Massachusetts Institute of Technology

3/1/2008

Inferring building and land use types from wireless network consumption.

Andres Sevtsuk
Carlo Ratti

Abstract:

Longitudinal data from wireless networks provides a novel opportunity to characterize places according to their network usage patterns. In a week's course, individual communication antennas have a unique bandwidth consumption and user count, which illustrate the daily activity patterns of their surrounding areas. By grouping areas with similar network usage together, we find that similarities in network usage correspond to similarities in land use and population density. This suggests that the gross functional and demographic indicators of urban areas could be approximated from the mere consumption statistics of wireless networks. We present findings on two different networks: a wireless Internet network on the MIT campus and a cellular phone network in the city of Rome. Using eigenvector and cluster analysis on week-long datasets, we compare usage traces from wireless antennas with empirical descriptions of cell-areas and show a strong correlation, as well as differences between the two.

Introduction

Recent developments in wireless communication have rendered mobile phones and WiFi enabled laptops increasingly popular worldwide. In some countries there are now more mobile phones than people[1] and several major cities have attempted to build citywide WiFi networks. Though the latter remain controversial[2], private WiFi networks are ubiquitous in many cities. As of the end of 2007, there were over 67,000 public hotspots available in the U.S. (JWire, 2007), roughly doubling every year. However, most Wifi networks remain private, with some larger institutions (e.g. universities, large corporate firms) employing thousands of access points.

The ubiquitous adoption of wireless communication networks has also introduced new research directions for spatial analysis. The vast popularity of communication networks makes them attractive for aggregate analysis of people's daily activity patterns. However, limited research is available on this subject, possibly due to the recent emergence of the data and the difficulties associated with obtaining it for academic study. Eagle and Pentland studied the longitudinal behavior of a small group of cell phone users, and proposed an interesting methodology coined "eigenbehaviors" for identifying structure in routinely repeating events. Reades, Calabrese & Ratti used a similar technique on individual cell-phone antennas in Rome to compare signal patterns at characteristic locations of the city with the presence of businesses in their corresponding areas (Reades, Calabrese & Ratti 2007). Their approach suggested that urban neighborhoods could be characterized by the typical weekly network usage patterns, but little evidence was found. This paper expands their approach using larger areas of analysis, and both business and residential distributions as comparisons. Mathew Jull did an analogous analysis of the WiFI network at MIT (Jull, Ratti forthcoming), by comparing WiFi consumption with building types. The present paper takes a very similar methodological approach, deriving eigenvectors from longitudinal data and using cluster analysis to differentiate urban areas with distinctly different network usage patterns. We first focus on the WiFi network at MIT and show how clustering of network usage

[1] CIA World Factbook 2005.
[2] Reality Bites, American cities' plans for ubiquitous Internet access are running into trouble. The Economist, Aug 30th 2007

patterns clearly distinguish between academic, residential and service buildings on campus. We then turn to a city-wide mobile phone network in Rome and find that areas that are similar in network usage are also highly correlated with areas that are similar in demographic as well as business composition. Our findings suggest that the gross functional and demographic indicators of urban areas could be approximated from their consumption statistics of wireless networks.

802.11 WiFi network at MIT

The MIT campus in Cambridge, Massachusetts covers 168 acres, a considerable portion of the city of Cambridge, MA. It consists of more than 190 buildings and houses 10,320 students and 9,414 employees. Since October 2005, full WiFi coverage is available in all academic and residential buildings, as well as most service buildings, employing over 3,000 active wireless access points.

As of 2005, the MIT wireless network infrastructure used the IEEE 802.11 protocol exclusively. The network currently uses three different types of access points with a signal radius from 130 to 350 feet indoors. This allows each antennas to serve one or part of a room, as well as neighboring rooms. Using wireless Internet is vastly popular in the MIT community. On a typical day, approximately 250 000 connections occur on the network (many users are no doubt counted several times in the same space because of their long connections). According to a study by Dal Fiore, Goldman, and Hwang in 2006, 73% of students bring their laptops either every day or some days of the week to campus.

