Regression Analysis-Based Load Modelling for Electric Distribution Networks

Gheorghe Grigoras and Bogdan Constantin Neagu

Abstract The decision making in the electric distribution systems is based on data collected from consumers and the various measurement points located in the network (transformer substations, supply points, branch points, etc.). The information obtained from a 100% integration of the smart metering to consumers comes to fill the data acquired through the Supervisory Control and Data Acquisition (SCADA) system, so that the Distribution Network Operator (DNO) can accurately estimate the state of the supervised system. However, the implementation of smart metering is in various implementation stages in different countries of the world, so that today it can be stated that there is no complete integration. The same aspect should be emphasized in the case of the SCADA system at the level of distribution networks, which is not 100% integrated in the low/medium voltage electric substations. In these conditions, the DNO should apply the mathematical tools that take into account the similarities between the consumers' behaviour and respectively the structure of the load supplied from the electric substations. The regression analysis-based approaches for the load modelling from the nodes of electricity distribution networks were treated in the chapter. The approaches refer to the estimation of the required powers in the supply points with a mixt structure of the load (i.e. residential, commercial, and industrial) at the hour when the maximum value of the load is recorded and the demand of residential consumers which represent the highest percentage from the load structure fed from the electric substations. The proposed approaches were tested in real operation conditions of the distribution networks from Romania.

Keywords Regression · Correlation · Peak load · Distribution networks · Residential consumers

G. Grigoras (✉) · B. C. Neagu
Electrical Engineering Faculty, Power System Department, Gheorghe Asachi Technical University of Iasi, Iasi, Romania
e-mail: ggrigor@tuiasi.ro

B. C. Neagu
e-mail: bogdan.neagu@tuiasi.ro

N. Mahdavi Tabatabaei and N. Bizon (eds.), *Numerical Methods for Energy Applications*, Power Systems,
https://doi.org/10.1007/978-3-030-62191-9_28

Abbreviation and Acronyms

SCADA	Supervisory Control and Data Acquisition
DNO	Distribution Network Operator
LV	Low Voltage
MV	Medium Voltage
HV	High Voltage
LR	Linear Regression
PR	Parabolic Regression
HR	Hyperbolic Regression
ER	Exponential Regression

1 Introduction

The loads from the nodes of electricity distribution systems (represented by the Medium Voltage/Low Voltage (MV/LV) electric substations) vary in time and have particular characteristics in each consumption point. Therefore, to solve the problems regarding the optimal network planning and operation, the demand management and the correct billing of consumers, the Distribution Network Operators (DNOs) need to know the dynamic behaviour of the loads in their networks [1–3]. On the other hand, the load variations are influenced by several factors, such as consumer type, time factor, climatic factors, other electrical loads correlated with the analysed load, historical values, and consumption profile [4–6].

The modelling of electricity consumptions is made using the records from the databases which describe the evolution of individual and aggregated loads. These data are recorded and processed systematically using appropriate methods. The following input information is frequently used in an analysis: the daily maximum value of load, the hourly power consumption, the daily/weekly electricity amount. For a better accuracy in the modelling process, a large database should be used, including the electricity consumptions for a long-time interval and, if possible, the evolution of demographic, climatic and economic activity indices for the geographical area and time interval of interest [5, 6].

Also, there are some restrictions which the Decision Makers must consider them in their analyses [7, 8]:

- The power flows must satisfy the fundamental laws of electrotechnics (Kirchhoff laws);
- The balance between the obtained loads in the estimation process and the measured values.
- The load does not depend by the structure of network.

The randomly selected working sample from the database must be subjected to a detailed analysis to identity the outliers, then following the correlation process to find

the relationships between the variables represented by the power/energy consumption and the climatic and weather factors [9, 10].

In the chapter, various approaches for the load modelling from the nodes of electric distribution networks, based on the correlation and regression analysis will be proposed. The support of the proposed approaches is represented by the processing process of the load profiles belonging to the MV/LV electric substations or LV consumers recorded with the help of smart meters using the statistical tools. The structure of chapter is divided in two parts: a short review about the correlation and regression analysis is made in the first part, and in the second part the regression analysis based-approaches are presented regarding the estimation of the powers in the MV/LV electric substations (at the hour when the maximum value of the load from the system is recorded) and the demands of the residential consumers.

