PROCEEDINGS OF ECOS 2023 - THE 36TH INTERNATIONAL CONFERENCE ON
EFFICIENCY, **C**OST, **O**PTIMIZATION, **S**IMULATION AND ENVIRONMENTAL IMPACT OF ENERGY SYSTEMS
25-30 JUNE 2023, LAS PALMAS DE GRAN CANARIA, SPAIN

Clustering and typification of urban districts for energy system modelling

Joseph Loustau[a, CA], Dorsan Lepour[a], Cédric Terrier[a] and François Maréchal[a]

[a] EPFL, Lausanne, Switzerland, joseph.loustau@epfl.ch

Abstract:

The interest in Urban Systems has been growing due to the necessary decarbonisation of city energy systems. Decision tools are developed using mathematical optimisation to enable proper decision-making in the transition process. The concept of energy communities - or district energy hub - is expected to have an impact on the energy system at both regional and national scales. However, the shift towards distributed energy systems complexifies the model due to more integrated subsystems and requires larger spatial boundaries to increase self-consumption and decrease grid stresses. The computational power required to model and optimise such systems is to rise drastically.

This work proposes to curtail the large computing needs by typifying the districts of a city, using clustering techniques. Accordingly clustered districts can be optimised by solving a typical district from the group and scaling its solution to the others. The clustering features considered are the districts energetic characteristics: the energy demands on one side, and the endogenous resources on the other. Data are normalised and a principal component analysis is conducted. Two clustering algorithms are investigated: a centroid-based (Kmedoids) and a density-based (GaussianMixture). The ideal number of clusters is determined by maximising the intra-cluster similarity and minimising the inter-cluster similarity, and the final clustering stability is evaluated through the Rand Index.

The method is applied on the case study of a typical European urban area and the two algorithms lead to two distinct typification. The clusterings are used to run an energy hub optimisation for the whole region and the results are compared to the one obtained without archetypes for validation. The results between the two approaches show no significant differences while a considerable computing time reduction is achieved.

Keywords:

Urban energy systems, clustering, energy modelling, energy communities.

1. Introduction

1.1. Background

Among all, the building and construction industry accounts for an estimated 37% of the global operational energy and 37% of the process-related CO_2 emissions, worldwide [1]. Additionally, the world urban population represents 55% of the total and is expected to grow to 6 billion people [2] by 2050 (70% of total). Today, two-thirds of the global energy consumption come from cities, which emit more than 70% of the total greenhouse emissions [2]. Priority needs then to be put on city energy systems decarbonisation. Those considerations lead to a growth of interest in Urban Energy Systems. According to the *Net Zero Emissions (NZE) by 2050* report [3],

https://doi.org/10.52202/069564-0288

decarbonisation should be driven by high electrification. The conventional energy systems (*e.g.* fossil-fuels and boiler for Space Heating (*SH*)) would be replaced by devices and technologies that require renewable energy vectors. In the *NZE* scenario, 1.8 billion heat pumps and 1.2 billion solar thermal systems combined to 7'500 TWh of building-integrated PV would need to be installed [1].
Consequently, the exploitation of local energy resources is expected to grow, shifting the current electrical grid to a more decentralised one; households or organisations will consume energy from the grid at times, while at other times they will produce surplus energy that they can injected into the grid. The installation of energy devices must then be carefully done so that it answers the grid constraints. To ensure the stable operation of the grid at nominal frequency, the grid must always be balanced. While the current system - where few plants are responsible for the production, from the higher voltage level - can easily adapt the plant turbines frequency to keep the balance, with non-drivable inputs on the grid at the low-voltage level, it would be much more difficult to reequilibrate the imbalances [4]. Moreover, high peak production power can also overload the transformers or cause transmission bottlenecks [5]. The grid restrictions make it challenging to determine what technologies should be installed for a building to meet its demand, as optimal solutions depend on the context. Thus, the energy problem must be solved on a case-by-case basis.

To enable good energy-wise policies to help the renewable technologies penetration, an important lever for action is to provide efficient decision tools, able to propose a set of solutions to a specific energy context [6]. The decision-making tools model the system design (*energy hub*) and its possible components and provide an analysis, leaning on mathematical optimisation or simulation. Traditionally, the building is considered as the energy hub perimeter.

1.2. Gaps and contribution

However, to obtain results at a city scale, solving each hub would be too demanding in terms of computational power. To tackle this issue, either the model has to lose accuracy or the data volume must be reduced. [7, 8] developed the use of Machine Learning techniques to cluster the buildings of a city. The building stock is thus simplified using archetypes (i.e. average reconstructed buildings or sample buildings). But, the change towards distributed energy systems encourages the extension of the *energy hub* spatial boundaries so that a higher self-consumption can be achieved, hence releasing some stress over the grid. Indeed, [9] shows that when maximising the electricity generated locally (e.g., solar panels), the interaction of the building energy system with the grid should be lowered. To do so, the *energy hub* considered in this research is the district, which allows interactions between buildings.

