applied sciences

MDPI

Article

Identification of Typical and Anomalous Patterns in Electricity Consumption

José Nuno Fidalgo [1,2,*] and Pedro Macedo [1]

[1] Institute of Systems and Computer Engineering, Technology and Science (INESC TEC), Campus da Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal; pedro.m.macedo@inesctec.pt

[2] Faculty of Engineering, Porto University, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

* Correspondence: jfidalgo@fe.up.pt; Tel.: +351-222094230

Abstract: Nontechnical losses in electricity distribution networks are often associated with a countries' socioeconomic situation. Although the amount of global losses is usually known, the separation between technical and commercial (nontechnical) losses will remain one of the main challenges for DSO until smart grids become fully implemented and operational. The most common origins of commercial losses are energy theft and deliberate or accidental failures of energy measuring equipment. In any case, the consequences can be regarded as consumption anomalies. The work described in this paper aims to answer a request from a DSO, for the development of tools to detect consumption anomalies at end-customer facilities (HV, MV and LV), invoking two types of assessment. The first consists of the identification of typical patterns in the set of consumption profiles of a given group or zone and the detection of atypical consumers (outliers) within it. The second assessment involves the exploration of the load diagram evolution of each specific consumer to detect changes in the consumption pattern that could represent situations of probable irregularities. After a representative period, typically 12 months, these assessments are repeated, and the results are compared to the initial ones. The eventual changes in the typical classes or consumption scales are used to build a classifier indicating the risk of anomaly.

Keywords: typical patterns; nontechnical losses; anomaly detection; energy theft; clustering; data mining

check for updates

Citation: Fidalgo, J.N.; Macedo, P. Identification of Typical and Anomalous Patterns in Electricity Consumption. *Appl. Sci.* **2022**, *12*, 3317. https://doi.org/10.3390/app12073317

Academic Editor: Hartmann Balint

Received: 28 January 2022
Accepted: 21 March 2022
Published: 24 March 2022

1. Introduction

The advent of smart grids and smart equipment in electricity distribution networks is expected to cause a substantial increase in the volume of available data, which will amplify the potential of data-driven knowledge extraction applications. New approaches such as the identification of archetypal consumption patterns and also the detection of the most abnormal ones, some of which might indicate illegal behavior (e.g., energy theft) start to be feasible under this new smart grid environment. The present article illustrates the exploitation of this new opportunity in terms of the characterization of consumers and the detection of nontechnical losses.

Consumer characterization has always been an important part of the electricity distribution business, increasingly so with the advent of smart grids. New energy resource such as distributed energy storage or load flexibility are essential components for the decarbonization of electricity systems. From this perspective, determining load profiles will be crucial to making the most of the new energy resources [1–3].

Energy losses in the distribution grid are usually classified into technical losses and commercial or nontechnical losses. The total annual losses are often assessed through the total energy balance: the sum of all the energy entered into the grid subtracted by all billed energy. Therefore, while global losses are easily calculated, their division into technical and nontechnical losses is often problematic [4–8]. The technical losses depend on the physical characteristics of the grid components and energy flows. The major component of these losses is related to the joule effect in conductors [6,8–11]. Commercial losses, however, are related to nonbilled energy, namely because of metering errors or energy theft. In Portugal, the cost of nontechnical losses is estimated to be over EUR 100 M per year.

The knowledge of consumer characteristics and behavior is also often used for nontechnical loss detection [12–15]. Sometimes, such as in [12], the analysis is based on hourly (or 15 min) data, usually to obtain average daily diagrams. In other studies [15], authors consider high-level features such as the average consumption and the maximum consumption in the last six months. The latter case was the most common before the dissemination of interval meters (in the scope of the transition to the smart grid paradigm), a process that began roughly 10 years ago in several European countries. In fact, these new meters can provide detailed information, far beyond the classical monthly energy consumption. The article [16] is an example of theft detection based on the new opportunities created by smart meter implementation. An alternative hardware-based approach for fraud detection is presented in [17]. In this case, the utility installed dedicated ammeters, upstream and downstream of the meter to check for differences. Customized hardware systems are quite expensive; thus, they are not feasible global solutions.

This paper describes the methodologies adopted for detecting anomalies in the energy consumption of high, medium and low voltage installations. This study involves two types of strategies:

1. Identification of patterns, consumption profiles and atypical consumers (outliers).
2. Analysis of the evolution of the energy consumption of each specific installation to detect changes in the consumption patterns that configure situations of potentially illicit behavior.

The innovation of the present article lies in the originality of the methodology. First, the consumption time series provided for this study has a daily time base, i.e., each item of the sequence is the consumption of a given day, for a given consumer. Consequently, the typical consumption patterns and atypical behaviors are based on daily consumption, or aggregation of this data (e.g., typical consumption distribution throughout the week) but never on intraday distributions (e.g., peak hours). Naturally, the hourly load distribution could provide a supplementary characterization of consumer behavior. However, that level of detail was not available. The most common approaches are in both extremes: either they use an hourly (or 15 min) base or a monthly base. Of course, this becomes a tradeoff between the level of information detail and data processing requirements. Another interesting and useful characteristic of the proposed methodology concerns the type of typical patterns to be identified: by weekday, month and logarithmic scale. Moreover, this is an integrated approach that provides both the identification of the most typical profiles and the most anomalous ones at the same time. Finally, to the best of our knowledge, the detection of anomalies based on a twofold analysis (distance to the classes of prototypes and behavioral changes) is also a novel proposal.

2. Methodology

The adopted approach is comprised of the following main steps:

1. Exploratory data analysis. The first goal is to acquire a general perspective of data distributions, according to the type of installation (industry, residential, etc.), voltage level (high, medium or low), geographical zone (region or district), type of energy contract and eventual changes in these contracts. Another important feature highlighted by this initial step is the identification of measurement gaps, i.e., occasional issues in the meter or in the communication system that cause an eventual lack of consumption records in certain intervals. In the case of a substantial number of gaps (larger than a given prespecified threshold) for a given meter, the respective time series is signaled as improper for analysis, and the DSO is informed about the need to perform a check-up of this meter.
2. Identification of typical consumption patterns.

3. Detection of changes in the consumption patterns and/or scale as a potential anomaly symptom.

2.1. Exploratory Data Analysis

This section describes the initial phase of data analysis and preprocessing, both concerning technical-commercial characteristics and daily consumption diagrams (time series). This allows a familiarization with the data of the facilities and the identification of potential inconsistencies.

The data were provided by the DSO in ".txt" format and hosted in a SQL Database (DB) "MySQL". The exploratory analysis was performed in R [18], which includes features for communication with the DB, allowing the extraction of the necessary data for further analysis and information treatment.

The exploratory analysis was developed with the following main purposes:

1. Data extraction and restructuring—This step was comprised of data withdrawal from different sources and reorganization of all the necessary information into a single but comprehensible DB. This process also permitted familiarization with the data.
2. Identification of gaps in the data, allowing for the identification of a lack of records, in both interday and intraday levels.

2.1.1. Data Organization

The information was structured across 8 tables: Installations, Contracts, Powers, Tariffs, Equipment, Anomalies, Nonanomalies and the daily consumption time series for each installation. A simplified schematic of the relational BD can be seen in Figure 1.

