Detection of anomalous energy consumption through clustering techniques: an application to large-scale food retailing activities

Alessandra Ghilardi[a], Guido Francesco Frate[a], Francesca Leonetti[a], Nicola Fredducci[b], Luca Brancolini[c], Lorenzo Ferrari[a]

[a] *Department of Energy, Systems, Territory, and Constructions Engineering University of Pisa, Pisa (PI), Italy, alessandra.ghilardi@phd.unipi.it, guido.frate@unipi.it, f.leonetti1@studenti.unipi.it, lorenzo.ferrari@unipi.it (**CA**)*
[b] *Unicoop Firenze, Firenze (FI), Italy, nicola.fredducci@unicoopfirenze.coop.it*
[c] *Inres Coop, Osmannoro (FI), Italy, luca.brancolini@inres.coop.it*

Abstract:

Anomalous energy consumption detection is a valuable strategy for pursuing energy efficiency. In commercial buildings, such as supermarkets, abnormal consumption can occur due to non-adequate equipment, such as lighting devices and refrigeration systems, or non-efficient HVAC plant management. Anomaly detection is usually performed on a single building by comparing its energy consumption to its usual behaviour and applying statistical or artificial intelligence-based techniques. Still, no anomaly emerges if its energy consumption is systematically high (or low). However, a more effective method for detecting anomalies would be to compare the energy consumption of a single building with that of others possessing similar characteristics. This paper then proposes an alternative approach based on clustering analysis. From this perspective, energy consumption data from a group of supermarkets are gathered in clusters to detect which presents abnormal behaviour compared to others with similar characteristics, such as the dimension and external weather conditions. An unsupervised density-based clustering algorithm for outlier detection (DBSCAN) is applied to a pool of 87 supermarkets located in Tuscany (Central Italy) to detect the abnormal ones, considering as input features the floor area, the electrical and thermal consumptions available from monthly bills, the type of the air-conditioning units, and the outdoor temperature. The analysis is performed over three years to detect recurring outliers on an annual and monthly scale to investigate possible seasonal effects. During the three years, approximately 15% of the supermarkets were consistently identified as outliers on both a monthly and annual basis. These findings were subsequently validated through an on-site inspection conducted by the energy manager of the supermarkets, revealing that 50% of the identified outliers exhibited exceptionally high thermal and electrical consumption due to improper plant operation.

Keywords:

Anomaly detection; DBSCAN; Energy efficiency; Unsupervised clustering.

1. Introduction

Pursuing energy efficiency in buildings, industry, transport, and energy supply sectors is one of the central policies imposed by the European Council to meet the 2030 emission targets [1]. Through the years, most attention has been dedicated to reducing energy consumption in the building sector since it is responsible for more than 30% of the CO_2 global emissions [2]. Beyond that, relevant emission cuttings can be achieved by actuating energy efficiency strategies also in the non-residential sector, which includes schools, offices and commercial activities. Non-residential buildings are characterised by higher energy intensities than residential ones and are responsible of the 5% of the total share of CO_2 emissions [3]. This highlights the potential for good energy efficiency practices to have a greater impact. Among this category, food-related commercial activities (i.e., supermarkets) are one of the most energy-intense because of the energy consumption related to space cooling and heating and refrigeration for food preservation [4].

Energy efficiency in existing supermarkets has been pursued mainly by replacing old devices with more efficient ones (HVAC units, refrigeration systems and lighting systems) [5], retrofitting the thermal insulation to minimise thermal losses [6] and implementing renewable generation sources [7]. Although widespread, these strategies often require a significantly high investment cost. On the other hand, efficient strategies for energy consumption management can be cheaper to implement. The most common management strategies involve applying optimal control schemes for HVAC units [8], and the refrigeration units [9] by replacing the traditional

https://doi.org/10.52202/069564-0295

PID control schemes with Model Predictive Control. In addition, the consumption optimization has also been pursued by implementing demand-side management paradigms [10,11].

Besides the cited traditional methods, data-based practices for energy consumption monitoring and fault detection are now gaining interest to make residential and non-residential buildings more efficient. In residential buildings, as well as in supermarkets, a large amount of data is usually collected and stored, thanks to the several measurement sensors spread over the energy plants. Using data for energy monitoring can be beneficial to spot abnormal energy consumption due to suboptimal operation of the HVAC or non-adequate equipment operation (i.e., lighting) [12]. Anomalous energy consumption can be identified using state-of-the-art statistical outlier detection methods, such as z-score [13]. Although statistical methods are effective and easy to implement and interpret, with these techniques, outlier detection is typically driven by a single feature (e.g., the energy consumption time series). More advanced data-driven techniques based on Machine Learning (ML) algorithms allow, instead, to perform the outlier detection over a dataset with multiple features, e.g., considering the weather conditions and the building characteristics beyond the only energy consumption data. So far, most academic research papers focus on using ML for anomalous consumption detection of residential buildings. Still, a similar methodological approach can be applied to commercial activities to investigate its potentiality [14].

AI-based energy monitoring techniques aim to detect anomalous energy consumption. Particularly, Unsupervised Clustering (UC) techniques are gaining interest in the energy sector thanks to their simplicity and the broad range of applicability [15]. Mainly, UC is useful to identify abnormalities only considering the intrinsic behaviour of the energy consumption without knowing a priori if it is normal or not. Several UC techniques for anomalous energy consumption exist, such as k-means clustering [18], Gaussian Mixture Model (GMM) [19] and density-based algorithms [20]. The latter category is particularly suitable for outlier detection since it is based on detecting anomalies by dividing the dataset into clusters with high-density data (regular consumption instances) and clusters with low-density data (abnormal consumption instances) [21]. Local Outlier Factor (LOF), Isolation Forest (IF) and Density-Based Spatial Clustering for Application with Noise (DBSCAN) are the most popular algorithms [22]. Unlike the k-means clustering approaches, density-based algorithms do not require initialising the number of clusters. DBSCAN is particularly suitable for outlier detection, since its capability of clustering data and identifying a specific cluster dedicated to outliers [23]. The clustering emerges from data by setting specific parameters defining each cluster's threshold radius of influence and the minimum number of points that define a cluster. outliers Although setting these parameters can be challenging since they cannot be optimised, some automatic strategies can be applied. DBSCAN found many applications for outliers detection in time series so far [24]. Regarding the energy sector, in [25], the authors propose the DBSCAN algorithm for anomalous energy consumption detection in residential buildings. In this case, the setting of the parameters is automatised, and the clustering results are explanatory and reasonable. Focusing only on time series can bring some limitations, though. Energy consumption patterns, indeed, naturally vary due to seasonal trends, making it hard to distinguish if the wrong operation is systematic or due to an extraordinary change in the boundary conditions (e.g., exceptionally hot or cold seasons). For this reason, authors in [26] perform anomaly detection of a building through a two-step procedure, firstly comparing the building energy consumption with past data and then comparing it with a pool of buildings with similar characteristics.

