



Orthogonal projection for anomaly detection in networking datasets

David Cortes-Polo¹ · Luis I. Jimenez² · Mercedes E. Paoletti³ · Jesus Calle-Cancho⁴ · Juan A. Rico-Gallego⁵

Received: 25 May 2022 / Accepted: 30 March 2023 / Published online: 21 April 2023
© The Author(s) 2023

Abstract

In recent years, the impressive growth of new wireless technologies, together with the appearance of new requirements in applications and services, is progressively changing the use of networks. Due to the high mobility required, the network must adapt to the infrastructure to meet the demands of the users. As a result, service providers currently have to over-provision network capacity, which is costly. In addition, considering efficient resource planning in advance involves a lot of labor-intensive efforts. Consequently, network usage analysis is a very useful tool that allows network administrators to find patterns and anomalies. Whilst pattern detection provides administrators the ability to define the infrastructure, anomaly detection provides rich and valuable information for certain applications, for example, to avoid network saturation in urban areas during peak hours. This article proposes a new methodology based on orthogonal projections over Call Detail Records (CDR) for anomaly detection to help in the dynamic management of the network in an urban area. The method is evaluated in a real scenario provided by an Italian telecommunications operator, considering different locations in the Milan metropolitan area, differentiated by the geographic resolution of the data, reaching F1 scores above 0.8. In addition, a new ground truth is presented, hoping it will become a reference data set for the community, in the form of a set of locations that have been corroborated for use in evaluating anomaly detection techniques.

Keywords B5G · Geographic information system · Mobile networks · Anomaly detection

1 Introduction

5 G and beyond (5 G and B5G) are expected to satisfy multiple challenges emerged in this mobile generation. Complex network architectures with ultra-dense deployments and multiple services with strict requirements are a common profile in this new generation of networks. The standard is developed to support 1000× the connection density,

3× the spectrum efficiency, 10× the experienced data rate and 10× the energy efficiency, among others (Zhang et al. 2019). These new network capabilities are supported by new technologies like virtual architectures, coexistence between traditional radio networks and new wave bands with massive MIMO, network slicing or the “network softwarization” (Barakabitze et al. 2020).

✉ David Cortes-Polo
david.cortes.polo@urjc.es

Luis I. Jimenez
nacho.jimenez@uva.es

Mercedes E. Paoletti
mpaollett@ucm.es

Jesus Calle-Cancho
jesuscalles@unex.es

Juan A. Rico-Gallego
jarico@unex.es

¹ Department of Signal Theory and Communication, University Rey Juan Carlos, Cam. del Molino, 5, 28942 Fuenlabrada, Madrid, Spain

² Intelligent Systems Group, Department of Computer Science, University of Valladolid, P.O de Belén, 15, 47011 Valladolid, Spain

³ Hyperspectral Computing Laboratory, Department of Technology of Computers and Communication, University of Extremadura, Av. de la Universidad, s/n, 10003 Cáceres, Spain

⁴ Extremadura Research Center for Advanced Technologies (CETA-CIEMAT), Calle Sola 1, 10200 Trujillo, Extremadura, Spain

⁵ Computer and Telematics Systems Engineering, University of Extremadura, Av. de la Universidad, s/n, 10003 Cáceres, Spain

These paradigms are focused on the suitability of the network to share data and enable the exchange of information and knowledge. They also combined with the ability to adapt to every service or user's requirement. As a result, there is an improvement in the user experience in terms of network speed, latency or reliability, and the deployment of new services on the network, such as the autonomous car, tactile Internet or remote surgery (Paolucci et al. 2021).

In this challenging context, the analysis of the network information such as that contained in Call Details Record (CDR), by means of Artificial Intelligence (AI) techniques, contribute to the improvement of the network performance, the Quality of Service (QoS) provided by the network and the Quality of Experience (QoE) experienced by users by analyzing their behaviour (Naboulsi et al. 2021). Therefore, CDR can be used to analyze human activity and mobility patterns, obtaining useful information in a wide range of fields, such as population distribution, city planning, tourism or event detection, among others (Jones et al. 2018; Zhao et al. 2016; Phithakitnukoon et al. 2012).

Moreover, B5G and 6 G standards aim to combine AI with the new technologies introduced in 5 G to transform the traditional network management into a self-adapting mechanism where the network reacts to the changing conditions with minimal human intervention (Zhang et al. 2019; Benzaid and Taleb 2020, Moysen and Giupponi 2018). In this context, the identification and analysis of unexpected or abnormal behaviours of the users is an important task that must be addressed to prevent network performance degradation or, even self-adapting available network and resources to manage these events. In this sense, anomaly detection can be considered a useful tool for optimizing network performance (Sultan et al. 2018).

In order to increase the efficiency of anomaly detection, the addition of network information metadata to the geographic coordinates provided by the CDR database offers an interesting mechanism to improve the precision of the data, revealing particular behaviours of the network. Furthermore, most proposed network-collected data analysis techniques in the literature (Garropo and Niccolini 2018; Zhu and Sun 2020; Sultan et al. 2019) focus on examining a single parameter or the sum of the set of characteristics, so the inclusion of additional information about a particular area should lead to a greater understanding of it and can be extracted through its analysis. With this view towards future progress, the independent data contained in a CDR can be represented as a set of different features in the form of a data cube to accurately describe the associated geographic area. This concept is widely used in other areas of knowledge (Racetin and Krtalić 2021). Nevertheless, it is very recently included in the analysis of communication networks (Cortés-Polo et al. 2019), hence it opens new paths in terms of its practical application. In particular, this work extends

the use of the Orthogonal Projection Network Analysis (OPNA) technique proposed in previous works (Cortés-Polo et al. 2019, 2021) to the detection of anomalous situations that may occur in the course of activity of an urban area to provide with a better understanding of what events could be considered a potential risk and coordinate the provision of services of a telecommunications network. In addition, this document provides a qualitative and quantitative study of the methodology used on a CDR of the Milan metropolitan area, adding new information about the events that occurred during the period in which this data set was collected. An additional objective of this work is to establish these facts collected as a reference in this data set for other works in this line of research.