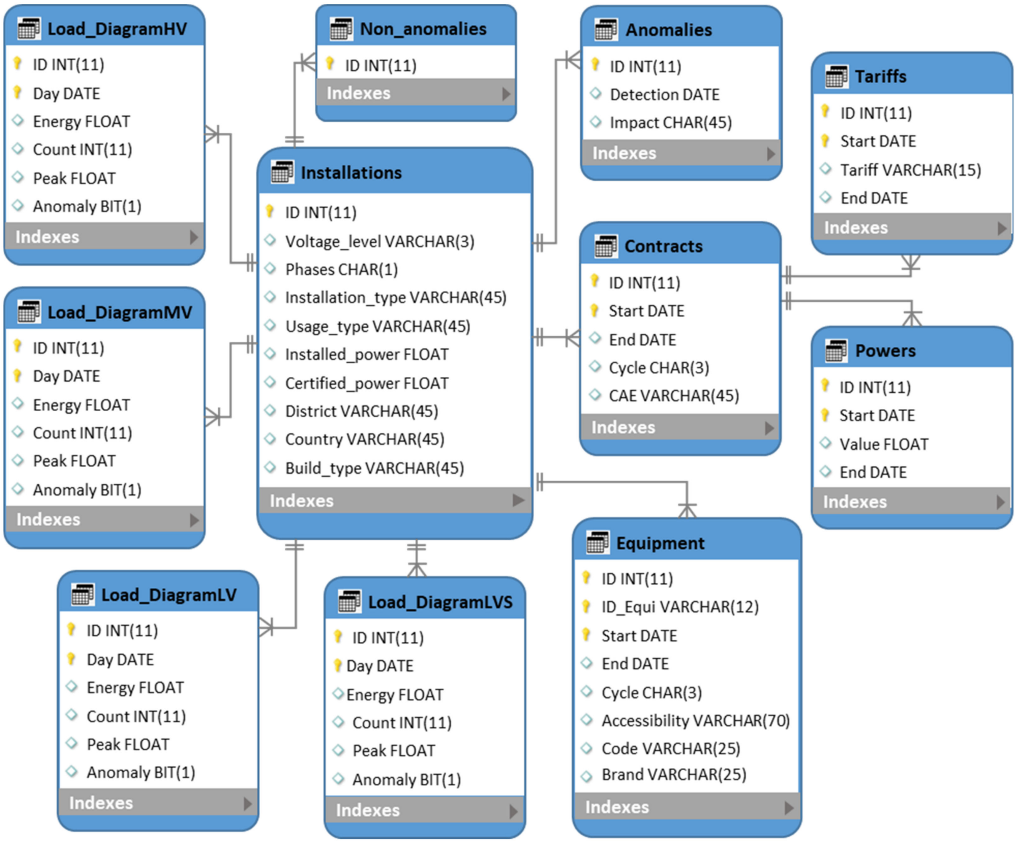


Figure 1. Structure of the DB.

A full description of the DB is outside the scope of this article. However, for a matter of illustration, the following figures are provided.

Figure 2 shows the distribution of consumers who changed their tariff during the period under analysis. In suspicious cases, the installation meter was inspected. If the meter was found to be not in conformity with the standards or if the seal was broken, the case was tagged as abnormal. Figure 3 shows the statistics of these initiatives recorded in the DB in 2017. This information is complemented with another tag that indicates whether or not the anomaly was considered to have impacted the consumption. This type of analysis contributes toward a deeper characterization of the system. For example, it was found that from the total identified anomalies, only 20% had an impact on the measured energy.

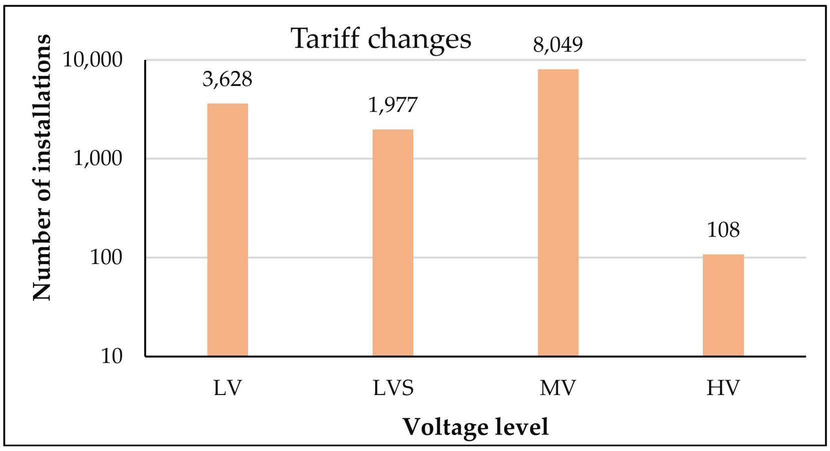


Figure 2. Distribution of consumers who change their tariff contract.

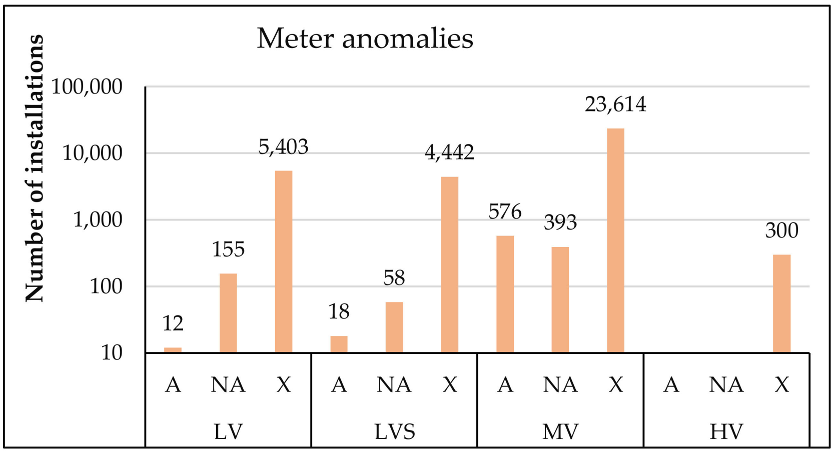


Figure 3. Meter anomalies (A—Abnormal; NA—Not-abnormal; X—Not tested.

2.1.2. Identification of Data Gaps

The identification of gaps in the records is a crucial step toward assessing the consistency and reliability of the available data. In this research, two types of data gaps were considered: interday and intraday.

Interday gaps refer to missing records in the daily time series between the start and the end dates of the analyzed period. Figure 4 shows a histogram of interday gaps in the consumption series: occurrence frequency versus the size of the gap.

This histogram shows that most installations have gaps of less than 50 days. However, a few installations have larger gaps, representing a substantial percentage of the whole time series.

The identification of these gaps is important not only to recognize registration flaws but also as a complementary piece of information in the processes of detection and identification of anomalies (e.g., when the meter is disconnected from the system with the purpose of meter tampering).

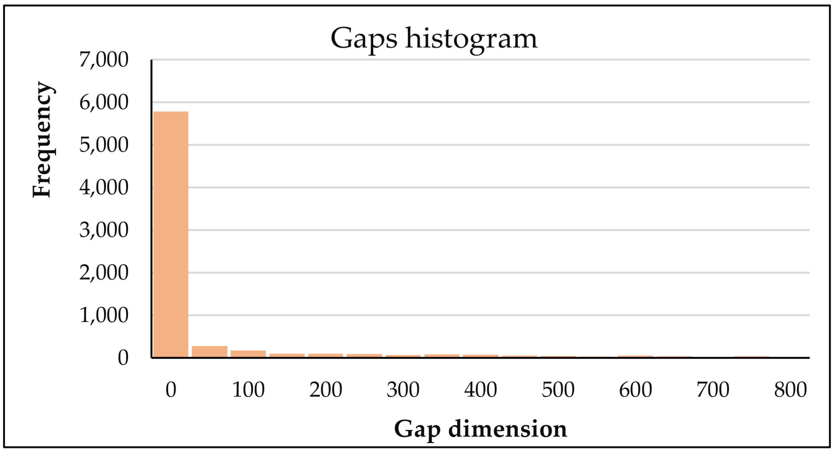


Figure 4. Histogram of interday gaps in the consumption series.

Intraday failures occur when the number of measurements throughout the day differs from the expected number for that date. The total number of 15-min measurements in one day should be 96, except for the daylight-saving time transition days. The histogram in Figure 5 shows the frequency of days with intraday record failures.

