# The independent component analysis with the linear regression – predicting the energy costs of the public sector buildings in Croatia

**Marinela Mokriš**[1,2,*]

[1] *Faculty of Economics in Osijek, University of Josip Juraj Strossmayer in Osijek, Trg Ljudevita Gaja 7, Osijek, Croatia*

[2] *Faculty of Organization and Informatics, University of Zagreb, Pavlinska 2, Varaždin, Croatia*
*E-mail: ⟨marinela.mokris@efos.hr⟩*

**Abstract.** In the European Union, the public sector buildings are considered significant energy consumers and are, thus, the subject of several directives that aim to ensure the renovation of existing and the construction of new buildings as nearly zero-energy buildings. Therefore, as part of the decision making, it is necessary to properly plan the renovation or construction. This research provides models for predicting the energy costs of the public sector buildings, which are dependent upon its characteristics (i. e., constructional, occupational, energy, etc.). For this purpose, a real data set of Croatian public buildings was used, which included 150 variables and 1724 observations. Since the data set consisted of a large number of variables, the motivation for the dimensionality reduction was addressed first. Then, the independent component analysis, the principal component analysis, and the factor analysis were performed as the dimensionality reduction methods for variable extraction. The results of these analyses were used as inputs for modelling the energy costs of the public sector buildings. The obtained models were compared to the model built on original variables. The obtained models show the application potential in decision making for building renovation and construction in the public sector of Croatia, whereas the best performance of prediction in terms of RMSE and SMAPE was achieved by the model that integrated the independent component analysis with the linear regression.

**Keywords**: energy cost, factor analysis, independent component analysis, prediction, principal component analysis, public sector buildings

Received: April 14, 2022; accepted: July 19, 2022; available online: December 22, 2022

DOI: 10.17535/crorr.2022.0013

## 1. Introduction

Excessive energy consumption causes both large financial costs and environmental pollution, which contributes to climate change and global warming (greenhouse gas emissions, etc.). The European Union is investing considerable effort in energy, which is evident from its following directives. In 2012, under the Energy Efficiency Directive 2012/27/EU, the EU set an objective to increase energy efficiency and to achieve saving 20 % of the Union's primary energy consumption by 2020 compared to projections made in 2007 [13]. In the Directive 2018/2002, the EU set the goal to improve energy efficiency by 2030 by decreasing energy consumption for at least 32.5 % [15]. Furthermore, according to the Directive 2010/31/EU, 40 % of total energy consumption in the EU belongs to the ever-expanding building sector. Within that same Directive, EU set the goal that all new public buildings constructed after 2018, and residential buildings constructed after 2020, should be nearly zero-energy buildings (nZEB) [12]. Moreover, in the Directive 2018/844, the EU set the goal to ensure the renovation of existing buildings

[*] Corresponding author.

into nZEB [14]. Therefore, the motivation for the energy consumption modelling of the public sector buildings, which is dependent upon its characteristics, is evident.

The sample for this research is based on a real data set obtained from Croatian Energy Management Information System (EMIS), and consists of 1724 observations on 150 variables. Even though the data set used for this research is not high-dimensional, since [21] consider a data set as high-dimensional if the number of variables $p$ is larger or much larger than the number of observations $N$, the data set is undoubtedly large. Consequently, a need for a dimensionality reduction will be justified in the following section of the paper.

The aims of this paper are: (1) to consider difficulties caused by the large number of variables in the model, (2) to reduce the dimensionality of the given data set by using the ICA, the PCA and the FA, and (3) to create models for predicting the energy costs (as a measure of energy consumption) of the public sector buildings and to compare their performances. By reducing multicollinearity and other negative aspects of a large number of correlated variables, it is assumed that a model with input obtained by the ICA, the PCA and the FA will ultimately achieve better performance than a model built on original variables. Finally, the expected scientific contributions of this research are: (1) the explanation of the effects that can be caused by a large number of variables, (2) the explanation of the advantages and disadvantages of the ICA method regarding the PCA and the FA, and (3) their application in building models for predicting energy costs and evaluation of their prediction performance.

The rest of the paper is structured as follows: Section 2 provides the theoretical framework and previous research, Section 3 focuses on the independent component analysis as the dimensionality reduction method and gives a comparison between the ICA, the PCA and the FA. Further, the linear regression is explained and the data are presented. The results and the discussion are provided in Section 4. Finally, the implications and the conclusion of the research are given in the last section.

## 2. Theoretical framework and previous research

Nowadays, many measurements and variables are being collected. In that sense, there is a clear discrepancy in the standard statistical methodology where one has dealt with many observations for several carefully selected variables based on certain theoretical or scientific knowledge [10]. Discovering relevant and non-redundant information from the collected data simultaneously is an important task [35].