Figure 1 - WiFi access points at their locations on the MIT campus

In the data that was made available to us by Information Services & Technology (IS&T), we observed wireless traffic in 3053 unique access points in 134 buildings on the campus. Data about some access points were not available to us, as they belong to networks operated privately by individual departments. Two large independent networks whose data we lacked belong to the MIT Media Laboratory and the Computer Science and Artificial Intelligence Laboratory (CSAIL). Others we are not able to map, because the GIS data we have about the campus does not yet include some recently constructed buildings. Each of the 3053 unique access points that were available to us were queried at a fifteen minute interval during 14 weeks of the 2006 spring semester, from Monday February 6th (first day of classes) to Monday May 22 (last day of classes). Each measurement showed the current number of WiFi users at each access point, the access points' identifier, and a Unix timestamp indicating the time of the query.

From this 14-week dataset we then obtained the average weekly activity log for each antenna. In order to avoid the activities of any particular week dominating, we determined a pattern for a typical Monday, Tuesday, Wednesday etc. Our final dataset thus showed the number of WiFi users on each access point in a seven day period at 15

minute intervals (672 counts per each antenna). Figure 2 below shows the average weekly user counts across campus.

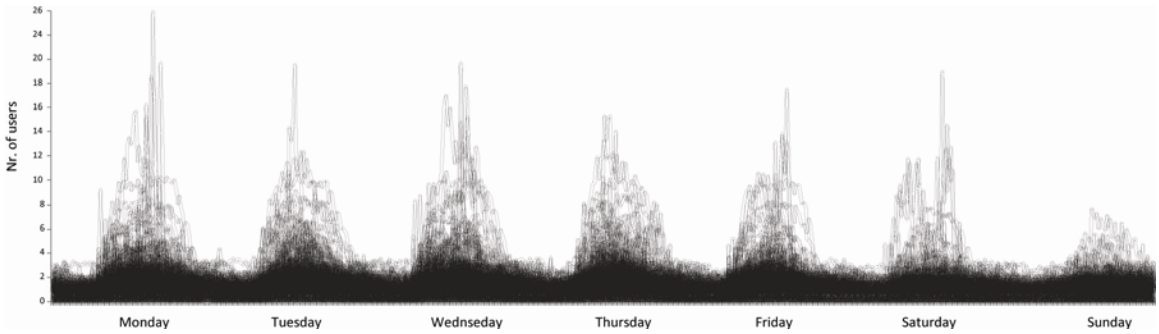

Figure 2: Average weekly WiFi user counts on access points across MIT.

Using principal component analysis on the 672 discrete measurements, we found that the dominant weekly trends in WiFi usage could be captured by the first three eigenvectors, presented in Figure 3 below (a detailed explanation of eigen-decomposition can be found in Reades, Calabrese & Ratti 2007). The first eigenvector captures the prevailing weekly trend with the highest variance across all the access points. Curiously enough it shows flat activity peaks right before midnight, which might be explained by most wireless activity occurring outside of classroom time, during late evening hours. The second eigenvector, uncorrelated with the first one, illustrates a typical daytime usage pattern. The third eigenvector shows more distributed daytime activity with equal popularity on weekdays and weekends.