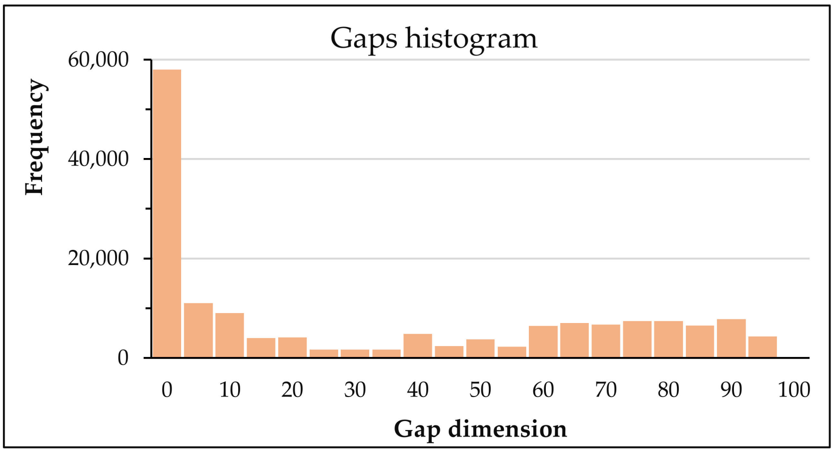


Figure 5. Histogram of intraday gaps in the consumption series.

Most of the problematic cases have less than 5 days of intraday lacks. The identification of these cases and their preprocessing is essential, as they directly affect the energy value since the daily energy consumption is derived from the sum of energy measured in these intervals.

2.2. Identification of Typical Patterns

The search for abnormal consumption patterns assumes that it is possible to identify a set of patterns considered typical. The first approach to anomaly detection consists of the identification of installations with atypical consumption patterns, which refers to installations with consumption characteristics that differ from the typical prototypes in their class.

In this work, clustering algorithms were used to group consumers according to their consumption characteristics (specified in the next section), to identify the prototypes (typical behaviors) and the installations that deviate most from those prototypes (outliers).

2.2.1. Feature Engineering

One of the most important phases of data-driven approaches is the identification of a set of relevant features able to characterize the system under analysis and, at the same time, provide meaningful inputs for the clustering algorithm and other datamining tools.

The adopted strategy involves a large volume of preliminary tests to identify the most suitable set of variables to be used as inputs of the clustering algorithms. As each consumer time series is a sequence of daily energy consumption, several transformations of these data were considered as potential features for consumers characterization, such as weekly feature (percentage of total consumptions that occur on Sundays, Mondays, etc.) or monthly feature (percentage of total consumptions that occur at January, February, etc.). A clustering algorithm is then applied to each feature, providing a classification tool for each consumer from different perspectives.

2.2.2. Consumers Representation

Three types of consumption aggregation were considered during this phase. The outcome of this procedure is a set of three vector features to characterize each consumer:

1. P_week is a vector with seven elements, with the percentage of the consumption that occurs on each weekday.
2. P_month is a vector with 12 elements, with the percentage of the consumption in each month of the year.
3. Log_E is the logarithm of the annual consumption. As the range of consumption scales is enormous, the logarithm attenuates that difference, making the clustering output more interesting. In the same way, the installations were previously split according to voltage levels.

Tables 1 and 2 illustrate the P_week and P_month vectors. As an example, the first consumer in Table 1 has a larger consumption on weekends (Saturday and Sunday with 18.9% and 20.3%, respectively). The same type of analysis can be made for P_month and Log_E, providing the DSO with some significant insights about consumption distributions.

Table 1. Examples of P_week instances.

Mon.	**Tue.**	**Wed.**	**Thu.**	**Fri.**	**Sat.**	**Sun.**
0.123	0.119	0.118	0.121	0.127	0.189	0.203
0.143	0.144	0.144	0.144	0.144	0.142	0.139
0.159	0.166	0.168	0.165	0.159	0.103	0.079
…	…	…	…	…	…	…

Table 2. Examples of P_month instances.

Jan.	**Feb.**	**Mar.**	**Apr.**	**May**	**Jun.**	**Jul.**	**Aug.**	**Sep.**	**Oct.**	**Nov.**	**Dec.**
0.09	0.08	0.08	0.08	0.08	0.09	0.08	0.08	0.09	0.08	0.08	0.08
0.02	0.01	0.03	0.06	0.08	0.12	0.22	0.22	0.14	0.06	0.02	0.03
…	…	…	…	…	…	…	…	…	…	…	…

Afterwards, the vectors are subjected to clustering, integrating a process that is summarized in the next section.

2.3. Detection of Primary Atypical Consumption (Outliers)

Initially, a set of filters was applied to the data with the purpose of filtering the time series with low data quality and categorizing the most discernible abnormal cases. Within this phase the following abnormality filters were implemented:

a. Data gaps—permits the identification of the lack of quality in the analyzed series. If a consumer data series has too many gaps, the consumer will be labelled as "Large gaps" and the classification for these cases stops here.
b. Low or no consumption—identifies installations with very low or no consumption. If the consumption is N times smaller than the first quartile limit, the consumer is classified as "Low consumption". This analysis and the quartiles set are independently attained for each voltage level. N is specified by the user.
c. Low-frequency usage—identifies the consumers that have a large percentage of very low consumption (below a minimum threshold). This minimum is defined individually for each installation and is based on the average consumption. The user is allowed to modify these parameters. The installations identified by the defined filters in points "b" and "c" are labelled "low or no consumption" and "low-frequency usage" respectively.
d. Concentrated consumption (week)—identifies the installations with a considerable percentage of the total consumption on a specific day of the week. If the consumption on Wednesdays is 50% of the week, then the consumer is classified as "CC_w". The same applies to other weekdays or other percentages (specified by the user). For example, it was found that some consumers have approximately 60% of their consumption on Tuesdays. This set of consumers is immediately classified as CC_w and extracted from the set of consumers that will be subjected to weekly consumption clustering.
e. Concentrated consumption (month)—identifies the installations with a large part of the consumption in a specific month of the year. For example, if the consumption in September is 40% of the whole year, then the consumer is classified as "CC_m" and extracted from the set of consumers that will be subjected to monthly consumption clustering. The same applies to other months or other percentages (specified by the user).

These cases represent certain aspects of abnormalities. The consumers with these tags are filtered and do not pass to the following phase. In this way, the clustering algorithm will be applied to a more homogenous set which makes the clustering much more effective.

2.4. Clustering

The remaining consumers are then moved to the clustering phase.

a. A clustering algorithm is applied to each feature (P_week, P_month and Log_E).
b. The output of the previous step is a classification tool to label each consumer according to the weekly consumption distribution (P_week), monthly consumption distribution (P_month) and total annual consumption (Log_E).
c. After new consumption data is collected during some months, this process is repeated, aiming at the detection of eventual alterations in these patterns. In general, class changes are considered an indicator of abnormality.

2.5. Detection of Scale Changes

A complementary analysis is applied to the consumption evolution to detect abnormal consumption patterns. For example, the DSO considers a consumer who has shown historically consistent high consumption, which suddenly drops by 70% suspicious. For this test, a reference date (Dref, in Figure 6) is set to divide the time series into two subperiods: DS0 and DS1. Then, three comparison tests are considered:

a. Compare DS0 to DS1.
b. Compare DS0 to DS0 + DS1.
c. Compare A to B.

