Load Profiles Clustering and Knowledge Extraction to Assess Actual Usage of Telecommunication Sites

(Article begins on next page)

Load Profiles Clustering and Knowledge Extraction to Assess Actual Usage of Telecommunication Sites

Simone Eiraudo*, Luca Barbierato*, Roberta Giannantonio†, Alessandro Porta†,
Andrea Lanzini*, Romano Borchiellini*, Enrico Macii*, Edoardo Patti*, and Lorenzo Bottaccioli*
*Energy Center Lab, Politecnico di Torino, Turin, Italy. Email: name.surname@polito.it
†TIM S.p.A., Italy, Email: name.surname@telecomitalia.it

***Abstract*—Deep awareness of a particular industry sector represents a fundamental starting point for its energy efficiency enhancement. In this perspective, a huge amount of industrial facilities' energy measurements are collected thanks to the widespread usage of monitoring systems and Internet-of-Things infrastructures. In this context, data mining techniques allows an effective exploitation of data for knowledge extraction to automatically analyse such enormous amount of data. This paper investigates a large data set including real telecommunication sites' aggregate electrical demand provided by the largest telecommunication service provider in Italy. The goal is the assessment of the actual usage category of telecommunication sites, aiming at supporting the facility management of the company and the energy knowledge discovery of each site category. A novel methodology is proposed that includes *i)* a proper normalisation method focused on energy Key Performance Indicators for telecommunication network energy management, *ii)* a time series decomposition tool to extract trends and periodical fluctuation of telecommunication sites' aggregated electric demand, and *iii)* the application of a k-Means clustering algorithm to assess sites' actual usage. The proposed methodology results in accurate outcomes, which witness the potential for practical application and discloses opportunities for further developments.**

***Index Terms*—Energy Efficiency, Telecommunication Service Provider, Clustering, Machine Learning.**

I. Introduction

The industry sector energy efficiency enhancement is a crucial challenge for the clean energy transition strategies of the European Community [1]. In fact, energy efficiency plays a fundamental role regarding this sector, representing a practical measure aimed at reducing costs and enhancing companies economical competitiveness and sustainability [2]. In particular, Information and Communication Technology (ICT) branch is experiencing an exponential growth of energy demand which increased by an annual rate of about 10% over the last decade [3]. Telecommunication service provider companies, which are part of the ICT sector, are responsible for 7% of the global electrical demand [4]. Data centres and data transmission networks represent two of the most important contributions to the telecommunication service provider energy consumption, reaching a share of about 1% each of the worldwide total electrical demand in 2019 [5]. For instance, the telecommunication industry in our country was responsible for 3.863 GWh of electrical demand in 2018 [6], which corresponds to 1.27% of the total consumption.

Over the last years, the research community has pointed out many issues related to the energy efficiency enhancement of telecommunication service provider industrial premises, such as innovative thermal management strategies [7] (e.g. free cooling application), structural transformation, and renewable energy sourcing [8]. Accordingly to what reported in the latter document, smart monitoring systems and the Internet-of-Things (IoT) paradigm represent the starting point to enhance awareness over energy behaviour of telecommunication sites. In this perspective, acquisition and investigation over huge amounts of data have become possible thanks to the widespread of measurement collection and data storage systems. At the same time, data mining techniques gained popularity to the purpose of an effective cognitive enhancement. Machine Learning (ML) emerged as the most promising discipline not only with the perspective of investigating data and extracting knowledge from them, but also as a tool to take practical actions in support of a more energy efficient facility management.

For these reasons, this paper deals with the challenge of enhancing knowledge extraction regarding energy behaviour of telecommunication sites. Particularly, it discusses the application of a ML-based methodology applied to a massive telecommunication sites load profiles data set. In fact, this work focuses on real-case data, provided by the largest telecommunication service provider in Italy. The investigated data set includes Central Offices (CO), Data Centers (DC) and Radio Base Stations (RBS). The proposed methodology includes *i)* the application of a time series decomposition model to obtain trends and periodical components of aggregated electric profiles, and *ii)* a clustering algorithm to retrieve periodical components similarities between groups of telecommunication sites. The aim is to provide a robust computational tool able to detect whether a facility has been wrongly labelled and to extract characteristic features from each usage category (i.e. CO, DC, or RBS).

