# Applicability of Multivariate Linear Regression in Building Energy Demand Estimation

Tamas Storcz[1*], István Kistelegdi[2], Kristóf Ronald Horváth[3], Zsolt Ercsey[1]

[1] Department of Systems and Software Technologies, Faculty of Engineering and Information Technology, University of Pécs, Pécs 7624, Hungary

[2] Energy Design Research Group, János Szentágothai Research Centre, University of Pécs, Pécs 7624, Hungary

[3] Marcel Breuer Doctoral School, Faculty of Engineering and Information Technology, University of Pécs, Pécs 7624, Hungary

Corresponding Author Email: storcz.tamas@mik.pte.hu

**ABSTRACT**

The vision of the research project is to find an energy optimal building configuration, suitable for specified requirements and restrictions. The first step on this way is to create a measure to compare building configurations, faster than explicit energetic simulations. The current study examines the applicability of multivariate linear regression to support the solution of building optimization problems. During the study, multivariate linear regression models were created to estimate the expected annual heating energy demand of building configurations and examined their accuracy Between examinations, the models were modified so that the complexity was increased only to such an extent that the approximation was still sufficiently accurate. The result was a multivariate linear model that estimated the expected output for unknown descriptive variables with a 0% relative error and a 1.6% standard deviation. The $R^2$ point of the estimates was 0.9884. Based on these, the model was considered applicable in the search space defined by the training patterns.

## 1. INTRODUCTION

The built environment represents one of the largest energy-consuming industry in the world [1, 2]. The main reason for this can be identified and traced back to the beginning steps of establishing a building; namely the design phase [3, 4]. At this stage numerous decisions are made that has enormous effects on the later realized buildings energy and comfort performance.

The research project targets to create an artificial intelligence-based model to advise the most energy efficient building configuration for predefined customer and designer requirements and restrictions of laws and orders. As a first step, a linear regression model is tested to estimate heating energy demand simulation results. Building configurations of the experiment was restricted to a 6-block residential building and simulations were made by using local meteorological data. When applying machine learning concepts, first used linear regression estimated the energy demand with R2=0.73. After providing non-linearity by extending the input set by multiplicative combinations of original inputs up to 3rd power. Resulting polynomial regression obtained 0% relative error average and 1.6% standard deviation. The R2 point of the estimates was 0.9884. This accuracy is not worse than the accuracy of weather statistics based energetic simulations, therefore accepted for further work by architect experts.

## 2. RELATED WORKS

It is crucial to fully understand the effects of (certain) design parameters to decrease this negative effect on our environment.

Certain types of sensitivity analysis methods are the most effective for this purpose [5, 6].

Regression models serve as an obvious solution for analyzing building properties, since building design variables can be easily translated or converted to numerical values that are in correlation. Linear regression is a widely accepted and applied mathematical approach/method, known for helping to analyze large databases in terms of the relationships between independent and dependent variables.

Several studies utilizing regression models focus on how natural lighting effects heating and cooling energy use [7], or lighting energy demand or concentrate on the effects of building design parameters on daylighting [8].

Numerous studies have proved the predictive power of regression models. Building repair time estimation is also representing high importance considering the increase in deteriorating buildings [9]. Sajjad et al. [10] proposed three buildings energy consumption prediction by a unique multi-output (MO) sequential learning model predicting heating and cooling loads also. A multiple linear regression model was developed [11] with an analysis of variance method (ANOVA) for predicting the annual heating and cooling energy demands in the three climate regions.

Another study identified a total of 12 key building design variables through parametric analysis [12], and considered as inputs in the regression models. A pseudo-random number generator based on three simple multiplicative congruential generators was employed to generate random designs for evaluation of the regression models. The comparative analysis showed that the margin of error at these building cases are 10%. Based on these results an estimation of energy savings can be made.

Research was focused on indoor environmental quality in zero energy buildings to predict the expected variation of IEQ according to various standards [13].

A study about urbanized areas [14] aimed at developing a surrogate model-based integrated optimization system to obtain energy-optimal thermal designs for residential buildings in the most urbanized cities in Turkey under different levels of budget constraints.

Kudabayev et al. [15] studied the thermal system of a room in a building. Using a proposed a mathematical model the results show that thermal insulation and thermal capacity of walls have effect heating and cooling energy demand.

According to the idea above, another study describes numerical simulations, made using ANSYS/FLUENT 16 software, to select thermal agents for building parts [16].

A further study aimed at creating tools to assess the relationship between heating energy use and indoor temperatures at different levels of occupant behavior in residential buildings [17].

According to Zou et al. [18] three steps are required to provide architects with robust and accurate design references when conducting design tasks. The first step is to create a database by generating the building objects randomly and performing building simulations on them. The next step is to train artificial neural network (ANN) models as a surrogate for demanding building simulation to predict the building performance accurately and quicker than a simulation. The last step is the optimization based on actual design constraints.

