Title: Inferring land use from mobile phone activity and points of interest

Abstract:
Understanding the spatiotemporal distribution of people within a city is crucial to many planning applications. However, obtaining data to create required knowledge currently involves costly survey methods. Data collected from pervasive mobile devices and large databases archiving points of interest in cities are providing new insights on urban systems. The locations and communication patterns of millions of individuals are recorded alongside information about the function of the places they go. This work uses dynamic data to quantify the relationship between activity within an area (measured via mobile phones) and land use. First, we implement a machine-learning algorithm to assess the ability of mobile phone data to predict land uses as designated by municipal governments. Finding modest success, we perform a detailed analysis of errors that suggests official zoning may be insufficient to understand activity within a place. To test this further, we incorporate additional data on points of interest crawled from a large online database, boosting predictive accuracy and supporting our theory. This analysis suggests measured activity and points of interest areas are inconsistent with officially zoned uses. Results provide a temporal dimension to our understanding of land use and suggest new data sources that may give a more accurate description of activity in a place.

Keywords:

Land Use, Big Data, Classification, Spatial Data

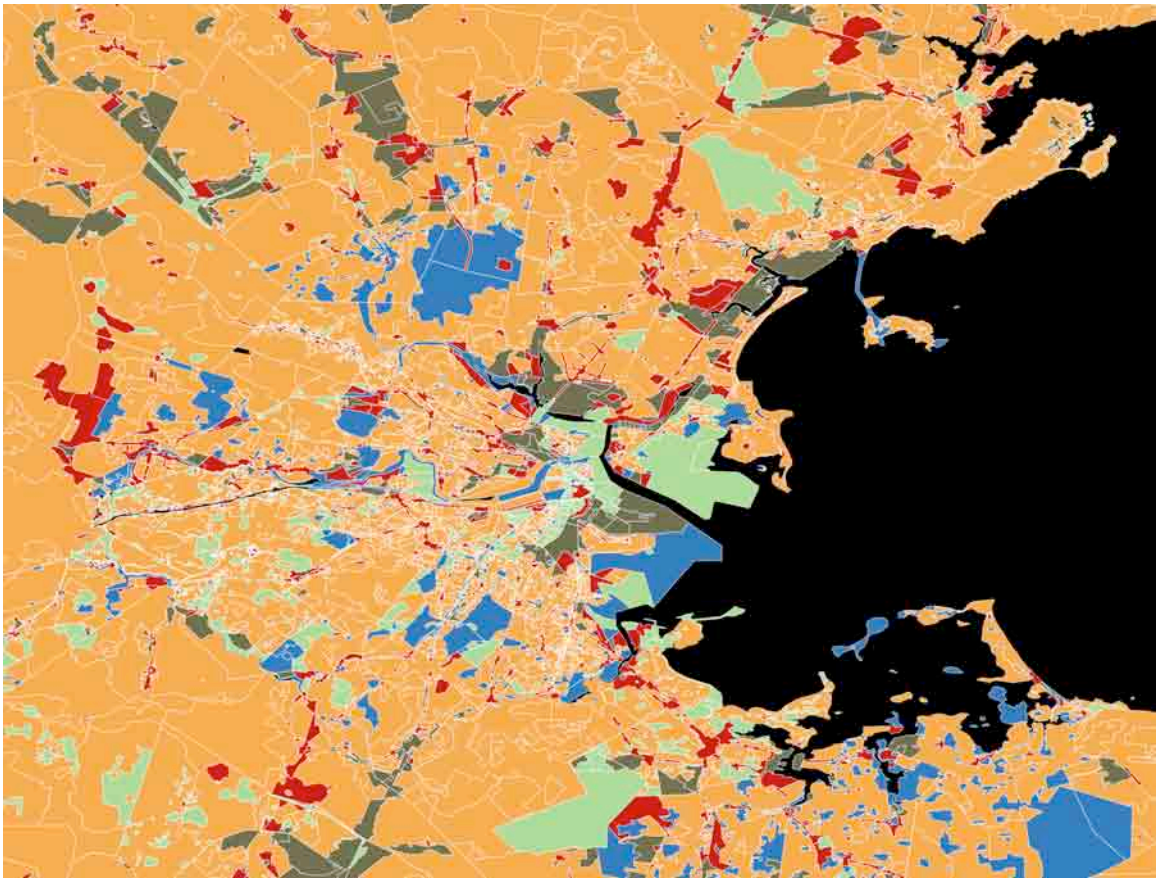

Figure 1: Zoning regulation for the Boston area. Color code: orange - Residential, red - Commercial, gray - Industrial, blue - Parks, green - Other.

Introduction:
In describing the ``organized complexity'' of cities, Jane Jacobs notes that a "park's use depends, in turn, on who is around to use the park and when, and this in turn depends on uses of the city outside the park itself." (Jacobs, 1961) Where people live, work, and play is intimately related to the time and distance required to move to and from places (Geurs, 2004). Understanding how individuals are distributed in space and time is crucial to making effective and efficient planning decisions within cities. The locations of public facilities and private businesses are influenced by and determine the demand for mobility.

How a particular area of a city is used is decided, in part, by the zoning regulations implemented and enforced through local governments. These regulations impact the structure of a city, dictating where housing or office space can be located. Zones of a kind share common usage. The central business district (CBD), for instance, is populated during office work hours whereas when offices are closed, relatively few people are found in these zones. Thus land use is closely is related to differences in population presence to be found at any given time in the zone. In practice, however, many zones feature a variety of uses that may differ from the official designation. As an example, zoning information for the Boston area is shown in Figure 1 Note, that zoning areas are not only restricted to land but also cover parts of rivers, lakes and the sea.

There is a large body of work dedicated to understanding the spatiotemporal dynamics of population and its relation to land use (Maat, 2005), (Banister, 1997), (Cervero, 1996). Measurements of human mobility within cities have traditionally been made via travel surveys. These surveys require subjects to record where they move to and from over an observation period (typically one day or a whole week), how they do so, and why. However, because surveys typically feature in-person interviews and demand a high workload from each subject, this method of data collection is expensive and limited.

Given these limitations, travel surveys suffer from relatively small samples (usually below tens of thousands of individuals), capture only short periods for each individual, and are updated infrequently. Fortunately, over the past decade a new type of measurement instrument has made its way into the pockets of people in nearly every culture and country. Each of the roughly 6 billion mobile phones currently in use[1] is capable of recording the location of calls, SMS, and data transmissions to within a few hundred meters. Moreover, these data are also collected and stored centrally by mobile phone providers for billing purposes.
With these data come enormous opportunities to improve our understanding of human mobility patterns.

In particular, call detail records (CDR), which provide information on the location of mobile phones any time a call is made or a text message is sent, provide large amounts information on the distribution of persons in a region at low costs compared to surveys. With such detailed information on the movement and communication patterns of individuals, privacy is an immediate concern. For the purpose of this study, we aggregate data to measure only the number of active phones in a given area during a given time interval. This method of data collection provides much higher levels of anonymity and reduces the risk that any breach of individual information while still revealing insights into the city.

In parallel with the growth in mobile phone use, services such as Google Places and Yelp allow anyone to log and access information on local points of interest (POIs) detailing the exact location and type of businesses, parks, and other important locations across a metropolitan area. These databases provide low cost, current, and detailed information on the places individuals go and what they do when they get there. Given these new data sources, the question arises as to whether the distribution of the numbers of active mobile phones and POIs can be used to quantify the how activity varies with zoning.

To have such a measurement method would be very advantageous. Zoning regulations exist, in part, to control the structure and function of cities. New data sources may provide a method of testing if the actual activity patterns of locations match the desired affect of zoning. In the event that activity patterns are not uniform across similarly designated zones, quantifying difference in usage may help better understand demand for mobility infrastructure across space and time compared to current analysis. Monitoring

[1] http://www.itu.int/net/ITU-D/index.aspx

the usage over time allows to detect changes in habits of the population as well as shifts in usage, which may indicate ongoing regional developments.

