Croatian Operational Research Review
CRORR **9**(2018), 255–268

Searching for an Optimal Partition of Incomplete Data with Application in Modeling Energy Efficiency of Public Buildings

Rudolf Scitovski[1], Marijana Zekić Sušac[2,*] and Adela Has[2]

[1] *Department of Mathematics, University of Osijek, Trg Ljudevita Gaja 6, 31000, Osijek, Croatia*
E-mail: ⟨scitowsk@mathos.hr⟩

[2] *Faculty of Economics, University of Osijek, Trg Ljudevita Gaja 7, 31000, Osijek, Croatia*
E-mail: ⟨{marijana, adela.has}@efos.hr⟩

Abstract. In this paper, we consider the problem of searching for an optimal partition with the most appropriate number of clusters for an incomplete data set $\mathcal{A} \subset \mathbb{R}^n$ in which several outliers might occur. Special attention is given to the application of the Least Squares distance-like function. The procedure of preparing the incomplete data set and the outlier elimination procedure are proposed such that the clustering process gives acceptable solutions. Appropriate justifications with proof are provided for these procedures. An incremental algorithm for searching for optimal partitions with $2, 3, \ldots$ clusters is applied on the prepared data set. After that, by using the Davies-Bouldin and the Calinski-Harabasz index the most appropriate number of clusters is determined. The whole procedure is organized as an algorithm given in the paper. In order to illustrate its applicability, the above steps are applied on the real data set of public buildings and their energy efficiency data, providing clear clusters that could be used for further modeling procedures.

Key words: clustering, incomplete data, missing data, optimal partition, energy efficiency of public buildings

Received: October 2, 2018; accepted: October 22, 2018; available online: December 13, 2018

DOI: 10.17535/crorr.2018.0020

1. Introduction

Let $\mathcal{A} = \{a^i = (a^i_1, \ldots, a^i_n) \in \mathbb{R}^n : i = 1, \ldots, m\} \subset \mathbb{R}^n$ be the set of m data points with n features $f_1, \ldots, f_n$. For the i-th datum a^i the value of the j-th feature is a^i_j. If for a datum a^i the values of some features are missing, then the datum a^i is said to be incomplete, and the whole set $\mathcal{A}$ is called an incomplete data set (see e.g. [15, 20]).

The problem of incomplete data is very frequent in real data sets, and standard methods of replacing missing data often do not provide a necessary level of precision and can therefore significantly bias the results in modeling procedures. This paper deals with the problem of incomplete data in the context of modelling energy efficiency of public buildings. It is part of a wider project of intelligent data analysis of energy efficiency in public buildings. In order to conduct such analysis, a real data set of public buildings in Croatia is collected with input space describing construction and energy-related characteristics of buildings. Due to a large data set, we aimed to perform a cluster analysis in the first phase of research in order to segment buildings according to their similarity, and to create energy efficiency prediction models for each of the obtained clusters in the second phase of research. The problem of incomplete data occurred in the preprocessing phase, and this paper suggests methods appropriate to replace missing values, normalize data and deal with outliers, in order to obtain most representative clusters.

*Corresponding author.

[8] and [16] were among the first important papers that dealt with clustering problems for incomplete data. General statistical methods dealing with incomplete data are based on the expectation-maximization algorithm (see e.g. [7, 34]). In [15] and [42], the authors focus on the problem of clustering incomplete data into the algorithm of fuzzy clustering. Various approaches to this problem can be found in [4]. [2] is a doctoral thesis which, inter alia, considers all aspects of incomplete data treatment. Methods for clustering data with missing values are also considered in the Master's thesis by [38]. In the paper [20], missing values are estimated in the form of intervals using the nearest neighbor method. Several papers about image analysis in case of data missing in an image or a partial missing edge image can be found in the proceedings [1]. In [21] and [36], the problem of finding guidance information from the crop row structure in case of missing plants and weeds is considered.

Incomplete data will be treated such that their internal structure is altered as little as possible. Particularly, we will keep in mind that the result of the clustering process differs as little as possible from the results we would get with complete data (see Section 2).

In our paper, in addition to the aforementioned problem with incomplete data, special attention is also paid to the problem of outliers occurring among the data (see e.g. [23, 26, 37]). Before dealing with the clustering process, we will eliminate outliers from the data set and, if necessary, normalize the data.

For such data set $\mathcal{A}$ and for some distance-like function $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, the set $\mathcal{A}$ should be grouped into $1 \leq k \leq m$ unempty and mutually disjoint clusters (see e.g. [11, 17, 22, 26, 29, 32, 34, 40]).

Example 1. *The methodology is illustrated on the set of* 3 766 *public buildings with two attributes which describe heated surface of the building (in* m^2*) and the heated volume area of the building (in* m^3*) that were selected as the most important features according to sensitivity analysis. The results show that the proposed preprocessing methods and optimal partitioning enable a clear distinction of three characteristic clusters of buildings which could be used for creating separate prediction models for each cluster. The methods are to be tested on more attributes and the modeling procedure is to enable a decision base for the process of planning and implementing measures for improving energy-related characteristics of buildings and for decreasing energy consumption.*

The main contribution of this paper is to suggest an algorithm with appropriate steps of data preprocessing and clustering. The steps include procedures to replace missing values, normalize data, and deal with outliers in order to obtain most representative clusters of public buildings. Differently from previous papers, incomplete data will be treated in a way that their internal structure is altered as little as possible so that the results of clustering procedures differ as little as possible from the results that would be obtained with complete data.

The paper is organized as follows: In the next section, we propose the procedure for preparing the data and give justification for such procedure. Furthermore, we also propose the outlier elimination method as well as the data normalization procedure. In Section 3, for an incomplete data set we propose an efficient algorithm for searching for an optimal partition with the most appropriate number of clusters. In Section 4, the suggested methodology is applied on a real data set in the area of energy efficiency of public buildings, with possible implications on further research. Finally, the conclusions and future work are discussed in Section 5.