Harish and Kumar have conducted a review of all the significant modeling methodologies which have been developed and adopted to model the energy systems of buildings [19].

The complex correlations or relations between design variables cannot be described with a simple mean average based weighted order system. Equal values can occur, and no order of equal values is properly established, or nearly identical values may receive the same score. Cases with different energy or comfort score compositions may also give the same total score.

## 3. BASELINE DATA AND BOUNDARY CONDITIONS

Creating a database is essential to discover correspondences and their margins related to the connection of the input and output parameters. It is also crucial in developing a regression model.

The generated configurations scale and the introduced modular system was based on a real award-winning active house. This conscious decision was made to ensure later validation steps with the help of measured data from a building monitoring system. Furthermore, this fact is forming the basis of an 80% energy saving [20, 21] potential via utilizing passive design components.

During an exemplary modelling uniform building blocks were used to form a generic family house. From all possible options 167 building configurations were selected by experts, based on various architectural design rules. These configurations were transformed into building cases by applying several different structures, wall- window ratios, and orientations to each configuration.

It is possible to deliver detailed, complex analysis in annual, hourly resolution about the time dependent daylight, comfort and energy behavior of the buildings with the help of dynamic

thermal simulation calculations, taking into account the local climate conditions. IDA ICE 4.8 indoor climate and energy dynamic thermal simulation software provided the calculation engine. Further, IDA ICE is capable of high-level visual representation and post processing considering various standards.

To create the model of the present paper, we created 5010 simulation samples using the IDA ICE dynamic thermal simulation program. The geographic location of the site and the local climate conditions were considered at the thermal simulations (ASHRAE IWEC2 Climate Database). Artificial illumination, equipment and occupants were modelled according to standard usage of typical single-family houses. The heat transfer system of the interior and the heating central system was scaled with appropriate performance. The air handling unit (AHU) system provides regular satisfactory air change (ACH) rates. The same boundary conditions were applied to the 5010 building cases' simulations: climate and location data, HVAC and operation settings. One of the purposes of the simulations is a detailed comparative analysis of heating energy demand. Based on the identic boundary conditions and settings, differences in the energy simulation results were expected, because of diverse building shape.
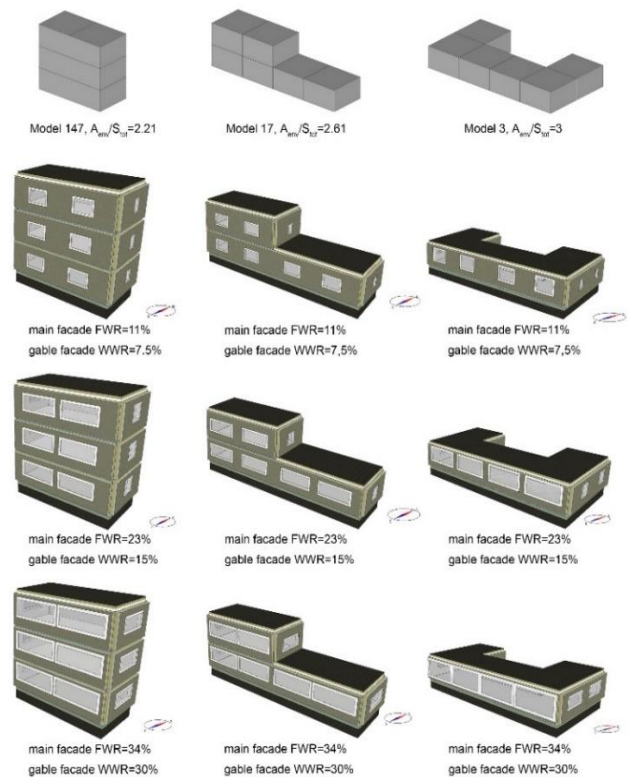
## 4. SIMULATED MODELS



**Figure 1.** Extension of configurations to building models

The following building design variables were considered, to provide the simulation engine with the necessary building properties (Figure 1).

• Two different structures, one meeting the minimum standards, while the other is almost meeting or exceeding the passive house standards.

• Three different window wall ratios were applied to the generated geometry configurations. Namely 30%, 60% and

90% ratios were prepared on the main facade (-area representing the largest surface facing the same direction).

• Five different orientations were chosen to enable the simulation engine applying the chosen climate weather dataset. The orientation directions are 90, 135, 180, 225, 275 degrees, where 0 degrees represents North.

Five orientations were considered in the investigation, including the most solar radiation dominant South, and two surroundings on each side: East, South-East, South-West and West directions. The main facades of the models were turned in the five different directions, because these have the most effect on (can decrease) heating energy demand.