Consequently this work investigates the potential of aggregated CDR data and point of interest data to infer dynamic land use, i.e. to understand how the population of different areas of a city changes with time and the type of businesses that exist there according to specific zoned land uses. The work centers on supervised classification of regions according to given zoning regulations. We demonstrate that CDR data can be used in order to classify zones of different types with reasonable accuracy. We explore the limitations of this accuracy in comparison to alternative measurements of land use. Through this process, normalization techniques are discussed to highlight differences between zones. The application and result of random forests for the classification is described in detail.

Mobile phones and human mobility:
Mobile phones have proven exceptionally good instruments to measure human behavior with. In one of the first studies utilizing these devices, Eagle and Pentland, 2006 were able to decompose mobile phone activity patterns of university students and employees into regular daily routines. Moreover, these patterns were found to be predictive of an individual's characteristics such as their major or employment level (i.e. graduate student). Subsequent research has built upon this work, scaling up in both geographic extent and sample size. Gonzalez et al, 2008 studied data from nearly one hundred thousand anonymous mobile phone users to reveal persistent regularities in the statistical properties of human mobility. Highlighting the remarkable predictability of human behavior, Song et al, 2010 estimated that it is theoretically possible to predict individual movements of users with as high as 93% accuracy using only data from mobile phones.

Mobile phone data has also provided insights on how space is used over time. For example, Reades et al, 2009 link mobile phone activity to commercial land uses in Rome, Italy. Measuring activity in 1km by 1km grid cells, they employ a form of principal component analysis to identify the dominant activity patterns. The authors qualitatively interpret areas of the city exhibiting this signal as commercial, though actual zoning information is not introduced. They then decompose activity across the city to identify regions with similar patterns of usage. Soto et al, 2011 use CDR mobile phone data at the cell tower level to identify clusters of locations with similar activity. Qualitative agreement between these clusters and land uses were observed. Applegate et al are able to identify clusters of call activity pertaining to work or recreation hours (Applegate, 2011).

Calebrese et al, 2010 apply similar decomposition and clustering techniques to classify locations on a university campus as classrooms, dormitories, etc. By analyzing Wi-Fi activity across 3000 Wi-Fi-access points, the authors used unsupervised, non-parametric techniques to identify clusters of locations with similar Internet use. These locations naturally fit into location profiles such as "lecture hall" or "dormitory." Finally, CDR data have proven useful to detect movement at the census tract scale (Calabrese, 2011). Location data from calls helped to measure origins and destinations for trips across the

Boston Metropolitan area. However, no attempt was made to associate such trips with land uses.

Other data sources have also been studied. Frias-Martinez et al, 2012 use unsupervised learning techniques on geo-tagged Tweets to cluster locations, finding strong correlations between found groups and actual land use patterns in Manhattan. Points of interest (POIs) and GPS data collected from taxi fleets have been combined with unsupervised learning algorithms to identify the rich structure of different functional sections of Beijing (Yuan, 2012).

To date, however most studies choose unsupervised over supervised learning techniques to combine traditional data sources on land use such as zoning regulations and CDR data or POI data. While unsupervised learning techniques impose fewer constraints on patterns within the data, their output is often difficult to compare with traditional planning tools like zoning. Zoning codes are a primary tool used by planners to encourage or discourage growth and usage patterns across space. Explicitly linking activity patterns measured from new digital sources to intended zoning use would provide a new, more efficient tool for planners to assess the effectiveness of these policies.

This study aims to investigate the link between zoned land use and mobile phone activity on a common spatial partitioning of the greater Boston area into regions of homogeneous land use. For each region the temporal profile of active phones is used in supervised classification techniques in order to identify patterns characteristic for a specific zoning classification. The corresponding patterns will be interpreted in detail.

Data sources:
Three data sources are used in this work: mobile phone activity records, digital point of interest logs, and official zoning regulations. For the Boston metro region, anonymized CDRs provide the location of a mobile phone by triangulating signal strengths from surrounding cell towers. Note that this differs slightly from traditional CDR data in which record the location of a call as the location of the mobile phone tower. This provides slightly higher accuracy and allows us to measure calls continuously across space rather than at points where towers are located. Triangulation by this method is accurate to within a few hundred meters depending on the tower density. These data make it possible to measure the amount of phone activity (counts of the number of calls and texts) that occurs within a given area and time window. In this study we use three weeks of CDR data for roughly 600,000 users in the Boston region home to nearly 3 million people. Though mobile phone data come from specific set of carriers, the market share of these carriers is between 30% and 50%. Moreover, for the same data, Wang et al, 2012 have shown that the ratio of census population to unique mobile phone users is nearly uniform across the region.

Points of interest (POIs) for the greater Boston region were crawled from the Google Places API. This web application allows businesses to place information such as address, descriptions, hours, etc. on into Google's database. When users search a spatial region for points of interest matching keywords such as "restaurant" or "post office," they are

given a list of locations matching those criteria in their area. Each point of interest consists of a location (latitude and longitude) as well as a number of tags describing the business or place (note a location may have more than one tag). The vast majority of POIs are establishments such as businesses or government facilities. Other semantic places such as markers for political zones or important transportation junctions also exist. For the Boston region, our dataset contains roughly 180,000 POIs with 104 unique descriptive tags.

In addition to novel data sources such as mobile phones and online POI databases, we obtain zoning classifications for the Boston metropolitan area. The Massachusetts Office of Geographic Information (MassGIS) aggregates uses into five categories: Residential, Commercial, Industrial, Parks, and Other. We are careful to note our assumption that actual land use and zoning classification are closely related while acknowledging that zoning regulations are only proxies for actual land use imposing restrictions.

Common spatial representation:
The first obstacle to studying the relationship between phone activity and land use is the reconciliation of the spatial dimensions of the data: While the location of the phone activities are recorded as coordinate pairs, zoning data is provided in polygons at roughly the parcel scale. The spatial partitioning of phone and population data is rarely the same as zoning parcels. To reconcile all data sources as well as to reduce the influence of noise (due to inter alia sources localization estimation noise) in the data, we transform both to the same uniform grid. A lattice is laid over the analysis region such that every cell in the lattice measures 200 by 200 meters. Different grid sizes have been tested. A size of 200 meters proved coarse enough to reduce the noise level and detailed enough in order not to mix many parcels of different zoning areas.

In order to reduce the high noise level average hourly time series of phone activity are computed. Here, the average is computed for each hour within a day of the week. Only cells with mobile phone activity above a certain threshold are used in the analysis. Similarly, the number of POIs tagged with each of the 104 unique descriptors is counted for each grid cell. With respect to zoning data, each cell is given a single zoning classification based on the most prevalent (in terms of fraction of area covered) use within the area.

Potential pitfalls of this method arise due to large heterogeneity in population density. Downtown areas are much more densely populated than the suburbs, a characteristic that is reflected in other spatial divisions like census tracts. This leads to sparse mobile phone activity in rural regions. However, the small grid size used in this analysis retains detailed information about block to block zoning regulations in dense urban areas. Figure 2 displays actual zoned parcels versus the gridded approximations. Table 1shows the frequency of each zoning class in the grid. The vast majority of land, nearly 75% of cells, is zoned as Residential. Other uses appear in roughly equal fractions.

Table I. Tabulation of Boston zoning.

Zone Use	Category Index	Count	Percentage
Residential	1	23322	74.28
Commercial	2	1854	5.90
Industrial	3	2236	7.12
Parks	4	1941	6.18
Other	5	2045	6.51

The land use profile of the city is dominated by residential use accounting for nearly 75%. Other uses share roughly the same percentage of remaining land.

