Hierarchical clustering of complex energy systems using pretopology

Loup-Noé Lévy[1,2], Jérémie Bosom[2,3], Guillaume Guerard[4], Soufian Ben Amor[1], Marc Bui[3], and Hai Tran[2]

[1] LI-PARAD Laboratory EA 7432, Versailles University, 55 Avenue de Paris, 78035 Versailles, France {f_author. s_author}@uvsq.fr
[2] Energisme, 88 Avenue du Général Leclerc, 92100 Boulogne-Billancourt France {f_author. s_author}@energisme.com
[3] EPHE, PSL Research University, 4-14 Rue Ferrus, 75014 Paris, France {f_author. s_author}@ephe.psl.eu
[4] De Vinci Research Center, Pole Universitaire Léonard de Vinci, 12 Avenue Léonard de Vinci, 92400 Courbevoie, France
[5] email{f_author. s_author}@devinci.fr

Abstract. This article attempts answering the following problematic: How to model and classify energy consumption profiles over a large distributed territory to optimize the management of buildings' consumption?
Doing case-by-case in depth auditing of thousands of buildings would require a massive amount of time and money as well as a significant number of qualified people. Thus, an automated method must be developed to establish a relevant and effective recommendations system.
To answer this problematic, pretopology is used to model the sites' consumption profiles and a multi-criterion hierarchical classification algorithm, using the properties of pretopological space, has been developed in a Python library.
To evaluate the results, three data sets are used: A generated set of dots of various sizes in a 2D space, a generated set of time series and a set of consumption time series of 400 real consumption sites from a French Energy company.
On the point data set, the algorithm is able to identify the clusters of points using their position in space and their size as parameter. On the generated time series, the algorithm is able to identify the time series clusters using Pearson's correlation with an Adjusted Rand Index (ARI) of 1.

Keywords: Artificial intelligence · data analysis · clustering algorithms · pretopology

1 Introduction

In 2015 was signed the Paris agreement in which government from all over the world undertook to keep global warming behind a 2°C increase compared to the

temperatures of 1990. The year of the Cop21, the worldwide buildings sector was responsible for 30% of global final energy consumption and nearly 28% of total direct and indirect CO_2 emissions. Yet the energy demand from buildings and building's construction still rises, driven by improved access to energy in developing countries, greater ownership and use of energy-consuming devices and rapid growth in global buildings floor area, at nearly 3% per year [6]. The International Energy Agency's Reference Technology Scenario (RTS), which accounts for existing building energy policies and climate-related commitments, shows that final energy demand in the global buildings sector will increase by 30% by 2060 without more ambitious efforts to address low-carbon and energy-efficient solutions for buildings and construction. As a result, buildings-related CO_2 emissions would increase by another 10% by 2060, adding as much as 415 $GtCO_2$ to the atmosphere over the next 40 years – the half of the remaining 2°C carbon budget and twice what buildings emitted between 1990 and 2016.[7] Yet there are significant opportunities for improvement, as in the United States where 16% of energy savings could be achieved by reducing performance deficiencies [23]. Energy actors such as Trusted Third-Party for Energy Measurement and Performance can play a role in identifying the most relevant actions to optimize energy consumption by exploiting the massive energy data now available [6].

There are many ways to decrease buildings' energy consumption [9]: social programs, incentive programs, new energies, energy efficiency, dynamic pricing, demand-response programs. But it is challenging to identify precisely what action to take.

Furthermore, the energy systems are not necessarily buildings. They can be a building floor or simply a place inside a building. In consequence, it is more accurate to talk about **sites** [6].

The scales of analysis are various both in time (consumption time series are analyzed on a 24h profile as well as on a yearly profile) and space (the studied system can go from one room to a group of buildings across a country). Because of that, there is no universal performance scale on which to compare a site to another.

Because sites present an important heterogeneity both in intrinsic properties and geographic situation [22] only a comparison between similar sites might be meaningful to understand the performance of a new site. By investigating the works that were effective on a certain site, one can deduce what programs will probably be efficient for sites of similar nature. Hence, clustering sites based on their characteristics and consumption will enhance their evaluation and the recommendations system.

Therefore the topic of our paper is as following: *How to cluster a large number of heterogeneous sites based on their energy consumption profiles to recommend the most relevant energy optimization solution possible?*

In this article, we will consider that the energy consumption profile encompasses all the physical characteristics of a site as well as the external factors and

[6] http://www.eia.gov/
[7] https://www.iea.org/topics/energyefficiency/buildings/

the consumption data (time series, categorical data and numerical data). The latter is considered as a time series.

Our goal is to study a group of sites to optimize their consumption thanks to recommendations done on similar sites. This can be assimilated to portfolio analysis. Portfolio analysis represents a domain in which a large group of buildings, often located in the same geographical area or owned or managed by the same entity, are analyzed for the purpose of managing or optimizing the group as a whole [22].

The key contribution of this paper is to provide a clustering method adapted to portfolio analysis based on a pretopological framework. - new definitions, properties, and demonstrations - detailed explanations of the algorithms and their pseudo-codes

Compared to the previous paper [17] this paper gives greater theoretical understanding of pretopology through added definitions, properties, and demonstration. It demonstrates how the pretopological framework used for the algorithms allows for the clustering of any finite set of items. It also explains the algorithms in greater details as well as presenting the pseudo-code of the algorithms. It also discusses the future work to exploit clustering for energy performance

The paper is structured as follows: the section 2 introduces clustering methods and some relevant examples on energy systems. The section 3 presents the pretopology theory and the different types of pretopological spaces. The section 4 explains in details the algorithms developped in the python library with pseudo-code, demonstrating how all finite set of items can be hierarchically clustered. The section 5 presents the clustering of different types of datasets. We discuss the results and futur work in the section 6 We conclude in the section 7.

2 Literature review

In this section, we present clustering methods and their application on energy systems. Clustering is a set of unsupervised machine learning methods that group unlabeled items into clusters. The journal paper of Iglesia et al. in Energies [12] presents a deeper analysis of clustering in energy system. To consult an exhaustive list of clustering algorithms, please read Xu et Al. survey [25].

There are four classes of clustering algorithms. Each of them having pros and cons: density-based clustering, centroid-based clustering, hierarchical clustering, distribution-based clustering. Let us present each class and their application to portfolio analysis in energy system.

Centroid-based clustering: In these methods, a cluster is a set of items such that an item in a cluster is closer to the center of a cluster than to the center of any other cluster. The center of a cluster is called the centroid, the average of all points in the cluster, or medoid, the most representative point in a cluster. The well-known centroid-based algorithm is the *K-means* algorithm and its extensions. The *K-means* algorithm is a powerful tool for clustering, but it requires to determine in advance the number of clusters that the algorithm should find.