This research proposes a method to typify urban districts using clustering techniques, thus combining the benefits of buildings interactions and typification. An application of the method is proposed using the canton of Geneva in Switzerland as a case study.

2. Methods

The method's premise is that by modelling a reduced number of districts, elected as *representative districts*, the results obtained draw the same conclusion as the ones that would be obtained by modelling every districts.
From this premise, the workflow proposed to enable large scale urban energy modelling is described in Figure 1. This paper covers the methods to find the representative districts.

https://doi.org/10.52202/069564-0288

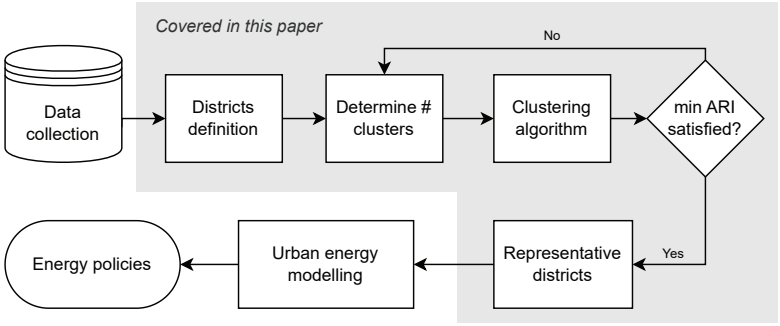


Figure 1: Workflow diagram for large scale urban energy modelling for integrated and distributed systems.

2.1. Definition of a district energy hub

Intuitively, a district is an area. Defining the *District Energy Hub (DEH)* consists in defining its spatial boundaries. A spatial attribute that must be common to all the buildings to consider within the same district is investigated. Because the initial target of the *DEH* is to limit exchanges with the grid, it is necessary that the buildings within the district boundaries can interact. Also, the aggregated load that the transformer has to handle must be taken realistically into account. Hence, the low-voltage transformer is the common attribute within a district in this research.
To be relevant, the clustering is designed so that its input features are the same as the input variables that describe the energy system in the model. A *DEH* is characterised by its infrastructures, the energy imported, the resources, and the on-site energy demands, as detailed in Fig. 2.

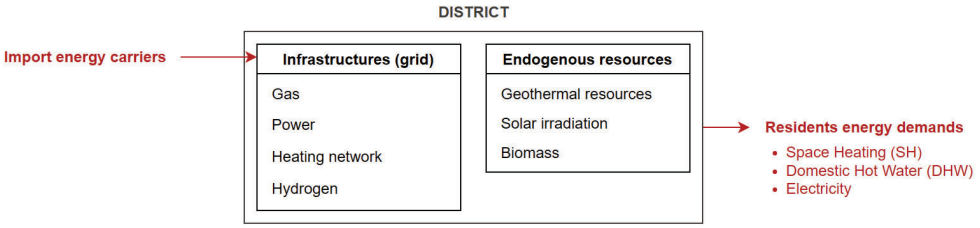


Figure 2: Key characteristics describing a district.

To restrain the number of features and facilitate the solving, the data go through a *Principal Component Analysis (PCA)* before the clustering. A PCA allows reducing a dataset dimensionality while preserving as much of the variability (*i.e.* the statistical information) as possible [10]. It also helps to interpret which features have the most impact on the clustering result. It is implemented in Python, using the *scikit* library.

2.2. Clustering

In *Urban Building Energy Modelling (UBEM)*, the algorithm that comes up most often and that has proven its effectiveness is *Kmeans* [11]. In this method, a derivative of Kmeans is implemented: the *Kmedoids* algorithm. It uses the concept of medoids instead of the mean, which diminishes the influence of the outliers, making it more robust. To make it operational to elliptical distribution, a second algorithm is proposed, the *GaussianMixture* algorithm. Those two algorithms are implemented in Python, using the *scikit-learn* library [12].