2. Preparing the data

In this section, we discuss how data should be prepared such that results of the clustering process are as similar as results that would be obtained by means of complete data. This particularly means that for incomplete data it is necessary to know how to determine the centers of clusters as accurately as possible and implement the minimal distance principle as

correctly as possible. In so doing, we consider different approaches from the literature mentioned earlier in Introduction. We will especially analyze the application of the Least Squares (LS) distance-like function $d_{LS}(a,b) = \|a-b\|_2^2$ and the ℓ_1-metric function $d_1(a,b) = \|a-b\|_1$, and give the mathematical proof for the approach used in the real data set of public buildings and their energy efficiency data

2.1. Determining the center of the incomplete data set

Generally, the center of the set $\mathcal{A}$ regarding the distance-like function d is defined as (see e.g. [17])

$$c = \operatorname*{argmin}_{x\in\mathbb{R}^n} \sum_{a\in\mathcal{A}} d(x,a). \tag{1}$$

Specially, if d is an LS-distance-like function, the center of the set $\mathcal{A}$ is called the centroid and can be obtained as

$$c_{LS} = \tfrac{1}{m}\sum_{a\in\mathcal{A}} a, \tag{2}$$

and if d is an ℓ_1-metric function, the center of the set $\mathcal{A}$ can be obtained as

$$c_1 = \operatorname*{med}_{a\in\mathcal{A}} a. \tag{3}$$

2.1.1. Determining the center of the incomplete data set with one feature

If $\mathcal{A} = \{a^i \colon i = 1,\dots,m\} \subset \mathbb{R}$ is the set where its element a^{i_0} is missing, then the center $\tilde{c}$ of the incomplete data set $\mathcal{A}$ by using the ℓ_1-metric function can be approximated as a median $\operatorname*{med}_{a^i\in\mathcal{A}\setminus\{a^{i_0}\}} a^i$ of available data (see e.g. [17, 26, 32]), whereby the error of approximation E is given by

$$E = \tilde{c} - c = \Big| \operatorname*{med}_{a^i\in\mathcal{A}\setminus\{a^{i_0}\}} a^i - \operatorname*{med}_{a^i\in\mathcal{A}} a^i \Big|.$$

The following lemma solves this problem for the LS-distance-like function. The lemma shows that the centroid of the set $\mathcal{A}\setminus\{a^{i_0}\}$ can be approximated by the arithmetic mean of the data set $\mathcal{A}\setminus\{a^{i_0}\}$ or by the arithmetic mean of the set

$$\tilde{\mathcal{A}} = \big(\mathcal{A}\setminus\{a^{i_0}\}\big) \cup \{\tfrac{1}{m-1} \sum_{a^i\in\mathcal{A}\setminus\{a^{i_0}\}} a^i\}.$$

Note that the set $\tilde{\mathcal{A}}$ emerged from the set $\mathcal{A}$ such that the element a^{i_0} was replaced by the element $\frac{1}{m-1}\sum\limits_{a^i\in\mathcal{A}\setminus\{a^{i_0}\}} a^i$. It is shown that the centroid will be the same in both cases.

Lemma 1. *Let $\mathcal{A} = \{a^i \colon i = 1,\dots,m\} \subset \mathbb{R}$ be the set with the centroid $c = \sum\limits_{i=1}^{m} a^i$.*

If the element $a^{i_0} \in \mathcal{A}$ is missing, then the approximation of the centroid c can be determined as

$$\tilde{c} = \tfrac{1}{m-1} \sum_{b\in\mathcal{A}\setminus\{a^{i_0}\}} b \qquad or \qquad \hat{c} = \tfrac{1}{m}\sum_{b\in\tilde{\mathcal{A}}} b. \tag{4}$$

There holds $\tilde{c} = \hat{c}$, and the error of approximation E will be

$$E = \tfrac{1}{m}(\tilde{c} - a^{i_0}). \tag{5}$$

Proof. Without loss of generality, we can assume that the last m-th element a^m of the set $\mathcal{A}$ is missing. Then according to (2), there holds

$$\begin{aligned}\hat{c} =& \tfrac{1}{m}\big(a^1+\cdots+a^{m-1}+\tfrac{1}{m}(a^1+\cdots+a^{m-1})\big)\\ =& \big(\tfrac{1}{m}+\tfrac{1}{m(m-1)}\big)a^1+\cdots+\big(\tfrac{1}{m}+\tfrac{1}{m(m-1)}\big)a^{m-1}\\ =& \tfrac{m}{m(m-1)}(a^1+\cdots+a^{m-1})=\tilde{a},\end{aligned}$$

what corresponds to (4).

For the error E we obtain

$$\begin{aligned}E=\tilde{c}-c =& \tfrac{1}{m-1}\sum_{i=1}^{m-1}a^i-\tfrac{1}{m}\sum_{i=1}^{m}a^i\\ =& \big(\tfrac{1}{m-1}-\tfrac{1}{m}\big)a^1+\cdots+\big(\tfrac{1}{m-1}-\tfrac{1}{m}\big)a^{m-1}-\tfrac{1}{m}a^m\\ =& \tfrac{1}{(m-1)m}(a^1+\cdots+a^{m-1})-\tfrac{1}{m}a^m=\tfrac{1}{m}(\tilde{c}-a^m),\end{aligned}$$

what corresponds to (5). □

Remark 1. *It is easy to show the generalization of the claim of Lemma 1 in case the set $\mathcal{A}$ has several missing data.*

2.1.2. Determining the center of the incomplete data set with two or several features

If $\mathcal{A}=\{a^i=(a^i_1,\ldots,a^i_n)\in\mathbb{R}^n\colon i=1,\ldots,m\}\subset\mathbb{R}^n$ is an incomplete data set (some elements have missing values of some features), the center $\tilde{c}$ of set $\mathcal{A}$ by using an ℓ_1-metric function can be obtained by calculating each component as the median of available data (see e.g. [17, 26, 32]).

