Discovering urban and country dynamics from mobile phone data with spatial correlation patterns

Roberto Trasarti[a]*, Ana-Maria Olteanu-Raimond[b], Mirco Nanni[a], Thomas Couronné[b], Barbara Furletti[a], Fosca Giannotti[a], Zbigniew Smoreda[b], Cezary Ziemlicki[b]

a. *KDD lab, Istituto di Scienza e Tecnologie dell'Informazione, CNR Pisa, Italy*
b. *Sociology and Economics of Networks and Services dept., Orange Labs, Paris, France*
*. Corresponding author: *roberto.trasarti@isti.cnr.it*

Abstract

Mobile communication technologies pervade our society and existing wireless networks are able to sense the movement of people, generating large volumes of data related to human activities, such as mobile phone call records. At the present, this kind of data is collected and stored by telecom operators infrastructures mainly for billing reasons, yet it represents a major source of information in the study of human mobility. In this paper, we propose an analytical process aimed at extracting interconnections between different areas of the city that emerge from highly correlated temporal variations of population local densities. To accomplish this objective, we propose a process based on two analytical tools: (i) a method to estimate the presence of people in different geographical areas; and (ii) a method to extract time- and space-constrained sequential patterns capable to capture correlations among geographical areas in terms of significant co-variations of the estimated presence. The methods are presented and combined in order to deal with two real scenarios of different spatial scale: the Paris Region and the whole France.

Key words: mobile phone, location data, mobility patterns, urban dynamics

1 Introduction

In recent years massive mobile phone location data have been studied and shown to have great potential to model human mobility (González et al., 2008; Song et al., 2010). In particular, studies on long-term mobility data proved how this kind of data could be important for urban planning and transportation studies (Reades et al., 2007; Calabrese et al., 2011). The difficulty to conduct classical travel surveys using self-report records (diaries or questionnaires) as well as the growing need to collect longitudinal data have drawn attention to automatic mobile phone data collection systems (Wang et al., 2010). What is clearly interesting for mobility analysis based on mobile phone data are the number of people concerned (almost the entire population is already equipped) and their longitudinal character (theoretically, there is no limit of time of observation). Moreover, the existing technical network offers the possibility of real-time and continuous localization monitoring. In fact, it can profoundly change the transportation research based on one-shot and infrequent surveys.

However, one of the main difficulties with mobile phone data is the incompleteness of the users' traces. In fact, people are localized only when they are using their phone (calling or sending SMS); it leads to several problems in finding mobility patterns by means of classical data mining algorithms.

The information provided by mobile phone infrastructures has been recently exploited in several interesting directions. One popular line consists in trying to map hand-overs (Bar-Gera, 2007) or individual Call Data Records (Tatem et al., 2009) to the road map, especially to estimate speeds, travel times and traffic state. The localization and the mobility detection are important information that can be used to estimate the population distribution (Ahas et al., 2010), to identify important locations such as home and work (Isaacman et al. 2011), and to support health care and social mobile applications. For instance, it can help understanding the spread of diseases (Tatem et al., 2009), as well as the behavior of tourists (Ahas et al., 2011; Olteanu et al., 2011) or the real-time monitoring of population density in urban areas (Ratti et al., 2005).

Further, trying to exploit the sequential pattern-mining paradigm some authors have also analyzed population density variations (Cao et al., 2005; Giannotti et al. 2007) to discover correlations between different geographical areas. In the field of population ecology (Bjornstad et al., 1999) proposed an approach essentially based on the computation of spatial covariance between time series. Following on from this research, we propose an event-driven approach, where the frequent co-occurrence of some events in different places and possibly different moments of time (though under some temporal constraints) is used to point out possible correlations, while covariance-based measures are able to spot only correlation that hold synchronously and for the whole time window.

The objective of our work is to provide exploratory tools for spotting statistically significant, yet potentially hidden, regularities in the way the population density deviates from expected values on different areas of the city (or other geographical context). In particular, we aim at discovering groups of regions that consistently *behave in a coordinated way*, suggesting the existence of some kind of connection among them. We stress the fact that our objective is to provide to the domain expert hints about patterns in the life of the city/geographical area that emerge directly from data and that might be unknown to him or simply unexpected. That contrasts with other approaches to human mobility study that start from a pre-conceived framework or model, and aim either at positioning the geographical unit (city, province, etc.) within the framework or at estimating some parameters to fit the model.

The patterns we look for in the data can take three forms: the first case is a *set of events* that frequently happen at the same time, intuitively representing regions that all react to some external factor. For instance, people might tend to concentrate in specific areas during leisure time whenever the weather conditions are exceptionally good. The second case is a *sequence of events* that frequently happen in a specific order, suggesting the existence of a reaction chain (one event causes successive events) or an external factor that is answered with different reaction times by the regions involved. For instance, a large increase of people at a central train station might be frequently followed by an increase in an other station within a few hours (a chain of events), or the delay of a bus line might generate effects on several locations along its route, yet at different times (different reaction times). Finally, a third case is the combination of the previous two, namely a sequence of sets of events, that might represent a reaction chain with effects on a wider zone, each step of the chain covering multiple areas at the same time, as well as some more complex phenomenon generated by both external factors and a cascade of effects. For instance, a large increase of people at a central train station might have effects on several stations, possibly with different reaction times. All the examples mentioned above involve simple phenomena and connections between regions that are probably already known to the domain expert. However, what a data-driven process like ours can provide is a way to systematically extract all the *promising* cases of connections for which there is statistical evidence in the data, providing to the domain expert a detailed and up-to-date overall picture, among which there might be non-trivial and useful bits of knowledge.

Our contribution differs from most works on the analysis of (human or vehicular) traffic flows based on phone data, which focus on detecting the movements of people among areas. While such movements clearly represent a source of connections between regions, of the kind we are tackling, our approach is much more general, and the patterns we aim to discover are not limited to cases of people migration between regions.

The paper is organized as follows. First, we introduce a new pattern definition (Section 2), *correlation pattern*, which tackles the problem of finding correlations between areas using the presence of cellphone users. In Section 3, we present an algorithm to extract such patterns together with the overall analysis process needed to prepare the raw input data and tools to help the analyst to organize automatically and navigate the results. The proposed algorithm and tools are integrated in an existing mobility data analysis platform: M-Atlas (Giannotti et al., 2011), thus allowing the analyst to take advantage of all the pre-existing features. Finally, a wide scale experimentation is carried out (Section 4), where the new methods are applied to three real case studies using Orange France data at two different granularity level: urban level (Paris region) and national level (France).

2 Objectives of analysis

Recent experiences using mobile phone data have shown that they can provide a good understanding of how the density of population changes during the day in various regions of a given (urban or larger) area, exhibiting periodic patterns during regular periods (working days, etc.), and some anomalies in specific locations during exceptional occasions (concerts, festivals, etc.) (Traag et al., 2011). However, the existing body of research appears to be mostly focused on the discovery of local phenomena, such as increases of population, or simple flows of population between pairs of regions. On the contrary, the dynamics of a city naturally create links of several different natures between regions of the city, sometimes even very far apart: some regions tend to get congested together as response to some external event (e.g. intense precipitations); others might be connected through a cause-effect chain, where the population of a region flows from one to the other in exceptional measure when the former exceeds some levels of saturation; an event that induces a loss of population in a region (for instance a closed subway station) might indirectly lead to variations somewhere else, positive (people taking the next station) or negative (people not taking the metro at all). Our claim is that the state-of-art in mobile phone-based human mobility studies still did not provide satisfactory means for inferring a map of the city (or other kind of area) that highlights such logical links between areas. We propose an approach based on mobile phone data analysis that tries to provide (partial) answer to such request.

2.1 Mobile phone data

In this study Call Detail Records (CDR) data are used. The CDRs are mobile phone logs collected for billing purposes, where location information (cell id) is generated at the start of a communication event: incoming and outcoming calls and SMA.. The CDR contains the timestamp, call duration and type of events (call, SMS), as well as the code of the cell in which the communication started. The location data (cell id) have to be decoded to obtain a geographic position. In our case, the cell tower geographic coordinates are used. Each cell tower is composed by at least three antennas (directional or omni-directional[1]). The CDRs were collected in 2007 and anonymized by Orange France before transferring to the research team. The popularity of CDRs in the research is mainly due to their huge size (months and months of communication and mobility behavior traces of millions of people) and their relative ease way of extraction and use (e.g. they have a standard formatand are recorded in highly secured databases) rather than the location data itself. The spatiotemporal information provided by CDRs is rather poor since the mobile phone positions are related to the antennas locations having heterogeneous spatial distribution and are generated by user dependent communication activities.

Figure 1 shows the temporal distribution of mobile phone events (calls and SMS) aggregated by 30 minutes. Some activity peaks can be observed at 12.30 pm, 3.30 pm, 4.30 pm and 7.30 pm. These peaks are related to lunch break, end of classes and commuting, coherent with the French daily activity rhythms.

