Application of Pretopological Hierarchical Clustering for Buildings Portfolio

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*
{*f_author. s_author*}*@devinci.fr*

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

Abstract: Our paper deals with the problem of the comparison of heterogeneous energy consumption profiles for energy optimization. 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 in order to establish a relevant and effective recommendations system. Comparing sites to extract similar profiles refers to a machine learning set of methods called clustering. To answer this problematic, pretopology is used to model the sites' consumption profiles and a multi-criteria hierarchical clustering algorithm, using the properties of pretopological space, has been developed using a Python library. The pretopological hierarchical clustering algorithm is able to identify the clusters and provide a hierarchy between complex items. Tested on benchmarks of generated time series (from literature and from french energy company), the algorithm is able to identify the clusters using Pearson's correlation with an Adjusted Rand Index of 1 and returns relevant results on real energy systems' consumption data.

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 continues to rise, 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[1].

There are various ways to decrease buildings' energy consumption (Guerard et al., 2017): social programs, incentive programs, new energies, energy efficiency, dynamic pricing, demand-response programs. Most of the time, buildings having the same profile of consumption are sensitive to similar programs.

However, the systems we study are not always buildings. They can be a building floor or simply a place inside a building. In consequence, it's more accurate to talk about **sites**.

Sites present an important heterogeneity both in intrinsic properties and geographic situation (Miller, 2016). In addition, the scales of analysis are various both in time (consumption time series are analysed 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

[1]http://www.eia.gov/

compare a site to another.

Unfortunately, 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.

A comparison between similar sites might be meaningful to understand the performance of a new site. Comparing different sites to categorize them by proximity is called clustering. 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 optimisation 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 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 (Miller, 2016).

The key contribution of this paper is to provide a clustering method adapted to portfolio analysis based on a pretopological framework.

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 its application as a clustering method. The section 4 shows a pedagogical example of the presented method. We conclude in the section 5.

2 LITERATURE REVIEW

Formally, clustering refers to a set of unsupervised machine learning methods which group unlabeled items in clusters. In this section, we present clustering methods and their application on energy systems. The journal paper of Iglesia et al. in Energies (Iglesias and Kastner, 2013) presents a deeper analysis of clustering in energy system. To consult an exhaustive list of clustering algorithms, we invite you to read Xu et Al. survey (Xu and Tian, 2015).

There are four classes of clustering algorithms with their pros and cons: centroid-based clustering, density-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 such methods, a cluster is a set of items such that an item in a cluster is nearest to the center of a cluster than to the center of any other cluster. The center of a cluster is called a centroid, the average of all the points in the cluster, or a medoid, the most representative point of a cluster. The most known centroid-based algorithm is the *K-means* algorithm and its extensions. *K-means* is a powerful tool for clustering but it requires to determine in advance how many clusters the algorithm should find.

Therefore, centroid-based algorithms are sensitive to initial conditions. Clusters vary in size and density and include outliers (isolated item) to the nearest cluster. Lastly, centroid-based algorithms don't scale with the number of items and dimensions. In those cases, centroid-based algorithms are combined with principal component analysis or spectral analysis to be more effective.

About the portfolio analysis in energy systems, Gao et al. (Gao and Malkawi, 2014) benchmark multidimensional energy use dataset using a *k-means* algorithm. Freischhacker et al. (Fleischhacker et al., 2019) design a spatial aggregation method, combined with *k-means*, based on city blocks' characteristics to reduce reductions due to energy use.

Density-based clustering: In density-based clustering, a cluster is a set of items spread in the data space over a contiguous region of high density of items. Items located in low-density regions are typically considered noise or outliers (Kriegel et al., 2011). The most known methods in this class are Density-Based Spatial Clustering of Applications with Noise (*DBSCAN*) and its extensions.

The formation of clusters is sensitive to two parameters: the density and the reachability. Hence, the clusters are distinct depending on those parameters. The main advantages are this density-based clustering algorithm does not require a-priori specification and it is able to identify noisy data while clustering. It fails in case of neck type datasets and it does not work well in case of high dimensionality data.

About the portfolio analysis in energy systems, Li et al. (Li et al., 2020) present a density-based

method with a particle swarm optimization of parameters of buildings portfolio. Their method forecasts next-day electricity usage thanks to the clustering. Marquant et al. (Marquant et al., 2018) use a density and loads based algorithm to facilitate large-scale modelling and optimisation of urban energy systems.