In case of applying the LS-distance-like function, the result from Subsection 2.1.1 can be generalized on the incomplete data set $\mathcal{A}$ with two or several features in the following way. Let us assume that $\mathcal{A}$ is an incomplete data set, i.e. among the elements of the set $\mathcal{A}$ there are such for which values of some features are unknown. Then the centroid of such set can be approximated

- such that for each feature we determine the arithmetic mean of known data, or
- such that for each feature instead of missing values in these places we put the arithmetic mean of known values for this feature and after that we determine the arithmetic mean of data reconstructed in this way.

According to Lemma 1, the centroid of the set $\mathcal{A}$ obtained in the first or the second way is equal.

Example 2. *Let $\mathcal{A}=\{(x_i,y_i)\colon i=1,\ldots,10\}\subset[2,5]^2$ be the set obtained by Wolfram Mathematica [39]:*

```
In[1]:= m = 10; SeedRandom[3]
        a = Round[RandomReal[{2, 5}, {m, 2}], -.02]
```

and shown in Fig. 1a and Table 1. Let us suppose that for the data of the set $\mathcal{A}$ (denoted orange in Fig. 1b) the first or the second coordinate is missing. If the first coordinate is missing, we put the arithmetic mean σ_1 of the remaining values for the first coordinates instead. Similarly, if the second coordinate is missing, we put the arithmetic mean σ_2 of the remaining values for the second coordinates instead. In such a way, the new corrected set $\tilde{\mathcal{A}}=\{(x'_i,y'_i)\colon i=1,\ldots,10\}$ (see Table 1) is defined.

i	1	2	3	4	5	6	7	8	9	10
x_i	3.44	3.04	2.54	3.74	3.22	4.92	2.3	4.26	3.04	3.94
x'_i	3.44	3.04	2.54	3.74	σ_1	4.92	2.3	σ_1	3.04	3.94
y_i	2.02	2.42	3.58	4.28	4.72	3.86	4.94	2.4	2.54	3.1
y'_i	2.02	σ_2	3.58	4.28	4.72	3.86	σ_2	2.4	2.54	3.1

Table 1: *The set $\mathcal{A}$ and the corrected set $\tilde{\mathcal{A}}$.*

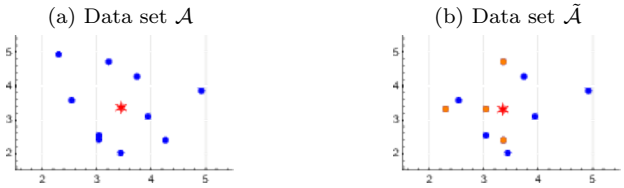

Figure 1: *The data set $\mathcal{A}$ and the corrected data set $\tilde{\mathcal{A}}$.*

As can be seen in Table 1 the datum for the first feature is missing in the data a^5 and a^8, and therefore in these places we put $\sigma_1 = 3.36$ (the arithmetic mean of the remaining values for the first feature). The datum for the second feature is missing in the data a^2 and a^7, and therefore in these places we put $\sigma_2 = 3.32$ (the arithmetic mean of the remaining values for the second feature). The centroid of the set $\mathcal{A}$ is $c = (3.44, 3.39)$, the centroid of the corrected set $\tilde{\mathcal{A}}$ is $\tilde{c} = (3.37, 3.31)$, and the error is $E = \|\tilde{c} - c\|_2 = 0.10469$.

2.2. The minimal distance principle

The second important issue we must pay attention to in the clustering process with incomplete data is the minimal distance principle. Let $a \in \mathcal{A} \subset \mathbb{R}^n$ be an arbitrary element of the set $\mathcal{A}$, let $c_1, c_2 \in \mathbb{R}^n$ be fixed centers and let $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ be a distance-like function. According to the minimal distance principle, the element $a \in \mathcal{A}$ is said to be closer to the center c_1 if there holds

$$d(c_1, a) \leq d(c_2, a). \tag{6}$$

The point of this principle is to group together similar objects by the criterion defined by the distance-like function d. If $a \in \mathcal{A}$ is an incomplete datum for which not all values of features are known, the question arises as to how to define a criterion by which the element a should be classified into the group with the center c_1 or the group with the center c_2. The application of the ℓ_1-metric function can be seen in [17, 26, 32].

In order to analyze this problem using the LS-distance-like function let us first consider the following lemma.

Lemma 2. *Let $c = (c_1, \ldots, c_n)$, $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$, whereby the component a_{i_0}, $i_0 \in \{1, \ldots, n\}$, is considered to be unknown. Furthermore, let $\hat{a} \in \mathbb{R}^{n-1}$ be the vector generated from the vector $a \in \mathbb{R}^n$ by dropping the components at i_0, and let $\tilde{a} \in \mathbb{R}^n$ be the vector generated from the vector $a \in \mathbb{R}^n$ such that its component is replaced with $\frac{1}{n-1} \sum\limits_{i \neq i_0} a_i$ at i_0. Then*

$$\tilde{E} := d_{LS}(c, \tilde{a}) - d_{LS}(c, a) < d_{LS}(c, \hat{a}) - d_{LS}(c, a) := \hat{E}. \tag{7}$$

Proof. Without loss of generality, we can suppose that the last n-th component a_n of the

vector a is missing. Then

$$\tilde{E} = d_{LS}(c,\tilde{a}) - d_{LS}(c,a) = \left(c_n - \tfrac{1}{n-1}\sum_{i=1}^{n-1} a_i\right)^2 - (c_n - a_n)^2,$$
$$\hat{E} = d_{LS}(c,\hat{a}) - d_{LS}(c,a) = c_n^2 - (c_n - a_n)^2,$$

from where immediately follows (7). □

Lemma 2 points to the conclusion that by the minimal distance principle for incomplete data using the LS-distance-like function at places with unknown components we should use the arithmetic mean of the remaining components. A reconstructed data set $\tilde{\mathcal{A}}$ with a new matrix $\tilde{A}$ is obtained in this way.