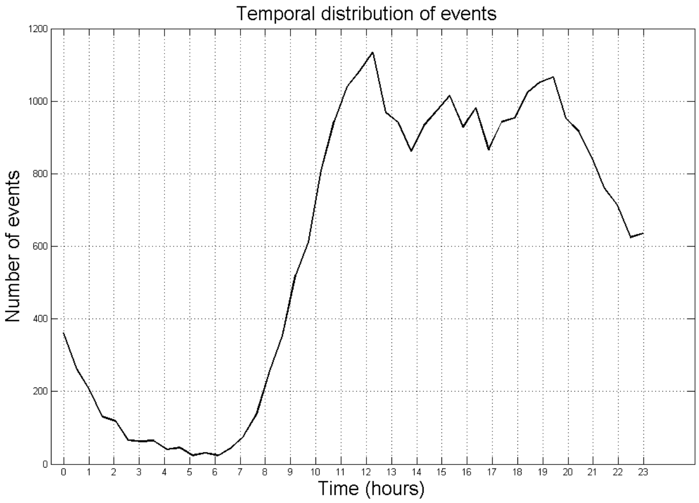

Figure 1 : The temporal distribution of events during a weekday

[1] An antena covers a specific geographical area depending on the location and population density. In low traffic locations such as rural areas, the anntenas will tend to be omni-directional covering a large circular area around the cell tower. In urban areas, three antennas will usually share a high site covering smaller areas. Each site will have directional antennas and cover a 120° arc away from the cell tower. In this study we use information on tower positions, so several cell id can have the same geographical coordinates.

2.2 Links between spatial regions

The goal of this work is to detect connections between different regions that can be inferred by the dynamic distribution of CDRs data in space.

Our approach is based on the observation that the population density distribution tends to be regular (periodic) in almost all regions. That is the result of several routine human activities such as going to work, going to schoolwhich are constantly generated by the same residents one the one hand of, and random activities due to tourism, use of services (e.g., shopping, leisure activities), generated by different people yet yielding overall stable densities, on the other hand. Therefore, finding evidence of connections between areas by simply looking at raw densities would essentially lead to link everything to everything.

Therefore, a more promising approach, , consists in looking for exceptions to regular behaviors. Following this idea, we start by searching *events* that represent significant deviations from regular trends, and locate them in space and time. Then, we detect recurrent combinations of events, by extracting frequent *patterns* of deviations. The types of patterns we look for essentially have the form of sequential patterns and are defined as follows:

Definition 1. Correlation patterns.

A correlation pattern (C-pattern) is a sequence of sets of events $D = \langle D_1, \dots, D_n \rangle$*, where* $D_i = \{d_1, \dots, d_n\}$ *is a set of events, each defined as* $d_j = (s_j, w_j)$*,* $s_j \subset \Re^2$ *is a spatial region,* $w_j \in \Re$ *is the weight associated to the event.*

A *C-pattern* describes a set of regions that often experience a (significant) deviation from their common behavior, and do that either at the same time (in the case the events belong to the same event set D_i) or at different times (if they belong to different event sets) but always in the specific order described by the pattern.

3 Methodology

The goal of the proposed method is to discover groups of spatial regions that exhibit some recurrent pattern of population presence variation, thus suggesting the existence of some connection between the regions related to the underlying mobility. In particular, a pattern describes which variations of population presence appear at each region and the temporal order such variations follow (including the case of regions with synchronous variations). In order to support this kind of analysis, in this section we introduce methods and algorithms to achieve three objectives: first, detect the locations and times where relevant variations of population take place; second, infer from them more complex patterns that link regions where variations tend to appear together or in some constant sequence; finally, navigate the discovered patterns along the spatial and temporal dimensions, and enrich patterns with additional information (derived from raw data) to help their interpretation.

3.1 Population presence and stops

The main concept behind our work is the population presence in a spatial region. Such presence can be estimated through mobile phone data in various ways, mainly depending on whether the presence measure for a region (within a time window) should include only users that stopped in the region or it should also count users that simply crossed it while moving towards a different destination. The first case requires a preliminary analysis of data aimed to label each single observation as either *stop* or *move*. On the contrary, raw phone call data records can be used, as they are for the computation of presence in the second case. The analysis methods proposed in this paper can be applied to both cases, and are actually general enough to accommodate also other concepts of *people presence*.

Although the case studies shown in the experiment section use exclusively the second approach, our tools are equipped with a heuristic solution also for the first approach, aimed to detect stops. In particular, our stop detection criterion is inspired by Palma et al. (2008), where spatial positions and time intervals are used instead of the speed: a *stop* for user *uid* is any ordered pair (p_k, p_m) of mobile phone points such that their location is the same ((p_i.x, p_i.y) = (p_k.x, p_k.y)), between them there are no points in different locations, and the temporal distance between them is longer than a given minimum duration threshold.

3.2 Correlation Patterns

This section describes how spatiotemporal observations contained in the data (only stops or also including crossings) can be manipulated to detect single interesting events, and how to build from them complex and recurrent patterns.

Table 1 provides a pseudo-code description of all the steps illustrated in the next two sections.

Table 1: C-pattern pseudo code

```
C-PATTERN(TrainS, TestS, S, T, σ, mGap)
Input: training set TrainS, test set TestS, input spatial regions S,
       input temporal intervals T, minimum support σ,
       minimum correlation ω, max-gap constraint mGap;
Output: set of C-patterns CP;
{
    // Density estimation
    foreach $s \in S$ and $t \in timeSlots(TrainS, T)$
        $DP(s,t)$ = ComputeDensity(TrainS, s, t);
    foreach $s \in S$ and $t \in timeSlots(TestS, T)$
        $DP'(s,t)$ = ComputeDensity(TestS, s, t);
    // Events detection
    $M$ = ComputeAvgDay(DP, S, T);
    $E_D$ = DetectEvents(M, DP', S, T);
    // C-pattern extraction
    $CP$ = C-SPAM($E_D$, ω, σ, … );
    return CP;
}
```

The algorithm requires as input a dataset (*TrainS*) to be used to estimate reference densities for each region and for each time slot, and the main dataset to be analyzed (*TestS*). Both *TrainS* and *TestS* are obtained from the previous step of stop detection. The set of regions (*S*) and time intervals (*T*) (which granularity can be hours of the day, for example) are provided by the user, as well as the minimum frequency of C-patterns to be extracted and the allowed maximum temporal gap between consecutive events in a C-pattern.

3.2.1 Density estimation and events detection

The basic elements of correlation patterns are the single events, which represent all the relevant variations of population for a given region. In this work, in particular, events are computed by comparing the density of population within a region in a given time against the expected density in the same area and time.

The first step of the process consists in defining both the spatial and temporal dimensions. In our experiments we consider several spatial granularities that range from the urban area partitioned according to the coverage of an antenna, to a whole nation partitioned into administrative regions.Once a temporal window is chosen (e.g. the month of June), a finer subdivision is defined: the periods (e.g. days) and the smaller time slots (e.g. hours). Periods and time slots are described by a parameter *T*, while the regions that cover the study area are described by a parameter *S*. Both parameters have to be provided by the user. The two parameters allow defining a spatiotemporal grid, and each observation of an input dataset can be assigned to one of its cells. The number of observations that fall in a cell defines so called the density of the cell.

For both input datasets (*TrainS* and *TestS*), a spatiotemporal grid of densities is computed. The density is a measure of the presence of individuals in a stated space and time. While *TrainS* is used to compute the density in a given time period, *TestS* is used as comparison over the time period in order to detect significant deviations of observations.

Based on the densities obtained for each region and each time slot over the *TrainS*, an expected density value is computed for each region, by averaging the densities measured at the same time slot of all the periods in the time window covered by the dataset. For instance, we might obtain an expected density for each pair *(region, hour of the day)*, i.e., 24 values for each region, assuming 24 one-hour time-slots.

Then, for each region and each time-slot, the corresponding density is compared against its expected value: if the difference is significant, an event of form *(region, weight, time slot)* is produced. The event represents a spatiotemporal slot and a discretized measure (*weight*) of how strong was the deviation. In particular, events are built on the basis of three parameters:

- a granularity of deviations, expressed as a percentage relative to the expected density;

- a minimum relative deviation, also expressed as a percentage, used to select significant deviations;
- an absolute minimum deviation, expressed as an integer number, used to discard extreme cases with very low densities.

The weights used for defining the events are multiples of the granularity, and an event for a region and a time slot is built only if the deviation of its density w.r.t. the corresponding expected density is larger than the absolute minimum deviation, and if it is larger, in percentage, than the minimum relative deviation.

Steps 1-6 of the general algorithm cover the phase just described.

In detail, the first phase concerns the construction of a model *M* that provides the average density of each region and each time slot in the typical period. After having partitioned the area in spatial regions, and the time window in time slots (step 1), the process aggregates the raw observations (step 2) counting the number of observations per region in each time slot (i.e. each hour of the day in each day of the training dataset). A similar process is repeated for the test dataset (steps 3 and 4). Then, the aggregation is used to build an average day *M* (the model) eliminating outliers and noise (step 5), where an average density is associated to each time slice.