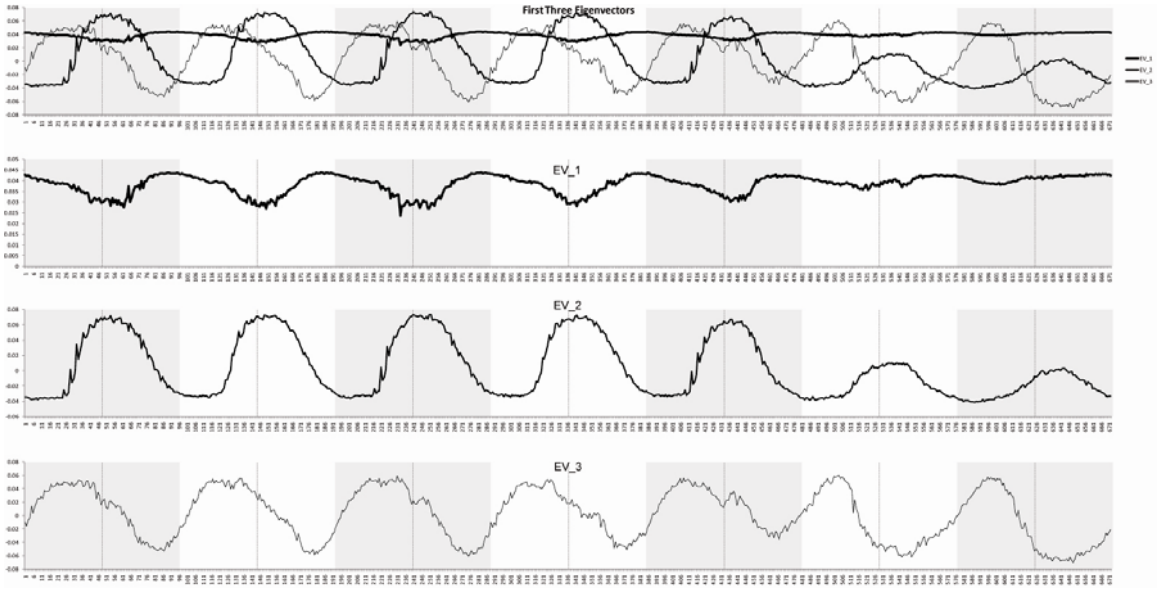

Figure 3: The first three eigenvectors of average weekly connections at MIT.

Using the three first eigenvectors, we performed an average distance cluster analysis on all 3053 access points to determine which groups of antennas had similar usage patterns over a week. The clustering fit analysis suggested that the most coherent structure was formed using 23 clusters. Many of the clusters in this solutions contained only one or a few antennas, which was expected since cluster analysis is known for efficiently distinguishing outliers. We selected the three largest clusters and compared them against the official MIT building type designations (academic, residential, service) in order to see whether similarities in WiFi signal correspond to similarities in building type. We expected the largest cluster to correspond to the most numerous building type on campus- the academic buildings (57 buildings). The second largest antenna cluster was expected to correspond to the second largest building group-residential (51 buildings), and the third largest to service edifices (26 buildings).

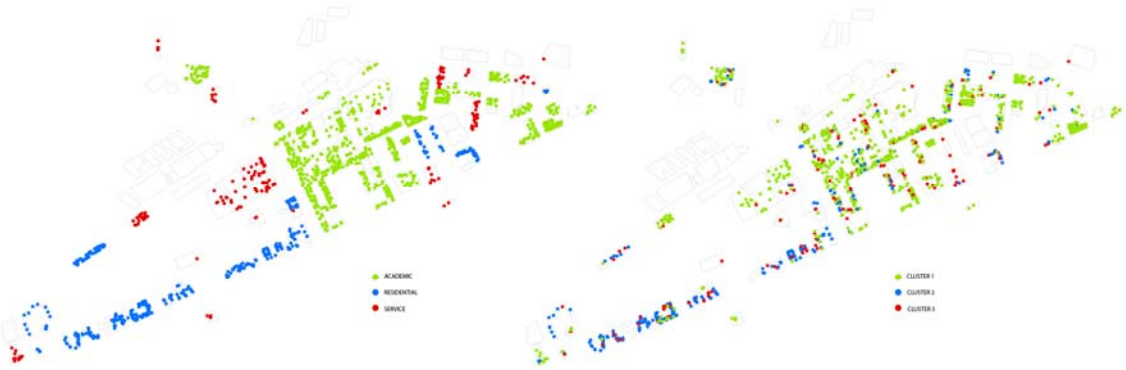

Figure 4: Left, functional designation of MIT buildings. Right, three largest clusters of the weekly WiFi usage.