2.2.1. *Kmedoids*

Kmedoids require an initial number of clusters to run. If no previous knowledge of the dataset allows knowing this parameter, performance measures are used to determine which number of clusters gives the best result. Three validity indexes are proposed in this method. To put in equations those indexes, the following notation is used:

- The set of clusters $\mathbf{C} = \{C_1, ..., C_I, ..., C_K\}$ with $1 \leq I \leq K$,
- The set of observations $\mathbf{y} = \{y_1, ..., y_i, ..., y_n\}$ with $1 \leq i \leq n$,
- The subset of observations attributed to the cluster $y_i \in C_I$, $i \leq n$, $\mathbf{y_I} = \{y_{I,1}, ..., y_{I,j}, ..., y_{I,n_{C_I}}\}$ with $1 \leq j \leq n_{C_I}$,
- $\bar{y}$ is the centroid of $\mathbf{y}$. Similarly, $\bar{y}_I$ is the centroid of $\mathbf{y_I}$.

- **Elbow index:** the elbow method consists in computing the sum of squared distances (distortion score or inertia (1)) for each point to its assigned centroid [13]. The optimal number of clusters is the one where the drop in the distortion score is the most important.
- **Silhouette index:** for a given observation and $y_{I,j}$, the silhouette score measures the mean distance to all points of its cluster (3a) and compares it to the mean distance to all points of the neighboring cluster (3b). $s(I, j)$ has a score of 1 if $y_{I,j}$ is a perfect match with cluster C_I and a score of -1 with the neighboring cluster. The silhouette index is the mean of all coefficients $s_{I,j}$ (3d).
- **Calinski-Harabasz index:** it estimates the cohesion of clusters. It evaluates the distance from points in a cluster to the centroids and the distance from the centroids to a global centroid (2).

  The elbow index is commonly used. However, according to [14], *CH index* and *silhouette index* are the ones performing the best to evaluate a Kmeans clustering.

$$SS(\mathbf{C}) = \sum_{C_I \in \mathbf{C}} \sum_{y_{I,j} \in \mathbf{y_I}} (y_{I,j} - \bar{y}_I)^2 \quad (1)$$

$$CH(\mathbf{C}) = \frac{n-K}{K-1} \frac{\sum_{C_I \in \mathbf{C}} n_{C_I} \cdot d(\bar{y}_I, \bar{y})}{\sum_{C_I \in \mathbf{C}} \sum_{y_{I,i} \in \mathbf{y_I}} d(y_{I,i}, \bar{y}_I)} \quad (2)$$

$$a(I, j) = \frac{1}{n_{C_I} - 1} \sum_{m \in C_I,\, j \neq m} d(y_{I,j}, y_{I_m}) \quad (3a)$$

$$s(I, j) = \frac{b(I, j) - a(I, j)}{max(a(I, j), b(I, j))} \quad (3c)$$

$$b(I, j) = \min_{J \neq I} \frac{1}{n_{C_J}} \sum_{y_{J,m} \in C_J} d(y_{I,j}, y_{J_m}) \quad (3b)$$

$$S(\mathbf{C}) = \frac{1}{n} \sum_{C_I \in \mathbf{C}} \sum_{y_{I,i} \in C_I} s(I, i) \quad (3d)$$

https://doi.org/10.52202/069564-0288

With this being defined, algorithm 1 is used to determine the optimal number of clusters k_{opt}.

Once k_{opt} is determined, the clustering in itself can be done. It is repeated 500 times and the one with minimal inertia is kept as the final clustering. If the right number of clusters was used, the clustering should be *stable*, *i.e.* observations should be labeled the same way from one iteration to the other. Rand [15] has created an index to evaluate this stability. When two clusterings (obtained from different algorithms or number of clusters) are to be distinguished, the one with the higher Rand index is kept. Moreover, for a clustering to be valid, it should have a minimum level of stability. [16] stated equation (4), defining it depending on the clustering sizes (r and s). If $ARI \leq min\ ARI$, clustering is rejected.

$$min\ ARI = \left[1 - \frac{1}{2}\binom{r+s-1}{2}\left[\binom{r}{2}^{-1} + \binom{s}{2}^{-1}\right]\right]^{-1} \tag{4a}$$

$$if\ r = s:\ min\ ARI = \frac{-r}{3r-2} \tag{4b}$$

2.2.2. *GaussianMixture*

The process is similar to the one for *Kmedoids*. The *GaussianMixture* algorithm needs a number of components and the shape of the covariance matrix to work. Again, if no previous knowledge helps to determine those parameters, scores are used to find them.
Because the *GaussianMixture* relies on probabilities, the best clustering is the one that maximises the likelihood. The likelihood function evaluates the joint probability of observed data as a function of the chosen statistical model. Given a set of n training vectors **y**, the GMM likelihood function can be written as in (5).

However, adding components helps increase the likelihood while it may lead to overfitting the data. Criteria introduce penalty terms on the number of parameters to solve the issue.

