Exploring Mobility of Mobile Users

B.Cs. Csáji[1,2,*], A. Browet[1], V.A. Traag[1], J.-C. Delvenne[1,4], E. Huens[1], P. Van Dooren[1], V.D. Blondel[1], Z. Smoreda[5]

1 Department of Electrical and Electronic Engineering, University of Melbourne, Australia
2 Computer and Automation Research Institute (SZTAKI), Hungarian Academy of Sciences
3 Department of Mathematical Engineering, Université catholique de Louvain, Belgium
4 Department of Mathematics, Facultés Universitaires Notre-Dame de la Paix, Namur, Belgium
5 Sociology and Economics of Networks and Services Department, Orange Labs, France
∗ **E-mail: Corresponding bcsaji@unimelb.edu.au**

Abstract

Mobile phone data has enabled researchers to examine human behaviour on an unprecedented scale. Many different aspects, such as their social network, temporal dynamics and mobile behaviour have been analyzed, mostly independently one from another. In this article we will explore the connections among various features of human behaviour related to these different aspects. We show that performing clustering and PCA on these features allows us to reduce the dimensionality significantly. In particular, it is possible to approximate the feature space using only 5 meta-features with only 5% error. The most important features seem to be geographical in nature, and in addition we observe that most people spend most of their time in only a few locations. Based on the weekly calling dynamics of these frequent locations we cluster them in order to obtain a better understanding of them. It seems that only home and office locations can be robustly identified, while other types of usages remain more marginal. Finally, we will provide some statistics concerning the geographical spread of these frequent locations, including an analysis of the commuting distance.

1 Introduction

The technologies of information and communication (phones, mobile phones, SMS, e-mails, search engines on the web, etc.) have been an important source of inspiration to sociology, especially these last years. Firstly because those technologies influence the behavior of people, which is a subject of study in itself e.g. [3,5,10]. Secondly because they provide massive amounts of data that can be used to shed some light on various aspects of human behavior. Phone and mobile phone data have been used to study people's social networks, sometimes in conjunction with some other features of the users, such as their gender and age [11], or some geographical positioning information [9].

Recording people's behavior through communication and information technologies is also useful for the providers of those technologies, for example for marketing purposes. Assuming people influence each other over these networks, one of the questions for example is which people are the best "influencers" [6]. The development and use of those techniques have stirred a debate about ethical issues and the protection of privacy [2]. Notwithstanding this critical stance, it might be possible to analyze properly anonymized datasets so as to obtain useful insights into human behaviour, without compromising individual privacy.

More recently, the mobile phone data available to sociologists have been enriched with new information: the position of the antenna used by the caller and the receiver. This allows us to know the approximate position of people when they receive or give a call. This has been used to analyze statistical regularities, or even laws, governing mobility of people in their everyday life [7]. Recently another study used both social network information and geographical positioning data to perform link prediction [20].

The data used in this paper, obtained from a mobile phone company, records all the communications between the mobile phones of this company in Portugal, over a period of 15 months. For every communication, we know the time of initiating and ending it, which user initiated it and the transmitting antennas used by the caller and the receiver. In addition, we also know the exact coordinates (longitude, latitude) of all antennas. All data have been properly anonymized.

We first provide a statistical analysis of the data. More precisely, we will endow every user with a set of various features. We show that the data are highly redundant using PCA and clustering analysis. That is, mobile phone users behaviour might be explained using only a set of five meta-features while incurring only a 5% error.

Observing that the most important features are geographical, we then pay specific attention to the frequent locations of users. By developing a procedure to extract those frequent positions, we observe that people spend most of their time in only a few locations. We then cluster the different calling pattern for each user and each location and we observe that only two types of locations are clearly identifiable, namely an home and an office location. These results match closely independent statistics obtained from the portugese national institute of statistics.

Finally, we analyze more in detail the behaviour of users having exactly one home and one office position. This allows us to predict the number of commuters between different regions of the country using a gravity-like model. More precisely, we observe that two different regimes exist: the first for distances smaller than 150 km (which is exactly the distance between the two larger cities); the other regime applies to larger distances and for that regime there is only a significant effect of the number of offices in the destination region.

2 Behavioral Analysis

In this section we are going to analyze the behavior of the customers based on suitably defined *features* or *statistics*. The features are designed to compactly summarize the calling and geographic behaviour of the users. They allow us to investigate interdependencies between various characteristics such as the durations and the distances of the calls, the size of the movements, the length as well as the frequency of the calls, etc. This can be done e.g. by analyzing the *correlations* between the features. In this section we will also consider *compressing* the data using statistical methods and identifying user classes via *clustering*.

2.1 Preprocessing

Data analysis usually starts with *preprocessing* the raw data. The most important preprocessing that was needed is that we applied a *moving weighted average* type filter on the calling positions of the users, see Figure 1.

This filtering was needed, since the positions of the users were given by the positions of the antennas they were connected to. Moreover, the closest antenna was not always the one serving the call. This can happen when the closest antenna is overloaded or when there is an obstacle between the user and the closest antenna. Therefore, the position of the antenna that served a call in itself should not be treated as the estimation of the user's position.

The filtering was computed as follows. The positions were smoothed independently for all users. Assume that a user has called at times $t(1), \ldots, t(n)$ and the coordinates of the antennas that served the calls were $x(1), \ldots, x(n)$. The smoothed positions of the user, denoted by $y(1), \ldots, y(n)$, can be calculated as

$$y(i) = \sum_{j \in B_\delta(i)} w(j)\, x(j) \qquad \text{with} \qquad B_\delta(i) = \{\, j : |t(j) - t(i)| \leq \delta \,\}\,, \tag{1}$$

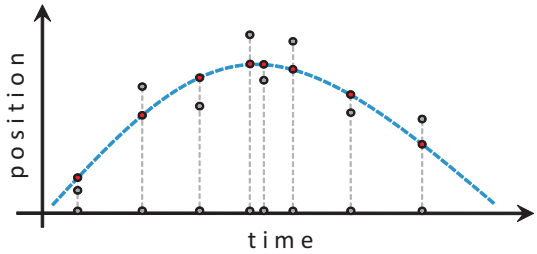


Figure 1. Applying a moving weighted average type filter on the positions, in order to decrease the effects of the noise. The gray dots represent the original positions of the user, while the red dots indicate the positions after smoothing.

where $B_\delta(i)$ denotes the indices of those calls which were initiated or received within maximum distance δ from the current time of the filtering. The parameter δ was chosen to be 30 minutes in our particular case. Positions which were further from the current time of the decision had proportionally smaller weights

$$w(j) = 1 - \frac{|t(j) - t(i)|}{\delta}, \tag{2}$$

where i denotes the current index of the call that should be smoothed.

Besides filtering, we only took those customers into account who had at least 10 calls during the given period (15 months). Moreover, in case of compression and clustering, we have *normalized* (scaled) and *centered* the data. Most of the analysis was performed on 100 000 randomly (uniformly) selected users. We performed (Student) *T-tests* to verify that our results are *statistically significant*.

2.2 Features

We defined 50 *features or statistics*, in order to summarize in a compact way the behavior of the users. Each considered feature measures with a single number one particular aspect of the user's behavior, such as the number of incoming or outgoing calls, the number of people who called or were being called by the customer, the position (coordinates) of the user (mean and deviation), the coordinates of the two most frequently used antennas, the durations of the incoming or outgoing calls (mean and deviation), the distances of the incoming or outgoing calls (mean and deviation), the directions of the incoming or outgoing calls (mean and deviation) and various movement measures (see later).

Even one feature alone can contain much information. For example, by analyzing the average locations of the users we can obtain information on the well-populated areas of the country, e.g., large cities can be clearly recognized as bright spots in the left part of Figure 2.

A novelty of our approach is that we propose, in addition to gyration [7], two more characteristics able to measure the movements of the customers. These measures were: the *diameter of the convex hull* and the *total line segment length*. All of these measures are based on the calling positions of the user and indicate with a single number how much a user has traveled, see part (2) of Figure 2. We take both incoming (received) and outgoing (initiated) calls into account. The order of the positions (calls) is not important for the first two measures, but it is significant for the line segment length.

The *gyration* measures the deviation (in a mean square-error sense) of the user's positions from his average location. It is the same as the *standard deviation* (square root of the second central moment) in statistics. We measure the distance between the calling positions with the Euclidean distance.