The second phase implements the events detection: the model *M* is compared to *TestS* in order to highlight the events E_D (step 6), computed as deviations from *M*. In particular, E_D will contain a sequence of sets of time stamped events for each day (or other spatial granularity, derived as temporal extension of *T*) having the form $\langle D_1, \ldots, D_n \rangle$, where $D_j = \{(s_j, w_j, t_j)\}_j$, s_j is a region, w_j is a weight for the event, and t_j represents the time slot.

3.2.2 Extraction of C-patterns

The extraction of *C-patterns* focuses on those patterns that appear frequently, i.e. they occur with some given minimum frequency among all. In particular, the *C-patterns* are a set of sequential patterns over the dataset of events E_D computed by using C-SPAM algorithm. C-SPAM (Constrained SPAM) is a new algorithm that extends SPAM algorithm (Agrawal & Srikant, 1995). Given a database of sequences, where each sequence is a list of "elements" ordered by time and each element is a set of items, SPAM discovers all sequential patterns (a list of sets of items) with a user-specific minimum support. The support is the percentage of data sequences that contains the pattern. The algorithm is based on the so-called *a priori* property, which states the fact that any super-pattern of an infrequent pattern cannot be frequent, and follows a candidate generation and verification process typical of the association mining paradigm. The first scan finds all the frequent items that form the set of single item frequent sequences. Then, in the following steps, new potential patterns are generated on the basis of the *a priori* assumption. Each candidate sequence contains one more item than the preceding step. By scanning the database the support for each set is computed. All the candidates with support no less than the minimum support is selected to form the new set of sequential patterns. The algorithm terminates when no new sequential pattern is found or when no new candidate sequence can be generated.

C-SPAM introduces several spatial and temporal constraints and allows extracting maximal and closed patterns. The most important constraint introduced and imposed in the *C-pattern* extraction is the minimum correlation value i.e. the threshold to be satisfied for producing a new C-Pattern. C-SPAM computes for each extracted *C-pattern* a correlation index (*c-index*), defined as follows. For a *C-pattern* $\langle D_1, \ldots, D_n \rangle$:

$$c - index(D) = \frac{supp(D)}{\prod_1^n \prod_{d \in D_i} supp(d)} \qquad (1)$$

where *supp*(*D*) (respectively, *supp*(*d*)) represents the support measure, i.e. the fraction of input sequences that contain the pattern *D* (resp., the single event *d*). This measure basically mimics the standard *lift* index defined for item sets, and expresses the ratio between the actual frequency of the pattern (D) and its expected frequency, computed under the assumption of complete independence between events. *C-patterns* with very high values of c-index can be considered surprising patterns, therefore potentially interesting, while *C-patterns* with a very low c-index are probably trivial. At step 7 of the overall algorithm, the frequent *C-patterns* satisfying a given support threshold (σ) and a correlation threshold (ω) are detected.

In addition, the user can specify other constraints such as:

- *Max/Min Size*: the maximum/minimum number of items in a pattern;
- *Max/Min Gap*: the maximum/minimum temporal gap between two consecutive items in a pattern;

- *Max/Min SpatialGap*: the maximum/minimum spatial gap between two consecutive items in a pattern;
- *Maximals*: if the resulting set of patterns must be maximal, which means that every pattern is not included in a biggest one, or not;
- *Closed*: if the resulting set of patterns must be closed, which means that every patter is not included in a biggest one with the same support value, or not.

All these constraints are included in the C-SPAM algorithm in a very efficient way taking advantage of their properties such as the monotonicity. To better explain how this property can be used in the algorithm we briefly present an example:

Example. Consider the dataset composed by three sequences:

a) 1:{A} 2:{B} 5:{C}
b) 1:{A} 2:{B} 3:{D}
c) 3:{A} 5:{F,G} 6:{D}
d) 2:{A} 3:{B} 5:{D}

where numbers represent the time and letters represent variations in a specific cells (e.g. A = +10% in the main train station). Executing the C-Pattern algorithm with the following parameters: Min_support = .5 (i.e. 2 sequences) and Max_gap = 2, the first step consists in finding the frequencies of each item removing the ones under the threshold: A, B, D are frequent. Then, in the second step the items are combined to generate the 2-items frequent sequences. In this case the Max_gap plays a significant role decreasing the support because it limits the valid subsequences, in this case the result is: A-B, B-D. The A-D is not allowed since there is only one sequence in the dataset and the time gap is less or equal than 2. In the third step the only sequence which is generated is A-B-D, since all constraints are satisfied. Due to the incremental way to generate each new pattern, at each step removing a pattern from the early stage of the process, reduces exponentially the cost in terms of memory and time.

3.3 Navigation and enrichment of C-patterns

The output of the *C-pattern* algorithm can be a very large set, which could be difficult to analyse. Thus, two automatic methods have been developed in order to organize the patterns by their temporal or spatial components. Furthermore, we enriched the *C-pattern by* using individual data in order to better understand the patterns, i.e. if there is a set of users moving between the areas that generate the events and then the pattern. All these tools are used in the experiments section showing how they can help in the interpretation of the patterns.

3.3.1 Temporal navigation

Each *C-pattern* represents a connection between areas with a specific variation in the number of users. A pattern may occur in different periods of the day and more than once in the same day. The temporal distribution of the occurrences of a pattern over the time window can be a useful means for navigating *C-patterns*. Here, we present a method that uses the temporal distributions of patterns to organize them in a tree structure by means of a hierarchical clustering method – more precisely, a single-link agglomerative algorithm. The Euclidean distance quantifying the distance between numbers of occurrences of the pattern in each hour of the day is used. The result is a *dendrogram* (basically, a tree structure) where each leaf represents a pattern and the intermediate nodes represent groups of similar patterns in terms of their temporal distribution. An example of dendrogram produced by the tool we implemented is shown in Figure 2 (left). The tool also allows visualizing the temporal distribution of each intermediate node, as shown in Figure 2(right). The graph associated to an internal node represents the typical distribution that applies in its sub-tree (bold red line) with additional information about the minimum and maximum values of the group for each time interval (pink shadow). The analyst can study the dendrogram at different levels in order to interpret group of patterns by their common temporal distribution, e.g., *morning patterns* as the patterns having all the occurrences in the morning.

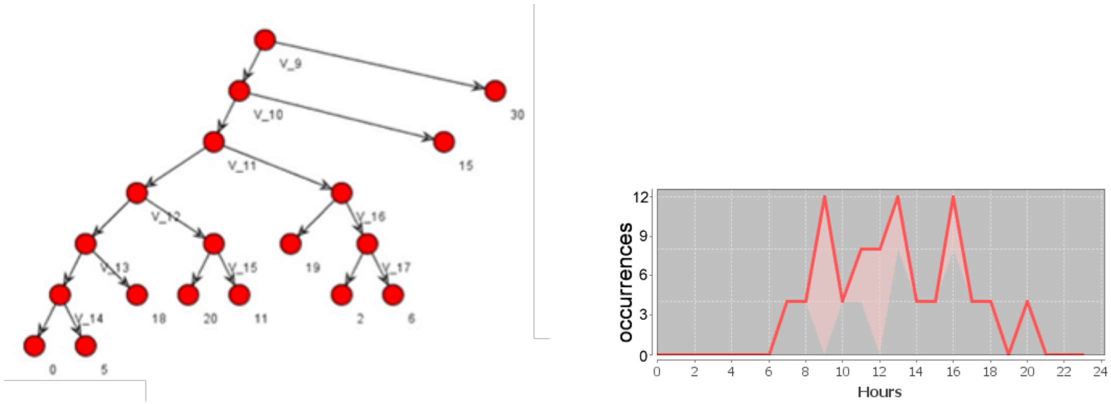

Figure 2 : (left) A dendrogram for temporal navigation. (Right) The graph showing the representative temporal distribution associated to a node (bold red line), and the minimum and maximum variation values (pink shadow).

3.3.2 Spatial navigation

Analogously to the temporal navigation, patterns can be organized by using their spatial component. In this case the tool groups the patterns composed by the same sequence of areas.

The spatial visualization can be very useful for understanding particular cases, especially when, depending on the granularity of the events, different variations of the same sequence appear in the results. An example is when sequences involving the same regions have different weights: these ones are actually treated as different patterns. The analyst can easily navigate the patterns and visualize the temporal distribution of each group as for temporal navigation. In fact, for each group the tool visualizes the representative temporal distribution and the minimum and maximum variations. It is important to notice that in this case the group can contain very different temporal distributions. In Section 5 an example of this tool will be illustrated.

3.3.3 C-pattern enrichment

As mentioned at the beginning of this paper, the connections between areas described by *C-patterns* can be generated by different phenomena. Some of them involve the propagation of effects from a region to other ones, some other phenomena involve exclusively external factors, and there is no direct interference among the regions involved. In this section, we present a method to enrich any *C-pattern* in order to quantitatively assess how much it fits a specific kind of connection, namely a *movement wave* of people: a sequence of people density deviations caused by the same group of persons moving from one region to the next one. Clearly, to perform this kind of evaluation it is not sufficient to know the population density of each region, since it requires to be able to identify the individuals that contribute to such density, in order to measure how many people appearing in a place at a given time also reappear later in other place.

