Data verification for forecasting of building energy consumption

Muzaffar Khudayarov[1*], *and Sarvar* Qurbonov[2]

[1]Department of Power Plants, Networks and Systems, Tashkent State Technical University named after Islam Karimov, 2 University Street, 100095 Tashkent, Uzbekistan
[2]JV LLC "SKB", 5 Majnuntol Street, Yashnabad district, 100204 Tashkent, Uzbekistan

Abstract. One of the necessary procedures prior to monitoring the energy consumption of buildings is the procedure of verification of the initial data on energy consumption. It is always applied in cases where there is even the slightest doubt in the correctness of the initial data. The article deals with the verification of data on energy consumption in buildings, which includes the implementation of procedures for detecting and replacing null, erroneous, absolutely equal, and recovery of lost data on energy consumption. To solve this problem the article presents a number of methods included in the program "Statistical data verification" developed in MATLAB environment. As an example, the process of verification of data on energy consumption in buildings, which include hospitals, polyclinics, rural health centres is presented.

1 Introduction

One of the necessary procedures in the management of energy consumption of buildings [1,2] is the procedure of verification of the collected database on energy consumption. Verification should always be applied if there is even the slightest doubt in the correctness of the initial data.
Verification is a logical-methodological procedure for detecting and replacing null, erroneous, absolutely equal data, as well as recovering lost data.
By no means always the initial data turn out to be quite correct. This greatly reduces the reliability of such procedures of buildings energy management as balance modelling, forecasting [3,4] and rationing [5].
The preliminary analysis of the collected data on energy consumption of more than 4000 buildings (hospitals, polyclinics, rural health centres) has shown, that not always the initial data used for calculations, are quite correct. They contain zero or absolutely equal data, as well as the so-called "outliers", which is the reason of incorrectness of the base. In some cases, it is necessary to increase the existing database for several years "backwards". Consequently, a preliminary verification of the energy database is required.
The main purpose of this work is to present methods of verification of data on energy consumption in buildings, as well as the use of, developed in MATLAB program " Statistical data verification" to solve this problem.

2 Methods

The main purpose of data verification is to correct the database obtained from the collection of information on the energy consumption of social facilities. Data verification involves performing the following procedures:

1) Zeros (null data) elimination;
2) Outliers (erroneous data) elimination;
3) Absolutely equal data elimination;
4) Recovery lost data.

2.1. Zeros elimination

Zero data is the first indication that the base is incorrect. The fact is that a really operated object cannot consume zero amounts of energy in any way. The presence of them in the database can only be due to their absence during the collection of information on individual objects.
In order to eliminate the zero data, they are first searched for, and then, if they are present within the database, they are restored by interpolation, and at the border of the database - by extrapolation. Interpolation and extrapolation are

*Corresponding author : muzaffar_hb@mail.ru

performed in 5 ways: 1) splines; 2) 3-degree polynomials; 3) cascade forward neural network with 10 neurons in the hidden layer; 4) feed forward neural network with 8 neurons in the hidden layer; 5) mean of the four methods.
Elimination of null data can be performed on both annual and monthly data.

2.2. Outliers elimination

Erroneous data (outliers) are atypical data that deviate significantly (without a correct explanation) from the centre of the distribution. Outliers appear due to gross errors in logging measurements, equipment failures, measuring in wrong units, etc. Also, a very common cause of outliers is an error in data entry into the computer.
To check and determine outliers, we used six statistical methods:
1) **Z-score method**. According to this method, Z-scores are defined as:

$$Z_{sc}(i) = \frac{x_i - x_{mean}}{x_{std}}, where\ x_{mean} = \sum_{i=1}^{n} x_i/n,\ x_{std} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - x_{mean})^2} \quad (1)$$

Data with $|Z_{sc}|$ greater than 3 are outliers [6], although the criteria may vary depending on the data set.
2) **Modified Z-score method**. On this method we replace x_{mean} by the sample median x_{median}, and s by the MAD (Median of Absolute Deviations about the median). According to this method, modified Z-scores are defined as:

$$modZ_{sc}(i) = \frac{0.6745(x_i - x_{median})}{median\{|x_i - x_{median})|\}} \quad (2)$$