The *diameter of the convex hull* measures the maximal distance between any two different position of the user during the given period. The convex hull of a finite set of points contains those points that can

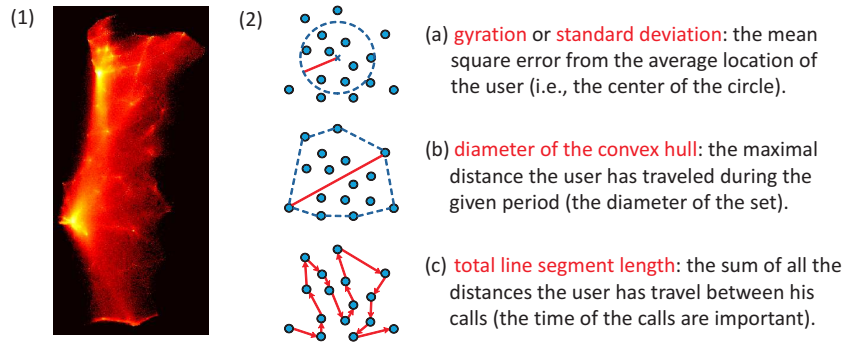


Figure 2. (1) The average locations of the users. Brighter colors indicate that a higher number of users have their average locations in that area. (2) Three different ways to measure the size of the movements of a customer: (a) gyration, (b) diameter of the convex hull and (c) total line segment length.

be achieved by convex combination of the original ones. This set is a polygon, and therefore its diameter is equal to the maximal distance between any two of its vertices, which correspond to calling positions.

Our third movement measure, the *total line segment length*, sums up all the distances between the consecutive positions of the customer's calls. For this measure it is very important that the positions of the calls are ordered according to the time of the calls (earliest call first). We also apply the Euclidean distance, in order to measure the size of the movement between the consecutive calls. Note that for this kind of movement measure the filtering operation (weighted average smoothing) explained earlier could have a large impact.

2.3 Correlation Analysis

After the features were computed for each user, we first analyzed the *interdependencies* between the features by a *correlation analysis*. As mentioned earlier, we took only 100 000 randomly (uniformly) selected users into account. However, the T-statistics demonstrate that this size was suitable for such analysis.

We now briefly recall the basic definition of statistical correlation. Let X and Y be two *random variables* defined on some probability space. Then, the correlation [8] of random variables X and Y is defined as follows

$$corr(X,Y) = \frac{\mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sigma_X \sigma_Y} \tag{3}$$

where μ_X, μ_Y and σ_X, σ_Y denote the *expected value* and the *standard deviation* of X and Y, respectively. Each such variable represent a feature of a user. The expected value and the standard deviation were approximated using the sample.

In some cases the correlations can be better analyzed in a *logarithmic* scale. We defined new (logarithmic scaled) versions of the variables as follows

$$\widehat{X} = \log(X - \min(X) + 1), \tag{4}$$

where "log" denotes natural logarithm and "$\min(X)$" is the minimum possible value of feature X. This quantity was also calculated from the sample, the minimum was taken over all users, see Figure 3 for an illustration of analyzing the correlations of two features on a normal and a logarithmic scale.

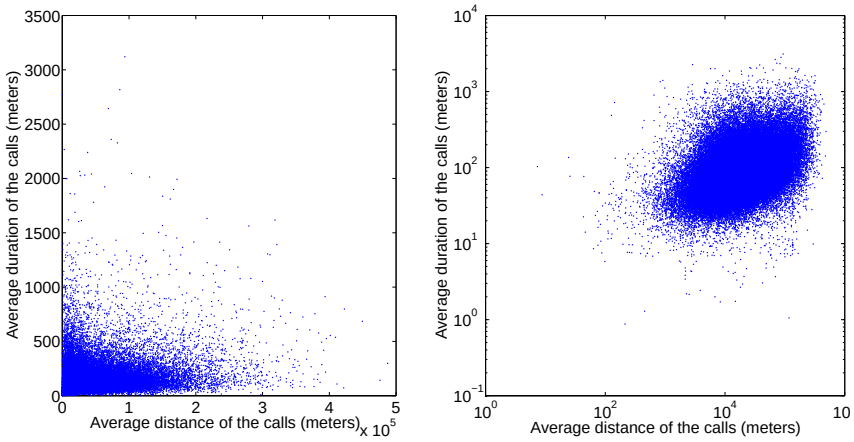


Figure 3. The correlations between the average distance of the calls (x axis) and the average duration of the calls (y axis). The left hand side shows a normal scale, while the right hand side demonstrates the correlations on a log-log scale.

2.3.1 Interdependencies of Features

Table 1 shows some examples of the correlations for 100 000 randomly selected users. It demonstrates the correlations between the features for a 95% probabilty confidence interval (with lower bound *L.Conf* and upper bound *U.Conf*), as well as the correlations between the logarithmic scaled variables. It shows that movement related features show correlations with several (but not all) features. However, it can also be seen that these correlations are, though significant, never too high. This phenomenon can be explained by the fact that we took the user without pre-selection into account. Analyzing the correlations between the features would show higher interdependencies if it would be done only on specific subclasses of users. Such classes could be identified using clustering techniques and highlight particular type of users such as men or women, frequent or infrequent callers, younger or older, blue collar or white collar employees, etc. This is left for future research.

Table 1. Some Examples of Correlation Analysis

Feature A	*Feature B*	*Correl.*	*L.Conf.*	*U.Conf.*	*LogCorr.*
NÂ°.Calls	NÂ°.Callers	0.91	0.91	0.92	0.90
Diam.Conv.Hull	NÂ°.Antennas	0.55	0.54	0.56	0.20
Avg.Duration	Avg.Distance	0.31	0.39	0.32	0.64
NÂ°.Antennas	NÂ°.Calls	0.60	0.59	0.61	0.68
Diam.Conv.Hull	Avg.Duration	0.05	0.03	0.07	0.18
Line.Segm.Len.	NÂ°.Antennas	0.45	0.43	0.47	0.75
Gyration	Std.Dev.Dist.	0.60	0.60	0.61	0.40

2.3.2 Identifying Sub-Models

We note that statistical correlation basically measures the *linear* dependencies between the random variables, viz., features. However, there could be *nonlinear* relationships, as well. It would be promising to investigate them, as well. As a further research direction we highlight that the problem of identifying (potentially nonlinear) sub-models in the structure. For example, if we denote feature i by X_i, which is a random variable, then we could identify relations, such as X_i, X_j and X_k are highly dependent, e.g., determined by a χ^2 probe. Moreover, from features $X_{i_1}, \ldots, X_{i_k}$, we can estimate (predict) up to some error the values of features $X_{j_1}, \ldots, X_{j_m}$. These questions belong to the field of *data mining* and could be handled through *density estimation*, but have many theoretical challenges. We could also apply regression techniques, such as an artificial neural network, in order to identify hidden dependencies [12], but it is nontrivial how to compute them efficiently in a large-scale dataset without trying out all potential combinations of the features (which explodes combinatorially).

2.4 Principal Component Analysis

In the previous section we showed the results of a correlation analysis. Now, we investigate the inter-dependencies of the features assuming (as is often done in practice) that the data have a (multivariate) *Gaussian distribution*. Then, we can use second-order methods to reduce their dimension, because all the information of (zero-mean) Gaussian variables is contained in the covariance matrix. Another reason for second-order methods is that they are computationally simple, often requiring only classical matrix multiplications. To further simplify the discussion we can assume that our variables are centered, which means that they have already been transformed by $x = x_0 - \mathbb{E}[x_0]$, where x_0 are the original non-centered variables. We are going to present a classical second-order method here.

2.4.1 Dimensionality Reduction

Principal Component Analysis (PCA) is widely used in signal processing, machine learning, data mining and neural computing. The basic goal of PCA is to reduce the dimension of the data. It can be proven that PCA is an optimal linear transformation for dimensionality reduction in the mean-square sense [1].