Internal walls and partitions are not included, only slabs were used to divide the levels from each other. This simplification was consciously undertaken. Slabs were necessary in order to avoid multiple storey-high indoor spaces, which are not typical for residential housing. Diverse sloping roofs, galleries, stairs, etc. represent all these – in first research step non-traceable and overcomplicating – issues, which must be considered in further research. All structure, glazing ratio and orientation combinations of every building configuration resulted a total of 5,010 building model sample variations for the simulation.

## 5. DESCRIPTIVE VARIABLES OF BUILDING CONFIGURATIONS

Strong dependency of building block coordinates is hard to be represented as independent inputs. Therefore, instead of representing the building configurations using the block coordinates, 14 descriptors were introduced into the system. Groups of number of different surfaces, edges and vertices serve as set of simple descriptive variables as shown in Figure 2 and Table 1.
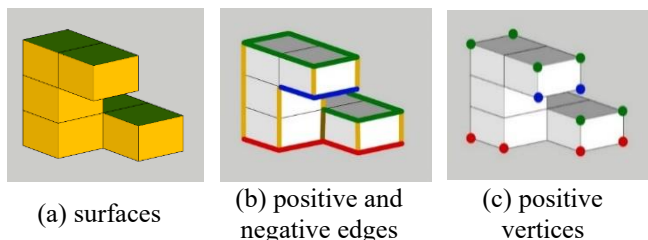


(a) surfaces    (b) positive and negative edges    (c) positive vertices

**Figure 2.** An exemplary geometry configuration illustrating introduced design variables

Further to the set of simple descriptive variables, a complex architectural descriptor is also introduced. The transmission heat loss surface to heated floor space ratio: $A_{envelope}/S_{total}$ is used, expressed in Eq. (1).

$$\frac{A_{env}}{S_{tot}} = \frac{A_{env-air} + A_{env-ground} \cdot 0.71}{S_{tot}} \qquad (1)$$

where $A_{env-air}$ means the roof and façade structures' surface in $m^2$, $A_{env-ground}$ means floor surface adjacent to ground in $m^2$ and $S_{tot}$ means the total net floor space in $m^2$. The factor 0,71 is a dynamic thermal simulation-based value, expressing an average of lower transmission heat loss rates related to ground surfaces in family house sized geometry cases. The smaller the $A_{env}/S_{tot}$ value, the higher the energy efficiency of the building geometry.

In the case of the current study the dependent variable (output variable) is the estimated annual heating energy demand.

In the present work, we propose the use of regression models to replace the simulation process in the search space under consideration. In the figure, the thick red arrow represents the generation of the expert input data. The green colored modelling is either simulation, or regression should the search space be adequately known.

**Table 1.** Parameters of models

| Name of the parameter | Explanation | Group |
|---|---|---|
| **Generic input** | | |
| structure | various | building design parameter |
| wall window rate | percent | |
| orientation | degree | |
| **Set of simple input variables** | | |
| g | connected to ground | surfaces |
| r | roof | |
| b | balcony | |
| w | external wall | |
| a | arcade (slab connected to air) | |
| g edge | ground edge_N.o. | edges |
| r positive edge | roof edge_N.o. | |
| r negative edge | roof edge_N.o. | |
| a positive edge | wall edge_N.o. | |
| a negative edge | wall edge_N.o. | |
| arcade positive edge | arcade edge_N.o. | |
| g vertex | g vertex_N.o. | vertexes |
| a air vertex | wall vertex_N.o. | |
| arcade vertex | arcade vertex_N.o. | |
| **Complex geometry descriptor input** | | |
| A/S | envelope surface floor surface | coefficient |
| **Output** | | |
| heating energy demand | [kWh/a] | energy results |

## 6. MACHINE LEARNING

According to Mitchel's illustrative, formalizable definition of machine learning [22], "a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." In other words, solutions of similar examples may improve the performance of problem solving together with its measure. Therefore, to use the machine learning procedure, the followings are required:

• Task class definition: Estimating the annual energy demand based on building properties.

• Experience from solving similar tasks: Analytical energy demand simulations using IDA ICE simulation software.

• Efficiency measurement: Quadratic and absolute differences.

With the above mentioned, let us create the hypothesis where the practical model solves unknown tasks within the task class with acceptable accuracy based on the previously solved similar examples. For the model creation and efficiency measurement, the IDA ICE simulation dataset is split into two parts. A major part of simulation results (75%) was used to build or train the model. This set is called "training data". While through the remaining 25% of experience, the efficiency of the model was measured or tested. This smaller set is called "test data." In the first step, model parameters are set based on training data, then model performance is

measured using test data, which is unknown for the model. The performance measure is the absolute and relative distance of the model output and the already known simulation result, calculated using Manhattan (L1) distance. The model is accepted or rejected by the experts based on average distances.

In the current paper, linear regression is applied. Its advantages are its speed and easy application, but disadvantage is the linear approximation, which will be detailed later.