### 2.1. Theoretical framework

The "blessings of dimensionality" and the "curse of dimensionality" are two common terms mentioned in the literature that are considered to be the "two sides of the same coin" [18]. Gorban et al. in [18] and [17] concluded that if a data set is high-dimensional, then, surprisingly, certain problems tend to be solved more easily, i. e., by simple and robust old methods and, ultimately, a fundamental trade-off between complexity and simplicity in high-dimensional spaces exists. They have also discussed the main benefits (blessings) of dimensionality that are founded on the measure phenomena "which suggest that statements about very high-dimensional settings may be made where moderate dimensions would be too complicated". On the other hand, the term "curse of dimensionality" relates to many problems that become exponentially difficult in high dimensions [17]. Dimensionality reduction is widely used as a preprocessing step in data mining and discovering knowledge from the data. The methods can usually be divided into the variable selection and the variable extraction approaches. The variable selection approach refers to methods that select the subset of relevant variables, while the variable extraction methods return new variables as a function of the original variables, i. e., transform data, potentially to a space of fewer dimensions. Dimensionality reduction is also necessary in order to facilitate the data visualization and understanding, reducing training and utilization times, and reducing measurements and storage requirements [19].

Hair et al. in [20] stated that the success of any multivariate technique depends on the decision of the variables which are going to be used as dependent (explained) and independent (explanatory) variables. The multiple regression analysis was performed after reducing dimensionality, and the importance of that decision within the multiple regression context follows. Namely, [20] emphasized three issues that researchers should keep in mind when making that decision: a strong theory, the measurement error and the specification error. A strong theory implies that the selection of the dependent and independent variables should be based on the conceptual or theoretical grounds. The measurement error refers to the degree to which the variable is an accurate and consistent measure of the concept being studied. If the variable used as the dependent measure has a substantial measurement error, then even the best independent variables may be unable to achieve acceptable levels of predictive accuracy. Nevertheless, the specification error is considered as the most problematic issue in the independent variable selection as it concerns the inclusion of irrelevant or the omission of relevant variables from the set of independent variables. An inclusion of an irrelevant variable may increase multicollinearity. Multicollinearity refers to an extent to which a variable can be explained by other variables in the analysis. If it is abundantly present, due to the independent variables' interrelationships, it complicates the interpretation of the regression model and it is more difficult to ascertain the effect of any single variable. For example, it can change the value or sign of the estimated coefficients, which could lead to certain statistically significant variables becoming insignificant. Therefore, multicollinearity effects the estimation of the regression coefficients, as well as their statistical significance, and it also effects the predictive ability of the regression model. What is more, with increasing multicollinearity the total variance explained decreases, which is the reason why multicollinearity may make the independent variable's statistical significance testing less precise and may reduce the statistical and practical significance of the analysis [20].

## 2.2. Previous research

With the advancement of machine learning, many methods from those fields are used for dimensionality reduction, such as convolutional neural networks [8], support vector machines [41], decision trees [42] and others. Information criteria such as joint mutual information maximisation [4], statistical methods such as random forest and its Gini importance [39], [48], correlations, $\chi^2$ test [48], and many other methods for dimensionality reduction are also used. One of the most commonly used statistical variable extraction methods is the principal component analysis (PCA) which transforms the starting space into a lower-dimensional one by converting it into a new set of variables - principal components. Those are obtained by solving the eigenvectors and eigenvalues of the covariance/correlation matrix and are uncorrelated and selected to retain most of the variations present in the original variables [28], [29]. In the context of energy consumption, [32] used the principal components regression method to vector autoregression model for the electricity consumption prediction.

The independent component analysis (ICA) "can be seen as a refinement of principal component analysis or factor analysis" [40]. What distinguishes the ICA among classical multivariate statistical methods, is the assumption of non-Gaussianity, which could capture the identification of original, underlying components in a multivariate data [25]. Therefore, it is not unexpected that it is quite commonly used, mainly for signal processing and data analysis. For example, [27] presented various applications of this method, including variable extraction, biomedical signal processing, image processing, telecommunications and econometrics. Various authors in biomedicine used this method for extracting main characteristics from gene expression [11], from EEG [25], and for functional magnetic resonance imaging [35]. Further, the ICA was also used in semantic purposes where it performed better than other methods, like hierarchical clustering, in order to find invisible information from the data [7]. It was also used combined with the PCA and neural networks for predictions in the stock market [36]. [25] also provided recent developments in the ICA, which included various topics, such as analysis of causal relations and others. Although the ICA has many different applications, this paper will focus on examining its application in the context of energy costs.

Additionally, the literature review shows that in the field of public sector building energy, authors have dealt with modelling energy consumption, energy consumption by individual energy source, energy costs, and energy intensity. Methods such as support vector machine [16], decision trees [16], [46], *Stochastic Impacts by Regression on Population, Affluence, and Technology* (STIRPAT) [37], ridge regression [37], partition trees [49], CART [51], random forest [49], [51], and linear regression [24], [49], [50] were used. Neural network was also commonly used, as in [16], [38], [1], [48], [43], [9], [49], and [50]. The methods were performed on the whole sample or on a sample divided into clusters, as in [38] and [22]. In addition to the methods mentioned above, various optimizations (e. g., [2]), simulations (e. g., [9]), and also different tools (such as the Energy Plus Software [9], [2], and others) were used.

## 3. Methodology and data

In this section, the independent component analysis is presented first and is followed by comparisons to the principal component analysis and the factor analysis. Then, the usage of the linear regression method within this context is explained and, finally, the data set is described.

