Exploring the Use of Urban Greenspace through Cellular Network Activity

Ramón Cáceres[1], James Rowland[1], Christopher Small[2], and Simon Urbanek[1]

[1] AT&T Labs – Research, Florham Park, NJ, USA
{ramon,jrr,urbanek}@research.att.com

[2] Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, USA
csmall@columbia.edu

Abstract. Knowing when and where people use greenspace is key to our understanding of urban ecology. The number of cellular phones active in a geographic area can serve as a proxy for human density in that area. We are using anonymous records of cellular network activity to study the spatiotemporal patterns of human density in an urban area. This paper presents the vision and some early results of this effort. First, we describe our dataset of six months of activity in the New York metropolitan area. Second, we present a novel technique for estimating network coverage areas. Third, we describe our approach to analyzing changes in activity volumes within those areas. Finally, we present preliminary results regarding changes in human density around Central Park. From winter to summer, we find that density increases in greenspace areas and decreases in residential areas.

1 Introduction

It is generally understood that urban greenspace is important to people. Greenspace provides recreational, health, and aesthetic benefits to citizens, as well as broader ecological benefits to cities. Understanding the interactions between people and greenspace is critical to our understanding of urban ecology.

Central to this understanding is knowing the extent to which people make use of greenspace. For example, knowing the timing, location, and magnitude of human presence in different green spaces is useful for managing these spaces. While the type and location of greenspace in urban areas is well documented, we lack accurate, quantitative measures of when and where people occupy it. Much of our current knowledge is anecdotal or based on limited survey data.

Cellular telephone networks can provide a wealth of objective information about human use of urban greenspace. Census data provide detailed maps of where people sleep, but tell us little about where people are, and are not, during their waking hours. In contrast, mobile phones are carried by a large portion of a city's population and are used throughout the day. A measure of how many phones are active in which geographic areas can thus serve as a proxy for human density in those areas.

Proc. of 2nd Workshop on Pervasive Urban Applications (PURBA), June 2012

In this work, we use anonymous records of cellular network activity to quantify the spatial and temporal patterns of human density within a major US metropolitan area. More specifically, we use counts of voice calls and text messages handled by cellular antennas as a measure of how many people are in the geographic areas covered by those antennas. Because of the close-knit spacing of antennas in urban areas, variations in these counts can shed light on the use of individual green spaces. We aim to characterize how the density of network activity changes over time, and how these density patterns relate to greenspace and microclimate. By aggregating activity into density maps at different times of day, week, and season, we hope to enhance our understanding of when people occupy different types of greenspace.

This paper presents the vision and some early results of this effort. First, we describe our dataset of six months of cellular network activity in the New York metro area. It contains more than 3 billion samples of activity at a 1-minute granularity over 6 months. Second, we present a novel technique for estimating cellular coverage areas. We have extended the previously used tessellation technique of creating one Voronoi region per cellular tower, and instead create one finer-grained region per collection of antennas residing on the same tower and pointing in the same direction. Third, we describe our approach to analyzing changes in activity volumes within those finer-grained regions. We are using Empirical Orthogonal Function analysis to identify spatial and temporal patterns of interest. Finally, we present preliminary results regarding changes in human density in the area around Central Park. From winter to summer, we find that density increases in greenspace areas and decreases in residential areas.

2 Dataset of Cellular Network Activity

We have gathered from a major US communications service provider a dataset of anonymous cellular network activity in the New York metro area. We began by identifying the set of ZIP codes within 50 miles of downtown Manhattan. We then obtained a list of cellular antennas that were active in those ZIP codes during the period of our study. We grouped into a *sector* the set of antennas that reside on the same cellular tower and that point in the same compass direction. We thus created a reference table of sector identifiers, locations, and directions.

For each of those sectors and for each minute of each day, we gathered counts of how many new voice calls and how many text messages were handled by the antennas in that sector. The contents of our data records are as follows:

Sector | Date | Hour | Minute | Voice Calls Started | Text Messages Handled

In subsequent analysis, we sum the number of voice calls and text messages to arrive at a single measure of cellular network activity that we term *call volume*. Similarly, we use the term *call density* to denote call volume per geographic area, and treat call density as a proxy for human density.

