Multi-Objectives, Multi-Period Optimization of district energy systems: I-Selection of typical operating periods

Samira Fazlollahi[a,b,*], Stephane Laurent Bungener[b], Pierre Mandel[c], Gwenaelle Becker[a], Francois Maréchal[b]

[a] *Veolia Environnement Recherche et Innovation (VERI), 291 avenue Dreyfous Ducas, 78520 Limay, France*
[b] *Industrial Process and Energy Systems Engineering, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland*
[c] *Veolia Environnement Recherche et Innovation (VERI), Chemin de la digue BP 76, 78603 Maisons-Laffitte, France*

Abstract

The long term optimization of a district energy system is a computationally demanding task due to the large number of data points representing the energy demand profiles.

In order to reduce the number of data points and therefore the computational load of the optimization model, this paper presents a systematic procedure to reduce a complete data set of the energy demand profiles into a limited number of typical periods, which adequately preserve significant characteristics of the yearly profiles. The proposed method is based on the use of a *k-means* clustering algorithm assisted by an ϵ-constraints optimization technique. The proposed typical periods allow us to achieve the accurate representation of the yearly consumption profiles, while significantly reducing the number of data points.

The work goes one step further by breaking up each representative period into a smaller number of segments. This has the advantage of further reducing the complexity of the problem while respecting peak demands in order to properly size the system.

[*] Corresponding author

Email addresses: Samira.Fazlollahi@epfl.ch, (Samira Fazlollahi), Stephane.Bungener@epfl.com, (Stephane Laurent Bungener), Pierre.Mandel@veolia.com, (Pierre Mandel), Gwenaelle.Becker@veolia.com, (Gwenaelle Becker), Francois.Marechal@epfl.ch (Francois Maréchal)

Two case studies are discussed to demonstrate the proposed method. The results illustrate that a limited number of typical periods is sufficient to accurately represent an entire equipments' lifetime.

Keywords: Typical periods, District energy systems, Mixed Integer Linear Programming, Evolutionary algorithm, Multi-objective optimization, Cluster analysis, $k - means$ algorithm

1. Introduction

Multi period mixed integer linear programming (MILP) is an effective method for designing distributed energy systems [1, 2]. It provides guidance for choosing optimal system configurations for minimizing costs and environmental impacts. The evaluation of the district energy system performance requires the estimation of the investment and of the corresponding operating costs. The calculation of the operating costs should consider the hourly, daily and seasonal variations of energy demand and the contribution of each production unit. It is therefore necessary to extend the optimization to the multi-period model. Due to the high number of variables, such a detailed description of the system requires excessive computational resources for solving the MILP optimization model.

Using the typical periods provides an efficient alternative for reducing the number of variables. The notion of the limited typical periods relies upon the assumption that a year can be accurately represented by a limited set of periods. The term period describes a portion of time of a certain duration. It can be a set of days, weeks, working days or weekends, defined by a sequence of time steps, over the life time of the equipment.

The problem of selecting typical operating periods based on the energy demand variations has been approached in different ways. Maréchal et al. [3] proposed an evolutionary algorithm optimization approach to select typical production scenarios for an industrial cluster. Ortiga et al. [4] developed a graphical method to select a reduced number of periods that reproduce the heating and cooling load duration curves. Lozano et al. [5] reduced the yearly demand profiles to 24 days by defining two typical periods for each month, one using averaged data for a working day and the other using averaged data of weekends. Casini et al. [6] used 3 typical periods for the thermal demands, one per season (winter, summer and intermediary) and 24 typical periods for electricity sold to the grid. Mavrotas et al. [7] used 12 typical periods, using monthly averages. Balachandra and Chandru [8] proposed 9 representative daily load curves for electricity demand. Initially they grouped 365 days according to the 12 months of the year and then used discriminant analyses to regroup the data into 9 typical periods. Dominguez et al. [9] used a partitioning *k-medoids* method and MILP model for selecting N_k representative days. The number of typical periods N_k is selected by the user and therefore not optimized in their method. They defined $n \times (n+1)$ binary variables in the MILP model, with n the total number of data points of the demand profile. Due to the large number of integer variables, solving the optimization model for the lifetime of the equipment becomes a computationally demanding task with a high resolution time. Marton et al. [11] also proposed an order-specific clustering algorithm for selecting the typical periods for electricity demand.

In this paper, a new method is developed to select the typical periods from the multiple time-dependent demand profiles. The method selects these periods by applying the partitioning *k-means* clustering method [10] and the ϵ-constraints known as a parametric optimization [12]. The goal is to minimize the number of typical periods and maximize their quality simultaneously. The yearly profiles are grouped into N_k clusters. The closest period to the center of each cluster is considered to be the typical period for representing the cluster. Extreme demand days are afterwards superimposed as insulated clusters. They are used to size the system properly. We go further by breaking up each representative period into a smaller number of segments.

The proposed method ensures that the selected typical periods accurately represent the important characteristics of the original data, namely the profile trends and peak demands. However, the *k-means* clustering method does not guarantee finding the global optimal point. To overcome this issue, five performance indicators, representing five ϵ-constraints, are proposed in order to guarantee reaching a local optimal with an acceptable quality.

The novelty of this work lies in the selection of the typical periods based on the clustering method and the ϵ-constraints optimization technique. The developed method converges very quickly, a major advantage compared to other techniques.

2. Typical operating periods

A mixed integer linear model (MILP) is developed by [1] for optimizing the design and the operating strategy of a district energy system. Due to the stochastic variation of the hourly demand profile, a multi-period approach is needed. However, the hourly profile with 8760 time steps makes the optimization model difficult to solve.

One way to reduce the size of the energy system optimization model is to represent the yearly profile using a limited set of typical periods. It provides an efficient alternative to reduce the number of variables and the size of the optimization model. The centroid clustering algorithm is used in the present work for selecting the typical periods.

2.1. Centroid clustering algorithm: k-means

Clustering is an unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters) [14]. Reflecting the high interest of clustering in data analysis, many algorithms can be found in the literature [15]. All these algorithms basically derive from two approaches; hierarchical and partitional approaches [16]. Hierarchical methods produce a nested series of partitions while partitional methods produce only one [14].

In this work the *k-means* partitioning algorithm is used to define the typical periods. The first researcher who proposed explicitly the *k-means* algorithm in the multidimensional case was Steinhaus in 1956 [17]. Alternatively, other authors (i.e. [18] and [19]) working in different fields also proposed their own versions of the algorithm. Even though *k-means* was first proposed over 50 years ago, it is still one of the most widely used algorithms for clustering [20].

k-means is a **greedy** optimization algorithm, which minimizes the squared error over all N_k clusters (Eq.1):

$$\min \left[\sum_{k=1}^{N_k} \sum_{i=1}^{N_i} d(\hat{\mu}_k, \hat{x}_i) \times z_{i,k} \right] \tag{1}$$

$$\sum_{k=1}^{N_k} z_{i,k} = 1, \quad \forall i \tag{2}$$

$$d(\hat{\mu}_k, \hat{x}_i) = \sum_{a=1}^{N_a} \sum_{g=1}^{N_g} (\hat{\mu}_{k,a,g} - \hat{x}_{i,a,g})^2 \quad \forall k, i \tag{3}$$

$$\hat{x}_{i,a,g} = \frac{x_{i,a,g}}{\max\{x_{i,a,g} \quad \forall i \in \{1, ..., N_i\}, \forall g \in \{1, ..., N_g\}\}} \quad \forall i, a, g \tag{4}$$

$X = \{x_{i,a,g}\}$ ($i \in \{1, ..., N_i\}$, $a \in \{1, ..., N_a\}$, $g \in \{1, ..., N_g\}$) is a set of N_i independent observations, such as 365 days of a year, to be grouped into a set of N_k clusters (groups of similar days). N_a presents the type of attributes such as heating, cooling and electricity loads. N_g refers to the number of values (measurements) for each type of attribute (i.e. $N_g = 24$ for hourly measurements in a day). The number of measurements, N_g, should be the same for all types of attributes. For instance, if the solar irradiation is measured every 15 minutes, while only hourly values of ambient temperatures are available, then the solar irradiation should be summarized into the equivalent hourly data. In Eq.1, the normalized value of the original observation is denoted by $\hat{x}$ (Eq.4), $\hat{\mu}_k$ refers to the center of each cluster for the normalized data, while μ_k denotes the center, which is transformed to the original scale of observations. A binary variable $z_{i,k}$ is equal to 1 if and only if observation $\hat{x}_i$ is assigned to the cluster k. Each observation is assigned to only one cluster (Eq.2). $d(\hat{\mu}_k, \hat{x}_i)$ is a distance between observation $\hat{x}_i$ and the center of cluster $\hat{\mu}_k$. The most common choice is to consider the squared euclidean distance (Eq.3).

The *k-means* algorithm takes the following steps:

1. Choose (randomly or not) N_k cluster centres;
2. Assign each observation to the closest cluster centre;
3. Recompute the cluster centres using the current cluster memberships (by simply calculating the mean).
4. If a convergence criterion is not met, go to step 2. Convergence criteria can be:
    - Relative: no (or minimal) reassignment of patterns to new cluster centers,
    - Absolute: minimal decrease in squared error,
    - Practical: maximal number of evaluations.