- **Akaike Information Criterion (AIC):** defined by (6a), it should be as small as possible. It is an efficient criterion when the model is very complex and is chosen in this context.
- **Bayesian Information Criterion (BIC):** defined by (6b), the BIC should also be as small as possible. It has consistency (meaning it would asymptotically select the candidate model having the correct structure), as its penalty term contains n.

$$p(y \mid \lambda) = \prod_{i}^{n} p(y_i \mid \lambda) \tag{5}$$

$$AIC = 2 \cdot ln(k_\lambda) - 2 \cdot ln(\hat{L}) \tag{6a}$$

$$BIC = k \cdot ln(n) - 2 \cdot ln(\hat{L}) \tag{6b}$$

where $\hat{L}$ is the maximised value of the likelihood function of the GMM defined in (5), k the numbers of estimated parameters of the model.

Algorithm 2 describes the steps to determine the key parameters. Once they are selected, as for *Kmedoids*, the Rand index is computed.

2.3. Case study: the Canton of Geneva

In this research, data are collected over the canton of Geneva, an area containing a typical mid-size European city, a peri-urban area, and some rural areas.
The districts are defined according to the LV transformers to which the buildings are connected.

Because the position of transformers is not publicly available, this research uses the synthetic grid approach from the work of [17], which estimates the position of transformers based on the buildings' energy demands. Most of the data concerning buildings comes from **QBuildings**, a database developed at the IPESE laboratory. The buildings data are aggregated over districts and normalised by the total *Energetic Reference Area (ERA)* by district.

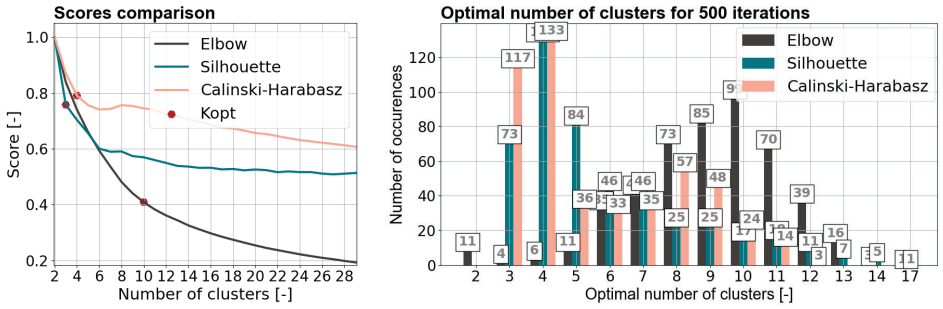
3. Results and discussion

The results and discussion seek to demonstrate the applicability of the method developed on a typical European city intending to subsequently run an energy system optimisation model. First, the clusterings obtained with the two algorithms are analysed. Second, the results of the optimisation with and without the clustering are compared to ensure the validity of its use for decision-making purposes.

3.1. Clustering results

3.1.1. Optimal number of clusters

As a first step, the optimal number of clusters should be determined to run the algorithm. Figure 3(a) shows the mean score obtained for each number of clusters for the three indexes and Fig. 3(b) shows the number of occurrences where each number of clusters gives the best score, with the *Kmedoids* algorithm.



(a) Mean scores obtained for the three indices.

(b) Occurrences where a number of clusters is found optimal.

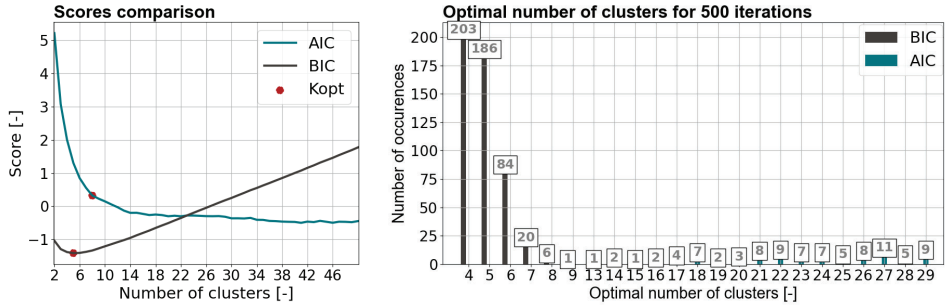
Figure 3: Results of the optimal number of clusters investigation, Kmedoids, repeated 500 times.

The silhouette and the CH index agree on 2 as the optimal number. However, 2 clusters are insufficient to discriminate between the different optimisation problems that may be posed. Therefore, the next optimum is looked at. According to Fig. 3(a), for CH, the line breaks at 4 and then there is a new local maximum at 8 and for the silhouette index, the break is at 3. This is confirmed by Fig. 3(b). The elbow occurs the most between 8, 9, and 10.