Therefore, centroid-based algorithms are sensitive to initial conditions. Clusters vary in size and density and include outliers (isolated items) from the nearest cluster. Finally, centroid-based algorithms do not scale with the number of items and dimensions. In this case, centroid-based algorithms are combined with principal component analysis or spectral analysis to be more efficient.

Regarding portfolio analysis in energy systems, Gao et al. [8] compare a multidimensional energy consumption dataset using a *k-means* algorithm. Freischhacker et al. [7] design a spatial aggregation method, combined with *k-means*, based on block characteristics to reduce reductions due to energy consumption.

Density-based clustering: In density-based clustering, a cluster is a set of features distributed in the data space over a contiguous region of high feature density. Elements located in low density regions are generally considered noise or outliers [13]. The well-known methods in this class are Density-Based Spatial Clustering of Applications with Noise (*DBSCAN*) and its extensions.

Two parameters influence the formation of clusters: density and accessibility. Therefore, clusters are distinct according to these parameters. The main strength of this density-based clustering algorithm is it does not require apriori specification and that it is able to identify noisy data during clustering. It fails in the case of neck type datasets and does not work well for high dimensionality data.

Regarding portfolio analysis in energy systems, Li et al. [18] present a density-based method with particle swarm optimization of building portfolio parameters. Their method predicts the next day's electricity consumption through clustering. Marquant et al. [21] use a density and load-based algorithm to facilitate large-scale modeling and optimization of urban energy systems.

Hierarchical clustering: Hierarchical clustering is most often a procedure whose goal is to transform a proximity matrix into a sequence of hierarchically structured partitions.

The two methods of hierarchical clustering are the bottom-up method (or agglomeration) or the top-down method (or division). Bottom-up methods start from disjoint classes and place each of the elements in an independent class. From the proximity matrix, the procedure searches at each step for the two closest classes, merges them, then places them in a second partition. The process is repeated to build a sequence of nested partitions in which the number of classes decreases as the sequence progresses until a unique class contains all elements. Top-down methods perform the reverse process.

The key problem of these algorithms is to define the criterion for grouping or aggregating two classes, i.e. a distance measure. Sites are defined as complex systems: [1, 5, 6, 10]. They are defined with numerical and categorical data as well as time series. For this reason calculating a distance between two elements is challenging and does not allow to use every feature of the site in a relevant way. Another drawback is the difficulty of identifying a precise number of clusters, especially in a large data set.

Regarding portfolio analysis in energy systems, Wang et al. [24] analyze the spatial disparity of final energy consumption in China through hierarchical clustering and spatial autocorrelation. Li et al. [19] implement a strategy based on agglomerative hierarchical clustering to identify typical daily electricity usage patterns.

Distribution-based clustering: Application to large spatial databases requires from clustering algorithms to have no or minimal input parameters and arbitrarily shaped clusters. Distribution-based clustering produces clusters that assume concisely defined mathematical models underlying the items, a relatively plausible assumption for some item distributions.

Most of the time, the mathematical models are based on the Gaussian, multinomial, or multivariate normal distribution. Clusters are considered fuzzy, which means that an item can be found in several clusters at a defined percentage. The best known algorithm is the Expectation-Maximization (EM) clustering with Gaussian mixture models (GMM). Thus, the GMM algorithm provides two parameters to describe the shape of the clusters: the mean and the standard deviation. The main drawback of these algorithms is that they cannot work on categorical dimensions.

Regarding portfolio analysis in energy systems, Lu et al. [20] use GMM clustering for the identification of heating load patterns. Habib et al. [11] provide EM clustering to detect outliers in the energy building portfolio.

Conclusion about clustering methods: None of the methods described above can answer the specificities of the studied system, either because they require the definition of a distance between the items, or because they cannot return the hierarchical clustering necessary to apprehend the different scales of a complex system.

Relevance of pretopology-based clustering: A pretopological space is defined by a relationship between a set of items and a larger set of items. It is therefore suitable for creating a hierarchical structure. It is based on the concept of abstract space. In such a space, the nature of the element is not relevant, it is rather the relations and properties linking the elements together that are important. This allows us to manipulate heterogeneous and complex elements such as our sites. Therefore, pretopology can be considered as a mathematical tool to model the concept of proximity for complex systems [2]. Pretopology is therefore the approach chosen to build our hierarchical clustering.

3 Pretopology

In this section we will explain the key concepts and definition of pretopology, such as pretopological space and pseudo-closure. We won't go into detail on the origins of pretopology but it is important to understand that the concept of pretopological space is obtained by weakening the hypothesis of the topological spaces. It allows the modeling of discrete structures unlike topology [2].

3.1 Pretopological space

Central definitions and propositions

Definition 1. *A pseudoclosure function* $a : \wp(U) \to \wp(U)$ *on a set* U*, is a function such that:*

- $a(\emptyset) = \emptyset$
- $\forall A \mid A \subseteq U : A \subseteq a(A)$

where $\wp(U)$ *is the power set of* U

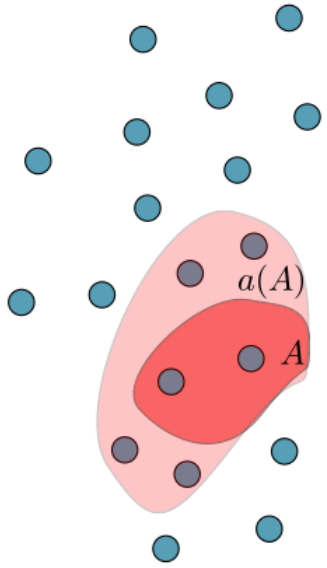


Fig. 1. Example of a pseudoclosure function [14]

Definition 2. *A tuple* $(U, a(.))$*, where* U *is a set of elements and* $a(.)$ *is a pseudoclosure function on* U*, constitutes a pretopological space.*

We note that a pretopological space is defined by establishing a relation between any set of elements and a bigger set. This is interesting in the construction of a hierarchy. The previous definition determines the most general pretopological space. By asking the function to fulfill some additional conditions we get more specific pretopological spaces:

Definition 3. *If* $\forall\ A, B \mid A \subseteq U,\ B \subseteq U : A \subseteq B \implies a(A) \subseteq a(B)$*, then we get a pretopological space of type* V*. This property is called isotony.*

Definition 4. *If* $\forall\ A, B \mid A \subseteq U,\ B \subseteq U\ : a(A \cup B) = a(A) \cup a(B)$*, then we get a pretopological space of type* V_D*.*

Definition 5. *If* $\forall\ A \mid A \subseteq U\ : a(A) = \bigcup_{x \in A} a(x)$ *then we get a pretopological space of type* V_S*.*

Given any pretopological space $(U, a(.))$, we can ask ourselves the question of what becomes of the concepts of closure classically defined in topology. In fact, the definition remains the same in pretopology [16].