2.1.1. Determining the optimal number of cluster

The *k-means* algorithm requires two user-specified parameters: firstly the cluster initialization or starting point (v), and secondly the number of clusters (N_k).

The result of the *k-means* clustering depends on the starting point (v), which relates to the non-deterministic character of the algorithm. One way to overcome this issue is to run the algorithm several times with different random starting points (i.e. $v \in \{1, ..., V_{max}\}$).

Various authors have proposed heuristics or statistical measures to set the optimal value of N_k (i.e. [21–23]). The general approach is to run *k-means* for a wide range of values of $N_k \in \{1, ..., N_{max}\}$ and to decide, based on some statistical measures and on domain expertise, which number of cluster works best.

In this study, the optimal number of clusters is assessed using three measures (Eq.5):

$$N_k^* : \quad [\min\{C(v^*_{N_k}), \forall N_k\}, \quad \max\{D(v^*_{N_k}), \forall N_k\}, \quad \min\{ESE(v^*_{N_k}), \forall N_k\}] \tag{5}$$

Subject to:

$$\textit{k-means}: \quad \min[\sum_{k=1}^{N_k}\sum_{i=1}^{N_i}(\sum_{a=1}^{N_a}\sum_{g=1}^{N_g}(\hat{\mu}^v_{k,a,g} - \hat{x}_{i,a,g})^2) \times z^v_{i,k}], \tag{6}$$

$$\forall v \in \{1, ..., V_{max}\}, \forall N_k \in \{1, ..., N_{max}\}$$

where;

$$C(v^*_{N_k}) = \min\{C(v_{N_k}), \quad \forall v\} \quad \forall N_k \in \{1, ..., N_{max}\} \tag{7}$$

$$D(v^*_{N_k}) = \max\{D(v_{N_k}), \quad \forall v\} \quad \forall N_k \in \{1, ..., N_{max}\} \tag{8}$$

$$ESE(v^*_{N_k}) = \min\{ESE(v_{N_k}), \quad \forall v\} \quad \forall N_k \in \{1, ..., N_{max}\} \tag{9}$$

$C(v_{N_k})$ denotes the average squared error, which evaluates the compact character of the clusters (the average intra-cluster distance);

$$C(v_{N_k}) = \frac{1}{N_k}\sum_{k=1}^{N_k}\sum_{i=1}^{N_i} z_{v,i,k} \times d_v(\hat{\mu}_k, \hat{x}_i), \quad \forall v, N_k \tag{10}$$

$D(v_{N_k})$ is the average inter-cluster distance, which evaluates the separation between the clusters;

$$D(v_{N_k}) = \frac{1}{N_k^2}\sum_{k=1}^{N_k}\sum_{j=1}^{N_k} d_v(\hat{\mu}_k, \hat{\mu}_j), \quad \forall v, N_k \tag{11}$$

$ESE(v_{N_k})$ is the statistical measure (Eq.12) proposed by [24], which evaluates the ratio of observed to the expected squared errors for N_k clusters.

The expected squared error is calculated by considering the squared error obtained with $N_k - 1$ clusters, under the assumption that the patterns have a uniform distribution. Obtaining a low value for the ratio means that the clustering obtained with N_k clusters is better defined than that obtained with $N_k - 1$ clusters. The measure is normalized, so that the values of k that yield small $ESE(v_{N_k})$ can be regarded as giving well-defined clusters, independently from the value of N_k [24].

$$ESE(v_{N_k}) = \begin{cases} 1 & if \quad N_k = 1, \forall v \\ \frac{N_k \times C(v_{N_k})}{\alpha_{N_k} \times (N_k - 1) \times C(v_{N_k - 1})} & if \quad C(v_{N_k - 1}) \neq 0, \forall N_k > 1, \forall v \\ 1 & if \quad C(v_{N_k - 1}) = 0, \forall N_k > 1, \forall v \end{cases} \tag{12}$$

$$\alpha_{N_k} = \begin{cases} 1 - \frac{3}{4 \times N_a \times N_g} & if \quad N_k = 2 \quad \& \quad N_a \times N_g > 1 \\ \alpha_{N_k - 1} + \frac{1 - \alpha_{N_k - 1}}{6} & if \quad N_k > 2 \quad \& \quad N_a \times N_g > 1 \end{cases} \tag{13}$$

$N_a \times N_g$ is the number of data set attributes and α_{N_k} is the weight factor. According to Eq.5, the best value for the number of clusters (N_k^*) should yield; a low value for the average intra-clusters distance ($C(v_{N_k}^*)$), a high value for the average inter-clusters distance ($D(v_{N_k}^*)$), and a low value for the $ESE(v_{N_k}^*)$ measure. It can be expressed by Eq.14 and Eq.15;

$$N_k^* : \quad R(N_k^*) = \min\{R(N_k), \quad \forall N_k\} \tag{14}$$

$$R(N_k) = \max\{R_C(v_{N_k}^*), \quad R_D(v_{N_k}^*), \quad R_{ESE}(v_{N_k}^*)\} \quad \forall N_k \tag{15}$$

In Eq.15, $R_C(v_{N_k}^*)$ is the rank of N_k clusters in the ascending order set of $C(v_{N_k}^*)$, $R_D(v_{N_k}^*)$ is the rank of N_k clusters in the descending order set of $D(v_{N_k}^*)$, $R_{ESE}(v_{N_k}^*)$ denotes the rank of N_k clusters in the ascending order set of $ESE(v_{N_k}^*)$, and $R(N_k)$ refers to the rank of N_k clusters. The N_k with the minimum rank (Eq.14) is chosen as the best option (N_k^*).

2.2. Typical periods' selection algorithm

The aim of the developed model is to minimize the number of typical periods and maximize their quality, which can be defined as a multi-objective optimization model (Eq.16).

$$\min_{\mu_k, z_{i,k}} \{\mathbf{N_k}\} \quad \max_{\mu_k, z_{i,k}} \{\textbf{Quality}\} \tag{16}$$

Subject to:

$$\textit{k-means} \quad : \quad \min \left[\sum_{k=1}^{N_k} \sum_{i=1}^{N_i} d(\hat{\mu}_k, \hat{x}_i) \times z_{i,k}\right]$$

Where $\mathbf{N_k}$ is the number of the typical periods, and the **Quality** is measured by using the following five performance indicators;

Profile deviation $\sigma^a_{profile,N_k}$: studies the accuracy of the original and typical period profiles compared to their averages (Eq.17). It is defined for each type of attribute such as hourly heating, cooling and electricity loads.

$$\sigma^a_{profile,N_k} = [\frac{1}{N_i \times N_g} \sum_{k=1}^{N_k} \sum_{i=1}^{N_i} \sum_{g=1}^{N_g} z_{i,k} \times [\frac{(x_{i,a,g} - \bar{x}_{i,a}) - (\mu_{k,a,g} - \bar{\mu}_{k,a})}{\bar{x}_a}]^2]^{\frac{1}{2}}, \quad \forall a \tag{17}$$

$\bar{\mu}_{k,a}$ (Eq.18) is the average value of the typical periods, $\bar{x}_{i,a}$ (Eq.19) is the average value of the i^{th} original observation, and $\bar{x}_a$ (Eq.20) is the average value for the entire original observations.

$$\bar{\mu}_{k,a} = \frac{1}{N_g} \sum_{g=1}^{N_g} \mu_{k,a,g} \quad \forall a, k \tag{18}$$

$$\bar{x}_{i,a} = \frac{1}{N_g} \sum_{g=1}^{N_g} x_{i,a,g} \quad \forall a, i \tag{19}$$

$$\bar{x}_a = \frac{1}{N_i \times N_g} \sum_{i=1}^{N_i} \sum_{g=1}^{N_g} x_{i,a,g} \quad \forall a \tag{20}$$

Deviation from the load duration curve of the average values of each period σ^a_{cdc,N_k}: compares the average values of the original observation and the typical periods for each type of attributes (Eq.21).

$$\sigma^a_{cdc,N_k} = \left[\frac{1}{N_i} \sum_{k=1}^{N_k} \sum_{i=1}^{N_i} z_{i,k} \times (\frac{\bar{x}_{i,a} - \bar{\mu}_{k,a}}{\bar{x}_{i,a}})^2\right]^{\frac{1}{2}}, \quad \forall a \tag{21}$$

Error in load duration curve deviation $ELDC^a_{N_k}$ of attribute a [9]: refers to the absolute deviation between the original and equivalent load duration curve for all data points (Eq.22), where LDC_o^a is created by sorting the original load profiles in a descending order. LDC_{e,N_k}^a is the load duration curve built with N_k typical periods for attribute a, and $p = 1, ..., N_g \times N_i$ is the number of data points with attribute a.

$$ELDC_{N_k}^a = \frac{\sum_{p=1}^{N_i \times N_g} |LDC_o^a(p) - LDC_{e,N_k}^a(p)|}{\sum_{p=1}^{N_i \times N_g} LDC_o^a(p)} \quad \forall a \tag{22}$$