This process can be carried out when the access to the individual data of the users (raw CDR data) is provided. The basic idea consists in considering each instance of a given *C-pattern* in the dataset of events, i.e. a sequence of regions and corresponding timestamps, and list the users that were detected in each region. Our objective is to measure the number of many users detected in each region of a given pattern. The measures obtained over each single occurrence of the C-pattern, then, are aggregated into a single score value, as described in the following definitions.

Definition 2. Shared users.

Given a C-pattern $P = \langle p_1, \dots, p_n \rangle$*,:, where* $p_j=\{(s_j, w_j)\}_j$*,, let* $u_1^k, \dots, u_m^k$ *represent the sets of users identified in the k-th occurrence of P in the events dataset D, each* u_j^k *corresponding to one of the events in P. Then,* number of shared users *of P is defined as* $U_{P,D} = \sum_{i=1\dots K} \left| u_1^i \cap \dots \cap u_m^i \right|$ *,where K is the total number of occurrences of P in D.*

The number of people shared along each instance of the *C-pattern* is then normalized by taking into account the density of each location involved, in the following way:

Definition 3. Transfer measure*.*

Given a C-pattern, P, and a set of events, D, the transfer measure among the regions in P is defined as:

$$\frac{U_{P,D}}{\sum_{i=1\dots K}|u_1^i|+\cdots+|u_m^i|} \qquad (2)$$

where m is the number of events in P and K is the number of occurrences of P in D.

When the transfer measure is high, the *C-pattern* describes the situation in which the variations in the correlated regions are likely to be caused by a movement of the same set of users.

Example. Let assume to have a *C-pattern* composed of three elements, $D = \langle\{d_1\ \},\{d_2\},\{d_3\}\rangle$, where: d_1=(S_a, 0.3), d_2=(S_b, 0.1), d_3=(S_c, 0.15). To simplify, let consider that we have a single occurrence of the pattern, i.e. K=1 in the definitions given above, and that the typical presence in each region described by the pattern is 100. In particular, that means that the measured population in the three regions is now, respectively, 130 (100 + 30%), 110 (100 + 10%) and 115 (100 + 15%). Consider now that the *shared users* is $U_{P,D} = 20$ users, i.e., 20 users are present in all three regions, while all others appear only in one or two of them. Then, we obtain that the *transfer measure* is equal to $\frac{20}{130+110+115} = 0.563 \cong 5.6\%$.

We notice that for privacy issues, mobile phone companies adopt anonymization techniques to preserve the identity of the users. For our purposes, the minimal requirement that guarantees both the privacy and the applicability of the method is that the user identifiers do not change within the time period (in the typical case, 24 hours). This allows having the set of data consistent in the minimum observation window (for instance, a day).

4 A study of Complexity and Performances

The proposed methodology consists in a set of data manipulations and pattern extractions, each one with its complexity. First the *Stop Computation* reads all the points of each trajectory in order to check if the constrains are satisfied for generating a stop. This process is O(*n*) where n is the number of points in the dataset. Second the *Density Estimation* counts the number of stops in each cell in each time slot, therefore the complexity is again O(*n*). Third the *Event Detection* compares the densities extracted for the test and the training set in order to discover the variations between them, this step is O(*t*s*) where t is the number of time intervals used (i.e. 24 hours for each day) and s is the number of cells. Fourth, the mining process: the *C-Pattern extraction* is exponential w.r.t. the number of items in the transaction *i* which is smaller than *t*s* (because the maximum size it is obtained if there is an event in each cell in each time interval), therefore the complexity is O(2^{t*s}). All this steps are executed sequencially ant then the overall complexity is composed by the two parts: O($n+2^{t*s}$), which is dominated by *n* if the number of observations are greater producing a small number of events, while it is dominated by the exponential term if several events are generated.

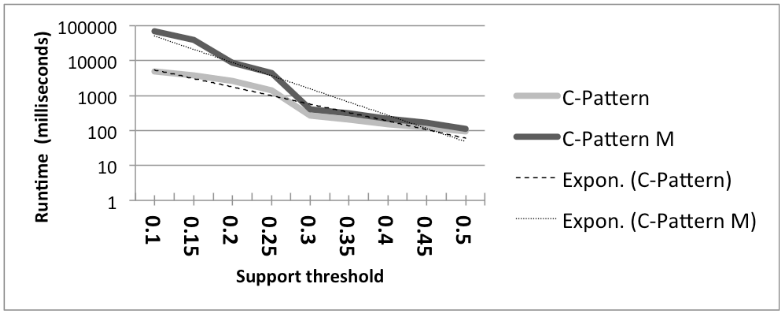

Figure 3: The runtime extracted by the *C-pattern* algorithm decreasing the support threshold using the Maximals (M) constraint and without it. The exponential trend-lines of the two are also reported.

Figure 3 shows the performances obtained during the experiments. The runtime of the C-Pattern is exponential with respect to the. In Figure 4, we can also see the effect of the maximal constraint applied (*C-Pattern M*) and, as expected, while it increases the runtime (Figure 3) it also drastically reduces the number of patterns discovered, since it removes the redundant ones.

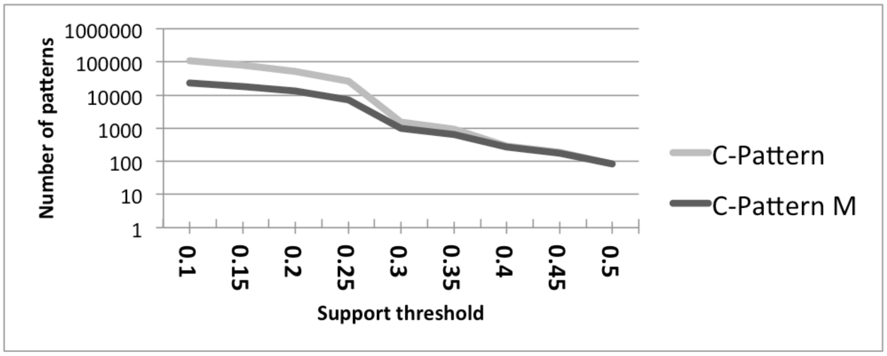

Figure 4: The number of patterns extracted by the *C-pattern* algorithm decreasing the support threshold.

To assess the meaningfulness of the C-Patterns extracted, we performed an experiment using a *Null Model* to generate the variations in the test set simulating a random behavior. The random variations are generated following two rules: (i) an event is generated every hour with a probability of 1/3 (derived by the average number of item per sequence in the real data); (ii) the intensity of an event is generated using a Gaussian distribution with average 0 and variance 1.49 giving then a high probability to generate small events and a low probability to generate a big events (the value are derived by the real data).

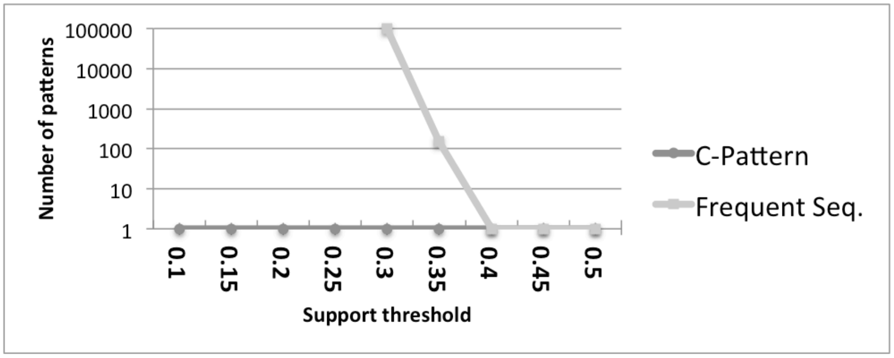

Figure 5: The frequent sequences and the *C-Patterns* extracted using a *Null Model* hypothesis on the data.

To better understand what happen using the null model, we report in Figure 5 the number of frequent sequences, without any constraint in the correlation, and the number of C-Patterns discovered using a correlation threshold equal to 50. Having no frequent sequences and no C-Patterns with high values of support proves that the patterns discovered in the real data are meaningful and not derived by random fluctuation. Moreover, for low values of support the number of frequent sequences becomes very high, but clearly the concept of correlation strongly discriminates meaningful patterns, reducing the result to 0 (which is the expected result).

5 Case studies

In this section, the presented methodology is applied to extract patterns from two different real datasets. The first experiment is applied on a urban area using data from the Paris area, when the stop computation algorithm is used to estimate the presence of people in the different city areas as a base for the correlation pattern algorithm. In the second experiment, the method is applied at a national level using data from the French territory. Both experiments have shown how the *C-pattern* algorithm can be used to find relevant mobility patterns.

In the experiments, a specific set of parameters for each step of data-preprocessing and *C-pattern* algorithm executions are used; these parameters are obtained from empirical studies and incremental tests in order to obtain satisfying results in terms of patterns quantity and semantic. As presented in the following experiments, other constraints exposed in section 3.2 are used, which help us to tune the computation, reducing the number of patterns and leading to a more understandable result.

5.1 Urban level: Paris

In this experiment the presented methods are applied on a dataset, which covers the Paris area. Figure 6 shows the position of the cell towers as well as the spatial partition of the space we have computed, using a Voronoi tessellation.