2 Correlation and Regression Analysis

To understand the operation of electric distribution systems, it is necessary to be studied the relationships between the state variables that characterize them (voltages, currents, powers, etc.). For these variables, the relationships can be analysed using the regression and correlation methods.

The regression methods allow the measurement and study of the relation between two or more variables, as well as the discovery of the connection laws between these. A mathematical expression can be obtained with the aim to estimate the values of one independent variable according to the values of other variables [11, 12].

Correlation analysis measures the intensity of the relationship between one or more variables. Depending on the regression model, the correlation can be treated as a single or multiple correlation [13, 14].

The following issues must be solved in a study which is based on the regression and correlation analysis [12]:

- Identify the existence of the relationship between variables. Solution: A logical analysis of the possibility of a relationship between the variables can be applied.
- Establishing the meaning and form of the relationship. Solution: Regression analysis methods can be used.
- Determining the intensity degree of relationship. Correlation analysis methods can be used.

2.1 Correlation Methods

2.1.1 Interdependent Parallel Statistical Series-Based Method

The analysis of statistical relationships takes into account the estimation of a regression model and measuring the intensity of the relationship between variables. The

Table 1 Cross-correlation matrix

x	y	
	$y_1 \ldots y_j \ldots y_p$	$\sum_j n_{ij} = n_i*$
x_1 … x_i … x_k	$n_{11} \ldots n_{1j} \ldots n_{1p}$ … $n_{i1} \ldots n_{ij} \ldots n_{ip}$ … $n_{k1} \ldots n_{kj} \ldots n_{kp}$	n_{1*} … n_{i*} … n_{i*}
$\sum_i n_{ij} = n_{*j}$	$n_{*1} \ldots . n_{*j} \ldots . n_{*p}$	$n** = \sum_i \sum_j n_{ij}$

analysis of the statistical relationship compares the terms of two interdependent parallel series x (independent variable) and y (dependent variable). For example, when two time series are compared, their elements are chronologically sorted, such that the existence and direction of the relationship can be easy identified. Thus, if both variables have a variation in the same direction, there is a direct relationship. If the variation is different, an inverse correlation is obtained. If the two time series vary independently, or one varies and the other remains constant, there is no relationship [8].

The method can be used for the time series with few variables, when there is a relationship between the pairs of variables (x_i, y_i, $i = 1, \ldots, N$).

2.1.2 Cross-Correlation Matrix Based Method

The principle of method is based on the grouping the elements of a data set using simultaneously both correlated variables (x and y). Equal intervals and an identical number of groups for both variables are recommended to be used. Thus, in the matrix, the existence, direction and intensity of the relationship can be appreciated using the distribution model of frequencies $n_{ij,}$ as it can be seen in Table 1.

If the frequencies n_{ij} are scattered relatively uniformly inside the matrix, there is no relationship between the variables considered. But, if they are concentrated around the diagonals, a stronger correlation can be identified between the variables x and y.

2.1.3 Graphical Method

The method involves the graphical representation of the pairs of values corresponding to the variables in a coordinate system, such that the existence, meaning, form and intensity of the correlation can be easy identified. The graph corresponds to the case where a relationship is defined in concordance with interdependent statistical parallel series-based method.

2.1.4 Analytical Methods

The analytical models allow determination of the mathematical relations and the numerical measurement of the intensity between variables. The regression models aim to represent the distribution type of correlated variables. The regression curves indicate the correspondence between the pairs (x_i, y_i). The following steps should be performed to establish and analyse a regression model:

- Building the correlation graph.
- Establishing the theoretical regression model of the relationship (based on the correlation graph adjustment) and identification of the equation corresponding to the chosen regression model.
- Determining the coefficients of the regression equation (with the least squares method) and interpreting the regression according to their sign and value.

