Effective Optimization of Billboard Ads Based on CDR Data Leverage

Imed Eddine Semassel
Department of Computer Science, Faculty of Sciences of Tunis, El Manar University, Tunis, Tunisia

Sadok Ben Yahia
Department of Software Science, University of Technology, Tallinn, Estonia

Abstract: Call Detail Records (CDRs) provide metadata about phone calls and text message usage. Many studies have shown these CDR data to provide gainful information on people's mobility patterns and relationships with fine-grained aspects, both temporal and spatial elements. This information allows tracking population levels in each country region, individual movements, seasonal locations, population changes, and migration. This paper introduces a method for analyzing and exploiting CDR data to recommend billboard ads. We usher by clustering the locations based on the recorded activities' pattern regarding users' mobility. The key idea is to rate sites by performing a thorough cluster analysis over the achieved data, having no prior ground-truth information, to assess and optimize the ads' placements and timing for more efficiency at the billboards.

Keywords: Call Detail Records, Rating scores, Outdoor advertising, Billboards

Introduction

In daily lives and around the world, mobile usage is exploding, and telecommunications services account for over 5 billion unique subscribers with access to mobile devices (Letouzé & Vinck, 2015). Indeed, almost ninety percent of people have mobile phones. These latter are moving, and arguably they are leaving tracks of their movements, which can generate extensive data and information. The data of telephone calls or other communication like Short Message Services (SMS) that passes through such devices are recorded by a Telecom Service Provider and referred to in the remainder of this paper as Call Data Record (CDR). The latter is a data structure storing information about a given telephonic activity involving a user of a telephonic network. A CDR usually contains spatial and temporal data.

It is a real challenge to obtain valuable data to elaborate accurate statistics. In this context, the emergence of new types of data, or metadata, collected passively from the population and

http://doi.org/10.18080/jtde.v10n2.527

anonymized to make them compliant for secondary use and the strict requirements of the General Data Protection Regulation (GDPR) of the European Union presents a genuine opportunity. Furthermore, they can complement self-reported data from interviews and questionnaires, which are time-consuming, laborious, and tricky to predict the dynamic changes. CDRs are among these passively collected data and are steadily being used in research and other big data forms (Cuzzocrea, Ferri & Grifoni, 2018).

On the socio-economic side, mobile phones are considered the preferred and most accessible means of communication (CTA, 2017). The lack of reliable landline infrastructure explains this preference. Besides, they play a pivotal role in socio-economic evolution (Arie, 2015). By deploying mathematical modelling techniques, researchers can explore issues that heavily rely on CDR data to gauge insights into the location of different populations and their evolution over time (Gore *et al.*, 2019; GSMA, 2017).

By their nature, CDRs are generated in large volumes. They can be seen as a wide-area sensor network as long as they offer a statistically accurate representation of the distribution of people in an area. They can be integrated with other sources to track large and heterogeneous groups of people (Bianchi *et al.*, 2015). Telecom operators continuously gather an immense quantity of CDRs, from which it is possible to extract additional information with low additional cost and generate valuable datasets. We can derive gainful knowledge from analyzing these data, which may help in city planning (Steenbruggen, Tranos & Nijkamp, 2015; Louail *et al.*, 2014), such as marketing, user profiling, disease spreading patterns, natural disasters, social event occurrence, and their impacts.

Mining mobile phone data offers mobile telecommunication operators the opportunity to have a two-sided business model. Firstly, they would generate revenues from their mobile phone users and, secondly, from upstream customers, such as advertising firms (Quercia *et al.*, 2011).

Outdoor advertising has a 500 billion USD global market; its revenue has grown by over 23% in the past decade to over 6.4 billion USD in the US alone (Zhang *et al.*, 2018). Billboards are the most used medium for outdoor advertising (about 65%), and 80% of people notice them when driving (Zhang *et al.*, 2018). A credible way to assess the performance of billboard ads should include both the number of people in front of the billboard and the likelihood that those people might like a specific ad shown on the billboard(Quercia *et al.*, 2011). Billboards have remained a crucial tool for spreading information to a target market. However, due to technological advances and consumer preferences, effective placement and innovative advertising methods are required to capture target customers' attention. Our proposition rates the zones based on the pattern of activities extracted from CDR data. We clustered these latter into 3 clusters, and we used results and statistical measurements to provide recommendations about the billboards and their ad content.

We structure the remainder of the paper as follows. First, in the section Related Work, we present scrutiny of the related works. Then, Section Methodology describes our contribution, the dataset we have used, and the results obtained. Finally, we present some concluding remarks and issues for the future work in the Conclusion.

Related Work

Sultan *et al.* (2019) introduced a clustering-based artificial neural network model (C-ANN) to study the operational efficiency of cellular networks and user action patterns using the CDR data, publicly available from Telecom Italia. The proposed solution focuses on the spatio-temporal analysis of CDRs and can model and classify the network traffic patterns. First, they convert the CDR's activity classification from an unsupervised to a supervised multi-class classification problem. Then, CDR activities are classified based on activity levels and spatio-temporal characteristics. The extracted insights shed light on a solid spatio-temporal relation with the actual network traffic patterns and would be gainful in monitoring and optimizing network traffic.

CDR data was also used to estimate the crowd on a free-to-view special event day in the study carried out by Sumathi *et al.* (2018) at the Indian Institute of Science campus. They applied exploratory data analysis and statistical techniques over the records to estimate the participant count. The CDRs related to people who have their home or work in that event location are removed. The obtained results underscored a positive correlation with participant count.

Starting from the CDR dataset of the D4D challenge, Bianchi *et al.* (2016) proposed a framework for identifying patterns and regularities and deriving meaningful information. The latter was of help in understanding users' habits and extracting their characterizing profiles, using two implementations of an unsupervised data mining procedure:

- LD-ABCD: an agent-based algorithm that extracts separated clusters in the data;
- PROCLUS: a subspace clustering algorithm that identifies clusters in a subset of the original feature space.

The method is greedy in computational space and time resources.

Mamei, Colonna & Galassi (2016) presented a method to identify a home, work, and other frequented places of users using CDRs. First, the method spatially clusters the collected CDR events of each user. Then, they weigh the clusters based on the user visiting pattern. Finally, based on a dynamic threshold, the authors select those weighted clusters associated with the relevant places. Thus, ground truth information from a fraction of users whose appropriate places were known beforehand was used.

To understand Estonian group mobility, Hiir *et al.* (2019) analyzed the CDR data to identify the social network patterns by modelling the call activity with a graph. The latter's nodes represent the set of call IDs in the dataset, and the edges represent the call activity between them. Furthermore, the authors studied the impact of natural and social events on the calling activity patterns. Thus, a descriptive analysis was carried out in different periods and between counties, followed by analyzing the effects of natural and non-natural events on call activity. They exploit natural events to inspect the impact of weather and the full moon. Non-natural events are used to inspect how a significant football match affected people's activity. Results showed that the calling activity depends on the calling time period, and that these events do impact it.