If $|modZ_{sc}(i)| > 3.5$, the data are considered outliers [7].
3) **Interquartile range (IQR) method.** On this method IQR is determined by the formula $IQR = Q_3 - Q_1$, where Q1 is the first quartile and Q3 is the third quartile.
The IQR bounds lie at a distance $Q_1 - k \cdot IQR$ and $Q_3 + k \cdot IQR$.
According to [8], the probability that one or more observations will be considered outliers at k ~ 2. The widely used value of k is 1.5. Observations outside the IQR bounds are considered as outliers.
4) **Adjusted IQR method**. The idea of the method is to calculate the limits of the "confidence interval" taking into account the asymmetry of the distribution, but so that for the symmetrical case it is still equal to 1.5 * IQR [9]. To determine some "asymmetry coefficient" we used the medcouple function (MC), which is defined as follows:

$$MC = \underset{x_i \le Q_2 \le x_j}{\text{med}}\ h(x_i, x_j) \text{ where } h(x_i, x_j) = \frac{(x_j - Q_2) - (Q_2 - x_i)}{x_j - x_i} \quad (3)$$

The adjusted IQR bounds lie at a distance [9]:

$$\text{for } MC \ge 0, \quad [Q_1 - k \cdot e^{-4MC} IQR; Q_3 + k \cdot e^{3MC} IQR\,] \quad (4)$$

$$\text{for } MC < 0, \quad [Q_1 - k \cdot e^{-3MC} IQR; Q_3 + k \cdot e^{4MC} IQR\,] \quad (5)$$

The value of k is 1.5. Observations outside the adjusted IQR bounds are considered as outliers.
5) **Grubbs test.** This test is known as the maximum normed residual test [10]. The Grubbs' test statistic is defined as:

$$G(i) = \frac{max|x_i - x_{mean}|}{x_{std}} \quad (6)$$

where $x_{mean}\ and\ x_{std}$ denoting the sample mean and standard deviation, respectively. The procedure for their calculation is presented in 1.
For the two-sided test, the hypothesis of no outliers is rejected if

$$G = \frac{(n-1)}{\sqrt{n}} \sqrt{\frac{t^2_{\alpha/2n,n-2}}{n-2+t^2_{\alpha/2n,n-2}}} \quad (7)$$

where $t^2_{\alpha/2n,n-2}$ - critical value *t* distribution with *(n-2)* degrees of freedom and a significance level of *α/(2n)*.
6) **Dixon test.** Dixons Q test identifies whether the smallest or largest data value is an outlier or not. This test is used for samples of small size (n≤ 30). Find the Q statistic using the following formula:

$$Q = \frac{x_n - x_{n-1}}{x_n - x_1} \quad (8)$$

where x_1 – value of the smallest observation, x_n – value of the largest observation, x_{n-1} – value of the second largest observation.
Then we find the Q_{crit} value. If $Q > Q_{crit}$ we reject the corresponding data point as an outlier.
Data are considered as outliers if any five of the six methods are met.
Further outliers' elimination is performed similarly to the zero's elimination, i.e. on the basis of interpolation and extrapolation 5 methods presented above.

2.3. Absolutely equal data elimination

The presence of absolutely equal data is also one of the signs of incorrectness of the base, which is obvious even in terms of physical sense. The fact is that it is very difficult to find two different objects, remote from each other, with different modes of operation, but with absolutely equal energy consumption. To eliminate absolutely equal data, first of all their search is performed. If available, they are eliminated by adding relatively small values (Δ=0.1) to the original data.

2.4. Recovery lost data

Recovering lost data is optional among these four procedures. It should be performed in case of missing data on individual objects.

Recovery of the lost part of the database can be done by extrapolation for each object. This is essentially a forecasting task, the other way around, where present data is used to determine past data. The data are extrapolated in the 5 ways given above. At the same time, you should specify the number of years for which you want to restore the database.