2.3. Elimination of outliers

Let us suppose that among the data set $\tilde{\mathcal{A}}$ several outliers can be expected that should be eliminated. For this purpose we will modify the idea of defining the parameter ϵ in the well-known `DBSCAN` algorithm (see e.g. [37]). First, according to [5, 9], for each $a \in \tilde{\mathcal{A}}$ we define the radius $\rho(a) > 0$ of the smallest circle containing $MinPts = 4$ elements of the set $\tilde{\mathcal{A}}$. The set of radii obtained in such a way will be grouped into two clusters by applying the LS-distance-like function and by using the efficient `SymDIRECT` method (see [13]). In such a way, we obtain two separate clusters, i.e. one with relatively small radii and one with relatively large radii. By analogy to trimmed k-means [6, 10], around the center of the cluster with relatively small radii (which is usually much more numerous) we will determine the smallest interval containing 95% elements of this cluster. Outside of that interval there are radii of circles of those elements of the set $\tilde{\mathcal{A}}$ identified as outliers. Note that here there are also naturally occurring elements of the set $\tilde{\mathcal{A}}$ corresponding to the cluster of relatively large radii. A subset of the set $\tilde{\mathcal{A}}$ from which outliers were dropped and the corresponding matrix will be denoted by $\mathcal{B}$ and B, respectively.

Example 3. *If the set from Example 1 is arranged in accordance with the above recommendations, we obtain the set $\tilde{\mathcal{A}}$ shown in Fig.2a. The arithmetic mean of the existing values of the first and of the second feature is $\sigma_1 = 2\,667.94$ and $\sigma_2 = 5\,564.48$, respectively.*

As can be seen from Fig. 2a, the set $\tilde{\mathcal{A}}$ has a significant number of outliers. For each element $a \in \tilde{\mathcal{A}}$, the radius $\rho(a) > 0$ of the smallest circle containing $MinPts = 4$ elements of the set $\tilde{\mathcal{A}}$ is shown in Fig.2b. The centroid of the cluster of relatively small radii is denoted by the red line, and the right edge of the interval containing 95% elements of this cluster is denoted by $\epsilon = 895.89$ (the orange line in Fig.2b). In this way, 3 572 elements of the set $\mathcal{B}$ were identified (Fig.2c). The remaining 194 elements of the set $\tilde{\mathcal{A}}$ are considered to be outliers.

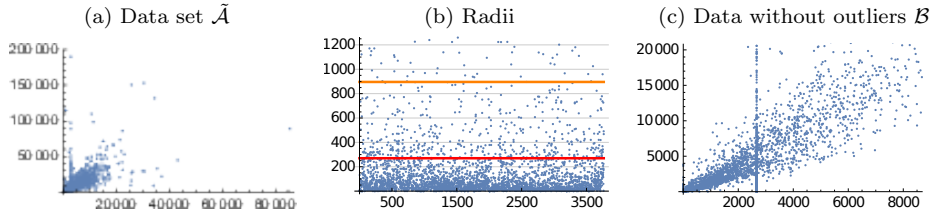

Figure 2: *Arranging a data set from Example 1.*

3. Searching for an optimal partition of an incomplete data set

Let us suppose that $\mathcal{B} = \{b^i = (b^i_1, \ldots, b^i_n) \in \mathbb{R}^n \colon i = 1, \ldots, m_B\}$ is the set in space $\mathbb{R}^n$ obtained from the set $\tilde{\mathcal{A}}$ by dropping the outliers. The set $\mathcal{B}$ can be interpreted by the matrix B with m_B rows and n columns. The rows represent the elements of the set $\mathcal{B}$, and the columns represent their features. Usually, $m_B \gg n$.

3.1. Normalization

The data set $\mathcal{B}$ is contained in a hyperrectangle $\mathcal{B} \subset [\alpha, \beta] \subset \mathbb{R}^n$, where lower and upper bounds of features are determined by vectors $\alpha = (\alpha_1, \ldots, \alpha_n)$, $\beta = (\beta_1, \ldots, \beta_n) \in \mathbb{R}^n$. If these bounds differ significantly by their features, it will be necessary to normalize the data set $\mathcal{B}$, i.e. the matrix B. This can be achieved (see e.g. [13, 22]) by the mapping $T \colon [\alpha, \beta] \to [0, 1]^n$, which transforms the set $\mathcal{B}$ into the set $\mathfrak{B} = \{T(b^i) \colon b^i \in \mathcal{B}\} \subset [0, 1]^n$, where

$$T(x) = D(x - \alpha), \qquad D = \operatorname{diag}\left(\tfrac{1}{\beta_1 - \alpha_1}, \ldots, \tfrac{1}{\beta_n - \alpha_n}\right). \tag{8}$$

After we group the data set $\mathfrak{B}$, the obtained results will be transformed back to the hyperrectangle $[\alpha, \beta]$ by applying the inverse mapping $T^{-1} \colon [0, 1]^n \to [\alpha, \beta]$, $T^{-1}(x) = D^{-1}x + \alpha$.

3.2. Searching for an optimal partition

Searching for an optimal partition of the set $\mathfrak{B}$ with the most appropriate number of clusters can be conducted by applying the `Incremental Algorithm` (see e.g. [3, 29]) by using the LS-distance-like function with correction by using the classical k-means algorithm (see e.g. [25]). The partition of the set $\mathfrak{B}$ with the most appropriate number of clusters will be determined based on the Davies-Bouldin index and the Calinski-Harabasz index (see e.g. [35]).

3.2.1. The choice of initial centers and the algorithm

The incremental algorithm starts by two carefully selected centers. For that purpose we use the idea described in [12].