The basic idea in PCA is to find the components $s_1, s_2, \ldots, s_n$ so that they explain the maximum amount of variance possible by n linearly transformed components. PCA can be defined in an intuitive way using a recursive formulation. Define the direction of the first principal component, w_1, by

$$w_1 = \arg\max_{\|w\|=1} \mathbb{E}\left[(w^T x)^2\right], \tag{5}$$

where w_1 is of the same dimension m as the random data x. The principal component corresponds thus to the direction in which the variance of the projection is maximized. Having defined the directions of the first $k-1$ principal components, the direction of the k-th principal component is determined as

$$w_k = \arg\max_{\|w\|=1} \mathbb{E}\left[\left(w^T(x - \sum_{i=1}^{k-1} w_i w_i^T x)\right)^2\right]. \tag{6}$$

The principal components are then given by $s_i = w_i^T x$. In practice, the w_i are simply obtained from the covariance matrix $C = \mathbb{E}\left[xx^T\right]$. The weights, w_i, are the eigenvectors of C that correspond to a given number of largest eigenvalues of the covariance matrix, C. If the data are non-Gaussian or if a mean-square type reduction is not adequate, one needs to use higher-order methods like independent component analysis [1].

Table 2. Compressing the Features by Principal Component Analysis

Variance Kept	*Mean Square Err.*	*Dimen. Needed*	*Compress. Rate*
99 %	1 %	24	48 %
98 %	2 %	13	26 %
95 %	5 %	5	10 %

2.4.2 Feature Reduction

By analyzing the features using PCA, we realized that they are highly *redundant*. Table 2 shows the results of this analysis. It can be seen that if we allow a 1 % (mean square) error in the variance, the number of features could be reduced by more than 50 %. Moreover, if the allowed error is raised to 5 %, we can reduce the number of feature to 5 (from 50), which means that we can achieve a compression rate of 90 %. Consequently, we could build 5 *super-features* by using the linear combination of the original features and by only storing the values of these 5 super-features we could determine the value of any of the 50 original features with a 5 % error (in the mean square sense). This implies that the features have many interdependencies and are highly redundant.

	Feature Name	Importance		Feature Name	Importance
1	Avg.Pos.Y	1,0000	26	Dev.In.Direction.Y	0,5580
2	1st.Antenna.Y	0,9866	27	Dev.Out.Duration	0,5463
3	2nd.Antenna.Y	0,9795	28	SVD.Sigma.2	0,5447
4	1st.Antenna.X	0,8131	29	Avg.Pos.Angle	0,5445
5	Avg.Pos.X	0,7979	30	Dev.In.Distance	0,5332
6	Avg.In.Direction.Y	0,7824	31	Dev.Out.Dir.Angle	0,5332
7	2nd.Antenna.X	0,7682	32	Diam.Conv.Hull	0,5297
8	No.Antennas	0,6943	33	Avg.In.Distance	0,5185
9	Avg.In.Dir.Angle	0,6694	34	Avg.Out.Dir.Angle	0,5185
10	No.Contacts	0,6316	35	Avg.Out.Distance	0,4906
11	Dev.Out.Direction.Y	0,6298	36	Avg.Pos.Length	0,4906
12	No.In(coming).Calls	0,6207	37	No.Antennas.50%	0,4791
13	Dev.Out.Distance	0,6169	38	Dev.Out.Direction.X	0,4611
14	Dev.Pos.Length	0,6169	39	Dev.Pos.X	0,4388
15	No.In.Callers	0,6168	40	Avg.In.TimeP	0,4351
16	No.Antennas.90%	0,6041	41	Dev.In.Direction.X	0,4311
17	Avg.In.Duration	0,5926	42	Dev.In.Dir.Angle	0,4210
18	No.Out(going).Calls	0,5922	43	Line.Segm.Length	0,3911
19	Dev.Pos.Y	0,5866	44	Avg.Out.TimeP	0,3774
20	Avg.Out.Direction.X	0,5812	45	SVD.Sigma.1	0,3763
21	Avg.Out.Duration	0,5777	46	Dev.In.TimeP	0,3743
22	Dev.In.Duration	0,5771	47	No.Out.Callers	0,3664
23	Avg.Out.Direction.Y	0,5764	48	Dev.Out.TimeP	0,3575
24	Avg.In.Direction.X	0,5657	49	Distance.1st.2nd.Ant	0,2852
25	No.Antennas.75%	0,5582	50	Dev.Pos.Angle	0,2842

Figure 4. The importance of the features according to PCA.

2.4.3 Important Features

As we saw in the previous section, the features are highly redundant and their number can be radically reduced if we allow their linear combinations as features (which combinations are called "super-features"). However, this may not always make sense, since, the weighted sum of, e.g., the number of calls the user made and the average duration of their calls, does not have a natural interpretation. Therefore, we did not replace the 50 original features with the newly achieved super-features, but we kept the original ones. Still, the super-features were proven to be very useful not only for showing that the original features were highly redundant, but also for ordering them according to their *importance*.

The importance of the features were determined as follows. It is known that PCA produces a set of orthogonal vectors $s_1, \ldots, s_n$, the *principal components*. They point toward the directions with maximal deviations. Note that the principal components are *not normalized* and that their length is important. Each original feature can be identified with an element of the standard canonical orthonormal basis. For example, Feature 1 can be identified with $e_1 = \langle 1, 0, \ldots, 0 \rangle^{\mathrm{T}}$. Now, the *importance* of Feature i can be defined as the *norm* of the transformed vector e_i after we have changed the basis to $s_1, \ldots, s_n$. In particular, we have applied the *max* norm and scaled it so that the most important feature has norm 1.

The list shown in Figure 4 presents the achieved order of the features. It can be observed that the most important features are geographic in nature, such as the average position of the user and the coordinates of the two most used antennas. This indicates that the locations of the customers and their calls are very important characteristics and should be seriously taken into account.

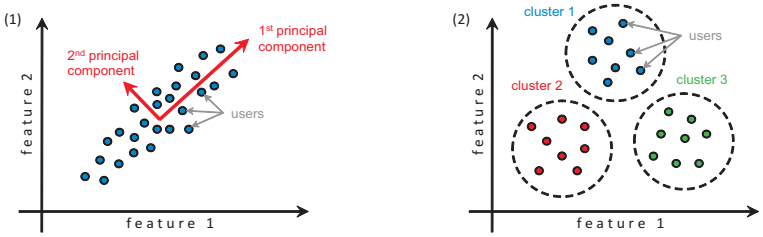


Figure 5. Processing by (1) principal component analysis and (2) clustering.

2.5 Cluster Analysis

After analyzing the data by using PCA, we applied *clustering*, in order to identify typical user classes based on their calling behaviors. The *subtractive* clustering method was used [4], which is a variant of the classical mountain method. An advantage of this method is that it can also identify the number of clusters needed. Figure 5 illustrates the processing of the data with clustering as well as with principal component analysis.

The application of subtractive clustering on the *normalized* data of 100 000 uniformly selected customers resulted in only 5 clusters. We point out that we have applied 0.5 as the initial radius for the clusters. Each such cluster is identified with its *central* element (a vector of feature values) and its range of influence.

2.5.1 Important Features

A natural question which arises is which are the *important* features for the clustering. More precisely, which are the features which separate the clusters the most ? As in the PCA section we ordered the features according to their importance, but instead of using the principal components as basis vectors, we used the vectors of the cluster centers as a new basis for the dominant feature subspace. The results of this ordering are presented in Figure 6. The achieved importance ordering indicates that, similarly to PCA, location and movement related features are also important characteristics for clustering.

Notice also that there is a clear difference between the x and y coordinates, while it is expected for an elongated country like Portugal.

