A Quantitative and Comparative Evaluation of Key Points Selection Algorithms for Mobile Network Data Sets Analysis

DAVID CORTÉS-POLO1, LUIS IGNACIO JIMÉNEZ GIL2, JOSÉ-LUIS GONZÁLEZ-SÁNCHEZ3, AND JAVIER CARMONA-MURILLO1

1Department of Computing and Telematics Engineering, Universidad de Extremadura, 10003 Cáceres, Spain
2Department of Technology of Computers and Communications, Universidad de Extremadura, 10003 Cáceres, Spain
3Research, Technological Innovation and Supercomputing Center of Extremadura (CénitS), 10071 Cáceres, Spain

Corresponding author: David Cortés-Polo (dcorpol@unex.es)

This work was supported in part by the EU, European Regional Development Fund, in part by the Regional Government of Extremadura, Spain, under Grant IB18003, in part by the Spanish Ministry of Science, Innovation and Universities under Grant RTI2018-102002-A-I00, and in part by the Spanish Ministry of Science, Innovation and Universities through the Project Go2Edge under Grant RED2018-102585-T.

This work did not involve human subjects or animals in its research.

ABSTRACT In recent years, network operators are receiving an outsize amount of data due to the increasing number of mobile network subscribers, network services and device signalling. This trend increases with the deployment of 5G that will provide advanced connectivity to wireless devices and develop new services. Network analytics must allow telecommunications operators to improve their services and the infrastructure extracting useful information from large amounts of data. A methodology based on orthogonal projections was developed in order to analyze the network information and facilitate the management and the operations to network providers. In the current study, different key points selection algorithms are investigated in order to make a quantitative and qualitative evaluation and analyze the performance of those algorithms which use different approaches to select these points, which will be utilized in the methodology. A novel synthetic data set has also been developed to statistically evaluate the effect of the key points selection algorithms in the clustering, as well as, measure the performance of the aforementioned methodology. Finally, these key points selection algorithms are used in a real scenario to evaluate the impact of the different approaches in the analysis.

INDEX TERMS 5G, data analysis, mobile networks, network analytic.

I. INTRODUCTION

Nowadays, it is estimated that networks generate and exchange around 2.5 exabytes of data daily [1]. With this massive amount of data, and the expected increase in 5G [2], network operators must examine the information extracted from their infrastructure in order to analyze and manage their networks efficiently. In the mobile wireless network, there are many measurements which are continuously exchanged between the mobile node and the network. This information usually is saved in a data set called Call Detailed Record (CDR) [3], [4]. With this information, the operators can take effective control of the network and optimize their resources[5]. This information also allows service providers to optimize the customer’s experience analyzing relevant historical data, and, finally, with this network analytics, efficient network planning can be done due to the analyzed information reveals the behaviour of the users. These are the reason which network operators monitor the network performance continuously and analyze all the data extracted from it [6].

In previous works [7], [8], a novel unsupervised methodology which exploits call detailed record information collected from the core network of a real mobile cellular network was presented. This methodology uses the linear mixture model [9]–[12] in order to describe a particular data set based on the network information in the metropolitan area of Milan. The spatio-temporal information contained within the CDR helps to analyze user-specific activity in a certain region...
at a particular time and date. The methodology proposed extracts this activity, which has been called comportment, using an algorithm called Orthogonal Subspace Projection (OSP) [13]. The results obtained showed that some characteristics could be inferred from the comportment of the network due to spatio-temporal relationships with mobile network usage, such as the network usage increment in a professional football match or the network usage in different Points of Interest (POIs) like Duomo, Milano Central or Politecnico di Milano.

The information extracted from the real data set was checked against public information of those POIs. Due to the lack of data set’s ground-truth [14], it is necessary to create a synthetic data set, similar to the real one, to validate the methodology and prove its efficiency. It is also important to analyze the algorithm which selects the key points, which are used as signatures in the methodology. In previous works, the methodology only was proved using the OSP algorithm. Other algorithms must be tested to analyze the methodology and to compare the obtained results. The main contributions of this paper are summarized below:

- A qualitative and quantitative comparison between three key points selection algorithms that have been introduced to extract the comportments is part of the methodology proposed. These algorithms use different approaches to extract the key points; two of them use the simplex method to extract these key points. On the other hand, the last approach minimizes the remaining error of the analyzed data set when a new key point is selected. To the best of our knowledge, no earlier studies address the key points algorithms’ extraction using these techniques, representing and analyzing real CDR data sets.

- Two new synthetic data sets have been developed in order to analyze in-depth the methodology proposed in previous works [7], [8] and to provide a valid reference for numerical measures. These data sets are based in a fractal algorithm that allows creating different scenarios and intervals that simulate the geographic distribution of urban regions, similar to the information contained in a real CDR data set.

- Along with the experiments carried out on the synthetic data, the results of the different variations of the proposed methodology in a real scenario are presented.

The rest of this paper is organized as follows. Section II summarizes the methodology used, and presents the three key points selection algorithms. A set of experiments based on the described methodology with the signatures extracted from the three key points algorithms are introduced in Section III in order to provide an analytical comparison. Finally, Section IV contains some conclusions and future research lines.

