

Linguistic Modeling and Synthesis of Heterogeneous Energy Consumption Time Series Sets

Sergio Martínez-Municio, Luis Rodríguez-Benítez, Ester Castillo-Herrera, Juan Giralt-Muiña, Luis Jiménez-Linares*

Department of Information and System Technologies, Escuela Superior de Informatica, University of Castilla-La Mancha, Paseo de la Universidad 4, Ciudad Real, 13071, Spain

ARTICLE INFO

Article History

Received 25 Jun 2018

Revised 07 Sep 2018

Accepted 02 Dec 2018

Keywords

Time series

Linguistic summaries

Fuzzy model

Clustering

Energetic consumption

ABSTRACT

Thanks to the presence of sensors and the boom in technologies typical of the *Internet of things*, we can now monitor and record the energy consumption of buildings over time. By effectively analyzing these data to capture consumption patterns, significant reductions in consumption can be achieved and this can contribute to a building's sustainability. In this work, we propose a framework from which we can define models that capture this casuistry, gathering a set of time series of electrical consumption. The objective of these models is to obtain a linguistic summary based on *y is P* protoforms that describes in natural language the consumption of a given building or group of buildings in a specific time period. The definition of these descriptions has been solved by means of fuzzy linguistic summaries. As a novelty in this field, we propose an extension that is able to capture situations where the membership of the fuzzy sets is not very marked, which obtains an enriched semantics. In addition, to support these models, the development of a *software* prototype has been carried out and a small applied study of actual consumption data from an educational organization based on the conclusions that can be drawn from the techniques that we have described, demonstrating its capabilities in summarizing consumption situations. Finally, it is intended that this work will be useful to managers of buildings or organizational managers because it will enable them to better understand consumption in a brief and concise manner, allowing them to save costs derived from energy supply by establishing sustainable policies.

© 2019 The Authors. Published by Atlantis Press SARL.

This is an open access article distributed under the CC BY-NC 4.0 license (<http://creativecommons.org/licenses/by-nc/4.0/>).

1. INTRODUCTION

Currently, society faces the challenge of using in a rational and efficient way the resources that it needs to carry out its activities. From among all these resources, energy is one of the fundamental pillars for its proper functioning. As quoted by the Energy Agency in [1], buildings represent more than a third of total energy and half of the world's electricity supply, which is the area of highest energy consumption. This leads to a direct responsibility for the emission of approximately one third of global carbon dioxide emissions. Nowadays, there is a growing interest in reducing energy consumption and greenhouse gas emissions in each sector of the economy [2]. However, with a projected population increase of 2500 million people by mid-century, and improvements in economic development and living standards, a dramatic increase in energy use in buildings is expected, which puts additional pressure over the energy system. Today, technological tools are in place and measures have been taken to help achieve greater energy efficiency and sustainability in buildings; however, these measures do not seem to be a priority due to their high economic cost and the difficulty of implementation in relatively short periods of time. Consequently, as is detailed in [3], there is a need to plan such policies in a framework of implementation and continuous review. The emergence of new

technologies, such as smart sensors or complex networks that give rise to the Internet of things (IOT), provides the ability to continuously monitor. This yields vast amounts of data, which allows us to gain a detailed understanding of how we use energy at a given point in time in quantitative terms; for example, the consumption of a building is *112 KW/h—11:00—20 February 2016*. Although this planning and further application of policies should not be based solely on this kind of solution, it is essential to incorporate models that characterize energy consumption patterns in qualitative terms, increasing their expressiveness and allowing assertions such as *most buildings have low consumption*.

Our work context is an organization made up of a set of geographically distributed buildings, whose electricity consumption data are available every hour and are provided by a meter, which are then aggregated to obtain the total daily consumption. This aggregation is carried out to obtain information on annual consumption per building, where each year per building contains 365 consumption data. A *clustering* algorithm will subsequently be applied to provide a categorization of consumption. These categorizations represent the annual consumption model for each building of the organization. This enables comparisons between individual consumption meters (buildings) within the organizational framework. Once the model has been set, the focus is now to establish linguistic descriptions that sum up a characterization of the daily behavior of each building and the error made by the model of each meter

* Corresponding author. Email: luis.jimenez@uclm.es

defined with phrases of the type: *the activity consumption of a building is large* or *the model underestimates the expected consumption*, respectively. For instance, the authors in [4] presents the adaptation of methodologies to generate customized linguistic descriptions of data.

The contributions of this work are the definition and generation of a model that summarizes the consumption behavior obtained in an organization in linguistic terms using *clustering* techniques. The linguistic component has been resolved through fuzzy linguistic summaries based on *y is P* protoforms. We propose their extension so that they have enough semantic capacity to express what consumption is, if we obtain little differentiated values. Finally, a *software* prototype has been designed and implemented to support the defined models and to enable these concepts to be presented visually through a graphical dashboard.

The rest of the paper is structured as follows: In Section 2, a review of the state of the art is presented on works related to the definition of electricity consumption models and linguistic summaries. An introduction to the concept of the meter model and also the dataset to be processed, together with a brief explanation of the techniques used for its definition, is shown in Section 3. Section 4 introduces the concept of organizational model and the definition of linguistic summaries are based on it. The error made by the models defined in linguistic terms is studied in Section 5. An extension of the classic summaries that enhances their semantic capabilities is presented in Section 6. Section 7 discusses the conclusions reached by applying the defined models to actual consumption data. Finally, Section 8 sums up the main findings and it makes a number of recommendations for future work.

2. BACKGROUND

The analysis of data to establish electricity consumption models has been discussed and studied in the literature. It has become a key tool when making informed decisions regarding the construction of new buildings or expanding existing ones. It has also been applied in different fields, such as determining the optimal size of ventilation systems for a specific building (heating, ventilating and air conditioning [HVAC]), identifying energy consumption balances in the design process, obtaining the best rates for existing buildings, or optimizing energy management systems for buildings [5]. For example, the authors in [6] propose an intelligent data analysis method for the definition of a daily electricity consumption model in buildings to enable the development of a building management system that is able to predict and detect abnormal energy uses. The detection of abnormal energy consumption has also attracted the attention of authors such as Capozzoli *et al.* [7], who propose a statistical pattern recognition model and a swarm of artificial neural networks along with *outliers* detection mechanisms for detecting these faults. We can also find works such as [8], where the authors use a model based on neural networks and waveform analysis to detect faults in different sensors. Meanwhile, the authors in [9] present a model, which they call *strip, bind and search* (SBS), that consists of identifying raw data from different consumption sensors to discover usage patterns between different devices and then monitor the behavior of these devices over time to detect abnormal uses of energy consumption. In [10], the authors use hourly consumption data to define a method for detecting abnormal energy consumption in buildings, which is divided into two parts: the first

is based on classification algorithms and regression trees to classify the data according to their attributes, and the second uses statistical techniques to detect erroneous consumption.

The detection of consumption patterns with which to segment the available data to establish consumption profiles is another of the applications of data analysis that has been widely used by the research community. Specifically, the techniques based on *clustering* appear to have been the most commonly used to address this issue [11–14]. For example, in [15], a cluster-based model is defined, which compares consumption data from different European countries to identify typical consumption patterns of different kinds of consumers between working days and holidays. In [16], the authors employ hierarchical agglomerative *clustering* algorithms (*bottom-up*) to identify days of the week with similar consumption profiles, which are subsequently used to define supervisory control strategies or define methods for detecting abnormal consumption in buildings. However, there are also models that are not based on *clustering* techniques. For example, in [17] the authors establish a framework that is able to establish profiles of energy demand in residential areas by means of a mathematical model that details the relationship between human activity and energy consumption. In addition, use an autoregressive moving average model (ARMA) to detect malicious consumption patterns due to electrical intrusions [18].

Finally, predicting consumer demand is another of the applications that is frequently studied and discussed in the literature. In this context, as indicated in [19], techniques based on artificial intelligence are the most popular, including the use of expert systems, genetic algorithms and systems based on artificial neural networks. In particular, we find a good deal of available work in this area [20–25]. In [26], the authors design a neural network that is based on a supervised multilayer perceptron to predict time series of electricity consumption, taking as a case study the monthly electricity consumption data in Iran. This demonstrates the superiority of the results obtained with this technique over more traditional models, such as the regression model. In [27], the authors propose a model to analyze energy demand in Jordan using a particle swarm optimization technique, which in turn compares the results with those obtained using the backpropagation algorithm and an ARMA.

Increasing the data storage capacity of these systems is a difficult and complex task because it requires an exploratory analysis to obtain a representation of the knowledge implicit in them. In these situations, it is appropriate to use techniques that allow a qualitative representation of the data, such as linguistic summaries. From an informal point of view, they can be seen as a phrase, which is usually short, or as a set of phrases that capture the essence of the data, which tends to be numerical and huge, which makes it difficult to understand [28]. Based on the concept of a protoform in the sense of Zadeh [29], can be formalized as

$$y \text{ is } P \quad (1)$$

where *y* is a syntagma (e.g., *the temperature of*), and *P* is a summary that has an associated linguistic value to *y* (e.g., *high*). Within this field, there are many works related to different disciplines, such as behavioral analysis or the quality in the way people walk [30, 31]; a study to describe the behavior of traffic over time [32]; or, its applicability to a generic set of time series [33]. In works more closely

related to electricity consumption [34], develop a method to generate linguistic summaries based on a set of time series of energy consumption that are provided by an energy supplier. The main difference with our proposal is that they generate these descriptions starting from the time series of a day, without defining a previous model that segments these time data into consumption patterns.