Maximum load duration curve difference Δ_{LDC,N_k}^a: The extreme values of the LDC^a are very important for sizing the urban system. This indicator describes the relative difference in maximum loads between the original and the typical periods (Eq.23):

$$\Delta_{LDC,N_k}^a = \frac{\max(LDC_o^a(p)) - \max(LDC_{e,N_k}^a(p))}{\max(LDC_o^a(p))} \quad \forall a \tag{23}$$

Number of periods whose relative error is higher than γ **(**$\Delta_{prod,\gamma,N_k}^a$**)**: corresponds to the number of periods where the total equivalent load during a period is higher or lower than the original value, by a margin of γ (defined by users as an assumption) (Eq.25).

$$\Delta_{prod,\gamma,N_k}^a = \sum_{i=1}^{N_i} \sum_{k=1}^{N_k} y_{i_k,a} \quad \forall a \tag{24}$$

where:

$$y_{i_k,a} = \begin{cases} 1 \quad if \quad \sum_{g=1}^{N_g} \frac{z_{i,k} \times |x_{i,a,g} - \mu_{k,a,g}|}{x_{i,a,g}} > \gamma & \forall a, i, k \\ 0 \quad\quad otherwise \end{cases} \tag{25}$$

The multi-objective optimization problem (Eq.16) is solved by applying the ϵ-constraints concept [12]. The application of the ϵ-constraints algorithm for multi-objective optimization of urban energy systems has been reviewed in [13]. The second objective, $\max\{$**Quality**$\}$, is therefore defined as a set of constraints with an upper limit of ϵ_j^a. The auxiliary model of Eq.16 is expressed as Eq. 26:

$$\min_{\mu_k, z_{i,k}} \{\mathbf{N_k}\} \tag{26}$$

Subject to:

$$\begin{array}{lcl} \textit{k-means} & : & \min \left[\sum_{k=1}^{N_k} \sum_{i=1}^{N_i} d(\hat{\mu}_k, \hat{x}_i) \times z_{i,k} \right] \\ \sigma^a_{profile,N_k} & \leqslant & \epsilon^a_1 \quad \forall a \\ \sigma^a_{cdc,N_k} & \leqslant & \epsilon^a_2 \quad \forall a \\ ELDC^a_{N_k} & \leqslant & \epsilon^a_3 \quad \forall a \\ \Delta^a_{LDC,N_k} & \leqslant & \epsilon^a_4 \quad \forall a \\ \Delta^a_{prod,\gamma,N_k} & \leqslant & \epsilon^a_5 \quad \forall a \end{array}$$

The proposed algorithm proceeds as follows:

- Step 1: Break down the energy profile into N_i observations made up of N_a attributes with N_g values.

$$\hat{x}_{i,a,g} \quad 1 \leq i \leq N_i \quad 1 \leq a \leq N_a \quad 1 \leq g \leq N_g \tag{27}$$

- Step 2: Set constraints on the maximum allowable values of the performance indicators (ϵ^a_j) and apply the *k-means* algorithm for selecting N^*_k periods. This step should be repeated with several random starting points as long as the constraints are not met. If this step does not converge into a feasible solution after V_{max} evaluations (i.e. $V_{max} = 1000$), it implies that the constraints set on the performance indicators are too constraining and they should either be relaxed by the user or the number of typical periods should be increased. Finally the result will be μ^*_k typical periods.

- Step 3: If the optimal number of periods, N^*_k, and the indicators' threshold, ϵ^a_j, are not known, run *k-means* for values of $N_k \in \{1, ..., N_{max}\}$, with $v \in \{1, ..., V_{max}\}$ random starting points (i.e. $V_{max} = 1000$, $N_{max} = 20$). The *k-means* clustering is very quick and running it with 1000 random initial points takes a matter of minutes.

- Step 4: Calculate the values of performance indicators, $ESE(v_{N_k})$ measure, the average intra-clusters distance ($C(v_{N_k})$) and the average inter-clusters distance ($D(v_{N_k})$) for $N_k \in \{1, ..., N_{max}\}$ and $v \in \{1, ..., V_{max}\}$ evaluations.

- Step 5: Draw the Pareto frontier of each performance indicator (the smallest value of the indicators over the V_{max} evaluations) and select the minimum accepted number of clusters N'_k, for which the indicators' improvement on the Pareto frontier from N'_k to $N'_k + 1$ is less than ξ (e.g. $\xi = 20\%$) (Eq.28). This implies that by increasing the number of clusters from N'_k to $N'_k + 1$ the improvement of the quality of the typical periods is not significant. N'_k is the minimum accepted number of clusters and not necessarily the best one.

$$\min N'_k \tag{28}$$

  Where:

$$\begin{aligned}
&|\frac{\sigma^a_{profile,N'_k} - \sigma^a_{profile,N'_k+1}}{\sigma^a_{profile,N'_k}}| && \leqslant 0.2 \quad \forall a \\
&|\frac{\sigma^a_{cdc,N'_k} - \sigma^a_{cdc,N'_k+1}}{\sigma^a_{cdc,N'_k}}| && \leqslant 0.2 \quad \forall a \\
&|\frac{ELDC^a_{N'_k} - ELDC^a_{N'_k+1}}{ELDC^a_{N'_k}}| && \leqslant 0.2 \quad \forall a \\
&|\frac{\Delta^a_{LDC,N'_k} - \Delta^a_{LDC,N'_k+1}}{\Delta^a_{LDC,N'_k}}| && \leqslant 0.2 \quad \forall a \\
&|\frac{\Delta^a_{prod,\gamma,N'_k} - \Delta^a_{prod,\gamma,N'_k+1}}{\Delta^a_{prod,\gamma,N'_k}}| && \leqslant 0.2 \quad \forall a
\end{aligned}$$

- Step 6: Select the best typical periods taking into account N'_k selected in step 5 (μ^*_k for $k \in \{1, ..., N^*_k\}$ in which $N^*_k \geqslant N'_k$) and the EES, intra and inter clusters distances (Section 2.1.1) as illustrated in Eq.5.

- Step 7: Once μ^*_k typical periods have been selected, add extreme typical period corresponding to the period of the year where the attribute "a" was highest ($\Delta^a_{prod,\gamma,N^*_k} = 0$). This extreme value ensures that the system can be properly sized. The existing typical periods index is modified so as to not take the extreme period into consideration twice.

- Step 8: Break up each representative period into a smaller number of segments, allowing for a further minimization of the data to be handled (Section 2.3).

Figure 1 illustrates the developed algorithm for selecting the typical periods.

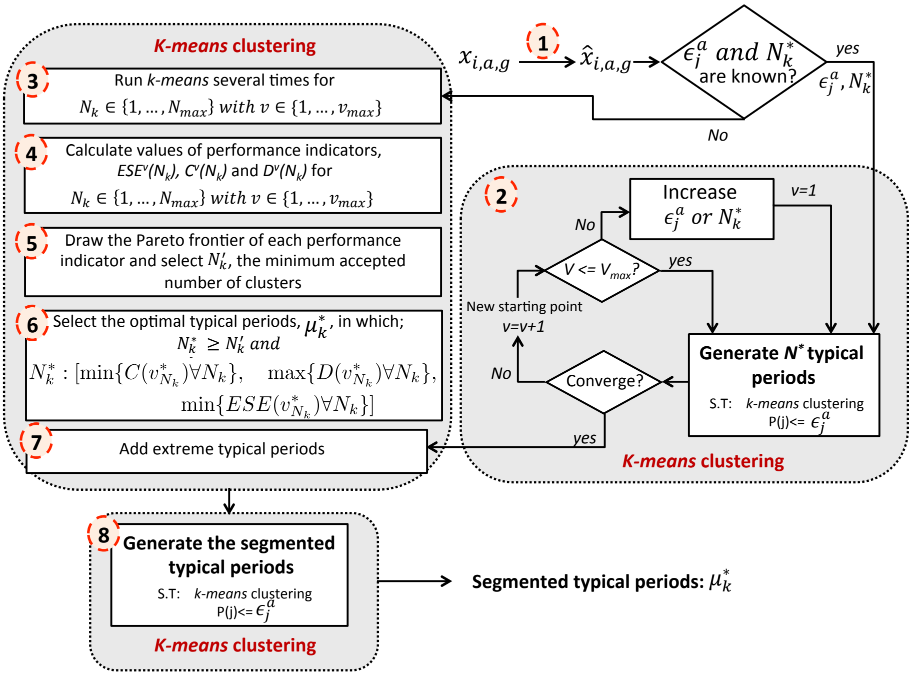
K-means clustering
3
Run k-means several times for
$N_k \in \{1,\dots,N_{max}\}$ with $v \in \{1,\dots,v_{max}\}$
4
Calculate values of performance indicators,
$ESE^v(N_k)$, $C^v(N_k)$ and $D^v(N_k)$ for
$N_k \in \{1,\dots,N_{max}\}$ with $v \in \{1,\dots,v_{max}\}$
5
Draw the Pareto frontier of each performance indicator and select N_k', the minimum accepted number of clusters
6
Select the optimal typical periods, μ_k^*, in which;
$N_k^* \geq N_k'$ and
$N_k^* : [\min\{C(v^*_{N_k})\forall N_k\}, \quad \max\{D(v^*_{N_k})\forall N_k\}, \min\{ESE(v^*_{N_k})\forall N_k\}]$
7
Add extreme typical periods
$x_{i,a,g}$
1
$\hat{x}_{i,a,g}$
ϵ_j^a and N_k^* are known?
yes
ϵ_j^a, N_k^*
No
2
Increase ϵ_j^a or N_k^*
v=1
No
$V <= V_{max}$?
yes
New starting point
v=v+1
No
Converge?
yes
Generate N^* typical periods
S.T: k-means clustering
P(j)<= ϵ_j^a
K-means clustering
8
Generate the segmented typical periods
S.T: k-means clustering
P(j)<= ϵ_j^a
K-means clustering
Segmented typical periods: μ_k^*

Figure 1: Illustration of the typical periods selection strategy

2.3. Segmented typical periods

Similarly to the typical periods algorithm, segmented periods are created with the help of the *k-means* clustering algorithm, with a performance indicator set as an additional constraint. Extreme values for each typical periods are always considered as a segment. Figure 2 presents an example of the segmented typical periods with 24 time steps.