II. KEY POINTS SELECTION ALGORITHMS
The methodology presented in [7] requires different key points which will be used to classify the network CDR, and is based on the linear model defined in Equation 1

\[ Y = U\alpha + e \]  

where \( Y \) is the complete data set, \( U \) is the comportment abundance in a particular area, \( \alpha \) is the weight matrix, and \( e \in \mathbb{R} \) is a matrix that represents the error introduced during the process. In this section, we briefly present a reminder of this methodology, its steps, and also a comparison of three methods applied in the key point extraction step, allowing us to visualize how accurate these variations are. Each method faces the problem from different approaches, enabling new applications in contexts of need, such as the available computing power.

A. METHODOLOGY TO ANALYZE MOBILE NETWORK DATA SETS
Due the new services deployed [15], [16] and the densification of the network brought by 5G, mobile communications currently provide a large amount of data. [17]. This will force the operators to track and analyze the information in order to assess the network usage. In mobile networks, part of this information is contained in the CDR data sets. These data set contains geolocated information and allows to analyze the network information to a concrete space and time evaluating the information obtained in different sub-areas covered by the network provider. These data sets can be organized as a three-dimensional data cube, as shown in Figure 1, where the first two coordinates are the spatial components and the third dimension is the network information collected in these areas and modified to ensure the anonymity of users and devices, leaving only network related information, by the network operator[14]. The first assumption made in the methodology used in this paper, concerns to the physical mobile network data...
infrastructure, supposing that there are a limited number of possible *comportments* and each sub-area is a linear combination of these limited number of signatures.

The first step in the methodology is to extract these key points, which will represent the signatures of the data set. These key points will be used as reference vectors which describe the classes found in the data set. In order to assign each sub-area to the closest *comportment* described by the extracted key point, a Euclidean distance-based technique is performed. With this technique, a relationship between each *comportment* and every sub-area is established.

The module sorts the key points to represent the results obtained, ordering the *comportments* from lowest to highest usage of the network.

Figure 2 illustrates an example of a resulting clustering in a map. In the example, the analyzed data is aggregated hourly, and hence, the number of classes analyzed is 120 [7]. The same information is superposed to the Milan city centre map, where the usage of the network can be analyzed in different sub-areas of the city. As could be observed, there are different levels of network usage, being the colored in red the highest network usage, located in the centre of the city. Middle network usage is located in the surrounding areas of the city center and the colours that represent this usage goes from green to yellow. Finally, the blue colour represents the areas with a low network usage.

**B. ITERATIVE ERROR ANALYSIS**

Iterative Error Analysis (IEA) algorithm, originally described in [18], aims to select a conjunction of cells as key points, denoted by $U$, which will be assumed as *comportments* of the data set, through the iterative reconstruction of the original data $Y$. The algorithm adds, one by one, *comportments* to the set $U$, accordingly the weights of each *comportment* $\alpha$ will vary to fit the equation. After that, it generates a reconstructed CDR $Y'$, which is compared with the original looking for differences among them, extracting the new values of $U$ in the next iteration.

Step by step, the algorithm starts selecting the first theoretical *comportment* of the set $u_0$, by calculating the average vector value of the data set (Eq. 2). This first element will not be part of the set $U$, but it is used to find the point of greatest average deviation of the set ($u_1$).

$$u_0 = \text{avg}(Y) = \frac{\sum^n_i y_i}{n} \tag{2}$$

where $n$ is the total number of sub-areas of the data set and $y$ refers to each sub-area. Using the subspace defined by its basis, $U$, for any vector $y_n = [y_0, y_1, \ldots, y_f] \in Y$, the projection of $y_n$ on $U$ is given by Equation 3.

$$\alpha_{c \times 1} = P(U)(y_n) = U(U^T U)^{-1} U^T y_n \tag{3}$$

Generalizing for all sub areas of $Y$, Eq. 4 describes the result in the weights $\alpha$ of each *comportment* selected in the set $U = \{u_0, u_1, \ldots, u_c\}$, being $c$ the number of *comportments*.

$$\alpha_{c \times n} = P(U)(Y) = U(U^T U)^{-1} U^T Y \tag{4}$$

To measure the error, the weights and the *comportment* set are used to solve the linear equation and obtain the reconstructed CDR, denoted by $Y'$, which is used to make a comparison with the original data $Y$ through the Root Mean Square Error described in Equation 5.

$$\text{RMSE}(Y, Y') = \sqrt{\frac{\sum^F_i (y_i - y'_i)^2}{F}} \tag{5}$$

As could be observed, the algorithm selects the key points, defined as *comportments* of the sub-areas. To select these key points different steps are involved to search the *comportments* that minimize the remaining error of the analyzed CDR of the network.
be the first of our selected targets set module point (HPM) in the data cube, which is going to as a vector.

The key points selection algorithm based on Orthogonal Subspace Projections [7]–[13] aims to find those extreme points in the \( f \)-dimensional space formed by the feature arrays collected for each sub-area of the data cube. Then, the variable space is the \( P \)-dimensional space, which covers one axis per variable, in which we can represent the comportment of a cell as a vector.

The first step of the algorithm starts looking for the highest module point (HPM) in the data cube, which is going to be the first of our selected targets set \( U \). For any vector \( y_u = [y_u, y_1, \ldots, y_f] \in Y \), and any subspace defined by its basis \( U \), the projection of \( y_u \) on \( U \) is also given by Equation 3.