Definition 6. *A part* F *of* U *will be a closure of* U *if and only if* $a(F) = F$

Proposition 1. *In a pretopological space of type* V*, the intersection of closures is a closure.*

Proposition 2. *In a pretopological* $V-$*type space, the closure and opening of any part of* U *still exists.*

Proposition 3. *In a pretopological space of type* V*, the closure of a part* A *of* U *is the smallest closure containing* A*. Denoted F(A).*

Proposition 4. *In a pretopological space of type* V*, every set has a closure. The proof can be found in [3].*

In a pretopological space of type V we can find the closure by repeatedly applying the pseudoclosure operator to the set and its subsequent images until it stops expanding. We can see an example of this in figure 2 [14].

If we have a pretopological space of type V_D and $\forall A \mid A \subseteq U\ : a(A) = a(a(A))$, then we get a topology. The pseudoclosure function here is said to be idempotent [14]. It's clear that in a finite space, $V_S = V_D$ [3]. Also, in pretopological spaces of type V_D the pseudoclosure of a set is completely defined by the pseudoclosures of its singletons. So if the space is also finite, we could draw an edge from an element to every element of its pseudoclosure, and the pseudoclosure would be equivalent to a particular graph. Figure 3 shows the relation between the two. This demonstrates that pretopology is also a generalization of graph theory [14].

There is a second way of characterizing pretopologies of type V and V_D . To understand it we need to give a few more definitions first:

Definition 7. *We say that a set* $\mathcal{F}$ *of* $\wp(\wp(U))$ *is a prefilter over* U*, if:*

$$\forall F \in \mathcal{F}, \forall H \in \wp(U), F \subset H \Longrightarrow H \in \mathcal{F} \tag{1}$$

Definition 8. *We say that a set* $\mathcal{F}$ *of* $\wp(\wp(U))$ *is a filter over* U*, if it is a prefilter stable under finite intersection, i.e.*

$$\forall F \in \mathcal{F}, \forall G \in \mathcal{F}, F \cap G \in \mathcal{F} \tag{2}$$

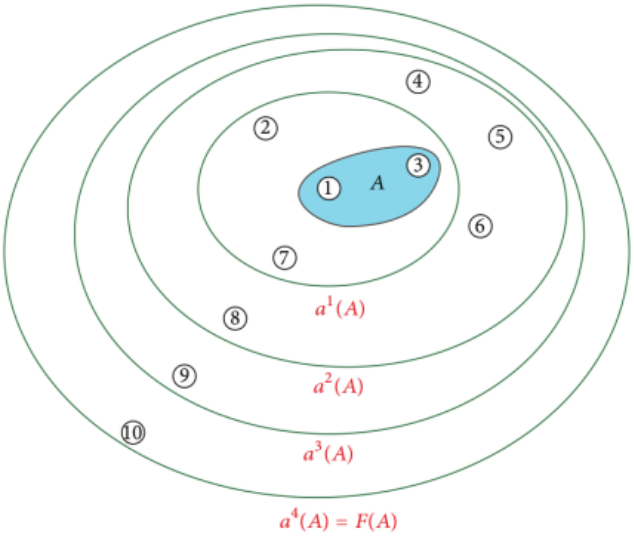


Fig. 2. Closure of set A [14]

In other words, and restricting ourselves to a finite space, a filter is the family of all supersets of a set $\mathcal{B}$, while a prefilter is the family of supersets of every member B_i of a family of sets $\mathcal{B}$. The family of sets $\mathcal{B}$ is called the basis of the prefilter. We can see in figure 4 an example of a filter and a prefilter with basis $B = 1, 4, 2, 4$ [14].

Now, if we have a set U, and for every $x \in U$ we have a prefilter $V(x)$ such that every member of $V(x)$ contains the element x, we can define a pseudoclosure function in the following way:

$$\forall A \subseteq U, a(A) = \{x \in U | \forall V \in V(x), V \cap A, \emptyset\} \tag{3}$$

We call the prefilter $V(x)$ the family of neighborhoods of x, and each set in the family is called a neighborhood of x. Figure 5 shows a graphical representation of this definition of the pseudoclosure. On the other hand, if we have a pseudoclosure function $a(.)$ in a pretopological space of type V, the family of sets given by:

$$V(x) = \{V \subset U | x \in i(V)\} \tag{4}$$

where $i(A) = a(A^c)^c$, is a prefilter. The following proposition shows that we can go from one definition to the other interchangeably [3]:

Proposition 5. *No two families of prefilters* $\{V(x_i)|x_i \in U\}$ *define the same pseudoclosure function* $a(.)$*, and no two pseudoclosure functions define the same family of prefilters* $\{V(x_i)|x_i \in U\}$*.*

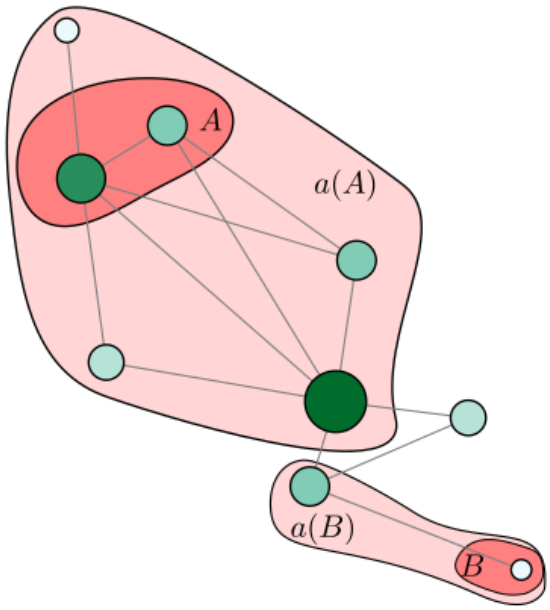


Fig. 3. Pseudoclosure function on a graph [14]

4 Hierarchical Clustering Algorithms

This section describes the algorithms developped in a Python library used for the construction of a closure and to build a hierarchical clustering of sites. This algorithm, whose pseudo-code is given in the source code 1, is organized in four phases:

- Determine a family of elementary subsets called seeds.
- Construct the closures of the seeds by iterative application of the pseudoclosure function.
- Construct the adjacency matrix representing the relations between all the identified subsets (even the intermediate ones).
- Establish the quasi-hierarchy by applying the associated algorithm on the adjacency matrix.