This section has contextualized our study, showing the current perspective of the related concepts, such as the analysis of energy consumption and its applications, and also the linguistic descriptions, the definition of the meter model is introduced in the next section, which is our main proposal to obtain information on consumption in linguistic terms.

3. METER MODEL DEFINITION

Our work is based on a set of data from a geographically distributed Spanish educational institution: the University of Castilla-La Mancha (UCLM). The UCLM consists of a series of buildings and we monitor their electricity consumption through a set of meters M . Each building has its own associated model, which will categorize its consumption and allow us to obtain the conclusions derived from it in linguistic terms. This categorization is resolved by the segmentation of consumption data into groups or classes using a partition-based *clustering* algorithm, which in this case will be k -means. Seen from a formal point of view, a model $M^i \in M$ of a meter i can be defined as a set of k groups that have an associated semantic:

$$M^i = \{G_0, G_1, \dots, G_{k-1}\} \quad (2)$$

where each G_j is defined by a centroid, c_j :

$$G_j = \{c_j\} \quad (3)$$

In our proposal, each meter model M^i will be composed of two groups that will characterize the following consumption patterns: periods with low levels or no activity, which will correspond to G_0 ; and, periods where there is a relevant activity, which will correspond to G_1 . Therefore, the terms {no activity, activity} will represent the semantics associated with the groups of the meter model M^i . To be able to carry out its definition, a study will first be made of the data obtained from each meter and the required temporality of the same will be used to define the yearly models searched. Later, the choice of the number of groups to constitute the models and the selected *clustering* algorithm will be justified. Finally, the obtained meter models will be shown.

3.1. Temporality and Data

Each meter records the hourly energy consumption in the format shown in Table 1: a consumption metric expressed in KW/h (*consumption*), the date on which consumption was recorded (*date*), and a numerical identifier of the meter itself (*id*).

Table 1 | Data format.

Consumption	Date	ID
29.36	20 June 2014 10:00:00	89
33.25	20 June 2014 11:00:00	89
	...	

It is essential to employ an appropriate time range that captures the actual workload obtained after one working day vs calendar day. For example, in the available consumption data, Figure 1(a) shows an interval of 30 hours of consumption of one of the available meters, in which it is possible to observe how this consumption is distributed over time. The workload does not begin at the beginning of the day as such (i.e., at 12:00 am), but rather there is some activity starting at 5:00 am. This situation tends to repeat itself in subsequent periods, as shown in Figure 1(b). Consequently, after appreciating a similar behavior in the other meters, it was decided to use a time window of 24 hours that would begin at 5:00 am, which is therefore our *day* of consumption, and this period is included in the hourly interval of [5:00 am, 4:00 am]. Once our concept of a *day* has been set, the next step is to aggregate the 24-hour set to get the total consumption, x , obtained in the day. First, it is necessary to carry out a cleaning process of the data at the hourly level to avoid any possible measurement errors made by each meter distorting the results obtained when generating the groups of the model M^i , which will be discussed in the following section:

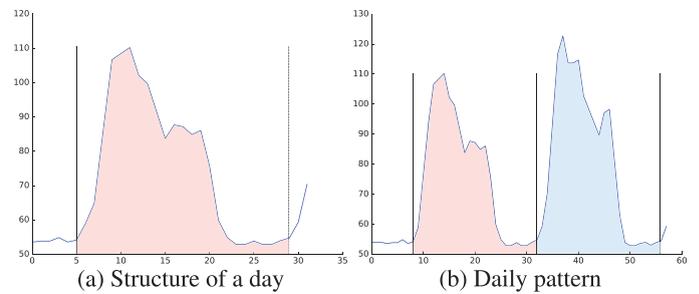


Figure 1 | Time series of hourly consumption.

3.1.1. Data processing

Noise is one of the characteristics that is inherent in practically all nonsynthetic datasets. In this particular case, negative values and consumption data were found to be too large in view of the consumption being recorded by a meter. To mitigate this effect, the first step is to replace all hourly observations whose recorded consumption is negative because this constitutes a failure in the meter reading that is easily detectable. At this point, two options were considered: either eliminate the consumption for that hour or replace its value. Bearing in mind that a day has 24 hours, if one of the hours is eliminated, then this means that this day will no longer be complete and, therefore, the remaining hours that are valid in principle will have to be discarded, thus dispensing with a day of consumption. If this situation occurs with a certain frequency, then there is the risk of losing too many data that are valid *a priori* due to an anomalous observation, which could make it difficult to generate the target groups; that is, G_0 and G_1 . Consequently, we consider that the best option is to replace these wrong observations at the time of capturing consumption, by a value that has a minimum incidence when it comes to grouping data to constitute M^i models. Because the starting data is a negative value, its replacement by a null value (zero) should be the least error to enter at constitute the day, which only serves to maintain the range of 24 hours needed without adding any value.

Detecting and replacing negative consumption data is a technique that can be applied to each meter in a trivial way; however, as has

already been introduced, it is not the only applicable technique. In addition to negative values, each meter may have recorded values where consumption is very high, depending on the type of activity carried out by the building that is monitoring its consumption or the season of the year to which it corresponds. This type of observation can also lead to errors when calculating the consumption of the day, obtaining the same problem as the negative consumption values. Therefore, it is necessary to mitigate the effect of these anomalous observations on each meter independently. This is important to point out because the consumption between buildings varies—a value that would be excessive in one building, may be normal in another. Consequently, it is necessary to establish a time interval, in hours, that is small enough to detect only alterations in consumption on dates that are related and is large enough to appreciate such variations. After experimentation, it was concluded that the most suitable period to detect these changes in consumption was 7 days, where each day will contain 24 hours of consumption in the interval [5:00 am, 4:00 am] or 168 hours. On this last data window, a statistical technique called interquartile ranges or IQR is applied to detect anomalous observations, which are replaced by a linear interpolation between contiguous observations because we assume that the consumption recorded at a specific time by a meter will be similar to the consumption recorded near that time.

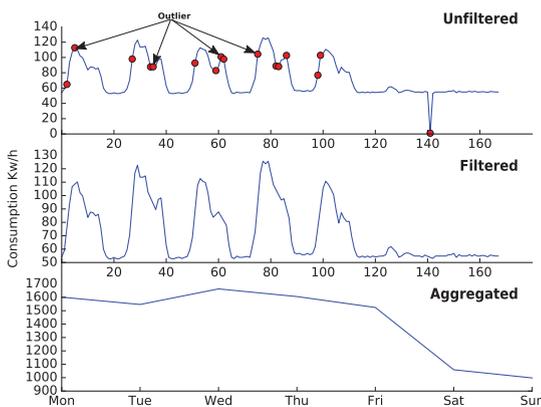


Figure 2 | Transformation of time series.

Figure 2 shows the different transformations that a time series undergoes. In the «unfiltered» view, raw data is displayed as they are captured by each meter. It illustrates the different points that are detected as *outliers* according to the IQR technique, such as in the daily peaks there is an abrupt decrease in consumption to pick up again (hours 38, 59, 85), or in the hour 140 there is a very far value with respect to the consumption obtained in the near future. In contrast, the «filtered» view shows the same series without these anomalous observations. The valleys that were previously highlighted now appear much softer, in accordance with the consumption that is obtained in near dates, and the peak obtained in the hour 140 is now much closer to how the consumption should have been. Finally, the «aggregated» view shows the aggregation of the filtered time series to obtain the total daily consumption, the latest type being the time series data set used to generate the models. These models will be obtained using the *k*-means algorithm. Both this and other partition-based algorithms have one common feature: to provide the number of *k* groups in advance to perform *clustering*. This section has already demonstrated that the proposed M^i meter models will have two groups, or what is the same, $k = 2$. Therefore, to

justify this decision, in the following section a probabilistic study is carried out whose conclusion is based on the available consumption data. The results will be compared with those produced by a series of heuristic techniques.

3.2. Hyperparameter Tuning for *k*-Means

The *k*-means is one of the best-known partitioning algorithms that is used in a wide variety of fields [35] due to its ease of implementation and the quality of the results obtained. However, despite its popularity, it suffers from certain limitations that should not be overlooked. As has already been explained, one is to choose the appropriate *k*, in addition to the incidence of the initialization process on the quality of the results obtained [36, 37]. A simple technique to try to mitigate this problem is to run the algorithm *n* times for a *k* given with different initial partitions, selecting the one that minimizes the sum of squared errors [38]. However, this carries with it the difficulty of discerning whether *n* is adequate or not. Among the existing variants of *k*-means, there is the so-called *k*-means++ [37], which tries to select *k* centroids randomly, one at a time, with a probability proportional to the squared distance with the nearest centroids already selected. Erisoglu *et al.* [37] point out that the overall performance is better than *k*-means pure, both in the quality of the result and in the execution. Consequently, it has been decided to use *k*-means++ to the detriment of *k*-means as an algorithm to generate the proposed meter models.