# 7. LINEAR REGRESSION

Regression is a wildly and commonly used statistical method, whereby details of dependencies of explanatory and response variables are to be explored and determined.

Linear regression [23] is a special case of general regression calculation where the dependent variable is obtained as a linear combination of descriptive variables, as shown in Eq. (2), with a first-order Taylor series.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon \quad (2)$$

$$h(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m \quad (3)$$

$$\varepsilon = y - h(x) \quad (4)$$

where, $y$ is the dependent or output variable, $x_1 \ldots x_m$ are descriptor or explanatory variables, $\beta_0 \ldots \beta_m$ are parameters of linear combination, $h(x)$ of Eq. (3) is the linear regression hypothesis function and Eq. (4) shows $\varepsilon$, the approximation error of linear regression.

## 7.1 Hypothesis

According to Eq. (3), there exists a set of $\beta_0 \ldots \beta_m$ parameters of the hypothesis, where approximation error of Eq. (4) would remain under a certain acceptance threshold specified by architects. The least squares analytical method [24] is used to determine appropriate $\beta_0 \ldots \beta_m$ parameters.

## 7.2 Model creation

When creating the multivariate linear regression model, it is supposed that the dependent variable is determined by many explanatory variables.

Values of the explanatory variables of all measures can be described by the matrix of Eq. (5)

$$X_o = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix} \quad (5)$$

where, $n$ is the number of measures, $m$ is the number of explanatory variables. Then $x_{ij}$ item of the matrix is the value of j-th explanatory variable during the i-th measure/observation.

Members of Eq. (3) hypothesis can be described as column vectors. Please note that the first column index of the matrix $X_0$ of Eq. (5) is 1, but the first row index of $\beta$ column vector of Eq. (6.b) is 0. In other words, there is no $x_0$ explanatory variable beside the constant parameter ($\beta_0$) in Eq. (3), therefore $X_0$ matrix must be extended with a column of constant 1 values, according to Eq. (6).

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & X_{nm} \end{bmatrix} \quad (6)$$

As a consequence, Eq. (2) of linear regression can be formulated as Eq. (7):

$$y = X\beta + \varepsilon \quad (7)$$

where, $\varepsilon$ is the column vector containing linear regression approximation errors of each measure/observation and its $\varepsilon_i$ element is the approximation error of the i-th measure.

In the calculation of $\beta$ parameters, to solve a linear equation with m unknowns, the method of least squares is applied. Here the $\beta$ parameter vector is calculated by Eq. (8).

$$\tilde{\beta} = (X'X)^{-1} X'y \quad (8)$$

The estimation of the newly created linear regression model, applying $\beta$ parameter vector can be given by Eq. (9).

$$\tilde{y} = X\tilde{\beta} \quad (9)$$

The absolute error of linear regression is given in Eq. (10).

$$\varepsilon_a = y - \tilde{y} \quad (10)$$

## 7.3 Model evaluation

In the current paper the $R^2$ metrics [25], detailed in Eq. (11), is used to evaluate the performance of the regression model.

$$R^2 = 1 - \frac{\sum (y - \tilde{y})^2}{\sum (y - \bar{y})^2} \quad (11)$$

The best score of $R^2$ measure is $1$, the lower results mean lower performance. Please note that, negative values can also be results of poor performance.

The performance is further measured by absolute error, given above, and the relative error, given in Eq. (12). Please note that this error is relative to the esteemed value.

$$\varepsilon_r = \frac{(y - \tilde{y})}{y} \quad (12)$$

## 7.4 Approximation results

During the performed architectural experiment, out of 5,010 simulations, 3,757 samples (75%) were used for model creation and remaining 1,253 samples (25%) were used for independent test of model performance. The selection of train and test samples was done by pseudo-random decisions with even distribution. Please note that should significant difference between training and test results occur, the model structure or the train-test sample selection method must be revised.

In the first case of the experiment, besides engineering inputs (structure, wwr, orientation), to describe a building configuration one complex descriptor (A/S) was used. In the second case, 14 simple descriptors were numerated from the building configuration (for example faces, edges, vertices) and these were used with engineering input. The inputs are detailed in Table 1.

