Identification and Characterization of Human Behavior Patterns from Mobile Phone Data

Pavlos Paraskevopoulos
University of Trento, Italy,
Telecom Italia - SKIL
p.paraskevopoulos@unitn.it

Thanh-Cong Dinh
University of Trento, Italy,
Telecom Italia - SKIL
thanhcong.dinh@unitn.it

Zolzaya Dashdorj
University of Trento, Italy,
Telecom Italia - SKIL,
Fondazione Bruno Kessler
dashdorj@disi.unitn.it

Themis Palpanas
University of Trento, Italy,
themis@disi.unitn.eu

Luciano Serafini
Fondazione Bruno Kessler
serafini@fbk.eu

ABSTRACT

The availability of datasets coming from the telecommunications industry, and specifically those relevant to the use of mobile phones, are helping to conduct studies on the patterns that appear at large scales, and to better understand social behaviors. This study aims at developing methods for enabling the extraction and characterization of normal behavior patterns, and the identification of exceptional, or divergent behaviors. We study call activity and mobility patterns to classify the observed behaviors that exhibit similar characteristics, and we analyze and characterize the anomalous behaviors. Moreover, we link the identified behaviors to important events (e.g., national and religious holidays) that took place in the same time period, and examine the interplay between the behaviors we observe and the nature of these events. The results of our work could be used for early identification of exceptional situations, monitoring the effects of important events in large areas, urban and transportation planning, and others.

1. INTRODUCTION

Starting with the assumption that important events affect the behavioral patterns of a significant number of people in such a way that these changes are reflected in their use of the mobile telephony, this study aims to develop methods for enabling the extraction, analysis, and evaluation of quantitative and qualitative information about the calling and mobility behavior patterns of users. We focus on the characterization of normal behavior patterns, the identification of exceptional, or divergent behaviors, the characterization of such behaviors (e.g., offering explanations for these behaviors), and the prediction of similar patterns. Examples of the situations we are interested in are national and religious holidays, as well as major events of local interest (e.g., sports events).

In order to achieve the above goals, we need to complement the information present in the D4D datasets with contextual information that describes the environment and context in which a user is making a phone call, and that can provide an additional set of feature for the characterization of the user behavior. Information about the context of a call can be extracted from other sources on the web, like event databases, weather forecast database, etc.

In this work, we study the call activity and mobility patterns, classify the observed behaviors that exhibit similar characteristics, and we analyze and characterize the anomalous behaviors. The results of our work can be used for early identification of exceptional situations, monitoring the effects of important events in large areas, urban and transportation planning, and others.

This paper is organized as follows. The next section briefly discusses related works. Section 3 describes D4D datasets and methods for preprocessing the data. We present our core analysis methodology in Section 4. Some experiments are conducted in Section 5. Finally, we summarize our work in Section 6.

2. RELATED WORK

Calls placed from mobile phone devices are traced in logs which can serve as an indication to understand personal and social behaviors. Researchers in the areas of behavioral and social science are interested in examining mobile phone data to characterize and to understand real-life phenomena [9, 5, 7, 15, 10], including individual traits, as well as human mobility patterns [1, 2], communication and interaction patterns[2, 11, 14].

Dynamics of call activity: Candia et al. [4] proposes an approach to understand the dynamics of individual calling activities, which could carry implications on social networks. The author analyzed calling activities of different groups of users; (some people rarely used a mobile phone, others used it more often). The cumulative distribution of consecutive calls made by each user is measured within each group and the result explains that the subsequent time of consecutive calls is useful to discover some characteristic values for the behaviors. For example, peaks occur near noon and late evening. The fraction of active traveling population

and average distance of travel are almost stable during the day. This approach can be applied for detecting anomalous events.

Moreover, a number of interesting approaches propose to analyze mobile phone data to understand personal movement patterns, in particular individual tracking and monitoring [13, 3, 16], and behavioral routines [6].

Human mobility: Furletti et al. [8] extract user profiles from mobile phone data. The authors analyze moving human behaviors which correspond to specific human profiles (such as commuter, resident, in-transit, tourist), inferred by profile assumptions. A classification technique based on neural networks, called self organizing map, is used to classify users by similar profiles that have temporal constraints based on their temporal distributions. The result shows that the percentage of residents was compatible with the customer statistics provided by the Telecom operator, and short-ranged temporal profiles like commuter and in-transit are significantly different and distinguishable from the profiles with a larger extent like resident. The authors tested their approach on a case study in the city of Pisa (Italy). The data consists of around 7.8 million call records during the period of one month. They identified a peak which was caused by the reporting of an earthquake news. The authors highlighted the necessity to align temporal call distributions with a series of high level observations concerning events and other contextual information coming from different data-sources, in order to have more specific interpretation of the phenomena.

Phithakkitnukoon et al. [12] analyze the correlation of geographic areas and human activity patterns (i.e., sequence of daily activities). pYsearch (Python APIs for Y! search services) is used in order to extract the points of interest from a map. The points of interest are annotated with activities like eating, recreational, shopping and entertainment. The Bayes theorem is then used to classify the areas into a crisp distribution map of activities. Identifying the work location as a frequent stop during the day from the trajectories of individuals, it derives the mobility choices of users towards daily activity patterns. The stop extractions are done in the same way as in [2]. The study shows that the people who have same work profiles have strongly similar daily activity patterns. But this similarity is reduced when the distance of work profile location of the people are increased. Due to the limitation of heterogeneity of activities in this paper, the result includes some strange behaviors, like shopping during the night in the shopping area, which cannot be explained by the ground truth of activities.

Anomaly detection: Candia et al. [4] propose a simple approach to detect exceptional situations on the basis of anomalies from the call patterns in a certain region. The approach partitions the area using Voronoi regions centered on the cell-towers, and computes the call pattern in the "normal situation". These patterns are compared with the actual data and anomalies are detected with the use of the percolation method.

Mobility patterns: In [2] the author analyze the mobility traces of groups of users with the objective of extracting standard mobility patterns for people in special events. In particular this work presents an analysis of anonymized traces from the Boston metropolitan area during a number of selected events that happened in the city. They indeed demonstrate that people who live close to an event are preferentially interested in those events.

Social response to events: The social response to events, and behavior changes in particular, have been studied by J.P.Bagrow et al. [1]. The authors explored societal response to external perturbations like bombing, plane crash, earthquake, blackout, concert, and festival, in order to identify real-time changes in communication and mobility patterns. The results show that from a quantitative aspect, behavioral changes under extreme conditions are radically increased right after the emergency events occur and they have long term impacts.

Crowd mobility: Calabrese et al. in [2] characterize the relationship between events and its attendees, more specifically of their home area. The consecutive calls are measured in the same manner as in [4], in order to determine the stop duration of the trajectories. Given an event, for each cell-tower of the grid, the count of people who are attending that event, and whose home location does not fall inside that cell-tower, describes the attendance of events in geo-space. Most of the people attending one type of event are most probably not attending other types of event and people who live close to an event are preferentially attracted by it. As a consequence, the approach could partly predict starting locations of people who are coming to the future events. This could be useful to determine anomalies and additional travel demands for the capacity planning considering the type of an event. Conversely, knowing event interests of people helps to detect the event. But estimating the actual number of attendees and validating the models is still an open problem due to the presence of noise in the ground truth data. So, it derives to other issues like refining mobility patterns belonging to the events which occurred in the similar region at a closer time, and distinguishing home locations for people who live in the same location where events are organized.