3.2.1. Selection of *k*

Determining the number of groups (*k*) enables us to associate a descriptive semantics with energy consumption.

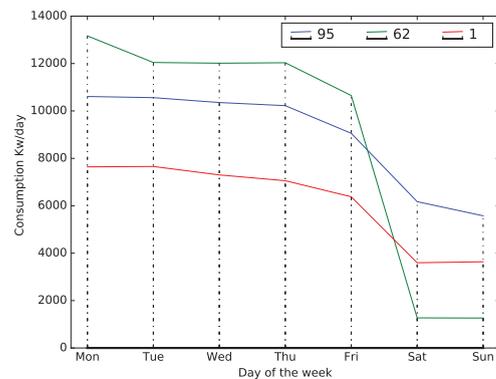


Figure 3 | One week's form of aggregate consumption per day.

The aim is to achieve a balance between the semantics searched and the error made in the process of *clustering*. It is well known that the greater the *k*, the more compact the groups will be and, therefore, the smaller the error will be. However, a high value of *k* does not always bring greater meaning to the final result. Figure 3 shows the distribution of aggregate consumption per day of the week for a full month of three consumption meters, whose identifiers (*id*) are 1, 62, and 95.

At first, we can see a pattern that is repeated throughout the week: a period where consumption is accentuated, drawing a «crest», followed by a «valley» where it decreases to a point where consumption is quite flat and there are no great variations. This suggests that a $k = 2$ can provide an adequate consumption categorization to the form described by the series. To confirm this hypothesis, let us consider the generation of two groups, with $k = 2$, where G_0 will refer to the categorization of consumption of no activity, and G_1 to that of activity. Let W the set of all working days:

$$W = \{Monday, Tuesday, Wednesday, Thursday, Friday\} \quad (4)$$

and H the set of holidays:

$$H = \{Saturday, Sunday\} \quad (5)$$

Also, consider the following set of conditional probabilities:

$$P(W/G_j) = P(Monday/G_j) + P(Tuesday/G_j) + \dots + P(Friday/G_j) \quad (6)$$

$$P(H/G_j) = P(Saturday/G_j) + P(Sunday/G_j), \quad (7)$$

where $P(W/G_j)$ refers to the probability that the set of working days is in a group G_j , and $P(H/G_j)$ to the probability of holidays, with $j \in \{0, 1\}$. Thus, our hypothesis can be considered valid if and only if

$$[P(W/G_0) < P(H/G_0)] \wedge [P(W/G_1) > P(H/G_1)] \quad (8)$$

To determine the effectiveness of the premise of Equation (8), we are going to use the result of the *clustering* of one of the meters shown in Figure 3, Meter 62, whose distribution of days per cluster is shown in Table 2. Therefore, the probability that the activity consumption (W) will be in the first group (G_0) will be

$$P(W/G_0) = \frac{11}{150} + \frac{10}{150} + \frac{11}{150} + \frac{13}{150} + \frac{17}{150} = \frac{62}{150} = \mathbf{0.413}$$

Table 2 | Distribution of days per group.

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	
G_1	34	35	34	32	28	1	1	165
G_0	11	10	11	13	17	44	44	150
	45	45	45	45	45	45	45	315

The remaining probabilities are listed in Table 3. Based on the results, we can conclude that as $[P(W/G_0) < P(H/G_0)]$ and $[P(W/G_1) > P(H/G_1)]$, our premise is fulfilled and, therefore, our initial hypothesis about the assumption of the number of groups to choose is valid.

Table 3 | Conditional probabilities.

$P(W/G_0)$	$P(H/G_0)$	$P(W/G_1)$	$P(H/G_1)$
0.413	0.586	0.987	0.012

However, this conclusion has only been shown to be valid for one of the meters. Therefore, we are going to extend our hypothesis Equation (8) for all available consumption meters, which in this case amounts to a total of 97. Let

$$\mathbb{E}_i [P(W/G_j)] \quad \forall i \in [0, 96] \quad (9)$$

the mathematical expectation for the probability that the set of workdays is in a G_j group for every meter i , and let

$$\mathbb{E}_i [P(H/G_j)] \quad \forall i \in [0, 96] \quad (10)$$

mathematical expectation for the set of holidays, with $j \in \{0, 1\}$. In this way, our extended hypothesis can be considered valid if and only if

$$\mathbb{E}_i [P(W/G_0)] < \mathbb{E}_i [P(H/G_0)] \wedge \mathbb{E}_i [P(W/G_1)] > \mathbb{E}_i [P(H/G_1)] \quad (11)$$

It is worth noting that mathematical expectation will be estimated through arithmetic mean. So, as in the previous case, to determine its validity, we will make use of the result of the *clustering* for each of the consumption meters, where a subset of the probabilities already calculated can be seen in Table 4. For example, following the three meters shown in Figure 3, for Meter 1, we have $P(W/G_0) = 0.148$ is less than $P(H/G_0) = 0.852$, and that $P(W/G_1) = 0.960$ is greater than $P(H/G_1) = 0.040$, conditions necessary and sufficient to confirm the hypothesis for selecting two groups. The same applies to Meter 95.

Table 4 | Subset of probabilities.

Meter	$P(W/G_0)$	$P(H/G_0)$	$P(W/G_1)$	$P(H/G_1)$
0	0.319	0.681	0.972	0.028
1	0.148	0.852	0.960	0.040
2	0.324	0.676	0.972	0.028
3	0.422	0.578	0.969	0.031
4	0.459	0.541	0.993	0.007
5	0.452	0.548	1.000	0.000
6	0.309	0.691	0.972	0.028
...				
90	0.415	0.585	0.969	0.031
91	0.363	0.637	0.924	0.076
92	0.392	0.608	0.970	0.030
93	0.381	0.619	0.970	0.030
94	0.324	0.676	0.972	0.028
95	0.289	0.711	0.973	0.027
96	0.000	1.000	1.000	0.000

Thus, the mathematical expectation that the activity consumption (W) will be in the first group (G_0) will be of

$$\mathbb{E} [P(W/G_0)] = \frac{0.319 + 0.148 + \dots + 0}{97} = 0.359 \quad (12)$$

The remaining expectations are shown in Table 5, where the new extended hypothesis Equation (11) is also fulfilled for the total of consumption meters and, therefore, we can conclude the number of groups we had assumed explains the semantics that we have defined.

Table 5 | Probabilities of expectations.

$\mathbb{E} [P(W/G_0)]$	$\mathbb{E} [P(H/G_0)]$	$\mathbb{E} [P(W/G_1)]$	$\mathbb{E} [P(H/G_1)]$
0.359	0.641	0.928	0.072

Through this probabilistic study, it has been demonstrated that the initial hypothesis of categorizing consumption data into two groups with which to generate the M^i meter models is adequate. To reinforce this conclusion, we will check if we obtain the same conclusions using two of the most widely used heuristic techniques in the literature [5, 39–43], *gap* technique [44] and *silhouette* [45]. Finally, we draw the same conclusions. To carry out this test, we compare the results obtained by each of these techniques applied on the same three consumption meters as in the previous case: 1, 62, and 95.

Table 6 | Optimal number of groups.

k	95		62		1	
	<i>sil</i>	<i>gap</i>	<i>sil</i>	<i>gap</i>	<i>sil</i>	<i>gap</i>
2	0.763	-16.08	0.650	-14.09	0.624	-14.68
3	0.765	-15.65	0.585	-14.23	0.595	-14.90
4	0.729	-15.71	0.628	-13.96	0.637	-14.66
5	0.717	-15.73	0.615	-14.01	0.580	-14.86
6	0.644	-15.86	0.616	-14.05	0.592	-14.93
7	0.650	-15.93	0.601	-14.10	0.601	-14.83

Highlighted in **bold** is the best value obtained for each technique in each meter.

Table 6 shows for each k the value achieved by the *gap* technique (column *gap*) and *silhouette* (column *sil*) for each consumption meter. For example, it can be observed that for Meters 1 and 95, both techniques show that the best possible partition is $k = 3$ and $k = 4$, respectively; however, for Meter 62, there is no agreement. Then, based on the results obtained, two problems can be seen:

1. There is no agreement between the techniques, so the choice of one or the other depends on the expert who is conducting the study.
2. Having solved the previous problem of choosing the heuristic technique, there is also no consensus when it comes to choosing the k for each consumption meter, which makes it impossible to extrapolate the k obtained for one of the meters to the other meters.