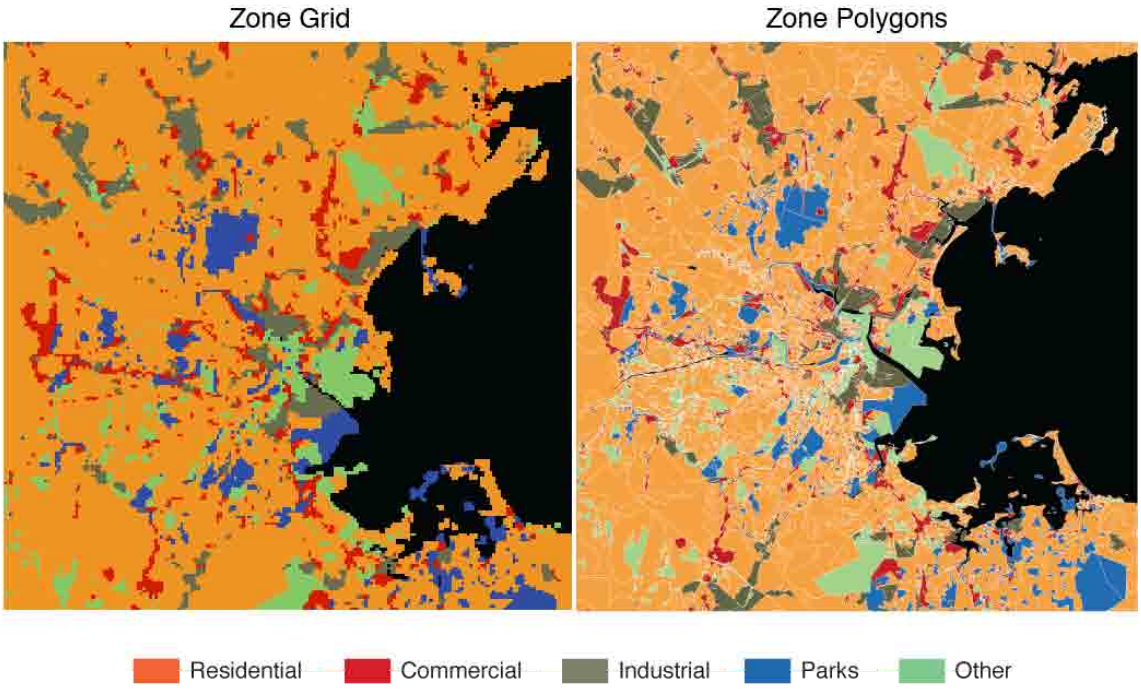

Figure 2: To improve computational efficiency and reconcile all mobile phone and traditional data sources, we create a uniform grid over the city. Zoning polygons (right), are rasterized to cells 200m by 200m in size (left). For cells where more than one zoning class exists, the most prevalent class is used. Given the small size of these cells, this data transformation provides an accurate map of the city while improving computational efficiency.

Descriptive Statistics:

We next examine the relationship between mobile phone activity and land use at the macro, citywide scale. Figure 3 displays time series of mobile phone activity averaged over all cells of a given zoning classification. Examining absolute counts of mobile phone events reveals that the average activity differs greatly between zoning classifications. While residential areas only show a maximum activity of roughly 50 events per hour, commercial cells reach approximately 100 events on average.

The spatial distribution of activity is also heterogeneous. The downtown area of Boston shows orders of magnitude higher activity levels than typical residential zones. In order to allow for classification based on relative mobile phone activity, time series are normalized using a z-score. By definition, the normalized time series have zero mean and unit standard deviation. Mathematically, the normalized activity of cell (*i,j)* is given by:

$$a_{ij}^{norm}(t) = \frac{a_{ij}^{abs}(t) - \mu_{ij}^{abs}}{\sigma_{ij}^{abs}}$$

The second row of Figure 3 (a) shows the average (over cells of one zoning class) normalized activity. These profiles are remarkably similar for all zoning classes showing the strong circadian rhythm of the city. Residents wake up, go to sleep, and wake again the next day. The rise and fall of activity in each zone, however, is not solely the result of users moving into and out of a region. Instead, it is largely due to an uneven distribution of phone use across the day. To account for this, during each hour, we subtract the average normalized activity of the entire metro region from the normalized activity at each given cell. The corresponding spatially de-meaned series will be referred to as *residual activity*. Residual activity can be interpreted as the amount of mobile phone activity in a region, at a given time, relative to the expected mobile phone activity in the whole city at that hour. Mathematically, it is calculated as follows:

$$a_{ij}^{res}(t) = a_{ij}^{norm}(t) - \bar{a}^{norm}(t)$$

where $\bar{a}^{norm}(t)$ is the normalized activity averaged over all cells at each particular time.

Averaging the residual activity for each zoning classification reveals patterns related to travel behavior. The last row of Figure 3 (a) and (b) provide the residual activity averages across zoning classes for weekdays and weekends. The most notable signal is the inverse relationship between residual activity in residential and commercial areas: While residential areas on average show higher than expected activity during the night and lower than expected during weekdays. As expected, the opposite is true for commercial zones. Somewhat surprisingly, the normalized activity does not show these features strongly. Only the residual activity demonstrates the expected behavior. There, also higher than average activity in parks on the weekend afternoons is visible.

Residential areas have higher residual activity in the early morning hours and late at night, while commercially zoned cells have a peak period during the day and show much lower activity levels late at night. These patterns most likely reflect the 9-to-5 business hours of offices and stores. More subtle patterns are also visible. In Boston, much of the CBD is zoned as Other or Mixed use. We see that residual phone activity in this zoning type has peaks in the early morning hours on Saturday and Sunday, suggesting these areas support nightlife on the weekends. These citywide time series show that mobile phone activity and land use are linked at the highest level of aggregation. By treating phone activity as a proxy for the spatial distribution of people at a given time period the expected patterns of concentration of people in the CBD and inner city region during the working day, and the shifts induced by the commuting behavior are visible in the residual activity levels.

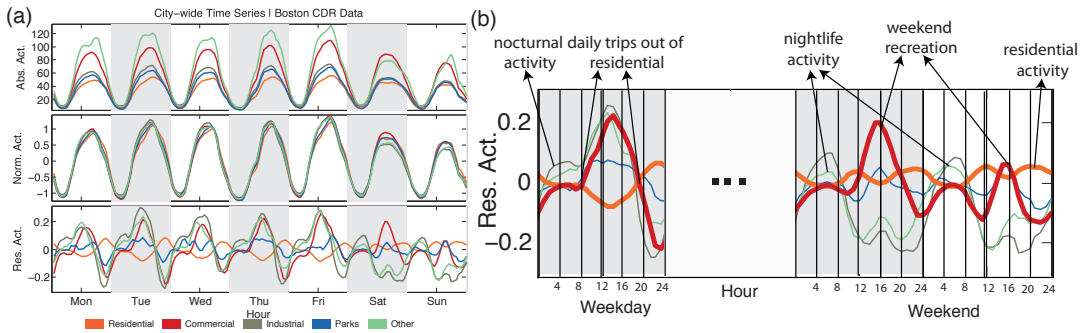

Figure 3: (a) Plots are shown for three different time series of average mobile phone activity within each of five land uses. The first plot shows absolute activity (number of calls and SMS messages). The second plot displays z-scored time series. The bottom plot shows residual activity. (b) More detailed view of average (over cells of the same zoning class) residual activity.

Figure 4 displays the spatial distribution of normalized activity (top row) and residual activity (bottom row) at three time instants. Not shown in the plots are the absolute activity levels that are distributed much like population density. The CBD of Boston has orders of magnitude more activity than the rest of the city. Mapping the logarithm of absolute activity over time once again only reveals the circadian rhythm of the city that strongly dominates the differences in land usage that consequently are not seen in these plots.

In the spatial distribution of the normalized activity the dominance of the CBD is less pronounced. Nevertheless, the circadian rhythm still dominates the differences between different zones. From this perspective, Boston appears as a mono-centric region, with small pockets of density located on an urban ring roughly 20km from the CBD.

By contrast, the spatial distribution of residual activity reveals a much richer structure. In the early morning hours, residual activity is located on the periphery of the region. During the day, this activity becomes heavily concentrated in the CBD or in small sub-centers on the urban ring. Later in the evening, activity again returns to residential areas on the periphery, away from centers. This suggests some correlation between commuting patterns and the spatial distribution of residual activity.