### 3.1. Independent component analysis

Independent component analysis (ICA) was originally designed to deal with the "cocktail party" problem [27], and although, today, the ICA is utilized for many various purposes, the "cocktail party" problem will be provided as the motivation for conducting the ICA.
Suppose that two people spoke at the same time on two microphones located in different locations. Two different signals, $x_1(t)$ and $x_2(t)$, were recorded, where $x_1$ and $x_2$ are amplitudes and $t$ is the time index. Each of the recorded signals is the weighted sum of the speech signals of the two people who talked, denoted by $s_1(t)$ and $s_2(t)$ respectively, or:

$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t)$$

$$x_2(t) = a_{21}s_1(t) + a_{22}s_2(t),$$

where $a_{11}, a_{12}, a_{21}$, and $a_{22}$ are unknown parameters which depend on the distance of the person from the microphone, while $s_1(t)$ and $s_2(t)$ are the original sources which are also unknown. It would be desirable to estimate the two original sources $s_1(t)$ and $s_2(t)$ from $x_1(t)$ and $x_2(t)$, which is known as the "cocktail party" problem. If the values $a_{11}, a_{12}, a_{21}$, and $a_{22}$ are known, the problem could be solved by simply inverting the linear system. Not knowing both the parameters $a_{ij}$ and the source $s_i(t)$ makes the problem much more difficult.

One way to solve this problem would be to use certain statistical properties of the signal $s_i(t)$ to estimate both $a_{ij}$ and $s_i(t)$. It is enough to assume that $s_i(t)$ are statistically independent at any time $t$. This is not an unrealistic assumption in many cases and may not be entirely accurate in practice. The independent component analysis can be used to estimate the parameters $a_{ij}$ based on the independence of $s_i(t)$, which then allows to extract the original signals $s_i(t)$ from their mixtures $x_i(t)$. In the rest of the paper, the time index $t$ will be omitted because it is assumed that every $x_i$ and $s_i$ are random variables, so then every $x_i(t)$ and $s_i(t)$ are realizations of those random variables.

The problem of the "cocktail party" is also an example of the blind source separation problem [26]. Within that problem, situations are observed in which a source emits a number of signals, for example, different areas of the brain emit electrical signals, or mobile phones emit radio waves. It is also assumed that there are several sensors or receivers which are in different positions. Each of them records a mixture of source signals with a different weight. “Source” means the source signal, while “blind” means that very little, if anything, is known about the mixing matrix and little is assumed about the source signals. ICA solves that problem.

### Definition and assumptions

To define the independent component analysis assume that $n$ linear mixtures $x_1, ..., x_n$ of $n$ independent components $s_1, ..., s_n$ are observed:

$$x_i = a_{i1}s_1 + a_{i2}s_2 + \cdots + a_{in}s_n = \sum_{j=1}^{n}(a_{ij}s_j), \forall i = 1, \ldots, n.$$

The independent components (ICs) $s_1, ..., s_n$ are latent variables, meaning that they cannot be directly observed. Mixing coefficients $a_{ij}, i, j = 1, \ldots, n$ are also assumed to be unknown. It is, therefore, necessary to estimate both latent variables $s_1, ..., s_n$ and mixing coefficients $a_{ij}, i, j = 1, \ldots, n$ using observations $x_1, ..., x_n$ [27], [26]. The IC model is a generative model, which means that it describes how the observed data are generated by the process of mixing the components $s_1, ..., s_n$.

Let us denote by $x$ the random vector whose elements are the $x_1, ..., x_n$, by $s$ the random vector with elements $s_1, ..., s_n$ and by $A$ the matrix with elements $a_{ij}, i, j = 1, \ldots, n$. Then, the IC model can be written as:

$$x = As.$$

The following theoretical assumptions and restrictions are required for the IC model to be estimated: (1) the components $s_i$ are mutually independent, (2) components $s_i$ come from a non-Gaussian distribution, (3) without loss of the generality, it can be assumed that the expectation of mixed signals $x_i$ and components $s_i$ is zero (if it is not zero, they should be centred), and (4) for simplicity, it will be assumed that matrix $A$ is square, i. e., there is an equal number of mixed and original signals, although this condition can be relaxed [26].
After estimating the matrix $A$, its inverse, say $B$, can be calculated and the independent components can be obtained simply by:

$$s = Bx.$$

Here, it is also assumed that matrix $A$ is invertible. If this is not the case, there are redundant mixtures that could be omitted, in which case matrix $A$ would no longer be square; which is again the case when the number of mixtures is not equal to the number of independent components [26].

Thus, according to the previous four assumptions (or at least the first two), the IC model can be identified, meaning that the mixing matrix and the independent components can be estimated up to some trivial uncertainties which will be discussed shortly.

### Estimation

Now, the question is how to get independent components, i. e., how to estimate the matrix $A$? Non-Gaussianity is the key to estimating the IC model, with a central limit theorem in the background that claims that the distribution of the sum of independent random variables tends towards the Gaussian distribution, under certain conditions. Thus, the sum of two independent random variables usually has a distribution that is closer to Gaussian than either of the two original random variables [26].

Suppose that all of the independent components have an identical distribution. To estimate one of the independent components, consider a linear combination of $x_i$ which can be denoted by $y$, i. e. $y = w^T x = \sum_i w_i x_i$ where $w$ must be determined. If $w$ is row of the matrix inverse to the matrix $A$, the linear combination, $y$, is just one independent component. How to use the result of the central limit theorem in this context?