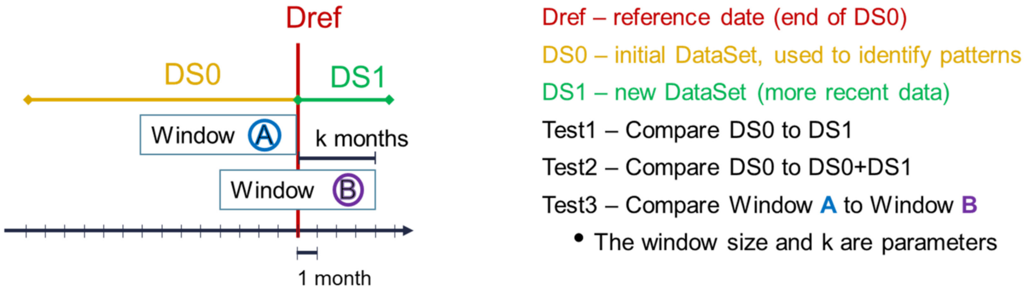

Dref
DS0
DS1
Window A
k months
Window B
1 month
Dref – reference date (end of DS0)
DS0 – initial DataSet, used to identify patterns
DS1 – new DataSet (more recent data)
Test1 – Compare DS0 to DS1
Test2 – Compare DS0 to DS0+DS1
Test3 – Compare Window A to Window B
• The window size and k are parameters

Figure 6. Subdivision of the consumption time series to perform the test scale changes.

2.6. Anomaly Detection and Ranking

When this procedure is concluded the consumer record is complemented with the following tags:

- Scale variation type (increase/decrease).
- Scale variation intensity.
- Change in week pattern (class).
- Change in month pattern (class).
- Distance between weeks shapes.
- Distance between months shapes.

These tags are considered potential indicators of some type of anomaly. For instance, a large-scale reduction in consumption might be an indicator of energy theft. On the one hand, this kind of circumstance is considered to have a high impact on the final anomaly score, while on the other, a small change in week or month shape has a smaller contribution to the final anomaly score.

The last phase consists of establishing a ranking of the consumers according to the anomaly degree. According to the DSO requirements, in the default implementation, a user-defined factor is associated with each tag, and the ranking is the sum-product of factors by tags. A final feature is used to build the final ranking: the data completeness. If the consumer record has no data gaps, the anomaly classification is assumed to be more reliable.

3. Results

This section presents the results according to the methodological sequence. In the following figures, a linear transformation was applied to the consumption time series for a matter of confidentiality, as requested by the DSO. The periods marked with yellow lines indicate missing values (gaps) in the consumer time series.

3.1. Primary Abnormality Filters (Outliers)

The following figures show an example of consumption time series for each primary filter (Figures 7 and 8).

Data Gaps

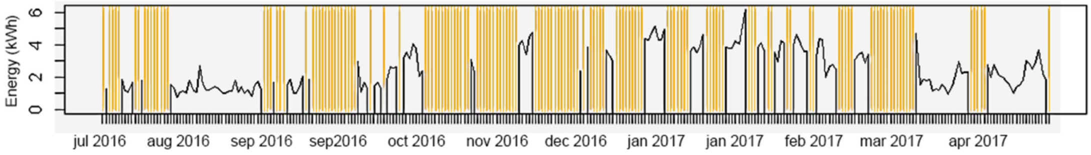


Figure 7. Example of installations labelled “Data Gaps”.

Low consumption & Low usage

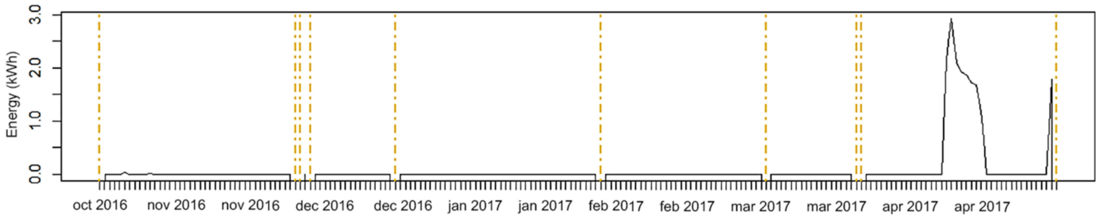
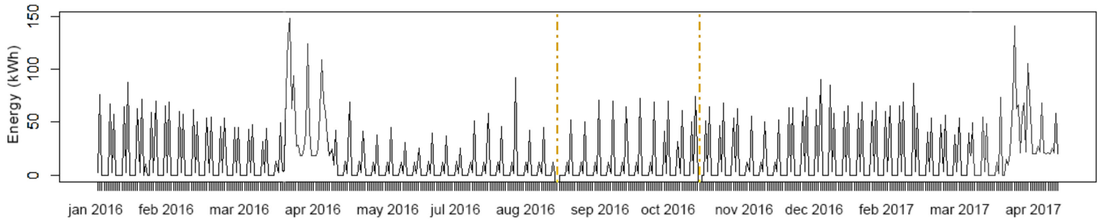
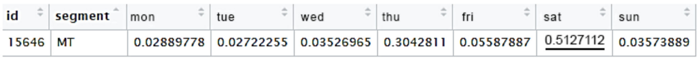


Figure 8. Example of installations labelled "Low consumption" and "Low usage".

Concentrated consumption (week)

Figures 9 and 10 show examples of Concentrated Consumption: more than half of the total consumption occurs on Saturdays and during November, respectively.

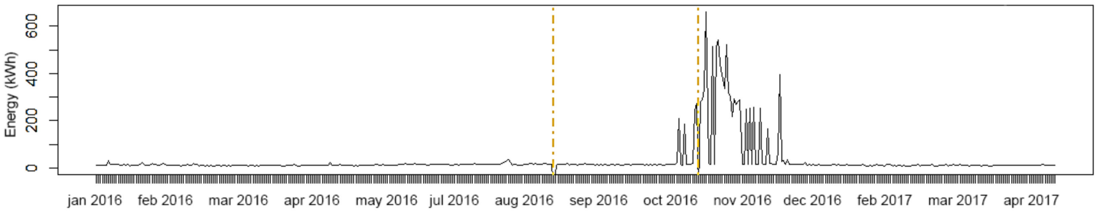


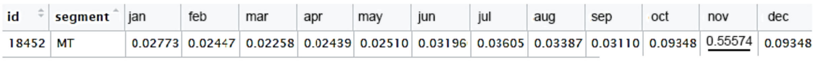


id	segment	mon	tue	wed	thu	fri	sat	sun
15646	MT	0.02889778	0.02722255	0.03526965	0.3042811	0.05587887	0.5127112	0.03573889

Figure 9. Example of installations labelled "Concentrated consumption (week)".

Concentrated consumption (month)





id	segment	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec
18452	MT	0.02773	0.02447	0.02258	0.02439	0.02510	0.03196	0.03605	0.03387	0.03110	0.09348	0.55574	0.09348

Figure 10. Example of installations labelled "Concentrated consumption (month)".

3.2. Clustering

In this work, the elected clustering tool was the Self-Organizing Maps (SOM) [19–25] followed by k-means [26–29]. These studies were developed on WEKA [30,31] and KNIME [32].

SOM performs a projection of the multidimensional space of variables into a two-dimensional map, where similar patterns are close to each other. The separation of classes and determination of an adequate number of clusters is made with the support of another clustering algorithm, k-means, which is fed with the outputs of SOM.

Several heuristics can be used to determine the "ideal" number of clusters. The most frequently used elbow method is based on the observation of the clustering performance curve as a function of the number of clusters, that is, to the extent that an increase in the number of clusters (k) can help in reducing the sum of variances "within the cluster"—var(k). Technically, given a k > 0, k groups of the dataset in question can be formed. Using k-means, the sum of variations "within the cluster" var(k) is calculated and a WSCC (Within Cluster Sum of Squares) var(k) curve is constructed according to the number of clusters k. The WSCC for the two cases of Figure 11 are shown in Figure 12.