Algorithm 1

Input: $\mathcal{A}$ {The set of data points}

1: Detect missing values of features and in their places put the arithmetic mean of known values as suggested in Section 2 – in that way, the set $\tilde{\mathcal{A}}$ is defined;
2: According to the procedure mentioned in Subsection 2.3, detect outliers in the set $\tilde{\mathcal{A}}$; Define the set $\mathcal{B} \subset [\alpha, \beta] \subset \mathbb{R}^n$ as the set $\tilde{\mathcal{A}}$ without outliers;
3: The set $\mathcal{B}$ should be normalized according to Subsection 3.1 – in that way the set $\mathfrak{B} \subset [0, 1]^n$ is defined;
4: According to Subsection 3.2.1, determine the first two initial centers c_1, c_2 of the set $\mathfrak{B}$, apply the Incremental Algorithm and find an optimal partition with $2, 3, \ldots$ clusters;
5: Applying the Davies-Bouldin and the Calinski-Harabasz index determine an optimal partition of the set $\mathfrak{B}$ with the most appropriate number of clusters;
6: The obtained results are transformed back in the area $[\alpha, \beta] \subset \mathbb{R}^n$;
7: Identify elements of clusters of an optimal partition as elements of the set $\mathcal{A}$.

Output: $\Pi^\star = \{\pi^\star_1, \ldots, \pi^\star_k\}$ {Optimal partition}

First, in a few random attempts we choose the center $c_1 \in \mathfrak{B}$ such that in its ϵ-neighborhood $\mathcal{O}(c_1)$ there are as many neighboring points from $\mathfrak{B}$ as possible. After that, the same procedure

is repeated on the set $\mathfrak{B} \setminus \mathcal{O}(c_1)$ determining in this way the second center $c_2 \in \mathfrak{B}$. The whole algorithm is described in detail in `Algorithm` 1 given below.

4. Application of the algorithm in energy efficiency of public buildings

To test previously described methods in the domain of energy efficiency of public buildings, a real data set of public buildings in Croatia is collected, which contains 3 766 buildings that provide education, health care, administration, sports, and other public services. The aim was to create a classification model that will be able to recognize the energy efficiency level of buildings on the basis of input space. Accurate recognition of the appropriate efficiency level is important for decision makers in institutions that decide on allocating investment resources to building reconstructions. In the preprocessing stage, it was necessary to deal with missing data and outliers in the data set, as well as to segment buildings into clusters that will be used as a basis for later modeling.

As an example, we have considered a problem with two building features; namely: heated surface of the building (in m^2) and the heated volume area of the building (in m^3). Those features are suggested based on previous research results obtained by [14] as important predictors of the energy efficiency level. The two observed features are dependent, and Pearson's correlation coefficient between them was 0.7782. Descriptive statistics of the two observed features is presented in Table 2.

Statistics	Heated surface of the building (in m^2)	Heated volume area of the building (in m^3)
Minimum (α_i)	0	0
1^{st} quartile	515.8	588.2
Median	1 335.8	2 085.7
Mean	2 667.9	5 564.5
3^{rd} quartile	3 507.0	6 853.6
Maximum (β_i)	85 500.0	189 825.0
Number of missing values	493	28
Number of data points (m)	3 766	3 766

Table 2: *Descriptive statistics of the observed feature vectors.*

Algorithm 1 suggested in Subsection 3.2.1 is performed, and in the second step of the algorithm the set $\tilde{\mathcal{A}}$ is generated with replaced missing data, as well as the data set $\mathcal{B}$ with 3 572 data points without outliers (see Fig. 2c).

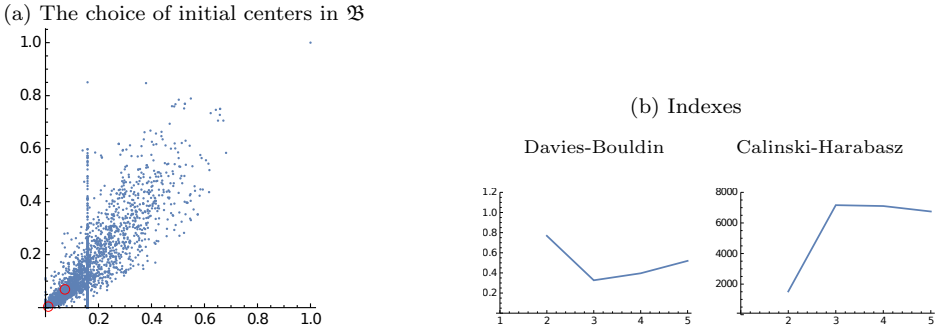

Figure 3: *The choice of initial centers in the normalized set $\mathfrak{B}$ and Davies-Bouldin and Calinski-Harabasz indexes.*

According to Subsection 3.1, by the mapping $T\colon [\alpha_1, \beta_1] \times [\alpha_2, \beta_2] \to [0,1]^2$ given by (8),

where $\alpha_1 = \alpha_2 = 0$, $\beta_1 = 16\,817.1$, $\beta_2 = 34\,415.1$ (see Table 2), the set $\mathcal{B}$ is transformed into a normalized set $\mathfrak{B}$ (Fig. 3a).

According to Subsection 3.2.1, two initial centers c_1, c_2 of the set $\mathfrak{B}$ (within red circles in Fig. 3a) are chosen. Incremental algorithm starts with these centers. Davies-Bouldin and Calinski-Harabasz indexes (see Fig. 3b) imply that an optimal partition of the set $\mathfrak{B}$ with the most appropriate number of clusters should have three clusters (see Fig. 4a). The largest number of buildings (2 271) is in the cluster $\pi_1^\star$, followed by the cluster $\pi_2^\star$ with 891 buildings, while the cluster $\pi_3^\star$ contains 410 buildings.