3. DATA DESCRIPTION AND PREPROCESSING

D4D provided four datasets, each dataset has different features, giving us the possibility to try more than one techniques on the available data. We now discuss the characteristics of these datasets, related event information during that period as well as some necessary preprocessing that we performed before applying our techniques.

3.1 Description of Datasets

The first dataset provided us that describes the aggregated communication between cell-towers. The second and the third datasets refer to mobility traces, fine and coarse resolution data, respectively. The fourth dataset contains data about the communication between the users, creating subgraphs. All the datasets contain data that covers the whole country of Ivory Coast and were collected from December 2011 to April 2012 (five-month period).

We concentrated mainly on the first two datasets. The first one consists of 175.645.538 rows, each dataset has a

record for each available column, while the second consists of 55.319.911 rows.

In preprocessing the data we observed that the volume of missing data was rather large, which made it difficult to make accurate predictions or connections with the events, during the subsequent analysis phases. This problem was created due to some technical problems, and as a result led to the loss of the origin, or the destination cell-tower id. The missing cell-towers were recorded as '-1'. More precisely, just for the first dataset, the amount of missing data was too big that for each cell-tower we had an average of 143.162 records, while at the same time the number of records for cell-tower '-1' were 1.846.084.

At this point we have to mention that even though the cell-tower ids range from 1 to 1238, there are some ids that don't belong to a cell-tower. Furthermore, there are some cell-towers that do not have any record during the whole five-month period. As a result we have just 1214 cell-towers with records plus one, the cell-tower '-1' that represents the missing data.

3.1.1 Aggregate Communication Between Cell-Towers Data

This dataset contains data about the number of calls and their total duration. The data was grouped by their origin and their destination cell-tower. Furthermore the dataset contains timestamps about the time that the calls were initialized, but not the time that they were terminated.

3.1.2 Mobility Traces: Fine Resolution Data

This dataset provides us data with the cell-tower ids that some specific users connected to for a predefined period and with the timestamps for each connection. The users that were "tracked" for the construction of this dataset were a sample that was changing every two weeks and was chosen every time at random. The id for each user is unique during this two-week period but after two weeks it is assigned to another user. This reduced the resolution of the data was necessary in order to protect the privacy of the users.

3.1.3 Events Data

In order to collect some interesting events that took part during the five-months period covered by our sample, we used the Google Search Engine and we manually extracted the most important events related to Ivory Coast. Examples of such events are public holidays, important festivals, sport events, concert shows, and news that could change the activity of a user.

The extracted events refer only to the time period between the beginning of December 2011 and the end of April 2012. These events are listed in Table 1, and include events of both both regional and national importance.

3.2 Preprocessing of Datasets

The datasets were structured in such a way that an immediate analysis was not possible in order to make clear conclusions about the changes of the calling activity. Before starting the development of our methods, we had to manipulate the data in a way that we would keep just the most useful (for us) data and turn them in a more usable form. In the following part of this section, we describe these preprocessing steps.

3.2.1 Useful Variables

The methods used in the first dataset have only two types of values. The first is the hourly number of calls for each cell-tower, and the second is the total duration of these calls. Due to the volume of the data, we decided to aggregate the 24 hourly values that each cell-tower has for each day into a single daily value. Even though this aggregation leads to some information loss, it allows us to perform an initial fast analysis, which can subsequently be refined, by using the hourly data values, for the cases in which we detect an abnormal behavior.

We note that many cell-towers did not contain 24 values for every day in the dataset (due to the missing data problem we discussed earlier). Moreover, some cell-towers did not have values for each day of the period that the available dataset was produced, but this did not cause a problem for our analysis.

Apart from the two variables provided in the dataset, we derived and used a third variable that helped us to perform our analysis. This variable is the "duration per call" (dpc) that can be extracted by the division of the daily duration of calls by the number of calls, for each cell-tower. The values for this variable were calculated according to Equation 1.

$$dpc_{i,j} = \frac{total\_duration_{i,j}}{number\_of\_calls_{i,j}}, i, j \in N \quad (1)$$

3.2.2 Normalizing the Data

There as some cell-towers that are in urban areas and some others that are in areas that don't have many citizens. This has as a result that the first group of cell-towers have a continuously high activity, with respect to both the number of calls and their duration. Furthermore there are some days, like public holidays, that have more calls than the days when there is not any special event. These two factors do not allow us to cluster the data because the days or the cell-tower that have this overhead would always be reported as outliers.

In order to eliminate this problem we normalize the data by using z-normalization. In statistics, the z-normalization ensures that all elements of the input vector are transformed into the output vector whose mean-μ is 0, while the standard deviation-σ (and variance) is 1. For this transformation, we used Equation 2.

$$x'_{i,j} = \frac{x_{i,j} - \mu_j}{\sigma_j}, i, j \in N \quad (2)$$

We normalize the values in two ways. First we normalize by day in order to have normalized data with respect to each individual day. This can be achieved by finding the mean value and the standard deviation for each day and then compute the new value according to each day. This

Date	Location	Event	Event type
Dec 25, 2011	Ivory Coast	Christmas Day	public holiday
Jan 01, 2012	Ivory Coast	New Year's Day	public holiday
Feb 05, 2012	Ivory Coast	Day after the Prophet's Birthday (Maouioud)	public holiday
Feb 13, 2012	Ivory Coast	Post African Cup of Nations Recovery	public holiday
Apr 09, 2012	Ivory Coast	Easter Monday	public holiday
Feb 05, 2012	Ivory Coast	Mouloud	public holiday
Feb 22, 2012	Ivory Coast	Ash Wednesday	public festival
Jan 14, 2012	Ivory Coast	Arbeen Iman Hussain	public festival
Jan 8, 2012	Ivory Coast	Baptism of the Losd Jesus	public festival
Mar 25- Apr 1, 2012	Bouake	Carnaval	public festival
Apr 1- May 1, 2012	Ivory Coast	Fete du Dipri	public festival
Apr 6, 2012	Ivory Coast	Good Friday	public festival
Feb 9, 2012	Ivory Coast	Mawlid an Nabi (Shia)	public festival
Feb 4, 2012	Ivory Coast	Mawlid an Nabi (Sunni)	public festival
Feb 5, 2012	Ivory Coast	Yam	public festival
Dec 7, 2011	Ivory Coast	Anniversary of the death of Felix Houphouet Boigny	public festival
Apr 13-14, 2012	Abidjan	Assine Fashion Days in Cote D'Ivoire	show concert
Apr 1-4, 2012	Yamoussoukro	Education international 22nd congress	conference meeting
Apr 25, 2012	Sakre	Violence attack in Sakre	emergency event
Dec 17-18, 2011	Yale	Violence	emergency event
Jan 7, 2012	Abidjan	Hilary Clinton's visit	news event
Jan 7-8, 2012	Abidjan	Kofi Annan's visit	news event
Mar 12-13, 2012	Abidjan	Election of National Assembly President and Prime Minister	news event
Dec 11, 2011	Abidjan	New parliament election	news event
Jan 30, 2012 19-20	Ivory Coast	ACNF 2012 match vs Angola	sport
Jan 26, 2012 20-21	IvoryCoast	ACNF 2012 match vs Burkino Faso	sport
Jan 22, 2012 17-18	IvoryCoast	ACNF 2012 match vs Sudan	sport
Feb 4, 2012 20-21	IvoryCoast	ACNF 2012 match vs Equatorial Gulnea	sport
Feb 8, 2012 20-21	IvoryCoast	ACNF 2012 match vs Mall	sport
Feb 12, 2012 20:30-21:30	IvoryCoast	ACNF 2012 final match vs Zambia	sport

Table 1: List of important regional and national events in Ivory Coast, for the time period between December 2011 and April 2012.

kind of normalization helps us to identify patterns across days. In this case we use Equation 2, where i is the cell id and j is the day.