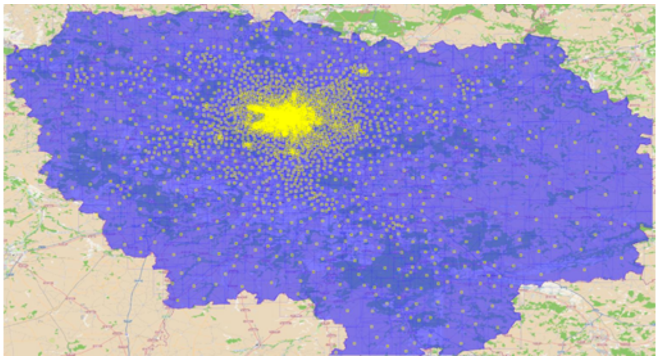

Figure 6: The distribution of the cell towers over the Paris area. We notice the presence of areas with high density of towers, centered in the city of Paris and including several smallest dense spots around it. A Voronoi tessellation is used to estimate the covered area for each cell tower.

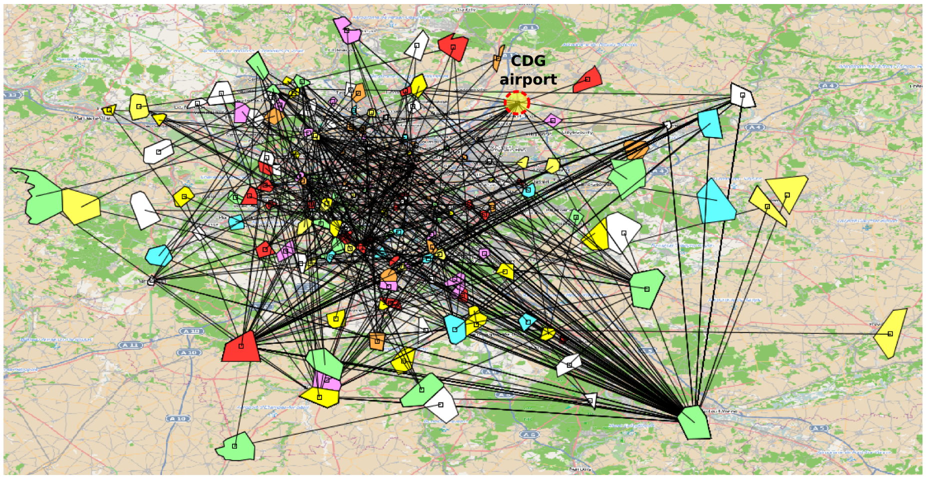

Figure 7: A set of *C-patterns* extracted over the Paris area. The patterns cover several regions well distributed both within the city and in the neighboring areas. The figure highlights the CDG airport, used later as case study.

Starting from this partition of the space, events are extracted by using a granularity of 5% with a relative minimum limit of 10%. Also, an absolute minimum limit of 50 calls was enforced, in order to eliminate noisy areas during the computation. Then, the *C-pattern* algorithm using the following parameters is executed:

$$\mathit{Min\_support}=.3 \wedge \mathit{Min\_size}=2 \wedge$$
$$\text{Max\_Gap} = 2 \wedge \text{Min\_correlation} = 50 \wedge$$
$$\text{Maximals} = \text{true}$$

The resulting patterns are shown in Figure 7. The lines connect the areas belonging to the same pattern. This representation gives both a general idea of the connections, and the areas which are involved in the result. Furthermore, we focus on a place that is important for a particular application and analyze all the patterns connected with it. This is the case of the *C-patterns* shown in Figure 8 where all the patterns starting from the Charles de Gaulle airport (CDG) are selected. This result shows how much all the major train stations are influenced by the airport, e.g., an increasing of 10% from in CDG airport impacts on Gare de l'Est train station by a 10% in 2 hours at maximum. For better understanding of the phenomena, the temporal dimension of patterns is analyzed since this chain of events may happen in different periods of a day.

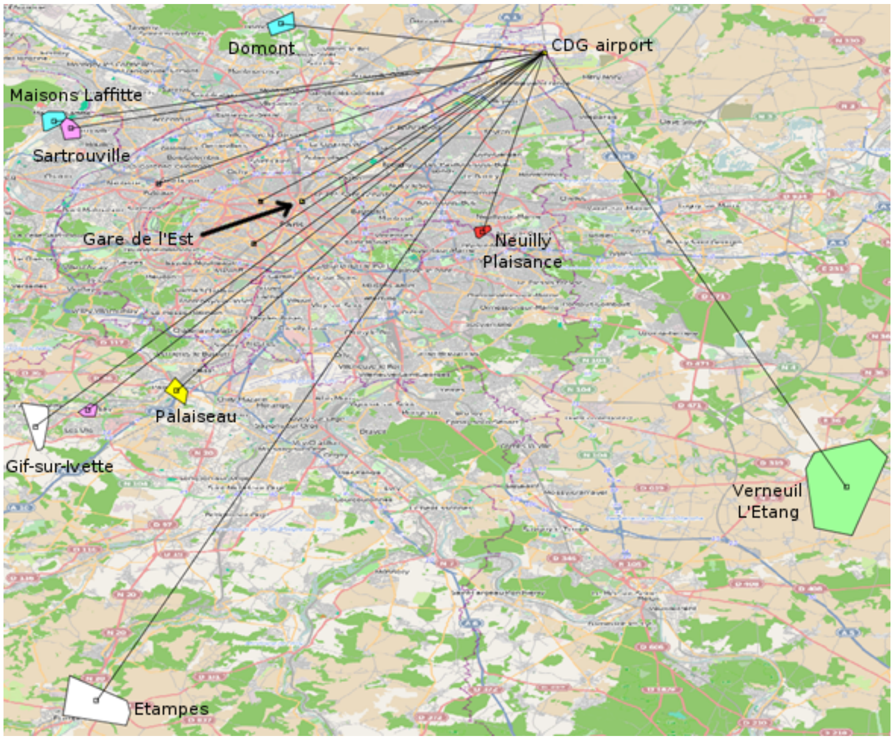

Figure 8: A selection of the C-patterns shown in Figure 7 containing all patterns starting from CDG airport.

Using the method presented in section 3, we built the dendrogram of the selected set of patterns shown in Figure 9 with the representative time distribution computed on the root of the tree. The time distribution shows that the patterns found occur mainly during daylight (the exceptions are moderately high values in the evening up to 9 p.m.), with major peaks at 10 a.m. and between noon and 1 p.m. Since the time distribution contains some variations around the early morning and the evening, we can inspect the tree, analyzing the nodes at the first level, i.e., V_14 (the one containing most patterns) and V_24. As shown (Figure 10), the distribution of V_24 is very similar to the root one, thus following the overall general trend, while the other one presents some differences. In particular, a large morning peak emerges at 8 a.m., followed by several smaller peaks throughout the day, the last being at 8 p.m., i.e. one hour earlier than the overall distribution. Also, this last peak is significantly higher than the corresponding one in in Figure 9. Figure 10 also reports the spatial distribution of the patterns corresponding to the two sub-trees selected before.

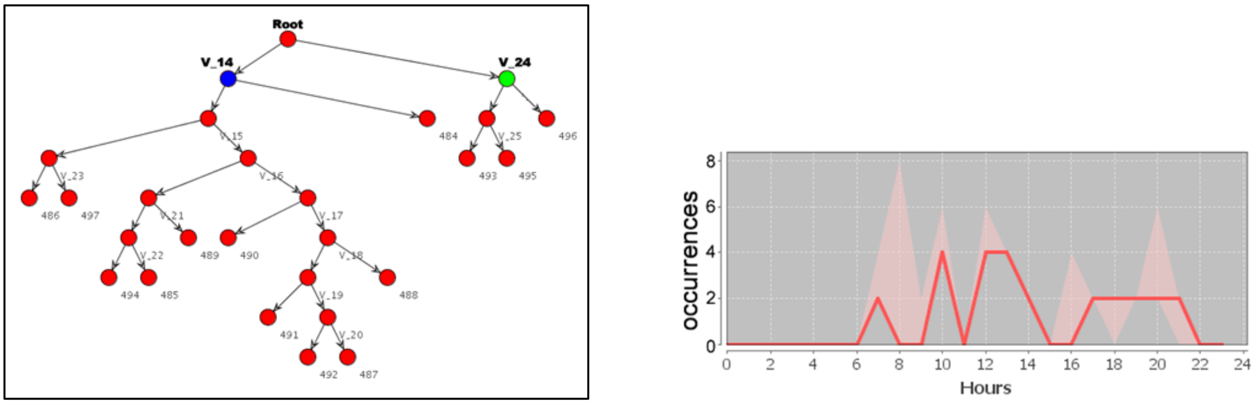

Figure 9: The dendrogram of the patterns shown in Figure 8 and the temporal distribution of the root node.

Figure 10 points out how the major train stations are mostly in the sub-tree routed in V_14 and characterized by the high peek in the morning (and, also, by a relatively high one in the evening). The last part of the analysis is the computation of Transfer measure. Here, all the patterns present a high value that clearly states that all the patterns are supported by real flow of people; e.g., the pattern between the CDG airport and Gare de l'Est has value of 0.18%. This experiment shows how the *C-patterns* can be used to understand the connections between different areas of the city and how the patterns can further be used to discover useful knowledge.