Several methods are possible to determine the seeds. Therefore, the algorithm is influenced by the following two hyperparameters:

- the $seedFunc(.)$ function which determines, for an element, a set of close elements which will constitute a seed,
- the degree d to specify the size of the seeds.

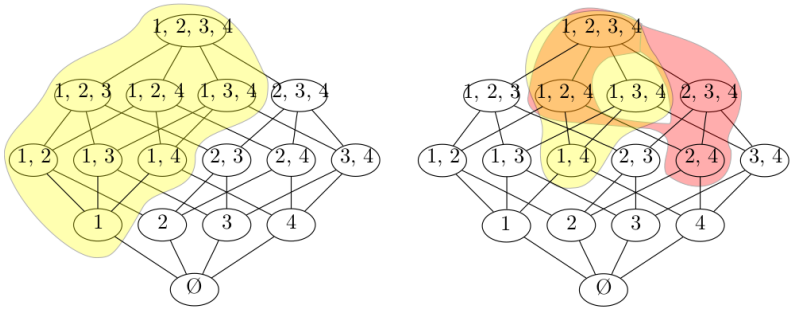
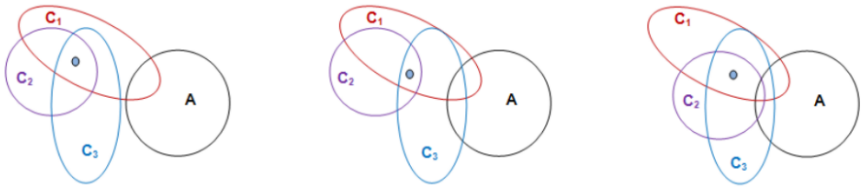


Fig. 4. Filters vs Prefilters [14]



(a) $x \notin a(A)$ (b) $x \notin a(A)$ (c) $x \in a(A)$

Fig. 5. Neighborhood definition of a pretopology [14]

The algorithm takes an additional hyperparameter, required by the *ExtractQuasihierarchy* algorithm in order to establish the quasi-hierarchy: th_{qh}, which corresponds to the threshold beyond which it is estimated that two elements are close. This number is generally between 0 and 1.

We now detail each phase of the algorithm.

Calculation of a family of elementary sets or seeds The aim here is to determine elementary subsets of size d called seeds thanks to the function $seedFunc(.)$ whose role is to find the d needed neighbors. To do so, we iterate on all the points of the set U associated to the pretopological space p. The pseudo-code of the resulting algorithm (named *ElementaryQuasiclosures*) is presented in the source code 2.

The algorithm 2 uses the function $FindNeighbors$ whose peudo-code is given in the source code 3. The latter takes as parameters an element of U, the number of neighbors sought d and the function determining the nearest neighbors $seedFunc(.)$. The $seedFunc(.)$ function usually takes as its value one of the following two functions:

Algorithm 1 QuasistructuralAnalysis: Algorithm for the analysis of the quasi-hierarchy of a pretopological space, based on the work of [14]

Require: $((U, a(.)), d, seedFunc(.), th_{qh})$
Ensure: $QF_{qh}, quasiHierarchy$
  $seedList \leftarrow ElementaryQuasiclosures((U, a), d, seedFunc)$
  $QF_e \leftarrow ElementaryClosedSubsets((U, a), seedList)$
  $Adj_{qh} \leftarrow ExtractAdjencyQuasihierarchy(QF_e)$
  $QF_{qh}, quasiHierarchy \leftarrow ExtractQuasihierarchy(QF_e, Adj_{qh}, th_{qh})$

Algorithm 2 ElementaryQuasiclosures: Construction of the seeds by applying the function $seedFunc(.)$ on all the elements of the set U, based on the work of [14]

Require: $((U, a(.)), degree, seedFunc(.))$
Ensure: $seedList$
  $seedList \leftarrow list()$
  for all $x \in U$ **do**
    $seedList \leftarrow list()$
    $seedList.append(seed)$
  end for

Algorithm 3 FindNeighbors: Determine the d neighbors of $firstNode$ using the $seedFunc(.)$ function, based on the work of [14]

Require: $(firstNode, d, seedFunc(.))$
Ensure: $path$
  $path \leftarrow list()$
  $lastTreatedNode \leftarrow firstNode$
  for all $i \in range(d)$ **do**
    $newNode \leftarrow seedFunc(lastTreatedNode)$
    $path.append(newNode)$
    $lastTreatedNode \leftarrow newNode$
  end for

- $ClosestNode(node)$ which identifies the closest nodes to an element. It is used in cases where a distance can be calculated, for example in the case where the studied relations are quantifiable.
- $RandomNeighbor(node)$ randomly browses the neighboring nodes. Its use is preferred when the relations are not quantifiable, for example in the case of values describing categories.

Construction of subsets by applying pseudoclosure To construct the subsets that will then be organized by the pseudo-hierarchy algorithm, *ElementaryClosedSubsets* uses the seed list *seedList* computed previously by *ElementaryQuasiclosures*. For each of the seeds in *seedList*, the membership function is applied iteratively until the pseudo-closure no longer gives bigger sets.

The intermediate and final subsets are stored in a list of unique element lists (*list* of *set*) named QF_{tmp} so that we don't have to reapply the membership later on the same sets. QF_{tmp} indexes the subsets according to the number of elements they contain. Since the membership of a set is always greater than or equal to its size, such indexing ensures that all elements are processed once and only once.

The list QF_e, constructed from the lists in QF_{tmp}, is then returned. The associated pseudo-code is presented in the source code 4.

Algorithm 4 ElementaryClosedSubsets: Computes the set of subsets by iterative application of the pseudo-closure function, algorithm inspired from [14]

Require: $((U, a(.)), seedList)$
Ensure: QF_e
  QF_{tmp} a list of $Size(U)$ sets
  for all $seed \in seedList$ **do**
    $QF_{tmp}[Size(seed)].append(seed)$
  end for
  for all $i \in range(1, Size(U) + 1)$ **do**
    for all $s \in QF_{tmp}[i]$ **do**
      $pseudoclosure \leftarrow a(s)$
      if $lastTreatedNode \leftarrow newNode$ **then**
        $QF_{tmp}[Size(pseudoclosure)].append(pseudoclosure)$
      end if
    end for
  end for
  $QF_e \leftarrow list()$
  for all $i \in range(Size(QF_{tmp}))$ **do**
    $QF_e.extend(QF_{tmp}[i])$
  end for

Construction of the adjacency matrix The objective of this algorithm is to establish the hierarchical relations between the graphs of QF_e identified by

ElementaryClosedSubsets. These relationships, between all QF_e sets, are represented as an adjacency matrix Adj_{qh}.