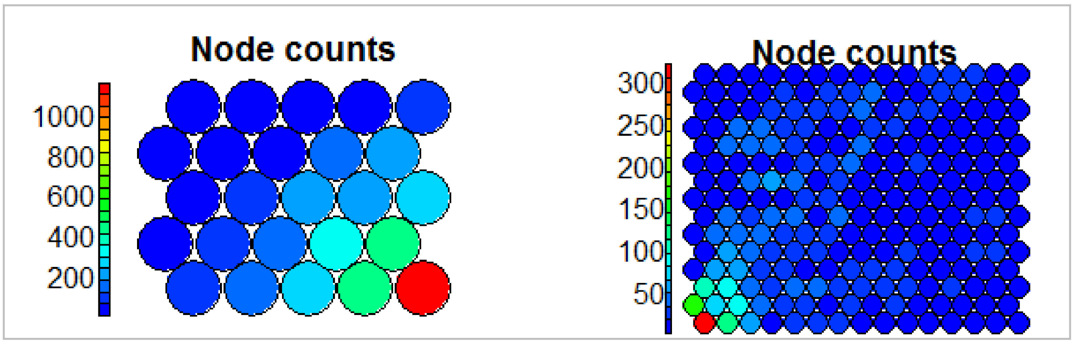


Figure 11. Clustering considering a 5 × 5 and 15 × 15 alternative maps.

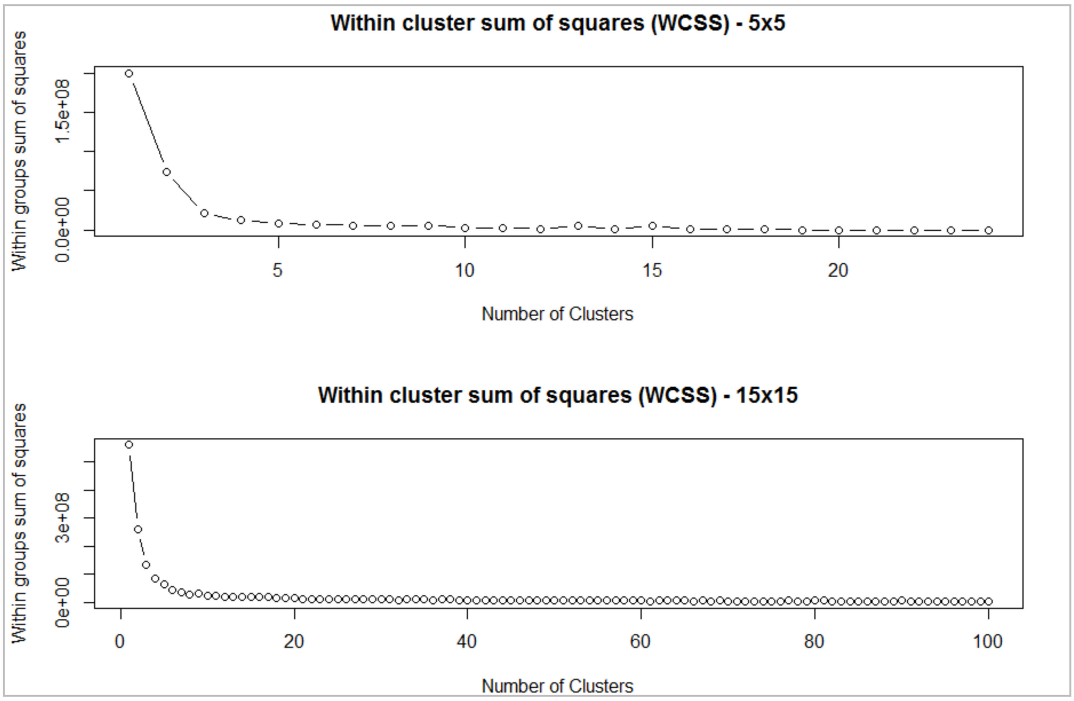


Figure 12. WCSS obtained for the maps 5 × 5 and 15 × 15.

Figure 13 shows the graphical result of the first part of the clustering algorithm for P_week. Each represents a subprototype of a set of consumers with similar weekly consumption distributions.

The final step of the clustering phase consists of applying a hierarchical clustering (k-means) to group the SOM nodes into N classes (specified by the user) as illustrated in Figure 14: defining three classes on the left side and five classes on the right side.

The clustering studies lead to four classes to define the prototypes of consumption evolution throughout the week and the months of the year. A total of five classes were stipulated to define the scale/amplitude of consumption prototypes.

Figure 15 shows that more than half of the consumers have a regular consumption throughout the week (Class 3). Class 4 represents the consumers with higher electricity consumption during the weekends, and Classes 1 and 2 show the consumers with lower consumption during the weekends. The pie chart on the right side shows that the distribution of consumers in Classes 1, 2 and 4 is rather balanced.

Conversely, in the month patterns (Figure 16), the number of consumers in each class is uneven. Most consumers belong to Classes 2 and 3, which is characterized by regular monthly consumption throughout the year, with a slight drop of consumption during the summer and a slight rise during the winter, respectively. Class 1 represents winter consumers. Finally, Class 4 is for summer consumers showing the highest variation throughout the months of the year with a large peak in July and August, most probably due to seasonal factors or activities mainly developed during the summer season (e.g., tourism).

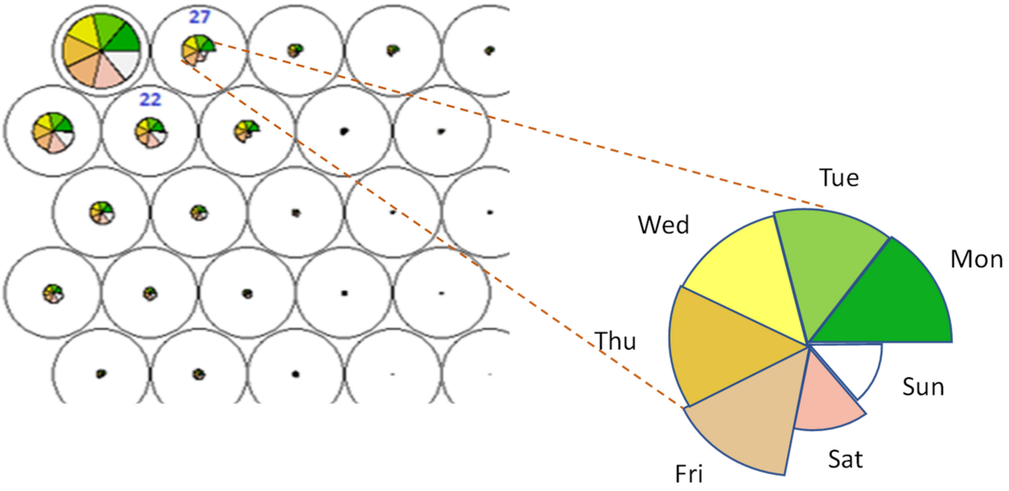


Figure 13. SOM projection for P_week.

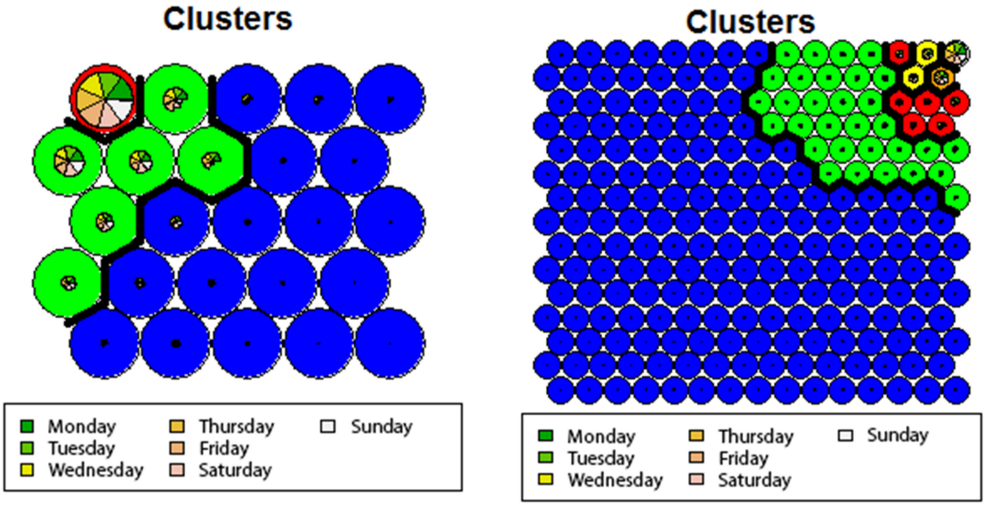


Figure 14. Gathering similar nodes into classes: cases 5 × 5 and 15 × 15).