In addition, we normalize by the cell-tower, using each individual cell-tower's mean value and standard deviation. This action helps us to identify patterns for the cell-towers. In this case we use Equation 1 again, but in contrast to the previous case, i is the day and j is the cell id.

3.2.3 Creating sequences of movement

In order to develop the third method described in this paper, we used the second dataset. The original data do not allow us to analyze movements of customer. Therefore, we need to preprocess the data using the following steps:

- Step 1: Grouping the close cell-towers to avoid spatial gaps.
- Step 2: Creating sequences of movement.
- Step 3: Data cleansing to correct the "lost" cell-tower named '-1'.

In Step 1, we determine the customers' trajectories. The trajectory $Traj$ of an arbitrary customer c is represented by an array of places with their associated time-stamp. The identifier of a place (Id) is the identifier of the cell-tower where the customer has network connection. Let p_i be the place i-th and t_i is its associated time-stamp, where $i \in [1, n_c]$, n_c is the maximum number of places which customer c has visited. We have:

$$Traj_c = \{(p_1, t_1) \rightarrow (p_2, t_2) \rightarrow ... \rightarrow (p_{n_c}, t_{n_c})\}$$

Next, for each trajectory, we group pairs of two consecutive places whose Harversine distance is within 500 meters. The pseudocode for this process is shown in Algorithm 1.

Algorithm 1 Group close cell-towers by Haversine distance

```
1: procedure GroupByHaversine(Traj_c)
2:     p_i ← Traj_c.p_1                          ▷ Obtain the first place
3:     while i < n_c do
4:         p_{i+1} ← Traj_c.p_{i+1}
5:         if Harversine(p_i,p_{i+1}) < 500 ∧ p_i ≠ p_{i+1} then
6:             loc ← p_i p_{i+1} ▷ Create representative location
7:             Traj_c.p_i ← loc
8:             Traj_c.p_{i+1} ← loc
9:         p_i ← p_{i+1}
10:        i ← i + 1
```

For Step 2, we split each trajectory, which was obtained from Step 1, into 140 sub-trajectories according to the 140 days of the dataset's observation period. Then, we create daily sequences of movements using those sub-trajectories. A sequence of movements is a 24-element array. In each element, there is an ordered list of locations where the customer visited during the hour that this element represents. Note that, a sequence has its identifier which made of the sample's Ids, that of the customer c and the day d in format yyyy-MM-dd. Thus, we have:

$$Seq_{s,c,d} = \{(< 0 >, l_0), (< 1 >, l_1), ..., (< 23 >, l_{23})\}$$

The i-th element is described by $(< i >, l_i)$ where i is the hour and l_i is the list of locations that the customer visited within hour i and $i+1$. For example, considering the following $SubTraj$ which is a sub-trajectory obtained from sample number 0 and the customer 1:

$SubTraj$ = {(264,'2011-12-06 16:59:00'), (264,'2011-12-06 21:00:00'), (264,'2011-12-06 22:36:00')}

Its sequence is obtained as below:

$Seq_{0,1,2011-12-06}$ = {(<0>),..., (<16>,264),..., (<21>,264), (<22>,264), (<23>)}

Finally, for each customer, Step 3 simply replaces the cell-tower '-1' in sequences of movements of that customer by the most frequent visited-cell-tower in the same hour range.

4. PROBLEM DESCRIPTION

In this work, we concentrate on three problems. First, we identify and investigate the anomalous behaviors discovered in some cell-towers, and examine the reasons that could cause such a behavior. The second problem we tackle is to characterize the way that a social event affects to the calling activity of a region, or to the entire country in general. Finally, we analyze the social response to some major events, and investigate how different events affect the mobility of users.

4.1 Analysis of Anomalous Behavior

In this part we present a method that allows us to analyze the calling activity of the entire country of Ivory Coast and to normalize data, creating clusters and extracting usage patterns. Furthermore, we can identify activities in specific regions or even in the entire country that are not normal.

4.1.1 Identifying Outliers

After normalization of the data we have to compare the values with respect to the day or the cell-tower. In order to compare these values we calculate the mean and the standard deviation for each cell-tower or for each day, depending on the analysis that we intend to do. Furthermore, we calculate the difference of each point from its neighbors and from the mean. If we have a point A that is much farther away from the mean than a point B that is the closest to A and between A and the mean, then the point A is marked as an outlier. In practice, we implement a simple density-based clustering algorithm to create one main cluster that contains the mean value and to separate this cluster from the points that are much different than the cluster. Such an example is the plot depicted in Figure 1.

In this figure, we have an analysis of the data points clustered by day after we normalized by the day. As we can see there are two days, the day 66 and 67, that have some cell-towers whose points are much farther from the mean than the rest of the points, creating a gap between them and the rest of the cluster. By analyzing these outliers, and after having set a threshold[1] of '3.5', we found that the

[1]We set this threshold-radius manually after observing the

cell-towers that have these values, never had such a calling activity during the rest of the period that covers our dataset.

The algorithm that implements our method is shown in Algorithm 2.

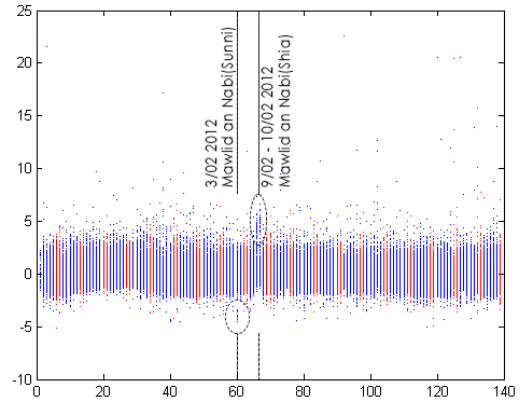


Figure 1: Daily plot for dpc normalized by day.

Algorithm 2 Grouping By Distances

procedure GROUPBYDISTANCES(threshold)
  $x'_{i,j} \leftarrow NormalizedValues$
  for $i = 1 \rightarrow MaxID$ **do** ▷ The MaxID is either the maximum ID of the cell-tower or the maximum ID for the days, depending on the analysis we intend to do. In most of the cases the day.
    $Distances \leftarrow allthedistances$ ▷ between each point that belongs to i and i's mean
    $array \leftarrow sortthedataaccordingtothedistances$
  for $i = 1 \rightarrow MaxID$ **do**
    for all *points* $\in i$ **do**
      if $distance \geq closerpoint$ **then** ▷ the point that is closest and between the examined point and the mean
        while $(NotTheEnd)$ **do**
          while $(NoPointWithGreaterDistance)$ **do**
            $checkNextPoint()$
            if it is close to the previous point **then** ▷ they propably form a sub-cluster **return** *point* ▷ as a possible outlier

4.1.2 Identifying Outliers Using the Standard Deviation

A second method that we used to identify the outliers is the comparison of the standard deviations. This method can be mainly applied on the daily values because each day has more or less the same features. More specifically each day has (almost) the same number of values and each value derives from a cell-tower that is every day at the same longitude and latitude. The only difference is that if an event is local then it will be hard to detect using the normalization by cell-tower. This results in the creation of datasets that have some steady main features plus some features that change and allow us to analyze them.