2.1.5 Regression Models with Two Variables

The relationship between two variables x and y can be expressed by a regression equation:

$$y_x = f(x) + e \tag{1}$$

where $f\ (x)$ represents a function which is dependent on the variable x, and e is the approximation error.

If the size of the database will grow, the approximation error e will decrease. Thus, a higher number of observations can lead at a stronger relationship. Function $f(x)$ can have different models depending by the data scatter.

Linear regression (LR) model

The LR model is most used in the practice. The relationship can be expressed using the following equation:

$$y_x = a + bx + e \tag{2}$$

The Eq. (2) can be plotted using a line. The variable e represents a random error given by:

$$e = y_i - y_{x_i}\ ; \quad i = 1, \ldots, N \tag{3}$$

where a and b are unknown coefficients, their values being determined using the least squares method.

The coefficient b from the expression (2) can have different signs which characterize the direction of the relationship between variables: "+", positive relationship; "null", no relationship, and "−", negative relationship.

The value of coefficient b shows the dependence degree between variables, namely how much the variable y increases or decreases when the variable x increases or decreases with one unit.

Parabolic regression (PR) model

In order to express this model, the second degree polynomial is usually used:

$$y_x = a + bx + cx^2 + e \tag{4}$$

where coefficients a, b and c are determined using the least squares method.

Hyperbolic regression (HR) model

$$y_x = a + \frac{b}{x} + e \tag{5}$$

Exponential regression (ER) model

In order to express this model, the following equation is used:

$$y_x = ab^x + e \tag{6}$$

For each sample, rel. (6) can be linearized by logarithm:

$$\log y_x = \log a + x \log b \tag{7}$$

2.2 Intensity of the Relationship Between Two Variables

The intensity of the relationship, if there is between two variables (x, y), indicates a concentration degree of or scattering of the values y around the regression model y_x. The intensity of the relationship can be measured based on the correlation coefficient and the correlation ratio.

2.2.1 Correlation Coefficient

The correlation coefficient is used to appreciate the intensity of relationship between the analysed variables. The calculation of this coefficient can be made using the relation:

$$\rho(x, y) = \frac{C(x, y)}{\sigma_x \cdot \sigma_y} = \frac{\sum_i (x_i - x_m)(y_i - y_m)}{n \cdot \sigma_x \cdot \sigma_y}, \quad i = 1, \ldots, N \tag{8}$$

where: $C(x, y)$—the covariance between analysed variables; x_m, y_m—the mean values of the variable; N—number of pairs of values; σ_x and σ_y—the standard deviation of variables x and y.

Between the regression coefficient b from relation (2) and the correlation coefficient, $\rho\ (x, y)$, there is the following relationship:

$$\rho = b \cdot \frac{\sigma_x}{\sigma_y} \tag{9}$$

The analysis of the relation (9) highlights that the sign of the correlation coefficient is identically with the sign of the regression coefficient, because σ_x and σ_y are positive or equal with zero. The value of the correlation coefficient is in the range $[-1, 1]$. These two extreme values represent a perfect linear relationship between the two variables ("positive" or "negative"). The missing of a relationship between the two variables can be recorded if $\rho = 0$.

2.2.2 Correlation Ratio

The correlation ratio η is defined by the relation:

$$\eta = \sqrt{\frac{\sigma_{y_x}^2}{\sigma_y^2}} \tag{10}$$

where

$$\sigma_y^2 = \frac{\sum (y_i - \bar{y})^2}{n};\ \sigma_{y_x}^2 = \frac{\sum \left(y_{x_i} - \bar{y}\right)^2}{n} \tag{11}$$

The correlation ratio have the values into the range [0, 1]. The value 1 indicates the existence of a relationship, namely the variation of the variable y depends only on by the variation of variable x.