The proposed work is organised as follows. Section II presents some recent research efforts on ML tools to investigate electric load patterns and presents the novelties of the paper. Section III introduces the proposed case study and includes an essential theoretical background about energy balance in telecommunication sites. Moreover, it presents the employed ML-based methodology, explaining the applied steps and algorithms. The outcomes from each step of the analysis are provided and discusses in Section IV to demonstrate the correctness of the obtained results. Final remarks and potential future developments are summarised in Section V.

II. Related works

Clustering techniques are one of the most widely adopted algorithms for unsupervised learning. Specifically, a significant research effort has been spent to make time series clustering effective. Clustering algorithms [9] and distance measures [10] have been mostly developed, such as k-Means and Euclidean distance that represent a widely adopted clustering method for time series. Nonetheless, novel partitional clustering algorithms have been designed to the specific purpose of handling time series [11]. Damayanti et al. [12] employ a k-Harmonic Means algorithm to effectively cluster electrical load profiles, achieving higher performances in terms of intra-cluster homogeneity and inter-cluster separation with respect to k-Means and Fuzzy c-Means. Yang et al. [13] instead employed the k-Shape algorithm, introduced by Paparrizos et al. in [14], in order to analyse energy consumption patterns from academic buildings in Singapore. The need for novel clustering algorithms is linked to the need of effectively comparing time series, by employing tools featuring shift, scaling and complexity invariance.

Similarly, the application of distance measures other than traditional Euclidean distance is gaining interest. In this perspective, Batista et al. introduced and tested a complexity-invariant distance in [15], proving its effectiveness for many applications. The aforementioned algorithms and distance measures enhanced the diffusion of time series clustering for many purposes, including pattern recognition, knowledge extraction, data reduction and outliers detection [16].

Clustering of load profiles has been widely employed for classification of buildings and market segmentation. In fact, many studies focused on clustering residential electricity customers. To this purpose, Motlagh et al. [17] converted electric load time series into signature objects by means of neural regression and subsequently performed clustering. Ryu et al. [18] instead reduced time series dimensionality by encoding year-round residential load profiles by means of convolutional auto-encoder. K-Means clustering algorithm was then employed, resulting in dramatically lower computational time with respect to clustering of the original time series. Moreover, Zarabie et al. [19] assessed that Affinity Propagation algorithm outperformed traditional k-Means, k-Medoids and spectral algorithm in residential load profiles clustering.

Different works pointed out the meaningfulness of periodical behaviour of time series. Sun et al. [20] decomposed time series by means of ensemble empirical model decomposition. This work made use of k-Means and Support Vector Regression in the successful perspective of providing accurate solar radiation forecasting. Giordano et al. [21] instead identified homogeneous electricity customers by employing a spectral-based approach, which revealed itself as suitable for clustering of time series featuring strong periodicity.

Considering the scarcity of industrial application and studies, this work addresses the fundamental task of investigating a real-case massive electric load data set pertinent to telecommunication sector provided by the largest telecommunication service provider in Italy. A novel methodology is proposed that aims at effectively analyse telecommunication sites aggregate load profiles to enhance the telecommunication service provider company facility management. Firstly, the employed normalisation method, based on telecommunication energy Key Performance Indicators (KPI), allows comparison within different sites' load size, preserving the information contained in the original data set. Then, it performs a decomposition model of time series, in accordance with the most significant periodical components, to highlights particular trends and periodical patterns of aforementioned normalised aggregated load profiles. Finally, the validation and comparison of different clustering tools applied to the periodical patterns are proposed to assess and identify the actual building category based on its actual usage.