	Feature Name	Importance
1	1st.Antenna.Y	1,0000
2	2nd.Antenna.Y	1,0000
3	Diam.Conv.Hull	1,0000
4	Avg.Pos.Y	1,0000
5	Avg.In.Dir.Angle	0,4914
6	Dev.In.Distance	0,4729
7	Dev.Out.Dir.Angle	0,4729
8	1st.Antenna.X	0,4620
9	2nd.Antenna.X	0,4597
10	Avg.Pos.X	0,4381
11	Dev.Out.Distance	0,3616
12	Dev.Pos.Length	0,3616
13	Dev.In.Direction.Y	0,3548
14	Dev.Out.Direction.X	0,3296
15	Dev.In.TimeP	0,3210
16	Dev.Out.Direction.Y	0,3184
17	Dev.Pos.X	0,3181
18	Dev.In.Direction.X	0,3147
19	Avg.Pos.Angle	0,3050
20	Dev.Pos.Y	0,2706
21	Dev.In.Dir.Angle	0,2659
22	SVD.Sigma.1	0,2440
23	Dev.Pos.Angle	0,2334
24	Avg.Out.TimeP	0,2051
25	SVD.Sigma.2	0,1771

	Feature Name	Importance
26	Avg.Out.Dir.Angle	0,1547
27	Avg.In.Distance	0,1547
28	Dev.In.Duration	0,1407
29	Avg.Pos.Length	0,1393
30	Avg.Out.Distance	0,1393
31	Avg.In.TimeP	0,1369
32	Dev.Out.TimeP	0,1126
33	No.Contacts	0,1111
34	Dev.Out.Duration	0,1034
35	No.In.Callers	0,0966
36	Distance.1st.2nd.Ant	0,0959
37	No.Antennas.90%	0,0872
38	No.Antennas	0,0829
39	Avg.In.Duration	0,0679
40	No.In(coming).Calls	0,0669
41	No.Antennas.75%	0,0648
42	Avg.Out.Duration	0,0560
43	Avg.In.Direction.Y	0,0502
44	No.Antennas.50%	0,0391
45	Avg.In.Direction.X	0,0388
46	Avg.Out.Direction.Y	0,0233
47	Line.Segm.Length	0,0164
48	Avg.Out.Direction.X	0,0158
49	No.Out(going).Calls	0,0155
50	No.Out.Callers	0,0017

Figure 6. The importance of the features according to clustering.

3 Frequent Locations

In the previous section we analyzed many different features, and concluded that the most important ones were related to geography. Additionally, we observe that most people spend most of their time in only a few locations. In this section we will focus on characterizing these frequent locations by analyzing the weekly calling patterns. Once the frequent locations have been characterized we will analyze them more in-depth.

As explained in section 2.1, the data are noisy, and often various antennas could be used for making a call from the exact same position. Hence, for frequent locations such as home and office, often multiple antennas will be involved, and we will first develop a method to estimate which antennas are relevant to characterize each single frequent location.

Once we have extracted the frequent locations and estimated a more precise position, we will provide various statistics. We will estimate the time people spent at work and at home. We will also characterize the different combinations of frequent locations (multiple 'homes' or 'offices') among people. We will also estimate the geographical density of homes and offices, and compare our estimates to indepent statistics. Finally, we will analyze commuting distances based on our estimates of home and office positions.

3.1 Frequent Location Detection

To start analyzing frequent locations, we will only consider users that make enough calls[1] so that we can actually say that somebody is frequently in a certain location. Otherwise, when somebody makes only a few calls, it is doubtful whether any location can be called "frequent". So, we select only users that make at least 1 call a day on average, and make consecutive calls within the next 24 hours 80% of the time. This last constraint is to induce a certain regularity of the user, and to exclude users who have a extremely high bursty behaviour [7]. Out of this selection, we take a random sample of 100.000 users.

[1]In this section, we include both call and text messages because we want to maximize the information about antenna usage. We usually talk about calls, but text messages are included in this analysis as well

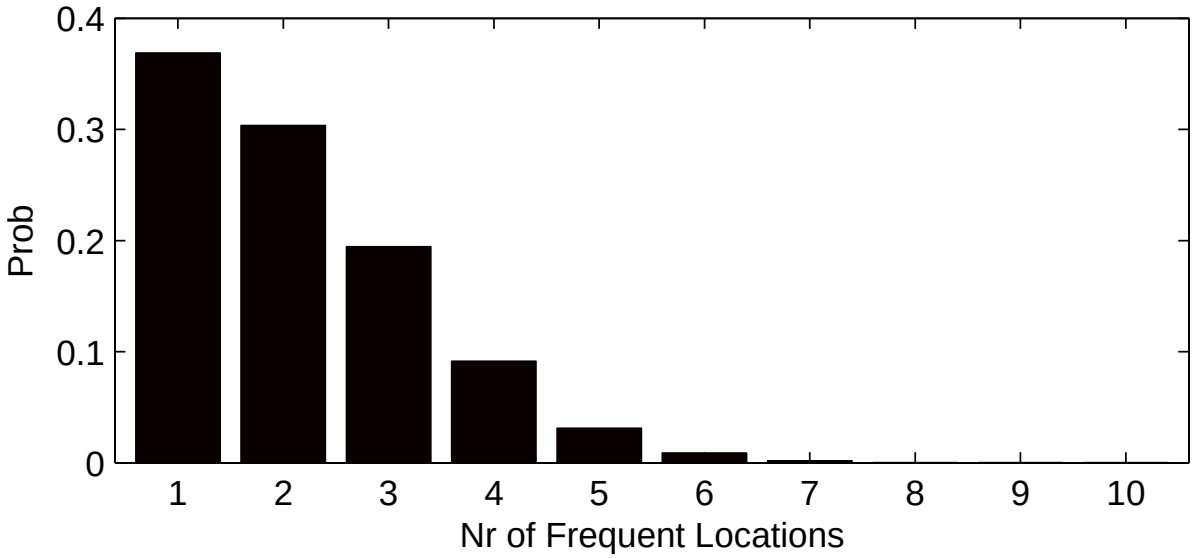


Figure 7. Histogram of the number of frequent location per user

For detecting frequent locations, it makes sense to start the analysis with the most frequent antenna (MFA). However, as stated earlier, often various antennas could be used for calls made from the same position, due to load balancing or random noise on the signal such as reflections on buildings, etc. So some antennas around the MFA might well have been used to also serve that frequent location. In order to deal with this we therefore have to group sets of antennas that are relatively close.

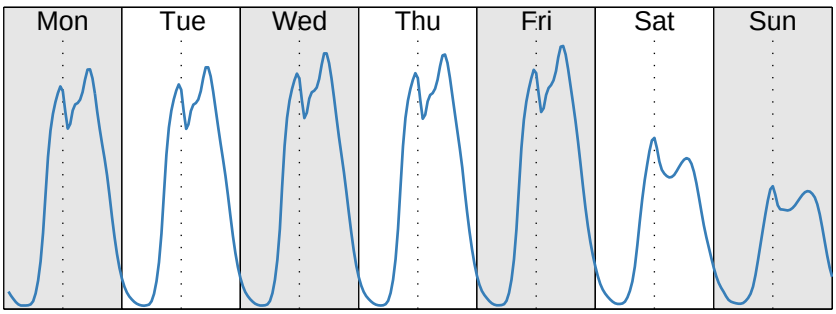
We will do so using the Delaunay neighbourship of antennas. This is based on the Voronoi tessalation, which provides a partition of the space such that each point in space is assigned to the closest antenna. In other words, each Voronoi cell covers the set of points which are closer to that antenna than to any other. Based on the Voronoi tessalation, one can create a dual graph such that all the neighbours of an antenna are composed of those antennas in adjacent Voronoi cells. These neighbours are called the Delaunay neighbours.

Finally, we group antennas around the most frequent antennas, based on Delaunay neighborship. More precisely, we define the Delaunay radius to be the largest distance between an antenna and any of its Delaunay neighbours (this is later used in the estimation of the position, see Section 3.3.2 for more information). We then include all antennas around the MFA that are within twice this radius[2].

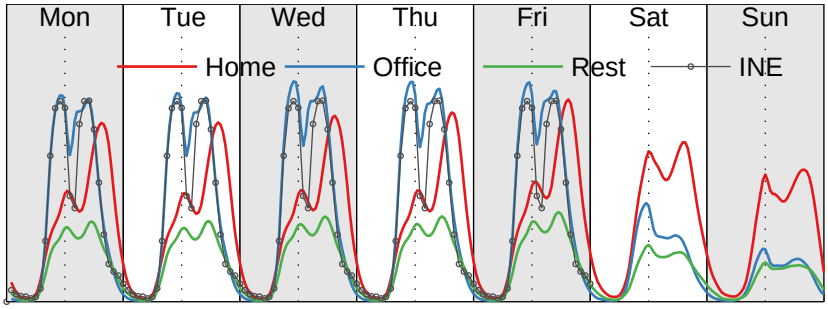
After having treated the first MFA, and merged the surrounding antennas, we move on to the remaining MFA. We then repeat the same procedure as described above, and keep iterating until a set of antennas around an MFA represents less than 5% of the calls. We repeat this for every user in our selection, so that in the end we have a number of frequent locations for all users.