The performance indicator $(\Delta^a_{sum,k,N_{s_k}})$ is defined as the maximum tolerated difference in total values of each type of attributes (i.e. total heat demands, total electricity consumptions, total solar irradiations) during each typical period, and is set as an additional constraint. The best number of segments, $N_{s_k^*}$, is not necessarily the same for all typical periods.

The *k-means* method is called to identify $N_{s_k^*}$ segments for each typical period k. As long as each segmented typical period does not respect the constraint, the clustering algorithm is called upon again with a new random starting point.

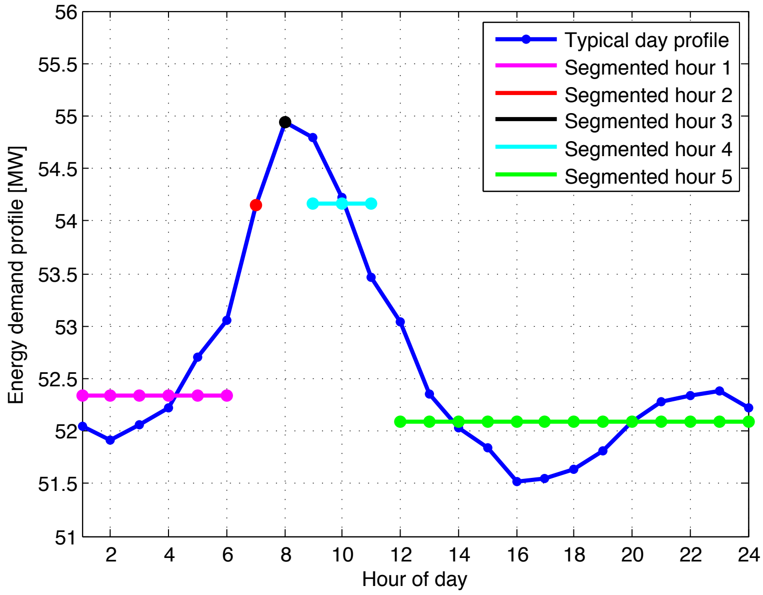

Figure 2: Segmented typical periods

Extreme values are then forced into the segments. Therefore, the end result is the segmented typical periods ($h_{k,a,N_{s_k^*}+1}$), made up of N_k^* typical periods with $N_{s_k^*}+1$ segments.

The performance indicator, which is here the maximum tolerated difference in the total demand during each period ($\Delta^a_{sum,k,N_{s_k}}$), is expressed as follows (Eq.29):

$$\Delta^a_{sum,k,N_{s_k}} = \frac{|\sum_{g=1}^{N_g} \mu^*_{k,a,g} - \sum_{s_k=1}^{N_{s_k}} d_{k,s_k} \times h_{k,a,s_k}|}{\sum_{g=1}^{N_g} \mu^*_{k,a,g}} \quad \forall a, k \in \{1, ..., N_k^*\} \quad (29)$$

d_{k,s_k} represents the duration (number of time step) of each segment, and h_{k,a,s_k} refers to the value of segment s_k in typical day k for attribute a (segmented typical periods).

If the best number of segments ($N_{s_k^*}$) and the indicator's threshold ($\Delta^a_{sum,k,N_{s_k}}$) are not known, the following steps are proposed to optimize the segmented typical periods;

- Step 1: Run *k-means* for values of $N_{s_k} \in \{1, ..., N_g\}$, with $v' \in \{1, ..., V'_{max}\}$ random starting points and calculate $\Delta^a_{sum,k,N_{s_k}}$ for each starting point.

- Step 2: Draw the Pareto frontier of $\Delta^{a}_{sum,k,N_{s_k}}$ and select the minimum accepted number of segments $N_{s'_k}$, for which the indicator' improvement on the Pareto frontier from $N_{s'_k}$ to $N_{s'_k}+1$ is less than 20% (Eq.30).

$$\min N_{s'_k}, \quad \forall k \tag{30}$$

Where:

$$|\frac{\Delta^{a}_{sum,k,N_{s'_k}} - \Delta^{a}_{sum,k,N_{s'_k}+1}}{\Delta^{a}_{sum,k,N_{s'_k}}}| \leqslant 0.2 \quad \forall a, k$$

- Step 3: Select the best segmented typical period taking into account $N_{s'_k}$ $(N_{s^*_k} \geqslant N_{s'_k})$ selected in step 2, the ESE, inter and intra clusters distances (Section 2.1.1).

- Step 4: Once segmented typical periods have been selected, extreme values are then forced into the segments, thereby adding one more segment to the segmented typical periods

- Step 5: Repeat steps 1 to 4 for $\forall k \in \{2, ..., N^*_k\}$ to calculate the segments of each typical period k.

The results of the algorithm may not always converge to the desired sequential segments. In order to reach sequential time steps, the proposed algorithm can be modified by considering the method developed by Balachandra and Chandru [8].

3. Illustrative examples

Two test cases are discussed to demonstrate the proposed method. The first case is a full-scale problem with a 23 years time horizon for supplying the heating demand of a district. The second case study aims to illustrate the proposed method by considering four type of attributes; the hourly solar irradiation, the electricity price, the heating and electricity demand profiles of a small district.

3.1. Test case 1

The multi-period MINLP optimization model is investigated in [1] in order to optimize the design and the operating strategy of district energy systems. The developed model is decomposed into a master and a slave optimizations. The master nonlinear model optimizes the system configuration and the size of conversion technologies. Meanwhile, the slave multi-period mixed integer linear model calculates the best operating schedule of selected conversion technologies.

The test case presented in [1] is used to illustrate the application of the typical periods selection method. The goal is to optimize the operating strategy of the fixed system configuration, in such a way as to supply the heating requirement of the urban area with optimal operating costs. The average annual heat demand is equal to 2100 GWh. In order to do so the slave multi-period MILP optimization model is applied [1]. The available conversion technologies are an incinerator with 160 MW_{th}, a 100 MW_{th} biomass boiler and a 130 MW_{th} coal boiler. In addition a natural gas boiler has to be sized to supply the peak loads. All units are assumed to be able to operate at any time with no limit on the availability of resources (Table 1).

Table 1: CO_2 intensity and price of available resources

Resources	$\triangle CO_2$:	Price: [25]
	$[kg/MJ]$	$[€/MJ]$:
Electricity	0.3071 [26]	0.0198
Natural Gas	0.0641	0.0105
Coal	0.0852	0.0030
Biomass	0	0.0036

The operating schedule of the system will be optimized by considering the different type of typical periods. In order to validate and demonstrate the proposed typical period selection method, the optimization results will be compared with a reference case (Section 3.1.4).

In the present work the hourly heating demand profile is estimated using meteorological data and the heating signature [27]. The heating signature is a linear model of the thermal power requirements as a function of the ambient temperature [27].

The ambient temperatures of the last 23 years from 1990 to 2012 are considered to estimate the hourly heating demand profile of the district (Figure 3) [27]. The first 20 years with 175200 (20 × 8760) time steps are used to select the typical periods, and the last 3 years data, from 2010 to 2012, are used for validation.

Form these data, a mean typical year with 8760 time steps is defined by considering the average values over 20 years (Figure 3).

