Country-Scale Exploratory Analysis of Call Detail Records Through the Lens of Data Grid Models

Romain Guigourès[1], Dominique Gay[2], Marc Boullé[2(✉)], Fabrice Clérot[2], and Fabrice Rossi[3]

[1] Zalando, Berlin, Germany
[2] Orange Labs Lannion, Lannion, France
[3] SAMM EA 4543, Univeristé Paris 1, Paris, France
marc.boulle@orange.com

Abstract. Call Detail Records (CDRs) are data recorded by telecommunications companies, consisting of basic informations related to several dimensions of the calls made through the network: the source, destination, date and time of calls. CDRs data analysis has received much attention in the recent years since it might reveal valuable information about human behavior. It has shown high added value in many application domains like e.g., communities analysis or network planning.

In this paper, we suggest a generic methodology based on data grid models for summarizing information contained in CDRs data. The method is based on a parameter-free estimation of the joint distribution of the variables that describe the calls. We also suggest several well-founded criteria that allows one to browse the summary at various granularities and to explore the summary by means of insightful visualizations. The method handles network graph data, temporal sequence data as well as user mobility data stemming from original CDRs data. We show the relevance of our methodology on real-world CDRs data from Ivory Coast for various case studies, like network planning strategy and yield management pricing strategy.

Keywords: Classification rule · Bayes theory · Minimum description length

1 Introduction

Telco operators' activities generate massive volume of data, mainly from three sources: networks, service platforms and customers data bases. Particularly, the use of mobile phones generates the so called Call Detail Records (CDRs), containing information about end-point antenna stations, date, time and duration of the calls (the content of the calls is excluded). While this data is initially stored for billing purpose, useful information and knowledge (related to human

Romain Guigourès was with Orange Labs when this work began.

A. Bifet et al. (Eds.): ECML PKDD 2015, Part III, LNAI 9286, pp. 37–52, 2015.
DOI: 10.1007/978-3-319-23461-8_3

mobility [1,23], social interactions [22] and economic activities) might be derived from the large sets of CDRs collected by the operators.

Recent studies have shown the potential added-value of analyzing such data for several application domains: United Nations Global Pulse [21] sums up some recent research works on how analysis of CDRs can provide valuable information for humanitarian and development purposes, e.g., for disaster response in Haiti, combating H1N1 flu in Mexico, etc. Also, leveraging country-scale sets of CDRs in Ivory Coast, the recent Orange D4D challenge (Data For Development [5]) has given rise to many investigations in several application domains [4] such as health improvement, analysis of economic indicators and population statistics, communities understanding, city and transport planning, tourism and events analysis, emergency, alerting and preventing management, mobile network infrastructure monitoring. Thus, the added-value of analysis of CDRs data does not need to be proved any longer.

Various classical data mining techniques [4] have been applied on CDRs data depending on the features and the task considered: e.g., considering network graphs from (source antenna, destination antenna) data or temporal sequences from (source antenna, date) data appeals for different clustering techniques for summarizing information in the data.

Contribution: in this paper, we suggest an efficient and generic methodology for summarizing CDRs data whatever the features are retained in the analysis. The method is based on data grid models [6], a parameter-free joint distribution estimation technique that simultaneously partitions sets of values taken by each variable describing the data (numerical variables are discretized into intervals while the categories of categorical variables are grouped into clusters). The resulting data grid – that can be seen as a coclustering – constitutes the summary of the data. The method is thus able to summarize various types of data stemming from CDRs: network graph data, temporal sequence data as well as user mobility data. We also suggest several criteria *(i)* to exploit the resulting data grid at various granularities depending on the needs of analysis and *(ii)* to interpret the results through meaningful visualizations. The whole methodology aims at demonstrating strong impacts on two key points on economic strategy: network planning and pricing strategy.

Outline: in the next section, we discuss further recent work related to CDRs and mobile phone trace data analysis as well as data grid models. In Section 3, we summarize the impacts of the various case studies on the economic development strategy related to the specific context of telecommunications in Ivory Coast. A brief description of the CDRs data characteritics is also given. Section 4 recalls the main principles of data grid models and introduces the tools for exploiting the resulting data grid. In section 5, we report the experimental results on the various case studies.

2 Related Work

CDRs data have received much attention in recent years. Famous applications of CDRs data analysis are for the benefit of social good: e.g., in the transportation

domain, [2] suggest a system for public transport optimization; in the health domain, e.g., [8] suggest a model for epidemic spread.

Mobile phones may also provide other types of data (e.g., the Nokia Mobile Data Challenge [15]), like applications events, WLAN connection data, etc. For instance, [13] pre-processed phone activities of one million users to obtain information about their approximative temporal location, then mined daily motifs from the spatio-temporal data to infer human activities. Finally, smart phones are or will be equipped with accelerometers and/or gyroscopes providing data about physical activities of users: [16] suggest a complete system of activity recognition based on smartphone accelerometers with potential application to health monitoring.

Research work related to data grid models: We are *not* coclustering data (objects $\times$ attributes) like pioneering work of Hartigan [12]. Data grid models are related to the work of Dhillon et al. [7] who have proposed an information-theoretic coclustering approach for two discrete random variables Y_1 and Y_2: the loss in Mutual Information $MI(Y_1, Y_2) - MI(Y_1^M, Y_2^M)$ is minimized to obtain a locally-optimal grid with a user-defined number of clusters for each dimension. This is limited to two variables and requires to choose the number of clusters per variable. Going beyond 2D matrices, recent significant progress has been done in multi-way tensor analysis [14,19]. Dealing with k-adic data, (also known as co-occurrence data, like contingency table), [17] suggest a coclustering method for social network and temporal sequence (with pre-discretization of time).