Following this method, we look at the standard deviation for each day and we compare it with the standard deviations of the 12 adjacent days, 6 before the day under examination and 6 after. This helps us to draw a conclusion on whether the calling pattern is more or less the same for this day as it should be, in respect to the period that we analyze. In case the standard deviation is not similar to the majority of the compared days, we can come to the conclusion that some event, such as a public holiday has taken place.

The pseudocode for this technique is shown in Algorithm 3.

Algorithm 3 Grouping By The Standard Deviation

procedure GROUPBYSTD(threshold)
  $x'_{i,j} \leftarrow NormalizedValues$
  for $j = 1 \rightarrow MaxID$ **do** ▷ The MaxID is either the maximum ID of the cell-tower or the maximum ID for the days, depending on the analysis we intend to do. In most of the cases the day.
    $Difference \leftarrow 0$
    $Similar \leftarrow 0$
    for $k = (j-6) \rightarrow j+13$ **do** ▷ Compare with the 6 previous days and the 6 days after
      if $i \neq j$ **then**
        if $(\sigma[i] \neq (threshold * \sigma[j]))||(\sigma[i] \leq (threshold * \sigma[j]))$ **then**
          $Difference \leftarrow Difference + 1$
        else
          $similar \leftarrow similar + 1$
    if $Difference \geq 6$ **then return** i

4.1.3 Correlated Abnormal Behaviors

We have already analyzed the cases that a value is an outlier for a cell-tower, or for a day. The problem that rises is the importance and the weight that the value has in general. If, for example, cell-tower 1 has for one day 100 calls and for the next day again 100, this could be possibly a normal pattern according to the cell-tower whose values remain more or less stable. What happens, though, if the values for all the cell-towers apart from this one are increased during the second day? This means that cell-tower 1 is an outlier. If we perform only a normalization with respect to the cell tower this outlier could be lost.

In order to avoid this situation, we have to correlate the two normalized values. This can be achieved by the subtraction of the two normalized values, and then look for outliers in this new space. This correlation can be achieved by using Equation 3, where x'_1 and x'_2 are the values derived from the two normalization procedures.

$$weight_{i,j} = x'_1 - x'_2, i, j \in N \quad (3)$$

4.2 Analysis of Social Response to Events

Here, we are interested in the shape, and the trend of call volume changes (social response) during events (e.g., public holidays, festivals etc.) at each cell-tower. Analyzing the social response to the events, we cluster the behaviors and the

data.

behavioral patterns in order to characterize the cells that have similar behavioral patterns as well as to identify normal, and abnormal behaviors in those cells. We assume that social response to events are similar at the scale of cell-tower area to that of the country. For example, a comparison of social responses to an event in two cell-towers is shown in Figure 2. The behaviors have a similar shape, pointing to the fact that during a public holiday people are in celebration and relaxing mood, which leads to reduced call volumes. A day after the holiday, the daily activities pick up again and the call volumes increased.

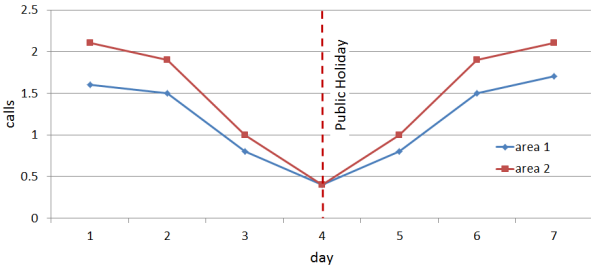


Figure 2: The behavioral patterns (social response) to a public holiday in the vicinity of the epicenter before (-3 < D_{event} < 0), during (-1 < D_{event} < 1), after (1 < D_{event} < 3) the holiday in two cell-towers. The dashed line describes a day-D_{event} when the event has been started.

To infer this type of information about social response, we analyze the call volume changes in the vicinity of the epicenter of an event before/during/after this event, in order to identify and extract behavioral patterns (social responses). As an event, we consider a public holiday which covers the entire population of the country.

To achieve this, we first annotate the information about public events that are described in Table 1 to daily call records at each cell. This event information is the context in which call volume is decayed or grown and it is described as a pair of location-time <l,t>. The location of the public events is Ivory Coast. Second, we extract call volume changes in the vicinity of the epicenter before/during/after the given event in order to extract behavioral patterns for the event from each cell. Third, we cluster these behavioral patterns to events by the call volume changes in order to find a similar pattern among all cell-towers.

4.2.1 Annotate the information about public events to daily call records

In order to prepare the daily call records, we group call records by day at each cell and normalize the daily call volume at each cell using the z-normalization procedure presented in Section 4.1. The equation to normalize the call volumes at each cell is described in Equation 4.

$$v_i' = \frac{v_i - \mu}{\sigma}, i \in N \quad (4)$$

1. v is a call volume
2. v' is a normalized call volume
3. μ is a mean value of call volume
4. σ is a standard deviation of call volume
5. i is a day parameter (i=1,2,3..,140)

Now, we need to annotate the information about events to the daily call records using the date. We create a data table that contains daily call records and event id in the following format.

```
CREATE TABLE DAILY_RECORDS (
date DATE,
originating_ant INTEGER,
total_duration_voice_calls INTEGER,
nb_voice_calls INTEGER,
event_id INTEGER
);
```

The following is an example of the resulting daily call records:

```
2011-12-05 2 175122 633 5
2011-12-06 3 3069220 310 2
2011-12-07 4 5665196 306 1
2011-12-08 5 5606249 303 0
```

4.2.2 Behavioral Pattern Extraction For Public Events

We analyze the call volume changes before/during/after the events. The behavioral patterns that occur in response to an event are represented as time-series of call activities in a given day-D_{event}, when the event has occurred, and in addition for d_1 days before (e.g., -3 < D_{event} < -1), during d_2 days (e.g., -1 < D_{event} < 1) and after d_3 days (e.g., 1 < D_{event} < 3). The pseudocode of behavioral pattern extraction is described in Algorithm 4. The algorithm returns the time-series of call volumes in certain periods in a given cell.

Algorithm 4 Extraction of behavioral pattern algorithm

```
1: function EXTRACT_PATTERN(D_event, d1, d2, d3, cell − tower)
2:     call_records ← daily_records(cell − tower)
3:     for d = 1 → 140 do
4:         if call_records[d, 1] = D_event then
5:             t ← 0                                  ▷ before the event
6:             for j = (D_event − d1) → D_event do
7:                 bpattern[t] ← call_records[j, 2]
8:                 t = t + 1
                                                      ▷ during the event
9:             for j = D_event → (D_event + d2) do
10:                bpattern[t] ← call_records[j, 2]
11:                t = t + 1
                                                      ▷ after the event
12:            for j = (D_event + d2) → (D_event + d3) do
13:                bpattern[t] ← call_records[j, 2]
14:                t = t + 1
       return bpattern
```

4.2.3 Clustering Behavioral Patterns

We cluster these behavioral patterns using hierarchical agglomerative clustering techniques in order to understand the most similar behaviors. To measure the dissimilarity between sets of call volumes, we use the Euclidean distance metric as described in Equation 5.

$$\mathrm{d}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (5)$$

1. x is a set of call volume 1
2. y is a set of call volume 2

Based on the dissimilarity between sets of call volumes, we cluster the pairs using average linkage clustering, a method for estimating the distance between clusters in a hierarchical cluster analysis. It specifies the distance between two clusters is computed as the average distance between objects from the first cluster and objects from the second cluster as presented in Equation 6.

$$D(C1, C2) = \frac{1}{N_{c1}N_{c2}} \sum_{i=1}^{N_{c1}} \sum_{j=1}^{N_{c2}} d(c1_i, c2_i) \quad (6)$$

1. $C1$ is cluster 1
2. $C2$ is cluster 2
3. N_{c1} is a number of values in cluster 1
4. N_{c2} is a number of values in cluster 2
5. $d(c1_i, c2_i)$ is the distance of cluster 1 and cluster 2

This way, we characterize the social responses to the events. Supervised classification and clustering can be used (e.g., k-means, k-nearest neighbor, SVM etc.).