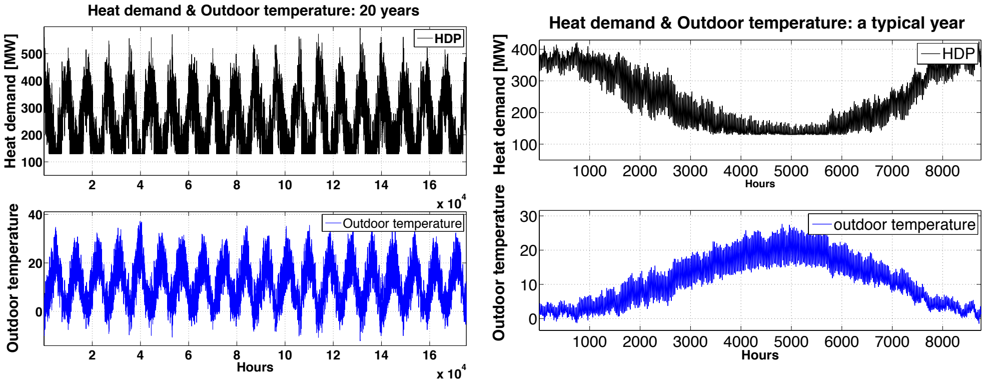

Figure 3: The ambient temperatures and estimated heat demands: 20 years and the mean typical year.

In order to reduce the optimization size, the heating demand data (series of 175200 values) is compressed to a limited number of typical periods by applying the following methods;

1. Empirical method: 13 typical periods, one per month as the average values and one extreme day [7].
2. Proposed *k-means* clustering method using the mean typical year data.
3. Proposed *k-means* clustering method using the original 20 years data equivalent to the lifetime of equipment.

3.1.1. Empirical periods

Figure 4 refers to the mean typical year with $N_i = 365$ observations, heating demand as an attribute ($N_a = 1$), 24 values for each observation ($N_g = 24$), and selected empirical periods with 312 (13×24) total time steps.

The five performance indicators proposed in Section 2.2 are used to calculate the qualities of the empirical periods for representing the original 20 years data as well as the mean typical year (Table 2). The effects of the empirical periods accuracy on the optimization results will be studied in Section 3.1.4.

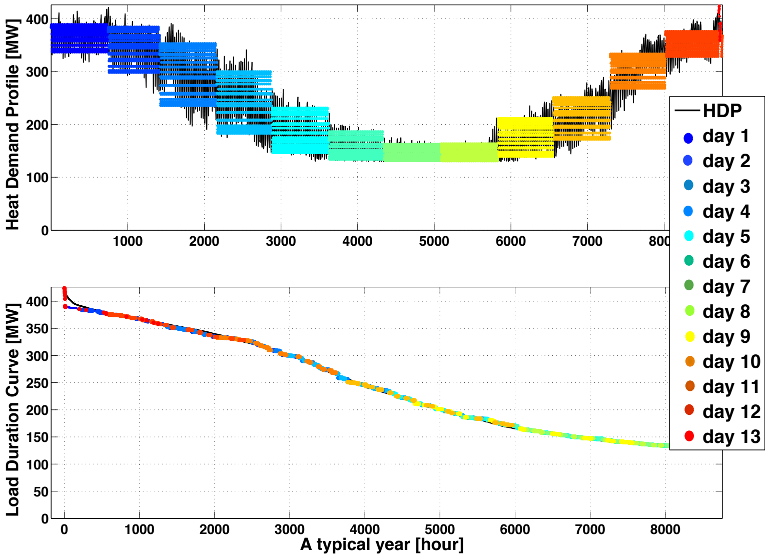

Figure 4: The empirical typical periods and the mean typical year heating demand.

Table 2: The qualities of the empirical periods compare to the original 20 years data, as well as the mean typical year data.

Quality indicators	The empirical periods and the mean typical year data	The empirical periods and the original 20 years data
σ_{cdc}	0.071	0.173
$\sigma_{profile}$	0.059	0.117
ELDC	0.056	0.153
Δ_{LDC}	0	0
$\Delta_{prod,0.07}$	113	4983

3.1.2. Proposed k-means clustering approach using the mean typical year data

In order to select the best representative typical periods from the mean typical year data, the proposed method in Section 2.2 is applied by considering $N_k \in \{1, ..., 15\}$ and $V_{max} = 1000$ random starting points.

According to Eq.5 and Eq.26, the best number of typical periods is chosen by considering the values of the intra and inter clusters distances, ESE and the Pareto frontiers of the five performance indicators (Figures 5 and 6).

Figure 5 indicates that for $N_k \geqslant 5$ the values of the performance indicators become constant and the relative differences between N_k and $N_k + 1$ are less than 0.2. The relative differences become close to zero by increasing the number of periods to $N_k = 15$. However, this leads to an increased size of the optimization model. A compromise between the optimization size and the quality of the typical periods is necessary. Therefore, the minimum accepted number of clusters is equal to 5 ($N'_k = 5$). The highest values of the average inter-clusters distance are obtained by N_k= 6 and 15 periods (Figure 6). The lowest ESE values are obtained for $N_k = 2$, $N_k = 6$ and $N_k = 14$ periods, and for more than 6 clusters the average intra-clusters distance tends towards zero (Figure 6). As a result, 6 periods plus one extreme period, are chosen as the best and qualified number of the typical periods ($N^*_k = 7$).

We go further by breaking up the 24 time steps of each representative period into smaller segments. The algorithm proposed in Section 2.3 is applied. The results indicate the optimal number of segments for the selected typical periods is equal to $N_{s^*_k} = 5$ for $\forall k \in \{2, ..., 7\}$ and $N_{s^*_k} = 4$ segments for the summer period ($k = 1$). Figure 7 illustrates how the mean typical year can be split up into its respective 7 typical periods with total 34 ($6 \times 5 + 4$) time steps. The 5-44% improvements of the quality of the 7 typical periods (see supplementary Table S1 and Table S2), with respect to the load deviation and variances, illustrates the advantages of the proposed *k-means* clustering method.

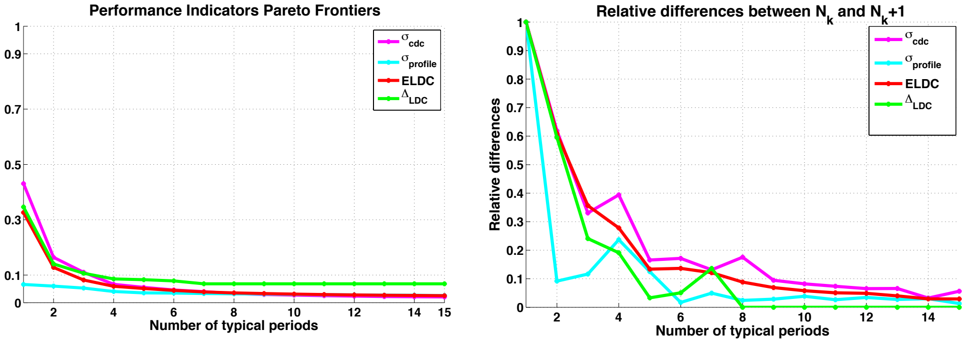

Figure 5: Pareto frontiers of typical periods' normalized performance indicators using the mean typical year data: Case study 1.

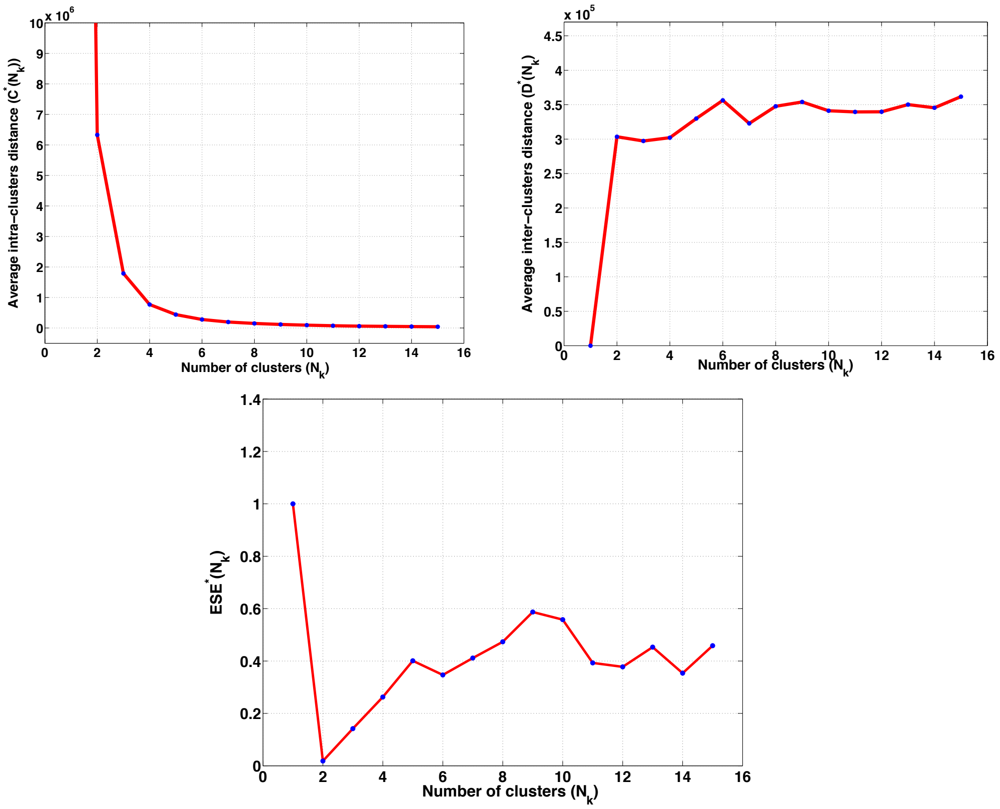

Figure 6: Intra and inter clusters distances and ESE measures as a function of the number of typical periods using the mean typical year data: Case study 1.