The Information Bottleneck (IB) method [20] stems from another information-theoretic paradigm: given the joint probability $P(X, Y)$, IB aims at grouping X into clusters T in order to both compress X and keep as much information as possible about Y. IB also minimizes a difference in Mutual Information: $MI(T, X) - \beta MI(T, Y)$, where β is a positive Lagrange multiplier. Wang et al. [24] build upon IB and suggest a coclustering method for two categorical variables. Extending IB for more than two categorical variables, Slonim et al. [18] have suggested the agglomerative multivariate IB that allows constructing several interacting systems of clusters simultaneously; the interactions among variables are specified using a Bayesian network structure.

To the best of our knowledge, our summarization approach is the only one to combine the following advantages: it is parameter-free, scalable and can be applied to mixed-type attributes (categorical, numerical, thus multiple types of time dimensions without pre-processing). Therefore, the same generic method can be used to analyze network graph, temporal sequence and mobility data.

3 Impacts on Economic Strategy

Besides the high-level knowledge extracted from country-scale data and confirmed by local sociologists from the University of Bouaké in Ivory Coast, these studies have also a strong impact on future economic development strategy, mainly in two identified branches:

- *Network planning strategy:* In 2014, there are around 20M inhabitants in Ivory Coast and the mobile service penetration rate is $\simeq$ 84% – with a still growing mobile phone market in a context of demographic growth. The analyses of the first two case studies and the resulting map projections (that can be seen as the network of calls available at various granularities, see Sections 5.1 and 5.2) are considered as an additional input for network planning and investment; for instance to help network designer in answering questions about how many and where the next antennas have to be set while preserving the quality of service at a reasonable cost.
- *Yield management pricing strategy:* a part of the pricing policy, called *Bonus Zone*, established in Ivory Coast offers discount prices (from 10% to 90%) to calling users depending on the location and hour of the emitting call. Maps and calendars resulting from the last two case studies on temporal distribution of output calls (see Section 5.3) and on mobility data (see Section 5.4) that are available at various granularities, provide valuable information to economic analysts in order to design optimized spatio-temporal pricing policy in the context of *Bonus Zone*.

Data Description and Studies. The CDRs data under study come from the Orange D4D challenge[1] (Data For Development [5]). We consider several case studies on two anonymized CDRs data sets from Ivory Coast, namely communication data and mobility data:

Case studies on communication data. Communication data consists in 471 millions mobile calls and covers a 5-month period (from 2011, December 1st to 2012, April 28th). The records are described by the four following variables: *emitting antenna* (1214 categorical values); *receiving antenna* (1216 categorical values); *time of call* (with hour precision); *date of call* (from 2011/12/01 to 2012/04/28). From this data set, we consider three subsets for:

1. *Analysis of call network between antennas.* Considering emitting antennas, receiving antennas and the calls made between antennas, the data set can be seen as a directed multigraph where nodes are antennas and links are the calls between antennas.
2. *Analysis of output traffic w.r.t. date of call.* We consider emitting antennas and the number of days for each call from referral to first day of recording. This data set can be considered as a temporal event sequence spanning over the whole observation period, where the time is the number of days passed and the events are the emitting antenna IDs.
3. *Analysis of output traffic w.r.t. week day and hour of call.* We consider emitting antennas, the day of the week (stemming from the date and considered as a numerical variable) and the hour of the day for each call. Here the time dimension is represented by two variables and the data of the whole period are folded up to week day and hour.

[1] http://d4d.orange.com/en/home

Case studies on mobility data. Mobility data consists in mobility traces of 50000 users over a 2-week period (from 2012 December 12th to 2012 December 24th), i.e. approximatively 55 millions records. The records are described by the four following variables: *anonymized user ID* (50000 categorical values); *connexion antenna* (1214 categorical values); *time of call* (minute precision); *date fo call* (from 2012/12/12 to 2012/12/24).
From this data, we consider the user trajectories (identified by user ID) inside the network for the following analysis:

1. *Analysis of user mobility w.r.t. week day and hour.* We consider the user ID, antennas, week day and hour. This data set can be considered as a set of spatio-temporal footprints, where each user ID is associated with a sequence of antenna usage over the time dimension. Here again, the time dimension is represented by two variables and the data of the whole period is folded up to week day and hour.

4 Exploratory Analysis through Data Grid Models

Data grid models aim at estimating the joint distribution between K variables of mixed-types (categorical as well as numerical). The main principle is to simultaneously partition the values taken by the variables, into groups/clusters of categories for categorical variables and into intervals for numerical variables. The result is a multidimensional (K-d) data grid whose cells are defined by a part of each partitioned variable value set. Notice that in all rigor, we are working only with partitions of variable value sets. However, to simplify the discussion we will sometime use a slightly incorrect formulation by mentioning a "partition of a variable" and a "partitioned variable".