As figure 4 suggests, a strong correlation exists between WiFi clusters and MIT building types. We found that 67.76% of WiFi cluster designations corresponded to building types they were located in (academic/residential/service). This confirmed that activity on the MIT WiFi network distinctly differs between building types and suggested that activity patterns observed on the WiFi network could be used as proxies to identify the buildings that the antennas are housed in.

Figure 5 shows the average weekly WiFi signal in clusters one, two and three. As expected, the average signal for the largest cluster (1) is dominated by daytime usage. The highest consumption of WiFi on weekdays typically occurs around 1 pm, whereas activity declines steeply after 4pm. Such a pattern seems characteristic to academic working spaces. The typical usage pattern for antennas in cluster two is quite different, more stable throughout the day showing minor peaks around 9am in the morning and 11pm in the evening. Unlike in the previous cluster, weekend usage is quite similar to weekday usage, which well matches our assumption of its residential nature. Lastly, cluster three presents some unexpected results. Rather than showing peak activity occurring during working hours, as we would expect with service buildings, peaks occur instead around 10am and more sharply right before midnight. This suggests that the third cluster might not in fact characterize service spaces, but instead activity spaces that resemble more to residential areas but have higher average usage rates. These could be for instance social working spaces or extracurricular activity spaces. The difference between the third cluster and service spaces also showed up in our correlations: whereas

clusters one and two matched well with academic and residential buildings, cluster three did not match well with service building designations.

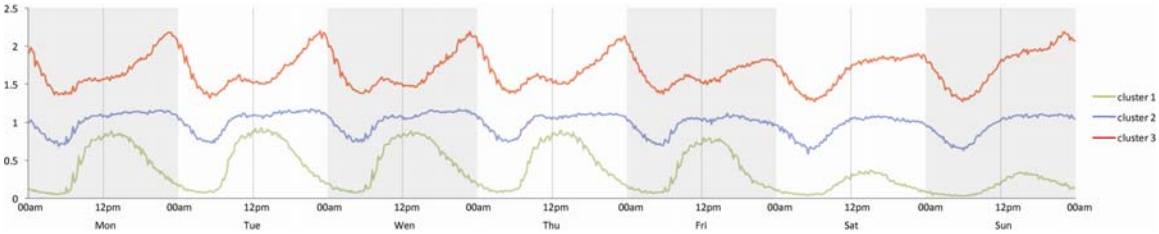

Figure 5: Average weekly trends in clusters 1, 2 and 3.

A closer look at Figure 4 shows that based on the observed WiFi activity, the use of space in the MIT buildings is in fact much more complicated and mixed than the official classification system of building-types designates. We can see from the figure, for instance, that many individual rooms inside academic buildings belong to non-academic clusters since their WiFi activity resembles more to residential or service spaces, than to academic spaces. Similar differences are found in almost all buildings throughout the campus, which implies that a WiFi based categorization of rooms leads to a time-of-use based categorization of MIT spaces, which differs from the traditional *academic, residential and service* designation of spaces.

The findings are thus twofold. On the one hand, a clear correspondence exists between WiFi usage and traditional building designations for academic and residential types. On the other hand, service buildings did not clearly match, which suggests that a use-based classification of spaces can depart from the static designations. Categorizing spaces by network usage reveals a much finer and more complex picture of MIT room types and opens new, interesting research opportunities. For instance, a time-of-use based classification of university spaces could lead greater efficiency in space usage and allow for more optimal classroom scheduling.

GSM wireless network in Rome

Paralleling the analysis of the WiFi network at MIT, we conducted a similar study on a different wireless network on a much larger urban territory in Rome, Italy. Our goal was to test whether a meaningful clustering of weekly network usage, which we saw at MIT, could also be found in a city-wide cellular phone network.