Due to a high level of collinearity of the features, the clusters of the normalized data set $\mathfrak{B}$ are distributed by the Voronoi diagram, which is in this case composed of two almost parallel lines (see Fig 4a)

$$\frac{x_1}{.25} + \frac{x_2}{.22} = 1, \qquad \frac{x_1}{.75} + \frac{x_2}{.5} = 1. \tag{9}$$

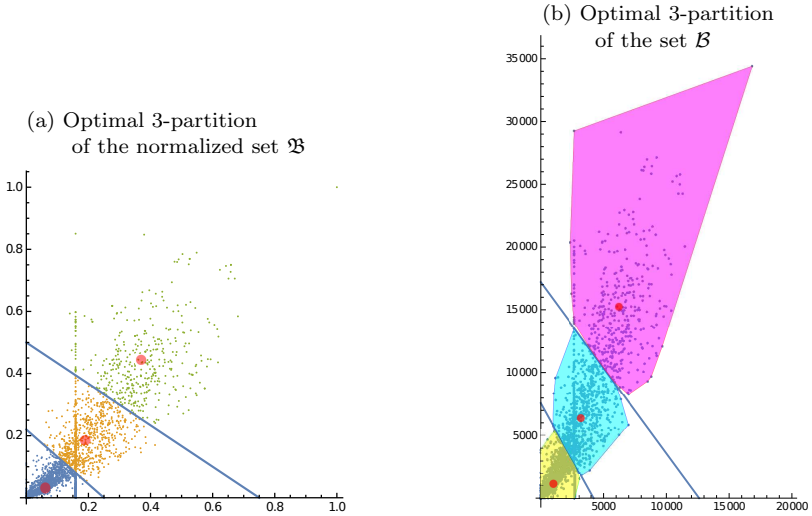

Figure 4: *Optimal 3-partition of the set $\mathcal{B}$ with two features.*

Taking into consideration that $\alpha_1 = \alpha_2 = 0$, by the inverse mapping $T^{-1}(x) = D^{-1}x$ the point $(x_1, x_2) \in [0,1]^2$ is transformed into a point in the set $[0, \beta_1] \times [0, \beta_2]$, and lines (9) are transformed into lines (see Fig 4b)

$$\frac{x_1}{.25\beta_1} + \frac{x_2}{.22\beta_2} = 1, \qquad \frac{x_1}{.75\beta_1} + \frac{x_2}{.5\beta_2} = 1, \tag{10}$$

which determine bounds of clusters of the set $\mathcal{B} \subset [0, \beta_1] \times [0, \beta_2]$ of original data values (see Fig. 4b).

Note that in our case it is easy to geometrically bound clusters. Let $b = (b_1, b_2) \in \mathcal{B}$. There holds (see Fig.4b)

$$\begin{aligned}
b \in \pi_1^\star \quad &\Leftrightarrow \quad \frac{b_1}{.25\beta_1} + \frac{b_2}{.22\beta_2} \le 1, \\
b \in \pi_2^\star \quad &\Leftrightarrow \quad \frac{b_1}{.25\beta_1} + \frac{b_2}{.22\beta_2} \ge 1 \quad \& \quad \frac{b_1}{.75\beta_1} + \frac{b_2}{.5\beta_2} \le 1, \\
b \in \pi_3^\star \quad &\Leftrightarrow \quad \frac{b_1}{.75\beta_1} + \frac{b_2}{.5\beta_2} \ge 1.
\end{aligned}$$

The center $c_1^\star = (1\,021.3, 1\,117.4)$ of the cluster $\pi_1^\star$ (heated surface of the building) is the building with heated surface of the building equal to 1 021.3 m^2, while the heated volume area of the building was 1 117.4 m^3. The center $c_2^\star = (3\,202.1, 6\,386.4)$ of the cluster $\pi_2^\star$ (heated surface

of the building) is the building with heated surface of the building equal to 3 202.1 m^2, while the heated volume area of the building was 6 386.4 m^3. The center $c_3^\star = (6\,229.1, 15\,311.6)$ of the cluster $\pi_3^\star$ (heated surface of the building) is the building with heated surface of the building equal to 6 229.1 m^2, while the heated volume area of the building was 15 311.6 m^3.

To create a profile of buildings that belong to each cluster, we have selected two points, i.e. two elements in each cluster that are close to the center of its cluster. By gaining insight into other available attributes of those buildings, we have identified specific characteristics of those buildings, as well as common characteristics of buildings within each cluster. The selected elements and their characteristics are presented in Tables 3, 4, and 5, for each cluster accordingly.

Element	Specific characteristics of each element	Common characteristics of elements within the cluster
(1027, 1131)	· belongs to the health sector; · operates 24 hours per day; · not cultural heritage; · age of building: 53 years; · renovated in 2008; · contains 1 floor.	· not cultural heritage; · share of window surface in total surface of the building ranges from 0.135 to 0.25; · construction thickness of external wall is 30 cm; · total power body heat of radiators ranges from 28 to 47 kW; · total installed thermal power of heaters ranges from 28.4 to 93.46 kW; · shape factor F_0 ranges between 0.89 and 0.9; · maximal coefficient of transmission heat loss per unit of heated area of the building ranges from 0.465 to 0.467.
(1007, 1122)	· belongs to the education sector; · operates 8 hours per day; · age of building: 8 years; · not renovated; · contains 2 floors; · 2 employees.	

Table 3: *Specific and common characteristics of selected elements in the immediate neighborhood of the centroid of the first cluster.*

Element	Specific characteristics of each element	Common characteristics of elements within the cluster
(3214, 6445)	· operates 8 hours per day; · age of building: 56 years; · renovated in 2008; · contains 2 floors; · installed additional gas-powered demand water heater (DHW); · 65 employees;	· belongs to the education sector; · share of window surface in total surface of the building ranges from 0.215 to 0.283; · construction thickness of external wall ranges from 35 to 43 cm; · installed electric heat pump; · installed electric-powered demand water heater (DHW); · total power body heat of radiators ranges from 211 to 310.92 kW; · total installed thermal power of heaters ranges from 237.8 to 795.192 kW; · total installed power of compact fluorescent lamps is larger than total installed power of incadescent lamps; · shape factor F0 ranges between 0.47 and 0.49 · maximal coefficient of transmission heat loss per unit of heated area of the building ranges from 0.601 to 0.6129.
(3037, 6336)	· age of building: 44 years; · renovated in 2005; · contains 3 floors; · electric cooling; · electric demand water heater (DHW); · 53 employees.	