In comparison, Figs. 4(a) and 4(b) indicate an optimal number of clusters between 4, 5, or 6 for the BIC index. The AIC obtains an elbow-like shape curve where the main change of curvature happens at 5 and 8. Note that only the full covariance matrix shape is shown, as it has always obtained the best score.

From those two results, 3, 4, 5, and 8 are the best number of clusters to investigate.

https://doi.org/10.52202/069564-0288



(a) Mean scores obtained for the two indices.

(b) Occurrences where a number of clusters is found optimal.

Figure 4: Results of the optimal number of clusters investigation, GaussianMixture, repeated 500 times.

3.1.2. Clustering

With the k_{opt} search results, algorithms are run to compute the Rand index and keep the most stable clustering. Table 1 displays the Rand index computed by repeating the clustering 1000 times, with the different numbers of clusters. The most stable clustering is obtained with the *GaussianMixture* algorithm and in particular when the requested number of clusters is 4 and 5.

Table 1: Rand index by clustering 1000 times, varying number of clusters.

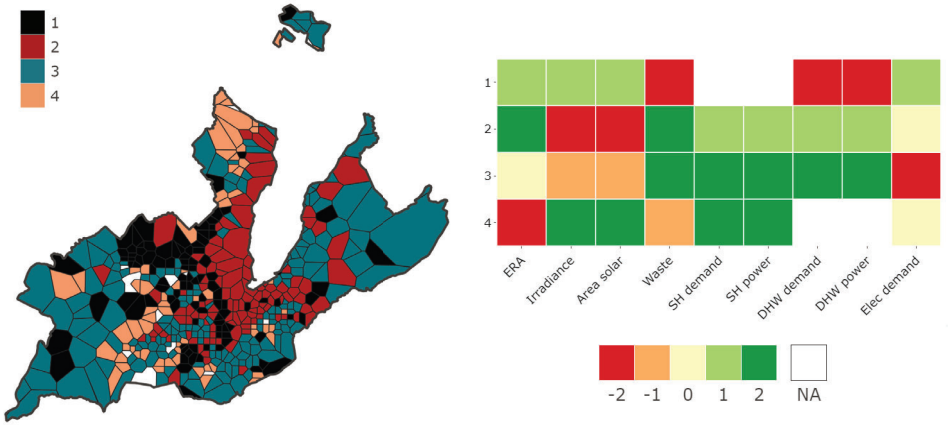
Number of clusters	3	4	5	8
Mixture	-	0.717	0.693	0.399
Kmedoids	0.331	0.383	0.287	0.388

Because they were the best options according to the BIC index and they have a satisfying Rand score, the 4^{th} and 5^{th} options are kept. While doing the clustering with 5 different clusters, it often resulted in an empty 5^{th} cluster. Therefore, 4 different clusters is the selected optimal number. The result of this clustering can be visualised in Fig. 5(a). A heat map from the input features (Fig. 5(b)) is used to understand what differentiates clusters from each other.

Looking at those elements, the clusters can be characterized and described in the following way:

Cluster	DESCRIPTION
1	Districts dominated by industrial, commercial, and administrative buildings
2	Residential belt around the lake, with high buildings density and low solar potential per capita
3	Peri-urban residential buildings, with low ERA density (i.e., single-family detached houses)
4	Peri-urban buildings with important solar availability (high roof on ERA ratio)

The hot water demand, along with the electrical demand, have the biggest influence on the clustering as they can vary a lot according to the building type.



(a) Map of the canton of Geneva coloured by clusters.

(b) Heat map evaluation of each feature for each cluster. Features are evaluated on a scale of -2 to 2. -2: Very low, -1: Low, 0: Average, 1: High, 2: Very High. Missing cells correspond to a non-decisive case, indicated by parsed values inside the cluster.

Figure 5: Clustering results and analysis on the canton of Geneva, operated with *GaussianMixture*, 4 components and full covariance matrix.