When running this procedure, we observe that the average number of frequent locations per user is about 2.14 and 95% of the population has less than 4 different frequent locations, as illustrated in Figure 7. This implies that only the top 3 or 4 frequent locations suffice for predicting relatively well the position of users most of the time [21]. There are however, quite a substantial number of users that only have one single frequent location, usually an office or home location, as we will see later on. This could be the result of having business and private phones, with the one being (almost) exclusively used

[2]We observe that taking twice the Delaunay radius produces an error of less than 0.1% for estimating positions, see Section 3.3.2 for more information



(a) Average calling dynamics



(b) Calling dynamics of the cluster

Figure 8. Weekly dynamics on average (a) and for the three detected clusters (b) using k-means clustering: home, office, and the remainder, compared to the independent statistics of the *Instituto Nacional de Estatistica* (INE). Using more clusters yields relatively similar results. The dotted line indicates noon of every day.

at work, and the other only at home.

3.2 Clustering weekly calling patterns

In general, there are two clearly identifiable periodic dynamics in the usage of mobile phones: a daily cycle, and a weekly cycle, as illustrated in Fig. 8. The daily dynamic follows largely the human circadian rythm, with a clear drop during the night, a gradualy increase in activity in the morning and decrease in the evening, with a small dip around lunch time. The weekly dynamic is related to the workweek, with different behavior in the weekends as opposed to the workdays.

Based on the procedure described in the previous section, we aggregate all the calls made using antennas associated to a frequent location. Since we have the time stamps of each call, we know at what times certain frequent locations are used. So, we obtain an idea of the temporal dynamics. The description of the temporal dynamics at a weekly scale seems to be especially suitable for this analysis. So, we divide the week into 168 hours, and then aggregate the temporal dynamics of the whole period. This results in a 168 long vector per frequent location with the calling frequency per each hour in each entry.

Based on the aggregated call vectors for all frequent locations, we perform k-means clustering. We ran

the k-means clustering for $k = \{2, \ldots, 10\}$ in order to see what possible types of usage we could discern. We observe that using $k = 3$ clusters leads to clear results as displayed in Figure 8. More precisely, we observe that one cluster clearly represents a pattern related to work. During the weekdays, we observe an increasing usage of mobile phone during the morning, a small drop around noon, supposedly representing lunch break, followed by a decreasing usage from around $6p.m.$ until the evening. During the weekend, these antennas are used far less. This behaviour is in excellent agreement with independent statistics from the *Instituto Nacional de Estatistica*[3] (INE) in terms of time spent at work, as shown in Figure 8b. Another cluster represents a usage that seems to be more associated to an home position. The usage of the antennas is lower during the day and the maximum is attained during the evening. The same antennas are also used more during the weekend. Finally, the third cluster seems to contain simply locations that do not follow the same dynamics of the previous two clusters, and follows the more general dynamic displayed in Figure 8a.

We observe that when using more then three clusters, they tend to be very similar to the results shown here. We expected that we could identify different types of usage, such as students having a different rythm from working people, or weekend houses that show no activity during the week, but we could not find them. Of course, such differences do exist, but they seem to be marginal when compared to the clear cadence of home and office. Hence, there does not seem to be an identifiable type of usage different from home and office. On the other hand, when using only two clusters, this obfuscates the result, and the separation between home and office positions is less clear.

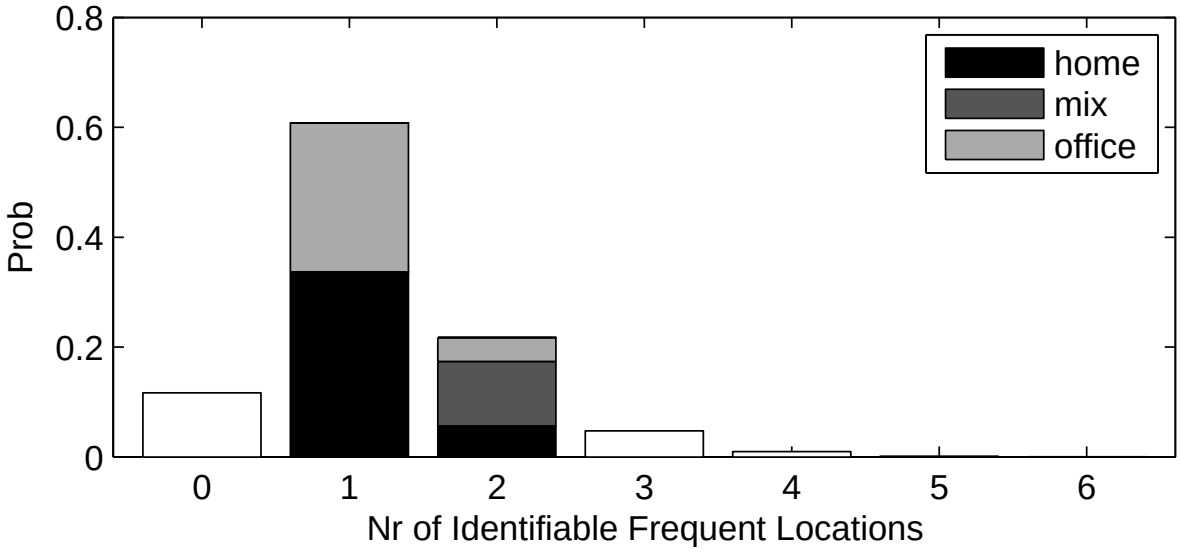


Figure 9. Histogram of the number of frequent location per user: a) all positions, b) only identifiable positions with a visual separation between only "home", only "office" or a mix between "home" and "office".

From all the frequent locations, about 60% of them are identifiable (classified as "home" or "office"). We observe that people tend to have less than 2 identifiable position as depicted in Figure 9. The majority of users have only 1 identifiable location, which is by definition either home or office. For the users with two identifiable locations, over 50% have both a home and an office, while the rest has either two homes or two offices.

[3] `http://www.ine.pt`

# of Home	# of Office	# of unclassified	Percentage
	1		16.8
1			16.5
1		1	9.1
1	1		6.6
	1	1	6.0
1		2	5.0
		1	3.5
1	1	1	3.5
		2	3.4
	1	2	2.7

Table 3. The 10 most frequent location combinations.

More detailed, we can examine all the different combinations we observe, and the top 10 most frequent combinations are displayed in Table 3. About 16% of the users has either a single home location or a single office location, whereas only 3.5% has a single unidentified location. For users with two frequent locations, the most common combination is one home location and one unidentified location, while 6.6% of the users show the more natural combination of one home location and one office location. About 85% of the user have at most 1 home and/or 1 office location, and about 12% of the users have exactly 1 home and 1 office location (and possibly multiple unidentified locations).

3.3 Estimating the position

3.3.1 Basic model

Since the frequent locations are represented by a multitude of antennas, we propose a model to estimate a more precise position (i.e. to determine more precisely the home and office). We consider a simplified version of the model proposed in [15], also used in [16]. The underlying idea is that users connect to antennas that have the highest signal strength, which is not necessarily the closest.

We start out by estimating the total signal strength of an antenna i at a certain position x. We assume, similar to [15], that the total signal strength consists of three components: the power of the antennas, the loss of signal strength over distance, and some stochastic fading of the signal, due to various scatterings and reflections in the environment. More specifically, we will use the following:

- The position of antenna i is X_i.
- The power is denoted by p_i, and we assume it to be constant and equal for all antennas, since we unfortunately have no information regarding the power of the antennas. So $p_i = p$ for all i.
- The path loss at position x for antenna i is modeled as:

$$L_i(x) = \frac{1}{\|x - X_i\|^\beta} \tag{7}$$

  where β is some parameter indicating how quickly the signal decays.