Then, using the data set \( Y \) and the conjunction of comportments \( U \) used as the axes of the new vector space, the Euclidean orthogonal projector to \( U \) can be derived, as shown in Equation 6.

\[
P_U^\perp = I_f - U(U^TU)^{-1}U^T
\]

being \( I_f \) the identity matrix of dimensions \( f \times f \).

With the Euclidean orthogonal projector, every point in the data set can be projected to the new vector space, transforming the data to the new coordinate system and allowing selecting the extreme points, which are going to be the reference vectors for each class for the future clustering process.

The OSP algorithm, originally described in [13], iteratively set those comportments to the most extreme points in the new coordinate system, ensuring, in the end, that every cell is inside the simplex formed by those points, as explained in Algorithm 2.

**Algorithm 1 Pseudo Code of Iterative Error Analysis Algorithm**

**INPUTS:** \( y_{nxf} \in \mathbb{R}^f \)

\[
u_0 = \text{avg}(Y)
\]

\[
U = [u_0 | 0 | \ldots | 0]
\]

for \( i = 1 \) to \( i <= t \) do

\[
\alpha = P_U(Y)
\]

\[
Y'_i = U\alpha
\]

\[
j = \text{argmax}(\text{RMSE}(Y,Y'))
\]

\[
U[i+1,:) = Y[j,:]
\]

end for

\[
U = U[1:t,:]
\]

**C. ORTHOGONAL SUBSPACE PROJECTIONS**

The key points selection algorithm based on Orthogonal Subspace Projections [7]–[13] aims to find those extreme points in the \( f \)-dimensional space formed by the feature arrays collected for each sub-area of the data cube. Then, the variable space is the \( P \)-dimensional space, which covers one axis per variable, in which we can represent the comportment of a cell as a vector.

The first step of the algorithm starts looking for the highest module point (HPM) in the data cube, which is going to be the first of our selected targets set \( U \). For any vector \( y_u = [y_u, y_1, \ldots, y_f] \in Y \), and any subspace defined by its basis \( U \), the projection of \( y_u \) on \( U \) is also given by Equation 3.

Then, using the data set \( Y \) and the conjunction of comportments \( U \) used as the axes of the new vector space, the Euclidean orthogonal projector to \( U \) can be derived, as shown in Equation 6.

\[
P_U^\perp = I_f - U(U^TU)^{-1}U^T
\]

being \( I_f \) the identity matrix of dimensions \( f \times f \).

With the Euclidean orthogonal projector, every point in the data set can be projected to the new vector space, transforming the data to the new coordinate system and allowing selecting the extreme points, which are going to be the reference vectors for each class for the future clustering process.

The OSP algorithm, originally described in [13], iteratively set those comportments to the most extreme points in the new coordinate system, ensuring, in the end, that every cell is inside the simplex formed by those points, as explained in Algorithm 2.

**Algorithm 2 Pseudo Code of Orthogonal Subspace projection Algorithm**

**INPUTS:** \( y_{nxf} \in \mathbb{R}^f \)

\[
x_0 = \text{HPM}
\]

\[
U = [x_0 | 0 | \ldots | 0]
\]

for \( i = 1 \) to \( i < t - 1 \) do

\[
p = P_U^\perp Y
\]

\[
j = \text{argmax}_{1,\ldots,r} = \text{max} = p[j,:]
\]

\[
U[i+1,:) = Y[j,:]
\]

end for

**III. EXPERIMENTAL RESULTS**

In this section, the results obtained by the methodology using the key points selection algorithms described in the previous section, are discussed. Firstly, the methodology is applied to a synthetic data set to validate the results obtained by the three analyzed algorithms. After that, the methodology is applied to a real data set to evaluate the network comportments extracted by each algorithm.$^D$

**A. SYNTHETIC DATA SET**

This article presents two synthetic data sets built in order to prove the methodology with known data to validate the presented algorithms. The algorithms used with real data set only can be compared between each other but not with an objective measurement. Consequently, with those synthetic data sets, the results can be measured and validated qualitatively and quantitatively and be extrapolated to other real datasets and methodologies.
Algorithm 3 Pseudo Code of Simplex Growing Algorithm

**INPUTS:** \( Y_{n \times f} \in \mathbb{R}^f \)

\[
\begin{align*}
    u_0 &= \text{arg} \left\{ \max_r \left| \det \begin{bmatrix} 1 & 1 \end{bmatrix} \right| \right\} \\
    U &= [u_0 | 0 | \ldots | 0] \\
    \text{for } i = 1 \text{ to } i \leq t \text{ do} \\
    V(u_0, \ldots, u_i, r) &= \left| \det \begin{bmatrix} 1 & 1 & \ldots & 1 \end{bmatrix} \right| \\
    u_{i+1} &= \text{arg} \left\{ \max_r [V(u_0, \ldots, u_i, r)] \right\} \\
    U[i+1, :] &= u_{i+1} \\
    \text{end for} \\
    U &= U[1 : t, :] 
\end{align*}
\]

These two data sets are similar to the real data set, described in Section III-B. Both of them have three dimensions: 10000 \( \times \) 5 \( \times \) 144 and 10000 \( \times \) 5 \( \times \) 24, where the first dimension is an index that locates the spatial coordinates using a linear transformation \( I_{xy} = xl + y \), where \( I_{xy} \) is the value of the index, \( x \) and \( y \) are the values of the spatial coordinates on the two-dimensional grid and \( l \) the size of the square grid, the second dimension describes each cell, and the third dimension represents the time intervals.