In a space of type V, two distinct closed elementary subsets F_x and F_y of QF_e :

- are either disjoint then $F_x \cap F_y = \emptyset$,
- either contain a nonzero intersection such that $\forall; z \in F_x \cap F_y$, $F_z \subset F_x \cap F_y$, where F_z is the closure of z.

Thus, if two subsets F_x and F_y overlap without one of them being contained in the other ($F_x \cap F_y \neq \emptyset$, $F_x \not\subset F_y$ and $F_y \not\subset F_x$), we know that a smaller set F_z contained in $F_x \cap F_y$ exists. The resulting hierarchical graph must therefore connect F_x and F_y as parents of F_z.

However, as we mentioned, the Laborde's algorithm [14] is intended to be applicable to non-V spaces as well. In such pretopological spaces, there is no guarantee that an element of $F_x \cap F_y$ will not grow beyond this intersection. This is illustrated in Figure 6. Furthermore, in the case of $d-n$ elementary sets, where n is the cardinality of U and d is the degree applied for creating the seeds, it is possible that none of the seeds are contained in the intersection. Thus, it is possible that no obvious structure emerges from the collection of quasi-closures.

To solve this problem, Laborde et Al. [14] generalizes the type of hierarchy constructed from quasi-closures so as to satisfy the following constraints:

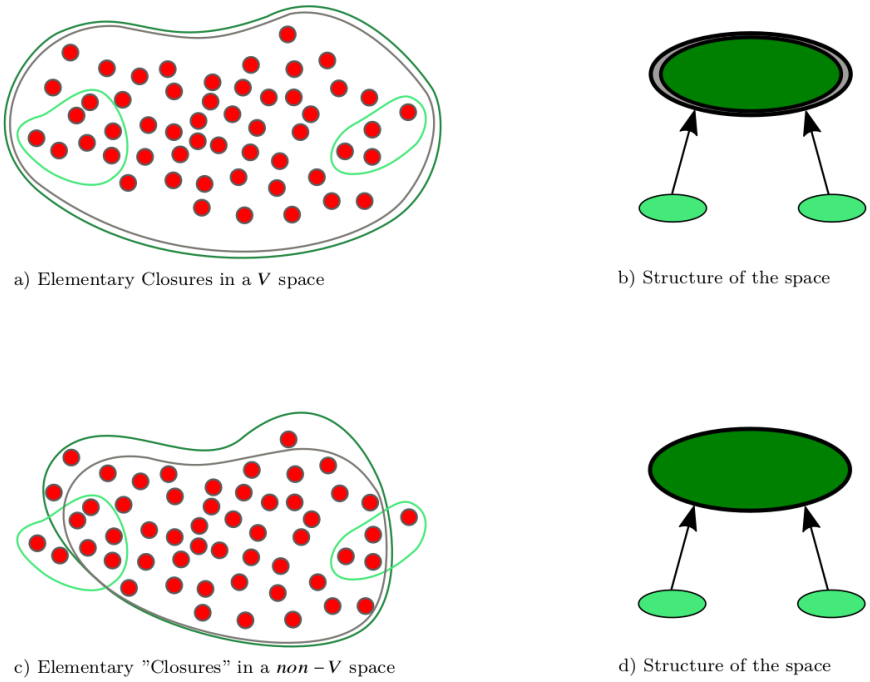
- Two subsets should be connected only if their intersection is nonzero ($F_x \cap F_y$),
- The larger the cardinality of the intersection $F_x \cap F_y$ is compared to that of F_x, the stronger the relation of F_x to F_y is,
- The larger the cardinality of the subset F_y compared to that of F_x, the less necessary it is that $F_x \cap F_y$ is large for the relation from F_x to F_y to be strong. In other words, a very large set will attract smaller sets even if their intersection is not very large.

The algorithm presented in the source code 5 implements this logic. It quantifies the relations between each pair of QF_e whose intersection is not empty and then returns the resulting matrix.

Construction of the quasi-hierarchy The quasi-hierarchy is built from the adjacency matrix by checking if the relations computed by *ExtractAdjencyQuasi-hierarchy* exceed the threshold th_{qh}. The new adjacency matrix thus obtained defines the quasihierarchy returned by the algorithm. The algorithm also returns the final list QF_{qh} of identified subsets for the set U. QF_{qh} corresponds to the list QF_e updated following the potential addition of new subsets during the construction of the quasi-hierarchy.

The quasi-hierarchy is established by applying the following rules on the values of Adj_{qh}:

- A link between two subsets is established in the quasi-hierarchy if their relationship exceeds the threshold th_{qh},



a) Elementary Closures in a V space

b) Structure of the space

c) Elementary "Closures" in a $non-V$ space

d) Structure of the space

Fig. 6. Construction of a quasi-hierarchy in a pretopological space of type V, according to the method of [15], and of type non-V, according to the method of [14], figures by this later author

- Two subsets that have strong mutual relations (exceeding the threshold th_{qh}) are considered equivalent. They are subject to a subsidiary treatment improving the resulting quasi-hierarchy.
- The resulting quasiclosures with the respective links determine the quasi-hierarchy.

Laborde et Al. [14] treats the case of equivalent sets by keeping the largest set and deleting the other. If the sets are of the same size then one of them is chosen randomly.

5 Model validation and visualization of results

Validation tool: To evaluate the pretopological hierarchical clustering, we also provide a set of tools to validate the model and show the results.

This program is developed to create a point dataset with the following parameters:

- the number of groups of dense items;

Algorithm 5 ExtractAdjencyQuasihierarchy: Construction of the adjacency matrix for the quasi-hierarchy, algorithm inspired from [14]

Require: (QF_e)
Ensure: Adj_{qh}
  $Adj_{qh} \leftarrow SquaredMatrixZeros(size(QF_e))$
  for all $F, G \in QF_e$ **do**
    $FhasG \leftarrow Size(F \cap G)/Size(G)$
    $GhasF \leftarrow Size(F \cap G)/Size(F)$
    $FbiggerG \leftarrow Size(F)/Size(G)$
    $GbiggerF \leftarrow Size(G)/Size(F)$
    $Adj_{qh}[Index(G), Index(F)] = GbiggerF * GhasF$
    $Adj_{qh}[Index(F), Index(G)] = FbiggerG * FhasG$
  end for

- the number of items of each group;
- the spatial dispersion of each group;
- the position of each group.

The size of an item is added as a second parameter, to evaluate multi-criteria clustering. Groups with different item size can be produced with the following parameters:

- the number of groups;
- the number of items of each group;
- the range of sizes of each group.

This program allows us to evaluate our method in different types of situations and to easily make adjustments or corrections.