Let us define $z$ with $z = A^T w$. $y$ is then $y = w^T x = w^T As = z^T s$, i. e., a linear combination of $s_i$ with weights $z_i$. According to the central limit theorem, $z^T s$ is more Gaussian than any $s_i$, and is least Gaussian when it is exactly equal to $s_i$, which is true when exactly one of $z_i$ is non-zero. Therefore, if $w$ is a vector that maximizes the non-Gaussianity of $w^T x$, it will correspond to $z$ which has exactly one component other than zero. This further means that $w^T x = z^T s$ is equal to one independent component. In that way, one independent component is obtained. The $n$-dimensional space of the vector $w$ has $2n$ local maxima, two for each independent component, which corresponds to $s_i$ and $-s_i$ (independent components can be estimated only up to the sign). To find all the independent components, it is necessary to find all those local maxima. Because the independent components are uncorrelated: one can always limit the search to a space that gives estimates uncorrelated with the previous ones. This corresponds to orthogonalization in a suitably transformed space.

Besides maximizing non-Gaussianity, minimization of mutual information and maximum likelihood estimation can also be used in order to estimate the IC model. All approaches are (approximatively) equivalent (more can be read in [27] and [26]).

The mixing matrix and independent components can be estimated with some uncertainties: (1) the variances (energies) of independent components cannot be determined, and (2) the order of independent components cannot be determined, which could be considered as a disadvantages of the ICA.

## Preprocessing

Before applying the ICA on the data, it is usually very useful to do some preprocessing techniques that make the problem of ICA estimation simpler. The techniques are centering and whitening [27], [26]. Centering of the observed vector $x$ is performed by subtracting its mean in order for it to be zero-mean. On the other hand, whitening of $x$ refers to a transformation after which a white vector is obtained, i. e. a vector with components that are uncorrelated and whose variances equal unity.

Furthermore, since it is mentioned that independent components cannot be naturally ordered, the fundamental issue is also to determine the number of generated independent components. There are several theoretical approaches to this. Some of them were originally developed for PCA analysis (e. g., the Kaiser rule [6]), whereas others use information theory (e. g., the Akaike information [6]) or are based on cross-validation [30], [5]. However, few methods have been proposed specifically for the ICA analysis. For example, the Bayesian information criterion (BIC) can be applied to the Bayesian formulation of ICA to select the optimal number of components [33]. [30] defined a novel criterion that considers the global properties of transcriptomic multivariate data. The Maximally Stable Transcriptome Dimension (MSTD) is defined as the maximal dimension at which the ICA still does not generate a large proportion of highly unstable signals. In addition, components can also be ranked according to the value of the non-Gaussianity measure used (e. g., kurtosis) or the variance explained by the components, etc.

Nevertheless, since whitening is a very simple and standard procedure, it is advisable to reduce the complexity of the ICA problem by reducing the dimension of the data at the same time as performing whitening [27]. In this research, whitening was performed with the PCA, as in [27], because the PCA is a whitening procedure that can also discard some of the principal components and, at the same time, reduce the dimensionality. Criteria such as the scree graph, the percentage of the variance of each principal component (the Kaiser rule), the cumulative percentage of total variance [28] and the parallel analysis [23] were compared in order to determine the number of retained principal components.

### FastICA algorithm

One of the most commonly used algorithms for independent component analysis is the FastICA algorithm with properties as follows: (1) convergence is cubic (or at least square) under the assumptions of the IC model, (2) it is easy to use since, unlike gradient-based algorithms, a step size parameter does not have to be selected, (3) finds the independent components of almost any abnormal distribution using any nonlinear function g, unlike in many algorithms where some estimate of the probability distribution function must be initially available, and then the nonlinearity selected accordingly, (4) the performance of the algorithm can be optimized by selecting a suitable nonlinearity g (especially, robust and/or minimal variance algorithms can be obtained), (5) independent components can be estimated one by one, which is roughly similar to the search for a projection, and (6) it is parallel, distributed, computationally simple, and requires little memory space [27].

### Comparison between the principal component analysis and the factor analysis (FA)

PCA is a linear transformation based, mostly, on the variance maximization representation that is not based on a generative model, although it can be derived from one. The PCA model is invertible if all principal components are retained. Once the principal components are found, the original observations can be expressed as their linear functions, and the principal components can also be easily obtained as linear functions of the observations. On the other hand, the FA model is a generative latent variable model; the observations are expressed by the factors, but the values of the factors cannot be calculated directly from the observations. Furthermore, the FA, like the PCA, is a purely second-order statistical method: only covariances between the observed variables are used in the estimation, which is due to the assumption of Gaussianity of the factors. Finally, the factors are assumed to be uncorrelated, which also implies independence in the case of Gaussian data. On the other hand, the ICA is a similar generative latent variable model, where the factors or independent components are assumed to be statistically independent and non-Gaussian, which is a much stronger assumption [26].

When considering the problem of blind source separation, it would be possible to find many different uncorrelated representations of the signals that would not be independent and would not separate the sources, i. e. uncorrelatedness by itself is not enough to separate the components which is the reason why the PCA or the FA cannot separate the signals. The PCA and the FA yield components that are uncorrelated, and independence is a much stronger property than uncorrelatedness [26].

In this research, all three mentioned dimensionality reduction methods were performed separately.

## 3.2. Linear regression

Linear regression is one of the most widely used method in social sciences [47]. The standard model can be written as:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon,$$

where $y$ is the explained (dependent) variable, $x_1, ..., x_p$ are explanatory (independent) variables, $\beta_0$ is the intercept, $\beta_1, ..., \beta_p$ are coefficients associated with variables $x_1, ..., x_p$ and $\epsilon$ is the error term. The ordinary least squares method is used to estimate the model coefficients [47].