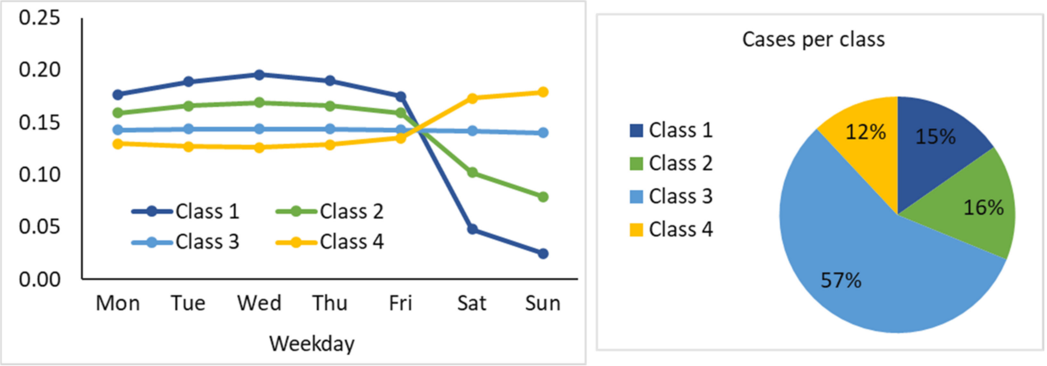


Figure 15. Four prototypes for week patterns defined by a range of 39,126 available installations.

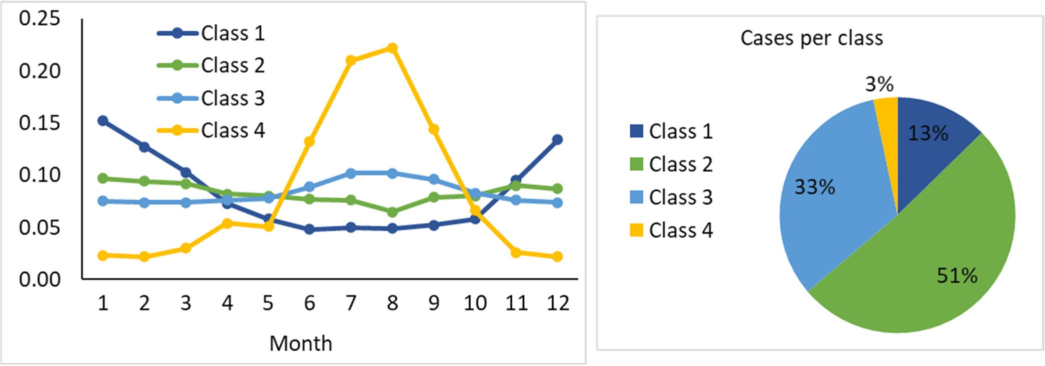


Figure 16. Four prototypes for month patterns defined by a range of 33,858 available installations.

One final clustering exercise was applied to the annual energy consumption (Figure 17). In this case, the cluster prototypes are presented on a logarithmic scale, due to the large differences among the classes' prototype magnitudes, even within the same voltage level. In this example, Class 5 consumption is roughly 300 times higher than Class 1.

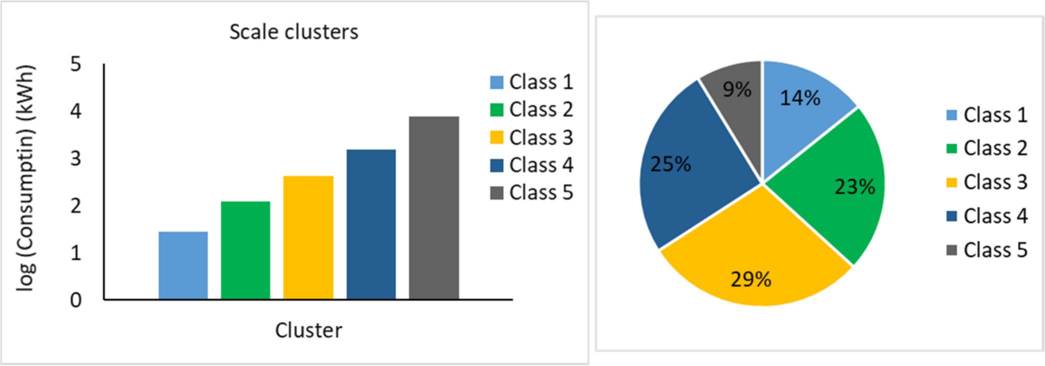


Figure 17. The scale classes for annual energy consumption defined by a range of 22,296 available MV installations.

The differences in the total available installations for each type of clustering is due to the previous filter for the concentrated consumptions. That is to say, more installations with concentrated consumption (month) were filtered.

3.3. Anomalies Detection

The previous section describes how a consumer is characterized according to different perspectives (classes of weekday and month distribution and consumption scale). These classes are only applied to the consumers who do not fall into the categories identified in the initial stage (see Section 2.3).

After this phase, each consumer is labelled with a set of tags that identify their classes. This is completed using the past consumption evolution during a given prespecified period (ideally, at least 12 months). The same classification procedure is repeated with a new dataset of measurements, in general with the most recently acquired data. It is assumed that a change in any class is a potential symptom of anomaly. These changes are then assigned to the consumer record.

Similarly, the tests of scale changes (Figure 6) contribute to the characterization of eventual changes in the "normal" consumption pattern. These tests aim to detect changes in average consumption as well as in the evolution of the consumption trend.

In summary, three tests are considered:

- Test1—to detect changes in the weekday consumption patterns and test changes in the average consumption.

- Test2—to detect changes in weekly, monthly and long-term trend consumption evolution.
- Test3—to detect changes in weekly and monthly consumption evolution.

The result of these tests is appended to the consumer record as illustrated in Figure 18.

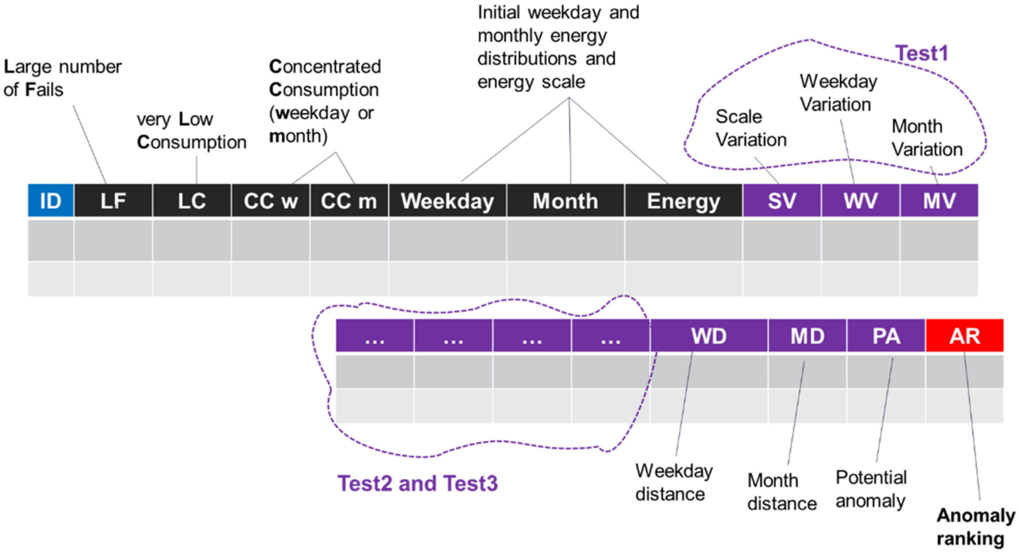


Figure 18. Building anomaly potential and ranking.