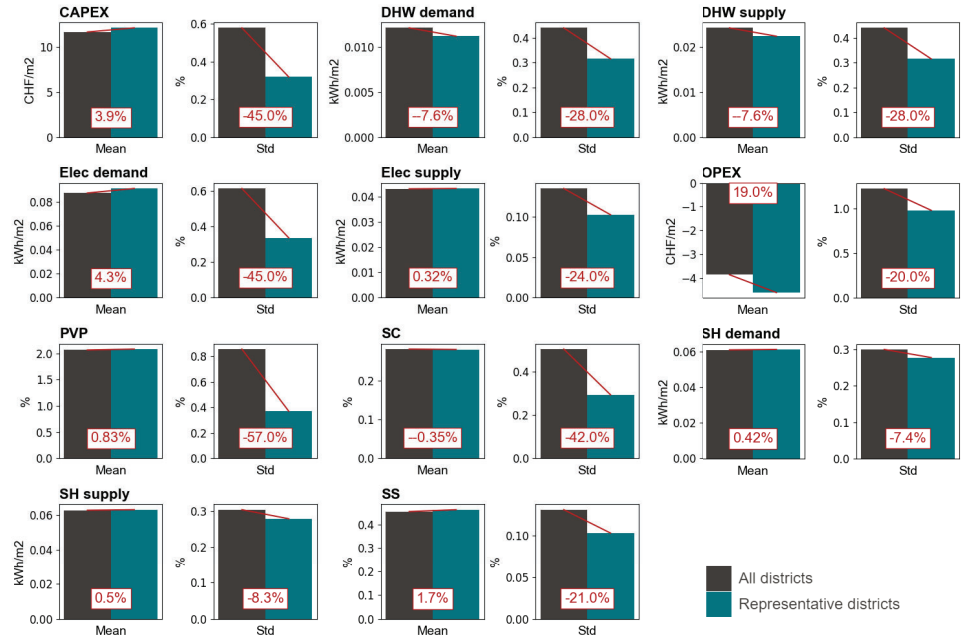
3.1.3. Comparison of optimisation results with and without district typification

To ensure the validity of the method, one should ensure that when using typical districts to do the optimisation (namely *Representative districts (RD)* approach in what follows), similar results are obtained as the ones where every district has been optimised specifically (*All districts (AD)*).

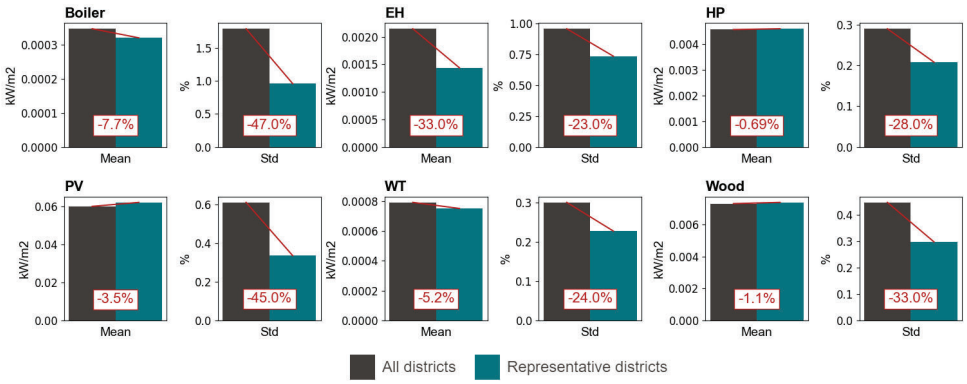
To do so, the decision variables of the model and the resulting *Key Performance Indicators (KPI)* are compared between the two approaches. The global mean obtained on the canton of Geneva should show no significant distinction. The mean is weighted by the total ERA that each cluster represents for *RD*. The intra-cluster variabilities are also compared to the standard deviation computed over the whole canton.
Using an optimisation tool that models an energy hub and determines the best energy system, finding the optimal energy system for the 468 districts of Geneva took 54h21m23s with *AD* and 48m18s with *RD*, i.e. a reduction in the calculation time of 98.5%. The comparison of their respective decision variables and *KPIs* are shown in Fig.6(a) and Fig.6(b). The major diminishing of the standard deviation for each evaluated metric confirms that the patterns among the inputs that participated to create the district clusters are meaningful as well for patterns in the optimisation results. Moreover, there was no significant difference between the mean obtained using the *AD* or the *RD* approach, with two exceptions: the OPEX and Electrical Heater installation size, which have high variability within the canton. Therefore, it is more difficult to accurately represent these indicators with typical districts.

Nevertheless, the overall results obtained show an excellent performance of the method to facilitate the optimisation of districts over a large scale.

https://doi.org/10.52202/069564-0288



(a) Main KPIs. *CAPEX*=Capital Expenditure, *DHW*=Domestic Hot Water, *Elec*=Electricity, *OPEX*=Operationnal Expenditure, *PVP*=Photovoltaïc Panel Penetration, *SC*=Self-Consumption, *SS*=Self-Sufficiency



(b) Installed capacity for the energy technologies. *Boiler*=Natural Gas Boiler, *EH*=Electrical Heater, *HP*=Heat Pump, *PV*=Photovoltaïc Panel, *WT*=Water Tank

Figure 6: Comparison of results obtained with the two approaches. *All districts* stands for the results obtained by optimising every district in the canton of Geneva, while *Representative districts* is obtained by running the representative districts and scaling their solution to the districts of their cluster. The figures in red indicate the representative difference between the two approaches.