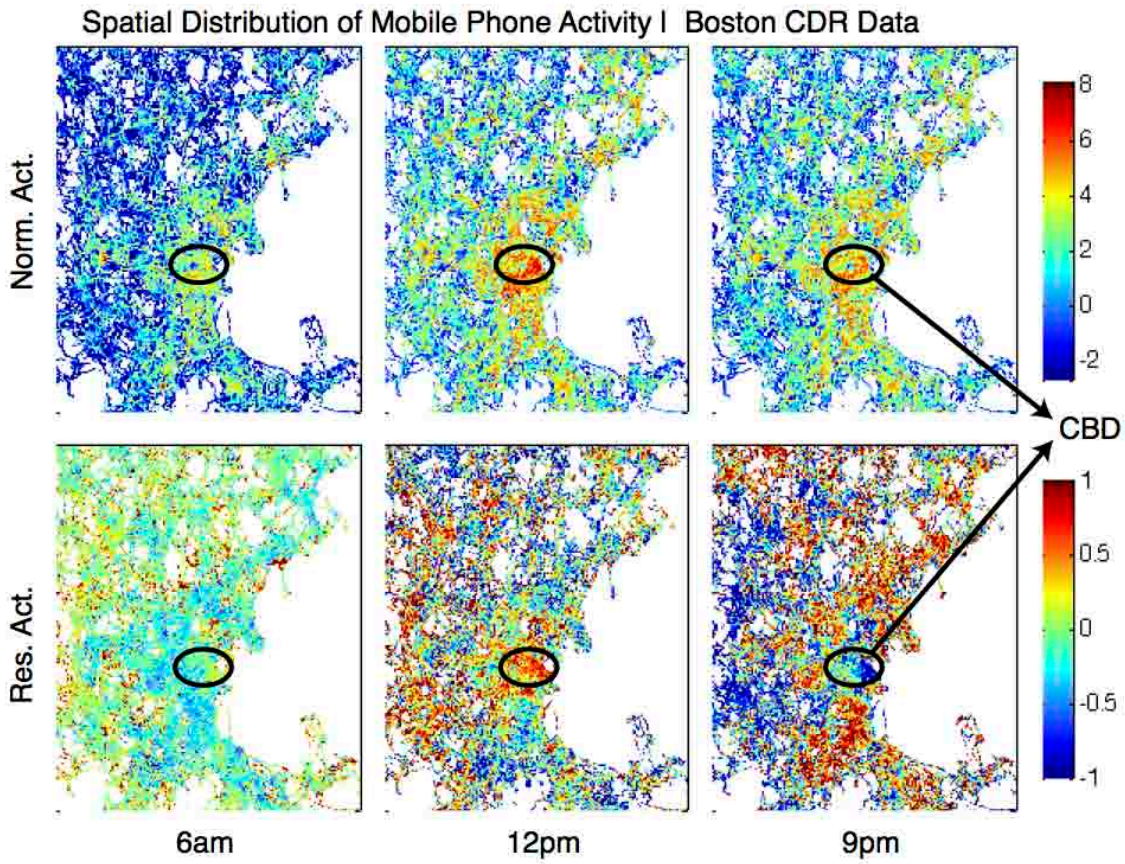

Figure 4: Spatial distribution of absolute and residual phone activity over the course of a day. While absolute mobile phone activity is dominated by population density with sleep and wake patterns, residual activity reveals flows into and out of the city center over the course of a day.

In addition to the spatial distribution of mobile phone activity, we also explore point of interest density for the metro region. Figure 5 displays counts for the twenty most common POI tags as well as a spatial density plot showing their distribution in space. More general tags such as `establishment' are featured prominently, while some tags like `park' match very closely with official zoning classifications. As expected, we find that the CBD has the highest density of POIs, with smaller secondary centers visible as seen in mobile phone activity.

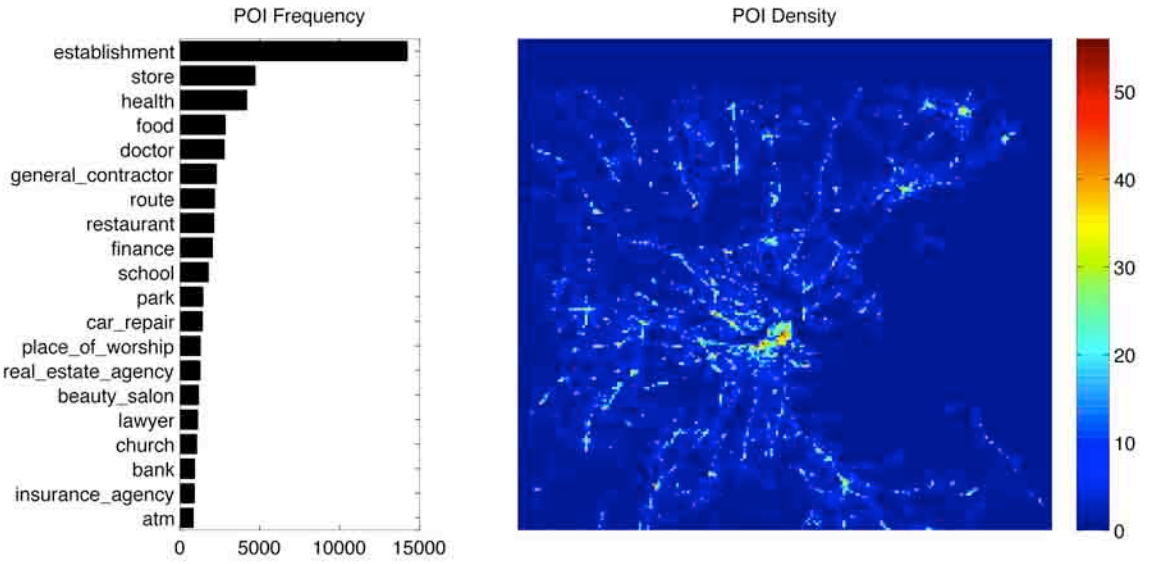

Figure 5: The left panel displays the region-wide frequency of the 20 most common point of interest tags. While some descriptors such as `establishment' are vague, others like `park' align closely with official zoning classifications. The right panel shows the spatial density of all POIs across the region. The central business district predictably shows the highest density, while other clusters are dispersed throughout the region.

Classifying Land Use by Mobile Phone Activity:

In the last section we observed correlation between residual mobile phone activity and land use on the macro scale. Fluctuations in mobile phone activity mimic our intuition of population changes related to commuting and recreational trips. In this section we investigate whether usage patterns in cell of a given class are homogeneous. This will be done by performing supervised classification based on features extracted from the residual activity time series and the classes provided by the zoning regulations as labels. Though previous work in this area has employed unsupervised learning techniques, access to extensive zoning data in a mature, regulated city such as Boston makes supervised learning an attractive option. Cross validation is used to test performance.

We implement the *random forest* approach described by Breiman, 2001. Other approaches including neural network based classifiers have been tested and led to similar results. Random forests are useful for their ability to efficiently classify data with large numbers of input variables (such as long time series). Rather than make comparisons for every feature of the data every time, a number of random subsets are chosen to more efficiently search the space. This does not come at the cost of accuracy as random forests have been shown to have high performance on a variety of datasets. Moreover, random forest classifiers allow weights to be introduced so that more frequently occurring classes do not overwhelm smaller ones. This feature will be exploited later to control for the large share of residentially zoned locations.