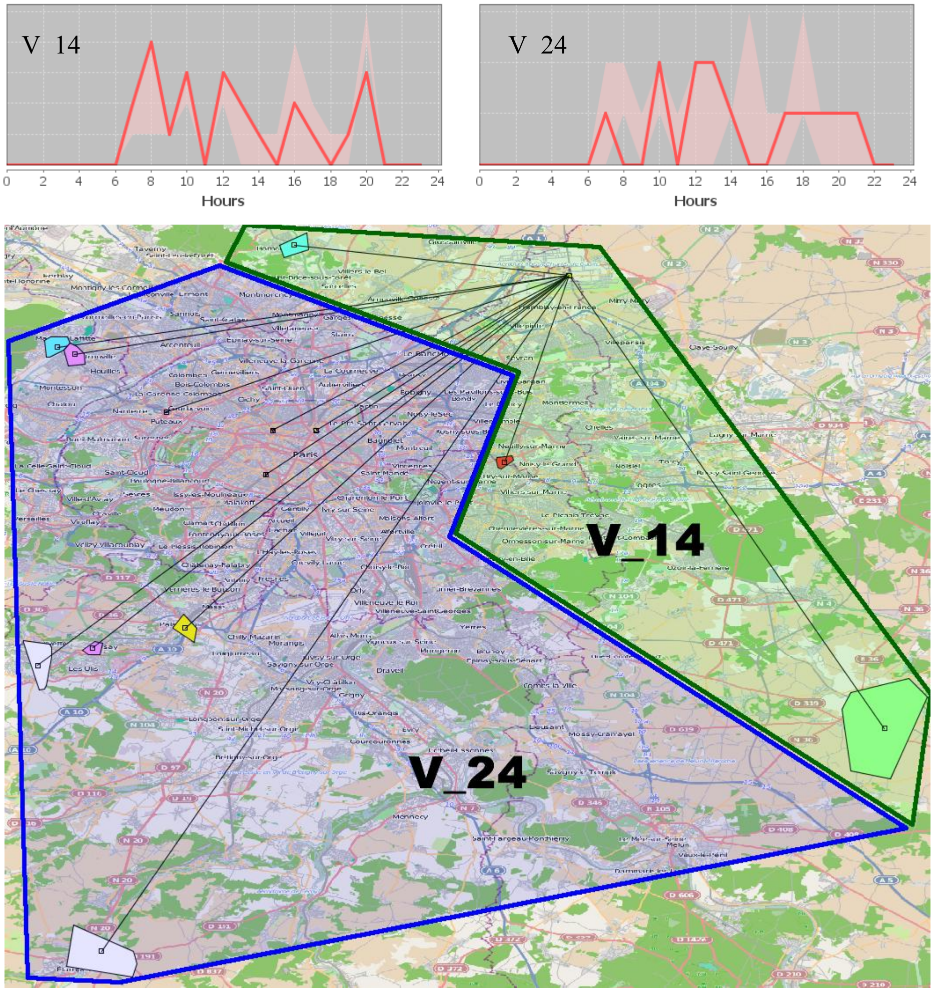

Figure 10: The distribution in time of the nodes V_14 (top left) and V_24 (top right), and the corresponding geographical areas covered by the patterns belonging to them.

5.2 National level: France

In this second experiment, a wider area (the French territory) is considered. Here, we look for patterns highlighting interconnections between different spatial units: departments or cities. As before, we start by defining the spatial division we are interested in. For this, the available information is represented by the geographical position of towers and the zip code of the area where they are located (Figure 11, left). The Voronoi tessellation, used in the first experiment (Paris Region), is not useful in this case because the fragmentation of the area would be too large for the analysis at this level. In fact, an important number of patterns with small spatial resolution would be obtained, being unable to bring out the connections between areas at national level. Thus, the analyses are carried out at two higher granularities: Departments and Cities.

5.3 Departments granularity

In France, the department is an administrative division; it is one of the three administratifs levels between the region and the municipalities ("commune") (French municipalities are roughly equivalent to civil townships in USA). The departments are identified by the first two digits of the postal code.

For each set of towers belonging to the same department, a convex hull is built (see Figure 11, right). Due to a spatial approximation, some areas intersect each other, but it is important to point out that the number of calls belonging to the set of towers in each area does not overlap. From this division of the space the events using a granularity of 0.1% with a relative minimum limit of 0.5% and an absolute minimum limit of 500 calls are defined.

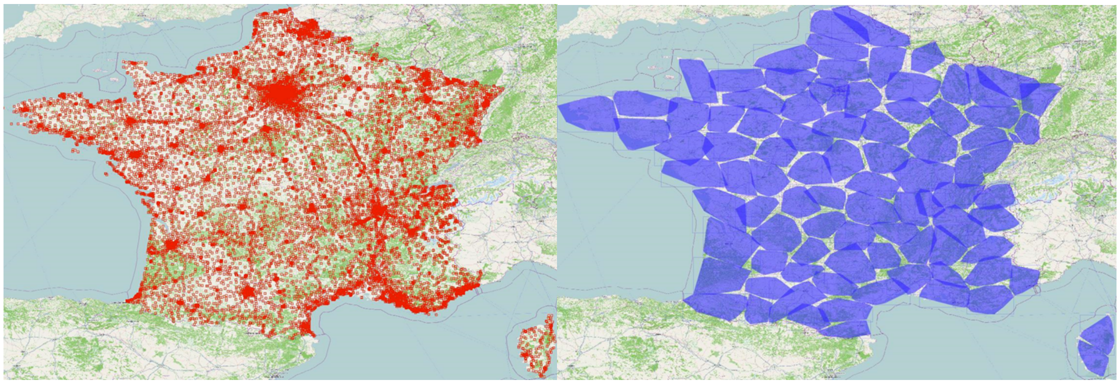

Figure 11: (left) distribution of cell towers over France; (right) convex hulls of the cell towers of each department.

Then, the *C-pattern* algorithm using the following parameters is launched:

Min_support=.3∧*Min_size*=2∧
Max_Gap = 4 ∧Min_Correlation = 100 ∧
Max_spatial_gap = 400 km ∧Maximals = true

The resulting patterns are shown in Figure 12. At this level, the discovered patterns follow the main roads and rails in France. Any pattern between departments having important airports was not found, which was an unexpected result. Apparently, this would suggest that air traffic tends to have only little abnormal fluctuations. On the contrary, we notice that the presence of an important airport in a department has some consequences on others departments that do not have airports. Figure 13, shows an example of this case. For example, by zooming on *Seine-Saint-Denis* department, a department in the immediate vicinity of Paris where Charles de Gaul airports is localized, we can see that it is connected with eight departments with different temporal distributions. We notice that more than one pattern exists for each pair of origin and destination. In fact they contain different values for the increment of calls, for this reason in the temporal distributions a representative and the variations are shown using the same technique shown above but grouping the patterns only by using the spatial component. Deepening the analysis, we can notice that these groups can be divided in two categories:

Incoming: the groups of patterns *a,b,c,d,e* move from the neighborhood departments to *Seine-Saint-Denis.*

Out-coming: the groups of patterns *f,g,h* move away from *Seine-Saint-Denis*.

These two categories have also similarity in the distributions: in fact the incoming one is more present during mid-day or early afternoon; the out-coming is more present during the evening. Since a C-pattern describes a sequence of events with a given temporal order, the distributions found suggest that the fluctuations in Seine-Saint-Denis that led to find the interconnections between these regions simply tend to occur between mid-day and afternoon. The temporal order mentioned above, then, naturally causes a slight anticipation when Seine-Saint-Denis is the starting point of the pattern, and a slight delay when it is the ending point. The groups *e* and *f* are a sort of exception, they have a two-peeks distribution which are typical of the systematic movements of an area.

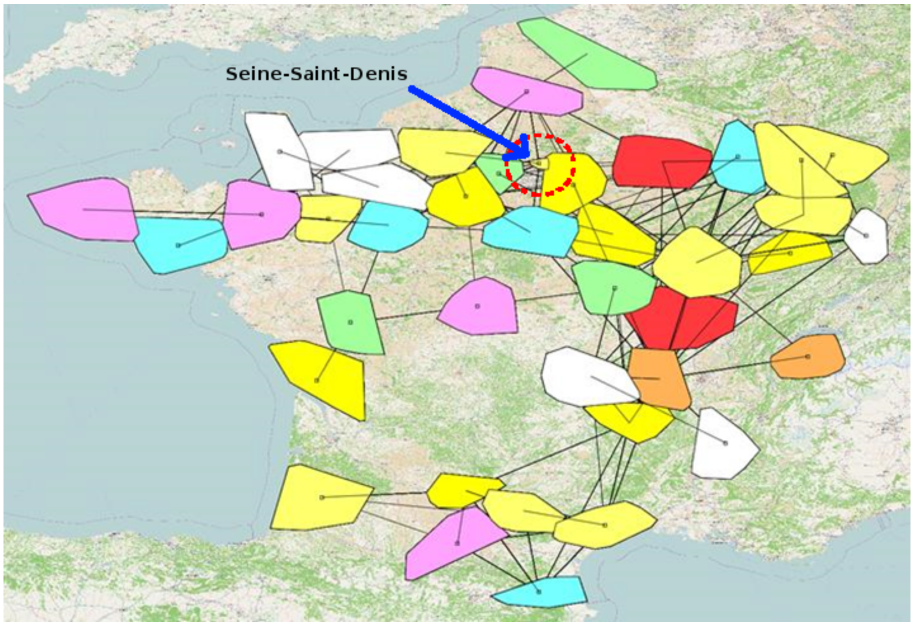

Figure 12: The set of patterns extracted on departments. Seine-Saint-Denis is highlighted.