The first data set is created following the same structure and size than the real one. The second data set reduces the number of intervals and simulates the real scenario where the intervals are grouped by hours, following the approach presented in [7]. The data set was created using a fractal distribution [20] to provide a baseline for simulating spatial patterns similar to a real environment.

This fractal distribution evolves in each interval following the time evolution as real data set. The classes used to generate each interval are selected, considering the represented spatial coordinate of the fractal and the interval. Each sub-area in the data follows the linear model described in Equation 1, and for this work, the cells combine five different classes, following the maximum number of comportments extracted in a single interval using the methodology [7]. These classes are extracted from a set of 720 or 120 unique classes depending on the number of time intervals of the created data set, which are 144 or 24, respectively.

With this method, the concentration of one class is not uniform through all the data set. The primary class is introduced in a proportion of \( \beta \), which defines the distribution of the other classes. Furthermore, the sub-areas are obtained following various premises: the zones of the map close to the border of a region are heavily mixed that in the centre of the region; the adjoining zones chose a primary class which is close, in the distance, to the primary class of the neighbour area; finally, to analyze various scenarios of comportments mixed in the zone, the data set is generated following different proportions, from cells mainly composed by the principal class, \( \beta = 90\% \), to a random proportion of each class in a cell.

This work intends to validate the methodology and the key points selection algorithms through a synthetic data set because it incorporates a reference to compare the results obtained, specifically to analyze the classes extracted and the accuracy of the clustering. Various metrics have been used to analyze the performance of the methodology with the synthetic data set. The first one is the Adjusted Cosine Similarity (ACS). As mentioned in Section II-A, the comportments can be defined as vectors formed with the analyzed components in the data set. As described previously, the Euclidean Distance is used to derive the similarity between two vectors, and this is used to classify the cell as a particular compartment. To measure the performance of the clustering algorithm, we require to use other tools, not only the length of two vectors, but also the angle that forms these vectors to measure the similarity. Equation 9 defines the measurement of the adjusted similarity between two vectors \( i \) and \( j \) [21].

\[
s(i, j) = \frac{\sum_{u \in U} (R_{u,i} - \bar{R}_u) \sum_{u \in U} (R_{u,j} - \bar{R}_u)}{\sqrt{\left( \sum_{u \in U} (R_{u,i} - \bar{R}_u) \right)^2} \sqrt{\left( \sum_{u \in U} (R_{u,j} - \bar{R}_u) \right)^2}} \tag{9}
\]

Being \( s(i, j) \) the similarity between the vectors. The value must be between \( 0 \leq s(i, j) \leq 1 \), which 0 means that both vectors are orthogonal each other and 1 is the same vector. The results extracted using the methodology are compared with the ground-truth, where each extracted key point is compared with every compartment of the ground-truth, using the ACS metric and selecting the closest match.

Figure 3 depicts the Cumulative Distribution Function (CDF) of the ACS between the synthetic data set taken as reference and the corresponding comportments obtained by the methodology using the key points selection algorithms described in Section II. The synthetic data set is generated using the fixed percentage of the main class (\( \beta \)), and a mixture percentage of the other classes. Those values go from \( \beta = 90\% \) to \( \beta = 25\% \) and the other classes are distributed from 10\% to 75\% depending on the value of \( \beta \). It has also been analyzed a random percentage of each class in every cell of the data set.

As can be observed in Figure 3, the IEA algorithm has the highest ACS among analyzed data sets. The algorithm looks the key points of the data set, which will be the comportments in the analysis, calculating the Root Mean Square Error between the selected points with other sub-areas of the CDR, choosing as new key points those with the highest
error. In every data set analyzed, up to 25% of the key points detected by the IEA algorithm are below 99% of similarity. That means that 75% of the key points extracted are comportments of the ground-truth. The other 25% of the key points, found by the algorithm, are vectors that are not in the ground-truth, but they are real data extracted from the data set. The algorithm chose those points due to the mean square error value is higher than other candidates.

Figure 3 also shows that there are around 5% of the analyzed points, using IEA algorithm, that are near-orthogonal. This means that the ACS value is near to 0 because the selected key points are not in the ground-truth, and form an
angle near to 90 degrees between the compartment extracted from the ground-truth and the key point selected by the IEA algorithm.

The SGA algorithm reduces the number of key points with high ACS when the mixture between the primary class and the others decrease. As shown in Figures 3(c) to 3(e), the number of classes with error are reduced compared with IEA algorithm. Figures 3(a) and 3(b) show the results of the SGA algorithm when the value of $\beta$ is small and the mixture of the different classes is high. The number of missed classes with low ACS is increased, and it is similar to the IEA algorithm, due to the SGA algorithm looks points which maximize the volume of the simplex, but these points are not close to the original class contained in the ground-truth. This is shown in Figure 3(a), where around 72% of the key points have been found in the ground-truth with a similarity of 99%. Meanwhile, around 13% of the key points have a similarity between 75% to 99% and the 3% of the classes are near-orthogonal.