We obtained data for this study from Telecom Italia Mobile (TIM), the largest service provider in the country. Besides TIM, there are three other large service providers in Rome: Omnitel Vodafone, Wind and Blue. TIM is currently market leader in the city, supplying about 40.3% of the share. This constitutes approximately one million users in Rome, less than half of the city's population. However, similarly as the Wifi counts at MIT only showed how many users were actively connected to the network at a specific access point, TIM's data used in this study did not describe the activity of all the registered users in Rome, but only those who were actively engaged in phone calls during the measurement periods in Rome. Alike MIT, we extracted a longitudinal data on 398 antennas within the ring road of Rome over 10 weeks and derived a "typical" weekly usage pattern for each antenna by averaging specific week's values.

Unfortunately the precise number of connections at each antenna over every 15minute measurement period was not available to us. We chose instead to use Erlang measurements, which are commonly used in mobile networks for assessing aggregate traffic in particular cells. An Erlang measure is essentially a use multiplier per unit time. The use of one mobile phone for one hour in a particular cell constitutes one Erlang, whereas the use of two phones for half an hour each also constitutes one Erlang. Since Erlang values are affected by both the amount of calls and each call's duration, then they do not tell us exactly how many users were connected to the network through a particular cell. Instead, the Erlang measurements present the bandwidth consumption at each cell.

Whereas at MIT, our reference category to WiFi clusters were individual building types, a similar approach in Rome was impossible since each cellular antenna had a much larger coverage area, encompassing many different building types and land uses. Figure 6 below shows the coverage area of the 398 mobile cells that we analyzed.