Our current dataset spans the six months between February 1 and July 31, 2011. It contains one record per minute for more than 12,000 sectors, yielding

more than 3 billion call-volume samples. We are currently gathering data for a full year, which will allow us to study a fuller range of seasonal and other temporal effects on human use of green spaces.

We have been careful to preserve privacy throughout this work. In particular, this study uses only anonymous and aggregate data. There is no personally identifying information in the data records described above.

3 Tessellation Based on Cellular Sectors

Our cellular network data gives us estimates of human activity levels, but we need a way to assign that activity to geographic areas. Voronoi tessellation has been used to associate spatial regions with cellular towers [3, 11]. Each tower is treated as a point p_i on a plane P, and a Voronoi region R_i is associated with each p_i. Each region consists of all points on the plane such that $R_i = \{x \in P : d(x, p_i) \leq d(x, p_k)\ \forall k\}$ with distance measure d. Euclidian distance is typically used for d, so that each region consists of points closer to the corresponding tower than to any other tower.

Voronoi tessellation has several important advantages: simplicity, manageable computational cost, and ease of interpretation. However, basing the tessellation only on tower locations results in coarse regions, and therefore coarse assignments of activity to geographic areas.

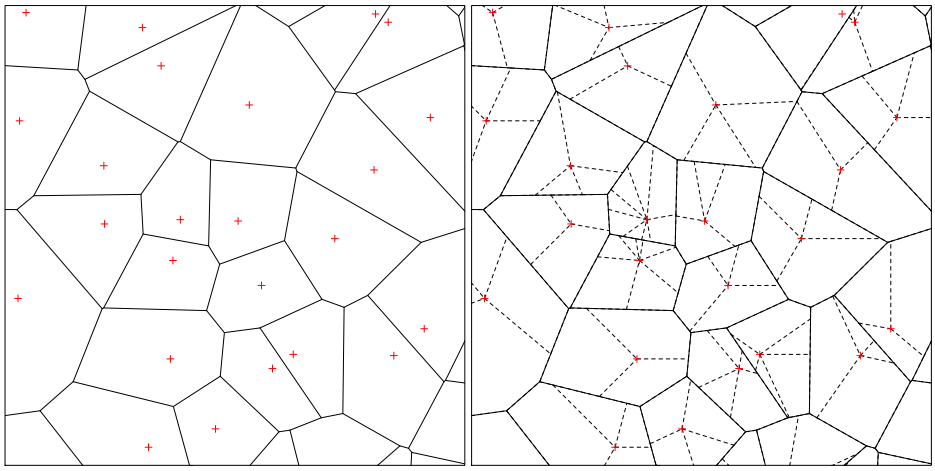
We have developed an algorithm that performs a finer-grained tessellation by making use of antenna directions in addition to tower locations. Typically, each tower holds multiple antennas that serve various technologies and frequencies, and that point in various directions. As we did when collecting the dataset described in Section 2, we group into a *sector* the set of antennas that reside on the same tower and that share the same compass direction, or *azimuth*.

The simplest way to use azimuth information is to further subdivide the Voronoi regions obtained using only tower locations. Conceptually, we can add edges that bisect the angles between sector azimuths. In practice, we obtain the same result by synthesizing virtual positions $v_{i,j}$ for each sector j of tower i as follows: Let p_i be the location of the tower and $\alpha_{i,1..s_i}$ the azimuths of the s_i sectors. Then

$$v_{i,j} = p_i + \varepsilon \begin{bmatrix} \sin \alpha_{i,j} \\ -\cos \alpha_{i,j} \end{bmatrix}$$

with a small $\varepsilon > 0$ and $j \in 1, \ldots, s_i$. A tessellation using the points $v_{i,j}$ will be equivalent to dividing the regions obtained from a tessellation using p_i by bisecting azimuth angles up to the error ε. This equivalence follows from the fact that the points defining the regions induced by p_i can only move at most by the distance ε as p_i is perturbed by that amount. Since the regions do not change by more than ε, the additional points can only partition the regions. Due to all $v_{i,1..s_i}$ being equidistant from p_i, the partitioning has to bisect the angle between two neighboring sectors.