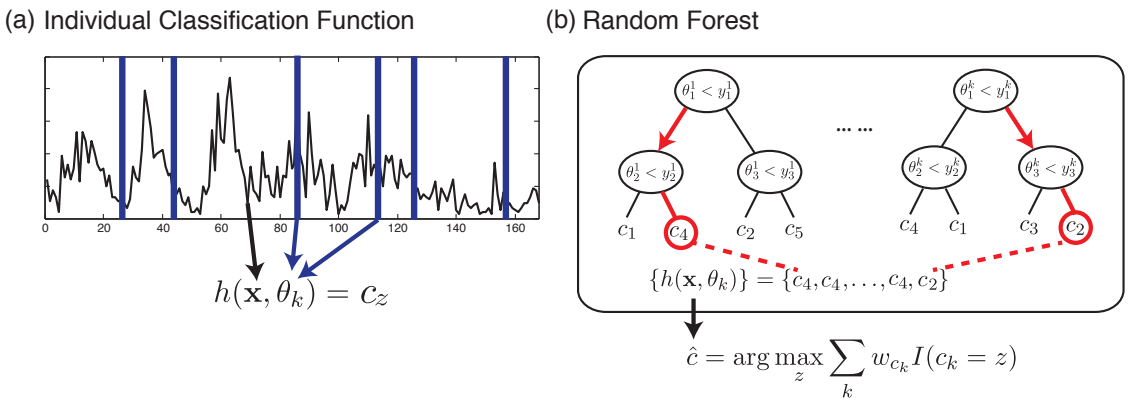

Figure 6: (a) Shows the inputs to each decision tree $(h(\boldsymbol{x}), \theta_k)$ A time series of residual phone activity, $\boldsymbol{x}$, is input and activity at a random subset of times , θ_k (denoted by the blue bars), is chosen to make comparisons. (b) A depiction of the random forest shows a number of different trees making predictions based on a different set of random times. Each tree casts a weighted vote for a certain classification. A final classification, $\hat{c}$, is made by counting these votes.

A random forest, $\{(h(\boldsymbol{x});\ \theta_k)$, k=1,...$\}$, is constructed from a set of decision trees as visualized in Figure 6. The training data determines the parameter vectors θ_k. Least squares or maximum likelihood estimation can be used to find these configurations. To obtain a single prediction for each input time series, a voting scheme is implemented. Each tree votes for a class based on its prediction. These votes can be weighted (weights denoted by w_{c_k}) so that votes for one class count more or less than votes for a different class. The weighted votes are summed and a single zoning class prediction, $\hat{c}$ is chosen for the original input time series.

For the calculations we use a MATLAB implementation of the random forest algorithm released by Jaiantilal[2]. Given the periodicity observed in the data, our initial approach uses 49 input features that are computed for each location as the input feature vector $\boldsymbol{x}$. These features include a 24-hour time series of residual mobile phone activity during an average weekday as well as a 24-hour time series of residual activity for an average weekend-day. The final feature is the mean of the location's absolute activity on any given day. Additional features such as the variance of mobile phone activity were tested, but none aided prediction. The output of the algorithm is a zoning classification for each location. Cross validation is used to test accuracy. We create 500 trees for each forest and define total accuracy as the fraction of correctly classified cells on the validation part of the sample.

Our first set of results include all five zoning classifications: Residential, Commercial, Industrial, Parks, Other. When all land use classes are included, however, we face a major challenge with classification. As noted above, nearly 75% of all cells are primarily residential. The next most common zoned use is Industrial at 7%. Because of our definition of total accuracy, the most naive classifier, simply assigning Residential to

[2] http://code.google.com/p/randomforest-matlab/

everything, will achieve 75% total accuracy, but will fail to capture any diversity in use. To guard against this, we weight the voting system to raise or lower the required votes in order to choose a given classification. The maximum of the weighted votes then provides the predicted class. Systematic variations of the weights on a (coarse) grid led to a choice of weights where the criterion applied was maximum classification accuracy for all classes but residential.

Finally, we note that the random forest classifier uses local information only to make a prediction. Given the size of our grid cells, it is reasonable to assume that land use does not differ greatly from each 200m by 200m tract of land to the next. To incorporate neighborhood information into our predictions, we implement a second pass algorithm. After the classifier has made a prediction for a cell, we examine the predictions for each of that cell's neighbors. If the majority of neighboring cells were predicted to be a land use that differs from the cell in question, that cell is switched to the majority use of its neighbors. In practice, this results in some spatial smoothing of noisy classification data. We find that performing the second pass provides gains of 2-10% overall accuracy for each classifier.

Even with vote weighting and the second pass algorithm, we achieve only modest results. Table 2 shows 54% accuracy over the whole city. This implies that striving for equal classification accuracy among prevalent classes reduces overall accuracy by about 20%. Figure 7 displays the spatial distribution of correctly and incorrectly classified locations. We note, however, that the algorithm does capture some spatial patterns in the data and that our intra-use accuracy is relatively high for Commercial and Industrial uses. Parks and Other mixed uses remain difficult to classify.

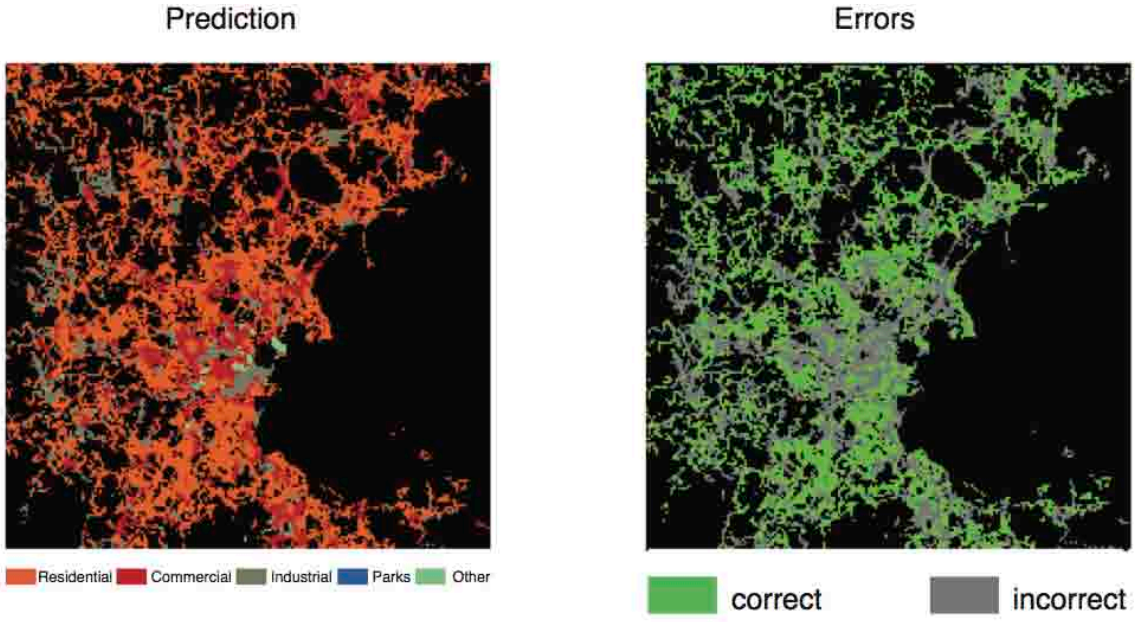

Figure 7: Left plot: zoning map as predicted from mobile phone data using the random forest classification algorithm. Right plot: spatial distribution of where the algorithm predicts land use correctly and where it fails. In general, these errors seem randomly distributed in space, suggesting that errors are not the result of some spatial correlations

such as population density. For comparison to actual zoning, see the left panel of Figure 2.

Table II. Random forest classification results.

Total Accuracy:	0.54				
	Res	Com	Ind	Prk	Oth
Land Share:	0.74	0.09	0.08	0.04	0.05
Vote Thresh:	0.60	0.10	0.10	0.10	0.10
	Confusion Matrix				
	Res	Com	Ind	Prk	Oth
Res	0.62	0.21	0.15	0.01	0.01
Com	0.30	0.48	0.19	0.00	0.02
Ind	0.33	0.27	0.38	0.00	0.02
Prk	0.52	0.26	0.18	0.02	0.02
Oth	0.37	0.28	0.25	0.00	0.10

Total accuracy is defined as the fraction of correctly classified cells. The share refers to the percentage of cells actually zoned for each class of use. Element (i, j) of the confusion can be interpreted as the fraction of actual zoned uses of class i that were classified as use j by the random forest. Thus the high percentages in the Res column can be interpreted as the algorithm heavily favoring classification as residential due to its overwhelming share of overall uses.

Table III. SVM and kNN classification results.