The remainder of this document is organized as follows. Section 2 presents the existing research related to this study. Section 3 enumerates and describes the different steps of the orthogonal projection-based methodology. Section 4 describes the CDR of the metropolitan area of Milano and, in particular, the selected cells to be analyzed, as well as the metrics used to evaluate the method. Section 5 describes the experiments intended to evaluate the accuracy achieved by our implementation in the context of multi-feature data analysis. Finally, Sect. 6 concludes this work with some remarks and hints as plausible future research lines.

2 Background

CDR provides invaluable information of around 4.9 billion Internet active users [(ITU) 2021] to the network providers, as it is used in many fundamental processes like value-added services development, customers' billing, network efficiency measurements, churn or fraud detection, etc. The information contained in CDR includes anonymous user data such as the start time stamp, duration and cellular sector of origin of each activity of voice traffic, text messages and data of each device among others. Moreover, CDR has the capability to organize this information in patterns, like a grid (Barlacchi et al. 2015). This may complicate the analysis as some behaviours are masked in the order of the data itself.

The data contained in CDR provides significant information to analyze the mobility patterns developing frameworks as in Zhao et al. (2016), to understand human mobility, or in Wang and Chen (2018) where the analysis reduces the location uncertainty and oscillation of the data. In addition, frameworks can be used to examine data from other perspectives, such as gendered urban mobility like in Gauvin et al. (2020); to extract other parameters like forensic analysis as in Abba et al. (2019) or the activities of the base stations in a mobile cellular network (Jiang et al. 2020). Recently, many machine learning techniques are included in frameworks oriented to analyze and classify the information of a

CDR (Sultan et al. 2019) or to apply it to concrete areas like churn prediction (Ahmad et al. 2019; Garimella et al. 2021).

Those frameworks, which analyse CDR using machine learning techniques, can also be applied for detecting unusual events of critical significance in the network and analysing the users' activities. As explained in the previous section, CDR contains spatial and temporal information about the network, providing an approximate location and temporal record, which are very useful when detecting particular events, an important peculiarity in anomaly detection analysis. In this sense, Sultan et al. (2018) describes a mechanism to identify anomalies using the K-means clustering algorithm and how to forecast future traffic for a user. Two different approaches are developed in Garroppo and Niccolini (2018) to detect on-the-fly traffic anomalies, using the Spatial Analysis and the Time Analysis separately. In Zhu and Sun (2020), the anomalies are detected using machine learning to extract traffic patterns through clustering algorithms and forecast the traffic of the network using autoregressive models. The anomaly detection research area is drawing more and more attention from academia and industry, due to the importance of detecting undesired events in the network (Dasari et al. 2021; Jiang and Liu 2022; Shajari et al. 2022).

This paper differs from the previous works due to the methodology used to analyze the CDR, which characterizes each area as a linear combination of different unknown network features or *compartments*. These are going to be unique descriptions of a particular activity in the network in a specific area. In addition, the sequence of different characterizations of a location, ordered according to its temporal record, provides significant information to identify associated behaviour patterns, being this activity analyzed using a statistical approach. Also in this work, a new set of points of interest are analyzed in order to validate the results obtained by the OPNA technique.

3 Orthogonal projection network analysis

The information contained in the CDR dataset provides valuable information, which can be used in many applications. The OPNA technique proposes a methodology to extract the descriptive *compartments* of each interval and use them to analyze network usage. Figure 1 depicts the two phases of the proposed methodology, composed of the OPNA technique and the analysis of some relevant locations to detect events which can be considered anomalous. This second phase will be covered in the following sections.

The OPNA methodology analyzes the CDR Database extracting the user's compartments of each zone and the time interval. We consider each CDR data in an interval as a linear combination formed by the characteristic vectors multiplied by their respective abundances and modified by

an error or correction factor, as shown in Eq. (1), where y_i is the measured value of the feature i , c_{ij} defines the value of the j th *compartment* in feature i , a_j is the fractional abundance of the j th *compartment*, and e_i represents the error or correction term for the feature i in this cell.

$$y_i = \sum_{j=1}^q c_{ij}a_j + e_i \quad (1)$$

The model which describes a single interval can be generalized for the other intervals $n = 1, 2, \dots, N$. Therefore, Eq. (1) can be rewritten in compact form as in Eq. (2), where Y represents a complete CDR interval, the *compartment* set is defined by C , A is a matrix containing the fractional abundances for each *compartment*, and E is the accumulated error in matrix form.

$$Y_n = C_n A_n + E_n \quad (2)$$

The second step of the OPNA methodology shown in Fig. 1 is the compartment extraction. The proposed methodology (Cortés-Polo et al. 2019) uses the Orthogonal Subspace Projection (OSP) algorithm (Harsanyi and Chang 1994) as one of the most efficient algorithms to extract the *compartment* contained in the CDR data set. Nevertheless, other algorithms can be used as presented in Cortés-Polo et al. (2021). The algorithm aims to find the extreme points of the I-dimensional space of each interval formed by the feature array for each sub-area of the data cube. This array represents the compartment of a cell at a particular time in the data set.

Each extreme element obtained is considered part of the set of *compartments* that will characterize the complete CDR. These *compartments* vectors are ordered by their modules in ascending order, and those for which similarity is likely to be considered representatives of a single intensity level are removed. Once ordered and reduced, the behaviours are compared with each cell contained in the CDR, selecting the behaviour that minimizes the Euclidean distance. With this information, a network usage map of each of the intervals is obtained, as shown in the third stage of the OPNA methodology shown in Fig. 1.

The OPNA methodology can be helpful in analyzing the CDR data entirely and locating different events, abnormal behaviours or resources misused. As can be observed in the flowchart, the proposed technique in this work is composed of two stages. Firstly, the OPNA analysis is applied as explained before, to extract the network usage of each cell of the data set. The second part of the methodology requires the selection of some locations, based on their expected behaviour, to detect events celebrated in those areas by statistical analysis of the median absolute deviation (MAD) in the second stage.