Another study by Scharff *et al.* (2015) analyzed CDRs to understand mobility patterns before, during, and after a large religious festival in Senegal called the Magal of Touba. The study investigated mobility and transport patterns by analyzing the changes in communication volume between the city of the festival and other locations in Senegal. It showed the primary routes used by the pilgrims and their respective travel times, with insights into the areas most pilgrims were coming from. Health officials and other stakeholders can underpin these findings to control and stop the spread of infectious diseases by providing information about health structures available during the festivals.

Nair, Elayidom & Gopalan (2020) proposed a method to estimate the traffic density using the global K-means clustering algorithm and K-nearest neighbour classification algorithm. The technique used the CDR data to distinguish traffic density in five different locations on weekdays and weekends. The method clusters the CDR data into a group of data that goes through a classification process to classify and sort the density of the traffic, whether empty, low, high, or complete.

The proposed approach of Leng, Zhao & Koutsopoulos (2021) infers an individual's home and workplace based on the behavioural patterns at various places. The method takes advantage of the mobile phone dataset's behavioural data. Unsupervised machine learning algorithms were utilized to uncover similarities across locations for multiple groups of individuals to infer home and workplace. The approach was tested on a real-world dataset in one of China's most populous cities.

The proposed unsupervised statistical method in DeAlmeida *et al.* (2021), Energy-based Flow Classifier (EFC), detects anomalies in CDR data using a classifier based on the inverse statistics of the Potts model. Abnormalities are not detected using the clustering algorithm.

However, K-means clustering is used to characterize and analyze the traffic patterns of each region. Telecom Italia provided actual traffic data from Milan, which was used to test the approach. The findings showed that taking into account the varying traffic patterns of different geographic locations is critical in determining the accuracy of anomaly detection in mobile networks.

All these works aim to gauge the crowds of places, either by observing the changes in the CDR activity or classifying the sites based on the users' behaviour and call activity. The results provide worthy insights that can improve many fields, including the effectiveness of outdoor advertising. Table 1 summarizes these works and application domains.

Table 1. Related work and applications

Reference	Year of publication	Dataset country	Dataset year	Method	Application
(Bianchi *et al.*, 2016)	2016	Ivory Coast	2012	Clustering algorithms	profiling users, marketing strategies
(Sultan *et al.*, 2019)	2019	Italy	2013	Clustering-based artificial neural network model	profiling users to monitor the mobile traffic
(Sumathi *et al.*, 2018)	2018	Bengaluru city	2013	Statistical techniques (linear regression)	measuring event success and outdoor advertising.
(Mamei, Colonna & Galassi, 2016)	2016	Italy	2012	Clustering algorithm (K-means).	identify users' relevant places, understand mobility patterns, manage mobility
(Hiir *et al.*, 2019)	2019	Estonia	2015	Social network analysis	identifying human behaviour, understanding the social fabric of events
(Scharff *et al.*, 2015)	2015	Senegal	2013	statistical techniques (frequencies with geographic plots)	Health, Transport, resource management
(Nair, Elayidom & Gopalan, 2020)	2020	India		Global K-means and K nearest neighbour	Density traffic
Leng, Zhao & Koutsopoulos, 2021	2021	China	2004	K-means, Fuzzy C-Means clustering	User Location Segmentation

Reference	Year of publication	Dataset country	Dataset year	Method	Application
DeAlmeida *et al.*, 2021	2021	Italy	2013	Unsupervised statistical method	Anomaly detection in mobile networks

In our work, we pay heed to two complementary aspects. We cluster the places based on the recorded activity patterns on the antenna level, which is quite similar to the clustering done by Sultan *et al.* (2019). The difference is that we rely on a mix of the two aspects, spatial and temporal, at the same time. We clustered the areas not only based on their instant activity levels as done in Sultan *et al.* (2019), but also on the patterns of these activities over time, allowing us to extract context information by categorizing them. Furthermore, we compute an hourly rating (RP) point for each place from the activity levels to reflect the spot importance and the effective content type on the ads. Finally, we use these latter data as input for a multi-criteria decision-making model, called TOPSIS, to sort and recommend the billboard places.

CDR Datasets

Mobile phones are ubiquitous in our daily lives as they convey and generate data of actual usage. Therefore, anonymous CDR data can provide dense and rich metadata. That data encapsulates the time and place where any person sends or receives a call and/or a text message. The Sonatel-Orange Telecom company has made CDR data available to the research community as part of the D4D challenge in 2015. The goal is to harness the potential of data from mobile calls to propel socio-economic development. Thus, researchers can investigate several points directly impacting development factors through anonymized CDR datasets.

The considered data are metadata covering the time and place where a user can make a telephone action (call or text message). For example, the Sonatel-Orange mobile operator in Senegal disclosed three triple-anonymized datasets from its mobile users between January 1, 2013, and December 31, 2013, at the occasion of the Data for Development (D4D) in Big Data 2014's challenge (de Montjoye *et al.*, 2014). Below, we briefly describe these datasets.

Dataset 1 represents the traffic per pair of antennas for the 1,666 antennas (sites) hourly. The dataset contains 24 files, of which 12 files contain the monthly voice traffic between the sites, structured as follows~:

- Timestamp: the day and time in YYYYY-MM-DD HH format (24-hour format).
- outgoing_site_id: id of the site from which the call originated.
- incoming_site_id: id of the site receiving the call.
- number_of_calls: the number of calls between both sites during the hour.

- total_call_duration: the total duration (in seconds) of all calls between the two sites during the hour.

A sample of this dataset is shown in Table 2.

Table 2. Excerpt of data (aggregated phone calls' metadata) from Dataset 1

Timestamp	outgoing site id	incoming site id	Number of calls	Total call duration
2013-01-01 00	1	1	1	54
2013-01-01 00	1	2	1	39
2013-01-01 00	1	24	1	2,957
2013-01-01 00	1	186	1	56
2013-01-01 00	2	2	22	418

Data about monthly traffic text messages between the antennas (sites) are structured as follows:

- Timestamp: the day and time in YYYYY-MM-DD HH format (24-hour format).
- outgoing_site_id: id of the site from which the text originated (SMS).
- incoming_site_id: id of the site receiving the text (SMS).
- number_of_calls: the number of texts (SMS) between the two sites during the hour.

We give an excerpt of this dataset in Table 3.

Table 3. Excerpt of data (aggregated text messages metadata) from Dataset 1

Timestamp	outgoing site id	incoming site id	Number of SMSs
2013-01-01 00	1	61	1
2013-01-01 00	1	340	1
2013-01-01 00	1	419	1
2013-01-01 00	1	420	1
2013-01-01 00	2	447	2

Dataset 2 represents fine mobility data spread per user over an interval of 2 weeks for an entire year. These data are unique in that we pair them with behavioural indicators computed by bandicoot python toolbox at the individual level for about 300,000 users randomly sampled. We give an excerpt of the second dataset in Table 4.

Table 4. Excerpt of fine mobility data (Dataset 2)

user-id	Timestamp	site id
1	2013-01-07 13:10:00	461
1	2013-01-07 17:20:00	454
1	2013-01-07 17:30:00	454
1	2013-01-07 18:40:00	327
1	2013-01-07 20:30:00	323

Dataset 3 summarizes the rounded one-year mobility data volume (country district level) with behavioural indicators at the individual level for approximately 150,000 randomly sampled users. Table 5 shows an excerpt of the third dataset (district-level mobility data).