Total Accuracy:	0.46					Total Accuracy: 0.69				
	Res	Com	Ind	Prk	Oth	Res	Com	Ind	Prk	Oth
Land Share:	0.74	0.09	0.08	0.04	0.05	...				
Vote Thresh:	0.60	0.10	0.10	0.10	0.10	N/A				
	Confusion Matrix					Confusion Matrix				
	Res	Com	Ind	Prk	Oth	Res	Com	Ind	Prk	Oth
Res	0.51	0.17	0.11	0.10	0.11	0.90	0.05	0.03	0.01	0.02
Com	0.24	0.40	0.14	0.07	0.05	0.72	0.17	0.04	0.01	0.05
Ind	0.22	0.22	0.35	0.04	0.17	0.67	0.10	0.15	0.00	0.08
Prk	0.34	0.28	0.11	0.16	0.11	0.78	0.10	0.05	0.03	0.04
Oth	0.27	0.28	0.16	0.10	0.28	0.68	0.12	0.06	0.00	0.14

We compare results from the random forest algorithm to results from supervised learning using a SVM with a Gaussian radial basis function and a k-nearest neighbors algorithm with $k = 10$. Both algorithms perform worse than the random forrest implementation.

To ensure that these results are not specific to the random forest algorithm, we also learn with support vector machines (SVM) and k-nearest neighbor (kNN) algorithms. Results are shown in 3.

Using a Gaussian radial basis function as a kernel, the SVM achieves 45% testing accuracy. Though less than the random forest, the SVM performs better on rarer classifications. Later we will show that a two-stage random forest implementation can be used to achieve even greater accuracy. The kNN algorithm achieves higher overall accuracy, but does so by over-predicting residential areas. This is likely because there are so many residentially zoned cells, it is difficult to find neighbors who are some other type. In general, the random forest algorithm provides the best performance when weighting over-all accuracy and accuracy within each zoning class.

To account for the tendency of the algorithm to over-predict residential use, we remove cells zoned as Residential from consideration. This leaves a nearly equal share of the remaining four uses: Commercial, Industrial, Parks, and Other. Table 4 and Figure 8 display results for this sub-classifier. Now, the zone with the largest share is commercial use, which only accounts for 33% of non-residential zones. Intra-use accuracy has improved significantly for Parks and Other mixed uses. Whereas the random forest including residential uses could only correctly classify 2% of zones classified for Parks, the sub-classifier, excluding Residential, correctly predicts 30% of park cells. A similar improvement from 10% to 34% is also observed for the Other or mixed-use category. The share of classes incorrectly classified as Residential roughly is distributed onto Parks and Others in the classifier without the Residential category, while commercial and industrial zones are not affected heavily. One hypothesis for this effect is that many cells while classified as Residential in rural areas are not fully developed and thus used as parks and in the city center show mixed usage. Including the large class of residential zones masks this effect.

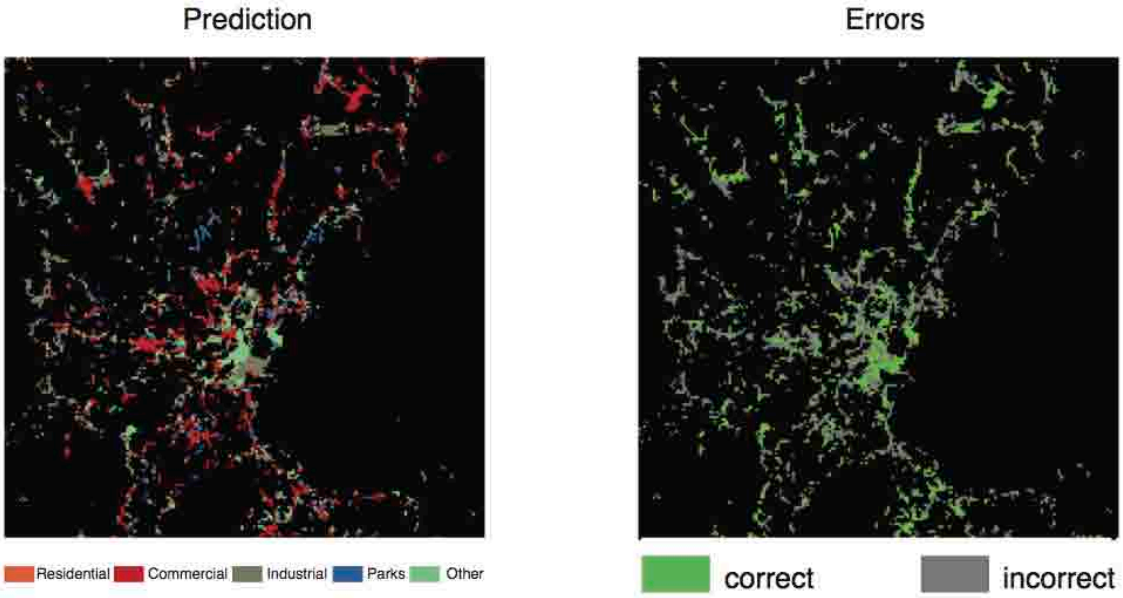

Figure 8: The left plot shows the city zoning map with residential areas removed as predicted from mobile phone data using the random forest classification algorithm. The right map displays the spatial distribution of where the algorithm predicts land use correctly and where it fails. Without residential areas to predict, the algorithm performs significantly better at predicting other uses. For comparison to actual zoning, see the left panel of Figure 2.

Table IV. Random forest classification results - less residential.

Total Accuracy:	0.40				
	Res	Com	Ind	Prk	Oth
Land Share:	0.00	0.33	0.31	0.16	0.20
Vote Thresh:	N/A	0.30	0.30	0.20	0.20
	Confusion Matrix				
	Res	Com	Ind	Prk	Oth
Res	N/A	N/A	N/A	N/A	N/A
Com	N/A	0.50	0.19	0.11	0.19
Ind	N/A	0.27	0.37	0.12	0.24
Prk	N/A	0.31	0.18	0.29	0.21
Oth	N/A	0.26	0.24	0.15	0.34

In this case, residential land has been removed from consideration. The algorithm is now able to correctly predict much larger fractions of rarer land uses.

The goal of the supervised learning algorithm is to make correct predictions of actual zoned use. Incorrectly classified cells are labeled as errors, but how an area is zoned is not necessarily the same as how it is used. As an example the area termed ``Back Bay'' containing some of Boston's most busiest shopping streets is classified as residential, as is the campus of MIT. Clearly these areas have a different usage than residential areas in the suburbs. A political and idiosyncratic process for setting and updated zoning regulations may lead to broad or unenforced development standards. In light of this, errors made by our classification algorithm may be due to incomplete zoning data rather than actual mistakes.
To examine this possibility further, we first analyze the prediction errors more closely and then introduce alternative data.

Figure 9 displays a detailed partitioning of classifier results. We compare average residual activity across three groups of cells: (I) All cells correctly predicted to be a given use. (II) All cells of another use incorrectly predicted to be the given use. (III) All cells of a given use incorrectly predicted to be some other use. Reviewing residential use, we see that Group I is defined as all residential cells correctly predicted to be residential. The average activity pattern is the most dominant pattern of residual activity for residential land use. We find that the residual activity in non-residential cells predicted to be residential (Group II) closely follows the pattern found in Group I. This strongly supports our hypothesis that though some zones are not classified as residential in the data, their phone activity patterns suggest they are used in similar ways. In contrast, the residual activity in residential cells incorrectly classified as some other use (Group III) displays the inverse pattern. This suggests our algorithm is identifying cells that are zoned as residential use but that do not share activity characteristic of that zoning class in reality.