4.3 Analysis of human mobility

To give an overview of the datasets, Figure 3 illustrates the total number of displacements, i.e., users changing a cell-tower, for all customers over time. There are ten time series that represent the entire dataset, where "series1" and "series10" are the first and the last consecutive two-week periods, respectively. Note that the beginning time of the first period is Monday 2011-12-05.

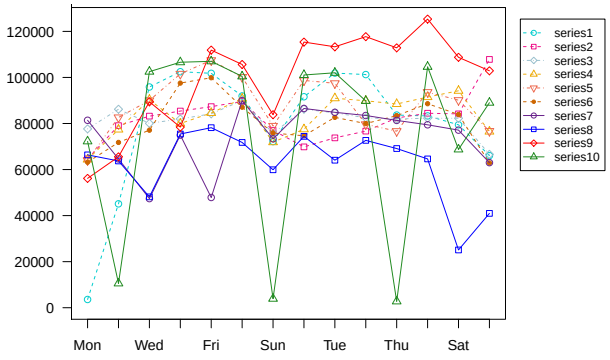


Figure 3: Total number of displacements

As shown in Figure 3, the time series in "series8" dips below other series, while coming to "series9", it keeps moving up until the beginning of "series10", it suddenly fluctuates. The interesting point here is that these strange behaviors of displacements might have correlation to the events happening during the same periods, i.e, Easter Monday on 2012-04-09.

In this section, we describe a method to analyze the correlation between movements of customers and the upcoming events. Thus, enabling us to detect in advance the locations (i.e., cell-towers) that are potentially related to some event. The proposed method consists of two steps: i) Inferring the cell-towers that show abnormal behavior in the number of distinct customers; ii) For each cell-tower obtained from the previous step, detecting potential cell-towers that may contain events.

In the first step, for each cell-tower, we count the number of distinct customers, who had network connection within the coverage of the cell-tower,considering a specific day d. Then, we use the sigma approach to extract all cell-towers that fall out of the following range R_d as outliers:

$$R_d = [\mu_d - \alpha \times \sigma_d, \mu_d + \alpha \times \sigma_d]$$

Where μ_d, σ_d and α is the mean, the standard deviation of number of distinct customers of all cell-towers in day d and the scale parameter, respectively.

Before going to the second step, we infer the home location of each customer by estimating the most frequent cell-tower where that customer stays during night hours (from 6PM to 8AM). Once the home locations of the customers are obtained, we use Algorithm 5 to detect the cell-towers that may have events among the cell-towers obtained from the previous step. Note that the sequence database S is obtained from Section 3.2.3, C and H are the lists of abnormal cells and home locations, and d is the day that cell-towers in C have abnormal behavior.

5. EXPERIMENTAL EVALUATION

5.1 Social response to events

In this section, we present the example of clustering the behaviors, and behavioral patterns that exhibit before/during/after an event by analyzing the social response. We characterize the social responses to the public event, Easter Monday (9 April, 2012). We extracted the daily call volumes at each cell-tower in the vicinity of the epicenter before ($-2 < D_{event} < 0$), during ($0 < D_{event} < 1$), after ($1 < D_{event} < 3$) Easter Monday as described in Figure 4. The figure shows that behavioral pattern for the Easter Monday is v-shaped and can fit inside 2 standard deviations. Before ($-2 < D_{event} < 0$) and during ($0 < D_{event} < 1$) the event, the call volumes are reduced to the lowest activity levels, suggesting that people are having a rest during public holiday and the urge to communicate is the weakest. After ($1 < D_{event} < 3$) the event, the call volumes increase again to reach the normal activity levels. We observe that some call volumes start decaying before the event. The reason could be that there are some areas where people do activities during the holiday such as concert place, relaxing place, eating place so on. Further, we investigate the call volumes by hour for these abnormal behaviors in those cell-towers.

However, the call volumes before/during the event ($-1 < D_{event} < 1$) are obviously divided into groups. Figure 4

Algorithm 5 Dectecting potential cell-towers

procedure DETECTEVENT($S, C, H, d, \alpha, minsup$)
  for all $c \in C$ **do**
    for $i \leftarrow 0, 23$ **do**
      $cnt \leftarrow$ count(S, d, i, c) ▷ Count the number of distinct customers have visited cell-tower c in hour i on the day d in database S
      $\mu \leftarrow$ getMean(S, d, i, c) ▷ Calculate the average number of distinct customers concerning the days before d
      $\sigma \leftarrow$ getStd(S, d, i, c)
      $lb \leftarrow \mu - \alpha \times \sigma$ ▷ Calculate lower bound
      $ub \leftarrow \mu + \alpha \times \sigma$ ▷ Calculate upper bound
      if $cnt > ub$ **then**
        $support \leftarrow$ getSupport(S, d, i, c, H) ▷ Calculate relative support level concerning only the sequences of customers who have visited cell-tower c in hour i of day d and their home location is not c
        if $support < minsup$ **then**
          $out \leftarrow c, d, i$
      if $cnt < lb$ **then**
        $support \leftarrow 1-$ getSupport(S, d, i, c) ▷ Calculate relative support level considering home location
        if $support < minsup$ **then**
          $out \leftarrow c, d, i$

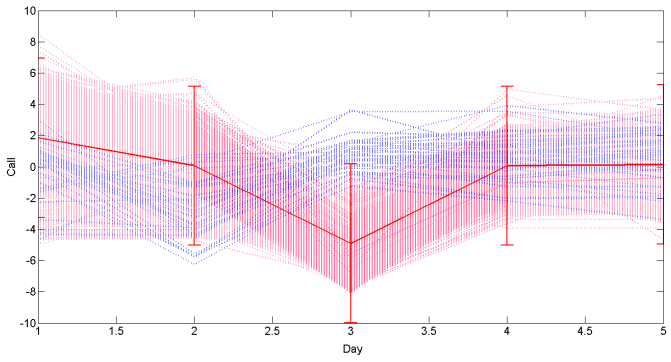


Figure 4: The call volume changes in the vicinity of the epicenter before (-2 < D_{event} < 0), during (0 < D_{event} < 1), after (1 < D_{event} < 3) Easter Monday. Pink and blue color represent daily call volumes in each cell-tower, red color represents for the μ (mean) $\pm$ 2*σ (standard deviation).

shows that call volumes (described in blue color) during the events are conversely increased. To cluster these behavioral patterns (-2 < D_{event} < 1) , we create a matrix that contains the call volumes in certain times at each cell-tower. The dataset can be seen as a collection of cell-tower vectors that contains the call volumes in certain times. We calculate the euclidean distance between all possible pairs of vectors to find dissimilarities. Then, we perform hierarchical clustering, starting from the two cell-towers that has the maximum distance as a cluster in first stage. In next stage, the two nearest clusters are merged into a new cluster. The process is repeated until the whole data set is agglomerated into one single cluster as represented in Figure 6. Figure 5 shows the classification of behavioral patterns before/during/after the event. After the clustering, we identified the most two groups; call volumes are increased during the event that covers 39 cell-towers out of 1214 cell-towers (total numbers of cell-towers which are identified from the experiment dataset) and call volumes are decreased during the event that covers 1173 cell-towers out of 1214 cell-towers.