Table 5. Excerpt of data from Dataset 3

user-id	Timestamp	district id
37509	2013-01-29 15:00:00	3
84009	2013-01-14 07:00:00	3
84009	2013-01-14 07:00:00	3
84009	2013-01-14 07:00:00	3
80150	2013-01-27 16:50:00	3

Methodology

This section describes in detail our new approach that aims to cluster places and then label them based on CDR data made available by the Sonatel-Orange Telecom company as part of the D4D challenge in 2015, and rate these places with scores that reflect their importance in an advertisement context. We rely on antenna positions and the fine mobility data spread per user over two weeks to generate activity patterns.

The overall system model presented in our contribution comprises three stages, and its workflow is flagged in Figure 1:

- Data pre-processing and analysis step,
- Clustering step,
- Results leveraging step.

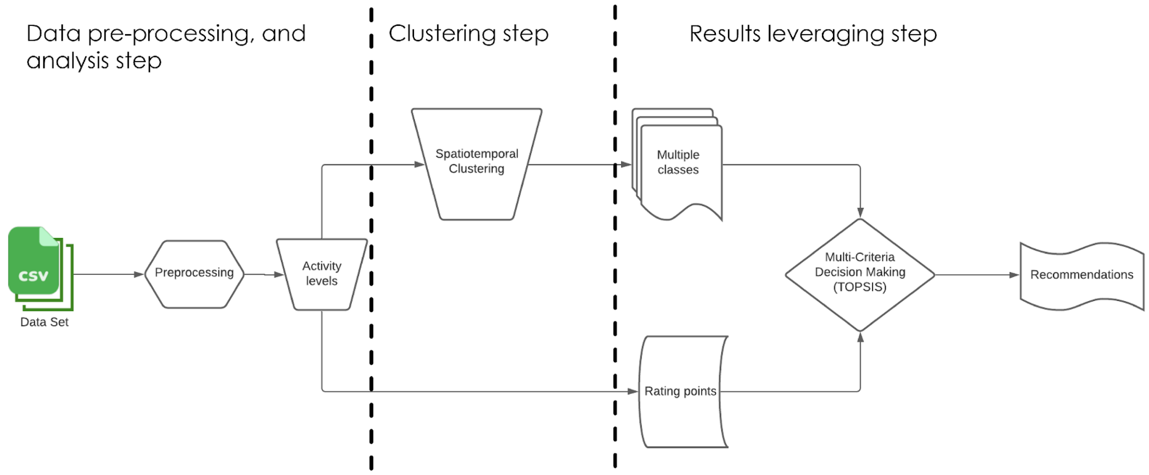

Figure 1. The workflow of our contribution

First, we pre-process and analyze the Dakar Region districts we focused on to get a global overview. We split the Dakar region into ten districts: each district showed a distinct pattern during the two-week period under analysis. According to the CDR activity level recorded for each day, the districts can be labelled by three different labels: high, medium, and low activity areas. Our fine-grained work applies to one district with a high activity level (e.g., the DAKAR PLATEAU, id=4 in Figure 2). We notice a sharp decrease in weekends compared to the rest of the week in the latter.

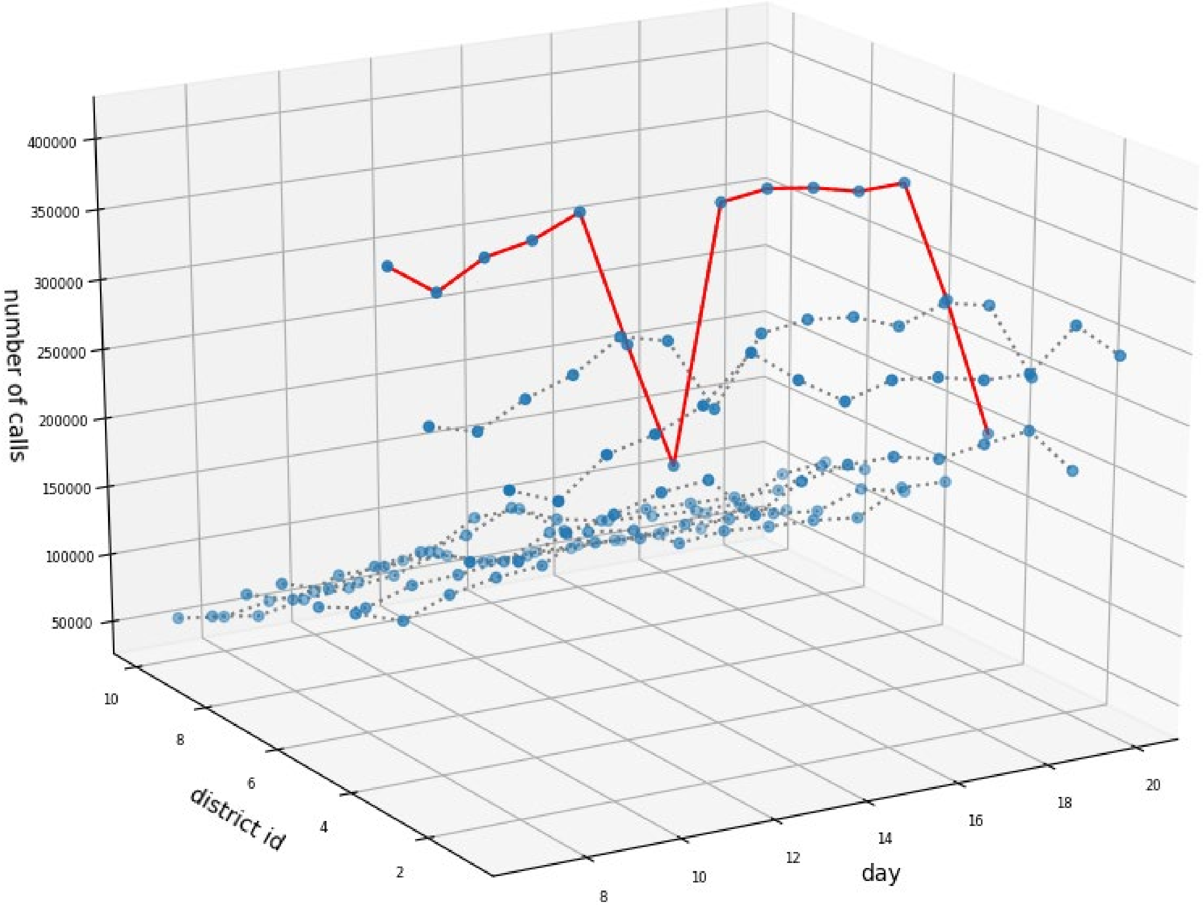

Figure 2. 3D plot of the relationship between the days and the activities at the district level

Then, as an outcome of the Clustering step, the antenna spots are clustered according to the CDR activities' patterns at the hours level into multiple clusters. Finally, the clustering results and the rating points computed from the activity counts are used as input for the TOPSIS model to sort and provide recommendations about the placement of ads and their content types. We perform the clustering step to categorize the antenna spots and transform the rating points from global to targeted rating scores.