4. Conclusion

The objective of this paper was to provide a method to enable urban energy modelling at a large scale and high accuracy, for integrated and distributed systems. It uses the concept of archetypes, introduced for buildings, and adapts it to the district scale by defining the low-voltage transformer as the reference energy hub.
The method suggests the use of two clustering algorithms - *Kmedoids* and *GaussianMixture*, depending on the data set - and provided as output the list of the districts labeled by cluster. It was validated on a case study (the canton of Geneva), with the following conclusions:

- It can be delicate to determine the optimal number of clusters;
- The model solving computation time over the whole region was reduced by 98.5%, with respect to the classic approach;
- The results between the two approaches show no significant difference with 4 clusters, although the high variability of certain decisions can lack a good representation with this approach.

The presented work opens up possibilities in the solving of more complex systems at an even larger scale. However, its main limitation is the selected number of clusters, this parameter being the most critical into the analysis of the energy system model results. Therefore, when planning urban energy systems with this method, it will be essential to carry out a correct analysis of the neighbourhood typology.

5. Fundings

The research published in this report was carried out with the support of the Swiss Federal Office of Energy SFOE as part of the SWEET project acronym. The authors bear sole responsibility for the conclusions and the results of the presented publication.

Appendix A Algorithms

Algorithm 1 Find the optimal number of clusters for Kmedoids

Input: A set of observations $\mathbf{y}$
Output: A vector $\mathbf{K_{opt}}$ counting the occurrences where a number of clusters has been determined as the best

$\mathbf{K_{opt}} \leftarrow \mathbf{0}$
$Bestscore \leftarrow 0$
for $i \leq 1000$ **do**
  for $k = 2 \leq 30$ **do**
    $\mathbf{C} \leftarrow Kmedoids(\mathbf{y}, n_{medoids} = k)$
    Compute $SS(\mathbf{C})$, $S(\mathbf{C})$ or $CH(\mathbf{C})$
    if $Index \geq Bestscore$ **then**
      $Bestscore \leftarrow Index$
      $k_{best} \leftarrow k$
    end if
  end for
  $\mathbf{K_{opt}}(k_{best}) \leftarrow \mathbf{K_{opt}}(k_{best}) + 1$
end for
return $\mathbf{K_{opt}}$

https://doi.org/10.52202/069564-0288

Algorithm 2 Find the optimal number of clusters for GaussianMixture

Input: A set of observations **y**
Output: Two vectors $\mathbf{K_{opt}}$ and $\mathbf{Cov_{opt}}$ counting the occurrences where a number of clusters has been determined as the best

$\mathbf{K_{opt}} \leftarrow \mathbf{0}$
$\mathbf{Cov_{opt}} \leftarrow \mathbf{0}$
$Bestscore \leftarrow 0$
for $i \leq 1000$ **do**
  for $k = 2 \leq 30$ **do**
    for $shape \in \{full, diag, tied, spherical\}$ **do**
      $\mathbf{C} \leftarrow GaussianMixture(\mathbf{y}, n_{components} = k, m_{cov} = shape)$
      Compute $AIC(\mathbf{C})$ or $BIC(\mathbf{C})$
      if $Index \leq Bestscore$ **then**
        $Bestscore \leftarrow Index$
        $k_{best} \leftarrow k$
        $shape_{best} \leftarrow shape$
      end if
    end for
  end for
  $\mathbf{K_{opt}}(k_{best}) \leftarrow \mathbf{K_{opt}}(k_{best}) + 1$
  $\mathbf{Cov_{opt}}(shape_{best}) \leftarrow \mathbf{Cov_{opt}}(shape_{best}) + 1$
end for
return $\mathbf{K_{opt}}$, $\mathbf{Cov_{opt}}$

References

[1] United Nations Environment Programme. *2022 Global Status Report for Buildings and Construction: Towards a Zero-emission, Efficient and Resilient Buildings and Construction Sector*. 2022-11. URL: https://wedocs.unep.org/20.500.11822/41133.

[2] World Bank. *Urban development overview*. World Bank. URL: https://www.worldbank.org/en/topic/urbandevelopment/overview (visited on 06/06/2022).

[3] International Energy Agency IEA. *Transition to sustainable buildings: strategies and opportunities to 2050*. Paris: IEA Publ, 2013. 284 pp. ISBN: 978-92-64-20241-2.