Therefore, the conclusion that emerges is that for a set of heterogeneous consumption time series, as is the case in this paper, trusting in the result of this type of algorithmic solutions is not appropriate because the choice of the k depends on the semantics that we want to assign to the models to be defined, which are shown in the following section:

3.3. Results

To define the M^i meter models, as introduced at the beginning of this section, the consumption data of an educational institution has been gathered from UCLM. Specifically, the period of data available is from 2011 to 2017. Based on this data, we can obtain up to a total of six different M^i meter models, one for each year from 2011 to 2016, because the consumption data corresponding to 2017 will be used to compare the last model (i.e., 2016) with the current consumption level of the organization when generating the associated linguistic summaries. A small sample of the prototypes obtained by the process of *clustering* for each of the consumption meters is shown in Table 7, which uses the aggregated series to obtain the daily total for each of the available annual periods. It can be observed that some of these periods, such as 2011 or 2012, have no consumption data available and, therefore, there is no generation

of groups. This is due to the gradual introduction of consumption meters over the years on the different buildings at UCLM. The only limitation that this implies is the impossibility of using the models of those years for certain meters as reference elements to find if the consumption that we obtained is what was expected.

Table 7 | Clustering result on each meter i .

	Meters							
	0		1		95		96	
	G_0	G_1	G_0	G_1	G_0	G_1	G_0	G_1
2011	411.46	1006.56	420.45	791.84	-	-	-	-
2012	518.99	965.09	455.23	837.04	-	-	-	-
2013	497.26	1044.59	603.39	1095.65	-	-	-	-
2014	535.37	1064.36	487.51	954.44	261.98	793.14	-	-
2015	735.50	1748.16	498.11	1164.04	297	1965.11	-	-
2016	500.44	1215.19	570.23	1213.66	276.31	1994.66	211.07	406.82

As already mentioned, the 2017 consumption data will be used to measure the error made by our models, which will be defined according to the *clustering* results of 2016. If we use the data from Table 7, then the model definition for Meter 0, for example, will make use of the model definition given in Equation (2), which will be given by $M^0 = \{G_0, G_1\}$, where G_0 has an associated semantics of no activity, and G_1 of activity, whose prototype or most representative consumption is $G_0 = \{500.44\}$ and $G_1 = \{1215.19\}$, respectively. This gives an actual daily consumption of x , which tells us which group it belongs to by means of a distance measurement. Depending on the associated semantics, we will be able to summarize in linguistic terms the consumption situation of a meter i with respect to the total obtained by the organization, whose definition and generation process will be discussed in the next section.

4. DEFINITION OF ORGANIZATIONAL MODEL

The concept of a meter model M^i has been introduced in the previous section. By themselves, they do not provide enough information to be able to draw conclusions regarding actual consumption being obtained in a given period. For example, if you have an actual consumption of $x = 675$ KW/day for Meter 0, whose model is determined by $M^0 = \{500.44, 1215.19\}$, the maximum that you can get is x belonging to the semantic group, but it is not possible to know if this level of consumption is high, moderate, and so on. Therefore, in fact, each M^i meter model is used as a base to establish an organizational model that can serve as a comparative environment for each of the meter models. Thus, an organizational model or metamodel is a model of models composed of each of the meter sub-models M^i , which by means of a linguistic description derived from it, provides a summary of the consumerist situation in which the organization is situated with respect to itself in previous years:

$$O = \left\{ \begin{array}{l} O_{G_0} = \{M_{G_0}^0, M_{G_0}^1, \dots, M_{G_0}^{|M|}\} \\ O_{G_1} = \{M_{G_1}^0, M_{G_1}^1, \dots, M_{G_1}^{|M|}\} \\ \vdots \\ O_{G_k} = \{M_{G_k}^0, M_{G_k}^1, \dots, M_{G_k}^{|M|}\} \end{array} \right\}$$

Based on this definition of the organizational model O , conclusions can be drawn about the consumption of each meter model M^i . On the basis of this, a linguistic summary based on fuzzy techniques can be established, which will be discussed below.

4.1. Linguistic Categorization of Each Meter in the Environment O

The organizational model O , for our case study, is defined by

$$O = \left\{ \begin{array}{l} O_{G_0} = \{M_{G_0}^0, M_{G_0}^1, \dots, M_{G_0}^{96}\} \\ O_{G_1} = \{M_{G_1}^0, M_{G_1}^1, \dots, M_{G_1}^{96}\} \end{array} \right\}$$

This metamodel O has an associated domain corresponding to the set of prototypes of no activity:

$$Dom(O_{G_0}) = \{500.44, 570.23, \dots, 211.07\}$$

and another domain corresponding to the set of activity prototypes:

$$Dom(O_{G_1}) = \{1215.29, 1213.66, \dots, 406.82\}$$

Thanks to the domains of the different groups that make up the organizational model, it is possible to define a characterization of the consumptions obtained by each meter through the calculation of percentiles because they allow us to know the positioning of each meter i in terms of consumption, with respect to the total of the organization. For example, if you have an actual consumption of $x = 750$ KW/day, and it is known that it is in the 60th percentile (P_{60}) of the organizational model O , then 60% of the meters consume the same or less than x . Therefore, the next step is to define this conclusion or characterization of consumption in linguistic terms to summarize the consumption of each meter i with respect to the organization's consumption in natural language using the definition of protoforms:

$$y \text{ is } P$$

Based on this definition of protoforms, y will be equal to *consumption*, and the P summary will be defined in fuzzy terms. Now with P , it can be compared the consumption of a meter with the consumption of the organization.

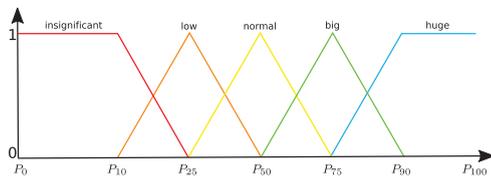


Figure 4 | Linguistic variable.

To formalize this summary P , we use a linguistic variable \mathcal{L}_v (Figure 4) composed of five values, which make up a fuzzy partition of the domain whose universe of discourse is given by all the prototypes that make up each domain (G_0 and G_1) of the metamodel O :

$$Dom(\mathcal{L}_v) = Dom(O_{G_j}), \tag{13}$$

with $j \in \{0, 1\}$. The values of the linguistic variable \mathcal{L}_v are defined by the fuzzy sets *insignificant*, *low*, *normal*, *big*, and *huge*, whose membership function is given by three parameters P_r which represent the value of the percentile r (Table 8). Figure 5 shows the membership functions considered for the definition of the values of \mathcal{L}_v .

Table 8 | Definition of fuzzy sets.

Label	μ
Insignificant	$R\{P_0, P_{10}, P_{25}\}$
Low	$T\{P_{10}, P_{25}, P_{50}\}$
Normal	$T\{P_{25}, P_{50}, P_{75}\}$
Big	$T\{P_{50}, P_{75}, P_{90}\}$
Huge	$L\{P_{75}, P_{90}, P_{100}\}$

Once the P summary is formalized, the last step before the linguistic summary can be built is to identify the domain of O to which a given x consumption of a specific i meter belongs to be able to apply the \mathcal{L}_v linguistic variable in the appropriate magnitudes. To do this, starting from the meter model M^i , we obtain the semantic group to which this consumption x belongs through the following formula:

$$\arg \min dist(x, G_j), \quad j \in \{0, 1\} \tag{14}$$

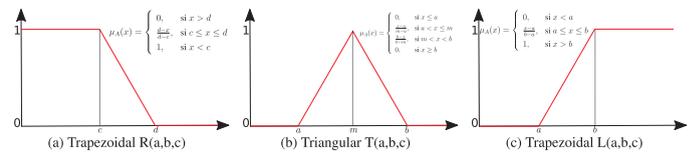


Figure 5 | Fuzzy membership functions used.

where $dist(x, G_j)$ is a distance measure that quantifies the similarity between consumption x and the prototype of the group G_j , typically the Euclidean. Knowing the semantic group to which x belongs, we can obtain the adequate domain of O and apply the linguistic variable \mathcal{L}_v to obtain the linguistic label that will define the summary P applying the operation of the maximum t-conorm. If the interest lies in obtaining information about the consumption of a meter i compared to the estimation determined by its model M^i , then the resulting protoforms will be of the form:

«Consumption is *insignificant*, *low*, *normal*, *big*, or *huge*.».

where y is the syntagma *consumption* and P is *insignificant*, *low*, *normal*, *big*, or *huge*. Meanwhile, if the interest lies in knowing how a given meter model M^i is placed in relation to the other submodels of the organization O , then the resulting protoforms will be of the type:

«My activity/no activity consumption is *insignificant*, *low*, *normal*, *big*, or *huge*.».

where y is the syntagma *My activity/no activity consumption* and P is *insignificant*, *low*, *normal*, *big*, or *huge*. To illustrate this construction process, a concrete example will be presented below. Let us assume that for Meter 1 (Table 7) we have an actual consumption of $x = 321$ KW/day and we want to understand this consumption with respect to the organization. To do this, it is first necessary to

identify the domain of O on which to apply the linguistic variable \mathcal{L}_v . Making use of Equation (14) we have

$$\arg \min \{dist(321, 570.23), dist(321, 1213.66)\} = 0$$

which indicates that the recorded consumption is in the semantic group of no activity (G_0). Knowing this semantic group, you get that the domain of O on which to apply \mathcal{L}_v is $Dom(O_{G_0}) = \{500.44, 570.23, \dots, 211.07\}$. In this way, we obtain x membership to each fuzzy set defined in \mathcal{L}_v , and by means of the maximum t-conorm we obtain the set to which x belongs:

$$\max \{ \mu_{insignificant}(x) = 0, \mu_{low}(x) = 0.16, \mu_{normal}(x) = 0.84, \mu_{big}(x) = 0, \mu_{huge}(x) = 0 \} = \mu_{normal}(x)$$