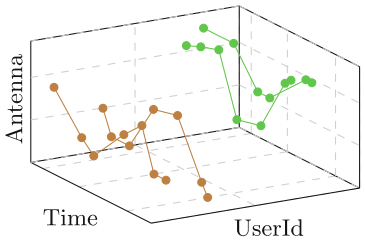
In order to choose the "best" data grid model M^* (given the data) from the model space $\mathcal{M}$, we use a Bayesian Maximum A Posteriori (MAP) approach. We explore the model space while minimizing a Bayesian criterion, called cost. The cost criterion implements a trade-off between the accuracy and the robustness of the model and is defined as follows:

$$cost(M) = -\log(\underbrace{p(M \mid D)}_{\text{posterior}}) \propto -\log(\underbrace{p(M)}_{\text{prior}} \times \underbrace{p(D \mid M)}_{\text{likelihood}})$$

Thus, the optimal grid M^* is the most probable one (maximum a posteriori) given the data. Due to space limitation, the details about the *cost* criterion and the optimization algorithm (called KHC) are available in appendix of [11]. Hereafter, we focus on the tools for exploiting the grid and their applications on large-scale CDRs data. The key features to keep in mind are: *(i)* KHC is parameter-free, i.e., there is no need for setting the number of clusters/intervals per dimension; *(ii)* KHC provides an effective locally-optimal solution to the data grid model construction efficiently, in sub-quadratic time complexity ($O(N\sqrt{N}\log N)$ where N is the number of data points). Figure 1 illustrates the input data and output results of KHC on an examplary mobility data stemming from CDRs.

IdUser	Date	Time	O-Antenna	D-Antenna
u1	14/07/2014	07:25:02	A1	A25
u1	14/07/2014	07:57:22	A2	A2
u1	14/07/2014	09:05:57	A2	A16
u1	14/07/2014	10:32:16	A2	A25
...	...	...	...	...
u4	11/11/2014	22:52:32	A11	A17

(a) Toy example of CDRs data. Here, mobility data (grey columns): UserId, Time, OriginAntenna.



(b) Resulting 3D data grid with 2 clusters of users, 4 time intervals and 3 clusters of antennas.

Fig. 1. From 3D mobility data, stemming from CDRs data, to data grid.

4.1 Data Grid Exploitation and Visualization

Because of the very large number observations in CDRs data, the optimal grid M^* computed by KHC can be made of hundreds of parts per dimension, i.e., millions of cells, which is difficult to exploit and interpret. To alleviate this issue, we suggest a grid simplification method together with several criteria that allow us to choose the granularity of the grid for further analysis, to rank values in clusters and to gain insights in the data through meaningful visualizations.

Dissimilarity Index and Grid Structure Simplification. We suggest a simplification method of the grid structure that iteratively merge clusters or adjacent intervals – choosing the merge generating the least degradation of the grid quality. To this end, we introduce a dissimilarity index between clusters or intervals which characterize the impact of the merge on the *cost* criterion.

Definition 1 (Dissimilarity index). *Let $c_{\cdot 1}$ and $c_{\cdot 2}$ be two parts of a variable partition of a grid model M. Let $M_{c_{\cdot 1} \cup c_{\cdot 2}}$ be the grid after merging $c_{\cdot 1}$ and $c_{\cdot 2}$. The dissimilarity $\Delta(c_{\cdot 1}, c_{\cdot 2})$ between the two parts $c_{\cdot 1}$ and $c_{\cdot 2}$ is defined as the difference of cost before and after the merge:*

$$\Delta(c_{\cdot 1}, c_{\cdot 2}) = cost(M_{c_{\cdot 1} \cup c_{\cdot 2}}) - cost(M) \tag{1}$$

When merging clusters that minimize Δ, we obtain the sub-optimal grid M' (with a coarser grain, i.e. simplified) with minimal *cost* degradation, thus with minimal information loss w.r.t. the grid M before merging. Performing the best merges w.r.t. Δ iteratively over the K variables without distinction, starting from M^* until the null model $M_\emptyset$, K agglomerative hierarchies are built and the end-user can stop at the chosen granularity that is necessary for the analysis while controlling either the number of clusters/cells or the information ratio kept in the model. The information ratio of the grid M' is defined as follows:

$$IR(M') = \frac{cost(M') - cost(M_\emptyset)}{cost(M^*) - cost(M_\emptyset)} \tag{2}$$

where $M_\emptyset$ is the null model (the grid with a single cell).

Typicality for Ranking Categorical Values in a Cluster. When the grid is coarsen during the hierarchical agglomerative process, the number of clusters per categorical dimension decreases and the number of values per cluster increases. It could be useful to focus on the most representative values among thousands of values of a cluster. In order to rank values in a cluster, we define the typicality of a value as follows.

Definition 2 (Typical values in a cluster). *For a value v in a cluster c of the partition Y^M of dimension Y given the grid model M, the typicality of v is defined as:*

$$\tau(v,c) = \frac{1}{1 - P_{Y^M}(c)} \times \sum_{\substack{c_j \in Y^M \\ c_j \neq c}} P_{Y^M}(c_j)(cost(M|c \setminus v, c_j \cup v) - cost(M)) \quad (3)$$

where $P_{Y^M}(c)$ is the probability of having a point with a value in cluster c, $c \setminus v$ is the cluster c from which we have removed value v, $c_j \cup v$ is the cluster c_j to which we add value v and $M|c \setminus v, c_j \cup v$ the grid model M after the aforementioned modifications.

Intuitively, the typicality evaluates the average impact in terms of *cost* on the grid model quality of removing a value v from its cluster c and reassigning it to another cluster $c_j \neq c$. Thus, a value v is representative (say typical) of a cluster c if v is "close" to c and "different in average" from other clusters $c_j \neq c$. Notice that this measure does not introduce any numerical encoding of the categories of the categorical variable under study.