Finally, the OSP algorithm finds around 87% of the compartments included in the ground-truth in the worst scenario, which is the random mixture data set (Figure 3(a)). With this algorithm, around 10% of the key points have an ACS between 75 and 99%. This result can be also observed with $\beta = 25\%$ and $\beta = 50\%$. When the mixture is reduced, using a $\beta$ greater than 50%, the OSP algorithm finds up to 96% of the compartments included in the ground-truth.

Analyzing all the scenarios, the average number of key points detected with an ACS greater than 99% using OSP algorithm is 91.83%, whereas with IEA algorithm the number of key points detected is around 77.83% and with SGA algorithm is 81.16%. The value obtained by the OSP algorithm is due to the point selection strategy followed, which enclose the rest of the analyzed points solving the simplex problem. The SGA algorithm calculates points which maximize the volume of the simplex, but under some conditions with a high mixture, these calculated points increase the ACS compared with OSP. The IEA key points selection algorithm selects the points of the data set with a higher ACS value than OSP because the algorithm selects points which have higher error compared with the rest of the points of the interval, but those points in some cases do not represent a compartment in the data set.

The second metric used to evaluate the key points selections algorithms is the overall accuracy. This measurement allows quantifying the clustering performance, evaluating the points well classified and the misclassified ones. The overall accuracy calculates the ratio between the correctly classified sub-areas and the total sub-areas in the data set. Numerically, the overall accuracy of the methodology is described in Table 1.

As could be observed the IEA algorithm has a low overall accuracy due to the number of misclassified sub-areas. In general, the clustering has missed around 35% of the 240,000 classified sub-areas using the IEA algorithm. Figure 4(a) shows the similarity of the compartments chosen to classify those sub-areas using the IEA algorithm. With this analysis, we can evaluate if the classes chosen by the methodology are near-orthogonal with the ground-truth, and do not exist in it, or they are extracted from the ground-truth, and measure the similarity with the original class. As could be observed, the methodology mainly has assigned compartments with similarity from 67% to 100% to the misclassified sub-areas. This means that some intermediate classes are not extracted using this algorithm and the sub-area is classified to the closest compartment extracted. It is remarkable that in the analysis with $\beta = 75\%$, the accuracy is increased. This can be observed in Figure 4(a) because the classes misclassified do not cover the same number of sub-areas that other data sets. This effect can be produced by various factors, the first one is because some intermediate classes are found, and many sub-areas are not misclassified with other closer compartment, and the second one is because the classes not found do not cover the same number of sub-areas, and this implies that the number of misclassified ones is reduced.

The SGA algorithm accuracy increases when the mixture of the data set is decreased. As could be observed in Figure 4(b), the reduction of the random generated points due to the mixture of the classes facilitates the calculation of the maximum volume. In the random data set, the overall accuracy decreases, reducing the similarity of the classes, as shown in Figure 3(a), which implies a reduction of the algorithm accuracy, similar to the IEA algorithm.

Finally, the OSP algorithm accuracy also increases when the mixture of the data set decrease. Table 1 shows the results of this algorithm when it is applied to different data sets.

### Table 1. Overall accuracy expressed in percentage of the methodology applied to the 24 and the 144 interval synthetic data set using the analyzed key points selection algorithms, SGA, OSP and IEA.

| Intervals | Algorithm | $\beta = 10\%$ | $\beta = 25\%$ | $\beta = 50\%$ | $\beta = 75\%$ | $\beta = \text{random}$ |
|-----------|-----------|----------------|----------------|----------------|----------------|----------------------|
| 24        | IEA       | 62.32          | 64.57          | 66.13          | 73.39          | 69.81                |
|           | OSP       | 99.44          | 97.81          | 92.70          | 92.98          | 87.07                |
|           | SGA       | 97.64          | 92.67          | 78.28          | 75.18          | 64.83                |
| 144       | IEA       | 61.89          | 58.93          | 63.09          | 64.45          | 61.92                |
|           | OSP       | 98.77          | 95.51          | 90.01          | 82.68          | 80.45                |
|           | SGA       | 95.62          | 84.75          | 73.28          | 68.36          | 64.62                |
FIGURE 4. Overall accuracy analysis which measures the number of sub-areas missed in the analyzed data sets. The classes are grouped by similarity calculated with ACS. The x-axis shows the similarity percentage with ground-truth. The y-axis describes the percentage of the cells with error. Figure (a) shows the results of the analysis with IEA algorithm. (b) describes the results with SGA algorithm and, finally, (c) shows the results with OSP algorithm.

As can be observed, this algorithm has the best overall accuracy of the three key point extraction algorithms analyzed. The number of failed sub-areas decreases even in the random mixture data set, which presents an 87% accuracy. As shown in Figure 4(c) the analysis of both data set, $\beta = 50\%$ and $\beta = 75\%$ presents similar results, around 8% of the 240,000 misclassified sub-areas, in contrast with the 25% of the failed sub-areas in SGA or around 35% of IEA misclassified sub-areas. In the $10000 \times 5 \times 144$ data set with 144 intervals, the number of classes is increased, which implies the number of failed sub-areas will increase due to this increment.