Hierarchical clustering: Hierarchical clustering is usually a procedure to transform a proximity matrix into a sequence of hierarchically structured partitions.

There are two methods of hierarchical clustering: ascending (or agglomerating) or descending (or dividing). The ascending methods begin with disjointed classes and place each of the items in an individual class. Based on the proximity matrix, the procedure searches at each step for the two closest classes, merges them, and then snaps into a second partition. The process is repeated to construct a sequence of nested partitions in which the number of classes decreases as the sequence progresses until a unique class contains all the items. The descending methods do the inverse process.

The primary problem with those algorithms is to define the criterion of grouping or aggregation criterion of two classes, i.e. a distance measure. Sites are defined as complex systems (Ahat et al., 2013; Bosom et al., 2018; Guérard et al., 2015). They are defined with numerical and categorical data as well as time series, calculating a distance between two items is challenging and doesn't allow to use each characteristic of the site in a relevant way. Another drawback is the difficulty to identify an accurate number of clusters, especially in a large dataset.

About the portfolio analysis in energy systems, Wang et al. (Wang et al., 2020) analyse the spatial disparity of final energy consumption in China thanks to hierarchical clustering and spatial autocorrelation. Li et al. (Li et al., 2019) implement an agglomerative hierarchical clustering-based strategy to identify typical daily electricity usage profiles.

Distribution-based clustering: The application to large spatial databases raises the following requirements for clustering algorithms: no input parameters or the strict minimum, clusters with arbitrary shape. Distribution-based clustering produces clusters which assume concisely defined mathematical models underlying the items, a relatively plausible assumption for some items distributions.

Most of the time, mathematical models are based on Gaussian distribution, multinomial or multivariate normal distribution. The clusters are considered fuzzy, which means an item may be in various clusters at a defined percent. The most known algorithm is Expectation-Maximization (EM) clustering with Gaussian mixture models (GMM). That way, the GMM algorithm provides two parameters to describe the shape of the clusters: the mean and the standard deviation. The chief drawback of those algorithms is that it cannot work on categorical dimensions.

About the portfolio analysis in energy systems, Lu et al. (Lu et al., 2019) use a GMM clustering for heating load patterns identification. Habib et al. (Habib et al., 2015) provide a EM clustering to detect outliers in energy buildings 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 relation between any set of items and a bigger set of items. It is therefore adapted to the creation of a hierarchical structure. It is based on the concept of abstract space. In such a space the nature of the item is not relevant, it is rather the relations and property linking the items to another that matters. This allows us to manipulate heterogeneous and complex items such as our sites. Because of that, pretopology can be considered as a mathematical tool for modeling the concept of proximity for complex systems (Auray et al., 2009). Pretopology is, therefore, the approach chosen to build our hierarchical clustering.

3 PRETOPOLOGY

In this section we explain the key concepts and definitions of pretopology, such as pretopological space and pseudo-closure. Then, we provide the main algorithm for the pretopological hierarchical clustering.

3.1 Pretopological space

Let us start with some definitions.

Definition 1. *A pseudoclosure function* $a : \wp(U) \rightarrow \wp(U)$ *on a set of items* 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

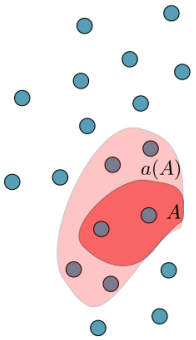

Figure 1: Example of a pseudoclosure function (Laborde, 2019).

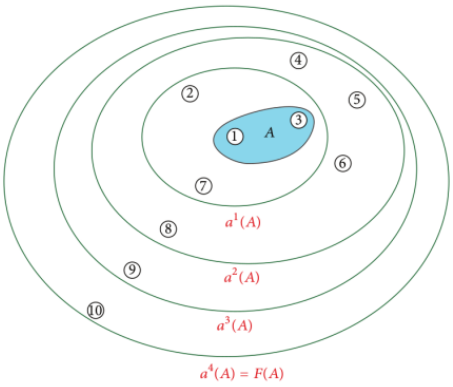

Figure 2: Closure of set A (Laborde, 2019).