Given this framework, this work aims to apply the DBSCAN clustering for anomalous energy consumption detection of a group of supermarkets. The paper contribution aims to cover some literature gaps, summarized as follows:

- Despite several papers contributing to this topic, most refer only to residential buildings even though commercial buildings are widespread and their energy consumption is more intense than residential ones. Identifying abnormal consumptions could foster significant energy savings in this context.
- The outlier detection is not performed over a single time series related to a single building but within a pool of supermarkets, defining, then, whether the consumption is normal or abnormal compared to the behaviour of the other supermarkets with similar characteristics (such as the floor area and weather conditions). The clustering is performed both on a monthly and yearly basis, repeating the analysis over three different years. The supermarkets, then, are marked as outliers if their anomaly shows systematically.
- The analysis is performed using consumption data extracted from monthly bills for electricity and gas consumption. Though buildings have several sensors, data collection and management can be expensive. Monthly bills are, instead, easy to collect, making the proposed outlier detection method attractive for situations in which detailed data are unavailable.

2. Case study

The selected case study comprises a group of 108 supermarkets in the Tuscany region, in the north centre of Italy. The group is heterogeneous, including mini-markets (floor area of up to 2000 m^2) and superstores (floor area of up to 5000 m^2). Most of the supermarkets, around 80%, use gas boilers for space heating during the

https://doi.org/10.52202/069564-0295

winter season, while the remaining 20% are equipped with electric heat pumps. The weather conditions mainly determine space heating consumption. The stores are located at latitudes between 42° and 44° N and correspond to specific climatic regions established by the Italian government regulation [27]. Most supermarkets are in the D zone, where heating space is allowed between November 1 and April 15. The remaining are in the E region, characterised by more demanding weather conditions, with an allowed heating season from October 15 to April 15.

The energy management division of the supermarkets provided data about consumption and weather conditions and some characteristics of the building over three years, from 2019 to 2021. The available data are as follows:

- **Electrical consumption data** E_{el}**:** electrical consumption data are provided from the monthly bills. The electrical consumption (kWh) is related to the space air conditioning during the summer and winter seasons (only for those supermarkets that are equipped with electric heat pumps), Medium Temperature (MT) refrigerators for fresh food conservation, Low Temperature (LT) refrigerators for frozen food conservation, lighting devices and food processes (e.g., grocery and bakery).
- **Thermal consumption data** E_{th}**:** The monthly bills also provide thermal consumption data measured with Standard cubic meters (Sm^3). The thermal requirement refers to the space heating (during the winter season) and hot water used by the employees for personal usage and food processing actions. Some supermarkets do not have thermal consumption because the air conditioning is electrified.
- **Buildings data:** The information about the building characteristics consists of the monthly opening hours H_o (for a total of about 4000 h/year) and the floor area A_f.
- **Weather data:** The weather conditions consist of two variables: the outdoor temperature and the outdoor humidity measured through sensors positioned outside the building location. These data are provided with a 15-minute timestep. The weather conditions will be expressed as Heating Degree Days (HDD) and Cooling Degree Days (CDD) (Eq. 1), which give a measure of the heating/cooling demand related to the outdoor temperature and the heating/cooling period. In Eq. (1) $T_{sp,i}$ is the indoor temperature set point for air conditioning, $\bar{T}_{ext,i}$ the outdoor temperature mean value over the day, and N is the number of considered days in the heating/cooling period.

$$HDD = \sum_{i=1}^{N} \max\left(T_{sp,i} - \bar{T}_{ext,i}, 0\right) ; \qquad CDD = \sum_{i=1}^{N} \max\left(\bar{T}_{ext,i} - T_{sp,i}, 0\right) \tag{1}$$

Figure *1* provides a summary of the sample data ranges. The histograms illustrate the distribution of the data across the group of supermarkets, indicating the number of buildings in the total population (n/n_{tot}) that share similar characteristics.

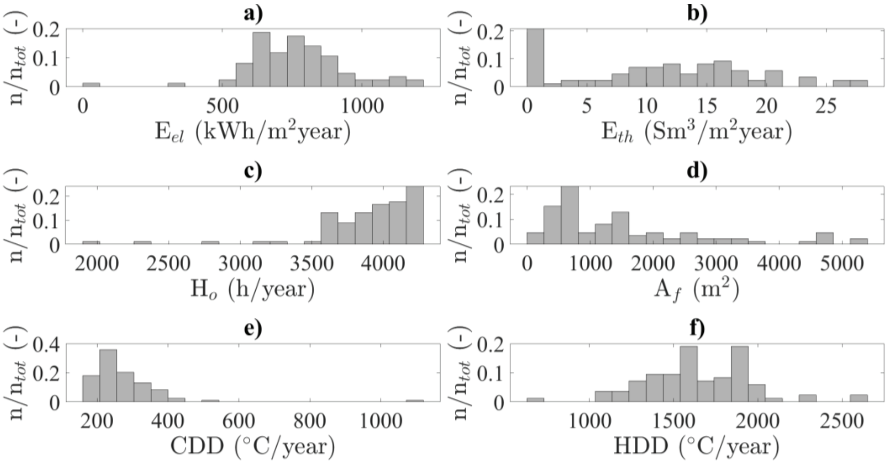

Figure 1. Characteristics of the pool of supermarkets plotted as a fraction of the total supermarkets n/n_{tot}. a) Specific annual electrical consumption E_{el} (kWh/m^2year); b) Specific annual thermal consumption E_{th} (Sm^3/m^2year); c) Annual opening hours Ho (h/year); d) Floor area A_f (m^2) e) Annual Cooling Degree Days CDD (°C/year); f) Annual Heating Degree Days HDD (°C/year).