Once the fuzzy set is available, the construction of the linguistic summary would look like consumption is *normal*. In contrast, if you want to know how the model of Meter 1, M^1 behaves with respect to the organization model O , then we must first choose the centroid or prototype c_j that we want to compare. Suppose that we choose the centroid in the no activity group ($G_0 = \{570.23\}$). In this case, it is not necessary to make use of Equation (14) to obtain the semantic group because it is already available with the selection of centroid. As previously, the domain of O on which to apply \mathcal{L}_v is $Dom(O_{G_0}) = \{500.44, 570.23, \dots, 211.07\}$. Again, by means of the maximum t-conorm we obtain the set to which c_j belongs:

$$\max \{ \mu_{insignificant}(x) = 0.31, \mu_{low}(x) = 0.69, \mu_{normal}(x) = 0, \mu_{big}(x) = 0, \mu_{huge}(x) = 0 \} = \mu_{low}(x)$$

As in the previous case, once we have obtained the fuzzy set, the construction of the linguistic summary would be as follows: consumption of *no activity* is *low*.

The linguistic summaries that we have seen up to now provide us with the following information for each meter i , during the period of validity of the model: 1. Regarding its actual consumption x , how it behaves as estimated by its model M^i . 2. Regarding your model M^i , how it behaves with respect to the other sub-models of O . However, we lack the necessary information about the error made in the estimation of the meter model M^i to be able to draw conclusions about its goodness; that is, if our model effectively reflects the underlying consumption patterns and, therefore, provides us with adequate conclusions in linguistic terms, or if, in contrast, it tends to be too pessimistic or optimistic in its estimations, drawing incorrect conclusions. Therefore, the following section will propose a mechanism to get a linguistic description of the error made by the models in linguistic terms.

5. LINGUISTIC DESCRIPTION OF THE ERROR

In this section, a framework will be described that will allow us to get the linguistic description of the error of each one of the M^i meter models defined, in addition to the organizational model O . In this way, it is possible to know if the defined model is able to capture the consumption patterns of each meter and of the organization and, therefore, to draw the appropriate conclusions in linguistic terms.

5.1. Description of the Meter Model M^i Error

To carry out the linguistic description of the error of each meter model M^i , it is necessary to associate the actual consumption x obtained in an instant of time in each meter i , with the semantic group G_j of the model that best defines it to obtain the error made in the estimation of the model by actual consumption day. This semantic group will be called G , and it will be the prototype that best describes actual consumption x . Formally, G is defined in general terms by the following function f , which combines the different membership degrees of different groups:

$$\tilde{G} = f(\mu_{G_0}(x), \mu_{G_1}(x), \dots, \mu_{G_{k-1}}(x)) \quad (15)$$

In the particular case that we are dealing with, two semantic groups ($k = 2$), G can be defined as G_h in terms of traditional set logic:

$$G_h \equiv f(\mu_{G_0}(x), \mu_{G_1}(x)), \quad (16)$$

where $h = \arg \max \{ \mu_{G_0}(x), \mu_{G_1}(x) \}$ and $\mu_{G_j} = dist(x, G_j)$; or it can also be defined in fuzzy logic terms:

$$f'(\mu_{G_0}(x), \mu_{G_1}(x)) = G_0 \times \mu_{G_0}(x) + G_1 \times \mu_{G_1}(x) \quad (17)$$

where $\mu_{G_j} = [0, 1]$. This membership is obtained by a fuzzy partition of the domain of the meter model M^i by means of two fuzzy sets (Table 9), one for each semantic group G_j .

Table 9 | Fuzzy sets over M^i .

Label	μ
No activity	$R\{0, G_0, G_1\}$
Activity	$L\{G_0, G_1, G_1 + G_0\}$

Note that while in Equation (16), \tilde{G} will be equal to one of the G_j groups defined in the meter model M^i , Equation (17) gives rise to a new prototype obtained by linear adjustment of the two existing groups, which gives us a finer-grain when cataloging the error made. For this reason, we use Equation (17) to obtain \tilde{G} . Once calculated, the error obtained when cataloging the real consumption x in one of the semantic groups will be given by

$$\epsilon = \tilde{G} - x \quad (18)$$

which gives us a value that is independent of the chosen scale, allowing us to compare the error obtained for each semantic group \tilde{G} . Consequently, to classify whether this error made with respect to the estimation of its meter model M^i is meaningful, with a confidence level of 95%, we consider the interval defined by $\pm 2\sigma$, with σ being the standard deviation the corresponds to the semantic group G_j to which belongs the actual consumption x , so that

- If $\epsilon > 2\sigma$, then we have predicted a higher consumption than the real one and, therefore, it is **overestimated**.
- If $-2\sigma \leq \epsilon \leq 2\sigma$, then we have predicted a consumption that is in line with the real one and, therefore, it is **adequate**.

- If $\epsilon < -2\sigma$, then we have predicted a lower consumption than the real one and, therefore, it is **underestimated**.

It is worth noting that with the application of Equation (17), because \tilde{G} is a semantic group that does not exist in the generation of the model, it lacks a standard deviation as such. For this reason, to classify the error made, we consider that σ will be the standard deviation of the semantic group G_j closer to \tilde{G} because it is considered to have a greater influence in the calculation of \tilde{G} , and thus the deviation of real consumption x with respect to the prototype G_j will be equivalent to the deviation between x and \tilde{G} .

With the ϵ error made by each real consumption x , the interest now focuses on categorizing the relevance of this error in terms of the meter model M^i —that is, if it is placed in the defined margins as adequate or if it is placed in the 5% of the remaining observations—to obtain a linguistic description that sums up the error of the estimated model. To do this, we will use the same linguistic variable \mathcal{L}_v (Table 10) that we defined in Figure 4 with one caveat: its domain.

Table 10 | Fuzzy sets on the goodness of the model.

Label	μ
Insignificant	R{0, 10, 25}
Low	T{10, 25, 50}
Normal	T{25, 50, 75}
Big	T{50, 75, 90}
Huge	L{75, 90, 100}

In this case, the domain will be given by the number of observations (errors) that fit into each category described above, according to the following formula:

$$L_j = \frac{|\epsilon_j|}{|\epsilon|} \times 100, \tag{19}$$

where $j \in \{overestimated, adequate, underestimated\}$, $|\epsilon_j|$ will be the number of errors that fit into one of the defined categories, and $|\epsilon|$ will be the total amount of error made.

Table 11 | Number of errors per category of the Meter 0.

Overestimated		Adequate		Underestimated		
L_j	μ	L_j	μ	L_j	μ	
	1	0	0	0	0	Insignificant
	0	0	0	0	0	Low
9%	0	91%	0	0%	0	Normal
	0	0	0	0	0	Big
	0	1	0	0	0	Huge

Given that there are three categories—overestimated, adequate, and underestimated—a linguistic description will be obtained for each, in addition to a global one that summarizes briefly and concisely if our model is still valid for the actual consumption obtained in the current period. For example, in Table 11 you can see the number of errors in each category for Meter 0 with the associated membership of each label defined by the linguistic variable \mathcal{L}_v for actual consumption in 2017 according to its model of the year 2016. Based on these data, the following linguistic summaries have been obtained:

«The model is **underestimating** the consumption in an insignificant way.»
 «The model is **adequate** in a huge way.»
 «The model is **overestimating** the consumption in an insignificant way.»

To obtain the linguistic description that summarizes in a global way the error made by the meter model M^i , we keep the one whose L_j is maximal. In this case, it would be

«The model is **adequate** in a huge way.»

The indicator to know if our model begins to suffer failures in the estimation of consumption, is determined by the category and the P linguistic summary associated to the description. This means that if the model is underestimating or overestimating consumption, we must look at the value of P to get an accurate conclusion. Thus, if the linguistic label associated with P is less than normal, then we can assert that the linguistic descriptions of the error provided are accurate; otherwise, the model should be updated. Until now, a framework has been established to get a linguistic description of the error made from each meter model present in the organizational model; however, nothing is known about the error made by the latter. Consequently, the following section will extend the concepts presented here so that they can be applied in the organizational model O .

5.2. Description of the Organizational Model O Error

The concepts seen in the meter models can be extended to determine the goodness of the organizational model O . To reach this goal, each category must be aggregated: *underestimated*, *adequate*, and *overestimated* for each meter model M^i that makes up the meta-model O and we must apply the same \mathcal{L}_v linguistic variable as in the M^i meter models. This aggregation, L'_j , is defined by the following equation:

$$L'_j = \frac{\sum L_j}{|L_j|} \tag{20}$$

Table 12 shows the error made in each meter model by category, including their aggregation according to Equation (20).