In this research, four linear regression models were produced to predict the energy costs of the public sector buildings. Model 1 uses the original variables as explanatory variables, Model 2 uses the independent components obtained from the ICA as the explanatory variables, Model 3 uses the principal components obtained through the PCA, whereas Model 4 uses the factors extracted by the FA as explanatory variables. Adjusted coefficient of determination ($R^2$), root mean square error (RMSE) and symmetric mean average percentage error (SMAPE) were used for the evaluation of the models' prediction performance and the comparisons between the models.

## 3.3. Data

The real data set that was used was obtained from the Energy Management Information System (EMIS), which is managed by the Agency for Legal Trade and Real Estate Brokerage (APN) in Croatia. It consists of the of the following characteristics of the public sector buildings in Croatia: constructional (share of windows, number of floors, energy coefficients of characteristic parts of the building, etc.), geospatial (region, county, etc.), meteorological, occupational (number of users, number of used days in a year, number of used hours in a day, etc.), energy data (energy costs), and heating and cooling characteristics (heated / cooled surface, internal heating / cooling temperature, etc.).

The total data set contains 1724 observations on 150 variables. 149 of them are explanatory variables, whereas the annual energy costs is an explained variable. For the ease of reading and clarity of the paper, the descriptive statistics of the variables are omitted. Before producing a linear regression models to predict the energy costs of the public buildings, the sample was first divided into a training and a testing part in the ratio 80:20, i. e., 1379 observations were in training, whereas 345 obsevations were in the testing part.

The Kaiser-Meyer-Olkin measure of sampling adequacy is used to check whether the correlation matrix of the sample is suitable for conducting the factor analysis, and is also used for the purposes of the principal and independent components analysis [31]. The Bartlett's test is used for the same purpose as it tests whether a correlation matrix is an identity matrix [3].

## 4. Results and discussion

The whole following analysis was performed in the R software. Firstly, the Kaiser-Meyer-Olkin (KMO) measure and the Bartlett's test were used in order to check the sampling adequacy. The KMO measure resulted in 0.85, and the Bartlett's test showed the $p$-value as significant ($p$-value $= 0$). Both of which justifiy the adequacy of sampling. As a first preprocessing step, prior the independent components analysis, the explanatory variables were centered by subtracting their means. The second step that precedes the ICA was whitening. Within that step, the principal component analysis was performed. The principal component analysis was chosen as a whitening procedure since the independent component analysis does not sort the components, and, in that way, does not perform the dimensionality reduction by itself, while the principal component analysis can discard some of the principal components according to certain criteria. In order to determine a number of retained principal components, the following criteria were compared: (1) the scree graph, (2) the Kaiser rule, (3) the cumulative percentage of total variance, and (4) the parallel analysis. The scree graph is shown in Figure 1., and the results according to all criteria are presented in Table 1.

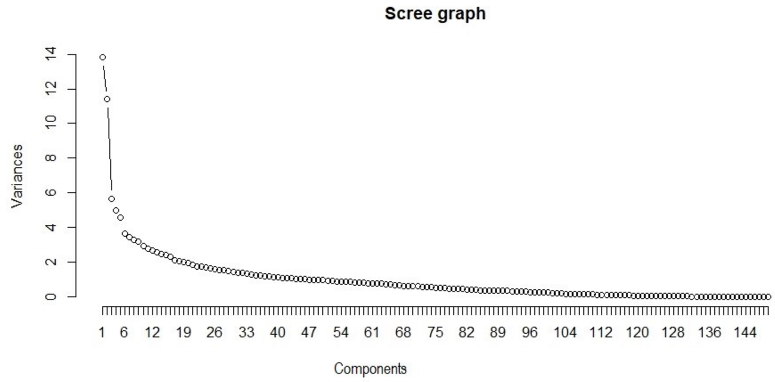

Figure 1: Scree graph

| Criteria | Number of retained principal components | Total variance explained (%) |
|---|---|---|
| Scree graph | 5 | 27.16 % |
| Kaiser rule | 45 | 76.48 % |
| Cumulative percentage of total variance | 37 | 70.69 % |
| Parallel analysis | 34 | 68.26 % |

Table 1: Number of retained principal components according to all used criteria and percentage of the total variance that is explained

The parallel analysis results were used, and 34 components were retained as they explain a substantial proportion of the total variance (68.26 %), while, at the same time, reduce the dimensionality the most.

The independent component analysis was then performed by the FastICA algorithm, which resulted in 34 components that are as independent and non-Gaussian as possible. Consequently, the original data set of buildings characteristics was significantly reduced, from 149 to 34 dimension. This finding is in conjunction with the literature that states that this method greatly contributed to reducing the dimensionality of the data set and extracting main underlying components from various data.

Moreover, for the purpose of a comparison, the factor analysis was also performed. The FA resulted in 39 factors (as suggested by the parallel analysis) that also explain a substantial proportion of the total variance (65.2 %).