Insightful Visualizations with Mutual Information. It is common to visualize 2D coclustering results using 2D frequency matrix or heat map. For KD coclustering, it is useful to visualize the frequency matrix of two variables while selecting a part of interest for each of $K-2$ other variables. We also suggest an insightful measure for co-clusters to be visualized, namely, the Contribution to Mutual Information (CMI) – providing additional valuable visual information inaccessible with only frequency representation. Notice that such visualizations are also valid whatever the variable of interest.

Definition 3 (Contribution to mutual information). *Given the $K-2$ selected parts $c_{i_3 \ldots i_K}$, the mutual information between two partitioned variables Y_1^M and Y_2^M (from the partition M of Y_1 and Y_2 variables induced by the grid model M) is defined as:*

$$MI(Y_1^M; Y_2^M) = \sum_{i_1=1}^{J_1} \sum_{i_2=1}^{J_2} MI_{i_1 i_2} \text{ where } MI_{i_1 i_2} = p(c_{i_1 i_2}) \log \frac{p(c_{i_1 i_2})}{p(c_{i_1 .})p(c_{. i_2})} \quad (4)$$

where $MI_{i_1 i_2}$ represent the contribution of cell $c_{i_1 i_2}$ to the mutual information, $p(c_{i_1 i_2})$ is the observed joint probability of points in cell $c_{i_1 i_2}$ and $p(c_{i_1 .})p(c_{. i_2})$ is the expected probability in case of independence, i.e., the product of marginal probabilities.

Thus, if $MI_{i_1 i_2} > 0$ then $p(c_{i_1 i_2}) > p(c_{i_1 .})p(c_{. i_2})$ and we observe an excess interaction between $c_{i_1 .}$ and $c_{. i_2}$ located in cell $c_{i_1 i_2}$ defined by parts i_1 of Y_1^M and i_2 of Y_2^M. Conversely, if $MI_{i_1 i_2} < 0$, then $p(c_{i_1 i_2}) < p(c_{i_1 .})p(c_{. i_2})$, and we observe a deficit of interactions in cell $c_{i_1 i_2}$. Finally, if $MI_{i_1 i_2} = 0$, then either $p(c_{i_1 i_2}) = 0$ in which case the contribution to MI and there is no interaction or $p(c_{i_1 i_2}) = p(c_{i_1 .})p(c_{. i_2})$ and the quantity of interactions in $c_{i_1 i_2}$ is that expected in case of independence between the partitioned variables.

The visualization of cells' CMI highlight valuable information that is local to the $K - 2$ selected parts and bring complementary insights to exploit the summary provided by the grid.

5 Exploration Results

Each application of KHC (available at http://www.khiops.com) for the various case studies data is achieved within a day of computation on a commodity computer – which confirms the efficiency of the method.

5.1 Analysis of Call Network between Antennas

The application of data grid models on the CDRs provides a segmentation with 1150 clusters, that corresponds to nearly one antenna per cluster. This is due to the large amount of data – 471 millions CDRs. Indeed, the number of calls is so high for each antenna that the distribution of calls originating from (resp. terminating to) each antenna can be distinguished from each other. In order to

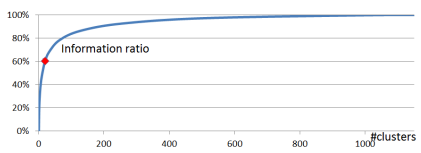


Fig. 2. Evolution of the information kept in the data grid model w.r.t. the number of clusters using the ascending hierarchical post-processing – from optimal data grid M^* (100%) to the null model $M_\emptyset$ (0%).

obtain a more interpretable segmentation, we apply the post-processing introduced in the Section 4.1. Figure 2 shows the information ratio (see definition 1) versus the number of clusters for all intermediate models obtained during the ascending hierarchical post-processing. Interestingly, the resulting Pareto curve shows that very informative models are obtained with few clusters. In our study, we decrease the number of clusters until keeping 60% of the model informativity – corresponding to 20 clusters, an admissible number for the interpretation. Throughout the simplification process, both partitions of source and target antennas stay identical. Thus we consider only the partition of source antennas for the rest of the study. Those clusters are projected on a map of Ivory Coast in Figure 3. Antennas are identified using dots, which color matches with the cluster they belong to.

The first observation is the strong correlation between the clusters and the geography of the country. Indeed, antennas from a same cluster are close to each other. The size of the clusters is almost the same in terms of area and match with the administrative zones of the country, except for Abidjan, the economic capital, which is split into four clusters. This is due to the high concentration of antennas in the city (32% of the ivorian antennas) and the dense phone traffic (34% of the calls).

We use the typicality (see definition 2) to rank the antennas of each cluster. The place, where the antenna with the highest typicality is located, is used to label the cluster. On the map in Figure 3, the size of the dots are proportional to the antenna typicality. Most typical antennas are located in the main cities of Ivory Coast. This phenomenon has already been observed in [3] and [10]: the discovered clusters match with the area of influence of the main cities of a country. We observe few exceptions: the cluster of the city of Sassandra contains the antennas of the city of Divo, while Divo is almost 4 times bigger than Sassandra (population wise) and is the sixth Ivorian city. Antennas in Divo are 40% less typical than the ones in Sassandra, meaning that allocating them to another cluster would be less costly for the criterion. Actually, calls emitted from Divo are significant in direction to other regions of Ivory Coast whereas calls from Sassandra are more internal to its region. In more formal terms, the calls distributions of the antennas in Divo are closer to the marginal distribution than to its cluster's distribution. This observation is not really surprising because Divo has experienced a recent growth of its population, due to migrations within the country [9]. Divo is also located in an area specialized in the intensive farming, that attracts seasonal workers from other parts of Ivory Coast.