Studying the shared users between the areas involved in the patterns, only some of them in the group *e* show a Transfer relation. In particular, the pattern with a variation of 1% in the department of Eure in the North-West and 2.5% in close to Paris Seine-Saint-Denis department has a measure value of 0.05%. Nevertheless it is clear that, at this granularity, the flow represents a very small percentage of the people compared to the presence of users. As described in section 3.3.3 this means that this pattern is consequence of a flow of people moving from one area to another. For other patterns the causes are not clearly identified, but this does not reduce the importance of the interconnection between the areas involved in the patterns.

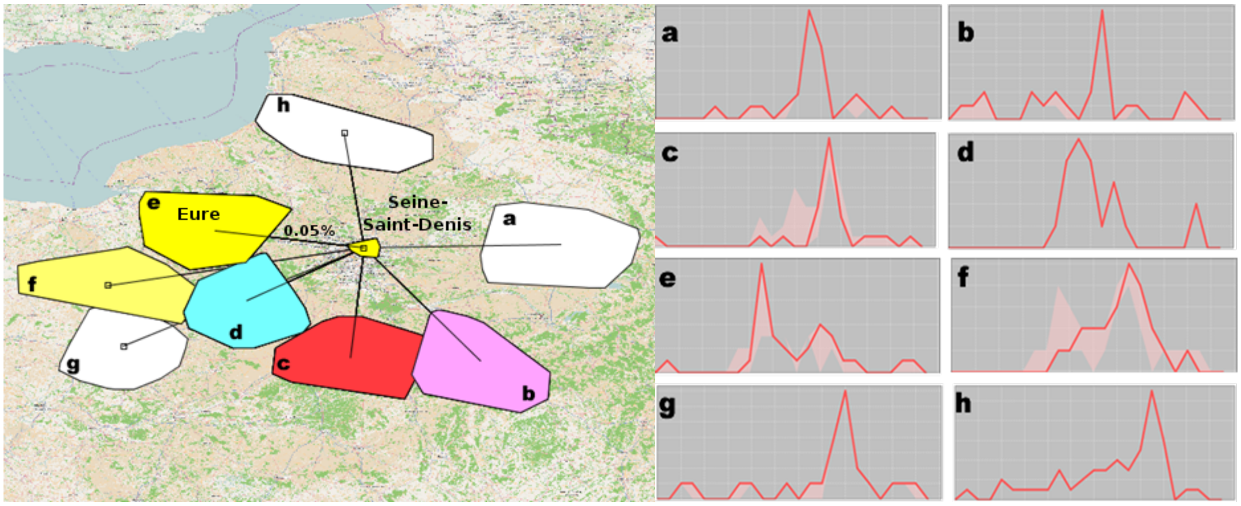

Figure 13: The patterns between Seine-Saint-Denis and the neighborhood departments. Each department is assigned to a letter, the same used on the right to represent the corresponding time distribution. The Eure department is also highlighted, together with the value of its corresponding Transfer relation from Seine-Saint-Denis.

5.4 Cities granularity

In the last experiment, our goal is to discover patterns related to cities instead of bigger regions. Thus, we focus on a different space partition. Since data do not contain information about the localization of the towers for each city, we implement a heuristic to calculate this correspondence. Considering the fact that the distribution of the towers in space is highly dependent on population density, we assume that a group of towers concentrated in a specific area represent a city or at least urban agglomerate. Starting with this assumption, in order to find these dense groups of towers a density based clustering method is used (Andrienko et al., 2009). The partition results are shown in Figure 14. It has been compared with the list of major cities in France confirming a good approximation. The events was built using a granularity of 0.1% with a relative minimum limit of 0.5% and an absolute minimum limit of 100 calls. The following parameters were used for *C-pattern* algorithm:

Min_support=.2∧*Min_size*=2∧
Max_Gap = 4 ∧Min_Correlation = 150 ∧
Max_spatial_gap = 400 km ∧Maximals = true

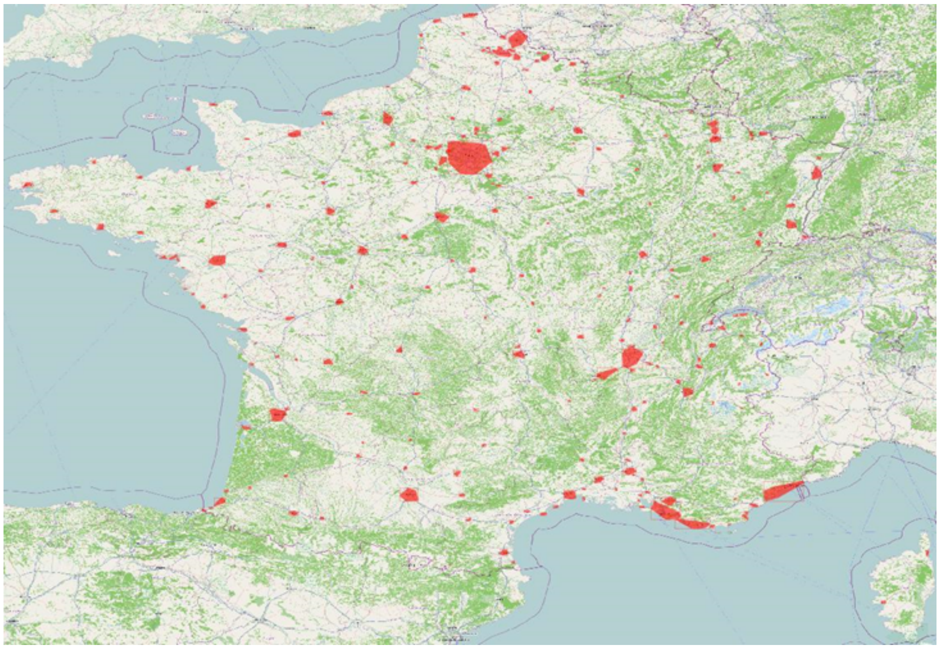

Figure 14: The convex hulls formed by the towers belonging to the same cluster of towers, (i.e. same city).

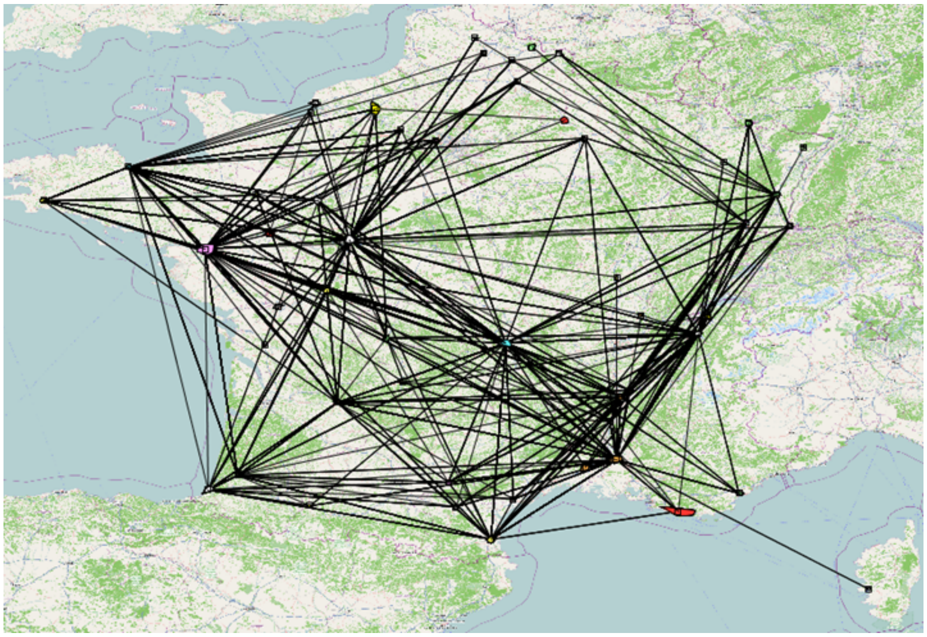

Figure 15 : An example of patterns extracted at city level.