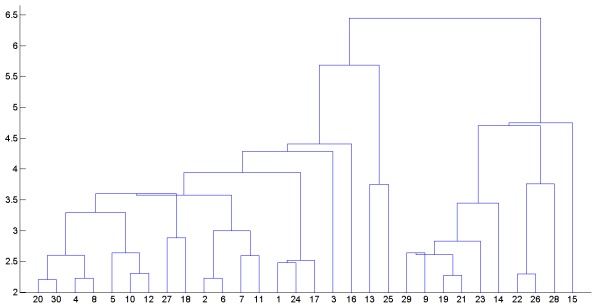
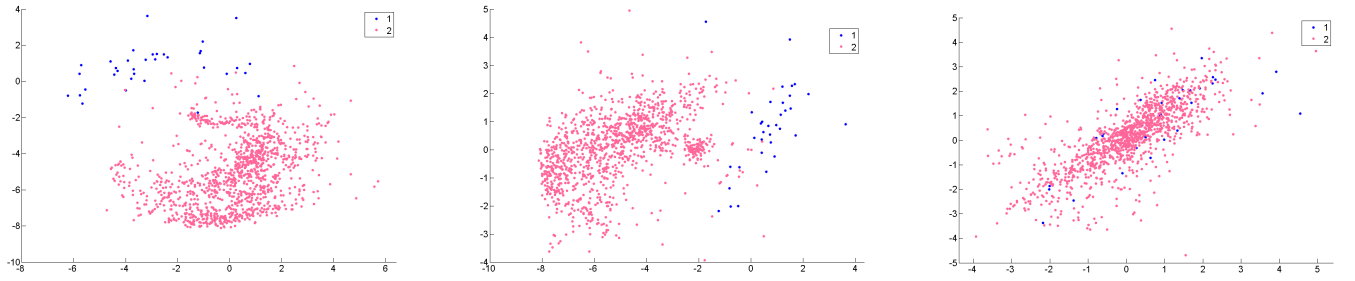


Figure 6: Groupings of behavioral patterns

The location of the cell-towers where the call volumes are increased during the event are shown in Figure 7. The call volume changes are concentrated in the center of the country (Daloa, Zuenoula, Bouafle) which is the main cocoa-growing region and south east of the country (Abidjan) which is the capital of the country. This could explain that during the Easter Monday, people work in the field as in working day at the cocoa-growing region, and other shops or restaurants are operated at the capital. Due to the fact that the event information is limited in web site for experiment period, we are not able to provide more accurate information to these call volume changes. Further, we can investigate the mobility patterns of users to check if there are movement from other cities to these area.

Similarly, we analyzed the social responses to all public holidays which are described in Table 1. The characterization of social responses to each public holiday is represented in Table 2. A day before the public holiday, Post African Cup of Nations Recovery there was the final match that Ivory Coast vs Zambia for 2012 African Cup of Nations. This affects to behavioral patterns during the public holiday, Post African Cup of Nations Recovery. Further, we investigate the mean pattern at each cell to investigate the magnitude of the call volume changes during the events. Also, the behavioral patterns during New Year's Day are overlapped with the public festival, New Year on 31 December. Some cells are not active



(a) A day before the event (-1 < D_{event} < 0)

(b) During the event (0 < D_{event} < 1)

(c) A day after the event (1 < D_{event} < 2)

Figure 5: Classification of behavioral patterns for the period (-1 < D_{event} < 2)

Public holiday	Pattern extraction period	Cluster count	Cell-towers at each cluster	not active cell-towers
Christmas Day	(-1 < D_{event} < 1)	3	[1],[572],[503]	138
New Year's Day	(-2 < D_{event} < 1)	4	[656],[8],[2],[411]	137
Day after the Prophet's Birthday (Maouioud)	(-1 < D_{event} < 1)	4	[943], [115], [2], [1]	153
Post African Cup of Nations Recovery	(-2 < D_{event} < 2)	2	[475],[548]	191
Easter Monday	(-1 < D_{event} < 1)	2	[39], [1173]	2

Table 2: Public holiday classifications

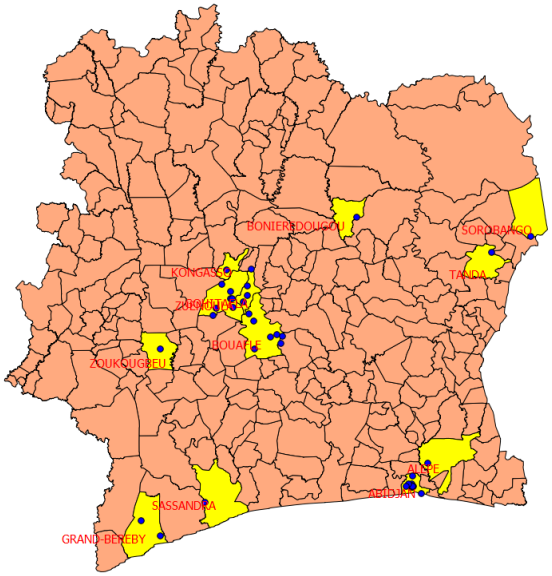


Figure 7: Location of the cell-towers where the call volumes are increased during the event (-1 < D_{event} < 1)

during the events

5.2 Anomalous behaviors

In this section we present some results after we applied the method described in section 4.1 on the first dataset.

The first step is to compute the duration per call (dpc) at each cell-tower. The next step is to normalize the data by day and by the cell-tower, as described in 3.2.2. After the normalization step we have six values, two for each initial variable, the daily number of calls (nb), the daily duration of the calls (dur) and the dpc for each cell-tower.

We cluster normalized data in the way described in section 4.1.1. This has a result to identify behaviors that are not normal. Such type of behaviors we can see in Figure 1. In this figure we can see the dpc normalized by each day values. At the x axis we have the day id and at the y axis we have the normalized value of the dpc by day for each day. With red color we can see the weekends and with blue color the weekdays. This difference at the colors makes it easier for us to achieve even a visualized comparison between the values. As we can see almost all the days follow the same pattern, having values that are in a small range plus some values that are (unique) outliers. Investigating these outliers, we found that are mostly the same cell-towers. This allows us to consider it as a normal pattern for these specific cell-towers and we don't analyze them more.

Although most of the days in the sample follow the same pattern, there are some days like the days with ids 60, 66 and 67 that have a sub-cluster of outliers. Investigating these outliers we found out that, for the days 66 and 67,

these cell-towers are only 36 and they are close to each other. Furthermore, we found out that the calling activity referring to the dpc for these cell-towers is unique for these days and they don't have such an activity for the rest of the five-month period. For the day 60 we have the same conclusions as we did with the days 66 and 67 with the difference that the outliers are negatives.