Finally, four different linear regression models were trained and their performance on the testing subsample was evaluated. Model 1 was trained on original variables, Model 2 on independent components, Model 3 on principal components and Model 4 on factors as input (explanatory variables). The models were compared in terms of adjusted $R^2$ coefficient, RMSE and SMAPE with results given in Table 2.

| | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| $R^2$ [%] | 80.06 | 71.97 | 68.21 | 80.06 |
| RMSE | 274 107.5 | 180 289.8 | 195 728.1 | 274 107.5 |
| SMAPE | 54.78 | 49.15 | 60.25 | 54.78 |

Table 2: The evaluation of models performance

It can be seen that Model 2, Model 3 and Model 4 reduced the dimensionality to a certain extent. In addition, Model 4 (FA - LR) had the same performance as Model 1 (LR) in terms of $R^2$, RMSE and SMAPE. Model 1 and Model 4 had the highest $R^2$, but also the highest RMSE, while they had SMAPE lower than Model 3 (PCA - LR). Model 3 also had the lowest $R^2$ coefficient. Finally, Model 2 (ICA - LR) had the $R^2$ coefficient lower than Model 1 and Model 4, but it resulted in the lowest RMSE and SMAPE errors. Therefore, it can be said that Model 2 outperformed other models in terms of prediction.

## 5. Conclusion

This paper initially addressed the motivation for building a model which predicts the energy costs of the public sector buildings. Considering that the building sector is one of the largest individual consumers of energy, in order to be able to implement the directives of the European Union to increase energy efficiency and achieve nearly zero-energy buildings, it is necessary to properly plan measures for the construction of new buildings or the renovation of existing ones. A real data set of public sector buildings in the Republic of Croatia was used, which contained a large number of variables (construction, occupational, energy, etc.). The effects that can be caused by a large number of variables in the model was discussed in terms of "blessings" and "curses" of dimensionality, with the emphasis on the latter. It is also mentioned that an inclusion of an irrelevant variable may increase multicollinearity and a technique's ability to fit the sample data, with a cost of overfitting the sample data and making the results less generalizable to the population, whereas the omission of relevant variables can bias the results and negatively affect any interpretation.

Secondly, the analysis of independent components was conducted and presented as one of the methods for reducing dimensionality, i. e., extraction of variables. It is a method that has recently become more common in order to find the hidden factors behind a set of random variables. The method estimates a linear representation from the data that would consist of components that are statistically independent (or at least as independent as possible) and non-Gaussian. It is often considered as an extension of the principal component analysis or the factor analysis. Therefore, a comparison between the principal component analysis and the factor analysis is provided, along with the motivation, definition, assumptions, estimation and ambiguities of the independent component analysis. In addition, the main properties of the FastICA algorithm, one of the most commonly used algorithms for the independent component analysis, were mentioned. Furthermore, since the ICA does not result in sorted components, in this paper, data whitening was performed by the principal component analysis that may, according to certain criteria, discard some of the components. In this paper, criteria such as a scree graph, the percentage of the variance of each principal component (the Kaiser rule), the cumulative percentage of the total variance and the parallel analysis were compared and 34 components were retained, as suggested by the parallel analysis, which explained around 68.26 % of the total variance. The FastICA algorithm was performed in order to estimate independent components and the initial set of explanatory variables were reduced from 149 to 34 dimensions. Moreover, the factor analysis was performed which resulted in 39 factors that explained 65.2 % of the total variance.

Finally, the ultimate goal of this research was to create a model for estimating and predicting the annual energy costs of the public sector buildings based on the mentioned characteristics. With that purpose, four different linear regression models were created and evaluated: Model 1 was conducted on original variables, Model 2 was conducted on independent components, Model 3 was conducted on principal components and Model 4 was conducted on factors as input. The purpose of all models is to support decision-making and planning for the implementation of energy renovation measures or construction of buildings. The models were compared in terms of adjusted $R^2$ coefficient, RMSE and SMAPE on testing the sub-sample, and Model 2 resulted in the lowest RMSE and SMAPE errors. Finally, Model 2 could be implemented in the information system of decision makers in order to be used in planning measures for the construction of new buildings or renovation of existing ones. This is an important applied contribution in the context of achieving energy efficiency and the objectives of the European Union directives. Further research should include many different dimensionality reduction methods, e. g., methods and optimization algorithms for variable selection, and methods for modelling this data set such as machine learning methods, in order to achieve the best prediction performance. However, more accurate models could be created if data on the characteristics of the buildings before and after renovation were available. It would then be possible to directly observe and model differences in energy costs over time. The limitation of the buildings' characteristics used in this research is that they were collected at one point in time and there is no information on the changes that occurred as a result of renovations. This could be solved by changes in the information system of the decision makers and by introducing the obligation to enter such data in the information system when a change in the buildings' characteristics occurs. Ultimately, that could have a significant impact not only on governmental financial costs, but also on the environmental care and sustainability.

## References

[1] Azaza, M., Eskilsson, A., Wallin, F. (2015). An open-source visualization platform for energy flows mapping and enhanced decision making. In Energy Procedia, 158, 3208-3214. doi:10.1016/j.egypro.2019.01.1006

[2] Banoczy, E., Szemes, P. T. (2014). Simulation-based optimization in energy efficiency retrofit for office building. In Proc. IEEE/SICE International Symposium on System Integration, 222-227.