Finally, a potential anomaly score is built on the base of all tags.

The following figures illustrate some cases of anomalies' detection. Figure 19 displays examples of cases tagged as anomalous because of large scale variations. Figure 20 includes some examples of consumers that present considerable changes in the weekday and/or month patterns.

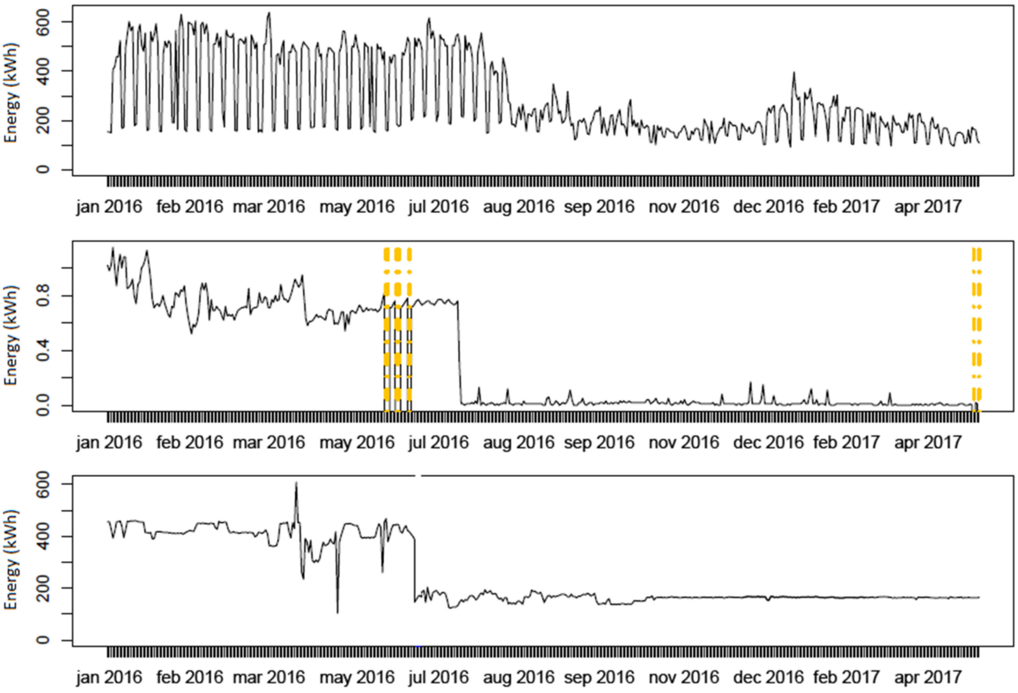


Figure 19. Three examples of anomalous detection (large scale variation).

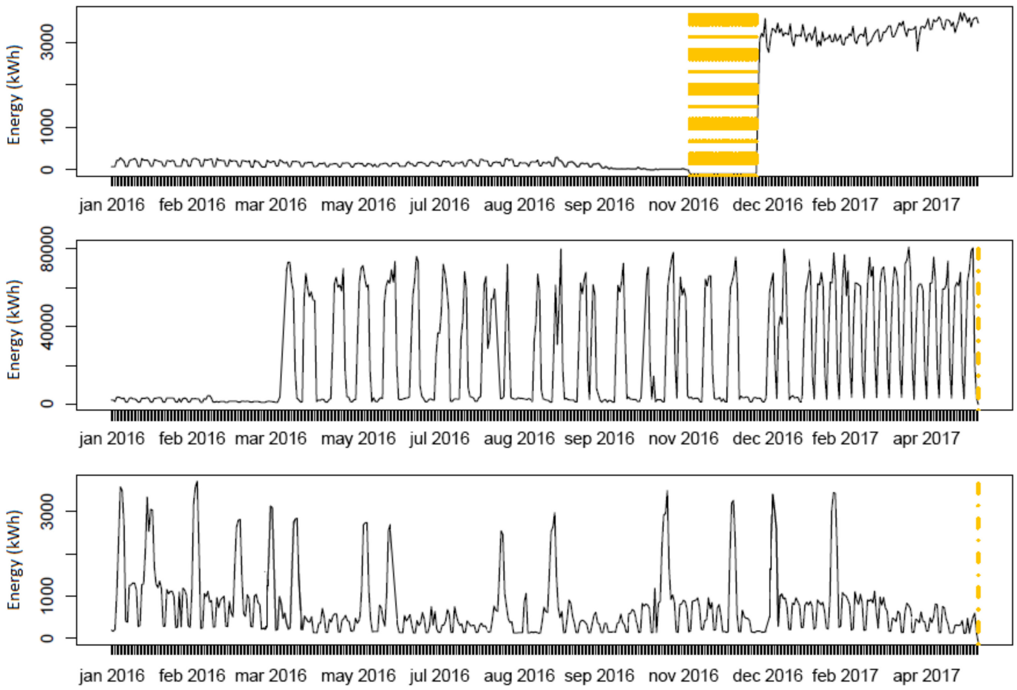


Figure 20. Three examples of anomalous detection (large changes on weekday/month patterns).

4. Conclusions

The developed approach showed very interesting and useful results. First, it provided a clear view of the consumers' most typical behaviors (week and month patterns) and consumption scale distribution. Second, it provided a classification of consumption anomalies according to some predefined settings, which can be changed by the user. The abnormal symptoms are detected, in the first stage, by a simple filtering procedure (e.g., a large percentage of the total consumption in a single month). In a second stage, a set of tests is performed to detect changes in week or month patterns, as well as in the consumption scale.

The analysis of the results confirmed that the implemented tool can effectively detect anomalous consumption cases, as it was projected to do. Several types of anomalies were discovered and characterized, leading to a ranking of suspicious cases, which will be later analyzed in detail by the DSO, including the examination of the meter installation.

Naturally, other types of anomalies cannot be perceived by the current approach, such as, if the meter was improperly installed or altered in the moment of its commissioning. In this case, the resulting consumption diagram would be consistently below what it should be, but with no changes in shape or scale of the diagram.

The typical consumption patterns and atypical behaviors were based on daily consumption, or aggregation of this data (e.g., typical consumption distribution throughout the week), but never on intraday distributions (e.g., peak hours). Naturally, the hourly load would have been more valuable, allowing, for example, an identification of characteristic load distributions within the day. Still, the present tool is certainly perceived as useful, and has been incorporated into the DSO set of tools to detect anomalous cases.