3 Results

The implementation of data verification procedures using the program " Statistical data verification", consider the example of electricity consumption of 150 buildings (hospitals, clinics, rural health centres) located on the territory of Tashkent region.

3.1. Zeros elimination

As noted above, the first step is to check for null data in the source database. The check showed the presence of 28 null data on five buildings. For example, there are 2 null values in the data of building No.6 (Fig.1). The null data replacement is performed by a forward propagation neural network with 10 neurons in the hidden layer.

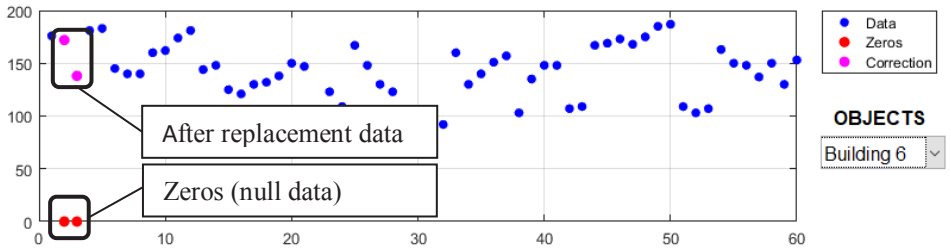

Fig. 1. Search and replacement of zeros (null data)

3.2. Outliers elimination

The results of the checking identified 68 erroneous data in 61 objects. For example, there are 3 outliers in the data of building No.9 (Fig.2). Outliers replacement is performed by a forward propagation neural network with 10 neurons in the hidden layer.

Fig. 2. Search and replacement of outliers (erroneous data)

3.3. Absolutely equal data elimination

The results of the checking identified 19 absolutely equal data (Fig. 3). They are eliminated by adding relatively small values (Δ=0.1).

All absolutely equal data - 19

	1	2	3	4	5	6	7
2018 year	2	55	91	0	0	0	0
3 equals	15	101	93	0	0	0	0
2019 year	8	18	57	147	0	0	0
4 equals	91	76	58	149	0	0	0
2020 year	73	147	0	0	0	0	0
2 equals	75	149	0	0	0	0	0
2021 year	26	63	63	64	67	73	147
7 equals	34	67	69	115	69	75	149
2022 year	8	73	147	0	0	0	0
3 equals	108	75	149	0	0	0	0

Fig. 3. Search and replacement of absolutely equal data

3.4. Recovery lost data

In the database, restore should be performed on building 4 (Fig.4). The data for the first year should be restored. The distribution of the annual value by months is performed randomly.

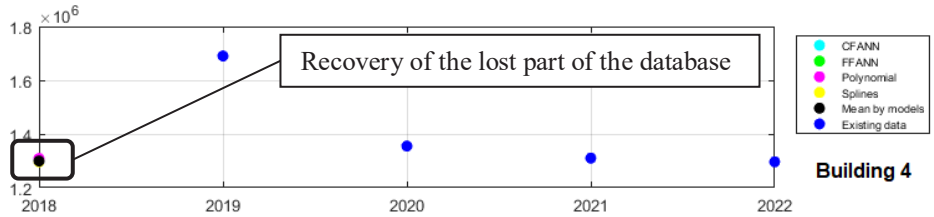

Fig. 4. Recovering first year data for building № 4

4 Discussion

Next, we evaluate the impact of verification procedures on the quality of forecasting models based on feedforward artificial neural networks. In this work, the following kind of model is used for energy consumption forecasting:

$$W_t = f(W_{t-2}, W_{t-1}, N_t) \tag{9}$$

where W_t is the energy consumption of the building at time t; W_{t-2}, W_{t-1} are the previous values of energy consumption of the building; N_t is the month of the year (1, 2, ..., 12);

For forecasting we use a model with one hidden layer consisting of 8 neurons. The hyperbolic tangent is used as the activation function. Training of neural networks is performed by Levenberg-Marquardt method.