Figure 1 illustrates the result of a regular tessellation based on tower locations p_i (left, locations denoted by red crosses) and our extended tessellation based



(a) Based on tower locations (b) Based on sectors and their directions

Fig. 1. Voronoi tessellations for a sample 60-km^2 area. Red crosses denote cellular tower locations. Taking into account sectors and their directions produces finer-grained estimates of network coverage areas.

on sector azimuths with $\varepsilon = 10^{-5}$ (right, added edges drawn as dashed lines). In both cases we have used a cylindrical projection to maintain near linearity in the tessellated area, with one degree of latitude as the unit. Therefore, the above ε corresponds to approximately 1 cm. The tower locations and sector azimuths are based on the actual configuration of a large US cellular network. The shown area is an excerpt of approximately 60 km^2 from a much larger tessellation.

Making use of the azimuth information clearly improves the granularity of the tessellation. For the complete New York metro area, the median area of a Voronoi region resulting from the extended tessellation is roughly one quarter of the median area resulting from the regular tessellation. At the same time, the number of regions increases by a factor of 3.5.

In ongoing work, we are experimenting with an adaptive version of our algorithm that chooses a different ε per tower to adjust for different distances between towers, e.g., closely packed in urban areas vs. farther apart in suburban areas. We also plan to conduct ground-truth experiments to quantify the accuracy of our estimates of cellular network coverage areas.

In this work, we use our extended tessellation technique to define one Voronoi region for each of the sectors present in the dataset of cellular network activity described in the previous section. We therefore approach the next phase of our study with an estimate of the coverage area of each sector, as well as a measure of call volume per coverage area per minute.

4 Empirical Orthogonal Function Analysis

In this study, we seek to quantify the spatiotemporal patterns of call volumes in order to infer the spatiotemporal distribution of people in the New York metro area. We can then analyse these patterns in the context of the spatial distribution of greenspace and temporal variations in weather conditions. We determine the greenspace distribution using vegetation maps derived from visible and infrared satellite imagery [8, 9]. We capture weather variations using data from a regional network of weather stations.

One objective of this coanalysis is to quantify the relationship(s) between outdoor ambient environmental conditions and the spatiotemporal distribution of people within the urban area. A primary challenge in this analysis is to distinguish between indoor and outdoor activity. A related challenge is to distinguish between regular patterns of activity (e.g. the dominant daily and weekly cycles) and the variations in these patterns that may be related to environmental conditions (e.g. indoors on cold days, outdoors on temperate days).

We will approach both of these challenges by mapping deviations from regular patterns as anomalies in time and space. We will accomplish this mapping using Empirical Orthogonal Function (EOF) analysis, a tool commonly used to quantify spatiotemporal patterns in meteorology and oceanography [12]. EOF analysis is a form of Principal Component (PC) analysis.

We will treat the call volume data as instantaneous spatial snapshots of call volumes, then analyse the spatiotemporal patterns in these time series of call volume maps. Our approach is similar to how PC analysis is used to reduce the dimensionality of multispectral imagery in remote sensing applications (e.g., [2, 4, 7]). Because variables in high-dimensional data are often correlated, PC transforms provide an efficient low-dimensional projection of the uncorrelated components of the data. The same property applies to temporal dimensions.

The utility of the PC transform for representing spatiotemporal processes is related to the fact that, for location x and time t, any location-specific pixel time series P_{xt} contained in an N-image time series can be represented as a combination of temporal patterns and their location-specific components as

$$P_{xt} = \sum_{i=1}^{N} C_{ix} F_{it}$$

where C_{ix} is the spatial Principal Component, F_{it} is the corresponding temporal Empirical Orthogonal Function, and i is the dimension. The EOFs are the eigenvectors of the covariance matrix that represent uncorrelated temporal patterns of variability within the data. The PCs are the corresponding weights that represent the relative contribution of each EOF to the corresponding series P_{xt} at each location x. The relative contribution of each EOF to the total spatiotemporal variance is given by the eigenvalues of the covariance matrix. The distribution of eigenvalues also gives an indication of the dimensionality of the data in terms of uncorrelated modes of variance.