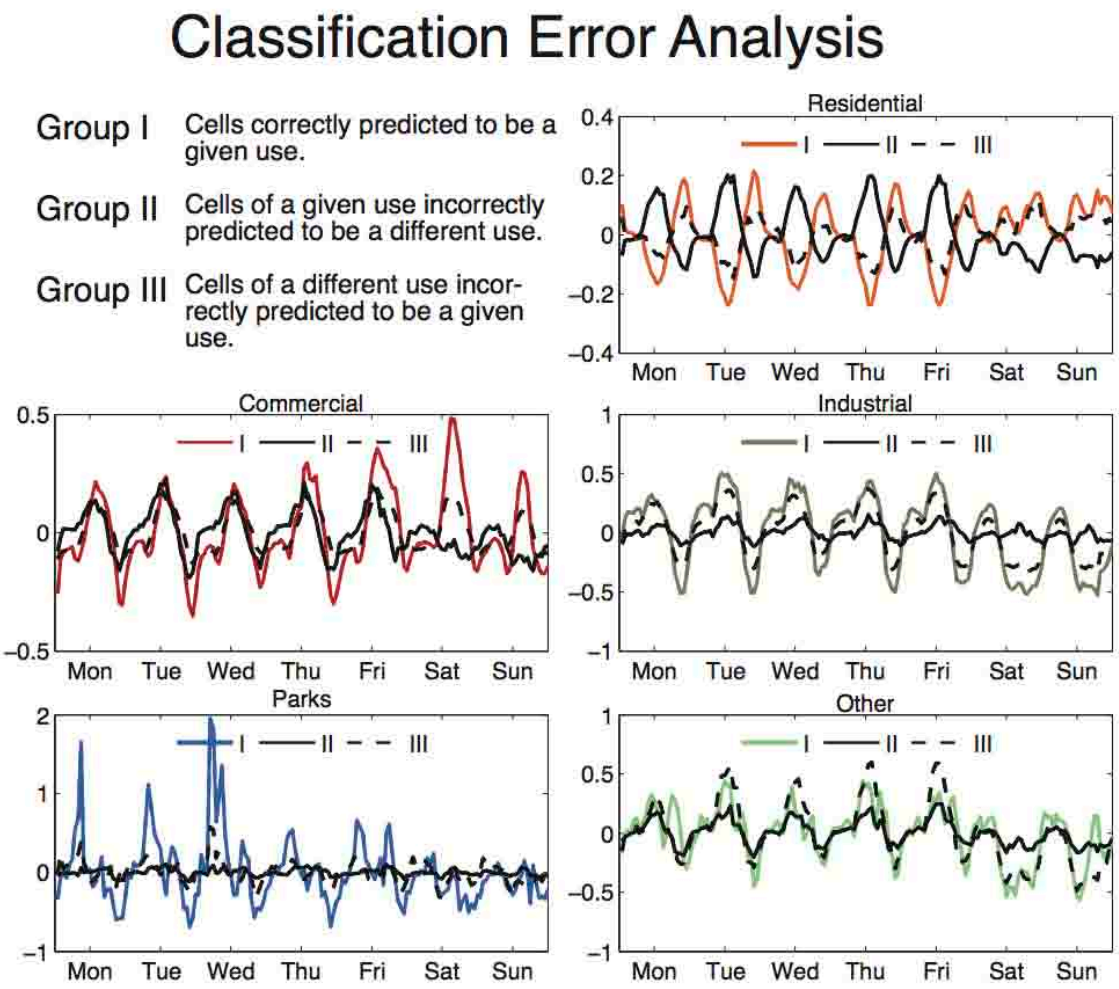

Figure 9: An analysis of classification errors. We consider three groups: (I) Cells correctly predicted to be a given use (II) Cells of a given use incorrectly predicted to be some other use (III) Cells of some other use incorrectly predicted to be a given use. For example, Group I includes all residential areas correctly predicted to be residential. Group II, residential cells predicted to be some other use (i.e. Commercial), have average activity that is the inverse of Group I, suggesting these locations were misclassified because they display fundamentally different activity patterns. Group III represent cells of other uses such as Commercial that behave like Residential.

Incorporating Points of Interest:

To further test our hypothesis that official zoning classifications may be incomplete, we incorporate point of interest (POI) data. Perhaps there is some behavior specific to when individuals use mobile phones that makes them unable to accurately proxy for activity in an area. POI data may provide a more direct measure of what is actually built in a region. With this in mind, we augment the feature vectors input into the random forest algorithm to contain both mobile phone activity and the composition of POIs in that grid cell.

For each cell, we create a feature vector with 104 elements (one for each unique tag) as counts of the number of POIs with that tag within the cell. These features may then be appended to the CDR data and input into the random forest algorithm.

To test the affect of adding POI information, we first attempt to predict official zoned uses using only the POI feature vectors, holding out mobile phone data. Total accuracy drops from 54% to 42%. Most of this decline is due to very high error rates in residential areas. Places that contain mostly private residences do not appear as places of interest on the map and are thus difficult for the algorithm to predict. Though overall accuracy falls, accuracy within other groups increases. Table 5 shows that success rates in commercial, park, and other/mixed areas rise by 8, 36, and 4 percentage points, respectively when compared to using CDR data (Table 2). These results are somewhat expected given the existence of tags directly indicating the existence of commercial establishments and parks.

Table V. Random forest classification results - POI Only.

Total Accuracy:	0.42				
	Res	Com	Ind	Prk	Oth
Land Share:	0.74	0.09	0.08	0.04	0.05
Vote Thresh:	0.60	0.10	0.10	0.10	0.10
	Confusion Matrix				
	Res	Com	Ind	Prk	Oth
Res	0.42	0.16	0.18	0.17	0.08
Com	0.09	0.54	0.27	0.05	0.06
Ind	0.15	0.25	0.42	0.09	0.10
Prk	0.24	0.14	0.12	0.38	0.12
Oth	0.19	0.29	0.16	0.22	0.14

Classification results using POI data only.

With mobile phone data able to separate residential from non-residential zones and POI data better suited for classifying non-residential types only, we next combine the two data sources. The combination is performed in two ways. The first runs our random forest algorithm with augmented feature vectors containing both mobile phone activity and POI data as input. The second performs a two-stage classification procedure. The two stage process first uses CDR data to predict residential or non-residential then takes all cells predicted as non-residential and classified them again as Commercial, Industrial, etc. using only POI feature data. Both methods give similar results (Table 6). The addition of POI information increases accuracy by roughly 15 percentage points to ~70% compared to CDR data alone. Moreover, improvements are made in mostly in commercial regions as apposed to residential.

Table VI. Random forest classification results - POI and CDR - One and Two Stage

One Stage						Two Stage					
Total Accuracy:	0.68					Total Accuracy:	0.71				
	Res	Com	Ind	Prk	Oth		Res	Com	Ind	Prk	Oth
Land Share:	0.74	0.09	0.08	0.04	0.05	Land Share:	0.74	0.09	0.08	0.04	0.05
Vote Thresh:	0.60	0.10	0.10	0.10	0.10	Vote Thresh:	0.60	0.10	0.10	0.10	0.10
Confusion Matrix						Confusion Matrix					
	Res	Com	Ind	Prk	Oth		Res	Com	Ind	Prk	Oth
Res	0.79	0.15	0.05	0.01	0.01	Res	0.89	0.05	0.01	0.02	0.03
Com	0.27	0.61	0.11	0.00	0.01	Com	0.58	0.31	0.05	0.02	0.05
Ind	0.35	0.30	0.34	0.01	0.00	Ind	0.54	0.23	0.13	0.03	0.07
Prk	0.66	0.22	0.06	0.03	0.03	Prk	0.72	0.09	0.02	0.09	0.07
Oth	0.51	0.27	0.13	0.01	0.08	Oth	0.56	0.23	0.05	0.05	0.11

Classification results using both mobile phone and POI data. Results for one and two stage classification algorithms are shown.

Significant portions of the city remain difficult to classify. The limited impact of additional detailed and current point of interest data provides further evidence that official zoning regulations does not guarantee uniform patterns of activity. There is good reason to suspect flaws in the zoning classification itself. Boston is a highly fragmented metropolitan region. The Metropolitan Area Planning Council (MAPC), a state created entity responsible for overseeing regional transportation and economic development includes 101 autonomous cities and towns[3]. Each city and town can set it's own zoning ordinances and classifications. To facilitate regional studies, the state of Massachusetts provides "Generalized codes have been added to make these data useful for regional display."[4] Evidence of these discrepancies is easy to find. For example, the Burlington Mall in Burlington, MA is zoned under "commercial use" according to the MassGIS zoning files. The same file zones the similarly sized Framingham Mall, in Framingham, MA, as "industrial".