- The so-called Rayleigh fading can be modeled as a multiplication of the signal strength by an exponential random variable with mean 1 [14]. Let therefore R_i be an exponentially distributed random variable with mean 1 for antenna i. The cumulative distribution function (cdf) for R_i is then

$$\Pr(R_i \leq r) = F(r) = 1 - e^{-r}, \tag{8}$$

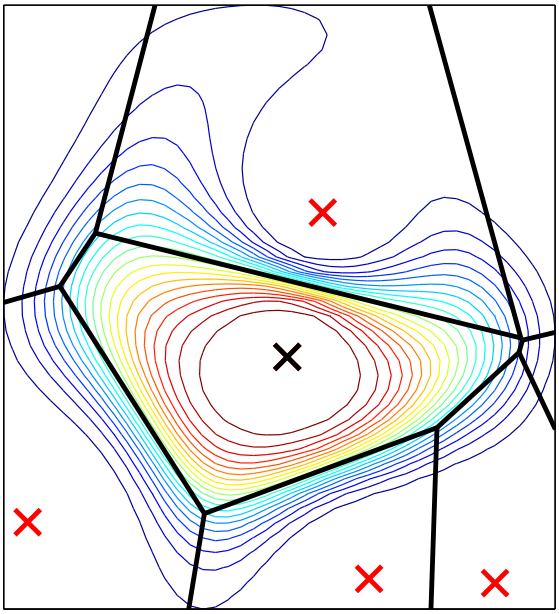


Figure 10. Probability density $\Pr(a = i|x)$ (represented by the level curves) for a particular antenna i (the central black 'X'), showing neighboring antennas (the red 'X's) and the local Voronoi tessellation (in dark lines). The density can be seen as a smoothened Voronoi tessellation, where there is also some (small) probability to be connected to antenna i when the user is in another Voronoi cell.

and the probability density function (pdf) is

$$f(r) = \frac{dF(r)}{dr} = e^{-r}. \tag{9}$$

Furthermore, we assume all R_i to be independent.

The signal strength $S_i(x)$ of an antenna i at location x is then defined as

$$S_i(x) = pL_i(x)R_i, \tag{10}$$

and we define the probability that a user, at position x, connects to antenna i, $\Pr(a = i|x)$, as the probability that the signal strength of antenna i is larger than the signal strength of any other antenna, such that

$$\Pr(a = i|x) = \Pr(S_i(x) > S_j(x), \quad \forall j) = \prod_j Pr(S_i(x) > S_j(x)) \tag{11}$$

Depending on the value of the parameter β, this probability can be seen as a smooth version of the Voronoi tessalation, as displayed in Figure 10.

3.3.2 Antenna neighborhood

As mentioned in the previous section, the probability that a user connects to a specific antenna depends on the position of the other antennas. The whole set of antennas $\mathcal{X}$ can be rather large, thereby slowing

down the method. By using a local approximation of the neighborhood, we might be able to speed things up, without affecting the results too much.

The idea of making a local approximation is tied to the decrease of probability to be linked to an antenna that is far away: only some antennas around a certain position are relevant. It seems then natural to construct local neighbourhoods of antennas, so as to make the method more efficient without inducing a large error.

We define the neighborhood $\mathcal{X}_i$ and domain $\mathcal{D}_i$ of antenna i to consist in the smallest circle enclosing at least all the Delaunay neighbors (and possibly more). As mentionned previously, the Delaunay neighbors are the antennas that are located in adjacent Voronoi cells. More precisely,

- for each antenna, we select all Delaunay neighbors and then select the maximum distance between the antenna and any of these neighbors, or

$$\rho_i = \max\{d(X_i, X_j) | j \text{ Delaunay of } i\}, \tag{12}$$

  where $d(X_i, X_j)$ is the distance between antenna i and j.

- we then define the domain

$$\mathcal{D}_i = \{x | \|x - X_i\| \leq r\rho_i\}, \tag{13}$$

  as the region within a radius $r\rho_i$ where r is a scaling factor. We observe that choosing $r = 2$ lead to a error of less than 0.1% in the computation of $\Pr(a = i|x)$ compared[4] to using the entire set $\mathcal{X}$;

- finally, the set of neighbors[5] is taken as all antennas within that region

$$\mathcal{X}_i = \{j | X_j \in \mathcal{D}_i \text{ for } j \in \mathcal{X}\}. \tag{14}$$

  Note that this set contains at least all the Delaunay neighbors but may also contain other antennas.

Finally, we compute the probability (11) as

$$\Pr(a = i|x) \approx \prod_{j \in \mathcal{X}_i} Pr(S_i(x) > S_j(x)) \tag{15}$$

leading to a large reduction of the computational time required.

3.3.3 Maximum Likelihood Estimation

We will now use the model explained above to estimate more accurately the position of a frequent location. For each frequent location, we know the number of calls k_i made using antenna i. The probability there were k_i calls using antenna i given position x is then $\Pr(a = i|x)^{k_i}$. Hence, the log likelihood of observing the call frequencies k, for the antennas in $\mathcal{X}_f$ where f is the MFA of a frequent location, for a certain position x is

$$\log \mathcal{L}(x|k) = \sum_{i \in \mathcal{X}_f} k_i \log \Pr(a = i|x). \tag{16}$$

The Maximum Likelihood Estimate (MLE) $\hat{x}$ of the position for a frequent location is then given by

$$\hat{x} = \arg\max_x \log \mathcal{L}(x|k). \tag{17}$$

For finding the MLE, we employ a derivative-free optimization scheme, since the gradient of the likelihood function is costly to evaluate. In particular, we used the Nelder-Mead algorithm [17] with

[4] average error on 1000 random points

[5] in order to deal with antennas near the border of the country (where the Delaunay neighbours can be quite far), we take into account the border, and create a slightly different neighbor set.

the weighted average position of the antennas associated to the frequent location as initialization. The average distance between the average position of the antennas and the MLE is 1.7 km and reaches at most distances up to around 35 km. This shows that although simply using the average position is a reasonable approximation, there can be significant differences.

3.4 Results

We will now analyze the results from the position estimation. First, we will show results about geographical distribution of frequent locations around the country and compare our results to independent statistics. Then, we will analyze the commuting distance, i.e. the distance travel between home and office, and fit a model for the number of commuters between counties.

3.4.1 Estimating population density

We can use the position estimates of the frequent locations to analyze the population distribution throughout the country. Using the county[6] level data, we count the number of homes for each county. We can then compare this to independent data obtained from the *Instituto Nacional de Estatistica*[7] (INE). Both results are displayed in Figure 11, from which we can see there is a good correspondence between the population size for each county and the estimate we make. The correlation between the two is 0.92. Hence, we can estimate quite well the population based on the mobile phone data. We also provide a more accurate density plot of the frequent locations, in which the cities light up more clearly compared to the county maps. Especially if we compare this to the distribution of the average position of users during the whole period (as displayed in Fig. 2), we see that the distribution of frequent locations follows much more the geography of the country. Probably the average position is distorted by commutes, which we will discuss now.

3.4.2 Commuting distance

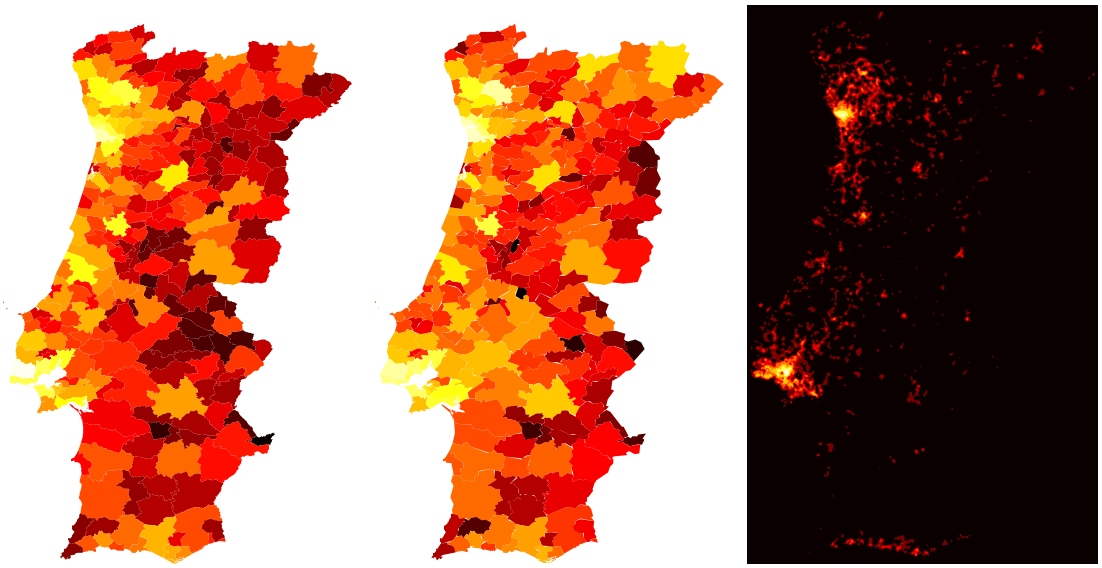
Since we have office and home positions, we can use these to estimate the commuting distances. Of course, when people have more than one home or one office, multiple commuting distances could be calculated, but it would be unclear which distance would be the "correct" one. Therefore, we decided to select only users who have exactly one home and one office (and possibly some unidentified frequent locations), which amounts to some 12% of the users. So, every user in this selection will have exactly one commuting distance. These commutes have been plotted in Figure 12, with smaller distances in more brighter colours. Two things stand out in this map. First, we clearly discern the two largest cities of Portugal: Porto and Lisbon. Secondly, most of the cities seem to attract mostly people living in the immediate surroundings. So, most of the commutes seem to go to a nearby city.