Table 12 | Number of errors per category of each M^i .

Meter	L_{over}	L_{adequate}	L_{under}
0	9%	91%	0%
1	4%	96%	0%
2	5%	95%	0%
3	12%	88%	0%
4	14%	82%	4%
5	7%	93%	0%
6	13%	87%	0%
		...	
90	10%	90%	0%
91	0%	61%	39%
92	9%	91%	0%
93	20%	80%	0%
94	17%	83%	0%
95	13%	87%	1%
96	23%	77%	0%
	9.35%	87.14%	3.55%

Highlighted in **bold** is the best value obtained for each technique in each meter.

The memberships of each error aggregated by category, L_j , to the linguistic tags defined in \mathcal{L}_v can be seen in Table 13. Based on these results, the following linguistic summaries are obtained:

- «The organizational model is **underestimating** consumption in an *insignificant* way.»
- «The organizational model is **adequate** to consumption in a *huge* way.»
- «The organizational model is **overestimating** consumption in an *insignificant* way.»

Table 13 | Number of errors by category of organizational model O .

Overestimated		Adequate		Underestimated		
L_j	μ	L_j	μ	L_j	μ	
	1		0		1	Insignificant
	0		0		0	Low
9%	0	87%	0	4%	0	Normal
	0		0.2		0	Big
	0		0.8		0	Huge

We highlight the use of the term *organizational* to distinguish the linguistic description of the O metamodel from those of each of the M^i meter models that make it up. Again, as in the previous case, the protoform that summarizes in linguistic terms the error made by the organizational model O will be given by that L_j added error that is maximal, which in this case turns out to be

- «The organizational model is **adequate** to consumption in a *huge* way.»

Throughout this section, a method has been suggested to evaluate each proposed model using fuzzy linguistic summaries. The use of fuzzy sets for their definition offers the possibility of working with little defined limits, so that the conclusion derived from them can be used with *some degree of membership* of several of these sets. When this is the case, the linguistic summaries based on the protoforms presented in the Section 2 are not sufficiently expressive to highlight this situation. Therefore, in the following section, we introduce a new concept of extended summary that allows us to draw conclusions in linguistic terms whose degrees of fuzzy membership are not very marked.

6. EXTENDED LINGUISTIC SUMMARIES

To reflect a case study in which classic protoforms do not provide enough capacity to capture the semantic hints of a given consumption situation, Table 14 provides a comparison of the membership to the fuzzy sets defined over the linguistic variable of Figure 6 of the same day of actual consumption for two different meters: 0 and 96.

In Meter 0, we have very marked membership levels to the two groups involved (normal and big), so the linguistic summary «consumption is *big*» provides us with a conclusion that does not give rise to doubt. However, the same does not apply to Meter 96. In this case, we have two sets whose memberships are very close (low and normal). If we were to follow the operation described above, then

Table 14 | Comparison of memberships in meters.

Label	Fuzzy membership (μ)	Meter
Insignificant	0	0
	0	96
Low	0	0
	0.455	96
Normal	0.003	0
	0.544	96
Big	0.996	0
	0	96
Huge	0	0
	0	96

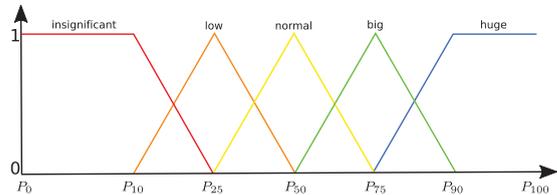


Figure 6 | Linguistic variable.

the most appropriate linguistic description would be «consumption is *normal*». However, linguistic summaries must be expressive enough to avoid masking information that leads to inaccurate conclusions, as would be the case here.

To try to capture this idiosyncrasy through linguistic summaries, we propose a modification of the protoforms shown in Section 2:

$$y \text{ is } P$$

allowing the addition of an absolute quantifier (*close to, near,*) [46], W , to the P summary, resulting in an extended summary P' :

$$P' = WP \tag{21}$$

This extended summary is able to model the linguistic description in terms of two linguistic labels with a *nexus* that provides the appropriate semantics. Returning to the example of Meter 96, if we use this extended summary, then the linguistic summary would be described as follows:

«consumption is being normal close to low.»

Thus, the protoform of Equation (1), making use of this extended summary P' , will, respectively, be defined as

$$y \text{ is } P' \tag{22}$$

The criteria to determine whether the addition of the W quantifier to the P summary is necessary to obtain a description that captures this kind of cases in linguistic terms will be given by a membership threshold δ associated with each fuzzy set, so that: if the membership value μ of a consumption element x is lower than this threshold, for example, $\delta = 67\%$, then it is necessary to use an extended summary P' ; otherwise, a summary P is enough.

$$\left\{ \begin{array}{l} P', \quad 0 \leq \mu(x) < \delta \\ P, \quad \delta \leq \mu(x) \leq 1 \end{array} \right\} \tag{23}$$

7. EXPERIMENTAL RESULTS

In this section, we will show the expressive capacity provided by linguistic summaries when summarizing the state of energy consumption of an organization, the UCLM, at a given moment starting from its model. This allows us to analyze specific situations in a legible and intuitive way with to undertake possible corrective actions.

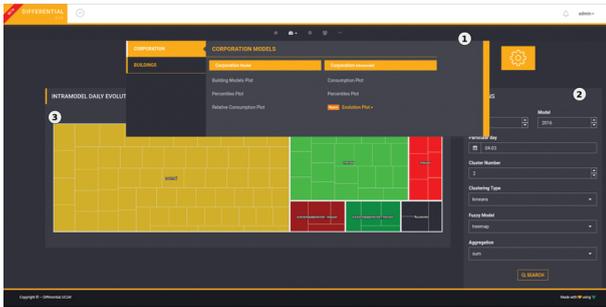


Figure 7 | Overview of the software prototype.

To support the models, a software prototype has been designed and implemented that enables us to monitor and notify the consumption obtained at different management levels of the institution, such as managers and those responsible for the administration of each building. This prototype has been visually structured in three well-differentiated areas, as can be seen in Figure 7: 1. This enables navigation between the different views generated from the models, and certain aspects of administration and management inherent to the application. 2. This shows the options available to select the time ranges with which the model and consumption are to be confronted, and also the type of aggregation resolved on the data, which at this point, only one aggregation is allowed to obtain daily total consumption, and the fuzzy model that will describe in linguistic terms the behavior of the model. 3. Displays different graphical views based on the defined models. The graphical view shown in the Figure 7 provides an organizational level view of the model's behavior for a particular consumption day. This view has been resolved by means of a treemap, in which each category, represented as big squared tiles, corresponds to the number of i meters grouped in each fuzzy set defined using the \mathcal{L}_t linguistic variable shown in Figure 8, where each nested tile within each category represents an i meter, and whose size is determined by the value of applying Equation (24) over \mathcal{L}_t , displaying itself in a proportional way with respect to the value of the rest of the tiles within the same category. Furthermore, the «squarified» shape is due to the algorithm used, which keeps each rectangle as square as possible in order to make a more compact view for displaying category hierarchies.

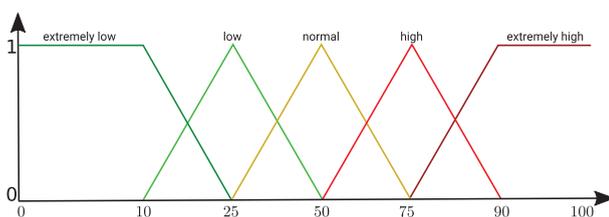


Figure 8 | Definition of the linguistic variable \mathcal{L}_t .

Thus, we have also added a new set that is not defined in \mathcal{L}_t to this variable, which is called absent. Its usefulness lies in categorizing those meters that do not have data of their real consumption for a given day, so that they can be excluded from the conclusions derived from the view. To categorize each meter i in \mathcal{L}_t , we calculate the error made when associating the real consumption obtained x in its meter model M^i according to Equation (18):

$$\epsilon = \bar{G} - x$$

In contrast to the proposal to describe the model error that was presented earlier, where three possible cases were defined depending on whether the error committed was adequate, overestimated, or underestimated, and linguistic summaries were constructed that defined the situation of consumption of the organization, in this case we let the fuzzy sets that expose this casuistry according to the membership values obtained in each case. To do this, we use the following equation:

$$\gamma = \frac{\epsilon - 2\delta}{4\delta} \times 100 \tag{24}$$

which maps each value of ϵ in the interval formed by $\pm 2\sigma$, where σ is the standard deviation of the semantic group G_j to which belongs the actual consumption x , and whose domain will be defined by $\gamma \in [0, 100]$. Thus, the value of γ is submitted to \mathcal{L}_t , which allows us to obtain the fuzzy set in which the meter i is categorized for a specific consumption day x :

$$x_i \in \mathcal{L}_t(\gamma) \tag{25}$$

and therefore, the organizational consumption will be determined by the category with the highest number of meters. For example, Figure 9 shows how the model for 22 June 2017 behaves, which was the day when the electrical consumption of the buildings was much higher than expected. This day corresponds to a day of normal work activity, and the buildings that reported the greatest consumption were those whose main activity is teaching and research. The cause of this spike may be due to the fact that very high temperatures were recorded on that day and, therefore, the air conditioning systems were working most of the day. Thus, if we apply the linguistic summaries that we defined in the previous section, we can conclude that

«The consumption of UCLM, with respect its model, is extremely high.»

because the category with the largest number of meters, according to the treemap, is the one defined as extremely high.