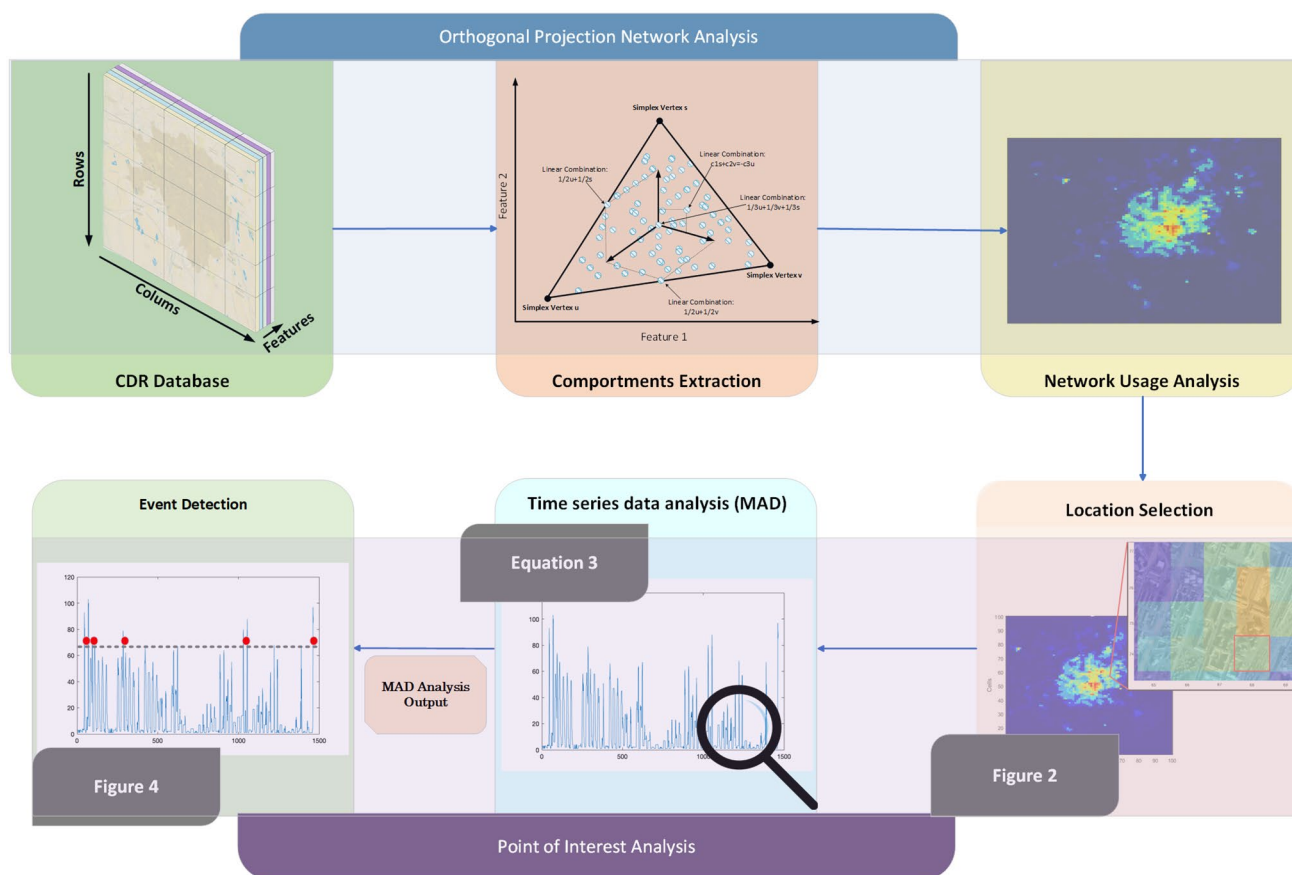


Fig. 1 Flow chart with the steps that compose each stage of the OPNA methodology and the process of detecting the events taken as reference in the analyzed cells

4 Data set and metrics

Telecom Italia and SpazioDati published the first open-source telecommunications data through the first Big Data Challenge in 2014 (Barlacchi et al. 2015), with the aim of promoting new ideas in the field through Big Data techniques. The data set used in this work summarizes the activity of Telecom Italy's telecommunications network during a period of two months, particularly November and December 2013, and specifically the activity of the city of Milano and its metropolitan area, as well as part of adjacent towns. The activity of the network is described through five representative characteristics of the use that was given to telecommunications in those years, which are described below.

- *Received / Sent SMS*: the information is generated each time a user receives or sends an SMS.
- *Incoming / Outgoing Call*: a new record is generated each time a user receives or issues a call.
- *Internet*: a record is generated each time a user initiates or finishes an Internet connection. During the same connection, a CDR is generated when the connection lasts

for more than fifteen minutes or when the user transferred more than 5MB.

Altogether, the data set has dimensions of 100×100 cells with five components in each interval, each of them being ten minutes long. The total number of intervals during the analyzed period is 8784 (144 intervals \times 61 days). For this work, the set of intervals has been reduced to a total of 1488, grouping them in complete hours.

4.1 Points of interest analyzed

Due to the very nature of the anomaly detection problem, in this work some key locations of the city of Milan have been compiled for analysis during the period studied, taking into account different locations where the occurrence of events is known. A large part of this work focuses on the presentation of several testable and verifiable events in three of the four analyzed locations, with the aim of extending the known ground truth of the data set, and expanding the functionality of the techniques that can be applied to it. Figure 2 depicts