[3] Bartlett, M. S. (1951). The Effect of Standardization on a chi square Approximation in Factor Analysis. Biometrika, 38, 337-344. doi:10.2307/2332580

[4] Bennasar, M., Hicks, Y., Setchi, R. (2015). Feature selection using Joint Mutual Information Maximisation. Expert Systems With Applications, 42(22), 8520-8532. doi:10.1016/j.eswa.2015.07.007

[5] Bro, R., Kjeldahl, K., Smilde, A. K., Kiers, H. A. (2008). Cross-validation of component models: a critical look at current methods. Anal Bioanal Chem, 390(5). doi:10.1007/s00216-007-1790-1

[6] Cangelosi, R., Goriely, A. (2007). Component retention in principal component analysis with application to DNA microarray data. Biol Direct, 2(2).

[7] Chagnaa, A., Ock, C.- Y., Lee, C.- B., Jaimai P. (2007). Feature Extraction of Concepts by Independent Component Analysis. International Journal of Information Processing Systems, 3(1). [doi:10.3745/jips.2007.3.1.033]

[8] Chen, Y., Jiang, H., Li, C., Jia, X., Ghamisi, P. (2016). Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks. IEEE Transactions on Geoscience and Remote Sensing, 54(10), 6232-6251.

[9] Cheung, C. T., Mui, Ky. W., Wong, L. T. (2015). A hybrid simulation approach to predict cooling energy demand for public housing in Hong Kong. Building simulation, 8(6), 603-611. [doi:10.1007/s12273-015-0233-8]

[10] Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. Aide-Memoire of the lecture in American Mathematical Society conference: Math challenges of 21st Centur. `https://www.researchgate.net/publication/220049061_High-Dimensional_Data_Analysis_The_Curses_and_Blessings_of_Dimensionality` [Accessed 16/03/22].

[11] Engreitz, J. M., Daigle, B. J. Jr., Marshall, J. J., Altman, R. B. (2010). Independent component analysis: Mining microarray data for fundamental human gene expression modules. Journal of Biomedical Informatics, 43(6), 932-944.

[12] European Parliament, Council of the European Union (2010). Directive 2010/31/EU of the European Parliament and the Council of 19 May 2010 on the energy performance of buildings. Official Journal of the European Union, 153, 13–35.

[13] European Parliament, Council of the European Union (2012). Directive 2012/27/EU of the European Parliament and the Council of 25 October 2012 on the energy efficiency amending Directives

2009/125/EC and 2010/30/EU and repealing Directives 2004/8/EC and 2006/32/EC. Official Journal of the European Union, 153, 13–35.

[14] European Parliament, Council of the European Union (2018a). Directive (EU) 2018/844 of the European Parliament and of the Council of 30 May 2018 amending Directive 2010/31/EU on the energy performance of buildings and Directive 2012/27/EU on energy efficiency. Official Journal of the European Union, 156, 75–91.

[15] European Parliament, Council of the European Union (2018b). Directive (EU) 2018/2002 of the European Parliament and of the Council of 11 December 2018 amending Directive 2012/27/EU on energy efficiency. Official Journal of the European Union, 328, 210.-230.

[16] Fu, Y., Li, Z. W., Zhang, H., Xu, P. (2015). Using Support Vector Machine to Predict Next Day Electricity Load of Public Buildings with Sub-metering Devices. In Proceedia Engineering, 1016-1022. doi:10.1016/j.proeng.2015.09.097

[17] Gorban, A. N., Makarov, V. A., Tyukin, I. Y. (2020). High-Dimensional Brain in a High-Dimensional World: Blessing of Dimensionality. Entropy, 22(82). 10.3390/e22010082

[18] Gorban, A. N., Tyukin, I. Y. (2018). Blessing of dimensionality: mathematical foundations of the statistical physics of data. Phil. Trans. R. Soc. A., 376(20170237). doi:10.1098/rsta.2017.0237

[19] Guyon, I., Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 3, 1157-1182.

[20] Hair, J. F. Jr., Black, W. C., Babin, B. J., Anderson, R. E. (2014). Multivariate Data Analysis (7th ed.). Harlow: Person Education.

[21] Hastie, T., Tibshirani, R., Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction (2nd ed.). New York: Springer.

[22] Hodak, K., Has, A., Mokriš, M. (2022). CLUSTER ANALYSIS FOR PROFILING PUBLIC SECTOR BUILDINGS OF CONTINENTAL CROATIA AS A SUPPORT FOR REGIONAL DEVELOPMENT. In RED 2022 Proceedings, 271-280.

[23] Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30(2), 179-185. doi: 10.1007/BF02289447

[24] Howard, B., Parshall, L., Thompson, J., Hammer, S., Dickinson, J., Modi, V. (2012). Spatial distribution of urban building energy consumption by end use. Energy and buildings, 45, 141-151. doi: 10.1016/j.enbuild.2011.10.061

[25] Hyvärinen, A. (2013). Independent component analysis: recent advances. Phil. Trans. R. Soc. A., 371(20110534). doi:/10.1098/rsta.2011.0534

[26] Hyvärinen, A., Karhunen, J., Oja, E. (2001). Independent Component Analysis. New York: John Wiley & Sons, Inc.

[27] Hyvärinen, A., Oja, E. (2000). Independent Component Analysis: Algorithms and Applications. Neural Networks, 13, 411-430.

[28] Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). New York: Springer.