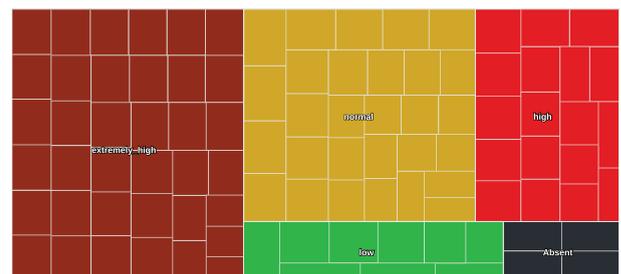


Figure 9 | Treemap of consumption on 22 June 2017.

In contrast, we find that on 10 August 2017 electricity consumption was much lower than expected (Figure 10). This day corresponds to the summer holidays period, which is characterized by the closure of all the buildings, so that there is hardly any activity beyond the residual consumption of the meters themselves, the servers, emergency lights, and so on.

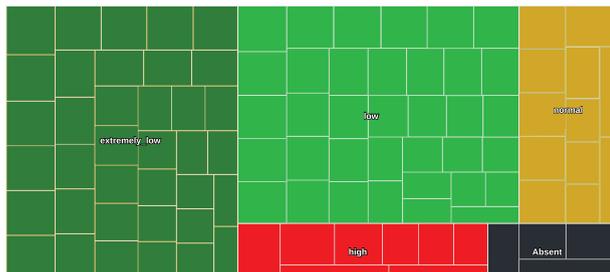


Figure 10 | Treemap of consumption on 10 August 2017.

In this case, if we apply the linguistic summaries that we defined in the previous section, we can conclude that

«The consumption of UCLM, with respect to its model, is extremely low.»

However, if we look at the distribution of meters by category (Table 15), we see that the number of meters that have been categorized with extremely low and low consumption are practically the same. For this reason, the conclusion obtained is not as precise as it should be, since there are meters whose fuzzy membership in these two categories is not very marked. Therefore, for this case, the most appropriate is to use an extended linguistic summary introduced in the previous section, leading to the result

«The consumption of UCLM, with respect to its model, is extremely low close to low.»

Table 15 | Distribution of meters by category for 10 August 2017.

Extremely Low	Low	Normal	High	Extremely High
38.14%	37.11%	12.37%	8.25%	0%

Furthermore, the model behaved best on 4 March 2017, when most of the buildings' consumption was cataloged as normal (Figure 11). In general, this is the most common trend in the organization for

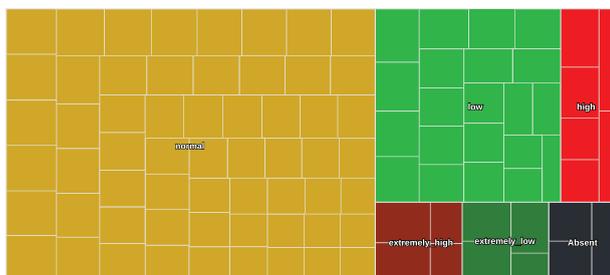


Figure 11 | Treemap of consumption on 4 March 2017.

the period 2017. Table 16 shows the distribution of buildings by category for that year. As in the previous cases, if we apply the linguistic summaries defined in the previous section, we can conclude that

«The consumption of UCLM, with respect its model, is normal.»

Finally, on 16 January 2017 and 15 April 2017, electricity consumption entries are higher and lower than expected, respectively. The first case (Figure 12) features a day of working activity preceded by a nonteaching period, Christmas, and a week of low work activity, given that it corresponds to a period of examinations, where the activity is limited to certain moments of the day. It is possible that this day concentrated most of the examinations in the different faculties that make up the UCLM, in addition to the use of heating systems due to the low temperatures experienced on a winter day. The second case (Figure 13) corresponds to a nonteaching day because it was in the Easter holiday period, which leads to relatively low consumption. However, it is worth noting that there is enough consumption to not be categorized as extremely low. Finally, by applying the linguistic summaries defined in the previous section, we conclude that in the first case

«The consumption of UCLM, with respect its model, is high.»

Whereas, in the second case

«The consumption of UCLM, with respect its model, is low.»

Table 16 | Distribution of meters by category of the model O.

Extremely Low	Low	Normal	High	Extremely High
3.60%	23.43%	42.01%	17.58%	9.41%

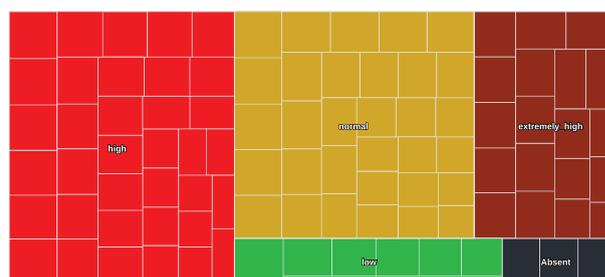


Figure 12 | Treemap of consumption on 16 January 2017.

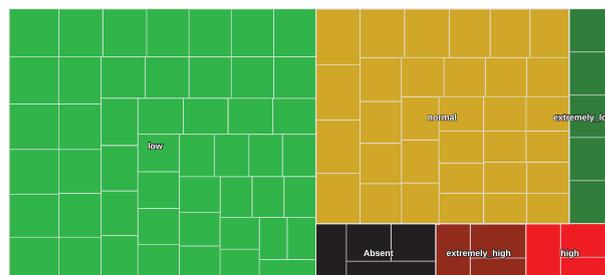


Figure 13 | Treemap of consumption on 15 April 2017.

8. CONCLUSIONS AND FUTURE WORK

This work proposes a new approach when analyzing and drawing conclusions from a set of time series of energy consumption data, by defining models that summarize the organization's consumption situation in linguistic terms. This will support decision-making by top managers when undertaking energy policies that contribute to the configuration of sustainable buildings. The definition of these models has been resolved by using cluster techniques. In particular, the k -means algorithm had a good performance and also good quality results. The choice of the number of groups has been based on the semantics that we wanted to associate with the model. In this work, we were motivated to detect consumption patterns (activity), and low or zero consumption (no activity). This allowed us to demonstrate their suitability for the dataset treated using probabilities and mathematical expectations rather than heuristic techniques. We have found that they do not give a uniform criterion of which k is most appropriate when dealing with a set of heterogeneous series. Linguistic summaries based on γ is P protoforms have been used for linguistic descriptions, where the summary has been modeled using a collection of fuzzy linguistic variables. The use of fuzzy sets when establishing this summary can lead to situations where it is more appropriate to present it in two labels if the threshold of membership is not very marked. In view of the lack of being able to describe this casuistry through the use of classic protoforms, in this work we propose an extension to model this idiosyncrasy by adding an absolute quantifier. We have also designed and developed prototype *software* where we support the models shown here, which is currently being used experimentally at the UCLM.

Future work should be aimed at the study of new data preprocessing techniques specific to a set of energy consumption data for electricity. This eliminates noise or mitigates the effect of possible *outliers* on the quality of the model obtained. To increase the performance of the model, we propose 1. To use the same technique by increasing the number of groups or making a finer segmentation of the identified groups (e.g., taking the data from the activity model (G_1) and defining a new submodel). 2. To use different models, such as those based on *deep learning* techniques. 3. To add more information to the model than just energy consumption data, such as *season of the year*, *weather forecast*, *building performance calendar*, *physical characteristics of buildings* (m^2 , *kind of activity*), and so on. Furthermore, we think it is possible to incorporate a system of alerts based on linguistic summaries into the model, confronting this proposal with data from other organizations, and develop a *big data* architecture that is based on microservices to support the definition and manipulation of models.

ACKNOWLEDGMENT

We have received support from the TIN2015-64776-C3-3-R project of the Science and Innovation Ministry of Spain, which is cofunded by the European Regional Development Fund (ERDF).