III. Methodology

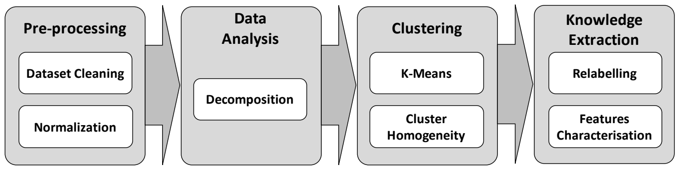

Fig. 1: Methodological pathway for telecommunication sites actual usage assessment and knowledge extraction

This work investigates a data set containing aggregated hourly electric load measurements for the whole year of 2019 regarding around two thousand buildings (or facilities) managed by the largest telecommunication service provider in Italy. The available data include active and reactive electrical load profiles. The vast majority of the buildings in the data set are CO, which are facilities containing telephone switches and other telecommunication hardware employed for the operation of the wired network. Furthermore, a few DC, RBS, and offices buildings are included. Many of these facilities, particularly for what concerns offices and CO, actually have a promiscuous usage. The geographical location, the Climatic Severity Index, the square footage of the areas devoted to offices and telecommunication equipment are included in the data set.

The proposed methodology in Figure 1 provides an automated tool for knowledge extraction from time series of aggregated load profiles of telecommunication facilities. The goal is to detect and fix wrong labels attached to buildings, in the perspective of supporting the facility management of the company. The methodology is presented in Figure 1 which involves four steps, namely ***Pre-processing***, ***Data Analysis***, ***Clustering***, and ***Knowledge Extraction***.

Pre-processing: Firstly, a data set cleaning step is carried out, to delete abnormal measurements and to reduce noise inside time series. Time series values were filtered by replacing values featuring gradients higher than three times the standard deviation calculated from the distribution of gradients from the time series itself. These values, as well as the missing ones, were substituted by linear interpolation. Furthermore, short time series with less than one year of measurements and those containing long lasting abnormal measurements (i.e. whenever 48 consecutive values were considered abnormal) have been filtered out. Hence, the cleaned data set is considered for the continuation of the study.

As the available data represent aggregate electrical demand, it is worth noting that telecommunication sites' energy balance includes 4 fundamental contributes:

$$E_{TOT}(t) = E_{TLC}(t) + E_{DISS}(t) + E_{CLC}(t) + E_{AUX}(t) \quad (1)$$

where E_{TLC} represents the electrical demand from telecommunication equipment, E_{DISS} takes into account the energy conversion losses, E_{CLC} is the contribution from the cooling system and E_{AUX} includes the electrical loads from auxiliaries and lighting systems.

To preserve the meaningfulness of time series and to effectively compare sites, a normalisation method has been proposed which relies on the theoretical background regarding telecommunication sites' energy behaviour [22], [23]. Since E_{TLC} and E_{DISS} may be assumed as constant values, their contribution well represents the time series base-load. Hence, the electrical demand fluctuations are due to the cooling load, lighting and auxiliaries systems. In fact, the ratio between total load by the first two contributions is strongly representative of the buildings' energy behaviour.

Since no occupancy is expected to affect the electrical demand of the sites during nights, and since the weather conditions in our country determine that the cooling system is not necessary during winter days colder hours, the electrical load measured at these time steps is representative of the aforementioned base-load, directly depending exclusively from the telecommunication equipment related consumption. Hence, the most suitable value for normalisation of load profiles is:

$$E_{min} \approx mean(min(E_i)_d),\ i \in [1,4],\ d \in winterdays \quad (2)$$

where E_i is the energy demand from the i_{th} hourly time step from winter day d. It is worth noting that the values obtained through the proposed normalisation method are analogous to the product of two of the most widely used efficiency metrics in the telecommunication branch, namely the Power Efficiency Index (PUE) and Utilization Factor (UF) [22]. In fact, these indexes are defined as:

$$(3) \quad PUE = E_{TOT}/E_{TLC} \qquad UF = E_{TLC}/E_{MIN} \quad (4)$$

Hence: $$E_{TOT}/E_{min} = PUE * UF \quad (5)$$

PUE is generally used to assess the efficiency of DC and CO. One may easily guess that this value will rise as the contribution of the cooling load increases. On the other hand, promiscuous sites and predominant offices buildings are expected to have high PUE values due to the strong contribution of E_{AUX}. UF takes into account the energy conversion efficiency, which is generally constant and depends on the installed conversion devices.