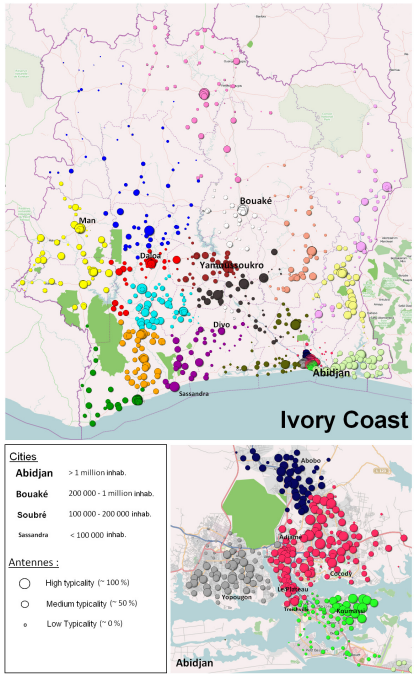


Fig. 3. Twenty clusters displayed on Ivory Coast map. There is one color per cluster.

Now, focusing on the segmentation of Abidjan: the city is divided into four parts with a strong socioeconomic correlation. The first cluster – in red in Figure 3 – covers central Abidjan, including the Central Business District (le Plateau), the transport hub (Adjamé) and the embassies and upper class area (Cocody). The second cluster – in light green – is located in the South of the

city. The covered neighborhoods are mainly residential areas and ports. Note that this cluster and the previous one are separated by a strip of sea, except for its North part that is included in the previous cluster. This very localized neighborhood matches with the party area of Abidjan. Finally, the last two clusters group antennas located in two areas with a similar profile: these are lower class neighborhoods. These clusters are separated not only because they are located in different parts of the city but especially because their call distribution differs: Abobo in dark blue and Yopougon in grey in the Figure 3.

Traffic between Clusters. Now, we analyze the distribution of calls between clusters of antennas using the contribution to the mutual information. We suggest to visualize the lacks and excesses of calls between the clusters, compared to the expected traffic in case of independence. Whatever the granularity level of the clustering, we observe a strong excess of calls from the clusters to themselves and weaker excesses and lacks between clusters. Studying the traffic within the clusters has a limited interest. We only focus on the inter-clusters traffic. To visualize the traffic between clusters, we use a finer clustering than previously. Here, we have 355 clusters for 95% informativity (see Figure 2). Figure 4 depicts the excesses of traffic between clusters – highlighted with red segments. The end points of the segments are drawn at the positions of the most representative antennas of the associated clusters (i.e with the highest typicalities). The opacity of a segment is proportional to the value of the contribution to mutual information and its width is proportional to the number of calls between clusters. The biggest cities – like Bouaké, San Pedro and Man – are clearly marked on the map: they are regional capitals, a fact that is confirmed and highlighted by the call traffic visualization. The case of Bouaké is particularly interesting: although it is not the country capital, its national influence seems bigger than the one of Yamoussoukro, the actual capital. Yamoussoukro is twice smaller than Bouaké (population wise) and is a quite recent city where there is no major economical activity, contrary to Bouaké. This fact can explain our observation.

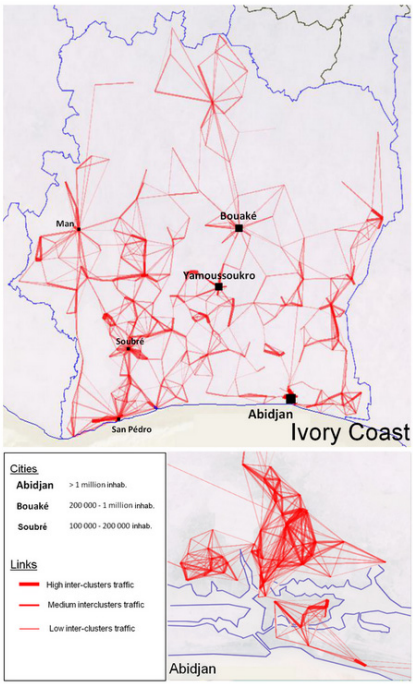


Fig. 4. Analysis of Excess of calls between clusters of antennas

We also observe that excess of traffic between major cities is a rare phenomenon. Cities are more like phone hubs, except in the West of the country around Soubré. This area is not a densely populated area but corresponds to a region with important migration flows. Finally, in Abidjan, we observe important excesses of traffic within neighborhoods, but not between neighborhoods.

5.2 Temporal Analysis of the Calls Distribution

From previous section, we learnt that the correlation between source and destination antennas is very high. The evolution of the calls distribution over time might be the same for both sets of antennas. Therefore, to track the evolution of traffic over time, we only study the evolution of the originating calls: one call is described by the emitting antenna and a day count (stemming from the date).

Again, the clustering of antennas resulting from the optimal data grid is also too fine for an easy interpretation (1051 clusters of antennas and 140 intervals for the day count). We coarse the grain of the grid with our hierarchical post-processing so that the informativity of the model is 80%, with ten clusters of antenna and twenty time segments. Since, missing values are abundant in this data, i.e., some antennas emitted no call during some time periods, consequently, we obtain time segments that are strongly correlated with missing data. For the same reason, antennas are grouped together because they experienced an absence of calls during one or several similar periods. In the Figure 5, the colored antennas belong to clusters having experienced simultaneous absences of calls. We observe that the green, orange, light blue and purple clusters are located in localized area. The missing data appear during short periods for these clusters. This grouping might be due to localized technical issues on the network. The antennas of the yellow cluster are spread over the country. These antennas are grouped because they have been activated at the same date. This use case provides a better understanding the dysfunctions in the network over the year.