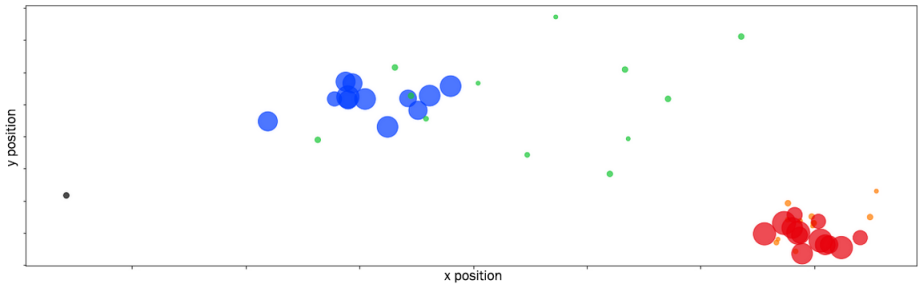


Fig. 7. The four clusters determined by our algorithm using both size and position as parameters, on a 2D disks dataset [17].

Visualization tool: The program colors each of the largest sets determined by our algorithm with a single color to make the clusters apparent. The validation

tool is tested with two groups of large and small elements and a two-dimensional position. The elements are shown in figure 7. In this example, four clusters were determined: blue, green, orange and red. The black dot at the far left of the figure 7 is an element identified as an outlier by the algorithms. For example, the red and orange elements are close to each other but separated into two clusters due to their different sizes, and the orange and green dots are similar in size but divided into two sets due to their different positions.

The program also displays the hierarchical classification consisting of the seeds, intermediate sets and final clusters. The hierarchical classification is displayed as a tree in which each set is identified by a number and is represented by a node.

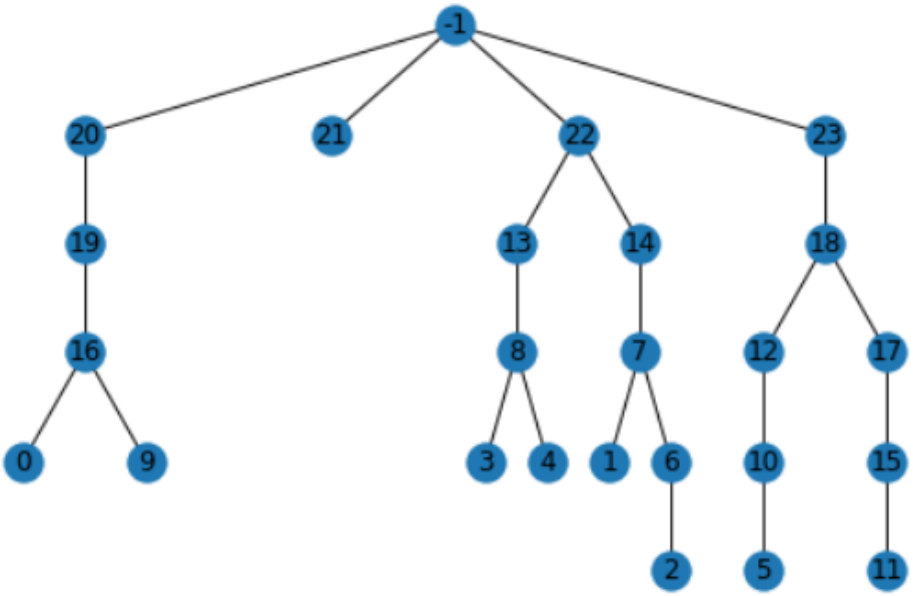


Fig. 8. A tree representing the pseudohierarchy relation between each intermediate set from the seed to the cluster [17].

For example, the hierarchy shown in Figure 8 shows the relationships between the sets determined by our algorithm applied to the dataset displayed in Figure 7. Only the sets with more than two elements are shown on this tree. We can recognize the four clusters that have been colored on figure 7, they are labeled 20, 21, 22 and 23. Figure 9 displays cluster 14 which is a child of cluster 21 (colored in green) in the hierarchical clustering. This hierarchy identifies large clusters of relatively similar items and provides more detail about small clusters of very similar items.

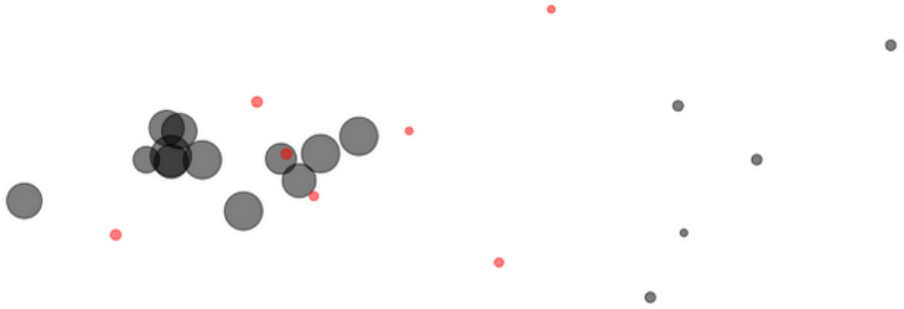


Fig. 9. The subset 14 in red representing a subgroup of the green clusters (subset 22) in figure 7 [17].

5.1 Results on different datasets

5.2 Benchmark dataset

Since the main data we have from the sites are time series of power consumption, we needed to test, visualize, and evaluate the clustering of a time series set. This section presents this test set and the results of our algorithm. The test set created, consisting of six clusters, is shown in Figure 10. Each cluster is composed of 30 time series of 60 points.

The similarity measure used to establish the value between two items is the Pearson's coefficient. The Pearson correlation coefficient measures the linear relationship between each pair of items, which in this case are time series.

Our program colored the time series based on the clusters it had identified (see figure 10).

5.3 Results analysis on benchmark dataset

The program identified exactly the same clusters as the ground truth given by the benchmark. To evaluate the validity of the clusters determined by the algorithm, our metric is the Adjusted Rand score, also called Adjusted Rand Index (ARI). As we have perfectly identified the clusters, the ARI of our clustering is 1. Figure 11 shows the confusion matrix between the cluster found by our method and the ground truth given by the benchmark.

Further experiments will be conducted in a future contribution.