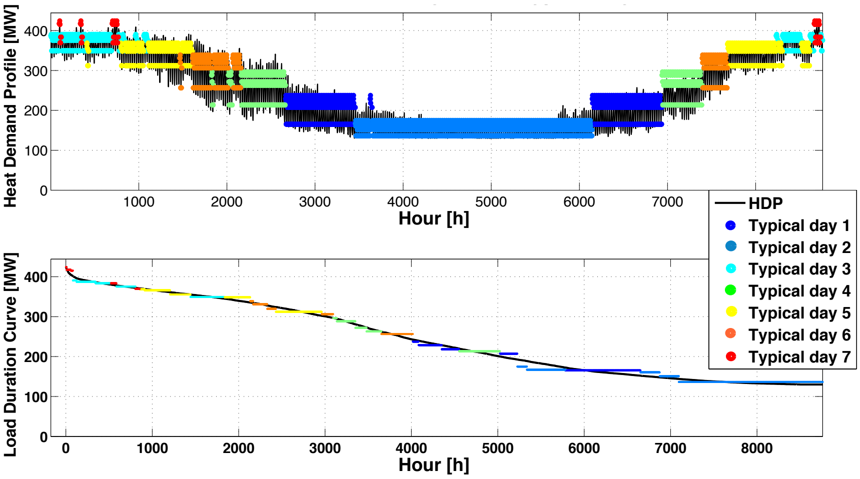

Figure 7: The mean typical year heat demand profile with 7 typical periods and corresponding 34 total time steps.

3.1.3. Proposed k-means clustering method using the 20 years data

Here, instead of the mean typical year, the heating requirements over 20 years (Figure 3) are used to select typical periods. Table 3 presents the values of the performance indicators for different number of typical periods from 5 to 13. The quality of the typical periods are improved by increasing N_k. However, the relative improvement for $N' \geqslant 5$ is less than 20%. Following the proposed method, the optimal number of typical periods, $N_k^* = 7$, was assessed using; Pareto frontiers of performance indicators, inter and intra cluster distances and ESE.

In Table 4 the column "Deviation" compares the quality of N_k^*=7 typical periods with corresponding 34 time steps (supplementary Figure S1) and the empirical periods with 312 (13 × 24) time steps (Figure 4). The five indicators present 8-63% higher qualities for $N_k^* = 7$, indicating that the proposed *k-means* clustering method provides the better approximation of the original data.

The qualities of the mean typical year for representing the original 20 years data are summarized in Table 3. Even though the classic mean typical year contains 8760 time steps, the qualities of the 7 typical periods are 35-60% higher (supplementary Table S3).

Table 3: The quality indicators of the typical periods using the original 20 years data

No. periods	N_k=5	N_k^*=7	N_k=9	N_k=11	N_k=13	The mean typical year
No.time steps	24	34	44	54	64	365 ×24
σ_{cdc}	0.096	0.063	0.062	0.058	0.050	0.157
$\sigma_{profile}$	0.115	0.108	0.100	0.098	0.096	0.102
ELDC	0.109	0.092	0.087	0.085	0.082	0.141
Δ_{LDC}	0	0	0	0	0	0.282
$\Delta_{prod,0.07}$	4191	2128	2044	1666	1200	4582

3.1.4. Validation and verification

The illustrative example is studied in order to identify the ability of the typical periods methodology for identifing an optimal operating strategy of the energy system. The typical periods selected using 1990 to 2009 data is applied on a period from 2010 to 2012 (a validation period) and compared to an accurate reference case. The total fuel consumption and the size of the peak boiler are considered as indicators to compare the results.

Table 4: The comparison between the quality of the 13 empirical periods and N_k^*=7 typical periods using the original 20 years data.

Indicators	N_k^*=7	13 empirical periods	Deviation*
No. time steps	34	312 (13 ×24)	
σ_{cdc}	0.063	0.173	63%
$\sigma_{profile}$	0.108	0.117	8%
ELDC	0.092	0.153	40%
Δ_{LDC}	0	0	-
$\Delta_{prod,0.07}$	2128	4983	57%

* The relative differences between the 13 empirical periods and 7 typical periods.

The reference case is calculated by applying a single period optimization model on each time step (3 × 8760 time steps). The goal is to optimize the operating schedule of the conversion technologies for supplying the heat demand of the 2010 to 2012 period, using a fixed size of conversion technologies. The available conversion technologies are an incinerator with 160 MW_{th}, a 100 MW_{th} biomass boiler and a 130 MW_{th} coal boiler.

The results of the reference case indicate 7515 GWh of municipal solid waste, 663 GWh of biomass, 1992 GWh of coal and 112 GWh of natural gas consumption. The peak natural gas boiler is set at 175 MW_{th} installed capacity. It is due to the systems' highest heating demand (565 MW_{th} in 2011).

The multi-period operation optimization [1] is applied using the selected typical periods. Three types of typical periods are considered:

1. 7 typical periods using the 20 years data corresponding to 34 total time steps
2. 7 segmented typical periods using the mean typical year with 34 total time steps
3. 13 empirical periods with 13 × 24 time steps

The objectives are to optimize the yearly operating strategy of the system and the size of the peak boiler. The results with regards to the yearly fuel consumptions and operating costs are multiplied by 3 to compare them to the 2010 to 2012 reference case (Table 5).

Figure 9 is a visual representation of the deviation between the typical periods and the original data in 2010. The closeness of the plots to the original data plot is an indicator of the accuracy of the method under study.

Table 5: The comparison between the reference case and the optimization results from 2010 to 2012 with regards to the size of the peak boiler, the fuel consumption and the operating costs.

	Ref.	13 Empirical	7 typical periods using	
	case	periods	the typical year	the 20 years
Municipal waste [GWh]	7415	7570(-2.1%)*	7593(-2.4%)	7489(-1.0%)
Biomass [GWh]	663	638(+3.8%)	654(+1.4%)	659(+0.6%)
Coal [GWh]	1992	2032(-2.0%)	2006(-0.7%)	1989(+0.15%)
Natural gas [GWh]	112	8.9(+92%)	6.72(+94%)	85(+24.0%)
Peak gas boiler [MW_{th}]	175	34(+80%)	34(+80%)	200(-14%)
Under estimated periods**	0	4	4	0
Operating costs [M€]	119.7	117.5(1.8%)	117(2.3%)	119.4(0.2%)
Resolution time [s]	2700	85	23	23
No. constraint	183×10^4	65320	7427	7427
No. variables	152×10^4	54423	6225	6225
No. integer variables	11×10^4	3756	432	432

*The relative differences between the reference case and the typical periods optimization.
**Number of time steps from 2010 to 2012 when the maximum original heat demands are higher than the maximum typical values.

3.1.5. Discussions

In this case study, the operation optimization of the system was studied to make precise conclusions on the quality of the typical periods.

Table 6: Performance indicators of the original 3 years data and the typical periods

	13 empirical	7 typical periods using	
	periods	the typical year data	the 20 years data
σ_{cdc}	0.103	0.083	0.056
$\sigma_{profile}$	0.094	0.090	0.086
ELDC	0.113	0.104	0.092
$\Delta_{prod,0.07}$	766	415	285

Table 6 refers to the qualities of the 13 empirical periods, 7 typical periods selected from the typical year as well as 7 typical periods selected from the 20 years data, for representing the original heating demand profiles of the 2010 to 2012 period.

Even though the 13 empirical periods contain more time steps ($13 \times 24 =$

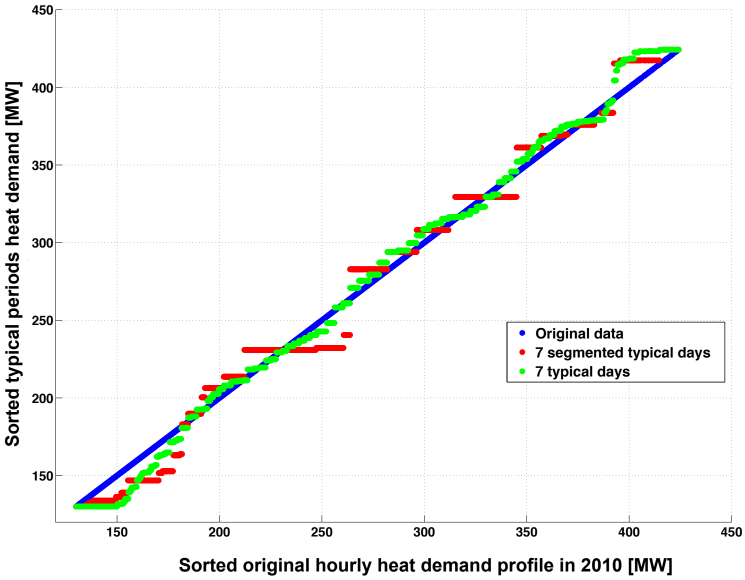

Figure 8: The sorted values of the original heat demand profile in 2010 versus the sorted heat demand profiles of 7 typical periods with 7×24 time steps and 7 typical periods with 34 $(6 \times 5 + 4)$ total time steps resulting from the mean typical year data. This figure illustrates the deviation between the original heat loads and the typical periods

312), the qualities of the results, with respect to the heat load deviation and variances, are higher with 7 typical periods. The 7 typical periods selected from the original 20 years data presented the most accurate results (Table 6).