Through the experiment cases, engineering inputs have not changed, therefore the difference of cases reflects usefulness of the applied building configuration descriptors. As shown in first row of Table 3 and Figure 3, $R^2$ score of multiple simple descriptors is slightly higher (0.7489) than score of single complex descriptor (0.7273), but both are far from the best value (1.0).
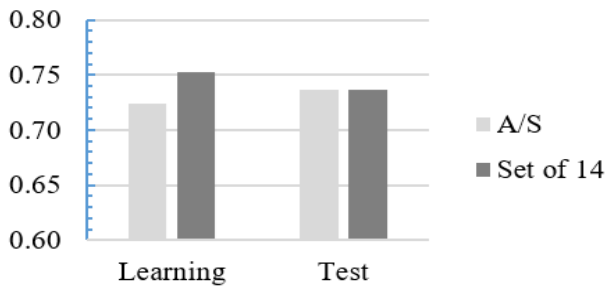


**Figure 3.** $R^2$ score of linear regression

Based on the $K^2$ test proposed by D'Agostino [26] and the Anderson-Darling [27] test performed on the distribution of error functions, it can be stated that absolute error function does not, but relative error function has normal distribution. Therefore, the mean and standard deviation are valid descriptors of the relative error. Applying these, it can be noted that even though relative error is under 1%, according to $2\sigma$ rule of normal distribution, the 8% std. dev (Table 2, Figure 4). results a ±16% interval for containing 95% of estimated values.

**Table 2.** Standard deviance of relative error [%]

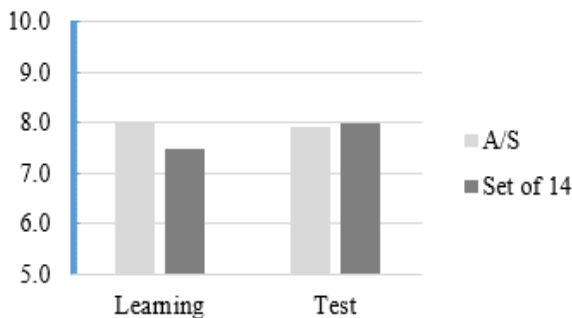| std.<br>dev. | A/S | | 14 descriptors | |
|---|---|---|---|---|
| | **Training** | **Test** | **Training** | **Test** |
| σ | 8.0 | 7.9 | 7.5 | 8.0 |



**Figure 4.** Absolute error $\sigma^2$ of linear regression

The absolute error is not normal, thus there is no information about its accuracy, please note that average of absolute approximation error was under 1.0 kWh/year.

Based on the above the following conclusions can be drawn. The function of annual heating energy demand cannot be estimated as a linear combination of the described input variables neither using single complex or multiple simple building structure descriptors. Understanding the extracted parameters of normal distribution, more than 30% of estimations are expected to be between 2 and 3 σ from the mean, in other words that results 16-24% error for more than 30% of estimations. In this form, the linear regression model is not suitable for replacing building energy requirement simulations.

# 8. NON-LINEAR REGRESSION

As presented in the previous paragraph, approximation of annual heating energy demand by linear regression is only possible with unacceptable error. This is obviously because the energy demand function is not linear. It would be necessary to apply a non-linear approximation method. All methods considered non-linear which do not use linear approximation e.g. logarithmic, exponential, trigonometric, etc. Non-linear transformations applied on input variables could also be part of non-linear regression. In the non-linear approximator polinom of the applied method, non-linearity comes from the higher degree of input variables. Such a regression is called polynomial regression [28]. Difficulty of this method is to specify the minimum sufficient degree of input variables and their multiplicative combinations to keep the model as simple as possible.

However, when looking back to the definition of linear regression, please note that there were no restrictions to the dependency of the input variables. That is when the solution of polynomial regression above is splitted into specifying degree of descriptors and their coefficients while on the other hand the exponential growth of model complexity together with the expansion of computation time of the analytical solution are accepted, then after adding second and third degree of input variables (and their multiplicative combinations up to 3rd degree), the solution is simplified to the previously explained multivariate linear regression model. Total numbers of input variables (including engineering parameters, building configuration descriptors and their multiplicative combinations) are listed in Table 3. As a result of the modifications, the new model is now able to approximate non-linear functions.

**Table 3.** Total number of input variables

| Input type | A/S | Simple<br>descriptors |
|---|---|---|
| engineering | 3 | 3 |
| building configuration | 1 | 14 |
| total 1st degree | 4 | 17 |
| total 2nd degree | 14 | 170 |
| total 3rd degree | 34 | 1139 |

## 8.1 Approximation accuracy

When accuracy of approximations is measured by $R^2$ score, shown in Eq. (9), Table 4 contains model performances. Upon these, it can be stated that extending input features with their maximum 3rd powers results the best approximation, but application of max. 2nd powers also results huge improvement in accuracy. Applying the set of simple building geometry descriptors besides engineering inputs performs slightly better ($R^2$=0.98) than the single, complex descriptor ($R^2$=0.95).