5.4 Real dataset

This dataset is built from Enedis (the French electricity network manager) consumption time series for 400 sites over one year. It is resampled with a time

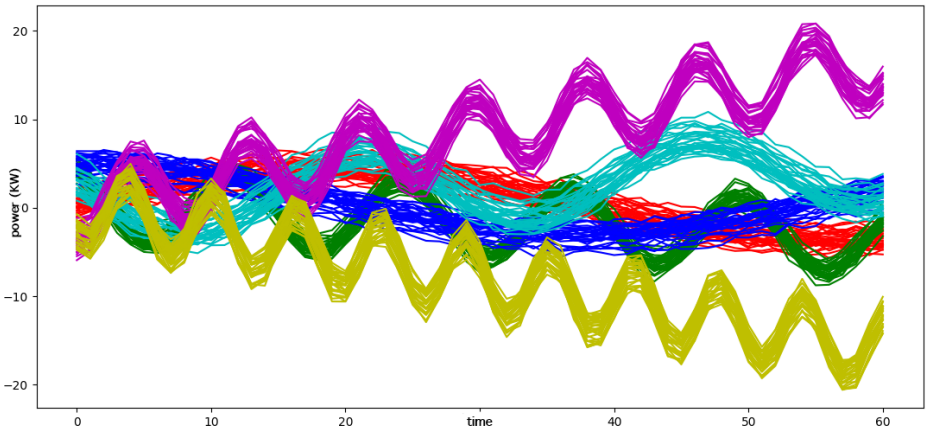


Fig. 10. The clusters identified by our algorithm [17].

step of half an hour, a day, a week and a month. The proximity between Enedis delivery points is evaluated on each resampled time series, each resampled time series corresponding to a characteristic of a site. After the Enedis dataset is constructed, the algorithm described in section 3 is applied on the time series.

5.5 Result Analysis on real dataset

Figure 12, displays the grouping of 50 Enedis time series representing all the clusters. Three clusters have been identified, in the green cluster there are two peaks per day, one in the morning, one in the evening, in the red clusters there is a single peak per day that lasts half the day, and in the blue cluster the consumption is constant during the day.

The algorithm identified relevant clusters in the sense that each items shares a characteristic with items in its cluster that it does not share with items in a different cluster.

6 Discussion and future work

The results we have shown on a real dataset are preliminary. To fully exhibit the potential of this algorithm, the clustering will have to be applied to a richer data set. This data set should include relevant features extracted from the consumption time series as well as physical characteristics of the buildings (such as the site's floor area, the type of heating, the insulation material, etc.). Correlation between the sites consumption and meteorological environment will also be a feature used for future works. By taking all these elements into account, the relevance of the clusters identified will be greatly improved.

There are two possibilities regarding the identified building clusters:

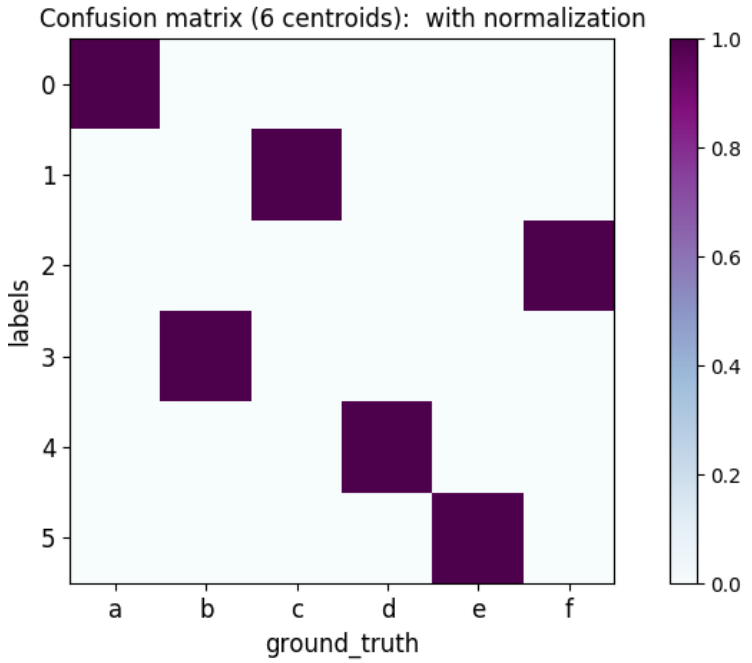


Fig. 11. Confusion matrix of the clusterization [17].

- The clusters correspond to an already defined classification i.e. the clusters can be compared to ground truth. For example, the clusters identified might correspond to the usage of the sites. In this case, we will implement semi-supervised learning and by using Machine Learning to tune the hyperparameters of the algorithm we will optimize the ARI index of our clustering.
- The clusters do not correspond to any known classification of buildings. In that case, we will have to apply knowledge extraction methods as well as energy experts' insight to give meaning to the newly found taxonomy of buildings.

Because the energy performance key indicators are not the same depending on the type of building [4], the insight given on building types will enhance energy performance evaluation and recommendation.

7 Conclusion

Building energy performance is a major challenge of the 21st century because of its important impact on climate change. Allowed by the growth of energy data, clustering of building based on consumption profile is a promising solution to efficiently identify the most relevant action to take for such complex energy systems. After a presentation of the state of art of the clustering methods, we

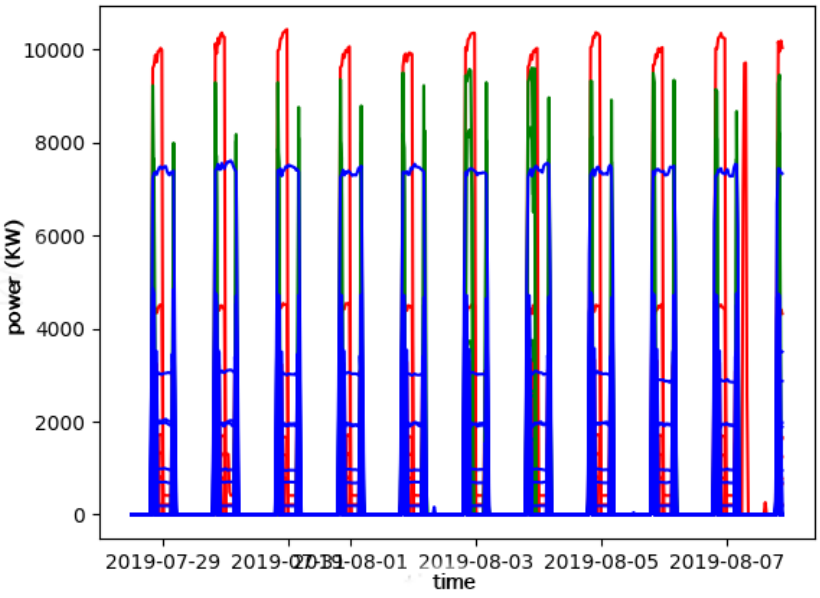


Fig. 12. Clustering of the Enedis time series [17]

propose a novel approach based on pretopology. The presented framework using pretopology allows for the multi-criteria hierarchical clustering of any finite set of items. Having a hierarchical structure gives insight into the similarities between building at different scales and therefore should provides a more refined understanding of the families and subfamilies of consumption profiles. The algorithm developed in Python for the construction of a Hierarchical Clustering of sites exploits this framework. The validation and visualization tools developed to test our algorithm allowed us to demonstrate visually and through ARI the relevance of the method on generated datasets as well as on real consumption time series dataset. The results demonstrate the potential of this solution for hierarchical clustering of heterogeneous systems.