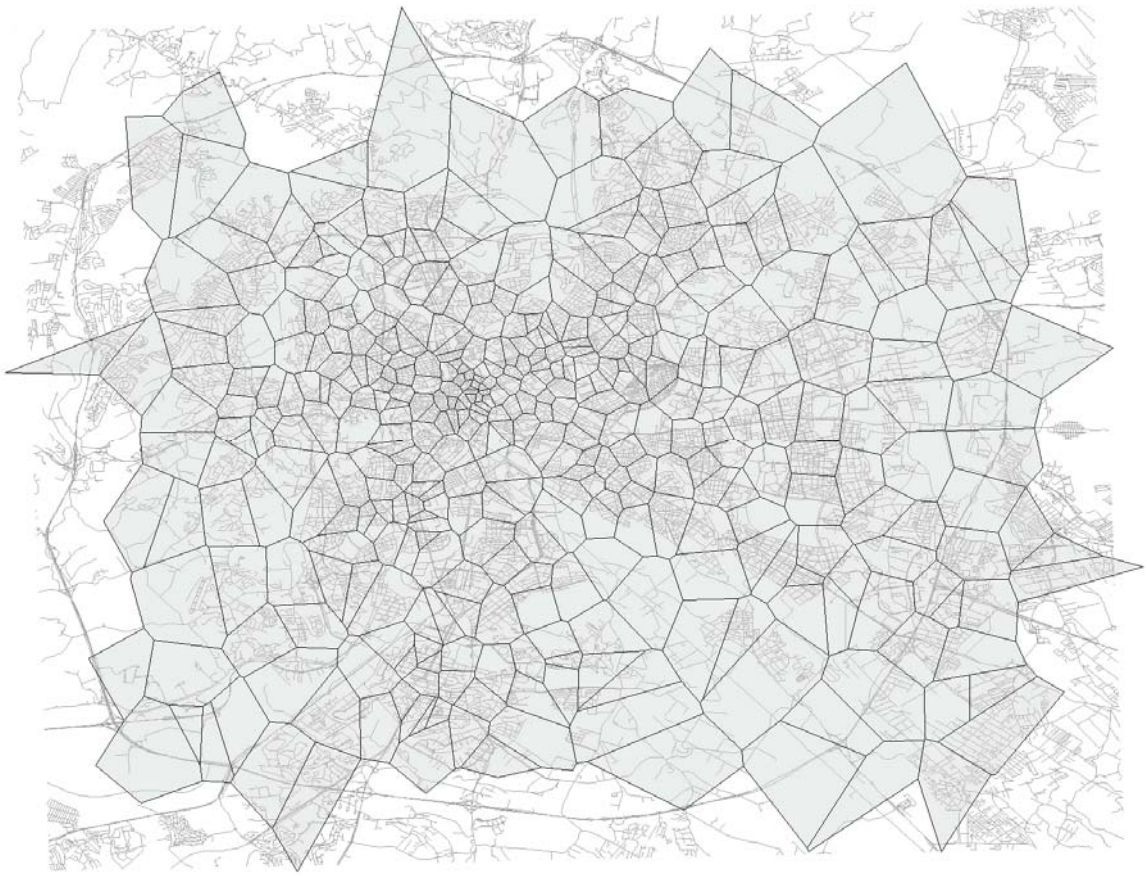

Figure 6: The 398 analyzed network cells in Rome.

Seeking to maintain a similarity with the MIT study despite the scale difference, we made three comparative assumptions in Rome: i) we hypothesized that a city-wide counterpart to the residential buildings of a university campus could be the city's residential density according to census tracts, ii) that the counterpart to service buildings could be the business distribution in the city and iii) a corresponding measure to academic work buildings could be the employment distribution. Our aim was to find out if areas of Rome that resemble to each other in cellular network usage, also resemble in residential distribution, employment distribution and business distribution. Unfortunately we were yet unable to find data for the employment distribution in Rome and reverted to only using residential and business distributions as reference categories. The business distribution was obtained from Yellow Pages in Rome, it therefore does not include all businesses in the city, but only those that were listed. The residential distribution

reflected the year 2000 census. Both reference categories were further broken down into specific demographic and business characteristics, as shown in Table 1 below.

	# people aged 0-9
	# people aged 10-19
	# people aged 20-65
	# people aged >65
	Government Services
	Clothing & Accessories
	Recreation & Hobbies
	Transportation
	Household Goods
	Travel
	Hotels & Accommodation
	Food
	Health Services
	Other Personal Services & Goods
	High-end Retail
	Financial Services
	Bars & Restaurants
	Beauty
	Entertainment & Culture
	Churches & Religious Buildings
	Business Services
Bu	Daily Retail & Services

Table 1: Demographic and business distributions used for clustering cell areas.

Based on these indicators in each network cell, we conducted a principal component analysis and used the first three eigenvectors to cluster cells with similar demographic and business distributions together. The fit test showed the most coherent solution was found with 16 clusters, from which only three contained more that 5 cells each. We thus used these three largest clusters as a reference for the comparison with the three largest Erlang clusters.

Figure 8: Left: Residential population density. Right: Distribution of firms listed on Yellow Pages.

Table 2 describes the typical census and business characteristics of the three different clusters. The average number of people in clusters 1,2 and 3 is 3632.81, 9760.74 and 2969.4 respectively, and the number of businesses, 76.20; 227.61 and 292.55 respectively. Cells in cluster one have the least amount of businesses but more residents than cluster three. These seem to be the relatively lower density residential areas of Rome. Cells in cluster 2, have more residents than the other two clusters, less overall businesses than cluster 3, but more transportation, food, health, beauty and daily retail establishments. Cluster 2 thus seems to represent typical dense residential areas, as shown in the residential distribution map in Figure 8. Cells in cluster three have the highest amount of businesses, most notably hotels, restaurants, business services, government services and specialized retail, but the least amount of residents. This suggests that cluster 3 characterizes dense CBD neighborhoods where few people live.