3 Case Studies in the Electric Distribution Networks

3.1 Power Correlation Problem

The quality and efficiency of complex problem-solving process regarding the optimal operation and planning of the electric distribution networks are largely determined by the accuracy of the load estimation methods. The estimation of the power demand and the electricity consumption is made starting from the historical data on the evolution of consumption, which is recorded systematically, processed by appropriate

methods. The main factors which can be taken into account are: daily peak load, hourly electricity consumption, and daily or weekly electricity [15, 16]. In order to have the most accurate estimation, a large-size database should be used including the hourly electricity consumptions for a sufficiently long period (minimum 1 year), the evolution of demographic and climatic factors, and economic indexes in the analysed areas [4–6]. These information must be subjected to a pre-processing stage to eliminate systematic, gross, and random errors, and then if it possible to find a relationship between variables represented by the electricity consumption and the climatic and weather factors [7, 8, 14, 17].

The practice applications have concluded that the success of an estimation method is based on the achievement of some appropriate conditions, such as: an accurate selection of estimation period, the applied method, the confidence of the initial data, the flexibility, and taking into account the climatic and weather factors. In the load estimation process (including the peak load), there are more mathematical methods developed in the literature. The most of the proposed approaches use the dependence between the maximum value of the load (peak load) and the annually/monthly/daily electricity consumption [7, 8].

Today, the most Distribution Network Operators (DNOs) from the European countries are in full process of implementing the smart metering system in the MV/LV electric substations and at the end consumers. The problem is that this process is slow and there are enough electric substations for which DNOs do not have yet information on their loading and the peak load to estimate the operation regime of the electric network. In this case, the loads, generally, and the peak load, particularly, can be estimated based on correlation studies, as will be shown in the following [6–8, 18].

If a simple linear regression model is used for the relationship between the mean values of the variables P and Q, then the following relations can be accepted (see Fig. 1):

$$Q = \rho_{PQ} \cdot \frac{\sigma_Q}{\sigma_P} \cdot P + k_{PQ} \tag{12}$$

$$\rho_{PQ} = \frac{C_{PQ}}{\sigma_P \cdot \sigma_Q} \tag{13}$$

$$C_{PQ} = \overline{P \cdot Q} - \overline{P} \cdot \overline{Q} \tag{14}$$

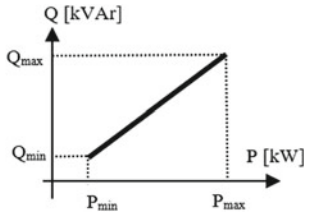

Fig. 1 The correlation between P and Q (direct variation)

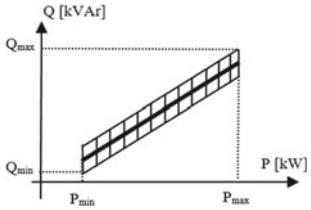

Fig. 2 The correlation between P and Q (opposite variation)

$$\sigma_P^2 = \overline{P^2} - \overline{P}^2 \tag{15}$$

$$\sigma_Q^2 = \overline{Q^2} - \overline{Q}^2 \tag{16}$$

where: P—the active power [kW], Q—the reactive power [kVAr], ρ_{QP}—the correlation coefficient between P and Q; C_{PQ}—covariance between P and Q; σ_P, σ_Q—the standard deviation of P and Q.

The overline indicates the mean value, and the coefficient k_{QP} is determined for each particular case, based on the correlation studies [7].

But, there are cases where the powers P and Q have an opposite variation. In these cases, a "variation belt" should introduced (see Fig. 2).

3.2 Peak Load Estimation Using Power Correlation

3.2.1 Solution Description

The estimation of the loads from the MV/LV electric substations at the hour when the maximum value (peak load) in the electric distribution system was recorded, will be made in this paragraph using a power correlation-based method.

In the initial step, a statistical analysis of the load profiles regarding to the active power from a database belonging a DNO in the MV/LV electric substations without the installed smart meters is performed. Different time frames can be used in this analyse, depending on the technical and load characteristics of the network. The length of the time frames (L_h with $h = 7$ or 24) could be chosen from the following: L_{24} *frame*, L_7 *frames* ($h_{PL} \pm 3\ h$), ($h_{PL} - 4\ h$; $h_{PL} + 2\ h$) and ($h_{PL} - 5\ h$; $h_{PL} + 1\ h$), where h_{PL} is the hour when the maximum value of load (peak load) from the system was recorded.