Author Contributions: Conceptualization, J.N.F. and P.M.; methodology, J.N.F.; software, P.M.; validation, P.M.; formal analysis, J.N.F. and P.M.; investigation, P.M.; resources, J.N.F. and P.M.; data curation, P.M.; writing—original draft preparation, J.N.F.; writing—review and editing, J.N.F. and P.M.; visualization, P.M.; supervision, J.N.F.; project administration, J.N.F.; funding acquisition, J.N.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work is financed by National Funds through the Portuguese funding agency, FCT—Fundação para a Ciência e a Tecnologia, within project UIDP/50014/2020.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Nomenclature

API	Application Programming Interface
DB	Database
DSO	Distribution System Operator
HV	High voltage
LV	Low voltage (normal)
LVS	Low voltage (special)
MV	Medium voltage
PAA	Piecewise Aggregate Approximation
SOM	Self-Organizing Maps
SQL	Structured Query Language

References

1. Le Ray, G.; Pinson, P.; Larsen, E.M. Data-driven demand response characterization and quantification. In Proceedings of the 2017 IEEE Manchester PowerTech, Manchester, UK, 18–22 June 2017; pp. 1–6. [CrossRef]
2. Rasouli, V.; Gomes, A.; Antunes, C.H. Characterization of Aggregated Demand-side Flexibility of Small Consumers. In Proceedings of the 2020 International Conference on Smart Energy Systems and Technologies (SEST), Istanbul, Turkey, 7–9 September 2020; pp. 1–6.
3. Khajeh, H.; Firoozi, H.; Hesamzadeh, M.R.; Laaksonen, H.; Shafie-Khah, M. A Local Capacity Market Providing Local and System-Wide Flexibility Services. *IEEE Access* **2021**, *9*, 52336–52351. [CrossRef]
4. Kumar, R.S.; Raghunatha, T.; Deshpande, R.A. Segregation of technical and commercial losses in an 11 kV feeder. In Proceedings of the 2013 7th IEEE GCC Conference and Exhibition (GCC), Doha, Qatar, 17–20 November 2013; pp. 76–79.
5. Viegas, J.; Esteves, P.R.; Melicio, R.; Mendes, V.; Vieira, S.M. Solutions for detection of non-technical losses in the electricity grid: A review. *Renew. Sustain. Energy Rev.* **2017**, *80*, 1256–1268. [CrossRef]
6. Lewis, F.B. Costly throw-ups': Electricity theft and power disruptions. *Electr. J.* **2015**, *28*, 118–135. [CrossRef]
7. De Oliveira, M.E.; Padilha-Feltrin, A.; Candian, F.J. Investigation of the Relationship between Load and Loss Factors for a Brazilian Electric Utility. In Proceedings of the 2006 IEEE/PES Transmission & Distribution Conference and Exposition: Latin America, Caracas, Venezuela, 15–18 August 2006; pp. 1–6.
8. Antmann, P. *Reducing Technical and Non-Technical Losses in the Power Sector (Background Paper for the World Bank Group Energy Sector Strategy)*; Technical Report; World Bank: Washington, DC, USA, 2009. Available online: https://openknowledge.worldbank.org/handle/10986/20786 (accessed on 8 July 2021).
9. Smith, T.B. Electricity theft: A comparative analysis. *Energy Policy* **2004**, *32*, 2067–2076. [CrossRef]
10. Depuru, S.S.S.R.; Wang, L.; Devabhaktuni, V. Electricity theft: Overview, issues, prevention and a smart meter based approach to control theft. *Energy Policy* **2011**, *39*, 1007–1015. [CrossRef]
11. Winther, T. Electricity theft as a relational issue: A comparative look at Zanzibar, Tanzania, and the Sunderban Islands, India. *Energy Sustain. Dev.* **2012**, *16*, 111–119. [CrossRef]
12. Nagi, J.; Mohammad, A.M.; Yap, K.S.; Tiong, S.K.; Ahmed, S.K. Non-Technical Loss analysis for detection of electricity theft using support vector machines. In Proceedings of the 2008 IEEE 2nd International Power and Energy Conference, Johor Bahru, Malaysia, 1–3 December 2008; pp. 907–912.
13. Nizar, A.H.; Zhao, J.H.; Dong, Z.Y. Customer Information System Data Pre-Processing with Feature Selection Techniques for Non-Technical Losses Prediction in an Electricity Market. In Proceedings of the 2006 International Conference on Power System Technology, Chongqing, China, 22–26 October 2006; pp. 1–7.
14. Nizar, A.H.; Dong, Z.Y. Identification and detection of electricity customer behaviour irregularities. In Proceedings of the 2009 IEEE/PES Power Systems Conference and Exposition, Seattle, WA, USA, 15–18 March 2009; pp. 1–10.
15. Ângelos, E.W.S.; Saavedra, O.R.; Cortés, O.A.C.; De Souza, A.N. Detection and Identification of Abnormalities in Customer Consumptions in Power Distribution Systems. *IEEE Trans. Power Deliv.* **2011**, *26*, 2436–2442. [CrossRef]
16. Kadurek, P.; Blom, J.; Cobben, J.F.G.; Kling, W.L. Theft detection and smart metering practices and expectations in the Netherlands. In Proceedings of the 2010 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT Europe), Gothenburg, Sweden, 11–13 October 2010; pp. 1–6.
17. Henriques, H.; Barbero, A.; Ribeiro, R.; Fortes, M.; Zanco, W.; Xavier, O.; Amorim, R. Development of adapted ammeter for fraud detection in low-voltage installations. *Measurement* **2014**, *56*, 1–7. [CrossRef]
18. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020. Available online: https://www.R-project.org/ (accessed on 8 July 2021).

19. Kohonen, T. *Self-Organizing Maps*; Springer: Berlin/Heidelberg, Germany, 2001.
20. Panapakidis, I.P. Clustering based day-ahead and hour-ahead bus load forecasting models. *Int. J. Electr. Power Energy Syst.* **2016**, *80*, 171–178. [CrossRef]
21. Biscarri, F.; Monedero, I.; García, A.; Guerrero, J.I.; León, C. Electricity clustering framework for automatic classification of customer loads. *Expert Syst. Appl.* **2017**, *86*, 54–63. [CrossRef]
22. McLoughlin, F.; Duffy, A.; Conlon, M. A clustering approach to domestic electricity load profile characterisation using smart metering data. *Appl. Energy* **2015**, *141*, 190–199. [CrossRef]
23. Tsekouras, G.J.; Hatziargyriou, N.D.; Dialynas, E.N. Two-Stage Pattern Recognition of Load Curves for Classification of Electricity Customers. *IEEE Trans. Power Syst.* **2007**, *22*, 1120–1128. [CrossRef]
24. Ramos, S.; Duarte, J.M.; Duarte, F.J.; Vale, Z. A data-mining-based methodology to support MV electricity customers' characterization. *Energy Build.* **2015**, *91*, 16–25. [CrossRef]
25. Lynn, S. Self-Organising Maps for Customer Segmentation Using R. 2014. Available online: http://www.shanelynn.ie/self-organising-maps-for-customer-segmentation-using-r/ (accessed on 17 February 2021).
26. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A K-Means Clustering Algorithm. *J. R. Stat. Soc. Ser. Appl. Stat.* **1979**, *28*, 100–108. [CrossRef]
27. Hamerly, G.; Elkan, C. Learning the K in K-means. *Neural Inf. Process. Syst.* **2004**, *16*, 281–284.
28. Tavakoli, K.; Pour-Aboughadareh, A.; Kianersi, F.; Poczai, P.; Etminan, A.; Shooshtari, L. Applications of CRISPR-Cas9 as an advanced genome editing system in life sciences. *BioTech* **2021**, *10*, 14. [CrossRef]
29. Brentan, B.; Meirelles, G.; Luvizotto, E.; Izquierdo, J. Hybrid SOM+k-Means clustering to improve planning, operation and management in water distribution systems. *Environ. Model. Softw.* **2018**, *106*, 77–88. [CrossRef]
30. Frank, E.; Hall, M.A.; Witten, I.H. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, 4th ed.; Morgan Kaufmann: Burlington, MA, USA, 2016.
31. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. *The WEKA Data Mining Software: An Update*; SIGKDD Explorations; Association for Computing Machinery: New York, NY, USA, 2009; Volume 11, Issue 1, pp. 10–18. [CrossRef]
32. Berthold, M.R.; Cebron, N.; Dill, F.; Gabriel, T.R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 319–326.