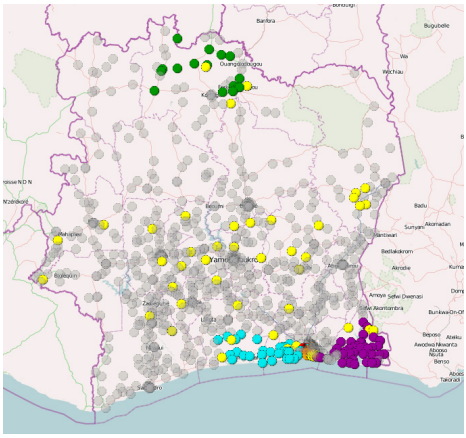


Fig. 5. Antennas activity clusters projected on Ivory Coast map. Colored clusters show inactivity periods while grey clusters indicate antennas whose traffic is complete over the period.

5.3 Analysis of Output Communications w.r.t. Week Day and Hour

Our objective is to build simultaneously a partition of the antennas, a partition of the week days and a discretization of the hour, i.e., a triclustering. For

the same reasons as previously, we only keep the emitting antennas.
At the finest level, we obtain a triclustering with 806 clusters of emitting antennas, 7 clusters of days and 22 time segments. Again, these results must be simplified to ease the interpretation. However, we fix the numbers of clusters of days and time segments, since they are acceptable for the analysis and we only reduce the number of clusters of antennas. With four clusters of antennas, we keep 51% of the informativity of the model.

Antennas are displayed on the map of Figure 6. We also build a calendar (see Figure 7) for each cluster with days in columns and time segments in lines. The color of the cells indicates the excesses (red) or the lacks (blue) of traffic emitted from the corresponding cluster. The lacks and excesses are measured using the contribution to the mutual information (see definition 3) between the cluster and the cross product of the cluster of weekday and the time segment: $MI(X_1^M; X_2^M \times X_3^M)$, with X_1^M the partition of the antennas, X_2^M the partitions of the weekdays and X_3^M the discretization of the time. Now we focus on the analysis of each cluster of antennas that we can easily label manually:

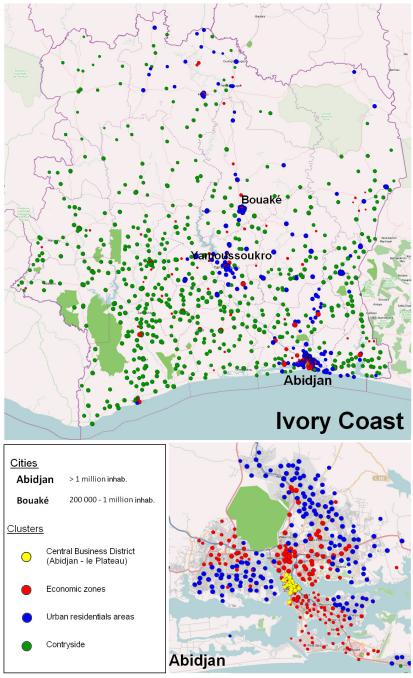


Fig. 6. Clusters on the map of Ivory Coast. Dots are antennas. There is one color per cluster.

Abidjan - Le Plateau (yellow). This cluster covers exactly the Central Business District of Abidjan. In the calendar of Figure 7, we observe an excess of calls from the Monday to the Friday, between 8-9am and 4-5pm. The rest of the time, there is a low lack of traffic emitted from this area. In other words, during the office hours, the phone traffic is higher than expected and lower the rest of the time. This is expected and representative of this type of area: a non-residential business district.

Economic Zones (red). The antennas of this cluster are located either in the commercial areas of the cities or in areas with a strong economic activity, like plantations or mines. In Abidjan, these antennas are located in industrial zones (South and North-West), the shopping districts (North of the business district) and the universities and embassies neighborhood (East). The traffic in these areas is mainly in excess from the Monday to the Saturday between 9 am and

5 pm. The correlation is very strong between the working hours and the calls traffic on these areas.

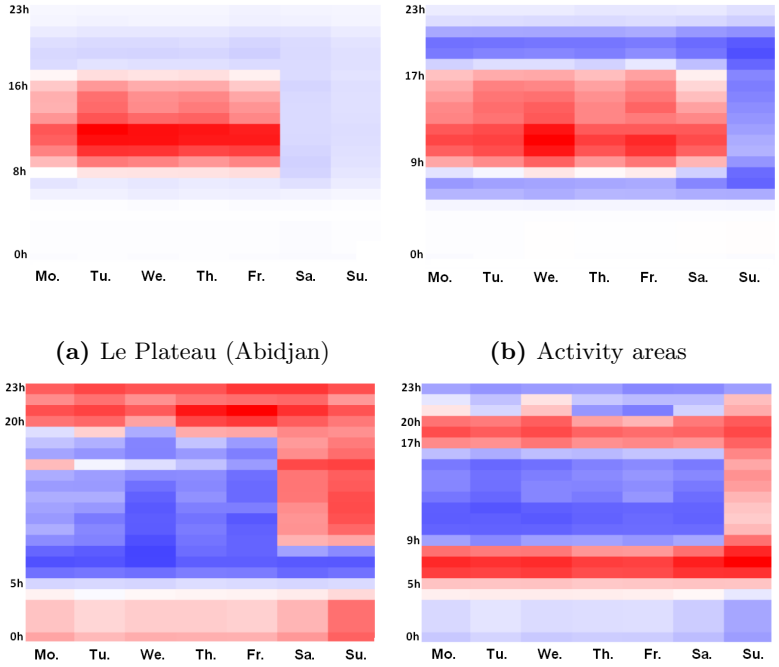
Urban Residential Areas (blue). The antennas belonging to this cluster are mainly located in the cities like Abidjan, Bouaké and Yamoussoukro. If we focus on Abidjan, we observe that the cluster covers the residential neighborhood located in the West and in the North-East of the city. At a finer level of partition of the antennas, this cluster would be split according to the socioeconomic class of the neighborhood: the upper class neighborhood in the East of the city is separated from the lower class neighborhoods, located in the North and the West. The calendar shows lacks of calls during the office hours and excesses the weekend, the night and the early morning during the week. This is correlated with the presence of people in residential areas. Note that the excesses of calls start around 8 pm, while it stops around 5 pm in the Central Business district or in economic areas. This time lag is due to the cheaper price of calls after 8 pm.