The input data for prediction is 60 "inputs-output" data pairs (over 5 years). The data is divided into a training sample (70%), and a test sample (30%). The training sample is used to train the model, and the test sample is used to check the quality of modeling. Let us compare four types of models derived from the sample:

Type 1 - Including zero data, absolutely equal data, and outliers;

Type 2 - where zero data are eliminated, but have absolutely equal data and outliers;

Type 3 - where zero data and outliers are eliminated but have absolutely equal data;

Type 4 - where zero, absolutely equal data and outliers are eliminated.

The criterion for the quality of models is the coefficient of determination, R^2:

$$R^2 = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - Y_{mod,i}\right)^2}{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \frac{\sum_{i=1}^{N} Y_i}{N}\right)^2} \tag{10}$$

where, Y_i is the actual value; $Y_{mod,i}$ is the value obtained from the ANN model.

The results of assessments of the impact of verification procedures on the quality of energy prediction models for randomly selected six buildings are presented in Table 1.

Table 1. The results of assessments of the impact of verification procedures

Buildings	Model type				Model Improvement		
	Type 1	Type 2	Type 3	Type 4	T2/T1	T3/T1	T4/T1
Building 1	0,01473	0,15334	0,5599	0,54866	10,4	38,0	37,25
Building 30	0,03364	0,18374	0,5984	0,58734	5,5	17,8	17,46
Building 90	0,05317	0,11366	0,62362	0,60347	2,1	11,7	11,35
Building 110	0,03256	0,17433	0,64752	0,61734	5,4	19,9	18,96
Building 134	0,04961	0,18658	0,69623	0,66553	3,8	14,0	13,42
Building 146	0,00408	0,11297	0,27727	0,27579	27,7	68,0	67,60
Average					9,1	28,2	27,67

5. Conclusions

It is not always the initial data that turn out to be quite correct. They contain zero, erroneous (outliers), absolutely equal data, which greatly reduces the quality of the models used to predict the energy consumption of buildings. In some cases, it is necessary to restore the lost data. Consequently, prior verification of the existing energy consumption database of buildings is required before performing forecasting.

The use of verification improves the quality of the models for the energy forecasting of buildings. So, replacing only zero data increased the quality of the models on average 9 times, replacing zero data and emissions increased the quality of the models on average 28 times, replacing zero data, emissions and absolutely equal data increased the quality of the models on average 27.7 times.

As the results of calculations have shown, the key procedures of data verification are the search and replacement of zeros and outliers. The procedure of search and replacement of absolutely equal data does not greatly affect the quality of modeling, and in our case it even slightly worsened the quality of models.

References

1. V.I. Gnatiuk, The Law of Optimal Construction of Technocenoses, Joint Publishing of TSU and the Center for Systems Research, Moscow (2005)
2. T.P. Salikhov, M.B. Khudayarov, Technique of control of power consumption in Buildings of Social Sites, *Energy Security and Energy Saving* **3**, 16-21 (2015)
3. T.P. Salihov, M.B. Khudayarov, Forecasting of energy resources consumption by social purpose objects, *Problems of Energy and Resource Savings* 161-167 (2014)
4. M.B. Khudayarov, A.T. Khabibulina, H.Kh. Karimkulov, Energy consumption forecasting methodology of a set of objects, *European Science Review* **11-12**, 240-242 (2016)
5. T.P. Salikhov, M.B. Khudayarov, Rationing of energy consumption of social facilities on the basis of cluster analysis, *Problems of Information and Energy* **5**, 71-76 (2014)
6. R.E. Shiffler, Maximum Z Scores and Outliers, *The American Statistician* **42**(1), 79–80 (1988)
7. B. Iglewicz, D.C. Hoaglin, How to Detect and Handle Outliers, ASQC Basic References in Quality Control, Wisconsin (1993)
8. M. Frigge, D.C. Hoaglin, B. Iglewicz Some Implementations of the Boxplot, *The American Statistician* **43**(1), 50–54 (1989)
9. M. Hubert, E. Vandervieren, An adjusted boxplot for skewed distributions, *Computational Statistics and Data Analysis* **52**(12), 5186 – 5201 (2008)
10. G. Brys, M. Hubert, P.J. Rousseeuw A Robustification of Independent Component Analysis, *Journal of Chemometrics* **19**, 364–375 (2005)