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

Definition 3. *In a pretopological space, we can find the closure by repeatedly applying the pseudoclosure operator to the set and its subsequent images until it stops expanding.*

Definition 4. *In a pretopological space the closure of a part A of U is the smallest closure containing A. Denoted $F(A)$ (see Figure 2).*

A pretopological space is defined by establishing a relation between any set of items and a bigger set. Each step of a pseudoclosure is interesting in the construction of a hierarchy. An example of pseudoclosure function is shown in Figure 1.

Now let us present our framework formalizing a pretopological space adapted from Julio Laborde works (Laborde, 2019). In this framework each pretopological space is characterized by a tuple $(G, \Theta, DNF(.))$, where:

- $G = \{G_1(V_1, E_1), G_2(V_2, E_2), ..., G_n(V_n, E_n)\}$ is a set of n weighted directed graphs.

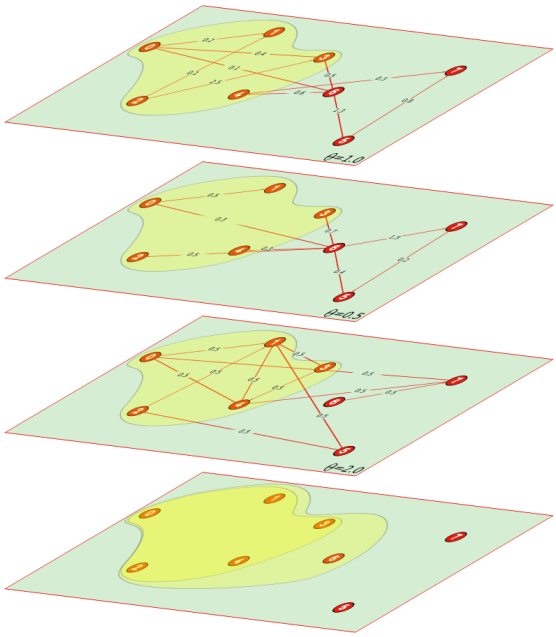

Figure 3: Example of a pseudoclosure under the framework (Laborde, 2019).

- $\Theta = \{\theta_1, \theta_2, ..., \theta_n\}$ is a set of n thresholds, each associated to one graph.
- $DNF(.) : (\wp(U), U) \rightarrow \{True, False\}$, where $\wp(U)$ is the power set of U, which is a boolean function expressed as a positive disjunctive normal form in terms of the n boolean functions $V_1(A, x), ..., V_n(A, x)$, each associated to a graph, and whose truth value depends on the set A and the item x.

We determine if an item $x \in U$ belongs to the pseudoclosure of a set A in the following way:

- $\forall V_i(A, x)$, $V_i(A, x) = True \iff \sum_{e_{xy} \in G_i, y \in A} w(e_{xy}) \geq \theta_i$, where e_{xy} is the edge going from x to y, and $w(e)$ is the weight of the edge e.
- The item $x \in U$ will belong to the pseudoclosure of $A \iff$ the $DNF(.)$ evaluates to True

Simply put, this checks in every graph if the sum of the weights of the edges going from the item x to the items inside A is bigger than the threshold associated to the graph. When this happens, the boolean variable associated to that graph acquires a value of True, otherwise it gets a value of False. If $DNF(.)$ evaluates to True with those values for the boolean functions $V_i(A, x)$, then the items belongs to the pseudoclosure. An example of this is illustrated in figure 3.

3.2 Algorithms

This section describes the algorithms used for the construction of a closure and to build a hierarchical clustering of sites.

The clustering procedure is structured in three phases:

1. Calculation of a family of elementary sets called seeds.
2. Construction of the subsets by applying pseudo-closure iteratively.
3. Establishing a structural relation among all the subsets using quasihierarchy.

Calculation of a family of seeds: The purpose of this procedure is to generate a small set from which the elementary closure subset will be calculated. Calculating those seeds from each item cause a lot of calculation. This can be avoided by starting with sets of 2, 3 or 4 items.

A seed of multiple items is calculated by proximity. The distance measure depends on the attributes (numeric attributes, binary attributes, nominal attributes, ordinal attributes, mixed-type attributes).

Construction of the subset: This algorithm applies the pseudo-closure on the seeds. That will produce bigger sets. The pseudo-closure is applied iteratively until providing closure. Since we have started applying the pseudoclosure on seeds, the closure we have determined are called closure subsets. By keeping the structure of all the pseudo-closure between the seed and the closure subset, the algorithm keeps a range of sets defining a hierarchy.