The Countryside (green). The antennas of this cluster are spread over the country, except in Abidjan and other cities in general. The calendar for this cluster is quite similar to the one of the urban residential areas, except that the excess periods are limited to the early evening and the whole Sunday.

5.4 User Mobility Analysis w.r.t. Week Day and Hour

Among the 50000 anonymized users, we focus on mobile users characterized by a frequent use of a large set of distinct antennas: after filtering, 6894 users are under study. For these 4-d data (user, antenna, week day and hour), KHC operates a tetra-clustering: as a result, users with the same mobility profile are grouped together, i.e., users who have connected to similar groups of antennas, on similar days of the weeks at similar time periods.
At the finest grain, we obtain 237 clusters of users, 218 clusters of antennas and three time segments ,while week days remain as singletons. Again, the granularity prevent us from an easy interpretation, and we simplify the model. We keep 50% of informativity, that enables a reduction of the numbers of clusters of users and antennas to 40, and the numbers of groups of week days and hour segments to two. The week is divided in two parts: the working days and the weekend. For the hour dimension, the split occurs around 6 pm. The intervals are 0 am - 6 pm and 6 pm - 12 am. Note that the bound at midnight is artificial, because the day start as this time. The cut at 6 pm is the last in the hierarchy of the time segmentation. Then it would have been more relevant to consider a day from 6 pm to 6 pm the next day. Nevertheless, it is easier to have an interpretations on a "usual" time period between 0 am and 12 pm. Therefore we keep the following segmentation: 0 am - 6 pm, 6 pm - 12 pm.

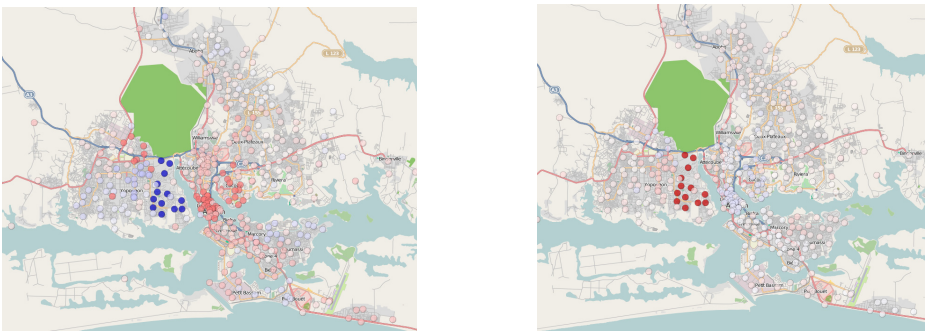
To illustrate the characterization of users' behaviors in terms of mobility provided by the grid, we focus on a group of users. The maps of Figure 8 shows the excesses and lacks of traffic in Abidjan during the week, for both periods of the day and for the selected group of users. The colors correspond to the mutual information $MI(X_1^M; X_2^M \times X_3^M \times X_4^M)$ where X_1^M is the partition of antennas;



(a) Le Plateau (Abidjan) **(b)** Activity areas

(c) Urban residential areas **(d)** Countryside

Fig. 7. Calendars of excesses (red) and lacks of calls emitted from each of the four clusters of antennas, in function of the weekday and the daytime.



(a) Working days before 6 pm **(b)** Working days after 6 pm

Fig. 8. For a group of user, excesses and lacks of uses of antennas according to the day of the week and the time of the day. Focus on Abidjan.

X_2^M, the partition of the weekdays; X_3^M the discretization of the daytime; and X_4^M, the selected partition of the users.

The selected group of users mainly connects to the antennas located in the East of Abidjan after 6 pm during the working days, while they rarely connect to the same antennas before 6 pm the same days. Then, it can be assumed that the selected cluster of users is composed people living in the same area. This hypothesis is reinforced by the socioeconomic nature of this part of Abidjan: it is a residential area. The contributions to mutual information of the other clusters of antennas are smaller. Three areas experience excesses of traffic before 6 pm and lacks after 6 pm. They correspond to the business district (Le Plateau), the embassies and universities neighborhood and the industrial zone located in the West of the city. The common feature of all these areas is their economic activity during the day. To sum up, we can assume that the users of the selected cluster are similar in that they live in the same area and work during the week in three localized area of Abidjan. Similar observations stand for several other clusters of users – thus we are able to summarize users' mobility behavior.