Finally we came to the conclusion that specific events or actions change the calling activity for these days for these specific cell-towers. You can see the cell-towers that are outliers for the days 66,67 and the day 60 in Figures 8,9 respectively.

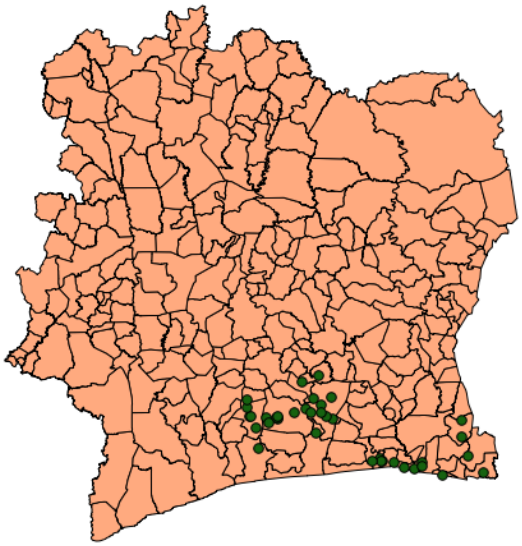


Figure 8: Cell-Towers Positive Outliers for 9-10 of February 2012

One more fact that evaluated our conclusions is the analysis that we did for the number of the calls for each day when this number was normalized by the day. In order to achieve this type of analysis we used the method described in section 4.1.2. By analyzing these values we found out that just a single event could cause a difference on the calling activity for just a region(when it is a local event) or even for the whole country.

Such an example is depicted in Figure 10 where we can see the difference of the calling activity for the whole country during the Christmas and the new year event, the easter, the days that we have unique events such as festivals and the rest of the days. We depict with blue color the weekdays and with red color the weekends.

Finally we evaluate the method described in 4.1.3 by analyzing the duration of the calls for each cell-tower while we have normalized it both by the day and the cell-tower. Subtracting the second value from the first we get a result that for cell-towers with ids 731 to 750 the weights for the weekends are mostly clustered at the positive values while the weights for the weekdays are clustered to the negative

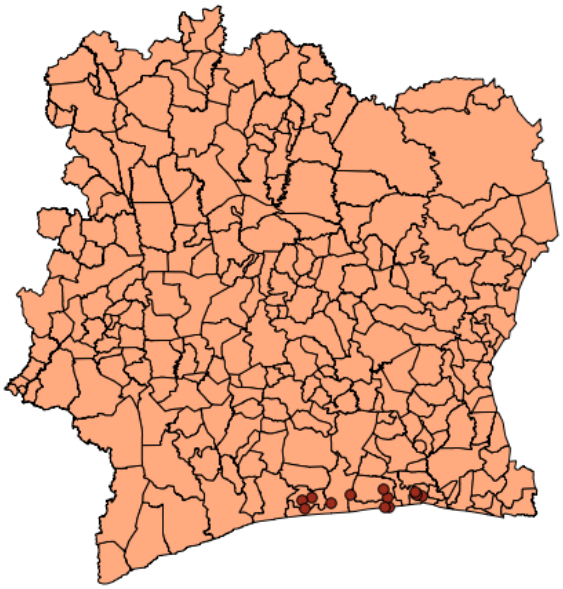


Figure 9: Cell-Towers Negative Outliers for 3 of February 2012

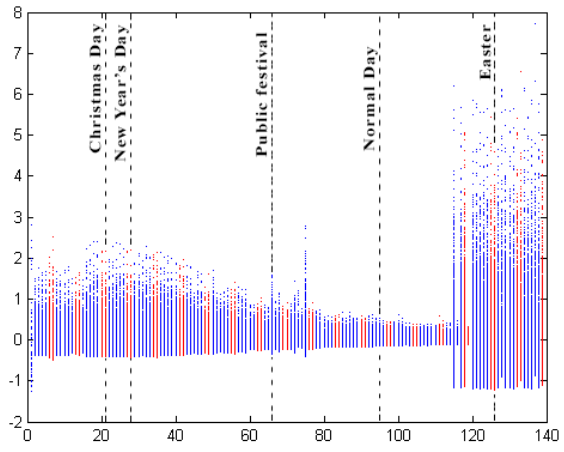


Figure 10: Number of calls for each day (normalized by day)

values. This makes it clear that these cell-towers have specific patterns for the weekdays and the weekends. In Figure 11 we have this analysis visualized. Again with blue color we have the weekdays and with red color the weekends. In x axis we have the cell-tower id and in y axis we have the normalized daily duration of calls.

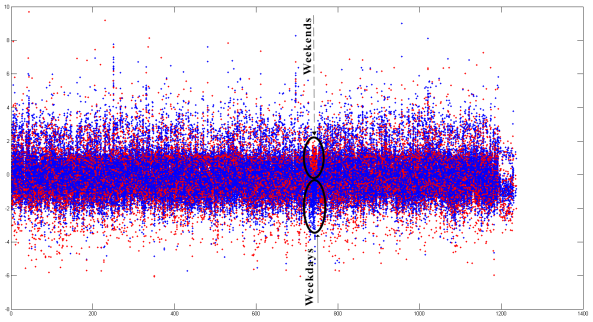


Figure 11: Weights For The Correlation of The Two Types of Normalized Values For The Duration (for each cell-tower)

5.3 Human mobility

Using the method described in section 4.3, we found the cell-towers that have abnormal behavior in movements of customers. As in Figure 12, there are significant signs of augment of movement of customers in the Christmas day, the day after the Prophet's Birthday (Maouioud) and the New Year holiday. While the movement of customers dramatically decreases before Carnaval.

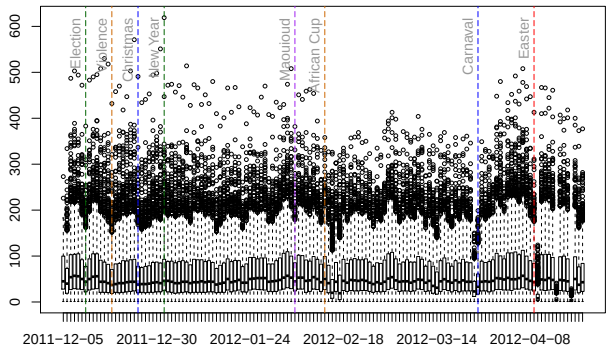


Figure 12: Result of inferring abnormal cell-towers. Points represent cell-towers and y-axis is the number of distinct customers visited the cell-towers

We choose the New Year day and Carnaval day as the interesting days for experimenting Algorithm 5. The potential cell-towers that have events on the New Year day, are shown in Figure 13. Almost cell-towers are located in Abidjan, which is the largest city of Ivory Coast. Among these cell-towers, we investigate one random cell-tower to illustrates Algorithm 5. As shown in Figure 14, we can infer that there may have special events during period from 0AM to 3AM and 9AM to 11PM at the location cover by cell-tower's Id '728'. During these events, there is significant increase in movement of customers arriving this location.