REFERENCES

- [1] Organisation for Economic Co-operation and Development, Transition to Sustainable Buildings: Strategies and Opportunities to 2050, Energy Technology Perspectives, OECD, Paris, 2013.
- [2] L.G. Swan, V.I. Ugursal, Modeling of end-use energy consumption in the residential sector: A review of modeling techniques, *Renew. Sust. Energy Rev.* 13(8) (2009), 1819–1835.
- [3] Natural language energy for promoting consumer sustainable behaviour. Accessed: 2018-09-11, https://cordis.europa.eu/project/rcn/195485_en.html.
- [4] P. Conde-Clemente, J.M. Alonso, G. Trivino, Toward automatic generation of linguistic advice for saving energy at home, *Soft Comput.* 22(2) (January 2018), 345–359.
- [5] R.E. Edwards, J. New, L.E. Parker, Predicting future hourly residential electrical consumption: a machine learning case study, *Energy Buildings.* 49 (2012), 591–603.
- [6] X. Li, C.P. Bowers, T. Schnier, Classification of energy consumption in buildings with outlier detection, *IEEE Trans. Ind. Electron.* 57(11), (2010), 3639–3644.
- [7] A. Capozzoli, F. Lauro, I. Khan, Fault detection analysis using data mining techniques for a cluster of smart office buildings, *Expert Syst. Appl.* 42(9) (2015), 4324–4338.
- [8] Z. Du, X. Jin, Y. Yang, Fault diagnosis for temperature, flow rate and pressure sensors in vav systems using wavelet neural network, *Appl. Energy.* 86(9) (2009), 1624–1631.
- [9] R. Fontugne, J. Ortiz, N. Tremblay, P. Borgnat, P. Flandrin, K. Fukuda, D. Culler, H. Esaki, Strip, bind, and search: a method for identifying abnormal energy consumption in buildings, in 2013 ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), IEEE, Pennsylvania, USA, 2013, pp. 129–140.
- [10] D. Liu, Q. Chen, K. Mori, Y. Kida, A method for detecting abnormal electricity energy consumption in buildings, *J. Comput. Info. Syst.* 6(14) (2010), 4887–4895.
- [11] G. Chicco, R. Napoli, F. Piglionne, Comparisons among clustering techniques for electricity customer classification, *IEEE Trans. Power Syst.* 21(2) (2006), 933–940.
- [12] C.S. Ozveren, C. Vechakanjana, A.P. Birch, Fuzzy classification of electrical load demand profiles—a case study, in 2002 Fifth International Conference on Power System Management and Control Conf. Publ. No. 488, 2002, 353–358.
- [13] Y.-H. Lin, M.S. Tsai, C.S. Chen, Applications of fuzzy classification with fuzzy c -means clustering and optimization strategies for load identification in NILM systems, in 2011 IEEE International Conference on, Fuzzy Systems (FUZZ) IEEE, Taipei, Taiwan, 2011, pp. 859–866.
- [14] G. Chicco, Overview and performance assessment of the clustering methods for electrical load pattern grouping, *Energy.* 42(1) (2012), 68–80.
- [15] A.K. Tanwar, E. Crisostomi, P. Ferraro, M. Raugi, M. Tucci, G. Giunta, Clustering analysis of the electrical load in European countries, in International Joint Conference on Neural Networks (IJCNN), 2015, IEEE, Killarney, Ireland, 2015, pp. 1–8.

- [16] J.E. Seem, Pattern recognition algorithm for determining days of the week with similar energy consumption profiles, *Energy Build.* 37(2) (2005), 127–139.
- [17] B. Aksanli, A.S. Akyurek, T.S. Rosing, User behavior modeling for estimating residential energy consumption, in: D.H.A. Leon-Garcia, R. Lenort (Eds.), *Smart City 360*, Springer, Bratislava, Slovakia and Toronto, Canada, 2016, pp. 348–361.
- [18] D. Mashima, A.A. Cárdenas, Evaluating electricity theft detectors in smart grid networks, in *International Workshop on Recent Advances in Intrusion Detection*, Springer, Amsterdam, The Netherlands, 2012, pp. 210–229.
- [19] K. Metaxiotis, A. Kagiannas, D. Askounis, J. Psarras, Artificial intelligence in short term electric load forecasting: a state-of-the-art survey for the researcher, *Energy Conver. Manag.* 44(9) (2003), 1525–1534.
- [20] G.K.F. Tso, K.K.W. Yau, Predicting electricity energy consumption: a comparison of regression analysis, decision tree and neural networks, *Energy*, 32(9) (2007), 1761–1768.
- [21] H.-T. Pao, Comparing linear and nonlinear forecasts for taiwan's electricity consumption, *Energy*, 31(12) (2006), 2129–2141.
- [22] M.L. Cam, A. Daoud, R. Zmeureanu, Forecasting electric demand of supply fan using data mining techniques, *Energy*, 101 (2016), 541–557.
- [23] D.C. Park, M.A. El-Sharkawi, R.J. Marks, L.E. Atlas, M.J. Damborg, Electric load forecasting using an artificial neural network. *IEEE Trans. Power Syst.* 6(2) (1991), 442–449.
- [24] J. Riquelme, J.L. Martínez, A. Gómez, D. Cros Goma, Load pattern recognition and load forecasting by artificial neural networks, *Int J. Power Energy Syst.* 22(2) (2002), 74–79.
- [25] B. Wang, S. Xu, X. Yu, P. Li, Time series forecasting based on cloud process neural network, *Int. J. Comput. Intell. Syst.* 8(5) (2015), 992–1003.
- [26] A. Azadeh, S.F. Ghaderi, S. Sohrabkhani, Forecasting electrical consumption by integration of neural network, time series and anova, *Appl. Math. Comput.* 186(2) (2007), 1753–1761.
- [27] M. El-Telbany, F. El-Karmi, Short-term forecasting of Jordanian electricity demand using particle swarm optimization, *Electr. Power Syst. Res.* 78(3) (2008), 425–433.
- [28] J. Kacprzyk, A. Wilbik, Temporal linguistic summaries of time series using fuzzy logic, in: E. Hullermeier, R. Kruse, F. Hoffmann (Eds.), *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Methods*, Dortmund, Germany, 2010, pp. 436–445.
- [29] L.A. Zadeh, A computational approach to fuzzy quantifiers in natural languages, *Comput. Math. Appl.* 9(1) (1983), 149–184.
- [30] M. Ros, M. Pegalajar, M. Delgado, A. Vila, D.T. Anderson, J.M. Keller, M. Popescu, Linguistic summarization of long-term trends for understanding change in human behavior, in *2011 IEEE International Conference on Fuzzy Systems (FUZZ)*, IEEE, Taipei, Taiwan, 2011, pp. 2080–2087.
- [31] A. Alvarez-Alvarez, G. Trivino, Linguistic description of the human gait quality, *Eng. Appl. Artif. Intell.* 26(1) (2013), 13–23.
- [32] A. Alvarez-Alvarez, D. Sanchez-Valdes, G. Trivino, Á. Sánchez, P.D. Suárez, Automatic linguistic report of traffic evolution in roads, *Expert Syst. Appl.* 39(12) (2012), 11293–11302.
- [33] R.M. Catillo-Ortega, N. Marín, D. Sánchez, A fuzzy approach to the linguistic summarization of time series, *J. Mult. Valued Log. Soft Comput.* 17 (2011), 157–182.
- [34] A. van der Heide, G. Triviño, Automatically generated linguistic summaries of energy consumption data, in *9th International Conference on Intelligent Systems Design and Applications*, 2009. *ISDA'09. IEEE 2009*, pp. 553–559.
- [35] J. MacQueen, Some methods for classification and analysis of multivariate observations, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Oakland, CA, 1967, vol. 1, pp. 281–297.
- [36] E. Keogh, J. Lin, Clustering of time-series subsequences is meaningless: implications for previous and future research, *Knowl. Info. Syst.* 8(2) (2005), 154–177.
- [37] M. Erisoglu, N. Calis, S. Sakalloglu, A new algorithm for initial cluster centers in k-means algorithm, *Pattern Recognit. Lett.* 32(14) (2011), 1701–1705.
- [38] A.K. Jain, Data clustering: 50 years beyond k-means. *Pattern Recognit. Lett.* 31(8) (2010), 651–666.
- [39] S. Petrovic, A comparison between the Silhouette index and the Davies-Bouldin index in labelling ids clusters, in *Proceedings of the 11th Nordic Workshop of Secure IT Systems*, Linkping, Sweden, 2006, pp. 53–64.
- [40] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, Understanding of internal clustering validation measures, in *2010 IEEE 10th International Conference on Data Mining (ICDM)*, IEEE, Linkping, Sweden, 2010, pp. 911–916.
- [41] F.M. Alvarez, A. Troncoso, J.C. Riquelme, J.S. Aguilar Ruiz, Energy time series forecasting based on pattern sequence similarity, *IEEE Trans. Knowl. Data Eng.* 23(8) (2011), 1230–1243.
- [42] C.A. Sugar, G.M. James, Finding the number of clusters in a dataset: an information-theoretic approach, *J. Am. Stat. Assoc.* 98(463) (2003), 750–763.
- [43] T. Pedersen, A. Kulkarni, Automatic cluster stopping with criterion functions and the gap statistic, in *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume: Demonstrations*, Association for Computational Linguistics, 2006, pp. 276–279.
- [44] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, *J. Royal Stat. Soc. Series B.* 63(2) (2001), 411–423.
- [45] P. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math.* 20(1) (November 1987), 53–65.
- [46] L.A. Zadeh, Fuzzy logic = computing with words, *IEEE Trans. Fuzzy Syst.* 4(2) (1996), 103–111.