Looking at the distribution of commuting distances, depicted in Figure 13, the presence of Porto and Lisbon seems to affect it. We can discern two different regimes, one regime with commuting distance less than 150 km, and the other regime with larger distances. This coincides with the distance of about 300 km between Lisbon and Porto. In fact, most of Portugal is within 150 km of one of these two cities. This suggests that most people tend to work no further away then the closest largest city, i.e. it is unlikely that people living close by to Porto will end up working in Lisbon. The first regime with distances less than 150 km can be reasonably well-fitted using a log-normal distribution with parameters $\mu = 2.35$ and $\sigma = 0.94$, as is displayed in Figure 13.

A common model when analyzing commuting distance is the so-called gravity model [18, 19], named after its resemblance to the ordinary law of gravity. This model formulates the number of trips w_{ij} made between two locations i and j as proportional to the population at the origin P_i and at the destination

[6] Technically, we used the NUTS-3 data, which for Portugal consists of groups of municipalities

[7] `http://www.ine.pt`



(a) Population Distribution (b) Estimated number of Homes (c) Distribution of frequent locations

Figure 11. (a) distribution of population size for each county throughout the country (statistics from INE); (b) estimated number of homes per county; and (c) the distribution of all frequent locations. Brighter (hotter) colors indicate higher population sizes.

P_j, and with some decay depending on the distance d_{ij} between i and j. More precisely, the model is formulated as

$$\hat{w}_{ij} \sim \frac{P_i^\alpha P_j^\beta}{f(d_{ij})}, \tag{18}$$

where $f(d_{ij})$ is usually taken as either a power law d_{ij}^γ or an exponential decay $e^{\gamma d_{ij}}$, with α, β and γ parameters to be estimated from the data.

We will formulate this in terms of the number of trips (commutes) between county i and county j. Instead of simply taking the population as P_i and P_j, we can take into account the fact that we know the distribution of both home positions *and* office positions. So, the probability to observe a trip from i to j can then be formulated in terms of number of homes at the origin H_i and number of office at the destination O_j.

Again, we will discern two regimes, the close-by regime with $d_{ij} < 150\,\text{km}$ and the far-away regime with $d_{ij} \geq 150\,\text{km}$. We have fitted both the power law decay as well as the exponential decay, and found the power law decay to be a slightly better fit. The results are displayed in Figure 14, and the fitting results in Table 4. Interestingly, the parameter for the decay distance γ for large distances is not significant, suggesting that for distances $d_{ij} \geq 150$ the number of trips no longer depends on the actual distance. In fact, the only coefficient remaining significant for large distances is the coefficient of the number of offices at the destination. So, for larger distances it seems only to be important how many opportunities there are at the destination.

The fit of the model using the number of homes and offices is better than the fit when using the population size. The R^2 when using number of homes and offices are 0.52 and 0.26 for the two regimes as displayed in Table 4, compared to 0.43 and 0.24 when using the population size. Hence, it is worth taking into account such factors when modeling commuting distances. Looking at how well the model

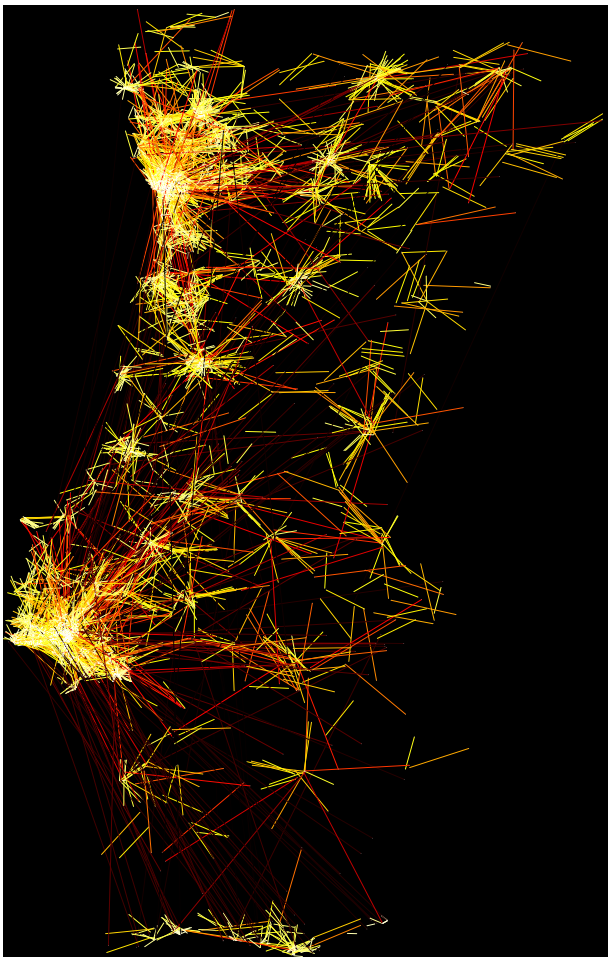


Figure 12. Commuting map for our sample of user. Brighter colours indicate smaller commuting distance. Most of the commutes only cover small distances, although some commutes cross half the country. The number of commutes decays approximately log-normally over distance.

predicts our data, we can see that for small distances, the prediction overestimates quite a bit the number of commutes. So there is room for improvement when analyzing small distances.

4 Concluding remarks

In this paper we analyzed the behavior of the customers based on their calling habits. We first sampled $100,000$ customers randomly and we filtered their locations, since those are based on the connected antenna locations, which are subject to disturbances. Then we defined and computed 50 features which compactly described the calling behaviors of the customers. We performed a correlation analysis on them which showed that movement and location related features are correlated with many other features. Then, we analyzed the data using principal component analysis (PCA). It showed that the features are highly redundant and can be efficiently compressed if we allow some error, e.g., $5\,\%$, after reconstruction. We also performed a clustering analysis and it found only a small number of typical user classes. The

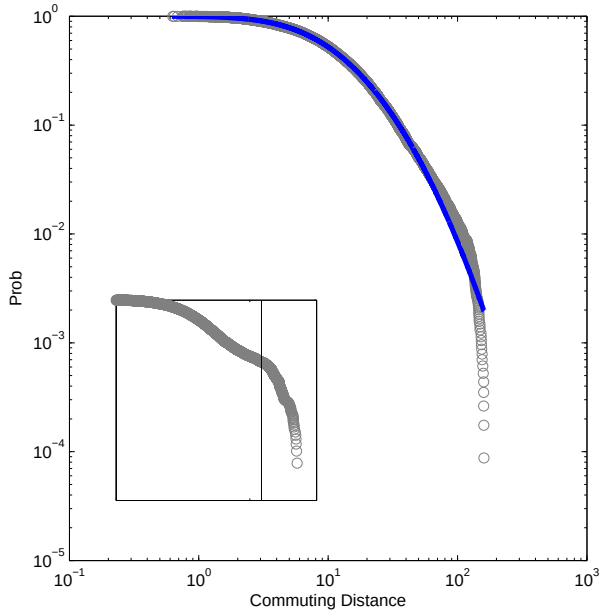


Figure 13. The distribution of commuting distance exhibit a log-normal distribution for distances $d < 150$ km. The inset shows the full distribution, with the line at 150 km separating the two different regimes. These two different regimes probably arises because almost the whole of Portugal is within 150 km of either Porto or Lisbon.

importance of the features, in case we applied PCA or clustering, was also defined and computed. This importance analysis demonstrated that location and movement related features are especially important for both cases. We therefore turned to the analysis of frequent locations of users.