Data Analysis: The contributions in Equation 1 affect the load profile, determining the base-load, a trend and periodic components. Hence, the proposed methodology takes advantage of a decomposition algorithm in the attempt of isolating the fluctuations linked to sites usage from the ones due to exogenous factors. Particularly, electrical demand is strongly affected by occupancy, daily and seasonal external temperature, and radiation fluctuations. These phenomena are expected to depict the characteristic periodicity of the electric load pattern. Specifically, the electrical load contribution due to occupancy is expected to depict a weekly periodicity, while daily and yearly periods characterise temperature and radiation fluctuations. The load profiles periodicity has been obtained through auto-correlation analysis of time series to detect the major periodical components (i.e. daily and weekly). To conclude, the Loess decomposition model has been applied to distinguishing the base-load, the seasonal and the above mentioned periodical components. This model represents an additive decomposition tool, whereas time series are interpreted as follows:

$$E_{TOT} = l + t + p + n \quad (6)$$

where l is the level component, which is constant over the whole time series; t is the trend representing the tendency of the time series to grow or decrease over consecutive periods; p is the periodic component; and n is the residual component, called noise, which is the difference between the sum of the previously mentioned components and the original time series. Since the data set includes measurements from the whole year of 2019, the seasonal component is obtained through the trend component. On the other hand, the weekly periodicity is found by setting a periodical parameter of the Loess decomposition equal to 24 time steps for daily periodicity and 168 time steps for weekly one.

Clustering: The load fluctuations linked to the electrical demand contribution of E_{AUX} and E_{CLC} are the most representative aspect of the time series concerning usage category. In fact, the building humans occupancy modifies the facility electrical consumption pattern generating specific periodic fluctuations. So, the proposed clustering algorithm should consider the time series daily and weekly periodical components identified in the previous step to differentiate between pure and promiscuous facilities. Several clustering algorithms have been tested based on the related works presented in Section II. More in detail, k-Means and k-Shape algorithms, and a density based algorithm, namely DBSCAN, were employed to group different time series by their daily and weekly periodical components. Since the normalised load values are significant from an energy point of view, Euclidean distances were adopted as the distance measure of abovementioned algorithms. The optimal number of clusters was selected in accordance to three clustering validity indexes,

Statistics	μ	$\mathbf{P_{25th}}$	$\mathbf{P_{75th}}$
CO, DC, RBS	1.20	1.15	1.23
Offices	1.46	1.24	1.64

TABLE I: Statistic description of the distribution of mean yearly normalised electrical load, in relation to provided usage category labels

namely Within Cluster Sum of Square Distances (WCSS), Silhouette coefficient and Davies-Bouldin Index (DBI), and to intra-cluster labels homogeneity.

Knowledge Extraction: Once intra-cluster usage category label homogeneity was achieved, a few sites resulted as discordant with respect to their cluster. Therefore, the telecommunication service provider energy managers checked if any change in intended usage or reduction/increase of the staff operating within those sites had occurred and had not been reported. Eventually, sites have been re-labelled in accordance with their actual usage category. Finally, the Knowledge Extraction step identifies the characteristic features from each building usage category. In particular, mean daily load fluctuations, energy behaviour differences within days of the week and the impact of external temperature and radiation have been investigated.

IV. Results

Results from each step included in the proposed methodology are discussed. To handle and analyse the data set, we took advantage of Python programming language exploiting scikit-learn and Keras libraries.

In the pre-processing step, the data set was reduced to $73,22\%$ of the overall amount of sites, whose provided usage categories percentages are: *i)* CO: $96,84\%$; *ii)* DC: $0,08\%$; *iii)* Offices: $2,25\%$; *iv)* RBS: $0,45\%$;*v)* Unlabelled: $0,03\%$; *vi)* Others: $0,08\%$.