**Table 4.** $R^2$ scores by different degree of inputs

| max.<br>degree | A/S | | 14 descriptors | |
|---|---|---|---|---|
| | **Training** | **Test** | **Training** | **Test** |
| 1st | 0,7240 | 0,7363 | 0,7525 | 0,7365 |
| 2nd | 0,9311 | 0,9329 | 0,9623 | 0,9632 |
| 3rd | 0,95244 | 0,9567 | 0,9899 | 0,9884 |

Rate of approximation accuracies, measured on test data, which is unknown for the model, is presented in Figure 5. Test

data is unknown for the model because it was not part of model creation or training.
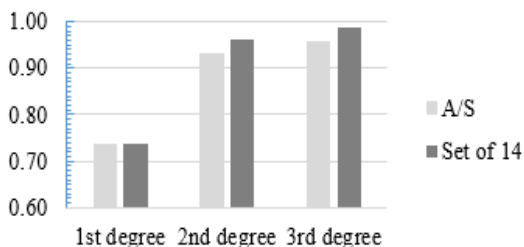


**Figure 5.** $R^2$ scores by different degree of inputs

Examining error function of the extended linear regression models applying higher degree of input variables, it can be stated that distribution of the absolute error arising from approximation is not, but the distribution of relative error can be considered as normal distribution. Therefore, the mean and standard deviation and histogram can be examined and used for confidence estimation similarly to the situation discussed earlier.

**Table 5.** Standard deviation of approximations [%]

| Input degree | A/S | 14 descriptors |
|---|---|---|
| max. 1st power | 7,9 | 8,0 |
| max. 2nd power | 3,9 | 3,1 |
| max. 3rd power | 3,3 | 1,6 |

Standard deviations of relative approximation errors measured on test (unknown for model) data is listed in Table 5. Differences and rates of standard deviations of relative approximation errors measured on test data is depicted in Figure 6.
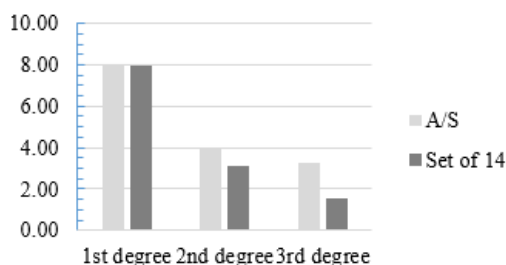


**Figure 6.** Relative error $\sigma^2$ [%]

Based on the above mentioned, it is visible that when testing a multivariate linear regression model which applies max. 3rd powers of input variables 0.9884 R2 score was reached, while standard deviation relative error of the approximation was 1.6%. Therefore, based on 3 σ rule of normal distribution, most probably the model will approximate 68% of the results with less than 1.6% relative error and 99% of the results will have less than 4.8% relative error.

This is further illustrated in Figure 7, where relative errors of approximations made by linear regression models with different degree of input variables on test (unknown for the model) data are collected into 4 bins. The bins are declared as the absolute value of the distance from the mean in a step of standard deviance. When describing numerically, the selected model estimates the 76.5% of the data with maximum 1 σ relative error and the rate of estimations with larger than 3 σ is less than 0.9%.
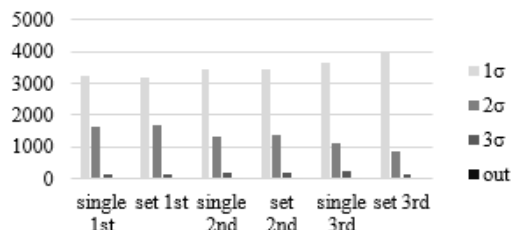


**Figure 7.** Histogram of absolute errors by reliability

The histogram shown in Figure 5 in accordance with executed normality tests fits the histogram of absolute value of the Gaussian bell curve. Rate of estimations out of 3 σ interval from the mean is less than 1% which is slightly higher than the rate declared by 3 σ rule (0.3%). Examination of these points are detailed in the next chapter.

### 8.2 Approximation error

Approximately 0.3% of estimation result error values farther than 3 σ from its average can be accepted in general, according to the behavior of normal distribution. It is visible in Table 6 and Figure 8 that when using 1st power of input variables, the count and rate of outsider points was 0.

**Table 6.** Number and rate of estimations out of 3σ

| Input degree | Out of 3σ | A/S | 14 descriptors |
|---|---|---|---|
| 1st | count | 0 | 0 |
|  | % | 0.0 | 0.0 |
| 2nd | count | 25 | 13 |
|  | % | 0.5 | 0.3 |
| 3rd | count | 37 | 49 |
|  | % | 0.7 | 1.0 |

But when using 2nd and 3rd power of the input variables the rate of such data points was almost 0.5% and 0.7%, which is almost two times more than expected.
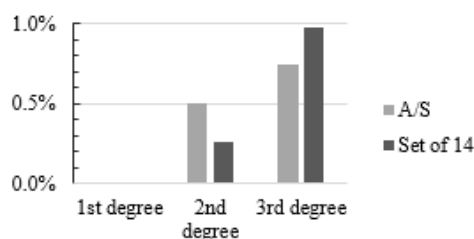


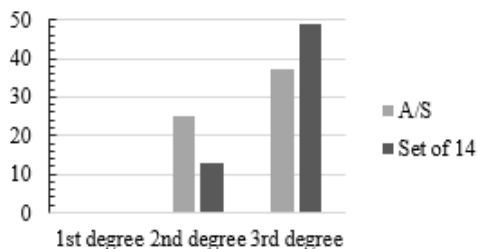**Figure 8.** Rate of estimations out of 3σ [%]



**Figure 9.** Number of estimations out of 3σ [pieces]

Increasing rate of similar outsider points may look wrong, but it is indeed good. The reason is that the rate is closely related to the standard deviation of the normal distribution.