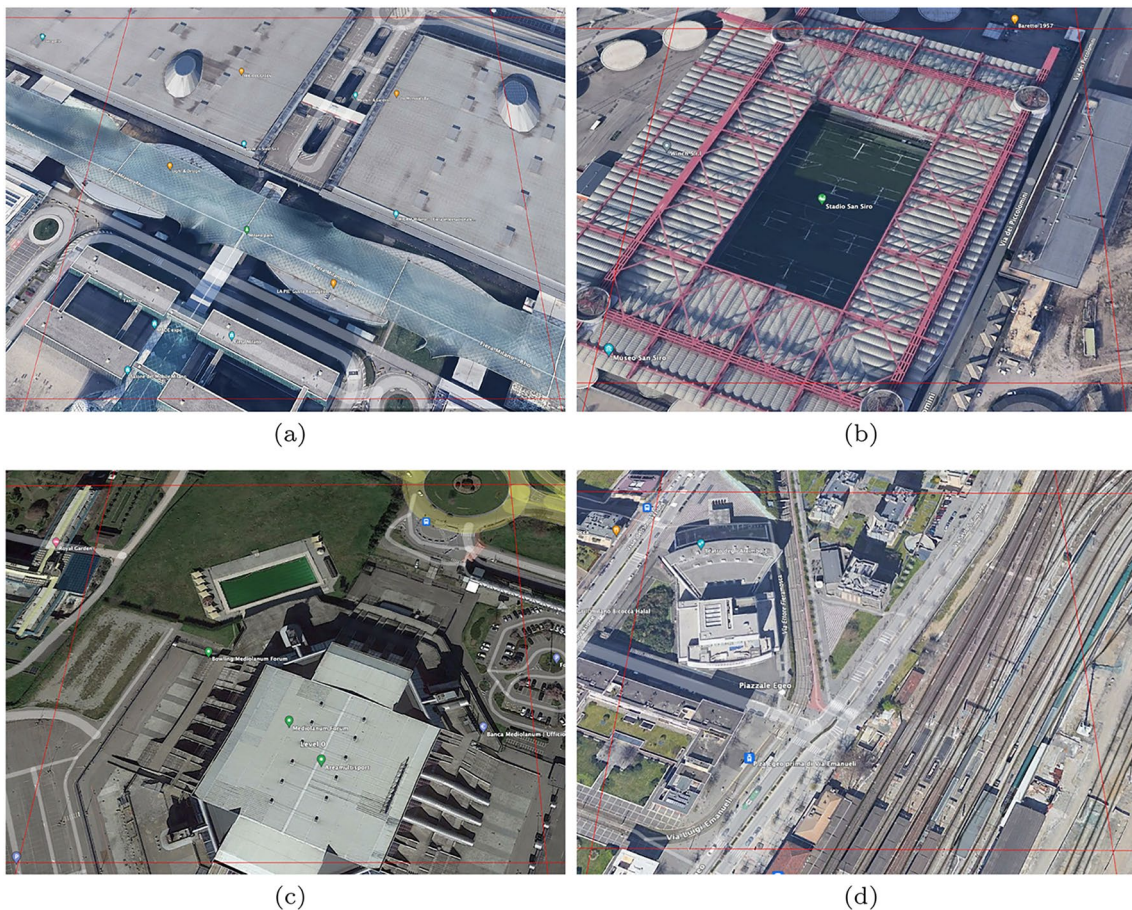


Fig. 2 In **a** is displayed Fiera Milano Rho cell, indexed with the value 7723. **b** shows the cell corresponding to the San Siro stadium the index of which is 5738. The Mediolanum Forum sports hall can be

seen in **c** and corresponds to cell 2144. In **d** at the bottom right is shown cell 7468 which is located Teatro degli Arcimboldi

the four locations that have been examined, which will be described in detail below.

The first cell of interest that we are going to consider in this work is the *Fiera Milano Rho*, one of the two complexes that comprise the Fiera di Milano, the main trade exhibition in Italy and one of the most important in the European Union. The cell analyzed for this study is the one located above the entrance of the enclosure, which is particularly interesting as its activity is distributed among several pavilions and depends on the exhibition. In this regard, some pavilions may not be used. In addition, these events have a longer duration in time compared to those examined in other cells which are directly related to the affluence of the event. Table 1 displays the events that occurred during the analyzed period, taking into account that the influx of people may vary greatly from one event to another.

Another interesting location that has been taken into account in this work is the *Giuseppe Meazza Stadium*, also known as the San Siro Stadium, with a capacity for 80,000 spectators. Indeed it is a sports venue located in the city of

Milan, where AC Milan and FC Internazionale play their home games. This point has been used in other works (Garroppo and Niccolini 2018; Parvez et al. 2017) with the same purpose of detecting abnormal behaviours as its use is very sporadic (only when there is a match since the training sessions of the Milan teams are held at the Milanello Sports Center, and the Angelo Moratti Sports Center) and the surrounding cells are devoid of activity unless there is an event in the stadium. In addition, the dates of the matches played during the network analysis period are well known as shown

Table 1 List of known events and corresponding dates during the analyzed period in the Fiera Milano Rho

Date	Event
07–10/11/13	71th Int. Motorcycle Exhibition
12–10/11/13	SINEI (Wine’s Fair)
15–18/11/13	CHIBIMART (Bijoux & accessories Fair)
30–08/12/13	Artigiano in Fiera (Christmas Market)

Table 2 List of known events and corresponding dates during the analyzed period in the San Siro Stadium

Date	Event
02/11/13	<i>NFM</i> AC Milan - AC Fiorentina
09/11/13	<i>NFM</i> FC Inter. Milano - Livorno
15/11/13	<i>IFM</i> Italy - Germany
23/11/13	<i>NFM</i> AC Milan - Genoa
01/12/13	<i>NFM</i> FC Inter. Milano - UC Sampdoria
04/12/13	<i>NFM-CI</i> FC Inter. Milano - Trapani
08/12/13	<i>NFM</i> FC Inter. Milano - Parma
11/12/13	<i>IFM-CL</i> AC Milan - Ajax
16/12/13	<i>NFM</i> AC Milan - AS Roma
22/12/13	<i>NFM</i> FC Inter. Milano - AC Milan

The *NFM* acronym corresponds to National Football Match, *NFM-CI* corresponds to Italy's Football Cup, *IFM* corresponds to International Football Match and *IFM-CL* corresponds to the Champions League competition

in Table 2, where the dates of the ten games played in the stadium by AC Milan, Inter Milan and the Italian national team are displayed.

Despite their differences, the previous locations, assume similar cell pattern, where the intensity of the use of the telecommunications network will affect the adjacent cells. That is why we will also analyze a couple of locations in which its influence is reduced to a single CDR cell. The first is the *Mediolanum Forum*, which is a sports hall located in the Italian town of Assago on the outskirts of Milan. With a capacity for 12,700 spectators, it is the court where the *Palacanestro Olimpia Milano* of the Italian national basketball league plays its home games. In terms of this analysis, it is interesting due to the large number of events held during the period examined, as it is shown in Table 3, including a variety of concerts and basketball games.