Subsequently, time series were normalised following the techniques introduced in Section III. After normalising time series, our expectations are confirmed by comparing the mean of normalised values from different usage categories, representing the product of the two aforementioned metrics on yearly basis. In fact, those sites devoted to the operation of telecommunication devices depict low mean normalised value over the year, while offices results in higher average load, in comparison to base-load, as it is synthesised in Table I. CO, RBS and DC are characterised by a low value of $PUE * UF$ indicator since 75% of their overall amount present values below 1.23 with a mean value of 1.20. Offices average load demand instead is higher with respect to base-load, as most of the sites are over 1.24, with a mean value of 1.46.

Periodical components can be easily detected by means of autocorrelation analysis. Values within time series depend on some extent from previous ones. This type of correlation can be calculated for the whole set of values, and the resulting autocorrelation coefficients can be represented in relation to lags, as shown in Figure 2. In this case, local minima are evident at lags 24, 48 and so on. This pattern indicates the presence of a daily periodic component, as values are strongly correlated to those at the same hour of previous days. Secondly, it may be observed that the value of peaks generally decreases as the lag increases. On the other hand, the seventh and fourteenth peaks overcome other local minima, proving that a value is strongly related to the one from the same hour and the same day of the week from previous weeks. Therefore, we can assess that time series has a weekly periodic component too. Finally, we extracted periodic components from time series taking advantage of Loess decomposition tools. Besides conveying meaningful information, handling the extracted 168-long time series determine important computational time savings, with respect to the original full load profiles.

Each load profile is characterised by different peak hours in relation to the building usage category. The maximum daily electrical demand for pure CO and DC is generally occurring during the afternoon, while many offices buildings show a peak demand between noon and 2 p.m.. This behaviour looks reasonable. In fact, since auxiliaries systems energy demand contributions are not expected to be of importance in pure CO and DC, the cooling load brings the most determinant contribution to the overall electrical load. As a consequence of this, the impact of outdoor temperature, which generally reaches its maximum during the afternoon, is expected to be determinant in what concerns CO and DC. Contrary, occupancy plays a key role for what concerns offices buildings. Peak electrical demand will likely occur in phase with peak occupancy within the buildings. For these reasons, we excluded phase-invariant clustering algorithms and distance measures suitable for matching time series with different phase or period. At the same time, and due to analogous reasons, k-Shape algorithm resulted in less meaningful outcomes, with respect to k-Means. On the other hand, DBSCAN was not able to achieve satisfactory results. This is likely due to the low equity in the density distribution of elements within the

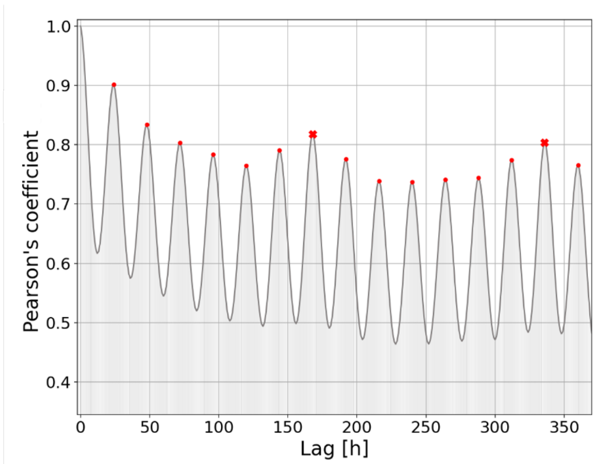

Fig. 2: Analysis of auto-correlation within values from a Central Office electrical load time series