Data pre-processing and analysis step

This first step aims to analyze and clean the data. We have analyzed the fine mobility data of the first two weeks recorded in Dataset 2 of the CDR data. We have paid heed to the Dakar Region; we made a 3D plot to underscore the relationship between days and activities at the district level (Figure 2). The continuous red line is the activity of the district DAKAR PLATEAU (id=4); the other dotted lines are the different districts of DAKAR; the points are the number of the records for each district per day. Observing the latter shows that the activity decreases during weekends compared to the other days of the week for the high activity districts, especially for the district with id=4 (DAKAR PLATEAU), where it shows five high counts that refer to the weekdays; and two low counts referring the weekends, which is similar to the work pattern.

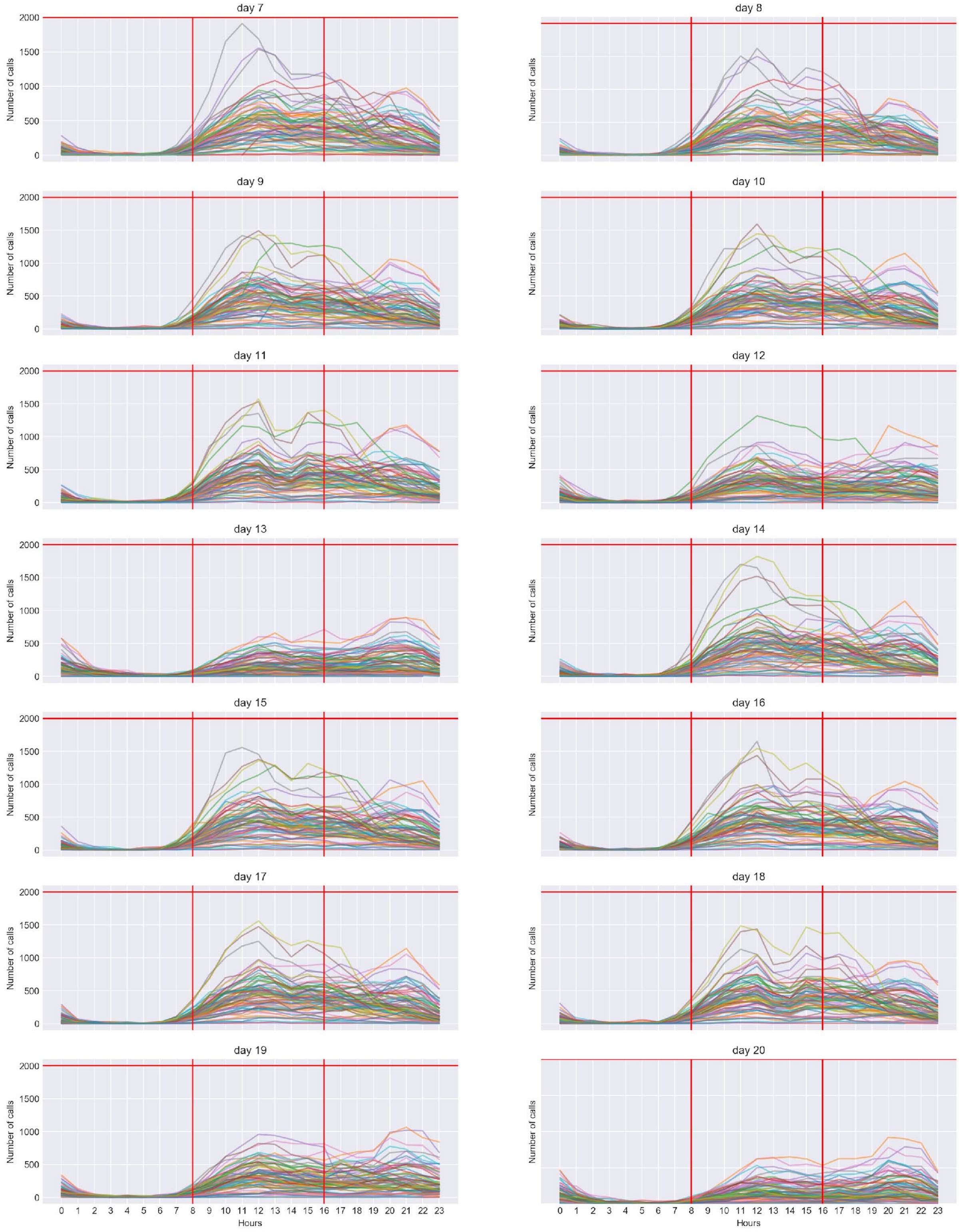

Figure 3. The activity of the district DAKAR PLATEAU of each antenna per hour. Each coloured line represents a different antenna plot.

We increased the granularity level by analyzing the activities per hour to have more fine-grained results. We have plotted for each site the number of activities/hours daily for both weeks (Figure 3). The vertical red line refers to the 8 am – 4 pm period for each day. (This interval is chosen to flag the differences in activities before, during, and after the working time.) The following facts are worthy of mention:

- shallow activity between 3 am and 6 am for all the antennas, where most people are asleep.
- Low activity at the weekend (day 13 and day 20) compared to other days.
- High activity between the vertical red line in Figure 3 represents 8 am – 4 pm, and several antennas showed their peak on all the days after 4 pm.

These observations assumed that the areas where the antennas belong could be clustered based on the activity pattern. Then, these latter can be labelled by type according to the average patterns of each cluster. For example, areas with top activity during the day and low at night on workdays can refer to a workplace, while areas with high activity at night can refer to urban places.

Clustering step

We cluster the sites according to the CDR activity pattern recorded in each antenna to group sites with similar patterns, using no ground-truth information about the location type of the antennas. We have used unsupervised machine learning algorithms, such as K-means, Gaussian mixture and Agglomerative clustering (Ficek *et al.*, 2012; Jin *et al.*, 2010; Murphy, 2013). These latter are the most popular clustering algorithms that can analyze and group similar unlabelled data into clusters. However, these algorithms require the number of clusters k to be generated; to determine the optimal number of clusters, we have used the silhouette score method (KMeans, 2020). The latter is used to measure the quality of clusters created using clustering algorithms in terms of how well samples are clustered with other samples that are like each other. We computed the Silhouette score for each sample of different clusters and chose its global maximum as the optimal k. For example, the Silhouette score-versus-k plot in Figure 4 shows a clear peak at k = 3. Hence, the data can be optimally clustered into 3 clusters.

In addition to the three algorithms that have shown similar results, we tried to use DBSCAN, which stands for Density-Based Spatial Clustering of Application with Noise. However, the latter did not perform the partition-based clustering and grouped all the antennas in only one cluster. Therefore, we tried an optimized version of DBSCAN named DBSCAN-GM. This combines Gaussian-Means and DBSCAN algorithms to cover the limitations of DBSCAN by exploring the benefits of Gaussian-Means. However, it provided the opposite of the DBSCAN results and clustered each antenna in a different cluster.