2.1 Analysed cases

Anomaly detection is performed by considering annual aggregated data and monthly data. The yearly analysis aims to detect which supermarkets have an overall anomaly in energy consumption over the year. Using yearly aggregated data makes this analysis computationally inexpensive and results in the preliminary identification of probable outliers. However, it could also be possible that some supermarkets have abnormal energy consumption only in specific months of the year due to seasonal effects. For example, anomalies could emerge

https://doi.org/10.52202/069564-0295

only during the winter months caused by non-efficient gas boiler usage, while the air conditioning units work correctly in the summer. Monthly analysis can overcome this limit and highlight seasonal effects. For this reason, monthly clustering is also performed even though it is a more computationally expensive analysis since it is repeated every month over three years.

3. Methodology

The methodology section includes the description and preliminary results of the data pre-processing, feature selection, dataset construction and clustering algorithm settings. All the analyses have been realised in Matlab version 2021b.

3.1 Data pre-processing

Pre-process procedures are essential to make raw data usable. Mainly, preliminary operations regarding missing data handling and data aggregation were performed as follows:

- Missing consumption data (both thermal and electric) were not replaced since outlier detection must be performed over actual data to spot anomalies. Supermarkets with missing consumption data were excluded from the analysis, reducing the pool from 108 to 87 supermarkets.
- If possible, missing weather data were interpolated or taken from external databases. Missing periods lower than 24 hours were interpolated linearly. Missing periods longer than one day were replaced with data from the database *IlMeteo.it* [28], which offers historical daily average values for outdoor temperature and humidity. In this case, missing data were replaced with new data, measured not in the proximity of the supermarket but at the closest weather station, usually located in the same municipality.
- Consumption data were aggregated starting from monthly values to obtain yearly global consumption. This operation is necessary to perform monthly and annual clustering to consider seasonal effects, as described in the Case Study Section. Weather data are aggregated into monthly and yearly values starting from the 15-minute time interval measurements.

3.1.1 Feature engineering and feature selection

Few additional features are created starting from the available data to help the clustering algorithm find patterns. The following two features are then created:

- A categorical variable that indicates the air-conditioning plant typology. Supermarkets which use electrical heat pumps for the summer and winter air-conditioning are marked with category E. In contrast, supermarkets which use the gas boiler during the heating season are marked with category EG. Following the standard procedure for categorical variables handling, this feature is processed as a dummy variable [29].
- HDD and CDD. Average values of outdoor air temperature are not particularly meaningful, mainly if the average refers to an extended period (month or year). HDD and CDD, instead, quantify the overall heating/cooling demand, involving the length of the heating/cooling season and the difference between the indoor set point temperature and the outdoor conditions.-Once the set of available variables is defined, feature selection is performed to remove unnecessary data and improve the clustering performance [30]. For this reason, a correlation analysis is performed to identify the more significant features. The feature selection is performed through the evaluation of the Spearman correlation coefficient ρ_s. Spearman correlation coefficient is a statistical measure of the correlation within the variable *x* and *y* based on rank assignment. For the calculation, the values of *x* and *y* are converted into ranks following an ascending or descending order. Then a high correlation score is assigned where a high rank of *x* corresponds to a high rank of *y*. Unlike the Pearson's coefficient, Spearman's is able to calculate the correlation between two variables even if their relationship is not strictly linear. The only assumption is that the relationship has to be monotonic [31]. The calculation of ρ_s is performed as shown in Eq. (2), where D_i^2 is the rank distance between variable *x* and *y*, and N the number of points which constitutes the cluster. The correlation is considered strong when $0.7 \leq |\rho_s| \leq 1$, moderate for $0.3 \leq |\rho_s| < 0.7$ and weak for $|\rho_s| < 0.3$.

$$\rho_s = 1 - 6 \cdot \frac{\sum_{i=1}^{N} D_i^2}{N(N-1)} \; ; \qquad D_i = x_i - y_i \tag{2}$$

Correlation analysis is not performed for the yearly aggregated values because no monotonic relationship emerges considering annual data. Correlation analysis is then performed only for the monthly values. Particularly, ρ_s is computed across the following variables:

- Electrical specific consumption E_{el} (kWh/m^2month)
- Thermal specific consumption E_{th} (Sm^3/m^2month)
- Degree Days DD $(°C/month)$. The generic symbol DD refers to HDD for winter months and to CDD for summer months
- Outdoor Relative Humidity RH (%)
- Opening hours H_o (h).

https://doi.org/10.52202/069564-0295

The preliminary results of feature selection are shown in **Figure *2***. The heatmap shows the correlation coefficients ρ_s across the combination of variables for one of the supermarkets as an example. The other supermarkets of the pool showed the same behaviour and are not shown for brevity. As expected, the correlation analysis shows that electrical and thermal consumptions are strongly related to the external temperature conditions (DD). RH showed no correlation with energy consumption, so it is removed from the analysis. Finally, since H_o showed only a moderate correlation with electrical consumption, it was included in the first tests. However, its impact was not significative for the outliers detection, so it was removed in the final dataset. In this plot, the correlation with the floor area was not computed because only one supermarket is considered (so the area does not vary).

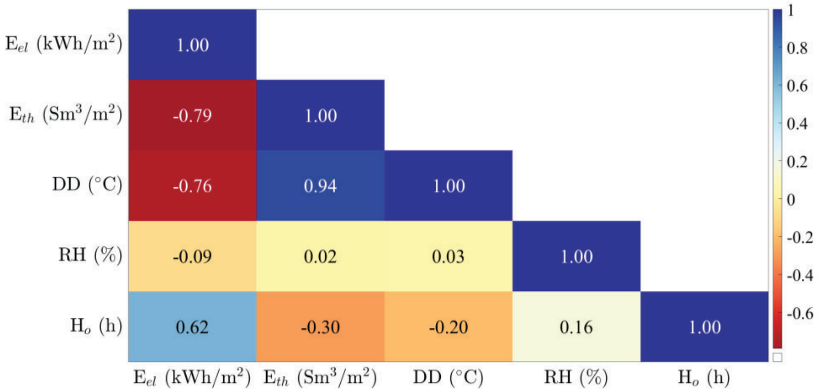

Figure 2. Spearman correlation values for the considered features E_{el}, E_{th}, DD, RH and opening hours. Data are related to one of the supermarkets of the pool as an example.