		Cluster 1	Cluster 2	Cluster 3
	# people aged 0-9	301.75	757.71	203.15
	# people aged 10-19	319.26	738.91	205.32
	# people aged 20-65	2316.03	6065.22	1943.77
	# people aged >65	695.77	2198.90	617.15
	Government Services	2.70	5.19	8.91
	Clothing & Accessories	6.14	19.18	36.18
	Recreation & Hobbies	2.99	9.33	9.73
	Transportation	8.98	25.36	13.64
	Household Goods	7.80	26.71	23.36
	Travel	2.34	6.47	8.09
	Hotels & Accommodation	2.90	3.17	18.73
	Food	4.72	17.57	14.82
	Health Services	7.57	28.13	18.55
	Other Personal Services & Goods	1.66	6.15	6.45
	High-end Retail	3.27	10.03	29.73
	Financial Services	3.24	8.68	10.45
	Bars & Restaurants	8.58	22.19	49.36
	Beauty	5.56	20.85	16.73
	Entertainment & Culture	0.98	2.06	7.73
	Churches & Religious Buildings	3.31	5.22	8.64
	Business Services	1.96	5.67	6.36
Bu	Daily Retail & Services	1.51	5.65	5.09

Table 2: Average demographic and business indicators of clusters 1,2 and 3.

Secondly, we also clustered the weekly Erlang patterns of network cells based on the three first eigenvectors and extracted the three largest clusters for comparison. Figure 9 illustrates the average weekly Erlang values in these three clusters. Network usage in the three clusters differs mostly in intensity, rather than time-of-day variations (Figure 9).

The weekday calling activity in all three clusters occurs predominantly during working hours, with typical peaks spiking at noon (lunch time) and between 6 and 7PM (end of workday). On the weekends there is over all less calling activity and the Sunday peaks occur earlier, around 10am (church time) and later around 8PM. While all the clusters have basically similar peak periods, they differ significantly in intensity: cluster one has the lowest Erlang values and cluster three the highest. These similarities and amplitude differences between the clusters suggest that Erlang clustering picks up more on density differences between urban areas than time-of-day activity differences.

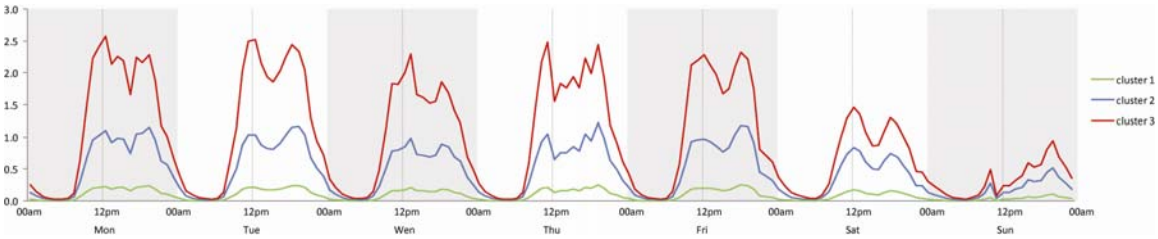

Figure 9: Average Erlang signatures for clusters 1,2 and 3.

A comparison between the Erlang and business/demographic clusters is compelling. Areas with a similar demographic and business distribution correspond to areas with a similar network usage. As shown in Figure 10, 63.14% of the cells that belong to the same Erlang cluster also belong to a similar demographic and business cluster. Comparing demographics and businesses separately with Erlang clusters, we found, however, that Erlang clusters match better with the business distribution (60.05%) than with the population distribution (51.29%). In the absence of employment data it is hard to conclude which socio-economic area characteristics are most clearly related to Erlang clusters. As noted above, we observed from the average signals of the three Erlang clusters that they mainly differed in amplitude rather than cycle. Lower amplitudes were associated with low-density residential neighborhoods, the medium amplitudes with high-density residential neighborhoods, and the highest amplitudes with areas that have a very dense business distribution, but few residents. However, we think that a more accurate neighborhood description could be found if employment data were also available.