In this study, dimensionality refers to the structure of the spatiotemporal patterns represented in the data—and their relative magnitude compared to the stochastic variance. The implicit assumption is that some number, $D \ll N$, of the low-order EOFs and their corresponding PCs represent deterministic processes, and that the higher-order dimensions represent stochastic variance. This property allows an observed pixel time series to be represented as a sum of deterministic and stochastic components in the following way:

$$P_{xt} = \sum_{i=1}^{D} C_{ix} F_{it} + \epsilon$$

Generally, EOFs are spatial patterns intended to represent spatially continuous modes of variability of physical processes, while the PCs are the weights representing the temporal contribution of the corresponding spatial patterns [5, 12]. In this study, we reverse the convention so that EOFs represent temporal patterns and PCs represent spatial weights. We consider daily, weekly, and seasonal trends that result from deterministic processes such as commuting, as well as higher-frequency day-to-day variability presumably related to ambient environmental conditions and isolated transient events. Additional details of the approach are given by [10].

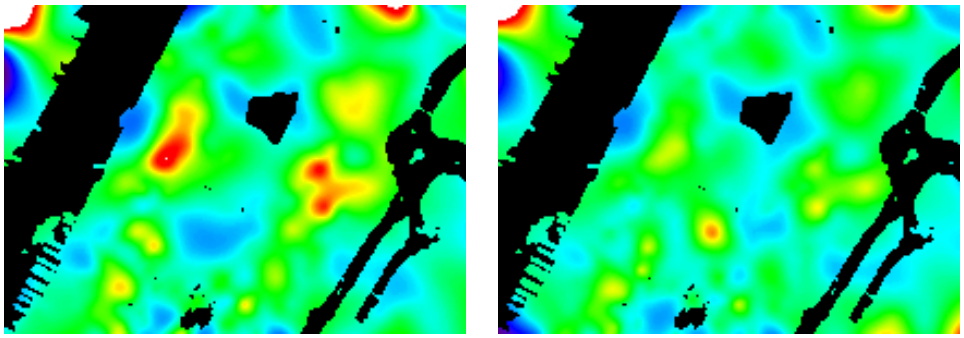
Our EOF analysis is mathematically related to the methods used in [1] and [6] to analyze cellular network data. However, we use the EOFs to identify and remove the dominant temporal periodicities in the data—thereby revealing any non-periodic spatial patterns related to greenspace and temporal patterns related to weather. In addition, we are experimenting with the combined use of EOF analyses and linear mixture models as described by [10].

5 Changes in Human Density around Greenspace

We are continuing to refine the analysis approach described in the previous section, and to apply it to the dataset described in Section 2. However, we can already see relevant patterns of human behavior emerge from preliminary analysis of selected subsets of the data.

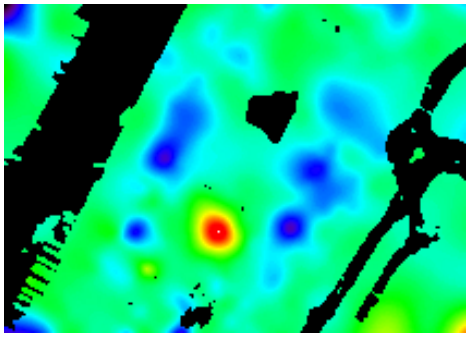
As an example of a seasonal change in human activity around greenspace, we compared Saturday-afternoon call density in central Manhattan between February 12 and July 9, 2011. We summed the per-minute call volumes between 2pm and 3pm for each sector within this area, then normalized by the sector area to produce density maps for each date. Figure 2 shows these two maps and their difference, along with a satellite image that highlights greenspace in the same area. We produced density surfaces using a 2D thin-plate spline to vary call density smoothly in space among the unevenly spaced sector centroids. By relaxing the tension on the spline, we minimized abrupt discontinuities between closely spaced centroids while preserving the larger-scale variations in density.