6 Conclusion

Motivated by two key points of economic development strategy of a telco in emerging countries, we have instantiated a generic methodology for exploratory analysis of CDRs data. Our method is based on a joint distribution estimation technique providing the user analyst with a summary of the data in a parameter-free way. We have also suggested several tools for exploring and exploiting the summary at various granularities and highlighting its relevant components. We have demonstrated the applicability of the method on graph data, temporal sequence data as well as user mobility data stemming from country-scale CDRs data. The results of the exploratory analysis are currently considered as valuable additional input to improve network planning strategy and pricing strategy.

References

1. Becker, R.A., Cáceres, R., Hanson, K., Isaacman, S., Loh, J.M., Martonosi, M., Rowland, J., Urbanek, S., Varshavsky, A., Volinsky, C.: Human mobility characterization from cellular network data. Commun. ACM **56**(1), 74–82 (2013)
2. Berlingerio, M., Calabrese, F., Di Lorenzo, G., Nair, R., Pinelli, F., Sbodio, M.L.: AllAboard: a system for exploring urban mobility and optimizing public transport using cellphone data. In: Blockeel, H., Kersting, K., Nijssen, S., Železný, F. (eds.) ECML PKDD 2013, Part III. LNCS, vol. 8190, pp. 663–666. Springer, Heidelberg (2013)
3. Blondel, V., Krings, G., Thomas, I.: Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone. Brussels Studies **42** (2010)
4. Blondel, V., de Cordes, N., Decuyper, A., Deville, P., Raguenez, J., Smoreda, Z.: Mobile phone data for development - analysis of mobile phone datasets for the development of ivory coast (2013). http://perso.uclouvain.be/vincent.blondel/netmob/2013/D4D-book.pdf

5. Blondel, V.D., Esch, M., Chan, C., Clérot, F., Deville, P., Huens, E., Morlot, F., Smoreda, Z., Ziemlicki, C.: Data for development: the D4D challenge on mobile phone data. CoRR abs/1210.0137 (2012)
6. Boullé, M.: Data grid models for preparation and modeling in supervised learning. In: Guyon, I., Cawley, G., Dror, G., Saffari, A. (eds.) Hands-On Pattern Recognition: Challenges in Machine Learning, vol. 1, pp. 99–130. Microtome (2011)
7. Dhillon, I.S., Mallela, S., Modha, D.S.: Information-theoretic co-clustering. In: KDD, pp. 89–98 (2003)
8. Frías-Martínez, E., Williamson, G., Frías-Martínez, V.: An agent-based model of epidemic spread using human mobility and social network information. In: SocialCom/PASSAT, pp. 57–64 (2011)
9. Gnabéli, R.: La production d'une identité autochtone en Côte d'Ivoire. Journal des anthropologues. Association française des anthropologues 114–115, 247–275 (2008)
10. Guigourès, R., Boullé, M.: Segmentation of towns using call detail records. In: NetMob Workshop at IEEE SocialCom (2011)
11. Guigourès, R., Gay, D., Boullé, M., Clérot, F., Rossi, F.: Country-scale exploratory analysis of call detail records through the lens of data grid models (2015). http://arxiv.org/abs/1503.06060
12. Hartigan, J.A.: Direct clustering of a data matrix. Journal of the American Statistical Association **67**, 123–129 (1972)
13. Jiang, S., Fiore, G.A., Yang, Y., Ferreira Jr., J., Frazzoli, E., Gonzàlez, M.C.: A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In: UrbComp@KDD (2013)
14. Kolda, T.G., Sun, J.: Scalable tensor decompositions for multi-aspect data mining. In: ICDM, pp. 363–372 (2008)
15. Laurila, J.K., Gatica-Perez, D., Aad, I., Blom, J., Bornet, O., Do, T.M.T., Dousse, O., Eberle, J., Miettinen, M.: From big smartphone data to worldwide research: The mobile data challenge. Pervasive and Mobile Computing **9**(6), 752–771 (2013)
16. Lockhart, J.W., Weiss, G.M.: The benefits of personalized smartphone-based activity recognition models. In: SDM, pp. 614–622 (2014)
17. Peng, W., Li, T.: Temporal relation co-clustering on directional social network and author-topic evolution. Knowledge and Information Systems **26**(3), 467–486 (2011)
18. Slonim, N., Friedman, N., Tishby, N.: Agglomerative multivariate information bottleneck. In: NIPS, pp. 929–936 (2001)
19. Sun, J., Tao, D., Faloutsos, C.: Beyond streams and graphs: dynamic tensor analysis. In: KDD 2006, pp. 374–383 (2006)
20. Tishby, N., Pereira, O.C., Bialek, W.: The information bottleneck method. In: Allerton Conference on Communication, Control and Computing (1999)
21. United Nations Global Pulse: Mobile phone network data for development (2013). www.unglobalpulse.org/Mobile_Phone_Network_Dat_for_Dev
22. Vieira, M.R., Frías-Martínez, V., Oliver, N., Frías-Martínez, E.: Characterizing dense urban areas from mobile phone-call data: discovery and social dynamics. In: SocialCom/PASSAT, pp. 241–248 (2010)
23. Wang, D., Pedreschi, D., Song, C., Giannotti, F., Barabási, A.L.: Human mobility, social ties, and link prediction. In: KDD, pp. 1100–1108 (2011)
24. Wang, P., Domeniconi, C., Laskey, K.B.: Information bottleneck co-clustering. In: Workshop TextMining@SIAM DM 2010 (2010)