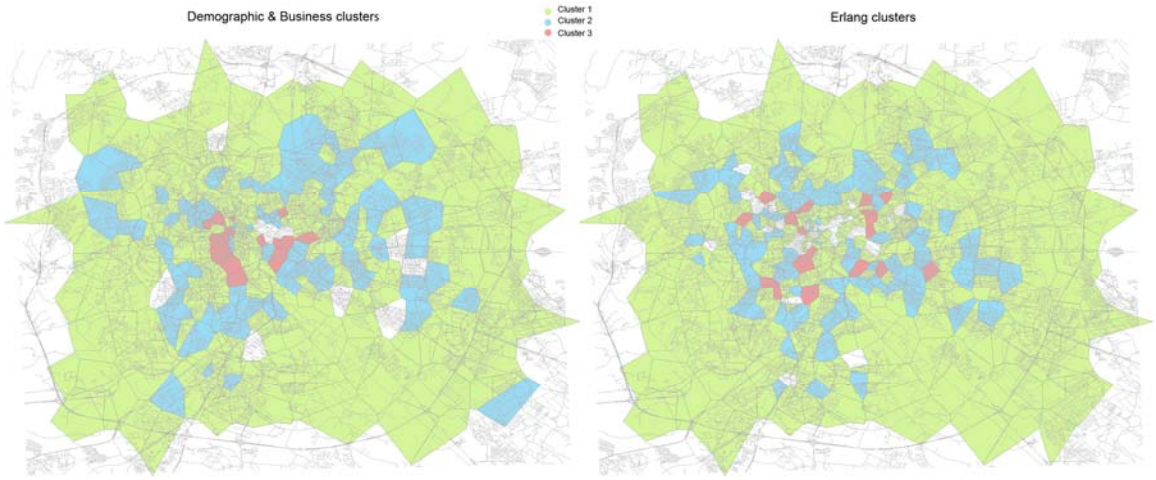

Figure 10: Comparison of business/demographic data clustering (left) VS Erlang data clustering (right).

Conclusion

Using longitudinal data from two different types of wireless communication networks, we have shown a correspondence between antenna usage patterns and the demographic and functional characteristics of location. Despite a generally good match, the correspondence between traditional land categorizations and wireless activity patterns is not exact, and should not be. The traditional categories of space usage typically describe the permanent aspects of places- their land use, the number of businesses, jobs or residents. The wireless network usage, on the other hand, reflects people's temporary presence and communications behavior in these spaces. We have found a clear relationship between the permanent attributes of place and its network usage but also demonstrated that the two differ in several regards.

Analyzing the WiFi network at MIT we clustered access points with similar use patterns into three basic clusters and found that 67% of them matched with the MIT building types. Academic and residential buildings had a very clear network use pattern, but contrary to or initial hopes, the third largest cluster's activity pattern did not match its expected "service" building type. Instead it seemed to characterize usage outside of

business hours, and we proposed a new hypothesis for its interpretation.

Secondly, analyzing the GSM mobile phone network in Rome we also found a strong correlation between areas that resemble in network usage and areas that resemble in their demographic and business composition (63%). The main distinguishing element in network clusters was in amplitude rather than cycle differences. The presence of a wide range of businesses that do not only cater to a local neighborhood, but a citywide population, typically relates to higher network use intensity. Residential density is slightly less determinant of high Erlang amplitudes, but also constitutes and important factor. We think that adding employment data to the comparison could offer more detailed area characteristics to distinguish locations with larger or smaller Erlang amplitudes that constitute difference network clusters.

References:

N. Eagle and A. Pentland, "Eigenbehaviors: Identifying Structure in Routine,"2006; http://vismod.media.mit.edu//tech-reports/TR-601.pdf.

Reades J., Calabrese F. & Ratti C., "Eigenplaces: analyzing cities using the space-time structure of the mobile phone network" 2007, Environment & Planning B (forthcoming)

Jull M, "Urban Rituals, analysis of the MIT campus" 2007, forthcoming.