3.1.2 Dataset construction

As a result of the correlation analysis and considerations given in the previous sections, the annual and monthly datasets are finalised, as **Table *1*** summarises. Finally, each dataset is normalised to ease the clustering procedure. Normalised data are all in the same range, so undesired effects due to different data scales will not affect the clustering results. In this work, data are normalised with the 2-norm approach, in which the Euclidean norm of the variable normalises each data observation (row). Eq. (3) shows the normalisation approach. The generic data x_i is normalised, becoming $x_{i,norm}$, dividing it by the 2-norm, where N is the number of rows of the variable.

$$x_{i,norm} = \frac{x_i}{\left[\sum_{k=1}^{N} |x_k|^2\right]^{1/2}} \tag{3}$$

Table 1. Dataset with selected features for annual and monthly clustering

	Features included in the dataset
Annual clustering	• $E_{el}\ (kWh/m^2year)$ • $E_{th}\ (Sm^3/m^2year)$ • $A_f\ (m^2)$ • Air-conditioning units type: E/EG
Monthly clustering	• $E_{el}\ (kWh/m^2month)$ • $E_{th}\ (Sm^3/m^2month)$ • $A_f\ (m^2)$ • $DD\ (°C/month)$ • Air-conditioning units type: E/EG

3.2 DBSCAN algorithm

DBSCAN algorithm is an unsupervised clustering technique that identifies high-density regions (normal behaviour) in the k-dimensional space representing the dataset and a few low-density regions where the outliers are located (abnormal behaviour). High-density areas, i.e., the clusters, are defined based on the neighbourhood concept, for which two points are in the same neighbourhood if their distance is below a threshold ε, called the neighbourhood parameter. The pairwise distance from a point to the surrounding ones can be defined in several ways, but the Euclidean distance is used in the analysis. Particularly the neighbourhood of a point *x*, $N_{\varepsilon}(\mathrm{x})$*,* is defined as in Eq. (4), where D is the set of points in the dataset, dist(x,y) is the Euclidean distance between two points *x* and *y,* and ε is the threshold distance [32].

$$N_{\varepsilon}(\mathrm{x}) = \{\mathrm{y} \in \mathrm{D} | \mathrm{dist}(x, y) \le \varepsilon\}; \qquad \mathrm{dist}(x, y) = \sqrt[2]{\sum_{i=1}^{N} (x_i - y_i)^2} \tag{4}$$

The cluster definition must include an additional parameter to identify high-density regions correctly. Particularly, $N_{\varepsilon}(\mathrm{x})$ cannot be constituted by less than a number of points equal to *minPts* (a scalar number ≥ 1). The DBSCAN algorithm operates as represented in **Figure 3**.

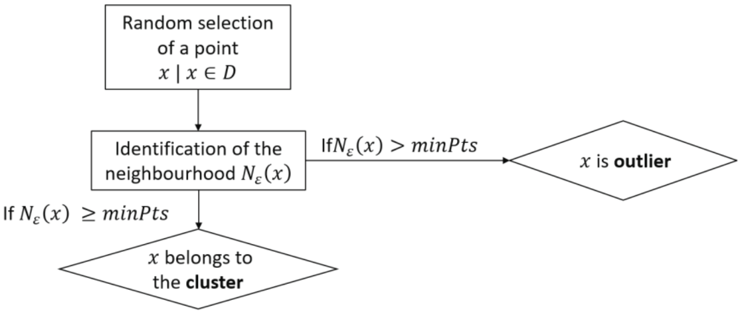
Random selection of a point $x \mid x \in D$
Identification of the neighbourhood $N_{\varepsilon}(x)$
If$N_{\varepsilon}(x) > minPts$
x is **outlier**
If $N_{\varepsilon}(x) \ \geq minPts$
x belongs to the **cluster**

Figure 3. Clustering process using DBSCAN algorithm.

3.2.1 Parameters setting

Since the abnormal energy consumption detection for this case study is unsupervised, the DBSCAN model cannot be trained based on the experience of some previously labelled data which constitute the benchmark (normal or abnormal). For this reason, the parameter setting cannot be optimised by minimising a specified loss function, as happens for supervised cases. The choice of the parameters ε and $minPts$, then, is driven by some specific considerations related to the dataset. However, some general guidelines can be followed to have a proper estimation of suitable values of ε and *minPts*:

- The choice of *minPts* depends on the configuration of the dataset. This work performs a sensitivity analysis varying minPts from 2 to 8. These values are selected considering the size of groups of supermarkets with similar characteristics, like the floor area or the geographical location, in the dataset.
- The choice of ε is based on the method proposed in [33]. To determine a meaningful ε value, the pair-wise distance is calculated from each point in the dataset to all the other *minPts* number of points. The obtained distances are then plotted against the sorted points. Since the number of anomalies is usually limited (i.e., less than 20%), the sorted distances show a typical linear trend for the points in the high-density areas. A fast slope change occurs in the proximity of the so-called "elbow", where the remaining sorted distances are substantial because they are related to the low-density areas (where the outliers are located and the points tend to be distant from each other). The selected ε value, then, is the distance value corresponding to the "elbow*".* Since the results can be affected by the value of ε (i.e., the larger ε, the lower the number of outliers), we performed a sensitivity analysis with different values near the *elbow*, and the results proved to be robust in this range*.* The value selected for ε varies for each simulated case, ranging from 0.0286 to 0.0853.

3.2.2 Key Performance Indicators

The quality of the unsupervised clustering is not quantitative but qualitative since data are not labelled a priori. Results will then be shown with graphical scatter plots to represent clusters and outliers. Despite the limitation of the unsupervised nature of the problem, some additional considerations can prove that one supermarket can be considered an outlier. The clustering analysis is repeated for each year in the dataset (2019-2020-2021) to support this thesis. The number of recurrent outliers is then collected over single or multiple years. Concerning the annual analysis, the fact that a supermarket has been marked as an outlier for more than one year suggests that an issue of some sort occurs systematically.