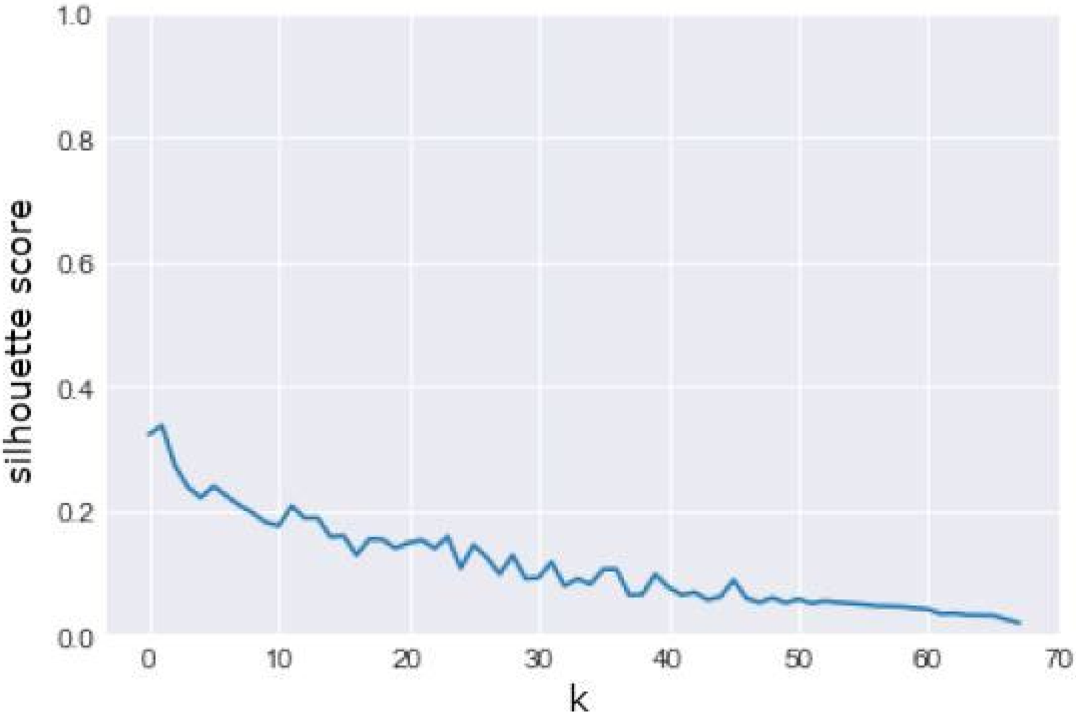

Figure 4. The Silhouette score-versus-k plot

To validate the clustering results, the antennas' activity is plotted as points in a 2D plane using a popular dimensionality reduction algorithm called t-SNE. We have selected this tool to validate our clustering, as it can visualize the antennas' activities, which are high dimensional data, in a two-dimensional map. Then, the antennas' activities are coloured according to the cluster they belong to (we show the results found by K-means).

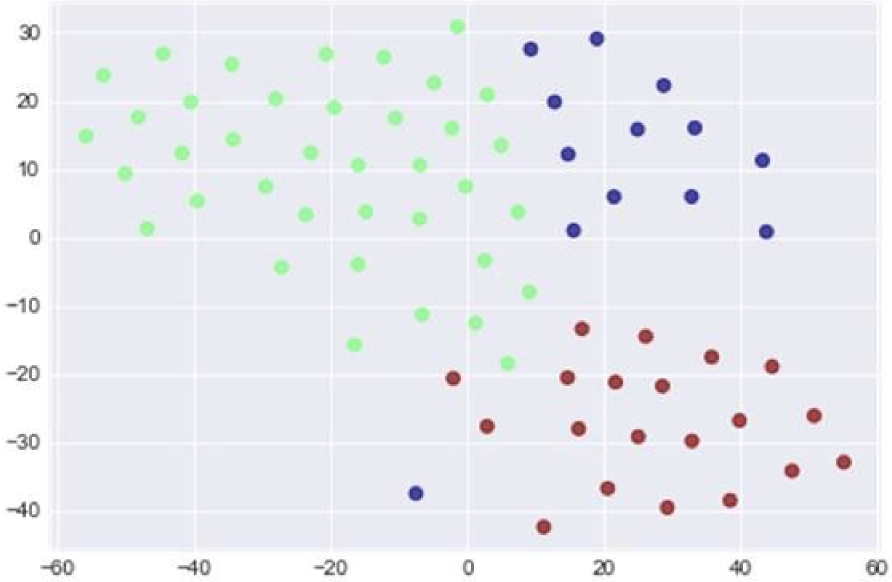

Figure 5. T-SNE Visualisation Validation of K-means results

In the plot (Figure 5), each point represents the activity of the antennas during the two weeks (336 hours). They were reduced from 336 to 2 dimensions. Theoretically, the distance between points in the higher dimensional space was preserved, so close points refer to similar antenna activity. The fact that most blue, red, and green points are close together indicates that the clustering worked well.

The clustering of the activity of the antennas during the two weeks into 3 clusters is shown in Figure 6 (we show the results found by K-means). Each curve in the plots represents the activity of an antenna during both weeks. The dashed curves represent the mean activity of each cluster, the vertical red line refers to the 8 am – 4 pm period for each day, and the horizontal red line refers to 1000 calls to flag the differences between plots. Each cluster shows a different global pattern:

- **Purple plot**: all the activities are average, and the spikes are after 16h during both weeks, including weekends.
- **Red plot**: where spikes are higher and within 8 am to 4 pm, and at weekends shows shallow activity compared to weekdays.
- **Green plot**: shows low activity compared to two of the other clusters.