Using the LR model, the steps of the estimation method are the following:

1. Consideration of a main variable in relation to which the correlation analysis will be performed. The main variable can be chosen as the HV/MV electric substation because the hourly powers P and Q are recorded all along using smart meters.
2. Determining the peak load and the hour when is recorded for the reference electric substation.

3. Calculation of the correlation coefficients between the profiles of the powers P and Q, recorded in each MV/LV electric substation, and the profile for the power P, recorded in the HV/MV electric substation chosen as reference. Also, the standard deviation of the powers P and Q recorded in the MV/LV electric substation will be calculated.
4. Determination of the values for the coefficients $b_{P_r}^{P_i}$, $b_{P_r}^{Q_i}$, a^{P_i}, and a^{Q_1} with the relations:

$$b_{P_r}^{P_i} = \rho_{P_r P_i} \cdot \frac{\sigma_{P_i}}{\sigma_{P_r}}; \quad i = 1, \ldots, N \tag{17}$$

$$b_{P_r}^{Q_i} = \rho_{P_r Q_i} \cdot \frac{\sigma_{Q_i}}{\sigma_{P_r}}; \quad i = 1, \ldots, N \tag{18}$$

$$a^{P_i} = \sum_{j=1}^{h} (P_{ij} - b_{P_r}^{P_i} \cdot P_{rj})/L_h; \quad i = 1, \ldots, N \tag{19}$$

$$a^{Q_i} = \sum_{j=1}^{h} (Q_{ij} - b_{P_r}^{Q_i} \cdot P_{rj})/L_h; \quad i = 1, \ldots, N \tag{20}$$

where: P_r—the active power corresponding to the HV/MV electric substation chosen as reference; P_i, Q_i—the active and reactive powers from the MV/LV electric substation i; N—the number of MV/LV electric substations from the analysed network; L_h—the length of time frame ($h = 7$ or 24).

5. Estimation of the powers P and Q from the MV/LV electric substations at the hour when the maximum value of load in the system was recorded can be made using the following LR models:

$$P_i = b_{P_r}^{P_i} \cdot P_{r\max} + a^{P_i} \tag{21}$$

$$Q_i = b_{P_r}^{Q_i} \cdot P_{r\max} + a^{Q_i} \tag{22}$$

where: $P_{r\,max}$—the peak load corresponding to the reference; P_i, Q_i—the estimated powers in the MV/LV electric substation $i = 1, \ldots, N$.

3.2.2 Testing the Solution

This paragraph presents testing the proposed method based on database belonging an electric MV distribution system (20 kV) with 34 MV/LV electric substations. The

Table 2 The estimated active powers in the MV/LV electric substations

No.	P_m (kW)	Frame L_{24}		Frame L_7 ($h_{PL} - 4\ h$; $h_{PL}+ 2\ h$)		Frame L_7 ($h_{PL} - 3\ h$; $h_{PL}+ 3\ h$)		Frame L_7 ($h_{PL} - 5\ h$; $h_{PL}+ 1\ h$)	
		P_e (kW)	Er_P (%)	P_e (kW)	Er_P (%)	P_e (kW)	Er_P (%)	P_e (kW)	Er_P (%)
1	240.5	239.87	−0.205	241.24	0.31	242.91	1.00	239.18	−0.54
2	216.3	224.75	3.90	220.51	1.95	221.72	2.50	214.72	−0.73
3	311.1	346.66	11.43	321.91	3.47	325.90	4.75	315.86	1.53
4	436	397.62	−8.80	442.13	1.40	448.77	2.92	430.6	−1.23
5	410.7	381.8	−7.03	414.33	0.88	404.67	−1.46	422.27	2.81
6	420.3	422.38	0.49	411.41	−2.11	415.96	−1.03	407.03	−3.15
7	600.6	614.11	2.24	604.11	0.58	597.28	−0.55	610.54	1.65
8	617.7	609.06	−1.39	611.49	−1.00	605.77	−1.93	622.15	0.72
9	561.2	560.48	−0.12	561.27	0.01	553.07	−1.44	567.94	1.20
10	208.7	213.18	2.15	205.07	−1.73	211.44	1.31	200.76	−3.8
11	617	607.32	−1.56	612.51	−0.72	598.1	−3.06	626.26	1.50
12	588.1	562.06	−4.42	588.08	−0.00	593.15	0.86	586.35	−0.29
13	404.9	408.13	0.79	389.69	−3.75	378.81	−6.44	406.06	0.28
14	357.2	397.83	11.37	360.54	0.93	362.83	1.57	359.61	0.67
15	360.4	384.7	6.74	356.33	−1.12	354.75	−1.56	363.06	0.73