Acknowledgements

This paper is the result of research conducted at the energy data management company *Energisme*. We thank *Energisme* for the resources that have been made available to us and Julio Laborde for his assistance with the conception of our pretopological hierarchical algorithm library.

References

1. Ahat, M., Amor, S.B., Bui, M., Bui, A., Guérard, G., Petermann, C.: Smart Grid and Optimization. American Journal of Operations Research **03**(01), 196–206 (2013). https://doi.org/10.4236/ajor.2013.31A019, http://www.scirp.org/journal/doi.aspx?DOI=10.4236/ajor.2013.31A019
2. Auray, J.P., Bonnevay, S., Bui, M., Duru, G., Lamure, M.: Prétopologie et applications : un état de l'art. Studia Informatica Universalis (Hermann) **7**, 27–44 (01 2009)
3. Belmandt, Z.: Manuel de prétopologie et ses applications. Hermès science publications (1993)
4. Boemi, S.N., Tziogas, C.: Indicators for buildings' energy performance. In: Energy Performance of Buildings, pp. 79–93. Springer (2016)
5. Bosom, J., Scius-Bertrand, A., Tran, H., Bui, M.: Multi-agent architecture of a mibes for smart energy management. Innovations for Community Services. I4CS 2018 **863** (2018)
6. Bosom, J.: Conception de microservices intelligents pour la supervision de systèmes sociotechniques: application aux systèmes énergétiques. Ph.D. thesis, Université Paris sciences et lettres (2020)
7. Fleischhacker, A., Lettner, G., Schwabeneder, D., Auer, H.: Portfolio optimization of energy communities to meet reductions in costs and emissions. Energy **173**, 1092 – 1105 (2019). https://doi.org/https://doi.org/10.1016/j.energy.2019.02.104, http://www.sciencedirect.com/science/article/pii/S0360544219303032
8. Gao, X., Malkawi, A.: A new methodology for building energy performance benchmarking: An approach based on intelligent clustering algorithm. Energy and Buildings **84**, 607 – 616 (2014). https://doi.org/https://doi.org/10.1016/j.enbuild.2014.08.030, http://www.sciencedirect.com/science/article/pii/S0378778814006720
9. Guerard, G., Pichon, B., Nehai, Z.: Demand-response: Let the devices take our decisions. In: SMARTGREENS. pp. 119–126 (2017)
10. Guérard, G., Ben Amor, S., Bui, A.: A context-free smart grid model using pretopologic structure. In: 2015 International Conference on Smart Cities and Green ICT Systems (SMARTGREENS). pp. 1–7 (2015)
11. Habib, U., Zucker, G., Blochle, M., Judex, F., Haase, J.: Outliers detection method using clustering in buildings data. In: IECON 2015-41st Annual Conference of the IEEE Industrial Electronics Society. pp. 000694–000700. IEEE (2015)
12. Iglesias, F., Kastner, W.: Analysis of similarity measures in times series clustering for the discovery of building energy patterns. Energies **6**(2), 579–597 (2013)
13. Kriegel, H.P., Kröger, P., Sander, J., Zimek, A.: Density-based clustering. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **1**(3), 231–240 (2011)
14. Laborde, J.: Pretopology, a mathematical tool for structuring complex systems: methods, algorithms and applications. Ph.D. thesis, EPHE (2019)
15. Largeron, C., Bonnevay, S.: A pretopological approach for structural analysis. Information Sciences **144**(1-4), 169–185 (2002)
16. Le, T.V.: Classification prétopologique des données: application à l'analyse des trajectoires patients. Ph.D. thesis, Lyon 1 (2007)
17. Levy, L.N., Bosom, J., Guérard, G., Amor, S.B., Bui, M., Tran, H.: Application of pretopological hierarchical clustering for buildings portfolio. In: SMARTGREENS. pp. 228–235 (2021)

18. Li, K., Ma, Z., Robinson, D., Lin, W., Li, Z.: A data-driven strategy to forecast next-day electricity usage and peak electricity demand of a building portfolio using cluster analysis, cubist regression models and particle swarm optimization. Journal of Cleaner Production **273**, 123115 (2020). https://doi.org/https://doi.org/10.1016/j.jclepro.2020.123115, http://www.sciencedirect.com/science/article/pii/S0959652620331607
19. Li, K., Yang, R.J., Robinson, D., Ma, J., Ma, Z.: An agglomerative hierarchical clustering-based strategy using shared nearest neighbours and multiple dissimilarity measures to identify typical daily electricity usage profiles of university library buildings. Energy **174**, 735 – 748 (2019). https://doi.org/https://doi.org/10.1016/j.energy.2019.03.003, http://www.sciencedirect.com/science/article/pii/S0360544219304074
20. Lu, Y., Tian, Z., Peng, P., Niu, J., Li, W., Zhang, H.: Gmm clustering for heating load patterns in-depth identification and prediction model accuracy improvement of district heating system. Energy and Buildings **190**, 49 – 60 (2019). https://doi.org/https://doi.org/10.1016/j.enbuild.2019.02.014, http://www.sciencedirect.com/science/article/pii/S0378778818308326
21. Marquant, J.F., Bollinger, L.A., Evins, R., Carmeliet, J.: A new combined clustering method to analyse the potential of district heating networks at large-scale. Energy **156**, 73 – 83 (2018). https://doi.org/https://doi.org/10.1016/j.energy.2018.05.027, http://www.sciencedirect.com/science/article/pii/S0360544218308478
22. Miller, C.: Screening Meter Data: Characterization of Temporal Energy Data from Large Groups of Non-Residential Buildings. Ph.D. thesis, ETH Zurich (2016)
23. Mills, E.: Building commissioning: a golden opportunity for reducing energy costs and greenhouse gas emissions in the united states. Energy Efficiency (2011)
24. Wang, S., Liu, H., Pu, H., Yang, H.: Spatial disparity and hierarchical cluster analysis of final energy consumption in china. Energy **197**, 117195 (2020). https://doi.org/https://doi.org/10.1016/j.energy.2020.117195
25. Xu, D., Tian, Y.: A comprehensive survey of clustering algorithms. Annals of Data Science **2**(2), 165–193 (2015)