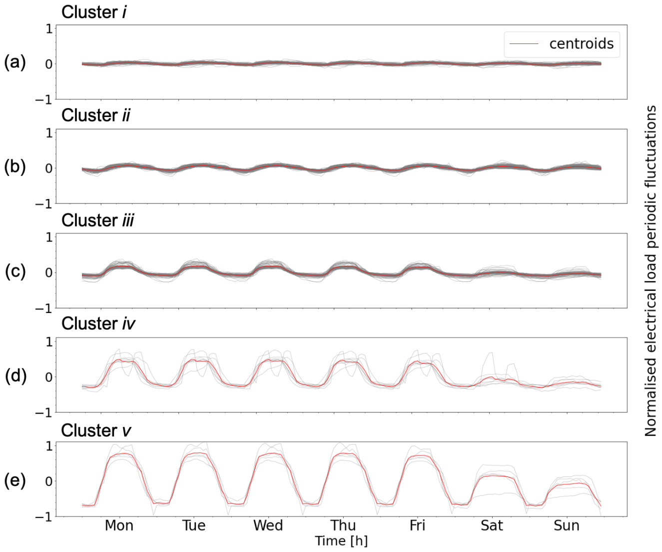

Fig. 3: Clusters resulting from k-means algorithm, setting k=7. Two clusters are not included in the figure, since they solely contain few anomalous sites. The grey lines indicate single sites' weekly periodic components

data set. On this basis, k-Means was adopted as the proper clustering algorithm. In order to achieve optimal intra-cluster homogeneity and inter-cluster separation, k-Means clustering algorithm was performed setting the different number of clusters, in the range 2-10 and considering the values for each clustering validity index. In particular, DBI should be low, Silhouette coefficient high [24], while the so-called Elbow's heuristic is considered for what concerns WCSS. Furthermore, we considered the resulting clusters' homogeneity with respect to the provided usage category labels.

On this basis, we selected seven as the optimal clusters' number, as it represented the most convincing trade-off between the aforementioned clustering accuracy and homogeneity parameters. The five most interesting clusters are reported in Table II. A more detailed investigation over load profiles of the two other clusters of cardinality three arouse suspicions about the reliability of these facilities. Hence, they were filtered out. The remaining five clusters include a number of sites, with a strong prevalence of CO in clusters i and ii (Figure 3-a and 3-b). On the other hand, offices are predominant in clusters iv and v (Figure 3-d and 3-e), while cluster iii features a more heterogeneous composition (Figure 3-c).

Clusters	*i*	*ii*	*iii*	*iv*	*v*
CO, DC and RBS	99.12%	98.66%	83.10%	37.50%	20.00%
Offices	0.50%	1.12%	15.49%	62.50%	80.00%
Unlabelled	0.38%	0.22%	1.41%	0.00%	0.00%

TABLE II: Composition of the most significant clusters in accordance to the provided usage category labels

These groups of sites not only have a strong correlation with the provided usage labels, yet they depict characteristic features. By having a look at Figure 3, one may easily observe important differences in daily fluctuations. Furthermore, weekends load demand reduction is noticeable in the last two plots, which we remark being those with a predominance of offices, hence whose load profiles are expected to depend on human occupancy.

Assuming that most of the sites are correctly labelled, clusters i and ii are considered as representative of pure or predominant CO, RBS or DC, clusters iv and v should include offices and promiscuous sites are expected to be included in cluster iii. A few sites labels are hence detected as abnormal. Particularly, three sites labelled as CO (0.23% of the COs) and one as RBS (16.7% of the RBSs) are included in clusters that represent offices. On the other hand, nine sites with label office were identified in clusters which include a vast majority of CO, RBS and DC. A field verification on these sites reported that the actual usage category of three out of four sites was correctly identified by the clustering algorithm. Similarly, six out of nine of the sites labelled as offices were actually predominant CO. It is worth noting that most of the sites abnormally labelled were building containing both telecommunication equipment devoted areas and offices. Many of these occurred in a change of usage category due to an increase or reduction of employees within the site. Finally, under the reasonable assumption that those sites whose label coincide with the usage category represented by a cluster are correctly labelled, the clustering algorithm achieves an accuracy of 99.7%.