Construction of the hierarchy from subsets: Our objective is now to determine a hierarchy between the subsets, called quasihierarchy. The algorithm is build following these rules:

- Two subsets are connected only if their intersection is not empty.
- The more of a set A is contained in a set B, the stronger the relation from A to B.
- The bigger the set B is compared to A, the lesser the part of A that should be contained in B to have a strong relation going from A to B. In other words, a very big set will attract smaller ones even if their intersection is not very large.
- Two sets that have a mutualy strong relation are considered equivalent, unless one is contained in the other, in which case the bigger of the two is a parent of the other in the quasihierarchy.

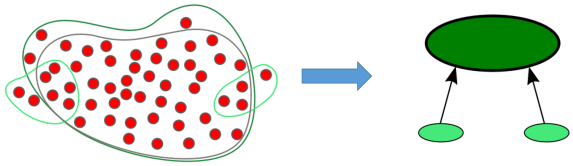

Figure 4: Construction of the quasihierarchy (Laborde, 2019).

The algorithm takes as input a set of subsets and a threshold, and returns a quasihierarchy by (see Figure 4):

- Quantifying the relation between each pair of sets determined with non-empty intersection.
- Creating a link in the quasihierarchy when the value of the relation is above the threshold.
- Sets having links going in both direction are considered equivalent and one is selected randomly.
- The resulting closures with the respective links determine the quasihierarchy.

3.3 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 to show the results.

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

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

To evaluate multi-criteria clustering, the size of an item is added as a second parameter. 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 helps to evaluate our method in different kinds of situations and make corrections or adjustments easily.

Visualization tool: To observe the results of the classification, the program colors each of the biggest sets determined by our algorithm in a unique color. The validation tool is tested with two groups of items with both big and small size and a 2-dimensional position. Items are shown on figure 5. In this example,

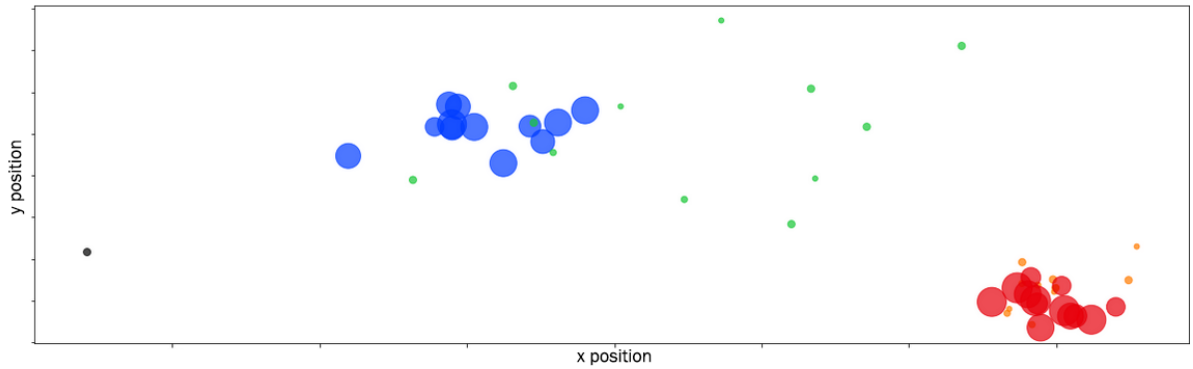

Figure 5: The four clusters determined by our algorithm using both size and position as parameters, on a 2D disks dataset.

four clusters have been determined: blue, green, orange and red. The black dot at the leftest side of the figure 5 is an item identified by the algorithms as an outlier. For example red and orange items are close to one another yet separated into two clusters because of their different sizes and orange and green points are similar in size yet divided into two sets because of their different positions.

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

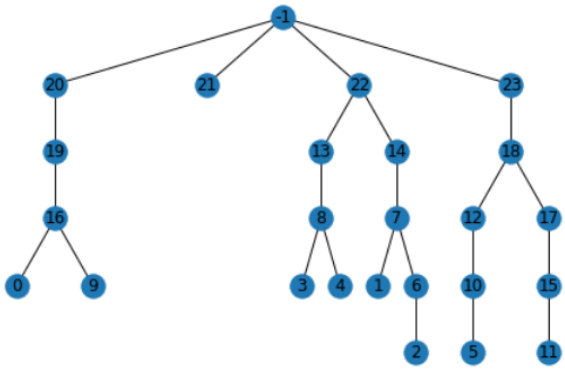

Figure 6: A tree representing the pseudohierarchy relation between each intermediate set from the seed to the cluster.