On the other hand, supermarkets marked as outliers for a single year probably faced some extraordinary operative conditions which never repeated. The same considerations are valid for the monthly clustering. In this case, supermarkets were marked as outliers over one year if they were outliers for at least three months. This value is reasonable because it can reflect seasonal effects due to the heating or cooling periods. Two Key Performance Indicators (KPIs) are then evaluated to quantify the cited results:

- Percentage of outliers n_{out} over 1, 2 or 3 years compared to the total number of supermarkets n_{sup} (Eq. 5):

$$k_{t,sup} = \frac{n_{out}}{n_{sup}} \cdot 100\% \tag{5}$$

- Percentage of outliers over 1, 2 or 3 years compared to the total number of outliers $n_{out,tot}$ (Eq. 6):

$$k_{t,out} = \frac{n_{out}}{n_{out,tot}} \cdot 100\% \quad (6)$$

4. Results and discussion

This session will provide at first step results about the clustering quality by visualising scatter plots with non-anomalous clusters and outliers. After that, the analysis of the KPI will help to understand the supermarkets which showed a systematic anomaly, which can be classified as actual outliers. Finally, the outliers found by the algorithm are verified by plant inspections conducted by the energy manager to validate the clustering results.

4.1 Qualitative clustering

The results of the annual clustering in **Figure *4*** are scattered in a three-dimensional plot that includes the three features defined in the dataset (E_{el}, E_{th} and A_f). The clustering is sensible since the main characteristics of the supermarkets are represented. The regions of the dataset with the highest density are clustered in Group 1 and Group 2. The two clusters are related to supermarkets with electric pumps (category E) and supermarkets with gas boilers (category EG). Smaller clusters indicate limited groups of supermarkets which are similar. Group 3 represents superstores (i.e., most extensive floor areas) with electric heat pumps, while Group 4 concerns superstores with gas boilers for space heating. Outliers emerge beyond regular clusters and are located mostly in low-density regions. Among the whole group of outliers, some show low consumption compared to the average of the surrounding clusters, while others show high consumption. The latter outliers were verified through an on-site plant inspection, finding some explanations for the abnormalities. For example, supermarkets 29, 74 and 51 are outliers because the electrical consumption is abnormal despite the A_f and E_{th} being similar to other supermarkets. In these cases, the anomaly was found to be due to obsolete lighting devices or thermal losses of LT and NT refrigerators that are not closed. The electrical consumption anomalies related to the air-conditioning units are rare since supermarkets of category E are recent, efficient, and monitored. Supermarkets 7, 6 and 45 have abnormal thermal consumption. In these cases, the anomaly is primarily due to sub-optimal operation or out-of-range setpoint temperatures on the heating boilers. Supermarkets of category EG are indeed older than the ones in category E, so the heating plants are not monitored in real-time.

On the other hand, a few supermarkets are marked as anomalous because of lower energy consumption than the neighbourhood, so the on-site inspection was not performed. In some cases, this behaviour is due to the low-quality data, which are partially missing (supermarket 31). In other cases, the low consumption is due to favourable climate conditions during the considered year or the wrong consumption estimation on the bills. The weather conditions effects emerge through the monthly analysis, in which HDD and CDD are used as predictors. **Figure *5*** can also be interpreted qualitatively as for the annual clustering results. In this case, some anomalies can be interpreted thanks to the additional information about the climate conditions. Exceptionally high DD are responsible for some detected abnormalities. For example, supermarkets 76 and 7 have similar DD to many supermarkets of the pool but significantly higher thermal consumption. The heating plants' sub-optimal management set point temperatures can be responsible for these abnormalities. Compared to these, Supermarket 48 is less alarming because it has higher E_{th} due to higher DD. Finally, supermarkets 18 and 25 have lower thermal consumption than the neighbourhood, but this is due to lower DD, indicating that the outdoor temperature was exceptionally lower than usual that month.

4.2 Systematic outliers

The qualitative considerations help support the clustering results, which cannot be compared to a benchmark. Despite this, the interpretation of the results may be impractical as it must be supported by visuals, and every outlier requires some additional effort to be confirmed or denied. Automating this process can be challenging, but iterating the analysis over multiple years can help to strengthen the outlier labelling with more reliability.

Figure *7* highlights the capability of the monthly clustering in understanding systematic seasonal effects that did not emerge from the annual analysis. The analysis is valid for all the months of the year except for September. In this case, data concerning E_{th} of September 2019 were missing, so the statistics could not be computed for this month since the data quality still was not good enough despite data pre-processing. However, the analysis is still relevant for the other eleven months. This graph shows that during winter months, from November to March, a high percentage of supermarkets are outliers over the whole three years. These systematic abnormalities highlight an improper operation of the heating systems, so abnormal consumption in the cold season is highly probable.

https://doi.org/10.52202/069564-0295

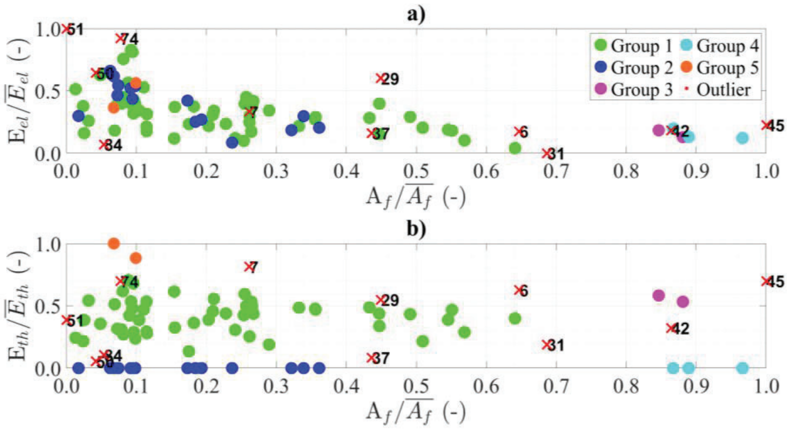

Figure 4. Annual clustering results for the year 2021, as an example. a) Specific electrical consumption Eel versus floor area A_f. b) Specific thermal consumption E_{th} versus floor area. All the variables are normalised with the maximum values $\bar{E}_{el}$, $\bar{E}_{th}$, $\bar{A}_f$ for data privacy. Outliers are marked with red crosses and are associated with a numerical ID.