Last but not least, the cell where the *Teatro degli Arcimboldi* is located will be analyzed. The theatre has a capacity for approximately 2400 spectators hosting opera performances and other cultural events. The cell is particularly interesting given its proximity to the University Milano Bicocca and to a theatre school, which can generate distortions for the detection of the events listed in Table 4 due to the differences between the typical school and the cultural leisure schedule.

All the events mentioned in the Tables 1, 2, 3, 4 have been corroborated through the different official websites of each location and people involved, audiovisual elements on the Internet and user records on Twitter.

4.2 Metrics

To analyze the cells and extract those intervals where an event is produced, the MAD is used. The metric involves

Table 3 List of known events and corresponding dates during the analyzed period in the Mediolanum Forum

Date	Event
02/11/13	Thirty seconds to mars concert
03/11/13	Queens of the stone age concert
07–10/11/13	Disney on ice
12/11/13	Alter bridge concert
13/11/13	Artic monkeys concert
15/11/13	<i>IBM</i> O. Milano–Strasbourg
16/11/13	Negramaro concert
18–19/11/13	Max pezzuli concert
20/11/13	Emma morrone concert
22/11/13	<i>NBM</i> O. Milano–Anadolu EFES
23/11/13	Five Fingers Death Punch Concert
24/11/13	<i>NBM</i> O. Milano–Reggiana
28/11–01/12/13	Cirque Du Soleil
03/12/13	Alessandra Amoroso Concert
05/12/13	<i>IBM</i> O. Milano–Real Madrid
08/12/13	<i>NBM</i> O. Milano–S. Montegranaro
12/12/13	X Factor Show Final
13/12/13	<i>IBM</i> O. Milano–Brose Bamberg
16/12/13	Laura pausini concert
18–19/12/13	Laura pausini concerts
21–22/12/13	Laura pausini concerts
26/12/13	<i>NBM</i> O. Milano–Vanoli Cremona
29/12/13	<i>NBM</i> O. Milano–Sidigas Avellino

NBM corresponds to National Basketball Match and *IBM* corresponds to International Basketball Match

finding the median of absolute deviations from the median, and can be defined as follows:

$$MAD = b \cdot \text{Median}(x_i - \text{Median}(X)), \quad (3)$$

where the *Median*(*X*) is the median of the analyzed cells' observations, and x_i is each analyzed interval of the cell. Finally, b is a constant related to the normality of the data, disregarding the abnormality induced by outliers (Rousseeuw and Croux 1993). In particular, to select the value of b , the distribution of the data must be assumed, where b is calculated as $b = 1/Q(75)$, considering $Q(75)$ as the 0.75 quantile of the underlying distribution. In case that the data is assumed that follows a normal distribution, $1/Q(0.75) = 1.4826$ (Leys et al. 2013). The selected value is a key aspect because defines the outlier values of the distribution. Figure 3 describes an example, where one of the aforementioned cells is analyzed, depicting the *comportment* of a *Regular Day*, with the network usage inside the MAD interval, and an *Event Detection*, where the MAD analysis shows an abnormal behaviour compared to other days.

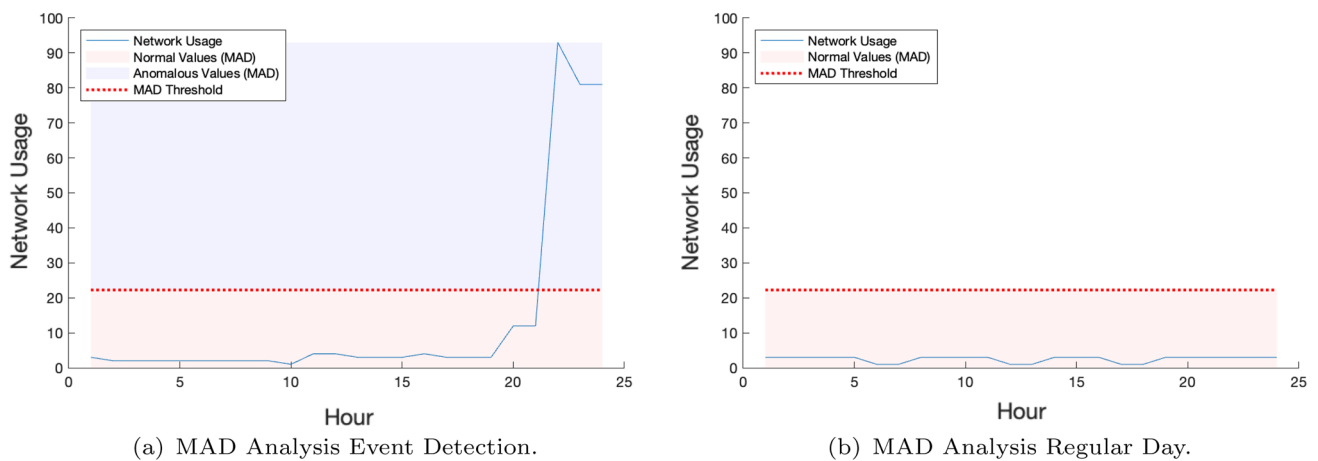


Fig. 3 **a** shows the detection of an anomaly in a single day where the MAD analysis calculates the threshold where the network usage is normal (the red zone of the graphic) and where there is an anomaly in

the network usage (the blue zone). **b** shows the same cell on another day where there are no anomalies in network usage

Table 4 List of known events and corresponding dates during the analyzed period in the Teatro degli Arcimboldi

Date	Event
02–04/11/13	Bob dylan concerts
12/11/13	Opera seconda in tour
16/11/13	Omaggio a morricone concert
21– 23/11/13	Siddhartha musical
25–26/11/13	Alessandro siani concerts
27–28/11/13	Balletto yacobson di sanpietroburgo
07–8/12/13	Ludovico einaudi concerts
09/12/13	Massimo ranieri concert
10/12/13	Max gazze concert
13–14/12/13	Mario biondi concert
17–22/12/13	Saint-Petersburg state ballet
28/12/13	Strauss festival orchestra
31/12/13	Johnny dorelly concert