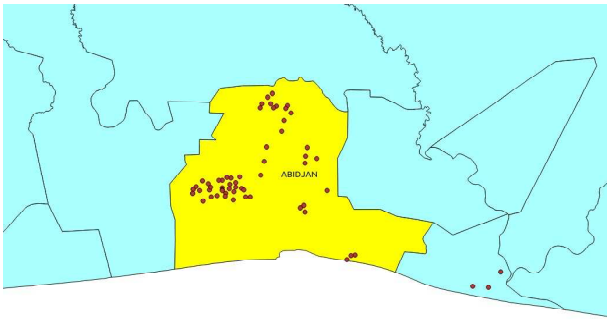


Figure 13: Result of detecting event's location. Points represent cell-towers and the day is 2012-01-01

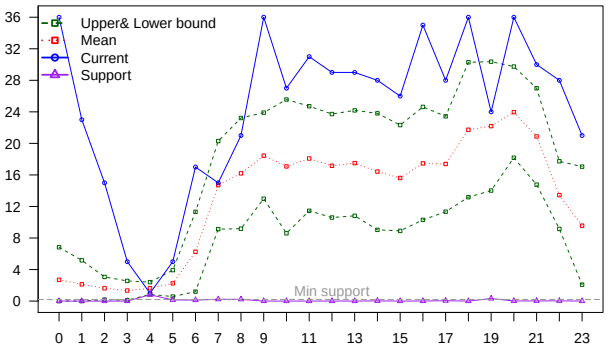


Figure 14: Number of distinct customers who visited cell-tower's Id '728', 2012-01-01, x-axis is hours. Minimum support is 0.2, and α is 1.5

In another context, Figure 15 plots the cell-towers that potentially have events among the abnormal cell-towers in the day before Carnaval event day. The cell-towers are mainly located in Bouake and Abidjan. We also investigate one random cell-tower as shown in Figure 16. There may have special events during period from 4PM to 6PM of the day 24th March, 2012 at coverage of cell-tower's Id '642'. During this event, the significant customers, who visited this cell-tower, may move to other locations or may not use their mobile phone as usual.

6. CONCLUSION AND DISCUSSION

In this paper, we presented three main methods for identifying patterns and outliers in the behavior and mobility of mobile phone users, by analyzing the data recorded by cell-towers. Our methods can be used to predict events and actions that are possible to happen if some specific circumstances exist, for example, to predict activities when we have dry-periods, or important events, such as festivals and public holidays.

We are currently extending our techniques in order to become more targeted. The goal is to enable a more detailed analysis, focusing on particular regions of the country, or examining in detail the patterns that occur in finer time scales (e.g., by analyzing the data at the hourly level).

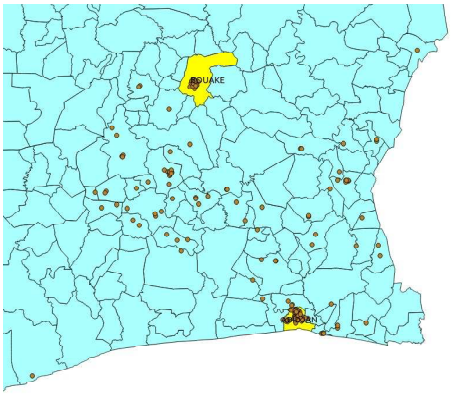


Figure 15: Result of detecting event's location. Points represent cell-towers and the day is 2012-03-24

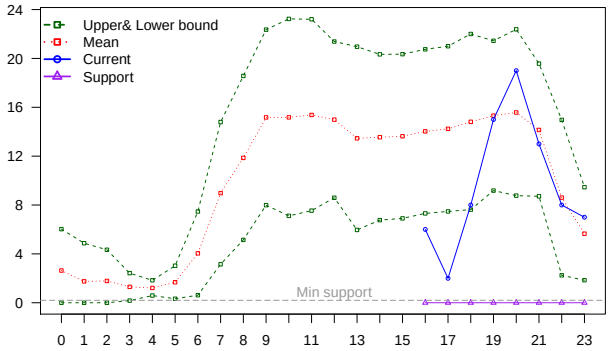


Figure 16: Number of distinct customers who visited cell-tower's Id '642', 2012-03-24. Minimum support is 0.2, and α is 1.5

7. REFERENCES

[1] J. P. Bagrow, D. Wang, and A.-L. Barabási. Collective response of human populations to large-scale emergencies. *CoRR*, abs/1106.0560, 2011.

[2] F. Calabrese, P. F. C., G. Di Lorenzo, L. Liu, and C. Ratti. The geography of taste: analyzing cell-phone mobility and social events. In *the Proc. of the 8th international conference on Pervasive Computing*, Pervasive'10, pages 22–37, Berlin, Heidelberg, 2010. Springer-Verlag.

[3] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti. Real-time urban monitoring using cell phones: A case study in rome. *Trans. Intell. Transport. Sys.*, 12(1):141–151, Mar. 2011.

[4] J. Candia, M. C. Gonzalez, P. Wang, T. Schoenharl, G. Madey, and A. L. Barabasi. Uncovering individual and collective human dynamics from mobile phone records, 2007.

[5] D. Choujaa and N. Dulay. Tracme: Temporal activity recognition using mobile phone data. In *Embedded and Ubiquitous Computing, 2008. EUC '08. IEEE/IFIP International Conference on*, volume 1, pages 119 –126, dec. 2008.

[6] N. Eagle and A. (Sandy) Pentland. Reality mining: sensing complex social systems. *Personal Ubiquitous Comput.*, 10(4):255–268, Mar. 2006.

[7] D. Fox. Location-based activity recognition. In *the Proc. of the 30th Conf. on Advances in Artificial Intelligence*, pages 51–51, 2007.

[8] B. Furletti, L. Gabrielli, C. Renso, and S. Rinzivillo. Identifying users profiles from mobile calls habits. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, UrbComp '12, pages 17–24, New York, NY, USA, 2012. ACM.

[9] M. Kwan, C. Arrowsmith, and W. Cartwright. Visualizing population movements within a region, 2011.

[10] M. A. Muhammad Awais Azam, Laurissa Tokarchuk. Human behavior detection using gsm location patterns and bluetooth proximity data. *The 4th international conference on mobile ubiquitous computing, systems, services and technologies-UBICOMM 2010*, pages 428–433, 10 2010.

[11] J. P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A. L. Barabasi. Structure and tie strengths in mobile communication networks, 2006.

[12] S. Phithakkitnukoon, T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti. Activity-aware map: identifying human daily activity pattern using mobile phone data. In *the Proc. of the 1st Intl. Conf. Human Behavior Understanding*, pages 14–25, 2010.

[13] C. Ratti, A. Sevtsuk, S. Huang, and R. Pailer. Mobile landscapes: Graz in real time. In G. Gartner, W. E. Cartwright, and M. P. Peterson, editors, *Location Based Services and TeleCartography*, Lecture Notes in Geoinformation and Cartography, pages 433–444. Springer, 2007.

[14] C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, and S. H. Strogatz. Redrawing the map of great britain from a network of human interactions. *PLoS ONE*, 5(12):e14248, 12 2010.

[15] T. Sohn, A. Varshavsky, A. LaMarca, M. Y. Chen, T. Choudhury, I. Smith, S. Consolvo, J. Hightower, W. G. Griswold, and E. de Lara. Mobility detection using everyday gsm traces. In *8th International Conference on Ubiquitous Computing (UbiComp)*, Irvine, CA, September 2006.

[16] J. Steenbruggen, M. Borzacchiello, P. Nijkamp, and H. Scholten. Mobile phone data from gsm networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. *GeoJournal*, pages 1–21, 2011.