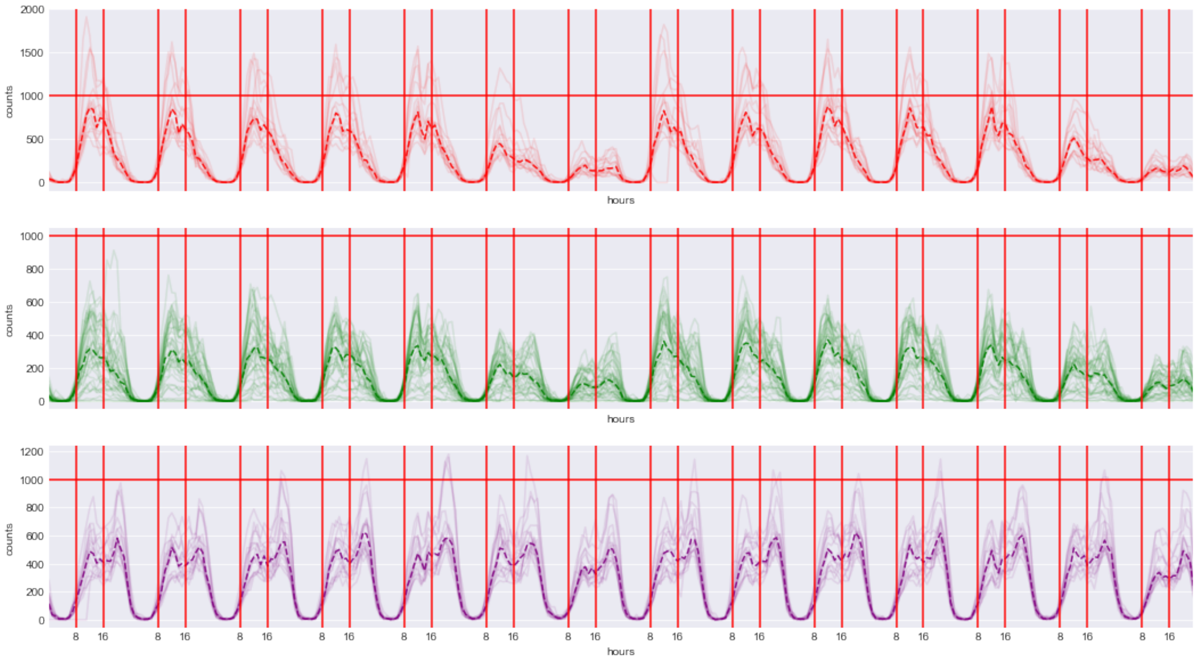

Figure 6. Activity grouped by the identified clusters

According to the patterns shown by the curves in the plots, we can label the clusters as follows:

- **Urban places** for the purple plot,
- **Workplaces** for the red plot,
- **Other places** having infrequent visits for the green plot.

Results leveraging step

Our next step is to leverage these results. First, we created a map that shows the places based on their type (work, urban, other types), as shown in Figure 7. Since we do not have the antenna coverage area, we have chosen to apply the Voronoi diagram to have a cell for each antenna that simulates its coverage area. Its colour represents the cluster where the antenna belongs.

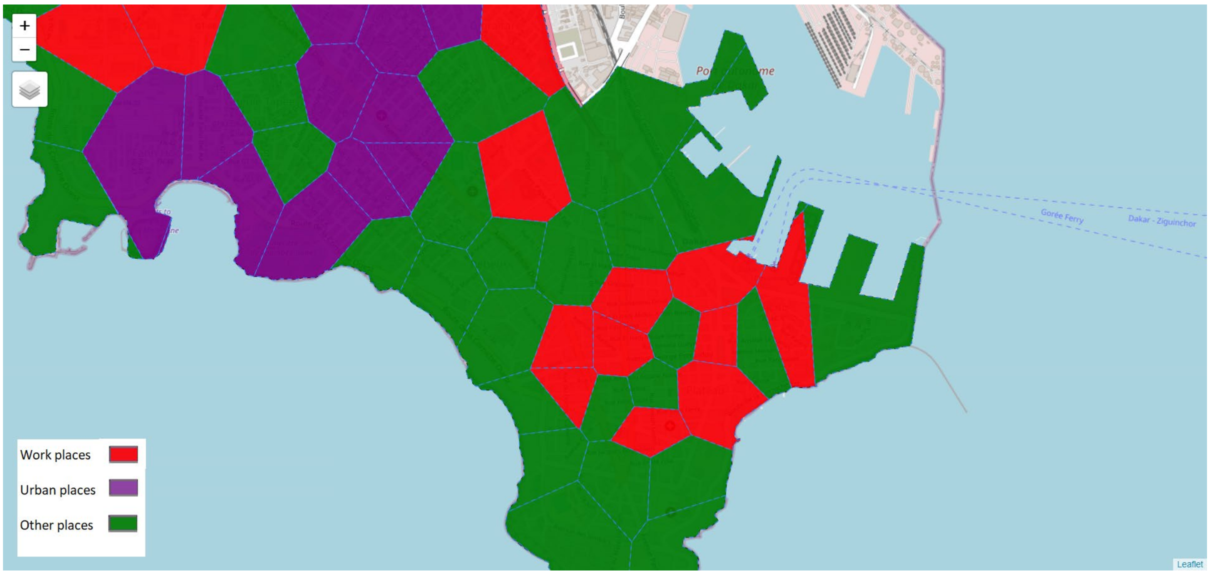

Figure 7. Labelling of the areas according to the identified clusters using the Voronoi diagram to assess the antenna's coverage

Compute rating points

To compute the audience size that reflects the area's importance to the billboard ads, we rely on the Gross Rating Points (GRP) measurement. This latter measures the impression concerning the number of people in the audience for an advertising campaign (Gross Rating Point (GRP), 2020). It is computed by multiplying the percentage of the audience reached by an advertisement with the frequency they see it. Based on this measure and using data extracted from the antenna activity, we compute these scores as follows:

$$RP = Reach\ (\%\ of\ audience\ reached) \times Exposed\ (number\ of\ User\ exposure) \\ \times Frequency\ (number\ of\ ad\ impressions)$$

An RP score for each antenna includes the following parameters:

- **Reach value** stands for the number of users at a given hour, divided by the total number of users that passed by these antennas.
- **The exposed value** is the average number of how many recorded CDRs there are for users at a given hour and antenna. (We added this value to gauge the number of times the user is exposed to the ads.)

- **The Frequency value** is the number of exposures of the ad, which can vary from 1 to the number that the advertiser allocates.

Using technique for order preference by similarity to ideal solution (TOPSIS)

After computing rating points for each antenna per hour using the formula presented in the previous section, the rating points are used as input to a Multi-Attribute Decision Making (MADM) method named TOPSIS. The TOPSIS model is a MADM method widely used for ranking alternative candidates based on multiple attributes. In our work, we set the hours as attributes and the antennas, which represent areas as alternatives, for sorting and selecting the best placements for the billboards for each cluster based on the scores computed.

Table 6 shows the matrix with the computed scores (rating points) for each antenna per hour, which is the input for the TOPSIS model. For each antenna, we calculate 24 values using the formula. Each value reflects the rating point of an antenna for one hour based on its recorded activities. Table 7 shows the ranked antennas with the scores computed with the TOPSIS model.