For instance, the hierarchy presented in figure 6 shows the relations between the sets determined by our algorithm applied to the dataset displayed on figure 5. This tree presents only the sets of more than two items. We can recognize the four clusters that were colored on figure 5, they are entitled 20, 21, 22 and 23. The figure 7 displays the set 14 which is a child of the set 21 (colored in green) in the hierarchical clustering. This hierarchy determines large groups of relatively similar items and provides more details about smaller groups of very similar items.

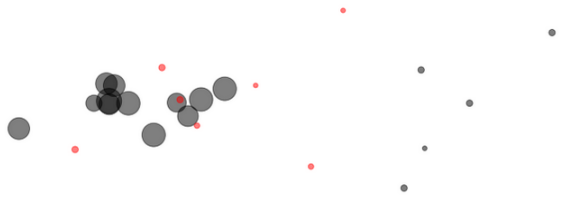

Figure 7: The subset 14 in red representing a subgroup of the green clusters (subset 22) in figure 5.

4 EXPERIMENTS AND RESULTS

4.1 Benchmark dataset

Because the main data we have on sites are power consumption time series, the clustering of a set of time series had to be tested, visualized and evaluated. This section presents this test set and the results of our algorithm. The created test set, composed of six clusters is presented on figure 8. Each clusters is composed of 30 time series of 60 points.

The similarity measure used to establish the value between two items is 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 according to the clusters it had determined (see figure 8).

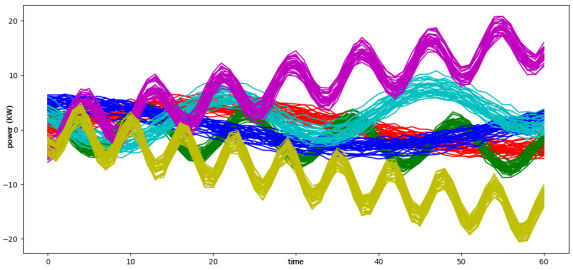

Figure 8: The clusters identified by our algorithm.

4.2 Results analysis on benchmark dataset

The program identified the exact 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). Since we perfectly identified the clusters the ARI of our clustering is 1. The figure 9 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.

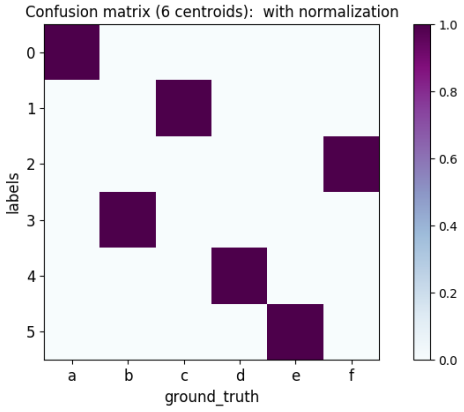

Figure 9: Confusion matrix of the clusterization.

4.3 Real dataset

This dataset is build from Enedis (Power Grid Operator in France) consumption times series of 400 sites over a year. It is resampled with a time step of half an hour, a day, a week and a month. The proximity between the Enedis delivery points is evaluated on each resampled time series, each resampled time series corresponding to one characteristic of a site. Once the Enedis data set is build, the algorithm described in section 3 is applied on the time series.

4.4 Result Analysis on real dataset

Figure 10, displays the clustering of 50 Enedis time series representing the whole clusters. Three clusters were identified, in the red clusters there is a single peak per day that lasts for half the day, in the green cluster there are two peaks a day, one in the morning, one in the evening, and in the blue cluster the consumption is constant during the day.

The algorithm has identified relevant clusters in the sense that each items shares one trait with the items of their clusters that they don't share with items of a different cluster.