The July-minus-February map shows a conspicuous density increase in the greenspace of southern Central Park, with pronounced decreases in the residential areas of the Upper East and West Sides adjacent to the park. These changes

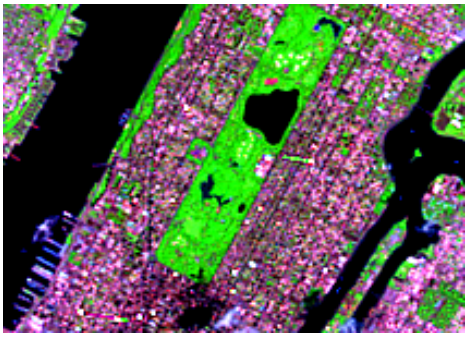


(a) Density: February 12, 2011, 2-3pm

(b) Density: July 9, 2011, 2-3pm



(c) Change in density: July minus February



(d) Landsat 5: July 7, 2011, 11am

(e) Normalized density scales: Call volume (top, 0 to 1) and change in volume (bottom, -0.005 to 0.005) per km^2.

Fig. 2. Spatio-temporal change in Saturday afternoon call density for central Manhattan. From winter to summer, call density increases in the greenspace of Central Park, but decreases in residential areas on the Upper East and West Sides. The visible-infrared satellite image shows parks and other greenspace as shades of green.

are consistent with the tendency of many New Yorkers to spend summer weekends outside the city, while many of those who remain visit Central Park.

Our analysis strategy for the complete New York metro area is based on the identification of spatiotemporal regularities and anomalies. We will use EOF analysis to quantify the spatial form of the dominant daily and weekly cycles associated with commuter migration, as well as any seasonal components that emerge in the low-order dimensions. Once identified, we will remove these components by inverse transformation of the remaining dimensions to produce a spatiotemporal representation of any anomalies that are distinct from the dominant periodicities. We can then directly compare the spatial components of these anomalies to maps of greenspace and thermal microclimate. We can likewise compare the temporal components of the anomalies to time series of air

temperature, precipitation, and humidity, in order to quantify whatever relationships may exist between the residual call volumes and the spatiotemporal variations in microclimate and ambient environmental conditions.

6 Conclusion

We have presented our ongoing exploration of how people use urban greenspace. We base our study on anonymous and aggregate records of cellular network activity in the New York metropolitan area. We developed a new tessellation technique to estimate the geographic coverage areas of individual cellular sectors. We are applying Empirical Orthogonal Function analysis to identify spatiotemporal patterns in the volume of cellphone activity in those areas. Our preliminary results indicate that our approach identifies relevant patterns of human behavior. We are continuing to refine our approach and apply it to larger-scale datasets. We also plan to carry out ground-truth studies to validate our results.

References

1. F. Calabrese, F. Pereira, G. DiLorenzo, L. Liu, and C. Ratti. The geography of taste: analyzing cell-phone mobility and social events. In *International Conference on Pervasive Computing*, 2010.
2. A. Green, M. Berman, P. Switzer, and M. Craig. A transformation for ordering mutispectral data in terms of image quality with implications for noise removal. *IEEE Transactions on Geoscience and Remote Sensing*, 26:65–74, 1988.
3. T. Horanont and R. Shibasaki. An implementation of mobile sensing for large-scale urban monitoring. In *Workshop on Urban, Community, and Social Applications of Networked Sensing Systems*, 2008.
4. J. Lee, A. Woodyatt, and M. Berman. Enhancement of high spectral resolution remote sensing data by a noise-adjusted principal components tranform. *IEEE Transactions on Geoscience and Remote Sensing*, 28:295–304, 1990.
5. R. Preisendorffer. *Principal component analysis in meteorology and oceanography*. Elsevier, Amsterdam, 1988.
6. 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.
7. A. Singh and A. Harrison. Standardized principal components. *International Journal of Remote Sensing*, 6:883–896, 1985.
8. C. Small. Estimation of urban vegetation abundance by spectral mixture analysis. *International Journal of Remote Sensing*, 22:1305–1334, 2001.
9. C. Small. High spatial resolution spectral mixture analysis of urban reflectance. *Remote Sensing of Environment*, 88:170–186, 2003.
10. C. Small. Spatio-temporal dimensionality and characterization of multitemporal imagery. *To appear in Remote Sensing of Environment*, 2012.
11. V. Soto and E. Frias-Martinez. Robust land use characterization of urban landscapes using cell phone data. In *Workshop on Pervasive Urban Applications*, 2011.
12. H. von Storch and F. Zwiers. *Statistical Analysis in Climate Research*. Cambridge University Press, Cambridge, UK, 1999.