For each analyzed cell, precision, recall and F1 scores are compared. In this regard, the precision is the fraction of correct predictions compared with the overall predictions made. The recall is the true positive rate and the F1 Score is the weighted average of precision and recall as it helps in balancing both parameters. The metrics are calculated as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = 2 \cdot \left(\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \right), \tag{6}$$

Table 5 Established ranges for the qualitative assessment of the results obtained from the F1 score metric

F1 score	Interpretation
> 0.9	Very good
0.8 – 0.9	Good
0.6 – 0.8	OK
< 0.6	Not good

where *TP* is the number of true positives, *FN* is the number of false negatives and *FP* is the number of false positives. The F1 score ranges from 0 to 1, where 0 is the worst possible score and 1 is a perfect score indicating that the model predicts each observation correctly. A good F1 score depends on the type of data you are working with. In this work we consider that a good value of F1 score is established above 0.8. By extension, we have established the ranges detailed in Table 5 to assess the results obtained qualitatively.

5 Experiments results

The experiments are conducted according to certain parameters in the validation of the results. The events taken as a reference are bounded from the start to end hours marked by the organization *a priori*. By dividing the data into one-hour time slices, any continuous block detected as an anomaly by the method, which totally or partially coincides with any of the established references, will be considered a hit; in case there is an anomaly detected by the method that does not have its peer in the established references, it will be considered as an unverified success; on the contrary, when the technique does not detect an anomaly in a time segment

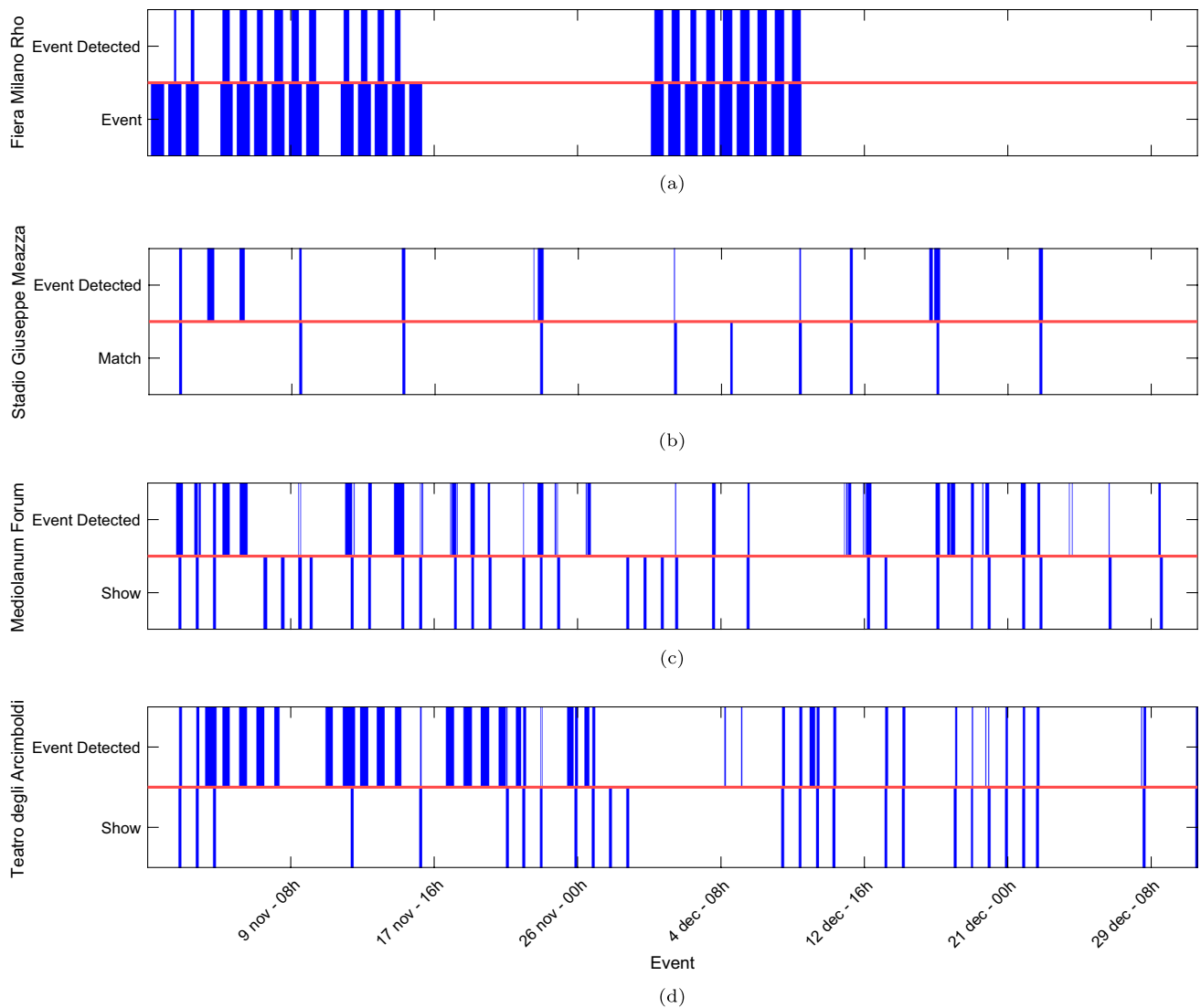


Fig. 4 Every detected event for each location in relation to the references mentioned in the document. The thickness of reference events is estimated according to the official records. **a** Fiera Milano Rho, **b** Stadio Giuseppe Meazza, **c** Mediolanum Forum and **d** Teatro degli Arcimboldi

where there is an established reference, it will be considered a failure. In this sense, the analysis of the four cells described in Sect. 4.1 has been performed, evaluating their results individually and obtaining the *precision*, *recall* and *F1* measurements that will be studied as a whole to validate the effectiveness of the method.

