Ratti, C., and Zhou, S. L. S. C. (2014). A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*, 28:1988–2007.
- Reades, J., Calabrese, F., and Ratti, C. (2009). Eigenplaces: Analysing cities using the space-time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, 36:824–836.
- Soto, V. and Frias-Martinez, E. (2011a). Automated land use identification using cell-phone records. In *HotPlanet '11 Proceedings of the 3rd ACM International Workshop on MobiArch*, pages 17–22. ACM.
- Soto, V. and Frias-Martinez, E. (2011b). Robust land use identification using cell-phone records. In Proceedings of the 1st Workshop in Pervasive Urban Applications, in conjunction with 9th Int Conf Pervasive Computing.
- Toole, J. L., Ulm, M., Gonzalez, M. C., and Bauer, D. (2012). Inferring land use from mobile phone activity. In Proceedings of the ACM SIGKDD International Workshop on Urban Computing -URBComp. ACM.
- Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., and Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. *Science*, 338:267–270.