Table 6. The input matrix of urban places for the TOPSIS model
Part 1

Attributes (time)	**Alternatives (Sites)**							
	82	**104**	**106**	**109**	**128**	**130**	**132**	**140**
0h	62.18	56.42	31.97	35.66	18.98	34.01	36.44	15.39
1h	24.12	21.67	13.95	15.14	8.88	12.25	14.99	6.40
2h	12.44	9.60	5.93	5.52	4.77	3.98	5.09	3.31
3h	5.86	3.66	3.05	3.80	3.13	2.32	1.95	1.47
4h	2.95	2.73	1.95	3.02	2.05	1.41	1.59	1.20
5h	2.65	1.78	1.31	2.59	1.91	1.57	1.23	0.91
6h	5.18	3.48	3.08	3.26	2.14	2.55	4.50	1.43
7h	17.76	13.43	9.34	9.70	8.65	8.21	11.57	5.22
8h	35.39	26.69	21.22	17.15	23.79	15.73	20.22	11.28
9h	57.28	44.19	37.42	29.45	45.98	31.89	35.12	22.19
10h	77.52	71.40	54.29	42.23	68.35	48.54	53.94	34.19
11h	96.13	88.72	68.15	53.05	78.90	57.51	67.64	43.70
12h	123.12	111.15	80.94	64.60	91.67	72.14	79.97	51.81
13h	112.75	110.82	76.58	64.14	79.98	73.79	81.11	49.07
14h	97.22	100.79	71.36	65.23	69.21	73.00	81.82	44.28
15h	93.87	99.99	76.34	67.98	73.11	78.64	81.47	45.16
16h	90.53	99.07	75.74	64.97	62.87	74.21	74.63	44.04
17h	90.53	100.64	76.89	74.38	60.60	74.98	77.34	49.73
18h	104.42	111.91	80.93	80.07	58.45	78.71	86.71	49.92
19h	130.14	130.99	89.68	87.87	64.00	96.32	111.32	56.09
20h	148.03	165.37	106.49	95.35	73.53	111.72	140.68	59.79
21h	157.33	167.41	107.22	100.08	71.25	111.26	153.02	50.90
22h	148.42	151.96	92.18	87.60	58.76	99.71	130.16	41.65
23h	106.12	110.66	65.28	70.93	39.90	72.01	87.67	29.22

Part 2

Attributes (time)	**Alternatives (Sites)**							
	147	**155**	**161**	**163**	**165**	**170**	**177**	**188**
0h	27.13	29.24	35.06	17.97	18.68	15.60	27.45	14.81
1h	10.36	11.88	12.48	7.07	7.30	5.02	10.54	4.74

Attributes (time)	**Alternatives (Sites)**							
	147	**155**	**161**	**163**	**165**	**170**	**177**	**188**
2h	3.19	3.82	3.69	2.74	3.31	1.82	4.25	1.92
3h	1.72	2.46	2.00	1.32	1.84	1.09	1.40	0.77
4h	1.07	1.17	1.14	0.92	1.25	1.07	0.77	0.64
5h	0.89	1.38	1.39	0.84	1.03	0.72	1.03	1.12
6h	2.45	2.26	2.93	1.57	1.55	1.54	2.33	2.23
7h	7.89	8.20	9.36	5.56	5.83	6.83	8.20	9.81
8h	14.39	17.56	23.34	16.24	12.59	13.39	16.07	20.11
9h	27.79	32.98	36.15	33.45	24.34	23.25	31.13	37.08
10h	39.22	51.73	54.98	50.43	38.55	34.77	42.96	49.58
11h	51.72	63.00	66.81	65.12	48.33	42.60	54.65	59.84
12h	64.45	67.21	76.00	74.15	53.68	51.56	59.96	66.51
13h	62.77	63.66	75.02	72.48	53.23	46.37	61.19	64.17
14h	60.62	60.29	69.34	61.95	49.24	41.90	54.28	55.08
15h	65.10	66.04	75.36	69.39	52.68	41.63	57.30	57.27
16h	60.61	60.84	69.82	65.82	51.99	39.87	55.59	54.32
17h	65.78	68.51	77.93	69.10	56.08	42.70	56.57	55.10
18h	65.29	67.96	78.58	68.65	55.95	42.55	59.54	51.28
19h	71.67	87.32	96.14	69.66	64.68	51.60	72.15	54.74
20h	83.67	111.22	126.07	71.26	65.12	58.09	91.13	61.70
21h	83.14	115.95	132.09	66.65	65.72	54.64	95.91	58.90
22h	76.41	98.06	113.53	58.95	54.40	47.72	79.71	47.40
23h	54.35	63.54	78.82	40.30	37.15	33.70	52.83	31.22

Table 7. Urban places ranked antennas based on TOPSIS scores

Antenna	**82**	**104**	**132**	**109**	**106**	**161**	**130**	**128**
Score	0.95	0.76	0.50	0.48	0.45	0.41	0.38	0.37
Rank	1	2	3	4	5	6	7	8
Antenna	**155**	**177**	**147**	**163**	**188**	**165**	**140**	**170**
Score	0.33	0.24	0.23	0.20	0.17	0.14	0.09	0.06
Rank	9	10	11	12	13	14	15	16

Discussion

The TOPSIS results for each identified cluster can help in setting up a reasonable pricing system, where the price that advertisers pay depends on two aspects (the type of place according to the clustering and their rank according to the size of the audience), which is a more precise valuation. Moreover, our clustering is quite similar to the clustering done by Sultan *et al.* (2019). The difference is that we cluster the places based on the patterns of the recorded activity on the antenna level, while paying heed to the spatial and temporal aspects at the same time. The antennas and their estimated coverage handle the spatial aspect. The temporal aspect is handled by the number of the recorded activities for each hour. Unlike Sultan *et al.* (2019), we rely on combining the two aspects. We clustered the areas considering their instant activity levels and the patterns of these activities during a period of time, allowing us to classify the areas as urban, work, and other places.

Furthermore, the final results can also provide recommendations that can optimize the efficiency and the impact of the ads and help in the control and placing of the billboards through a dashboard. Among these recommendations are:

- As we ranked places according to the times and spots with a high density of people, we can recommend the places for billboards and advertising;
- As we have the type of places (work, home, other spots), we can recommend how and where the advertisers should place their advertising;
- As we know the transition areas and time and places with low density, we can recommend a set of rules on how the ads should be in these places, especially on highways, to avoid accidents;
- As we have the type of places, we can recommend the type (animation, 3D, etc.) of the content of the ads based on the places and the peak time to increase the efficiency;
- Using rating points along with the type of places, we can recommend the subject of the content of the ads based on the time and spots (return from work, going to work, etc.) to recall products to the consumers and increase the efficiency of the ads;
- Recommend planning the time allocated to the ads and how much they cost based on each place's type and rating point.

Following these recommendations helps advertisers target their audience more effectively by designing attractive ads that consider several parameters. For example, an ad related to food and restaurants in a place labelled as a workplace, around midday, would be very efficient, since thoughts in that area are about lunch, which increases the probability of noticing and being attracted by these ads.

Conclusion

Based on all the observed results, it becomes clear that CDR data analysis can help identify and characterize user trends. In this paper, we presented a method for analyzing and exploiting mobile data. The key idea is to cluster and label the locations based on the pattern of the users' activities recorded by the antennas for two weeks. Besides the clustering, we computed rating points from each antenna's registered user activities. This latter reflects the importance of areas. We have chosen the marketing context, one of several contexts where our method can apply. We focused on the Dakar region district antenna places. The clustering produced three clusters. The clustering results have been used along with RP scores to provide a reasonable pricing system and a set of recommendations. The latter is shown to be beneficial in optimizing the efficiency of the outdoor advertisement. It is also worth mentioning that, owing to the contextual information extracted by this method for each location, we might guide the content of the shown ads on the billboards to be catchier and increase their reachability impact.

We can further develop this work to improve clustering accuracy using multiple data sources. For example, we can use the traffic between antennas to find the relation between places and cluster the records according to the user activity level and their visited sites before applying the proposed method.

Moreover, we can add additional attributes to the TOPSIS model to rank places with more context-related conditions. Further, it can be extended by adding event aspects to find their correlation with user mobility patterns. In addition, filtering the users by classifying them according to their movements can improve the results and suggest more recommendations.