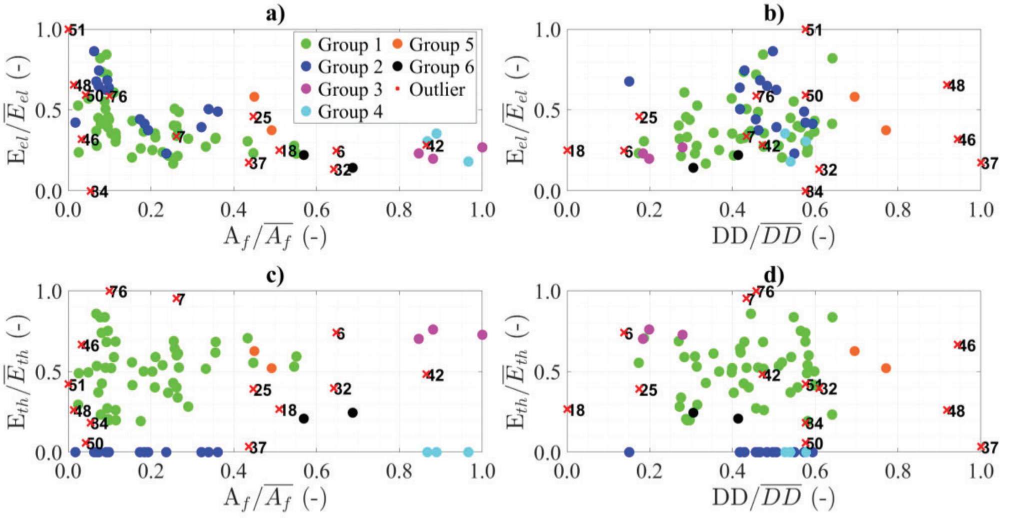

Figure 5. Monthly clustering for January 2021, as an example. a) Specific electrical consumption E_{el} versus floor area A_f. b) Specific electrical consumption E_{el} versus Degree Days DD. c) Specific thermal consumption E_{th} versus floor area A_f. d) Specific thermal consumption Eth versus Degree Days DD. All the variables are normalized with the maximum values $\bar{E}_{el}$, $\bar{E}_{th}$, $\bar{A}_f$ and $\overline{DD}$ for data privacy. Outliers are marked with red crosses and are associated with a numerical ID.

Figure *6* shows how many supermarkets were marked as outliers for 1/3, 2/3 or 3/3 years of the investigated period. Around 10% of the supermarkets had abnormal consumption for only one year over three, highlighting that these abnormalities are not systematic but exceptional. Instead, the green stacked bar highlights the percentage of supermarkets that were marked as outliers for all three years. In this case, since the anomaly is systematic, these supermarkets probably have operating issues related to the consumption of their subsystems. This result is strengthened by comparing the annual and monthly $k_{t,s}$ and $k_{t,out}$ indicators. Regarding 1/3 years and 2/3 years outliers, there are a few differences. The monthly analysis in **Figure *7*** highlighted occasional outliers (orange stacked bar), thanks to its capability of catching seasonal anomalies that cannot appear in the annual analysis, so the results are slightly different. Regarding the 3/3 years abnormalities, yearly and monthly clustering showed almost the same percentage. This result highlights that around 10% of supermarkets in the pool have systematic anomalous consumption that emerges independently from the weather conditions, so it strictly concerns operational issues of the plants. This result is then valuable information for the energy management unit of the supermarkets since it suggests that a plant check is necessary. As a final remark, thanks to this analysis, systematic anomalous supermarkets are marked as outliers with sufficient certainty. The cost of the on-site plant inspection to verify these results and eventually fix the anomaly, then, is justified and can bring to potential savings. **Table 2** provides the validation of the results through the on-site plant inspection. The 50% of the recurrent outliers proved to have some issues that caused the abnormal consumption, such as the faulty operation of the HVAC units and old lighting equipment. Only in one case the abnormality was caused by a wrong estimation of the monthly bills. The remaining 50% were not found to be outliers, mainly because they were the cases that showed a positive abnormality (i.e., with consumptions lower than supermarkets with similar features).

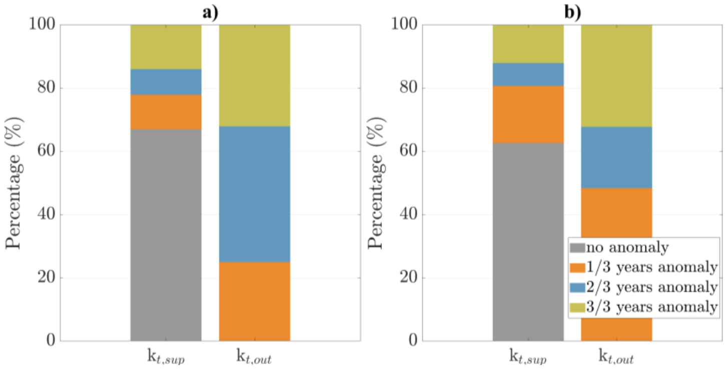

Figure 6. Abnormalities share for a) Annual analysis; b) Monthly analysis.

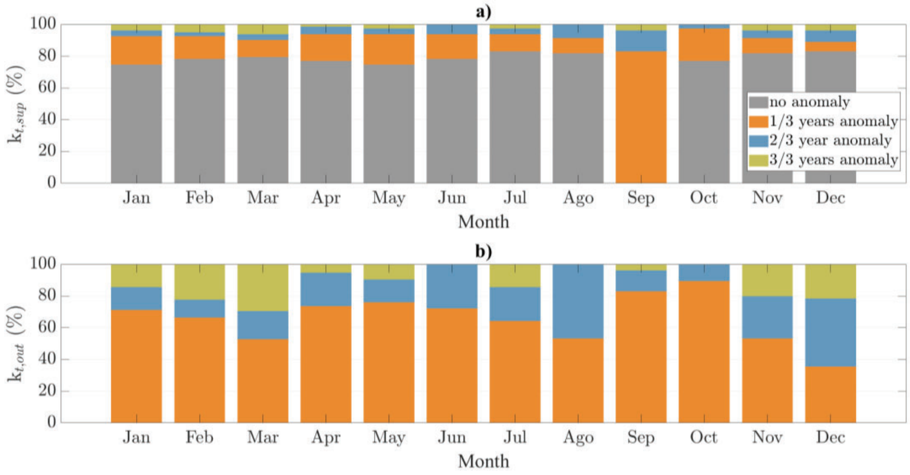

Figure 7. Abnormalities share for monthly analysis for KPIs: a) percentage of outliers among the total number of supermarkets; b) percentage of outliers among the total number of outliers.