Table 4: *Specific and common characteristics of selected elements in the immediate neighborhood of the centroid of the second cluster.*

It can be observed in Tables 3, 4, and 5 that selected buildings from each of the clusters share some common characteristics that differ from characteristics of buildings in other clusters. For example, the share of window surface in total surface of the building is lower in the cluster $\pi_1^\star$ than in the cluster $\pi_2^\star$, as well as total power body heat of radiators. Another construction attribute that is also important for energy consumption is construction thickness of the external wall, which is in this cluster lower than in other two clusters. Regarding the shape factor,

Element	Specific characteristics of each element	Common characteristics of elements within the cluster
(6088, 15327)	· belongs to the administration sector; · is cultural heritage; · age of building: 120 years; · contains 2 floors; · electric cooling; · electric demand water heater (DHW) and additional gas-powered DHW; · installed heat pump; · 1990 employees.	· share of window surface in total surface of the building ranges from 0 to 0.377; · construction thickness of external wall ranges from 40 to 64 cm; · total power body heat of radiators ranges from 0 to 400 kW; · total installed thermal power of heaters ranges from 400 to 216 038 kW; · shape factor F0 ranges between 0.39 and 0.41; · maximal coefficient of transmission heat loss per unit of heated area of the building ranges from 0.6593 to 0.9543.
(6295, 15078)	· belongs to the education sector; · not cultural heritage; · age of building: 33 years; · contains 2 floors; · 248 employees.	

Table 5: *Specific and common characteristics of selected elements in the immediate neighborhood of the centroid of the third cluster.*

buildings in the cluster $\pi_1^\star$ have a higher shape factor than buildings in clusters $\pi_2^\star$ and $\pi_3^\star$. The maximal coefficient of transmission heat loss per unit of the heated area is lower in the cluster $\pi_1^\star$ than in clusters $\pi_2^\star$ and $\pi_3^\star$. Both buildings in the cluster 2 belong to the educational sector. In addition to those common characteristics, each of the selected buildings has some specific characteristics, which are mainly present in the number of employees which varies across buildings, the sector they belong to, age of the building, the number of floors, etc.

For this preliminary research, we bring this comparison as an example of a possible application in energy efficiency management. Decision makers could use such building profiles and make decisions on investments in particular buildings and clusters.

The next step in the application will be to generate clusters by using more feature vectors, in order to obtain more accurate partitioning which takes into account more information about the buildings. On the basis of generated clusters, for each partition of the data set a prediction model based on neural networks will be created to classify buildings according to their efficiency level. Such procedure is expected to gain more accurate modeling results. Separate models could be incorporated in a software tool that could be used to support decision makers while allocating the funds to reconstruction of certain buildings. Also, characteristics of each cluster could reveal information on some common features that buildings in the same clusters share; for example, if the buildings that have hip roof are more energy efficient than the buildings with flat roof, or if the type of heating matters, or the isolation material, etc.

The same problem of recognizing the energy efficiency level of buildings is observed in [14] but with no replacement of missing data and outlier exclusion. By providing the appropriate methods of dealing with incomplete data and outliers in relation to clustering procedure, the algorithm suggested in this paper could improve the modeling of energy efficiency of public buildings.

5. Conclusions

The issues of incomplete data and outliers are common in real data sets, especially in modeling energy efficiency. The purpose of this paper was to provide a methodology that could efficiently deal with those issues such that their internal structure is altered as little as possible, i.e. to obtain the results that differ as little as possible from the results that would be obtained with complete data.

In this paper, we decided to apply the LS-distance-like function and an incremental algorithm for searching for an optimal partition since in that case we can naturally use Davies-Bouldin and Calinski-Harabasz indexes.

The outlier elimination problem is efficiently solved by one modification of the idea used for density-based clustering.

As working with a real data set is assumed, after preparing data, a normalized data set in the hyperrectangle $[0,1]^n$ was defined on which the clustering process was performed.

The suggested procedure is illustrated on a real data set containing construction and energy-related characteristics of public buildings that could be used to model the energy efficiency level in the next stage of research. Preliminary results on two selected feature vectors show that clear partitioning into three clusters of buildings can be obtained.

In future research we plan to create separate machine learning models for each cluster of buildings to recognize their efficiency levels. In addition to its methodological contribution, the paper can be used as a basis for providing models that could assist decision makers in allocating resources to measures for improving energy-related characteristics of buildings that could lead to savings in energy consumption and environmental protection.

Acknowledgements

This work was supported by Croatian Science Foundation through research grant IP-2016-06-8350 "Methodological framework for efficient energy management by intelligent data analytics" and research grant IP-2016-06-6545 "The optimization and statistical models and methods in recognizing properties of data sets measured with errors".

References

[1] Abi-Nahed, J. and Yang, M. P. J. G. Z. (1973). Robust active shape models: A robust, generic and simple automatic segmentation tool, in: Lecture Notes in Computer Science 4191.

[2] Äyrämö, S. (2006). Knowledge Mining Using Robust Clustering. Ph.D. thesis. University of Jyväskylä.

[3] Bagirov, A.M., Ugon, J. and Webb, D. (2011). An efficient algorithm for the incremental construction of a piecewise linear classifier. Information Systems 36, 782–790.

[4] Bezdek, J.C., Keller, J., Krisnapuram, R. and Pal, N.R. (2005). Fuzzy models and algorithms for pattern recognition and image processing. Springer, New York.

[5] Birant, D. and Kut, A. (2007). ST-DBSCAN: an algorithm for clustering spatial-temporal data. Data & Knowledge Engineering 60, 208–221.

[6] Cuesta-Albertos, J. A., Gordaliza, A. and Matrán, C., (1997). Trimmed k-means: An attempt to robustify quantizers. The Annals of Statistics 25(2), 553–576.

[7] Dempster, A., Laird, N. and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39, 1–38.