Finally, we perform unsupervised learning similar to that of Frias-Martinez et al, 2012 and use k-means clustering to find groups of locations with similar activity patterns. Though we find the same number of groups ($k=4$), the composition of these groups in terms of zoning classification is heavily skewed towards residential. Differences do appear, but they appear in the share of rarer zoning types (Table 7)

[3] http://www.mapc.org/about-mapc

[4] http://www.mass.gov/anf/research-and-tech/it-serv-and-support/application-serv/office-of-geographic-information-massgis/datalayers/massgis-data-zoning.html

Table VII. k-Means cluster composition.

No.	Size	Res	Com	Ind	Prk	Oth
1	4174	**0.85**	0.05	0.04	0.03	0.03
2	5338	0.71	**0.12**	0.07	0.05	0.05
3	3149	0.61	0.11	**0.15**	0.04	**0.09**
4	522	0.77	0.03	0.10	**0.06**	0.04

For each of the $k = 4$ clusters identified by the k-means, unsupervised learning, we calculate the fraction of cells in that cluster that are zoned as a particular class. All clusters contain a large fraction of residential zones, though differences can be found in the share of rarer zoning types. We also report the size (number of cells) of each cluster.

This shows that the unsupervised clustering is dominated by residential cells similar to the supervised classification. Contrary to the supervised techniques given above, however, the underlying reasons for this segmentation are not interpretable for the unsupervised clustering. Additional undocumented clustering based on hierarchical clustering techniques conducted on the daily profiles alone (i.e. ignoring the POI data) led to a larger number of clusters some of which can be given an interpretation (one cluster corresponds to smaller suburban malls). The big bulk, however, does not permit any interpretation in connection with classical land use categories. In this respect the supervised learning approach is more appropriate as it allows the investigation of mismatches between observed usage via mobile phone data and desired usage stated by the official land use classification.

Conclusion:

In this article, we examined the potential of CDR data to predict land usage. We demonstrated that aggregate data shows the potential to differentiate land usage based on temporal distribution of activities. While the absolute activity is dominated by the circadian rhythm of life, eliminating this rhythm reveals subtle differences between the five main land use categories Residential, Commercial, Industrial, Parks and Other. The addition of a temporal dimension to zoning classification may aid strategic planning decisions related to land use.

As the data are available at a high spatial resolution, we investigated the capabilities to infer land use on a fine grid of 200 by 200 meters. We found that supervised classification based on labeled zoning data provides estimated land use classifications that show better accuracy than random assignment. At the same time accuracy is worse than classifying every zone as Residential, the dominant category.

Reasons for this lack of accuracy might be found in the nature of the data used: actual usage might differ from the zoning regulations and Residential is often confused with Parks and Other zones. Omitting residential zones, the classification accuracy for Parks and Other zones greatly increases while industrial and commercial zones classification accuracies are not heavily affected. For rural areas where residential land might not be

fully developed this is plausible. For urban zones the distinction between Residential and Other zones might also be subject to temporal changes as mixed use is prevalent. Finally, analysis of prediction errors reveals that the algorithm fails to correctly classify areas because they have fundamentally different mobile phone activity patterns. This suggests that there may be heterogeneity in how land is actually used, despite its official zoned classification.

To investigate this further, we crawled and incorporated additional point of interest data for the region. These points of interest present a much more detailed and current picture of what type of businesses and services exist throughout the city. However, we find little correlation between POI data and official zoning classifications. Though predictions of zoning types other than residential are improved modestly using POI data, the mapping is far from perfect. A hybrid approach, using both mobile phone and POI data as feature vectors in classification provides the best performance. In total, we are able to correctly classify roughly 70% of locations in a city, provided enough call data is available. Classifying zones other than residential and commercial remains challenging. These results suggest that official zoning classifications fail to sufficiently describe how an area will be used. Further work will explore the relationship between mobile phone activity and points of interest more closely, leaving official regulations behind.

Thus the main conclusion is that the CDR data shows some potential to infer actual land use both on an aggregate level and on a higher spatial resolution. However, zoning data might not be the optimal data source to infer actual land use and hence act as ground truth to guide the supervised learning algorithm. In this respect, our analysis suggests that mobile phone activity may be used to measure the heterogeneity in how space is used that cannot be captured by simple and broad zoning classifications. Moreover, the incorrect predictions made by our algorithm with the addition of alternative data may still be useful. They suggest where updates to traditional zoning maps can be made so as to better reflect actual activity or highlight areas where more planning oversight is needed.

Collectively, these results provide a tool that can be used to augment static measures of population distributions with high-resolution spatiotemporal dynamics. We hope this information will be useful to make effective and efficient choices of locations for both public and private resources. In addition to potential applications, we hope that tools and techniques developed and applied above will prove useful to merging traditional and novel data.

Bibliography:

APPLEGATE, D., DASU, T., KRISHNAN, S., AND URBANEK, S. 2011.Unsupervised clustering of multidimensional distributions using earth mover distance. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’11. ACM Press, New York, New York, USA, 636.

BANISTER, D. 1997. Reducing the need to travel. Environment and Planning B: Planning and Design 24, 3, 437–449.

BREIMAN, L. 2001. Random Forests. Machine Learning 45, 1, 5–32.

CALABRESE, F., DI LORENZO, G., LIU, L., AND RATTI, C. 2011. Estimating Origin-Destination Flows Using Mobile Phone Location Data. IEEE Pervasive Computing 10, 4, 36–44.

CALABRESE, F., READES, J., AND RATTI, C. 2010. Eigenplaces: Segmenting Space through Digital Signatures. IEEE Pervasive Computing 9, 1, 78–84.

CERVERO, R. AND KOCKELMAN, K. 1997. Travel demand and the 3Ds: Density, diversity, and design. Transportation Research Part D: Transport and Environment 2, 3, 199–219.

EAGLE, N. AND PENTLAND, A. 2006. Reality mining: sensing complex social systems. Personal and Ubiquitous Computing 10, 4, 255–268.

FRIAS-MARTINEZ, V., SOTO, V., HOHWALD, H., AND FRIAS-MARTINEZ, E. 2012. Characterizing Urban Landscapes Using Geolocated Tweets. In 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing. IEEE, 239–248.

GEURS, K. T. AND VAN WEE, B. 2004. Accessibility evaluation of land-use and transport strategies: review and research directions. Journal of Transport Geography 12, 2, 127–140.

GONZALEZ, M. C., HIDALGO, C. A., AND BARABASI, A.-L. 2008. Understanding individual human mobility patterns. Nature 453, 7196, 779–782.

JACOBS, J. 1961. The death and life of great American cities. Vintage Books. ACM Transactions on Embedded Computing Systems, Vol. 9, No. 4, Article 39, Publication date: March 2010.

MAAT, K., VAN WEE, B., AND STEAD, D. 2005. Land use and travel behaviour: expected effects from the perspective of utility theory and activity-based theories. Environment and Planning B: Planning and Design 32, 1, 33–46.

READES, J., CALABRESE, F., AND RATTI, C. 2009. Eigenplaces: analysing cities using the spacey¨- y˙time structure of the mobile phone network. Environment and Planning B: Planning and Design 36, 5, 824– 836.

SONG, C., QU, Z., BLUMM, N., AND BARABA´SI, A.-L. 2010. Limits of Predictability in Human Mobility. Science 327, 5968, 1018–1021.

SOTO, V. AND MARTINEZ, E. F. 2011. Automated land use identification using cell-phone records. In Proceedings of the 3rd ACM international workshop on MobiArch. HotPlanet ’11. ACM, New York, NY, USA, 17–22.

WANG, P., HUNTER, T., BAYEN, A. M., SCHECHTNER, K., AND GONZA´LEZ, M. C. 2012. Understanding road usage patterns in urban areas. Scientific reports 2, 1001.

YUAN, J., ZHENG, Y., AND XING, X. 2012. Discovering regions of different functions in a city using human mobility and pois. KDD.