Table 2. Results validation for recurrent outliers emerged from the monthly clustering analysis.

Anomalous supermarket	Years of anomaly	Anomaly cause
6	3/3	Old lighting devices; Heating system
7	3/3	Heating system
19	2/3	Heating system
29	3/3	Old lighting devices; HVAC system
30	3/3	Old lighting devices
43	3/3	Heating system
56	3/3	Wrong bill estimation
75	2/3	Heating system

5. Conclusions

The paper investigated the potentialities of using an unsupervised clustering approach for anomalous energy consumption detection in a group of 87 supermarkets in Italy. The final goal of the study was to implement a computationally powerful strategy to identify anomalous supermarkets compared to the others in the group. Unlike standard statistical techniques, this approach helped to identify the outliers considering many features simultaneously, such as thermal and electrical consumptions, the dimensions and weather conditions. The analysis was performed using the DBSCAN algorithm, which is suitable for outlier detection for datasets mainly composed of high-density regions (regular consumptions) and a few low-density regions (abnormalities).

Although this methodological approach had already been applied for anomalous energy consumption detection, it was rarely applied to non-residential buildings, which can easily suffer from non-optimal energy usage because of their higher complexity and energy intensity. In addition, most of the papers performed the outlier detection by comparing one building to its past energy consumption, making it impossible to identify systematic anomalies. The work then tried to bridge these research gaps by observing numerous similar supermarkets, starting from monthly electrical and thermal consumption bills, which are data easy to collect and manage.

Firstly, the results proved that the proposed set of clustering parameters effectively represented the dataset adequately by identifying the main clusters in relation to the air-conditioning plan typology. Once the good quality of the clustering was verified, the annual clustering identified the anomalous supermarkets using consumption data and the floor area as features. After that, weather conditions were embedded in the monthly analysis, which was able to identify the supermarkets that behave anomalously only in particular months due to seasonal effects. Finally, the analyses went beyond the simple labelling as *normal* or *abnormal* consumption by repeating the clustering over the three available years of the dataset. By doing so, supermarkets labelled as outliers over the three years can be considered very likely outliers. Approximately 10 % of the supermarkets of the pool showed an anomaly over 3/3 years. This result was confirmed by comparing the annual and the monthly analyses so that supermarkets systematically affected by significant abnormal energy consumption are identified with improved confidence. This improvement is particularly meaningful since it overcomes the issues of the traditional unsupervised clustering, in which the non-optimal setting of the parameters brings uncertain results. In addition, the monthly analysis highlighted that around 50 % of supermarkets are outliers for only 1/3 years. These supermarkets are not particularly alarming since their abnormality is not systematic and is probably due to non-frequent wrong operation or extreme weather conditions. Repeating the clustering over different time scales and different years then helps to detect the *false positive* outliers and focus the effort on the systematic ones. This result is validated by the on-site inspection of the potential anomalous plants, which proved to have operational issues.

Further improvements could concern the integration of additional features, as far as the data are available (like the occupancy, the type of refrigeration units and so on). In addition, if the supermarkets change during the investigation period, additional characteristics of new equipment could improve the clustering quality. However, since this kind of data is usually difficult to collect, the paper proposes a methodology involving data which are easily accessible.

Acknowledgements

This research has received financial contribution from the Italian Operative National Plan (Piano Operativo Nazionale, PON) in the framework of the project Ricerca e Innovazione 2014−2020 (PON R&I) – Azione IV.6 “Contratti di ricerca su tematiche dell'innovazione e green” (DM MUR 1062/2022).

Prof. Lorenzo Ferrari and Dr. Guido Francesco Frate gratefully acknowledge the financial support of the University of Pisa (UNIPI), in the framework of the Research Project PRA 2022_31: MetOdi per riDurre gli usi di EneRgiA Termica ed Elettrica in ambito civile e industriale (MODERATE).

Nomenclature

AI	Artificial Intelligence
CDD	Cooling Degree Days
DBSCAN	Density-Based Spatial Clustering for Application with Noise
DD	Degree Days
GMM	Gaussian Mixture Model
HDD	Heating Degree Days
HVAC	Heating, Ventilation and Air Conditioning
IF	Isolation Forest
KPI	Key Performance Indicator
LOF	Local Outlier Factor
LT	Low Temperature
minPts	Minimum number of Points
ML	Machine Learning
MT	Medium Temperature
RH	Relative Humidity
UC	Unsupervised Clustering
UD	Unsupervised Detection

https://doi.org/10.52202/069564-0295

Symbols

A	Area
D	Distance
E	Energy consumption
H	Hours
k	Percentage parameter
N	Neighbourhood
n	Number
T	Temperature
x, y	Generic points

Greek symbols

ρ	correlation coefficient
ε	eps parameter

Subscripts and superscripts

el	electrical
ext	external
f	floor
$norm$	normalised
o	opening
out	outlier
s	Spearman
sp	setpoint
sup	supermarkets
th	thermal
tot	total

References

[1] Energy efficiency targets n.d. https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficiency-targets-directive-and-rules/energy-efficiency-targets_en (accessed February 27, 2023).

[2] Ngarambe J, Yun GY, Santamouris M. The use of artificial intelligence (AI) methods in the prediction of thermal comfort in buildings: energy implications of AI-based thermal comfort controls. Energy Build 2020;211:109807. https://doi.org/10.1016/j.enbuild.2020.109807.

[3] World Energy Balances Highlights - Data product - IEA n.d. https://www.iea.org/data-and-statistics/data-product/world-energy-balances-highlights (accessed March 2, 2023).

[4] Monforti F, Dallemand JF, Motola V. Energy use in the EU food sector: State of play and opportunities for improvement Development of long-term energy projections for African countries View project Coal regions in transition View project. 2015. https://doi.org/10.2790/158316.

[5] Ríos Fernández JC, Roqueñí N. Analysis of the potential of Spanish supermarkets to contribute to the mitigation of climate change. Sustain Prod Consum 2018;14:122–8. https://doi.org/10.1016/j.spc.2018.02.003.

[6] Gigoni L, Betti A, Crisostomi E, Franco A, Tucci M, Bizzarri F, et al. Day-Ahead Hourly Forecasting of Power Generation from Photovoltaic Plants. IEEE Trans Sustain Energy 2018;9:831–42. https://doi.org/10.1109/TSTE.2017.2762435.