The third tool to measure the performance of the three key points selection algorithms is the average accuracy which measures the accuracy of each comportment classified in the interval. This measurement allows quantifying the performance of the class analysis. The average accuracy calculates the ratio between the correctly classified sub-areas of each class and the total sub-areas of this class in the data set. Numerically, the average accuracy of the methodology is described in Table 2.

The results are similar to the overall accuracy, where the highest accuracy is observed in the OSP algorithm, followed by the SGA algorithm. Both of them increase the accuracy when the mixture of the data set decrease. IEA algorithm experiences the same behaviour of the overall analysis, and the accuracy does not increase when the mixture decrease. This is expectable because, as could be observed in Figure 3, up a 15% of the comportments found in the data set are not similar to the ground-truth.

To analyze the type of classes with low average accuracy, the comportments have been ordered using the Euclidean norm, which denotes the magnitude of the class, making those
TABLE 2. Average accuracy expressed in percentage of the methodology applied to the 24 and the 144 interval synthetic data set using the analyzed key points selection algorithms, IEA, OSP and SGA.

| Intervals | Algorithm | \( \beta = 10\% \) | \( \beta = 25\% \) | \( \beta = 50\% \) | \( \beta = 75\% \) | \( \beta = \text{random} \) |
|-----------|-----------|----------------|----------------|----------------|----------------|----------------|
| 24        | IEA       | 73.46          | 73.12          | 66.87          | 71.86          | 70.06          |
|           | OSP       | 97.34          | 95.59          | 92.15          | 90.64          | 87.83          |
|           | SGA       | 95.27          | 90.38          | 80.76          | 73.72          | 70.11          |
| 144       | IEA       | 70.73          | 67.28          | 65.35          | 62.93          | 63.44          |
|           | OSP       | 96.51          | 94.12          | 89.29          | 83.45          | 82.19          |
|           | SGA       | 93.53          | 85.19          | 74.49          | 68.62          | 66.01          |

FIGURE 5. Average accuracy analysis classifying the comportments with less than 90% of accuracy by network usage profile.

with higher magnitude considered higher network usage and those with lower magnitude considered lower network usage. These values are referred relatively to each other due we have to assume that a signature with zero values in every component represents no network usage, and we do not know the highest values those components can reach, as proposed in the methodology [7]. Figure 5 shows the average number of classes with an accuracy lower than 90%. As could be observed, from 60 to 80% of these classes are located in the low network usage profile, which corresponds with the lower classes profile with a small Euclidean norm value. This means that the components which conform the compartment vector have lower values than the others, and the modules of the vectors are close to each other.

Finally, as described in the previous evaluation metrics, the key points algorithms differ in some aspect among them, as can be observed in Figure 6. These figures depict the clustering using the analyzed key points selection algorithms. In the analyzed data set, there are some errors in the clustering, for example at the end interval using the IEA algorithm where some sub-areas are classified as a lower network usage profile, whereas in the ground-truth, the sub-areas are classified as a high network usage profile. In the beginning interval, using the SGA algorithm, some sub-areas are classified as high network usage profile wherein the ground-truth are low profile. There are also some errors in the intermediate interval analyzed by OSP algorithm where some sub-areas are classified like low network usage profile, and in the ground-truth are medium network usage profile.

In conclusion, we can assure that OSP algorithm obtains better results in the analyzed data set than IEA or SGA algorithms, obtaining SGA better results than IEA under certain conditions where the mixture is not very high. OSP obtain good results using different mixtures (different \( \beta \) values) and, as shown in Figure 4(c), this algorithm obtains the lowest percentage of error in the classification of the cells.

B. REAL DATA SET

At the beginning of 2014, Telecom Italia launched the first edition of the Big Data Challenge. This competition intended to stimulate the creation and development of innovate technological ideas in the field of Big Data. After the challenge, these data sets were released under the name of Open Big Data [14]. The data set contains around 745 Gb of data that includes telecom, social data, meteorological information, air quality information and news from Milan and Trento province. Therefore, these data sets include information of urban and rural areas, and covers two months (from November to December 2013).

The information provided by the data set includes the received and sent SMS, which is the information generated each time a user sends or receives an SMS, and it is recorded in the CDR. Also, the data set covers the incoming and outgoing calls where a new record is generated each time a user makes or receives a call. Finally, the data set also encompasses the Internet information where a record is generated each time a user starts or ends an Internet connection. The information will also be annotated in the CDR if the connection lasts for more than 15 minutes, or the user transferred more than 5 MB.

By aggregating the extracted data, the data set was created to provide the volume of SMS, calls and Internet traffic activity for different time intervals. This work analyzes the Milan data set which extension is divided into square sub-areas of \( 235 \times 235 \) m. All the information contained in the data set is organized as data cubes containing the spatio-temporal information of the sub-areas analyzed. This definition of the
FIGURE 6. Heat map representation of the synthetic data set analysis. Figures (a), (b) and (c) shows the clustering made by the ground-truth. Figures (d), (e) and (f) represents the results obtained using the IEA key points selection algorithm in three different intervals of the data set. Figures (g), (h) and (i) represents the results obtained using the OSP key points selection algorithm in the same intervals of the data set. Finally, Figures (j), (k) and (l) represents the results obtained using the SGA key points selection algorithm.
data set allows us to perform a set of experiments through the period of time recorded. The methodology presented in [7] is used to analyze the comportments of the sub-areas.