With respect to the size of the peak boiler, it was underestimated by both the empirical periods and the 7 periods selected from the mean typical year. The obtained size was 80% less than that found by the reference case. This is explained by the extreme period of 2011, with -10 ^oC ambient temperature and 565 MW_{th} heat load demand not being represented in the mean typical year. The frequency of such a high demand is only 4 periods over 3 years (Table 5), which is not significant. In the optimization with the 7 typical periods selected using the original 20 years data, the peak boiler capacity is 14% higher compared to the reference case. This is because the heat load of the extreme periods during the first 20 years is 590 MW_{th}, which is not the case from 2010 to 2012.

With respect to the total fuel consumption and operating costs, the relative differences between the reference case and 7 typical periods selected from 20 years' heat loads present the least error, especially for biomass and natural gas consumption (Table 5).

We can sum up that 7 typical periods selected using the 20 years data give an accurate picture of the system's operations.

The optimization and reference case resolution times are summarized in Table 5. The results pointed out that the resolution time increases significantly with respect to the time steps of the demand profiles. The optimization may reach more accurate results by extending the number of time steps. With increased accuracy comes increased computational costs, with associated memory problems and prohibitive resolution time. This is especially true for solving multi-objective optimizations with a MINLP model. A compromise should always be made between the resolution time and the number of time steps.

3.2. Test case 2

The second test case is proposed to illustrate the application of the typical periods to the heating demand, electricity demand, electricity price (eex.com) and solar irradiation data of a district with 30,000 inhabitants. The aim is to optimize the operating strategy of the fixed system configuration, in such a way as to supply the energy requirement of the urban area with optimal operating costs. The data of the last 4 years from 2009 to 2012 are available. The first 3 years are used to select the typical periods and the last year, 2012, is used to validate the selected typical periods. The period is defined as a day with 24 time steps.

The case comprises 5 conversion technologies (Figure 9); a 4 MW_{el} gas turbine, a 6 MW_{el} gas engine, a 30 MW_{th} biomass boiler, a 35 MW_{th} gas boiler and 50,000 m^2 of solar thermal, using economic data from [28]. A 41 MW_{th} peak natural gas boiler is sized for the systems highest demand, present on the extreme day (120 MW heating demand). The possibility also exists to import electricity from the main grid. The solar thermal plant requires accurate meteorological data to determine the capacity of this technology for each given period, reason for which the solar irradiation and ambient temperatures profiles are also included into the study.

Following the proposed algorithm in section 2.2, for $N_k \geqslant 6$ the values of performance indicators become constant and the relative differences between N_k and N_k+1 are close to zero (see supplementary Figure S2). This indicates that by increasing the number of clusters from N_k to N_k+1 the improvement

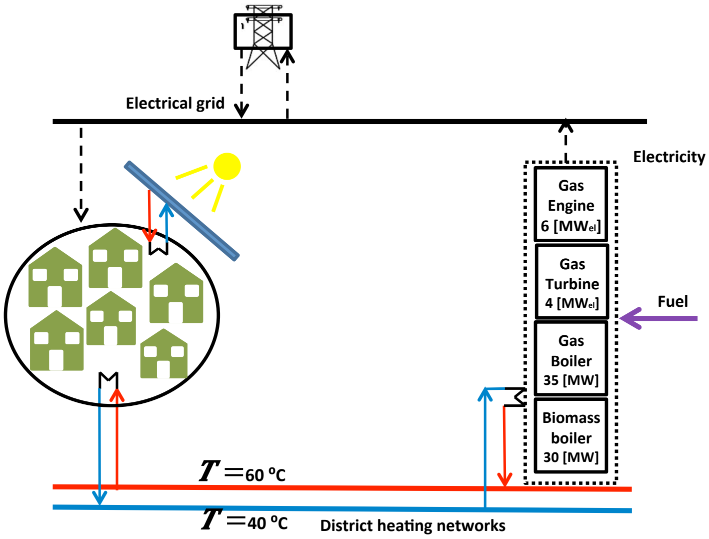

Figure 9: Test case 2 - an urban area with 30,000 inhabitants

of the typical periods quality is not significant. As a result, the minimum accepted number of clusters is equal to 6 ($N'_k = 6$). Based on the values of the three statistical measures for N_k ranging from 1 to 15 (supplementary Figure S3) and according to Eq.14, $N_k = 7$ has the lowest value for the average intra-clusters distance, the highest value for the average inter-clusters distance and the lowest value for ESE measure.

While electricity can be imported from the main grid, the central plant must supply all heating requirements, especially in the extreme period with the maximum heat demand. Therefore, the extreme period with the highest heat demand is added. To resume, 7 periods plus one extreme period are chosen as an optimal number of typical periods ($N^*_k = 8$).

We go further by breaking up the time steps of each representative period into 5 smaller segments (Section 2.3). Figure 10 presents the original data in 2012 and respective 8 typical periods with total 40 (8 × 5) time steps.

In order to make a precise conclusion on the quality of the typical periods, the reference case of 2012 were compared with the typical periods operation optimization results in terms of the operating cost, fuel consumption and the heat production of each equipment (Table 7). We investigate the 8 typical periods with 40 time steps, as well as 8 typical periods with total 192 (8×24)

time steps.

With respect to the heat share and operating costs, the 8 typical periods with 192 (8 × 24) time steps led to the most accurate results, as the relative error with the reference case shows. A period with the maximum heat demand is presented by all three type of typical periods. However, the total number of operating hours of the 41 MW_{th} peak boiler in 2012 is 36-59% under estimated by the typical periods (59% by the empirical periods, 36% by 8 typical periods with 192 time steps, and 52% by 8 typical periods with 35 time steps). Therefore, 28-53% errors in the peak boiler's heat production are pointed out in Table 7.

Apart from the peak boiler, the maximum relative differences between the optimization results of 8 typical periods with 40 time steps and the results of 8 typical periods with 8 × 24 time steps is only 2.2%. However, its resolution time is 60% less. The optimization may reach more accurate results by extending the number of time steps but this will increase the computational costs.

Table 7: Test case 2 - Comparison between the reference case and the typical periods optimization results in terms of the operating costs and the heat production

	Reference	Empirical periods	8 periods	8 periods
No. time steps	365×24	13×24	8×24	8×5
Solar thermal [GWh]	22.7	25.9 (-14.1%)*	23 (-1.3%)*	23.4 (-3.1%)
Biomass boiler [GWh]	134.8	141.0 (-4.6%)	136.5 (-1.3%)	136.5 (-1.3%)
Gas boiler [GWh]	48.3	36.7 (+24.1%)	46.6 (+3.5%)	46.3 (+4.1%)
Peak boiler [GWh]	3.2	1.5 (+53%)	2.3 (+28%)	1.7 (+46.7%)
Gas engine [GWh]	43.5	48.6 (-11.7%)	44.1 (-1.4%)	45.1 (-3.7%)
Gas turbine [GWh]	37.2	39.6 (-6.4%)	37.8 (-1.6%)	37.8 (-1.6%)
Electricity import [GWh]	57.9	52.0 (+10.2%)	57.1 (+1.4%)	56.2 (+2.9%)
Natural gas fuel [GWh]	232.1	232.8 (-0.3 %)	231.4 (+0.3 %)	232.8 (-0.3 %)
Biomass fuel [GWh]	170	178.5 (-5%)	173 (-1.8%)	173 (-1.8%)
Resolution time [s]	760	64	48	19
Operating costs [M€]	13.9	13.7 (+1.4%)	13.8 (+0.7%)	13.8 (+0.7%)

*The relative differences between the reference case and the typical periods optimization results