The results are shown in Figure 15. Compared with the previous experiment, it is interesting to see how some new patterns emerge and how some of them disappear. At this level, we notice that the pattern do not respect the transversal axis as at the department level and Paris does not polarize any more the territory. Figure 17 shows an example of pattern from *Nantes* and *Tours* with its time distribution associated. It highlights how the two cities are interconnected during the early morning with an increment of 5% for both of them. As in the previous analysis, at this level it is easy to find no *Transfer relation*. For the selected pattern, even with a high support value of 0.65 and a correlation of 319, the measure is equal to 0.02%. However, the relatively large distance between the two cities (around 200 km) and the presence of highways connecting them effectively, suggests that the most likely cause of the logical connection between these regions is the exchange of population flows, together with the presence of recurrent fluctuations of such population. No symmetric patterns were found with the parameter setting adopted here, meaning that either the flows in the opposite direction are less compact (for instance, more distributed during the day) or they are less subject to fluctuations.

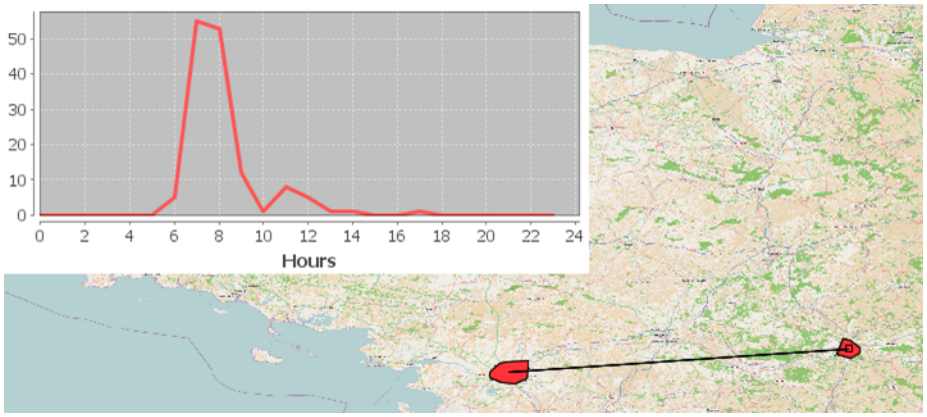

Figure 16: An example of patterns extracted at city level, connecting Nantes (on the left) and Tours (on the right), together with the time distribution of the pattern.

6 Conclusion

A new kind of pattern called *C-pattern*, aiming to discover hidden logic of connections between regions of a city (or other kinds of areas, at different scales), by analyzing frequently co-occurring changes in population densities is presented in this paper. Being based on spatiotemporal aggregations of presence, such approach overcomes typical completeness limitations of CDR data, due to the poor sample rate under which the locations of a single user are monitored. We have developed an algorithm to extract these patterns and efficient tools to organize the patterns in an automatic way, helping the analyst to browse the patterns and to better understand and interpret them. Furthermore, we have presented three different case studies using two different granularities: the urban level and national level, showing examples of data preparation, pattern extraction and pattern interpretation.

References

Agrawal, R. & Srikant, R. (1995). Mining sequential patterns. *Proc. 11th International Conference on Data Engineering,* Taipei, pp. 3-14.

Ahas, R., Silm, S., Jrv, O., & Saluveer, E. (2010). Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology,* 17, pp. 3-27.

Ahas, R., Tiru, M., Saluveer, E., & Demunter, C. (2011). Mobile telephones and mobile positioning data as source for statistics: Estonian experiences. *New Techniques and Technologies for Statistics conference,* Brussel, Feb. 2011.

Alvares, L.O., Bogorny, V., Kuijpers, B., de Macedo, J.A.F., Moelans, B., & Vaisman, A. (2007). A model for enriching trajectories with semantic geographical information. *ACM-GIS,* ACM Press, pp. 162-169.

Andrienko, G., Andrienko, N., Hurter, C., Rinzivillo, S., & Wrobel, S. (2011). From movement tracks through events to places: Extracting and characterizing significant places from mobility data. *Proc. of IEEE Visual Analytics Science and Technology (VAST 2011)*, IEEE Computer Society Press, pp.161-170.

Andrienko, G., Andrienko, N., Rinzivillo, S., Nanni, M., Pedreschi, D., & Giannotti, F. (2009). Interactive visual clustering of large collections of trajectories. *Proc. of IEEE Visual Analytics Science and Technology (VAST 2009)*, IEEE Computer Society Press, pp. 3-10.

Bar-Gera, H. (2007). Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel. *Transportation Research Part C*, 15, pp. 380-391.

Bjornstad, O.N., Ims, R.A., & Lambin, X. (1999). Spatial population dynamics: Analyzing patterns and processes of population synchrony. *Trends in Ecology and Evolution,* 14, pp. 427–432.

Buard, E. (2011). Pratiques spatiales des populations animales: analyses par les trajectoires. *Dixièmes Rencontres de Théo Quant,* Besançon, Feb. 2011. Retrieved from <http://thema.univ-fcomte.fr/theoq/pdf/resumes/TQ2011%20RESUMES.pdf>.

Calabrese, F., Di Lorenzo, G., Liu, L., & Ratti, C. (2011). Estimating origin-destination flows using mobile phone location data. *Pervasive Computing*, 10, pp. 36-44.

Cao, H., Mamoulis, N. , & Cheung, D.W. (2005). Mining frequent spatio-temporal sequential patterns. *Fifth IEEE International Conference on Data Mining*, Houston, TX, Nov. 2005.

Couronné, T., Olteanu-Raimond, A.M., & Smoreda, Z. (2011). Looking at spatio-temporal city dynamics through mobile phone lenses. *Proc.of IEEE International Conference of Network of the Future,* Paris, Nov. 2011, pp. 128-134

Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., & Trasarti R. (2011). Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal*, 20, pp. 695-719.

Giannotti, F., Nanni, M., Pinelli, F., & Pedreschi, D. (2007). Trajectory pattern mining. *Proc. of 13th ACM SIGKDD international conference on Knowledge discovery and data mining,* pp. 330–339.

González, M., Hidalgo, C., & Barabási, A.L. (2008). Understanding individual human mobility patterns. *Nature,* 453(7196), pp. 779–782.

Isaacman, S., Becker, R. , Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., and Varshavsky, A. (2011). Identifying important places in people's lives from cellular network data. *Pervasive Computing. Lecture Notes in Computer Science,* 6696, pp: 133–151.

Olteanu, A.M., Trasarti, R., Couronné, T., Giannotti, F., Nanni, M., Smoreda, Z. & Ziemlicki, C. (2011). GSM data analysis for tourism application. *7th International Symposium on Spatial Data Quality,* Coimbra, October 2011.

Palma, A. T., Bogorny, V., Kuijpers, B., & Alvares, L.O (2008). A clustering-based approach for discovering interesting places in trajectories. *ACMSAC,* ACM Press, pp. 863-868.

Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R., & Ratti, C. (2010). Activity-Aware Map: Identifying human daily activity pattern using mobile phone data. *Proc. of International Conference on Pattern Recognition, Workshop on Human Behavior Understanding,* Springer, Heidelberg, pp. 14-25.

Ramm, K & Schwieger, V. (2007). Mobile positioning for traffic state acquisition. *Journal of Location Based Services,* 1, pp. 133-144.

Ratti, C., Sevtsuk, A., Huang, S., & Pailer, R. (2005). Mobile Landscapes: Graz in real time. Retrieved from <http://senseable.mit.edu/papers/pdf/RattiSevtsukHuangPailer2005LBSVienna.pdf>

Reades, J., Calabrese, F., Sevtsuk, A., & Ratti, C. (2007). Cellular Census: Explorations in urban data collection. *IEEE Pervasive Computing,* 6, pp. 30-38.

Rocha, J., Oliveira, G., Alvares, L., Bogorny, V., & Times, V. (2010). A direction-based spatio-temporal clustering method. *Proc. of 5th IEEE International Conference Intelligent Systems, London,* pp. 114-119.

Song, C., Qu, Z., Blumm, N., & Barabasi, A.L. (2010). Limits of predictability in human mobility. *Science,* 327(5968), pp. 1018–1021.

Spaccapietra, S., Parent, C., Damiani, M.L., De Macedo, J.A , Porto, F., & Vangenot, C. (2008). A conceptual view on trajectories. *Data and Knowledge Engineering,* 65, pp. 126-146.

Tatem, A.J., Qiu, Y., Smith, D. L., Sabot, O., Ali, A. S., & Moonen, B. (2009). The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents. *Malaria Journal,* 8:287, doi:10.1186/1475-2875-8-287.

Traag, V.A., Browet, A., Calabrese, F., & Morlot, F. (2011). Social event detection in massive mobile phone data using probabilistic location inference. *Proc. of IEEE 3rd International Conference on Social Computing (Socialcom)*, pp. 625- 628.

Wang, H., Calabrese, F., Di Lorenzo, G., & Ratti, C. (2010). Transportation mode inference from anonymized and aggregated mobile phone Call Detail Record. *13th International IEEE Annual Conference on Intelligent Transportation Systems*, Madeira Island, Sept. 2010.

Yan, Z., Parent, C., Spaccapietra, S., & Chakraborty, D. (2010). A hybrid model and computing platform for spatio-semantic trajectories. *7th Extended Semantic Web Conference (ESWC)*, Heraklion, Greece, May 2010.

Zimmermann, M., Kirste, T., & Spiliopoulou, M. (2009). Finding stops in error-prone trajectories of moving objects with time-based clustering. *Communications in Computer and Information Science,* 53, pp. 275-286.