In this paper, some well-known Points Of Interest [7] are evaluated in order to analyze the performance of the key points selection algorithms used by the methodology. Due to
the lack of ground-truth in the real data set, the results of the analysis used as the reference will be those obtained with the OSP algorithm because this algorithm has the best results with different mixtures in the synthetic data set.

Figure 7 shows the comportment analysis of a random day using the three key points selection algorithms. The chosen day is Monday, December 2\textsuperscript{nd} 2013.

As could be observed, the classes used to describe the behaviour of the sub-areas are similar in the IEA and the OSP algorithms. Those comportments classified as low network usage profile in OSP are classified in the same profile in IEA algorithm. Medium-high network usage profile classes in IEA algorithm are classified in a higher network usage profile class as can be observed in the comportments which describes the behaviour of the Politecnico di Milano (5772) or in the Duomo (5059). The SGA algorithm classifies the comportments of all the POIs in a higher network usage profile. This is because the SGA algorithm extracts a high number of low network profile classes from the data set that they are not found in the other two algorithms. When the comportments are ordered by module, the colour assignment of these comportments is different from the other algorithms. As could be observed, it is not fair compare the three algorithms using the colour scale as shown in Figure 7. A new method is required to measure the similarity of the results obtained with the three key points algorithms.

The new analysis measures the similarity of the comportments using the adjusted cosine similarity presented in Equation 9. The analysis corresponds with the entire data set, analyzing each interval of every day, obtaining 24 measurements (24 intervals of a day) with 61 repetitions (from November to December 2013) with a confidence level of 95\%. The comportments obtained by the OSP algorithm will be compared with those extracted by the other two analyzed key points algorithms. As an example, Figure 8 shows the results of three analyzed POIs, which represents the comparison of the three key points selection algorithms.

Figure 8(a) compares the result of the three algorithms in the Duomo, cell (5059). As could be observed the similarity of the three algorithms is around a 98\% in this sub-area, which can be assured that the same comportments describe the behaviour of this sub-area and using any algorithm: OSP, IEA or SGA.

The same behaviour is observed in Figure 8(b), which analyzes the Politecnico di Milano, cell (5772). The average similarity is near 98\% in each interval except in the intervals from 4 to 6 a.m. where the similarity decreases to 97\% in the analysis of OSP and IEA and around 95\% in the analysis of OSP and SGA. In these intervals, the sub-area is usually classified as a low network usage profile, but in both cases, the similarity of these intervals decrease due to the class selected as low network usage profile differs in one or various components of the comportment chosen by OSP.

Figure 8(c) shows the distance analysis of the Mercato of Milan, cell (4874). As can be observed, in the intervals from 1 to 5 a.m. the similarity decreases in both algorithms, IEA and SGA. This is due to the profile of these intervals is one of the lowest network usage profiles. In the IEA the similarity decreases around 3\%, and in the SGA algorithm results, the decrement is around 30\%. Similarly, with the previous analysis, these intervals coincide with the lower activity of the network in this area and the key points chosen as low network usage profile in these intervals differs from chosen by OSP algorithm. In this sub-area, the other intervals have an average similarity around 91\% using SGA algorithm and 98\% using IEA algorithm.

IV. CONCLUSION

This work has evaluated in-depth the methodology proposed in previous works to analyze and extract knowledge of the next-generation networks. This information is important in order to plan and manage the network efficiently. Moreover, with the deployment of 5G and the integration of a huge number of devices, the mechanisms to analyze the network and take advantage of this information will be in a short time a key aspect.

This article presents a novel synthetic data set similar to a real CDR, to analyze in-depth the methodology. This synthetic data set provides the ground-truth necessary to evaluate the accuracy of the methodology and allows to analyze any new algorithm or technique added to it. As demonstrated, with this data set, three key points selection algorithms have been compared fairly concluding that the OSP key points selection algorithm obtain better results in the synthetic data set. A real data set has also been analyzed with these algorithms, analyzing different sub-areas, which comportments are well known. This analysis shows similar results in the distance analysis, except when the network usage is the lowest one where the algorithms can provide different results because of the closeness of the comportments.

In future research, the spatial component of the data set must be analyzed in-depth, evaluating the neighbour sub-areas in order to extract valuable information of contiguous sub-areas of the data set, which can approximate the comportment of the key points selection algorithms. In addition, the effectiveness of different classification methods like other clustering algorithms or traditional data mining methods, which will be integrated with the methodology, will be proved.

REFERENCES

[1] M. S. Hadi, A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, “Big data analytics for wireless and wired network design: A survey,” Comput. Netw., vol. 132, pp. 180–199, Feb. 2018.
[2] K. Sultan, H. Ali, and Z. Zhang, “Big data perspective and challenges in next generation networks,” Future Internet, vol. 10, no. 7, p. 56, Jun. 2018.
[3] V. D. Blondel, A. Decuyper, and G. Kriangs, “A survey of results on mobile phone datasets analysis,” EPJ Data Sci., vol. 4, no. 1, p. 10, Dec. 2015.
[4] M. S. Mahdizadeh and B. Baharak, “A regression framework for predicting user’s next location using call detail records,” Comput. Netw., vol. 183, Dec. 2020, Art. no. 107618.
[5] Y. Qiao, Z. Xing, Z. M. Fadullah, J. Yang, and N. Kato, “Characterizing flow, application, and user behavior in mobile networks: A framework for mobile big data,” IEEE Wireless Commun., vol. 25, no. 1, pp. 40–49, Feb. 2018.
[6] M. S. Parwez, D. B. Rawat, and M. Garuba, "Big data analytics for user-activity analysis and user-anomaly detection in mobile wireless network,” *IEEE Trans. Ind. Inform.*, vol. 13, no. 4, pp. 2058–2065, Aug. 2017.