[29] Jolliffe, I. T., Cadima, J. (2016). Principal component analysis: a review and recent developments. Philos. Trans. A. Math. Phys. Eng. Sci., 374(2065). doi: 10.1098/rsta.2015.0202

[30] Kairov, U., Cantini, L., Greco, A., Molkenov, A., Czerwinska, U., Barillot, E., Zinovyev A. (2017). Determining the optimal number of independent components for reproducible transcriptomic data analysis. BMC Genomics, 18(712). doi:10.1186/s12864-017-4112-9

[31] Kaiser, H.F. (1974). An index of factor simplicity. Psychometrika, 39(1), 31-36. doi:10.1007/BF02291575

[32] Kavaklioglu, K. (2019). Principal components based robust vector autoregression prediction of Turkey's electricity consumption. Energy Systems-Optimization modelling Simulation and Economic Aspects, 10(4), 889-910. doi:10.1007/s12667-018-0302-z

[33] Krumsiek, J., Suhre, K., Illig, T., Adamski, J., Theis, F. J. (2012). Bayesian independent component analysis recovers pathway signatures from blood metabolomics data. J Proteome Res, 11:4120–31. doi:10.1021/pr300231n

[34] Lee, J.A., Verleysen M. (2007). Nonlinear Dimensionality Reduction. New York: Springer.

[35] Lee, S., Shen, H. P., Truong, Y., Lewis, M., Huang, X. M. (2011). Independent Component Analysis Involving Autocorrelated Sources With an Application to Functional Magnetic Resonance Imaging. Journal of the American Statistical Association, 106(495), 1009-1024. doi: 10.1198/jasa.2011.tm10332

[36] Liu, H., Wang, J. (2011). Integrating Independent Component Analysis and Principal Component Analysis with Neural Network to Predict Chinese Stock Market. Mathematical Problems in Engineering, 382659. doi: 10.1155/2011/382659

[37] Ma, M., Yan, R., Cai, W. G. (2017). An extended STIRPAT model-based methodology for evaluating the driving forces affecting carbon emissions in existing public building sector: evidence from China in 2000-2015. Natural Hazards, 89, 741-756. doi:10.1007/s11069-017-2990-4

[38] Maki, S., Ashina, S., Fujii, M., Fujita, T., Yabe, N., Uchida, K., Ginting, G., Boer, R., Chandran, R. (2018). Employing electricity-consumption monitoring systems and integrative time-series analysis models: A case study in Bogor, Indonesia. Frontiers in energy, 12(3), 426-439. doi:10.1007/s11708-018-0560-4

[39] Menze, B. H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics, 10(213). doi:10.1186/1471-2105-10-213

[40] Nordhausen, K., Oja, H. (2018). Independent component analysis: A statistical perspective. Wiley Interdisciplinary Reviews-Computational Statistics, 10(5). doi:10.1002/wics.1440

[41] Pal, M., Foody, G. M. (2010). Feature Selection for Classification of Hyperspectral Data by SVM. IEEE Transactions on Geoscience and Remote Sensing, 48(5), 2297-2307. doi:10.1109/tgrs.2009.2039484

[42] Ruggieri, S. (2019). Complete Search for Feature Selection in Decision Trees. Journal of Machine Learning Research, 20(104). https://www.jmlr.org/papers/volume20/18-035/18-035.pdf

[43] Ruiz, L. G. B., Cuellar, M. P., Calvo-Flores, M. D., Jimenez, M. D. P. (2016). An Application of Non-Linear Autoregressive Neural Networks to Predict Energy Consumption in Public Buildings, Energies, 9(9). doi:/10.3390/en9090684

[44] Summerfield, A. J., Lowe, R. J., Oreszczyn, T. (2010). Two models for benchmarking UK domestic delivered energy. Building Research and Information, 38(1), 12-24. doi:10.1080/09613210903399025

[45] Theodoridis, S., Konstantinos, K. (2009). Pattern Recognition. London: Academic Press.

[46] Tonković, Z., Mitrović, S., Zekić-Sušac, M. (2018). Business Intelligence System for Managing Natural Gas Consumption of Public Buildings. In Proc. International Scientific Conference on Economic and Social Development (pp. 769-778.)

[47] Wooldrige, J. M. (2012). Introductory econometrics—a modern approach (5th ed). Boston: Cengage Learning.

[48] Zekić-Sušac, M., Knežević, M., Scitovski, R., (2020). Deep Learning in modelling Energy Cost of Buildings in the Public Sector. In Martínez-Álvarez, F., Tronosco Lora, A., Sáez Muñoz, J. A., Quintián, H, Corchado, E. (Eds.). Advances in Intelligent Systems and Computing (pp. 101-110). Cham: Springer International Publishing Ag.

[49] Zekić-Sušac, M., Mitrović, S., Has, A. (2021). Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities. International Journal of Information Management, 58. doi:10.1016/j.ijinfomgt.2020.102074

[50] Zekić-Sušac, M., Knežević, M., Scitovski, R., (2021). Modelling the cost of energy in public sector buildings by linear regression and deep learning. Central European Journal of Operations Research, 29, 307-322. doi:10.1007/s10100-019-00643-y

[51] Zekić-Sušac, M., Has, A., Knežević, M. (2021). Predicting Energy Cost of Public Buildings by Artificial Neural Networks, CART, and Random Forest. Neurocomputing, 439, 223-233. doi:10.1016/j.neucom.2020.01.124