When general accuracy is increasing, that will decrease the estimation error interval through the decrement of σ of normal distribution. This could bring a slight increase of the number (and rate) of estimations out of 3σ. Now it is understandable that 0 outsider points of the first model was because of unacceptably large σ. The increment of number and rate of outsider points does not decrement the acceptability of the estimation model. Nevertheless, analysis of such points as anomaly points and trying to identify the reasons of anomaly by architects and building engineer experts could significantly increase the reliability of the process.

It can be seen in Figure 9 that the number of points to be examined as anomaly point is around 40 for the whole experiment. Thus, the resources required for further anomaly analysis can easily be provided.

## 9. CONCLUSION

Based on analysis and interpretation of experiment results, it can be stated that applying a multivariate linear regression model using at most $3^{rd}$ degree of 14 simple building configuration descriptors besides engineering inputs can be used to estimate annual heating energy demand of buildings in conditions predefined by engineering parameters. The estimation accuracy measured by $R^2$ score acquired 0.9884. The relative error of expected output estimations had 0% average and 1.6% standard deviation. Therefor expected value of relative estimation error for more than 99% of the sample space is under 5%. As a consequence, in the predefined environment, the presented multivariate linear regression model can replace the simulations.

In other words, the linear regression model containing only the input variables in their original form is not suitable for replacing building energy requirement simulations. It was necessary to increase the model complexity up to the $3^{rd}$ degree to achieve applicable estimation results. Even though the complexity of the model became greater and thus the number of computational steps increased exponentially, it is worth mentioning, that the total computational time did not change significantly in the case of the experiment.

However, it is important to state that the experiment of approximation was made in the search space stretched by well known, discrete values of explanatory variables. Extension of this search space with new discrete values along existing dimensions and adding new dimensions or transferring values into the continuous domain requires modification in the model accordingly.

Furthermore, the approximated simulations use local weather statistics, therefore their approximations are also location dependent.

Although the proposed model can estimate heating energy demand accurately only in restricted manners, it is still useful as performance measure in energy demand comparison of building configurations.

## 10. FUTURE WORKS

Based on conclusions above, a neural network model is under construction, to extract required features and avoid using higher power of inputs. Using point cloud representation of the search space is also planned to make configuration size, structure and functionality flexible. These changes could result

a model suitable to compare a wide range of building configurations, letting to select the energy efficient.

## REFERENCES

[1] Cao, X., Dai, X., Liu, J. (2016). Building energy-consumption status worldwide and the state-of-the-art technologies for zero-energy buildings during the past decade. Energy and buildings, 128: 198-213. https://doi.org/10.1016/j.enbuild.2016.06.089

[2] IEA, UNEP. (2019). Global Status Report for Buildings and Construction, vol. 224, 2019.

[3] Menezo, C., Lepers, S., Depecker, P., Virgone, J. (2001). Design of buildings shape and energetic consumption. Build Environ, 36(5): 627-635.

[4] AlAnzi, A., Seo, D., Krarti, M. (2009). Impact of building shape on thermal performance of office buildings in Kuwait. Energy Conversion and Management, 50(3): 822-828. https://doi.org/10.1016/j.enconman.2008.09.033

[5] Heiselberg, P., Brohus, H., Hesselholt, A., Rasmussen, H., Seinre, E., Thomas, S. (2009). Application of sensitivity analysis in design of sustainable buildings. Renewable Energy, 34(9): 2030-2036. https://doi.org/10.1016/j.renene.2009.02.016

[6] Saltelli, A., Tarantola, S., Campolongo, F., Ratto, M. (2004). Sensitivity analysis in practice: A guide to assessing scientific models. Chichester, England.

[7] Moret, S., Noro, M., Papamichael, K. (2013). Daylight harvesting: a multivariate regression linear model for predicting the impact on lighting, cooling and heating. In Proceedings of the 1st IBPSA Italy Conference, Bolzano (Italy).