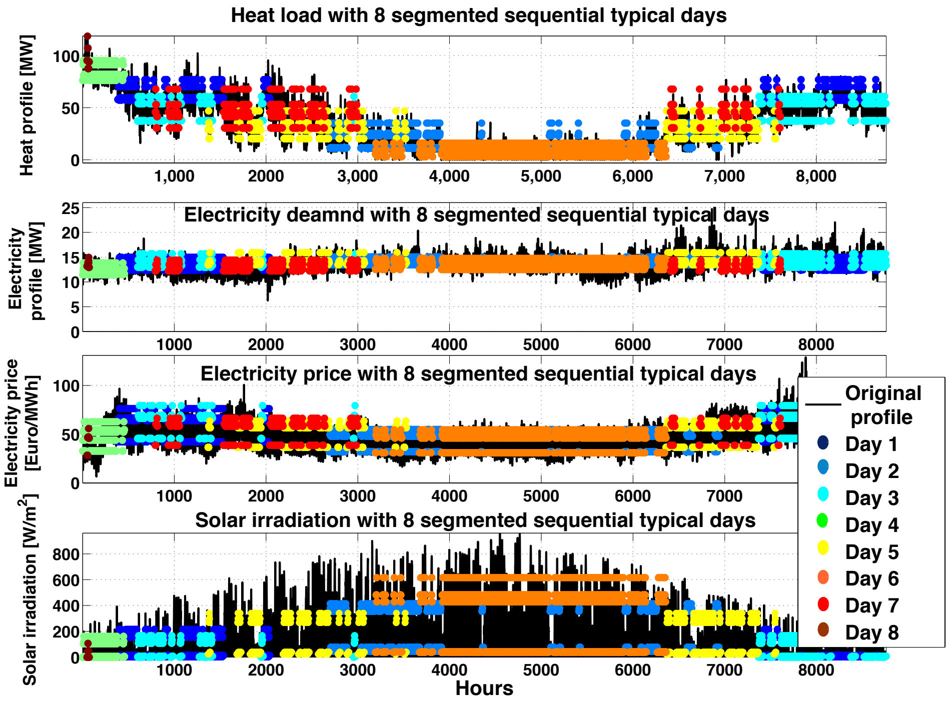

Figure 10: Test case 2- The heat and the electricity demand profiles and the solar irradiation with 8 typical periods and 40 time steps in 2012

4. Conclusions

In the present work, a new method has been developed to select the typical periods from the multiple time-varying demand profiles. The proposed method is based on the *k-means* clustering algorithm. It is developed by considering five performance indicators as additional constraints to guarantee reaching a qualified local optimal. In addition, three statistical measures are used for selecting the optimal number of typical periods.

We go further by breaking up the time steps of each representative period into smaller number of segments, further reducing the problem complexity, while respecting significant characteristics such as the peak demands and profile trends. The proposed method can easily be modified to work with typical weeks and also to accommodate other considerations such as a complex electric tariff structure or maintenance periods.

Two test cases are discussed to demonstrate the proposed method. The results of the first test case illustrate that the whole lifetime of conversion technologies can be considered for selecting the typical periods, and the proposed method can reduce a complete demand data with 20×8760 time steps into 7 segmented typical periods with total 34 time steps. The second test case illustrates the advantages of the proposed method for selecting the typical periods with respect to several attributes such as the hourly heating profile, the solar irradiation, the electricity demand and the ambient temperature.

Nomenclature

$MILP$ mixed integer linear programming

$C(v_{N_k})$ the average intra-clusters distance

$D(v_{N_k})$ the average inter-clusters distance

$x_{i,a,g}$ observation i with attribute a and measurement g

$z_{i,k}$ a binary variable equal to 1 if observation i placed in the typical period k

$\bar{x}_{i,a}$ the average value of observation i with attribute a

$\mu_{k,a,g}$ the centre of the cluster k

$\bar{\mu}_{k,a}$ the average value of the typical period k

N_k the number of the typical periods

N_k^* the optimal number of the typical periods

$N_{s_k^*}$ the best number of segment for the typical period k

$d(x_i, \mu_k)$ the distance between observation i and the center of the cluster

N_i the number of the observation

N_a the number of attributes

v the index for the random starting points

V_{max} the maximum number of random starting points

N_g the number of measurements

LDC_o^a the load duration curve of the original data of attribute a

LDC_e^a the load duration curve of typical periods of attribute a

$\sigma_{profile}^a$ the profile deviation of attribute a

σ_{cdc}^a the deviation from the load duration curve of the average values of each period and attribute a

$ELDC^a$ the error in load duration curve deviation of attribute a

Δ_{LDC}^a the maximum load duration curve deviation of attribute a

$\Delta_{prod,\gamma,N_k}^a$ Number of periods whose relative error is higher than γ

ξ threshold for performance indicators' improvement

$\Delta^{a}_{sum,k,N_{s_k}}$ the maximum tolerated difference in total values of attribute a in typical period k for N_{s_k} number of segment

h_{k,a,s_k} the value of segment s_k in typical period k corresponding to attribute a

d_{k,s_k} the duration (number of time step) in segment s_k

References

[1] S. Fazlollahi, F. Maréchal, Multi-objective, multi-period optimization of biomass conversion technologies using evolutionary algorithms and mixed integer linear programming (MILP), Applied Thermal Engineering 50 (2013) 1504 - 1513.

[2] R. P.Menon, M. Paolone, F. Maréchal, Study of optimal design of polygeneration systems in optimal control strategies, Energy 55 (2013) 134 - 141.

[3] F. Maréchal, B. Kalitventzeff, Targeting the integration of multi-period utility systems for site scale process integration, Applied Thermal Engineering 23 (2003) 1763 - 1784.

[4] J. Ortiga, J. Bruno, A. Coronas, Selection of typical days for the characterisation of energy demand in cogeneration and trigeneration optimisation models for buildings, Energy Conversion and Management 52 (2011) 1934 - 1942.

[5] M. A. Lozano, J. C. Ramos, M. Carvalho, L. M. Serra, Structure optimization of energy supply systems in tertiary sector buildings, Energy and Buildings 41 (2009) 1063 - 1075.

[6] M. Casisi, P. Pinamonti, M. Reini, Optimal lay-out and operation of combined heat & power (CHP) distributed generation systems, Energy 34 (2009) 2175 - 2183.

[7] G. Mavrotas, D. Diakoulaki, K. Florios, P. Georgiou, A mathematical programming framework for energy planning in services' sector buildings under uncertainty in load demand: The case of a hospital in Athens, Energy Policy 36 (2008) 2415 - 2429.

[8] P. Balachandra, V. Chandru, Modelling electricity demand with representative load curves, Energy 24 (1999) 219 - 230.

[9] F. Domnguez-Muoz, J. M. Cejudo-Lpez, A. Carrillo-Andrs, M. Gallardo-Salazar, Selection of typical demand days for CHP optimization, Energy and Buildings 43 (2011) 3036 - 3043.

[10] G.A.F Seber, Multivariate observations, New York: John Wiley & Sons, (1984).

[11] C.H. Marton, A. Elkamel, T.A. Duever, An order-specific clustering algorithm for the determination of representative demand curves, Computer and Chemical Engineering 32 (1999) 1373 - 1380.

[12] R.E. Steuer, Multiple criteria optimization: theory computation and application, Robert E. Krieger Publishing Malabar (Florida) (1989) .

[13] S. Fazlollahi, P. Mandel, G. Becker, F. Maréchal, Methods for multi-objective investment and operating optimization of complex energy systems, Energy 45 (2012) 12 - 22.

[14] A. K. Jain, M. N. Murty, P. J. Flynn, Data clustering: A review, ACM Computing Surveys 31 (1999) 264 – 323.

[15] G. Gan, C. Ma, J. Wu, Data clustering: Theory, algorithms, and applications, ASA-SIAM Series on Statistics and Applied Mathematics (2007).

[16] A. K. Jain, R. C. Dubes, Algorithms for clustering data, Prentice Hall College Div (1988).

[17] H. Steinhaus, Sur la division des corp materiels en parties, Bull. Acad. Polon. Sci 1 (1956) 801 – 804.

[18] G. H. Ball, D. J. Hall, Isodata, a novel method of data analysis and pattern classification, Stanford Research Institute (1965).

[19] J. B. MacQueen, Some methods for classification and analysis of multivariate observations, Fifth Berkeley symposium on mathematics, statistics and probability, University of California Press (1966) 281297.

[20] A. K. Jain, Data clustering: 50 years beyond k-means, Pattern Recognition Letters 31 (2010) 651–666.

[21] R. Kothari, D. Pitts, On finding the number of clusters, Pattern Recognition Letters 20 (1999) 405–416.

[22] S. Ray, R. H. Turi, Determination of number of clusters in k-means clustering and application in colour image segmentation, in Proceedings of the Fourth International Conference on Advances in Pattern Recognition and Digital Techniques (1999).

[23] R. Tibshirani, G. Walther, T. Hastie, On finding the number of clusters, Journal of the Royal Statistical Society (2001) 411–423.

[24] D. T. Pham, S. S. Dimov, C. D. Nguyen, Selection of k in k-means clustering, Mechanical Engineering Science 219 (2004) 103–119.

[25] IEA, Energy statistics 2011, relation with member countries poland, international energy agency (2011). Viewed 13 January.

[26] IPCC, Intergovernmental panel on climate change, IPCC Fourth Assessment Report, The Physical Science Basis, Geneva, CH, Switzerland (2007).

[27] L. Girardin, F. Maréchal, M. Dubuis, N. Calame-Darbellay, D. Favrat, Energis: A geographical information based system for the evaluation of integrated energy conversion systems in urban areas, Energy 35 (2010) 830 – 840.

[28] L. Gerber, Integration of Life Cycle Assessment in the conceptual design of renewable energy conversion systems, Ph.D. thesis, Ecole Polytechnique Federale de Lausanne, Switzerland, 2013.

Acknowledgments: The authors would like to acknowledge Veolia Environnement Recherche et Innovation (VERI) for the financial support.