Besides improving our method results by modifying inputs, we also use other strategies such as adding Fuzzy Clustering and additional steps in the workflow. The latter can be enhanced by considering the different types of communication (SMS, voice call) in the mobile networks.

Acknowledgment

This work was supported grants to TalTech – TalTech Industrial (H2020, grant No 952410) and Estonian Research Council (PRG1573)

References

Arie, S. (2015). Can mobile phones transform healthcare in low and middle-income countries? *BMJ*, 350:h1975. https://doi.org/10.1136/bmj.h1975

Bianchi, F. M., Rizzi, A., Sadeghian, A., & Moiso, C. (2016). Identifying user habits through data mining on call data records. *Engineering Applications of Artificial Intelligence*, *54*, 49–61. https://doi.org/10.1016/j.engappai.2016.05.007

Bianchi, F. M., Scardapane, S., Uncini, A., Rizzi, A., & Sadeghian, A. (2015). Prediction of telephone calls load using Echo State Network with exogenous variables. *Neural Networks*, *71*, 204–213. https://doi.org/https://doi.org/10.1016/j.neunet.2015.08.010

CTA [Consumer Technology Association]. (2017, July). *How mobile phones are changing the developing world?* Retrieved from https://www.cta.tech/News/Blog/Articles/2015/July/How-Mobile-Phones-Are-Changing-the-Developing-Worl.aspx

Cuzzocrea, A., Ferri, F., & Grifoni, P. (2018). Intelligent Sensor Data Fusion for Supporting Advanced Smart Health Processes. In L. Barolli & O. Terzo (Eds), *Complex, Intelligent, and Software Intensive Systems*, *611*, 361–370. https://doi.org/10.1007/978-3-319-61566-0_33

de Montjoye, Y.-A., Zbigniew, S., Romain, T., Cezary, Z., & D. Blondel, V. (2014). D4D-Senegal: The Second Mobile Phone Data for Development Challenge. *CoRR*, 1–9. https://doi.org/10.48550/arXiv.1407.4885

DeAlmeida, J. M., Pontes, C. F. T., DaSilva, L. A., Both, C. B., Gondim, J. J. C., Ralha, C. G., & Marotta, M. A. (2021). Abnormal Behavior Detection Based on Traffic Pattern Categorization in Mobile Networks. *IEEE Transactions on Network and Service Management*, *18*(4), 4213–4224. https://doi.org/10.1109/TNSM.2021.3125019

Ficek, M., & Kencl, L. (2012). Inter-call mobility model: A Spatio-temporal refinement of call data records using a Gaussian mixture model. *2012 Proceedings IEEE INFOCOM*, 469–477. https://doi.org/10.1109/INFCOM.2012.6195786

Gore, R., Wozny, P., Dignum, F. P. M., Shults, F. L., van Burken, C. B., & Royakkers, L. (2019). A Value Sensitive ABM of the Refugee Crisis in the Netherlands. *Proceeding 2019 Spring Simulation Conference (SpringSim)*, 1–12. https://doi.org/10.23919/SpringSim.2019.8732867

Gross Rating Point (GRP). (2020). Retrieved from https://marketing-dictionary.org/g/gross-rating-point/

Hiir, H., Sharma, R., Aasa, A., & Saluveer, E. (2019). Impact of Natural and Social Events on Mobile Call Data Records--An Estonian Case Study. *Proceeding International Conference on Complex Networks and Their Applications*, 415–426. https://doi.org/10.1007/978-3-030-36683-4_34

Jin, X., & Han, J. (2010). K-Means Clustering. In *Encyclopedia of Machine Learning* (pp. 563–564). https://doi.org/10.1007/978-0-387-30164-8_425

KMeans Silhouette Score Explained With Python Example. (2020). Retrieved from https://dzone.com/articles/kmeans-silhouette-score-explained-with-python-exam

Leng, Y., Zhao, J., & Koutsopoulos, H. (2021). Leveraging Individual and Collective Regularity to Profile and Segment User Locations from Mobile Phone Data. *ACM Transactions on Management Information Systems*, *12*(3). https://doi.org/10.1145/3449042

Letouzé, E., & Vinck, P. (2015). *The law, politics and ethics or cell phone data analytics*. Retrieved from http://datapopalliance.org/wp-content/uploads/2015/04/WPS_LawPoliticsEthicsCellPhoneDataAnalytics.pdf

Louail, T., Lenormand, M., Cantu Ros, O. G., Picornell, M., Herranz, R., Frias-Martinez, E., Ramasco, J. J., & Barthelemy, M. (2014). From mobile phone data to the spatial structure of cities. *Scientific Reports*, *4*, 5276. https:/doi.org/10.1038/srep05276

Mamei, M., Colonna, M., & Galassi, M. (2016). Automatic identification of relevant places from cellular network data. *Pervasive and Mobile Computing*, *31*, 147–158. https://doi.org/10.1016/j.pmcj.2016.01.009

Mobile policy handbook: an insider's guide to the issues. (2017). Retrieved from https://www.gsma.com/mena/wp-content/uploads/2018/10/Mobile_Policy_Handbook_2017_EN.pdf

Murphy, K. P. (2013). *Machine learning: a probabilistic perspective*. MIT Press.

Nair, S. C., Elayidom, M. S., & Gopalan, S. (2020). Call detail record-based traffic density analysis using global K-means clustering. *International Journal of Intelligent Enterprise*, *7*(1/2/3), 176–187. https://dx.doi.org/10.1504/IJIE.2020.104654

Quercia, D., Di Lorenzo, G., Calabrese, F., & Ratti, C. (2011). Mobile Phones and Outdoor Advertising: Measurable Advertising. *IEEE Pervasive Computing*, *10*(2), 28–36.

Scharff, C., Ndiaye, K., Jordan, M., Diene, A. N., & Drame, F. M. (2015). Human mobility during religious festivals and its implications on public health in Senegal: A mobile dataset analysis. *Proceedings of 2015 IEEE Global Humanitarian Technology Conference (GHTC)*, 108–113.

Steenbruggen, J., Tranos, E., & Nijkamp, P. (2015). Data from mobile phone operators: A tool for smarter cities? *Telecommunications Policy*, *39*(3), 335–346. https://doi.org/10.1016/j.telpol.2014.04.001

Sultan, K., Ali, H., Ahmad, A., & Zhang, Z. (2019). Call Details Record Analysis: A Spatio-temporal Exploration toward Mobile Traffic Classification and Optimization. *Information*, *10*(6), 192. https://doi.org/10.3390/info10060192

Sumathi, V. P., Kousalya, K., Vanitha, V., & Cynthia, J. (2018). Crowd estimation at a social event using call data records. *International Journal of Business Information Systems*, *28*(2), 246–261. https://doi.org/10.1504/IJBIS.2018.10012931

Zhang, P., Bao, Z., Li, Y., Li, G., Zhang, Y., & Peng, Z. (2018). Trajectory-Driven Influential Billboard Placement. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2748–2757. https://doi.org/10.1145/3219819.3219946