Table 6 describes the results obtained using the OPNA proposal to analyze the network usage of each cell of the data set and to extract the average distance between each data point and the mean of each cell using the MAD approach. Figure 4 depicts all the reference events in comparison to the anomalies detected for each location described above. The graphs show the events digitally and configuring the same time frame for easy understanding, each one of them will be analyzed in detail below.

Figure 4(a) depicts the results obtained in the cell Fiera Milano Rho. As could be observed, all events except two were detected by the proposed approach. Furthermore, there are no false positives or unverified events in this cell. The results are as expected, as the Fiera Milano Rho is divided

Table 6 Evaluation using precision, recall and F1 score of each analyzed cell

Cell	Precision	Recall	F1 score
Fiera Milano Rho	1	0.913	0.954
San Siro Stadium	0.818	0.900	0.857
Mediolanum Forum	0.806	0.781	0.793
Teatro Arcimboldi	0.649	0.923	0.762

into different well-defined pavilions and there are no other potential users in the surrounding areas.

As shown in Fig. 4(b), all but one of the reference events for the San Siro stadium are detected with the OPNA + MAD approach. According to Cortés-Polo et al. (2019), the cell usage is generally low in all matches, except in the International Football Match Italy vs. Germany on 15th of November and the derby match between FC Internazionale Milano and AC Milan on 22nd of December. In this context, it is noteworthy that users' behaviour during this type of event has changed since the dataset was obtained in 2013, due to the rise of various social networks, and only when attendance was close to full occupancy is the usage of the network relevant. It is also necessary to mention two unverified events on November 4th and 6th.

The previous locations have in common that the resolution of the mesh in which the data set is divided is adjusted to the space occupied by the enclosures in question. This fact results in a purer analysis of the method used. On the contrary, the following cells analyzed encompass an area where the analyzed location is more blurred. As a result, we can check how the introduction of "noise" affects the proposed method.

Figure 4(c), is focused on the events that occurred in the Mediolanum Forum Sports Hall, which comprises not only the stadium but also a sports area with a bowling alley, an indoor swimming pool and a gym. Events held in this multisport area can be detected as unverified events in the analysis and decrease the recall measure. In the figure, it can be clearly observed that there are six unverified events and seven missed events, mainly related to the Disney on Ice event, the Cirque du Soleil show and a Mario Biondi concert.

Finally, Fig. 4(d) shows the events detected in cell 7468, which contains the Teatro degli Arcimboldi. This cell has mixed usages, as the theatre only takes up around 40% of the cell. As is shown in Table 6, the precision obtained in this analysis is around 65% as a direct consequence of the unverified events, located in the weekdays from 4th to 8th, 11th to 15th, 18th to 22nd, 25th and 26th of November from 9 to 20 h. In addition to the theatre, the cell is also occupied by a music and theatre school called *Scuola del Teatro Musicale (STM)*. This school is opened in the intervals where the unverified events occur and mainly could be assigned to the operation of the school, as it could be seen in the Annual Report of the STM.¹ If those events were considered, the precision and the *F1* score of the algorithm will be increased up to 95% and 94% respectively. We consider that it is necessary to include this location, even though the results could

not be accurately confirmed, precisely due to the composition of this cell, and the wide margin for improvement that it may mean for new techniques that might be applied in the future.

6 Conclusions

This paper explores the application of a novel methodology to analyze and classify real user traffic, to detect network anomalies using a real CDR dataset from Telecom Italia. Leveraging the knowledge gained from previous analyses, extracting the *compartments* in each cell of the analyzed dataset provides an interesting way to find events that require some attention to handle the required network traffic. With this methodology, network operators can take advantage of the information contained in their CDR in order to improve both network performance and work efficiency, implement a new QoS rule or increase the QoE of a specific service in a zone, and analyse the behaviour of its users over time. Likewise, in this work new points of interest and contrasted references of events that occurred during the period included in the data set studied have been included in order to provide the community with the necessary elements for experimentation and testing of other techniques. A qualitative and quantitative evaluation of the proposed method has been carried out through a series of experiments that demonstrate its potential.

As future lines of action, further analysis will be performed to introduce new algorithms to detect anomalies in this dataset, which could lead to new techniques, such as supervised anomaly detection algorithms or deep learning algorithms. The results obtained by these applications should be applied to other data sets with similar characteristics such as that of the province of Trento.

Acknowledgements This publication has been possible thanks to the funding granted by the Ministry of Economy, Science and Digital Agenda of the Junta de Extremadura and by the European Regional Development Fund of the European Union under the projects IB18003, IB20040, GR21097 and GR21040, and it is also supported by the Spanish Ministry of Science and Innovation (Ref. PID2019-110315RB-I00 APRISA).

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data availability The data that support this study are available from the first author upon reasonable request (Email address of the first author: david.cortes.polo@urjc.es).

Declarations

Conflict of interest The authors declare no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation,

¹ <http://www.scuolateatromusicale.it/wp-content/uploads/2019/02/stm-report-annuale-2014.pdf>.

distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abba E, Aibinu AM, Alhassan JK (2019) Development of multiple mobile networks call detailed records and its forensic analysis. *Digit Commun Netw* 5(4):256–265. <https://doi.org/10.1016/j.dcan.2019.10.005>
- Ahmad AK, Jafar A, Aljoumaa K (2019) Customer churn prediction in telecom using machine learning in big data platform. *J Big Data*. <https://doi.org/10.1186/s40537-019-0191-6>
- Barakabitze AA, Ahmad A, Mijumbi R, Hines A (2020) 5g network slicing using sdn and nfv: a survey of taxonomy, architectures and future challenges. *Comput Netw*. <https://doi.org/10.1016/j.comnet.2019.106984>
- Barlacchi G, Nadai MD, Larcher R, Casella A, Chitic C, Torrisi G, Antonelli F, Vespignani A, Pentland A, Lepri B (2015) A multi-source dataset of urban life in the city of Milan and the province of Trentino. *Sci Data*. <https://doi.org/10.1038/sdata.2015.55>
- Benzaid C, Taleb T (2020) Ai-driven zero touch network and service management in 5g and beyond: Challenges and research directions. *IEEE Network* 34:186–194. <https://doi.org/10.1109/MNET.001.1900252>
- Cortés-Polo D, Gil LIJ, Calle-Cancho J, González-Sánchez J-L (2019) A novel methodology based on orthogonal projections for a mobile network data set analysis. *IEEE Access* 7:158007–158015. <https://doi.org/10.1109/ACCESS.2019.2949804>
- Cortés-Polo D, Gil LIJ, González-Sánchez J-L, Carmona-Murillo J (2021) A quantitative and comparative evaluation of key points selection algorithms for mobile network data sets analysis. *IEEE Access* 9:92030–92042. <https://doi.org/10.1109/ACCESS.2021.3092596>
- Dasari DB, Edamadaka G, Chowdary C, Sobhana M et al (2021) Anomaly-based network intrusion detection with ensemble classifiers and meta-heuristic scale (ECMHS) in traffic flow streams. *J Ambient Intell Humaniz Comput* 12(10):9241–9268
- Garimella B, Prasad G, Prasad M (2021) Churn prediction using optimized deep learning classifier on huge telecom data. *J Ambient Intell Humaniz Comput*. <https://doi.org/10.1007/s12652-021-03413-4>
- Garroppo RG, Nicolini S (2018) Anomaly detection mechanisms to find social events using cellular traffic data. *Comput Commun* 116:240–252. <https://doi.org/10.1016/j.comcom.2017.12.009>
- Gauvin L, Tizzoni M, Piaggessi S, Young A, Adler N, Verhulst S, Ferrer L, Cattuto C (2020) Gender gaps in urban mobility. *Humanit Soc Sci Commun*. <https://doi.org/10.1057/s41599-020-0500-x>
- Harsanyi JC, Chang C-I (1994) Hyperspectral image classification and dimensionality reduction: an orthogonal subspace projection approach. *IEEE Trans Geosci Remote Sens* 32(4):779–785
- ITU (2021) Measuring digital development facts and figures. Accessed: 2022-01-30. <https://www.itu.int/en/ITU-D/Statistics/Documents/facts/FactsFigures2021.pdf>. Accessed 30 Jan 2022
- Jiang Y, Liu G (2022) Two-stage anomaly detection algorithm via dynamic community evolution in temporal graph. *Appl Intell*. <https://doi.org/10.1007/s10489-021-03109-4>
- Jiang D, Huo L, Song H (2020) Rethinking behaviors and activities of base stations in mobile cellular networks based on big data analysis. *IEEE Trans Netw Sci Eng* 7(1):80–90. <https://doi.org/10.1109/TNSE.2018.2861388>
- Jones KH, Daniels H, Heys S, Ford DV (2018) Challenges and potential opportunities of mobile phone call detail records in health research: Review. *JMIR mHealth uHealth* 6:e9974
- Leys C, Ley C, Klein O, Bernard P, Licata L (2013) Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *J Exp Soc Psychol* 49(4):764–766. <https://doi.org/10.1016/j.jesp.2013.03.013>
- Moysen J, Giupponi L (2018) From 4G to 5G: self-organized network management meets machine learning. *Comput Commun*. <https://doi.org/10.1016/j.comcom.2018.07.015>
- Naboulsi D, Fiore M, Ribot S (2015) Stanica R large-scale mobile traffic analysis: a survey. *IEEE Commun Surv Tutor*. <https://doi.org/10.1109/COMST.2015.2491361>
- Paolucci F, Cugini F, Castoldi P, Osinski T (2021) Enhancing 5g sdn/nfv edge with p4 data plane programmability. *IEEE Netw* 35:154–160. <https://doi.org/10.1109/MNET.021.1900599>
- Parwez MS, Rawat DB, Garuba M (2017) Big data analytics for user-activity analysis and user-anomaly detection in mobile wireless network. *IEEE Trans Ind Inform* 13:2058–2065. <https://doi.org/10.1109/TII.2017.2650206>
- Phithakitnukoon S, Leong TW, Smoreda Z, Olivier P (2012) Weather effects on mobile social interactions: a case study of mobile phone users in Lisbon Portugal. *PLoS One*. <https://doi.org/10.1371/journal.pone.0045745>
- Racetin I, Krtalić A (2021) systematic review of anomaly detection in hyperspectral remote sensing applications. *Appl Sci*. <https://doi.org/10.3390/app11114878>
- Rousseeuw PJ, Croux C (1993) Alternatives to the median absolute deviation. *J Am Stat Assoc* 88(424):1273–1283. <https://doi.org/10.1080/01621459.1993.10476408>
- Shajari M, Geng H, Hu K, Leon-Garcia A (2022) Tensor-based online network anomaly detection and diagnosis. *IEEE Access* 10:85792–85817
- Sultan K, Ali H, Zhang Z (2018) Call detail records driven anomaly detection and traffic prediction in mobile cellular networks. *IEEE Access* 6:41728–41737. <https://doi.org/10.1109/ACCESS.2018.2859756>
- Sultan K, Ali H, Ahmad A, Zhang Z (2019) Call details record analysis: a spatiotemporal exploration toward mobile traffic classification and optimization. *Information*. <https://doi.org/10.3390/info10060192>
- Wang F, Chen C (2018) On data processing required to derive mobility patterns from passively-generated mobile phone data. *Trans Res Part C* 87:58–74. <https://doi.org/10.1016/j.trc.2017.12.003>
- Zhang Z, Xiao Y, Ma Z, Xiao M, Ding Z, Lei X, Karagiannidis GK, Fan P (2019) 6g wireless networks: vision, requirements, architecture, and key technologies. *IEEE Veh Technol Mag* 14:28–41. <https://doi.org/10.1109/MVT.2019.2921208>
- Zhao Z, Shaw S-L, Xu Y, Lu F, Chen J, Yin L (2016) Understanding the bias of call detail records in human mobility research. *Int J Geogr Inf Sci* 30(9):1738–1762. <https://doi.org/10.1080/13658816.2015.1137298>
- Zhu Q, Sun L (2020) Big data driven anomaly detection for cellular networks. *IEEE Access* 8:31398–31408. <https://doi.org/10.1109/ACCESS.2020.2973214>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.