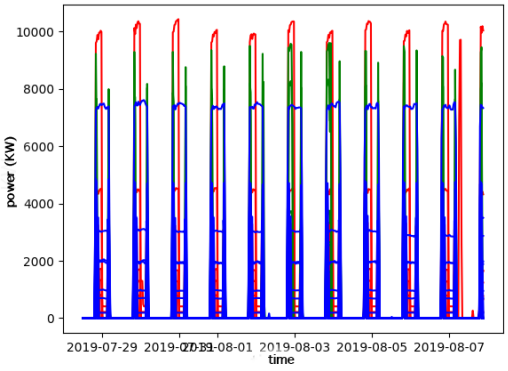

Figure 10: Clustering of the Enedis time series

5 CONCLUSION

Important energy savings can be made by acquiring better insight over building consumption profiles. To determine what savings can be made on a building, an important element is to compare its energy consumption with the one of other buildings. However, energy systems (building and sites) are heterogeneous, complex, and are described by numerical and categorical data as well as consumption time series and are therefore hard to compare to one another. Hence the need for an adapted clustering method. Studying the state of the art methods of clustering made us create a new hierarchical algorithm based on pretopology. Indeed, pretopology theory provides tools to determine relation of proximity between heterogeneous sets. These algorithms were developed in a Python library alongside tools of visualization and evaluation. Results on generated test data sets demonstrated the efficiency and the relevance of this library.

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

Ahat, M., Amor, S. B., Bui, M., Bui, A., Guérard, G., and Petermann, C. (2013). Smart Grid and Optimization. *American Journal of Operations Research*, 03(01):196–206.

Auray, J.-P., Bonnevay, S., Bui, M., Duru, G., and Lamure, M. (2009). Prétopologie et applications : un état de l'art. *Studia Informatica Universalis (Hermann)*, 7:27–44.

Bosom, J., Scius-Bertrand, A., Tran, H., and Bui, M. (2018). Multi-agent architecture of a mibes for smart energy management. *Innovations for Community Services. I4CS 2018*, 863.

Fleischhacker, A., Lettner, G., Schwabeneder, D., and Auer, H. (2019). Portfolio optimization of energy communities to meet reductions in costs and emissions. *Energy*, 173:1092 – 1105.

Gao, X. and Malkawi, A. (2014). A new methodology for building energy performance benchmarking: An approach based on intelligent clustering algorithm. *Energy and Buildings*, 84:607 – 616.

Guerard, G., Pichon, B., and Nehai, Z. (2017). Demand-response: Let the devices take our decisions. In *SMARTGREENS*, pages 119–126.

Guérard, G., Ben Amor, S., and Bui, A. (2015). A context-free smart grid model using pretopologic structure. In *2015 International Conference on Smart Cities and Green ICT Systems (SMARTGREENS)*, pages 1–7.

Habib, U., Zucker, G., Blochle, M., Judex, F., and Haase, J. (2015). Outliers detection method using clustering in buildings data. In *IECON 2015-41st Annual Conference of the IEEE Industrial Electronics Society*, pages 000694–000700. IEEE.

Iglesias, F. and Kastner, W. (2013). Analysis of similarity measures in times series clustering for the discovery of building energy patterns. *Energies*, 6(2):579–597.

Kriegel, H.-P., Kröger, P., Sander, J., and Zimek, A. (2011). Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):231–240.

Laborde, J. (2019). *Pretopology, a mathematical tool for structuring complex systems: methods, algorithms and applications*. PhD thesis, EPHE.

Li, K., Ma, Z., Robinson, D., Lin, W., and Li, Z. (2020). 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.

Li, K., Yang, R. J., Robinson, D., Ma, J., and Ma, Z. (2019). 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.

Lu, Y., Tian, Z., Peng, P., Niu, J., Li, W., and Zhang, H. (2019). Gmm clustering for heating load patterns in-depth identification and prediction model accuracy improvement of district heating system. *Energy and Buildings*, 190:49 – 60.

Marquant, J. F., Bollinger, L. A., Evins, R., and Carmeliet, J. (2018). A new combined clustering method to analyse the potential of district heating networks at large-scale. *Energy*, 156:73 – 83.

Miller, C. (2016). *Screening Meter Data: Characterization of Temporal Energy Data from Large Groups of Non-Residential Buildings*. PhD thesis, ETH Zurich.

Wang, S., Liu, H., Pu, H., and Yang, H. (2020). Spatial disparity and hierarchical cluster analysis of final energy consumption in china. *Energy*, 197:117195.

Xu, D. and Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193.