[8] Dixon, J. K. (1979). Pattern recognition with partly missing data. IEEE Transactions on Systems, Man, and Cybernetics SMC-9, 617–621.

[9] Ester, M., Krieogel, H. and Sander, J. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise, in: 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland. pp. 226–231.

[10] Fritz, H., García-Escudero, L. A. and Mayo-Iscar, A. (2013). A fast algorithm for robust constrained clustering. Computational Statistics and Data Analysis 61, 124–136.

[11] Gan, G., Wu, J. and Ma, C. (2007). Data Clustering: Theory, Algorithms, and Applications. SIAM, Philadelphia.

[12] Grbić, R., Grahovac, D. and Scitovski, R., (2016). A method for solving the multiple ellipses detection problem. Pattern Recognition 60, 824–834.

[13] Grbić, R., Nyarko, E. K. and Scitovski, R., (2013). A modification of the DIRECT method for Lipschitz global optimization for a symmetric function. Journal of Global Optimization 57, 1193–1212.

[14] Has, A. and Zekić-Sušac, M., (2017). Modelling energy efficiency of public buildings by neural networks and its economic implications, in: Zadnik-Stirn, L., Drobne, S. (Eds.), Proceedings of the 14th International Symposium on Operations Research in Slovenia.

[15] Hathaway, R. J. and Bezdek, J. C. (2001). Fuzzy c-means clustering of incomplete data. IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics 31, 735–744.

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

[17] Kogan, J. (2007). Introduction to Clustering Large and High-dimensional Data. Cambridge University Press, New York.

[18] Kogan, J. and Teboulle, M. (2006). Scaling clustering algorithms with Bregman distances, in: Berry, M.W., Castellanos, M. (Eds.), Proceedings of the Workshop on Text Mining at the Sixth SIAM International Conference on Data Mining.

[19] Leisch, F. (2006). A toolbox for k-centroids cluster analysis. Computational Statistics & Data Analysis 51, 526–544.

[20] Li, T., Zhang, L., Lu, W., Hou, H., Liu, X., Pedrycz, W. and Zhon, C. (2017). Interval kernel fuzzy c-means clustering of incomplete data. Neurocomputing 237, 316–331.

[21] Marchant, J. (1996). Tracking of row structure in three crops using image analysis. Computers and Electronics in Agriculture 15, 161–179.

[22] Morales-Esteban, A., Martínez-Álvarez, F., Scitovski, S. and Scitovski, R. (2014). A fast partitioning algorithm using adaptive Mahalanobis clustering with application to seismic zoning. Computers & Geosciences 73, 132–141.

[23] Rousseeuw, P. J. and Leroy, A. M. (2003). Robust Regression and Outlier Detection. Wiley, New York.

[24] Sabo, K. and Scitovski, R., (2008). The best least absolute deviations line – properties and two efficient methods. ANZIAM Journal 50, 185–198.

[25] Sabo, K. and Scitovski, R. (2015). An approach to cluster separability in a partition. Information Sciences 305, 208–218.

[26] Sabo, K., Scitovski, R. and Vazler, I. (2013). One-dimensional center-based l_1-clustering method. Optimization Letters 7, 5–22.

[27] Sabo, K., Scitovski, R., Vazler, I. and Zekić-Sušac, M. (2011). Mathematical models of natural gas consumption. Energy Conversion and Management 52, 1721–1727.

[28] Scitovski, R. (2017). A new global optimization method for a symmetric Lipschitz continuous function and application to searching for a globally optimal partition of a one-dimensional set. Journal of Global Optimization 68, 713–727.

[29] Scitovski, R. and Scitovski, S., (2013). A fast partitioning algorithm and its application to earthquake investigation. Computers & Geosciences 59, 124–131.

[30] Scitovski, R., Vidović, I. and Bajer, D., (2016). A new fast fuzzy partitioning algorithm. Expert Systems with Applications 51, 143–150.

[31] Scitovski, S. and Šarlija, N. (2014). Cluster analysis in retail segmentation for credit scoring. Croatian Operational Research Review 5, 235–245.

[32] Späth, H. (1983). Cluster-Formation und Analyse. R. Oldenburg Verlag, München.

[33] Tan, P.N., Steinbach, M. and Kumar, V. (2006). Introduction to Data Mining. Wesley.

[34] Theodoridis, S. and Koutroumbas, K. (2009). Pattern Recognition. Academic Press, Burlington. 4^{th} edition.

[35] Vendramin, L., Campello, R. J. G. B. and Hruschka, E. R. (2009). On the comparison of relative clustering validity criteria, in: Proceedings of the SIAM International Conference on Data Mining, SDM 2009, April 30 – May 2, 2009, Sparks, Nevada, USA, SIAM. pp. 733–744.

[36] Vidović, I. and Scitovski, R. (2014). Center-based clustering for line detection and application to crop rows detection. Computers and Electronics in Agriculture 109, 212–220.

[37] Viswanath, P. and Babu, V.S. (2009). Rough-DBSCAN: a fast hybrid density based clustering method for large data sets. Pattern Recognition Letters 30, 1477–1488.

[38] Wilson, S. E., (2015). Methods for Clustering Data with Missing Values. Ph.D. thesis. University of Leiden.

[39] Wolfram Research, I. (2016). Mathematica. Wolfram Research, Inc., Champaign, Illinois. version 11.0 edition.

[40] Zaki, M. J. and Jr., W. M. (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, New York.

[41] Zekić-Sušac, M., Šarlija, N. and Benšić, M. (2004). Small business credit scoring: A comparison of logistic regression, neural network, and decision tree models, in: Proceedings of the International Conference on Information Technology Interfaces, ITI, pp. 265–270.

[42] Zhang, L., Lu, W., Liu, X., Pedrycz, W. and Zhong, C. (2016). Fuzzy C-Means clustering of incomplete data based on probabilistic information granules of missing values. Knowledge-Based Systems 99, 51–70.