[4] Gideon A. H. Laugs, René M. J. Benders, and Henri C. Moll. "Balancing responsibilities: Effects of growth of variable renewable energy, storage, and undue grid interaction". In: *Energy Policy* 139 (Apr. 1, 2020), p. 111203. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2019.111203. URL: https://www.sciencedirect.com/science/article/pii/S0301421519307876 (visited on 06/23/2022).

[5] Karl-Kiên Cao, Johannes Metzdorf, and Sinan Birbalta. "Incorporating Power Transmission Bottlenecks into Aggregated Energy System Models". In: *Sustainability* 10.6 (June 2018). Number: 6 Publisher: Multidisciplinary Digital Publishing Institute, p. 1916. ISSN: 2071-1050. DOI: 10.3390/su10061916. URL: https://www.mdpi.com/2071-1050/10/6/1916 (visited on 06/20/2022).

[6] Shimeng Hao and Tianzhen Hong. "The Application of Urban Building Energy Modeling in Urban Planning". In: *Rethinking Sustainability Towards a Regenerative Economy*. Ed. by Maria Beatrice Andreucci et al. Future City. Cham: Springer International Publishing,

2021, pp. 45–63. ISBN: 978-3-030-71819-0. DOI: 10.1007/978-3-030-71819-0_3. URL: https://doi.org/10.1007/978-3-030-71819-0_3 (visited on 03/07/2023).

[7] Ina De Jaeger et al. “A building clustering approach for urban energy simulations”. In: *Energy and Buildings* 208 (Feb. 1, 2020), p. 109671. ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2019.109671. URL: https://www.sciencedirect.com/science/article/pii/S0378778819313271 (visited on 06/19/2022).

[8] Sara Torabi Moghadam et al. “A new clustering and visualization method to evaluate urban heat energy planning scenarios”. In: *Cities* 88 (May 1, 2019), pp. 19–36. ISSN: 0264-2751. DOI: 10.1016/j.cities.2018.12.007. URL: https://www.sciencedirect.com/science/article/pii/S0264275118307625 (visited on 06/19/2022).

[9] Luise Middelhauve. “On the role of districts as renewable energy hubs”. PhD thesis. 2022.

[10] Ian T. Jolliffe and Jorge Cadima. “Principal component analysis: a review and recent developments”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (Apr. 13, 2016). Publisher: Royal Society, p. 20150202. DOI: 10.1098/rsta.2015.0202. URL: https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202 (visited on 05/12/2022).

[11] Chao Wang et al. “Data acquisition for urban building energy modeling: A review”. In: *Building and Environment* 217 (June 1, 2022), p. 109056. ISSN: 0360-1323. DOI: 10.1016/j.buildenv.2022.109056. URL: https://www.sciencedirect.com/science/article/pii/S0360132322002955 (visited on 06/20/2022).

[12] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[13] M A Syakur et al. “Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster”. In: *IOP Conference Series: Materials Science and Engineering* 336 (Apr. 2018), p. 012017. ISSN: 1757-8981, 1757-899X. DOI: 10.1088/1757-899X/336/1/012017. URL: https://iopscience.iop.org/article/10.1088/1757-899X/336/1/012017 (visited on 05/13/2022).

[14] Olatz Arbelaitz et al. “An extensive comparative study of cluster validity indices”. In: *Pattern Recognition* 46.1 (Jan. 2013), pp. 243–256. ISSN: 00313203. DOI: 10.1016/j.patcog.2012.07.021. URL: https://linkinghub.elsevier.com/retrieve/pii/S003132031200338X (visited on 05/13/2022).

[15] William M. Rand. “Objective Criteria for the Evaluation of Clustering Methods”. In: *Journal of the American Statistical Association* 66.336 (Dec. 1, 1971). Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1971.10482356, pp. 846–850. ISSN: 0162-1459. DOI: 10.1080/01621459.1971.10482356. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1971.10482356 (visited on 05/19/2022).

[16] José E. Chacón and Ana I. Rastrojo. “Minimum adjusted Rand index for two clusterings of a given size”. In: *Advances in Data Analysis and Classification* (Feb. 9, 2022). ISSN: 1862-5355. DOI: 10.1007/s11634-022-00491-w. URL: https://doi.org/10.1007/s11634-022-00491-w (visited on 05/20/2022).

[17] “Countrywide PV hosting capacity and energy storage requirements for distribution networks: The case of Switzerland”. In: *Applied Energy* (2020). Ed. by Rahul Gupta, Fabrizio Sossan, and Mario Paolone. DOI: 10.1016/j.apenergy.2020.116010.

 https://doi.org/10.52202/069564-0288