[7] Franco A, Cillari G. Energy sustainability of food stores and supermarkets through the installation of pv integrated plants. Energies 2021;14. https://doi.org/10.3390/en14185678.

[8] Ge YT, Tassou SA. Control optimizations for heat recovery from CO2 refrigeration systems in supermarket. Energy Convers Manag 2014;78:245–52. https://doi.org/10.1016/j.enconman.2013.10.071.

[9] Hovgaard TG, Larsen LFS, Edlund K, Jørgensen JB. Model predictive control technologies for efficient and flexible power consumption in refrigeration systems. Energy 2012;44:105–16. https://doi.org/10.1016/j.energy.2011.12.007.

[10] Glavan M, Gradišar D, Moscariello S, Juričić Đ, Vrančić D. Demand-side improvement of short-term load forecasting using a proactive load management – a supermarket use case. Energy Build 2019;186:186–94. https://doi.org/10.1016/j.enbuild.2019.01.016.

[11] Coccia G, D'Agaro P, Cortella G, Polonara F, Arteconi A. Demand side management analysis of a supermarket integrated HVAC, refrigeration and water loop heat pump system. Appl Therm Eng 2019;152:543–50. https://doi.org/10.1016/j.applthermaleng.2019.02.101.

[12] Wang A, Lam JCK, Song S, Li VOK, Guo P. Can smart energy information interventions help householders save electricity? A SVR machine learning approach. Environ Sci Policy 2020;112:381–93. https://doi.org/10.1016/j.envsci.2020.07.003.

[13] Seem JE. Using intelligent data analysis to detect abnormal energy consumption in buildings. Energy Build 2007;39:52–8. https://doi.org/10.1016/j.enbuild.2006.03.033.

[14] Miller C, Nagy Z, Schlueter A. A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings. Renew Sustain Energy Rev 2018;81:1365–77. https://doi.org/10.1016/j.rser.2017.05.124.

[15] Himeur Y, Ghanem K, Alsalemi A, Bensaali F, Amira A. Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives. Appl Energy 2021;287. https://doi.org/10.1016/j.apenergy.2021.116601.

[16] Anil Kumar KS, Chacko AMMO. Clustering Algorithms for Intrusion Detection: A Broad Visualization. ACM Int Conf Proceeding Ser 2016;04-05-Marc:3–6. https://doi.org/10.1145/2905055.2905195.

[17] Ahmed M, Mahmood AN, Islam MR. A survey of anomaly detection techniques in financial domain. Futur Gener Comput Syst 2016;55:278–88. https://doi.org/10.1016/j.future.2015.01.001.

[18] Henriques J, Caldeira F, Cruz T, Simões P. Combining k-means and xgboost models for anomaly detection using log datasets. Electron 2020;9:1–17. https://doi.org/10.3390/electronics9071164.

[19] Ahmed SRA, Al-Barazanchi I, Jaaz ZA, Abdulshaheed HR. Clustering algorithms subjected to K-mean and gaussian mixture model on multidimensional data set. Period Eng Nat Sci 2019;7:448–57. https://doi.org/10.21533/PEN.V7I2.484.

[20] Bhattacharjee P, Mitra P. A survey of density based clustering algorithms. Front Comput Sci 2021;15. https://doi.org/10.1007/s11704-019-9059-3.

[21] Himeur Y, Alsalemi A, Bensaali F, Amira A. Smart power consumption abnormality detection in buildings using micromoments and improved K-nearest neighbors. Int J Intell Syst 2021;36:2865–94. https://doi.org/10.1002/int.22404.

[22] Pereira W, Ferscha A, Weigl K. Unsupervised detection of unusual behaviors from smart home energy data. Lect Notes Comput Sci (Including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2016;9693:523–34. https://doi.org/10.1007/978-3-319-39384-1_46.

[23] Fong S, Rehman SU, Aziz K, Science I. DBSCAN : Past , Present and Future 2014:232–8. doi: 10.1109/ICADIWT.2014.6814687.

[24] Jain P, Bajpai MS, Pamula R. A Modified DBSCAN Algorithm for Anomaly Detection in Time-series Data with Seasonality. Int Arab J Inf Technol 2022;19:23–8. https://doi.org/10.34028/iajit/19/1/3.

[25] Yao G, Guo C, Ge Q, Ait-Ahmed M. A practical building energy consumption anomaly detection method based on parameter adaptive setting DBSCAN. Cogn Comput Syst 2021;3:154–68. https://doi.org/10.1049/ccs2.12015.

[26] Arjunan P, Khadilkar HD, Ganu T, Charbiwala ZM, Singh A, Singh P. Multi-user energy consumption monitoring and anomaly detection with partial context information. BuildSys 2015 - Proc 2nd ACM Int Conf Embed Syst Energy-Efficient Built 2015:35–44. https://doi.org/10.1145/2821650.2821662.

[27] Gazzetta Ufficiale n.d. https://www.gazzettaufficiale.it/eli/id/1993/10/14/093G0451/sg (accessed May 11, 2023).

[28] Che tempo faceva a Firenze - Archivio Meteo Firenze » ILMETEO.it n.d. https://www.ilmeteo.it/portale/archivio-meteo/Firenze (accessed March 3, 2023).

[29] Alkharusi H. Categorical Variables in Regression Analysis: A Comparison of Dummy and Effect Coding. Int J Educ 2012;4:202. https://doi.org/10.5296/ije.v4i2.1962.

[30] Chicco G. Overview and performance assessment of the clustering methods for electrical load pattern grouping. Energy 2012;42:68–80. https://doi.org/10.1016/j.energy.2011.12.031.

[31] Hauke J, Kossowski T. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. Quaest Geogr 2011;30:87–93. https://doi.org/10.2478/v10117-011-0021-1.

[32] Ester M, Kriegel H-P, Sander J, Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proc 2nd Int Conf Knowl Discov Data Min 1996. https://doi.org/10.11901/1005.3093.2016.318.

[33] DBSCAN - MATLAB & Simulink n.d. https://www.mathworks.com/help/stats/dbscan-clustering.html#mw_4aa35c21-70f7-43a8-b310-1db43ea97eae (accessed March 7, 2023).