The final step of this study regards the identification of each category's load profiles characteristic features. The feature which better represents the usage category is the amplitude of mean daily fluctuations. These results are synthesised in Figure 4. Offices depict in Figure 4-c significant fluctuations, with noticeable difference within days of the week. More in detail, the 53.8% of offices depict mean working days fluctuations over 100% with respect to E_{min}, and the whole set of sites

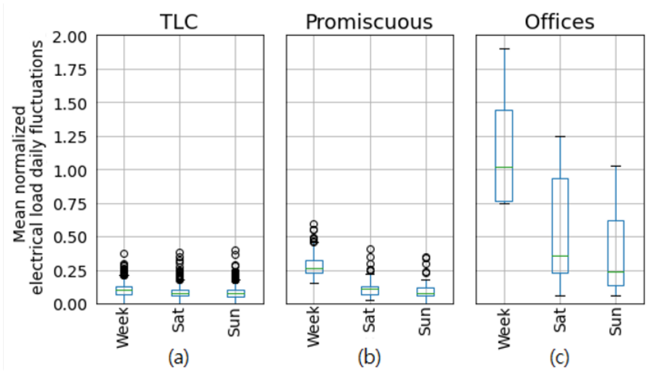

Fig. 4: Distribution of mean normalised electrical load fluctuations over different building usage category and day of the week

have values beyond 70%. These fluctuations are significantly lower during Saturdays and Sundays, whereas 85% of the sites are characterised by values in ranges from 15% to 125% and from 10% and 103% respectively. Differently, promiscuous sites in Figure 4-b do not depict noticeable difference within day types. A slight difference may be observed within weekdays and holidays daily fluctuations. Specifically, almost 90% of promiscuous sites have fluctuations in ranges from 18.8% to 50.6% during working days and from 4.0% to 26.7% during holidays. Finally, mean fluctuations from the vast majority of sites devoted to telecommunication equipment operation (see Figure 4-a) are below 19% with respect to base-load, regardless of the day of the week. It is worth noting that the amplitude of fluctuations from holidays in promiscuous sites is pretty close to those from any day in predominant and pure CO, RBS and DC. These witnesses that, as soon as no occupancy affects electrical demand from the promiscuous site, their energy behaviour is analogous to the one from pure telecommunication sites. Other characteristic aspects related to usage categories are *i)* the electrical load peak hours, *ii)* the correlation between load and outdoor air temperature, and *iii)* the correlation between load and radiation, which are not deepen in this work.

V. Conclusions

In this paper, we discussed the application of some widely used ML techniques to a real and large data set containing the electrical load profiles of telecommunication sites characterised by different usage categories. The present work proposed a methodological path aimed to highlight characteristic features and to assess their correct actual usage, in the perspective of supporting large company facility management. Data normalisation and load profiles decomposition made it possible to effectively apply a partitional clustering algorithm to the data set. The proposed methodology provided accurate identification of building usage categories since 99.7% of the sites were grouped in clusters representative of their actual usage. At the same time, the methodology revealed itself as an effective tool for assessing the actual usage category of the sites detected as abnormally labelled. Finally, it was possible to identify some characteristic features from each category. Specifically, mean daily normalised load fluctuation revealed itself as the most representative feature regarding usage category. Based on the outcomes of this study, a classification model shall be designed in the next future with the perspective of providing more efficient usage-based labelling of telecommunication sites.

References

[1] "Energy efficiency directive — Energy." [Online]. Available: https://ec.europa.eu/energy/topics/energy-efficiency/targets-directive-and-rules/energy-efficiency-directive_en

[2] J. Malinauskaite, H. Jouhara, L. Ahmad, M. Milani, L. Montorsi, and M. Venturelli, "Energy efficiency in industry: Eu and national policies in italy and the uk," *Energy*, vol. 172, pp. 255–269, 2019.

[3] S. Lambert, W. Van Heddeghem, W. Vereecken, B. Lannoo, D. Colle, and M. Pickavet, "Worldwide electricity consumption of communication networks," *Optics express*, vol. 20, no. 26, pp. B513–B524, 2012.

[4] M. Avgerinou, P. Bertoldi, and L. Castellazzi, "Trends in data centre energy consumption under the european code of conduct for data centre energy efficiency," *Energies*, vol. 10, no. 10, p. 1470, 2017.