Based on the weekly calling patterns of frequent locations, we clustered them and found that only home and office locations could be clearly identified. Surely other types of usages, such as weekend houses are present in the data, but they are marginal when compared to the clear pattern of home and office locations. We have characterized how many people have how many frequent locations, and in what combinations they appear most likely (e.g. multiple houses or offices). Finally, we estimated the position of these frequent locations based on a probabilistic inference framework. Using these positions we could provide a fairly accurate estimate of the distribution of the population, which showed a correlation of 0.92 with independent population statistics. These positions also allowed us to analyze the commuting distance, and the data seems to be explained reasonably well by the so-called gravity model. This model works better when taking into account separately the number of homes and offices instead of working simply with population sizes. So, when considering commuting distances, it is worth to take this into account.

The present study is only an exploratory analysis of the data. Further research into the frequent locations and their behavior should be undertaken. Especially since this data set has both geographical data as well as social network data, it would be interesting to analyze the interaction between these two different aspects further.

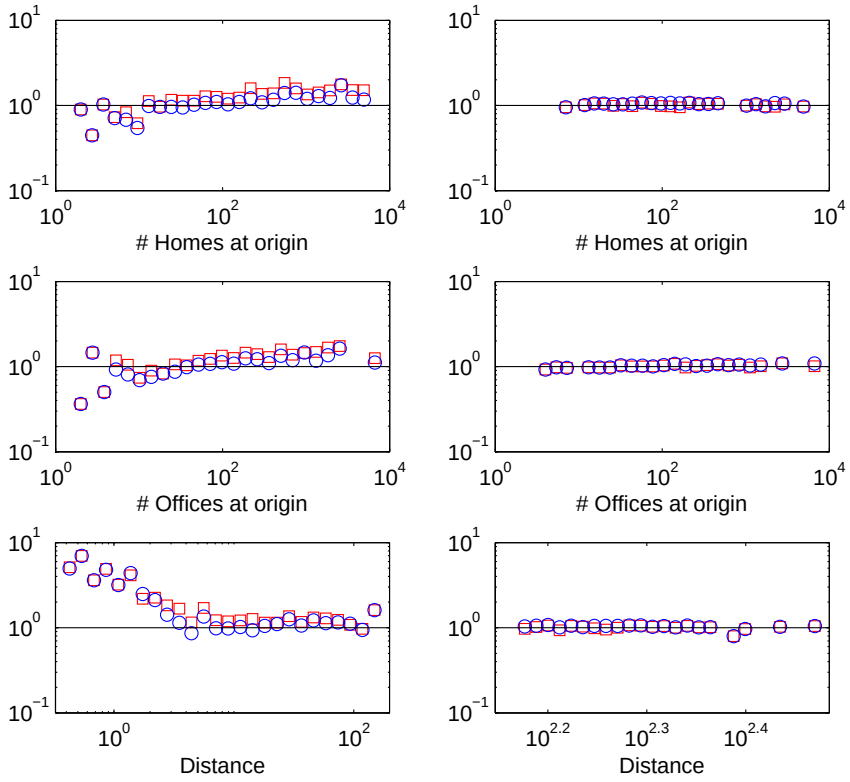


Figure 14. Plot of the prediction ratio $\hat{w}_{ij}/w_{ij}$ for the regime with $d < 150\,\text{km}$ at the left and $d_{ij} \geq 150\,\text{km}$ at the right. Red squares indicate the mean value and blue circles the median.

Acknowledgments

References

1. H. Aapo, Survey on Independent Component Analysis, *Neural Computing Surveys*, Vol.2, pp.94–128, 1999.

2. A. Adams and M.A. Sasse, Privacy issues in ubiquitous multimedia environments: Wake sleeping dogs, or let them lie ? in Proceedings of *INTERACT' 99*, Edinburgh, pp.214–221, 1999.

3. Donald Ball, Towards a Sociology of Telephones and Telephoners, in Truzzi, M., Ed. *Sociology and Everyday Life*, Englewood Cliffs, N.J., Prentice Hall, 1968.

4. S. Chiu, Fuzzy Model Identification Based on Cluster Estimation, *Journal of Intelligent & Fuzzy Systems*, Vol.2(3), 1994.

5. T. de Baillencourt, T. Beauvisage and Z. Smoreda, Are social networks technologically embedded ? *Réseaux*, Vol.145-146, pp.81–114, 2007.

6. C. de Kerchove, E. Huens, P. Van Dooren and V. Blondel, Social Leaders in Graphs, in Lecture Notes in Control and Information Sciences, *Positive Systems*, Vol.341, pp.231–237, 2006.

Coefficient	Variable	$d_{ij} < 150$	$d_{ij} \geq 150$
α	Number of homes at origin	$0.17^{**} \pm 0.013$	0.018 ± 0.013
β	Number of offices at destination	$0.21^{**} \pm 0.013$	$0.030^{*} \pm 0.012$
γ	Distance	$0.37^{**} \pm 0.018$	0.13 ± 0.11
	R^2	0.52	0.26

Table 4. Fitted parameters and R^2 of the gravity model, with standard errors reported. ** $p < 0.001$, * $p < 0.05$.

7. Marta C. González, César A. Hidalgo and Albert-László Barabási, Understanding individual human mobility patterns, *Nature*, Vol.453, pp.779–782, June 2008.

8. R.W. Hogg and A.T. Craig, *Introduction to Mathematical Statistics*, Macmillan Publishing Co., 1978.

9. R. Lambiotte, V.D. Blondel, C. de Kerchove, E. Huens, C. Prieur, Z. Smoreda, P. Van Dooren, Geographical dispersal of mobile communication networks, *Physica A*, Vol.387, pp.5317–5325, 2008.

10. C. Licoppe, Z. Smoreda, How networks are changing today with changes in communication technology, *Social Networks*, Vol.27(4), pp.317–335, Oct. 2005.

11. Z. Smoreda and C. Licoppe, Gender-Specific Use of the Domestic Telephone, *Social Psychology Quarterly*, Vol.63(3), pp.238–252, 2000.

12. Zs.J. Viharos and L. Monostori, Training and Application of Artificial Neural Networks with Incomplete Data, *Proceedings of the 15th International Conference on Industrial & Engineering Application of Artificial Intelligence & Expert Systems*, Cairns, Australia, pp.649–659, 2002.

13. Instituto Nacional de Estatistica, Inquérito à Ocupa cão do Tempo, 1999.

14. D. Tse and P. Viswanath, Fundamentals of wireless communication, *Cambridge University Press, 2005*

15. H. Zang, F. Baccelli, and J. Bolot, Bayesian inference for localization in cellular networks, *Proc 29th Conf Info Comm*, pp. 1963–1971, IEEE Press, 2010.

16. V.A. Traag, A. Browet, F. Calabrese and F. Morlot Social Event Detection in Massive Mobile Phone Data Using Probabilistic Location Inference, submitted to *IEEE SocialCom'2011.*

17. J.A. Nelder and R. Mead, A Simplex Method for Function Minimization, *The Computer Journal, vol. 7, 1965.*

18. D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J.J. Ramasco, and A. Vespignani, Multiscale mobility networks and the spatial spreading of infectious diseases,*Proceedings of the National Academy of Sciences of the United States of America*, pp 21484–9, vol. 106, 2009.

19. S. Erlander and N. F. Stewart, The Gravity Model in Transportation Analysis, *Utrecht, VSP*, 1990.

20. Wang, D., Pedreschi, D., Song, C., Giannott, F., and BarabÃ¡si, A.-L. Human mobility, social ties, and link prediction. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining KDD 11, 1100.* ACM Press. doi:10.1145/2020408.2020581, 2011

21. Song, C., Qu, Z., Blumm, N., & Barabasi, A.-L. Limits of Predictability in Human Mobility. *Science*, 327, 1018. doi:10.1126/science.1177170