[7] D. Cortés-Polo, L. I. J. Gil, J. Calle-Cancho, and J.-L. González-Sánchez, “A novel methodology based on orthogonal projections for a mobile network data set analysis,” *IEEE Access*, vol. 7, pp. 158007–158015, 2019.

[8] D. Cortés-Polo, L. I. J. Gil, J.-L. González-Sánchez, and J. Calle-Cancho, “A methodology for network analysis to improve the cyber-physicals communications in next-generation networks,” *Sensors*, vol. 20, no. 8, p. 2247, Apr. 2020.

[9] L. Drumetz, J. Chanussot, and C. Jutten, “Variability of the endmembers in spectral unmixing,” in *Hyperspectral Imaging* (Data Handling in Science and Technology), vol. 32, J. M. Amigo, Ed. Amsterdam, The Netherlands: Elsevier, 2020, pp. 167–203.

[10] L. He, J. Li, C. Liu, and S. Li, “Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1579–1597, Mar. 2018.

[11] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, “Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 2, pp. 354–379, Apr. 2012.

[12] A. Plaza, G. Martin, J. Plaza, M. Zortea, and S. Sánchez, “Recent developments in endmember extraction and spectral unmixing,” in *Optical Remote Sensing*. Berlin, Germany: Springer-Verlag, 2011, pp. 235–267.

[13] J. C. Harsanyi and C.-I. Chang, “Hyperspectral image classification and dimensionality reduction: An orthogonal subspace projection approach,” *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 4, pp. 779–785, Jul. 1994.

[14] G. Barlacchi, M. De Nadai, R. Larcher, A. Casella, C. Chitic, G. Torrisi, F. Antonelli, A. Vespiagnani, A. Pentland, and B. Lepri, “A multi-source dataset of urban life in the city of milan and the province of trentino,” *Sci. Data*, vol. 2, no. 1, pp. 1–15, Dec. 2015.

[15] A. Ghosh, A. Maeder, M. Baker, and D. Chandramouli, “5G evolution: A view on 5G cellular technology beyond 3GPP release 15,” *IEEE Access*, vol. 7, pp. 127639–127651, 2019.

[16] J. G. Andrews, S. Buzzi, S. V. Hanly, A. Lozano, A. C. Soong, and C. and J. Zhang, “What will 5G be?” *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[17] J. Liu, M. Sheng, L. Liu, and J. Li, “Network densification in 5G: From the short-range communications perspective,” *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 96–102, Dec. 2017.

[18] R. Neville, K. Staenz, T. Szeredi, J. Lefebvre, and P. Hauff, “Automatic endmember extraction from hyperspectral data for mineral exploration,” in *Proc. 21st Can. Symp. Remote Sens.*, 1999, pp. 21–24.

[19] C.-I. Chang, C.-C. Wu, W. Liu, and Y.-C. Ouyang, “A new growing method for simplex-based endmember extraction algorithm,” *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 10, pp. 2804–2819, Oct. 2006.

[20] G. Martin and A. Plaza, “Region-based spatial preprocessing for endmember extraction and spectral unmixing,” *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 4, pp. 745–749, Jul. 2011.

[21] S. Roman, S. Axler, and F. Gehring, *Advanced Linear Algebra*, vol. 3, New York, NY, USA: Springer, 2005.

---

**David Cortés-Polo** received the degree in computer science from the University of Extremadura, Spain, and the Ph.D. degree in telematics from University of Extremadura, in 2015. From 2011 to 2014, he worked as a Researcher and a Teaching Assistant with the University of Extremadura. Since 2020, he has been an Associate Professor with the Department of Computing and Telematics System Engineering, Universidad de Extremadura. His main research interests include IP-based mobility management protocols, performance evaluation, and network CDR analytics.

**José-Luis González-Sánchez** received the Engineering degree in computer science form the Polytechnic University of Cataluña, Barcelona, Spain, and the Ph.D. degree in computer science from the Polytechnic University of Cataluña, in 2001. He has worked for years at several private and public organizations as a system and network manager. He is currently a full time Associate Professor of the Computing Systems and Telematics Engineering Department, University of Extremadura, Spain. He is also the General Manager of CénitS.

**Javier Carmona-Murillo** received the Ph.D. degree in computer science and communications from the University of Extremadura, Spain, in 2015. From 2005 to 2009, he was a Researcher and Teaching Assistant. Since 2009, he has been an Associate Professor with the Department of Computing and Telematics System Engineering, Universidad de Extremadura. During the past years, he has spent research periods with the Centre for Telecommunications Research, King’s College London, U.K., and Aarhus University, Denmark. His current research interests include 5G networks, mobility management protocols, performance evaluation, and the quality of service support in future mobile networks.

---