[5] "Data Centres and Data Transmission Networks – Analysis - IEA." [Online]. Available: https://www.iea.org/reports/data-centres-and-data-transmission-networks

[6] "Pubblicazioni Statistiche - Terna spa." [Online]. Available: https://www.terna.it/it/sistema-elettrico/statistiche/pubblicazioni-statistiche

[7] H. M. Daraghmeh and C. C. Wang, "A review of current status of free cooling in datacenters," *Applied Thermal Engineering*, vol. 114, pp. 1224–1239, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.applthermaleng.2016.10.093

[8] "The case for committing to greener telecom networks — McKinsey." [Online]. Available: https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-case-for-committing-to-greener-telecom-networks

[9] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, 2017.

[10] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, "Time-series clustering - A decade review," *Information Systems*, vol. 53, pp. 16–38, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.is.2015.04.007

[11] X. Huang, Y. Ye, L. Xiong, R. Y. Lau, N. Jiang, and S. Wang, "Time series k-means: A new k-means type smooth subspace clustering for time series data," *Information Sciences*, vol. 367, pp. 1–13, 2016.

[12] R. Damayanti, A. G. Abdullah, W. Purnama, and A. B. Nandiyanto, "Electrical load profile analysis using clustering techniques," in *IOP Conference Series: Materials Science and Engineering*, vol. 180, no. 1. Institute of Physics Publishing, 2017, p. 012081. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2017MS&E..180a2081D/abstract

[13] J. Yang, C. Ning, C. Deb, F. Zhang, D. Cheong, S. E. Lee, C. Sekhar, and K. W. Tham, "k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement," *Energy and Buildings*, vol. 146, pp. 27–37, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.enbuild.2017.03.071

[14] J. Paparrizos and L. Gravano, "K-shape: Efficient and accurate clustering of time series," in *In Proc. of SIGMOD/PODS'15*, vol. 2015-May. Association for Computing Machinery, may 2015, pp. 1855–1870.

[15] G. E. Batista, E. J. Keogh, O. M. Tataw, and V. M. De Souza, "Cid: an efficient complexity-invariant distance for time series," *Data Mining and Knowledge Discovery*, vol. 28, no. 3, pp. 634–669, 2014.

[16] S. Zolhavarieh, S. Aghabozorgi, and Y. W. Teh, "A review of subsequence time series clustering," 2014.

[17] O. Motlagh, A. Berry, and L. O'Neil, "Clustering of residential electricity customers using load time series," *Applied Energy*, vol. 237, pp. 11–24, mar 2019.

[18] S. Ryu, H. Choi, H. Lee, H. Kim, and V. W. Wong, "Residential Load Profile Clustering via Deep Convolutional Autoencoder," in *In Proc. of SmartGridComm 2018*. Institute of Electrical and Electronics Engineers Inc., dec 2018.

[19] A. K. Zarabie, S. Lashkarbolooki, S. Das, K. Jhala, and A. Pahwa, "Load profile based electricity consumer clustering using affinity propagation," in *In Proc. of EIT 2019*. IEEE, 2019, pp. 474–478.

[20] S. Sun, S. Wang, G. Zhang, and J. Zheng, "A decomposition-clustering-ensemble learning approach for solar radiation forecasting," *Solar Energy*, vol. 163, pp. 189–199, 2018.

[21] F. Giordano, M. L. Rocca, and M. L. Parrella, "Clustering complex time-series databases by using periodic components," *Statistical Analysis and Data Mining: the ASA Data Science Journal*, vol. 10, no. 2, pp. 89–106, 2017.

[22] M. Sorrentino, M. Bruno, A. Trifirò, and G. Rizzo, "An innovative energy efficiency metric for data analytics and diagnostics in telecommunication applications," *Applied Energy*, vol. 242, no. March, pp. 1539–1548, 2019. [Online]. Available: https://doi.org/10.1016/j.apenergy.2019.03.173

[23] F. Genova, M. Gaspardone, A. Cuda, M. Beoni, G. Fici, and M. Sorrentino, "Optimal energy management of tlc plants using low cost wireless sensor network based monitoring system and a simulation model for data analysis," in *In Proc. of INTELEC 2009*. IEEE, 2009, pp. 1–6.

[24] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, no. 1, pp. 243–256, 2013.