[8] Lee, J., Boubekri, M., Liang, F. (2019). Impact of building design parameters on daylighting metrics using an analysis, prediction, and optimization approach based on statistical learning technique. Sustainability, 11(5): 1474. https://doi.org/10.3390/su11051474

[9] Kwon, N., Ahn, Y., Son, B.S., Moon, H. (2021). Developing a machine learning-based building repair time estimation model considering weight assigning methods. Journal of Building Engineering, 43: 102627. https://doi.org/10.1016/j.jobe.2021.102627

[10] Sajjad, M., Khan, S. U., Khan, N., Haq, I.U., Ullah, A., Lee, M.Y., Baik, S.W. (2020). Towards efficient building designing: Heating and cooling load prediction via multi-output model. Sensors, 20(22): 6419. https://doi.org/10.3390/s20226419

[11] Aghdaei, N., Kokogiannakis, G., Daly, D., McCarthy, T. (2017). Linear regression models for prediction of annual heating and cooling demand in representative Australian residential dwellings. Energy Procedia, 121: 79-86. https://doi.org/10.1016/j.egypro.2017.07.482

[12] Lam, J.C., Wan, K.K., Liu, D., Tsang, C.L. (2010). Multiple regression models for energy use in air-conditioned office buildings in different climates. Energy Conversion and Management, 51(12): 2692-2697.

https://doi.org/10.1016/j.enconman.2010.06.004

[13] Danza, L., Belussi, L., Salamone, F. (2020). A multiple linear regression approach to correlate the Indoor Environmental Factors to the global comfort in a Zero-Energy building. In E3S Web of Conferences, 197: 04002. https://doi.org/10.1051/e3sconf/202019704002

[14] Yigit, S. (2021). A machine-learning-based method for thermal design optimization of residential buildings in highly urbanized areas of Turkey. Journal of Building Engineering, 38: 102225. https://doi.org/10.1016/j.jobe.2021.102225

[15] Kudabayev, R., Suleimenov, U., Ristavletov, R., Kasimov, I., Kambarov, M., Zhangabay, N., Abshenov, K. (2022). Modeling the thermal regime of a room in a building with a thermal energy storage envelope. Mathematical Modelling of Engineering Problems, 9(2): 351-358. https://doi.org/10.18280/mmep.090208

[16] Al-Tajer, A.M., Basem, A., Khalaf, A.F., Jasim, A.K., Hammoodi, K.A., Hussein, H.Q. (2022). A numerical simulation to select the optimal thermal agents for building parts. Mathematical Modelling of Engineering Problems, 9(5): 1393-1398. https://doi.org/10.18280/mmep.090530

[17] Magalhães, S.M., Leal, V.M., Horta, I.M. (2017). Modelling the relationship between heating energy use and indoor temperatures in residential buildings through Artificial Neural Networks considering occupant behavior. Energy and Buildings, 151: 332-343. https://doi.org/10.1016/j.enbuild.2017.06.076

[18] Zou, Y., Zhan, Q., Xiang, K. (2021). A comprehensive method for optimizing the design of a regular architectural space to improve building performance. Energy Reports, 7: 981-996. https://doi.org/10.1016/j.egyr.2021.01.097

[19] Harish, V.S.K.V., Kumar, A. (2016). A review on modeling and simulation of building energy systems.

Renewable and sustainable energy reviews, 56: 1272-1292. https://doi.org/10.1016/j.rser.2015.12.040

[20] Sadoughi, A., Kibert, C., Sadeghi, F.M., Jafari, S. (2019). Thermal performance analysis of a traditional passive cooling system in Dezful, Iran. Tunnelling and Underground Space Technology, 83: 291-302. https://doi.org/10.1016/j.tust.2018.09.024

[21] Ochoa, C.E., Capeluto, I.G. (2008). Strategic decision-making for intelligent buildings: Comparative impact of passive design strategies and active features in a hot climate. Building and Environment, 43(11): 1829-1839. https://doi.org/10.1016/j.buildenv.2007.10.018

[22] Mitchell, T., Keller, R., Kedar-Cabelli, S. (1986). Explanation-based generalization: A unifying approach. Machine Learning. https://doi.org/10.1007/978-1-4613-2279-5

[23] Long, J.S. (1997). Regression Models for Categorical and Limited Dependent Variables.

[24] Heiberger, R. M., Holland, B. (2015). Linear Regression by Least Squares. In Statistical Analysis and Data Display. Springer, New York, NY, pp. 235-262.

[25] Cameron, A.C., Windmeijer, F.A. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. Journal of Econometrics, 77(2): 329-342. https://doi.org/10.1016/s0304-4076(96)01818-0

[26] D'agostino, R.A.L.P.H., Pearson, E.S. (1973). Tests for departure from normality. Empirical results for the distributions of b 2 and√ b. Biometrika, 60(3): 613-622. https://doi.org/10.1093/biomet/60.3.613

[27] Stephens, M.A. (1974). EDF statistics for goodness of fit and some comparisons. Journal of the American Statistical Association, 69(347): 730-737. https://doi.org/10.1080/01621459.1974.10480196

[28] Ostertagová, E. (2012). Modelling using polynomial regression. Procedia Engineering, 48: 500-506. https://doi.org/10.1016/j.proeng.2012.09.545