(continued)

peak load in this system is recorded at the hour 15. Following the steps of method, the LR models for different time frames were used in the analysis. The values of the active powers at the hour when the load peak was recorded in the analysed system, for the time frames L_{24} and L_7, are presented in Table 2.

The RL models obtained for all considered time frames in the case of a MV/LV electric substation (no. 28) from the analysed system are represented in Figs. 3, 4, 5 and 6 to observe the estimation accuracy for some time frame.

The errors were calculated with the relation:

$$Er_p = \frac{P_e - P_m}{P_m} 100 \quad [\%] \tag{23}$$

where: P_e—estimated active power; P_m—measured active power.

It can be observed that the errors are smaller in the case of the time frame $7\ h$ ($h_{PL} - 5\ h$; $h_{PL} + 1\ h$) than in the others frames, the average error being 1.48%.

Table 2 (continued)

No.	P_m (kW)	Frame L_{24}		Frame L_7 (h_{PL} − 4 h; h_{PL}+ 2 h)		Frame L_7 (h_{PL} − 3 h; h_{PL}+ 3 h)		Frame L_7 (h_{PL} − 5 h; h_{PL}+ 1 h)	
		P_e (kW)	Er_P (%)	P_e (kW)	Er_P (%)	P_e (kW)	Er_P (%)	P_e (kW)	Er_P (%)
16	365.6	371.42	1.59	376.21	2.90	388.43	6.24	356.33	−2.53
17	545.7	596.06	9.22	577.8	5.88	596.63	9.33	543.04	−0.48
18	254.9	257.03	0.83	258.62	1.46	257.09	0.86	260.77	2.30
19	191.3	213.61	11.66	196.52	2.73	202.63	5.92	191.78	0.25
20	168.1	196.46	16.87	175.47	4.38	183.60	9.22	162.46	−3.35
21	667.5	622.26	−6.77	633.51	−5.09	618.86	−7.28	662.08	−0.81
22	421.4	409.54	−2.81	423.75	0.55	414.28	−1.68	428.87	1.77
23	440.4	388.9	−11.69	419.91	−4.65	411.62	−6.53	436.36	−0.91
24	637	642.21	0.81	620.27	−2.62	633.43	−0.56	619.42	−2.75
25	452.3	444.02	−1.82	453.28	0.21	459.63	1.62	448.10	−0.92
26	623.7	534.84	−14.24	622.70	−0.16	615.64	−1.29	616.91	−1.08
27	402.2	392.49	−2.41	403.55	0.33	394.5	−1.91	408.35	1.53
28	671.8	677.4	0.83	682.12	1.53	686.71	2.21	679.72	1.18
29	634.1	643.68	1.51	638.69	0.72	640.00	0.93	628.87	−0.82
30	594.2	569.71	−4.12	606.74	2.11	613.92	3.32	594.83	0.10
31	388.1	437.63	12.76	392.07	1.02	386.38	-0.44	399.55	2.95
32	329.8	341.49	3.54	333.66	1.17	332.31	0.76	330.87	0.32
33	571	568.94	−0.36	551.65	−3.38	563.85	−1.25	549.30	−3.80
34	635.4	667.39	5.03	637.9	0.39	630.39	−0.78	655.06	3.09

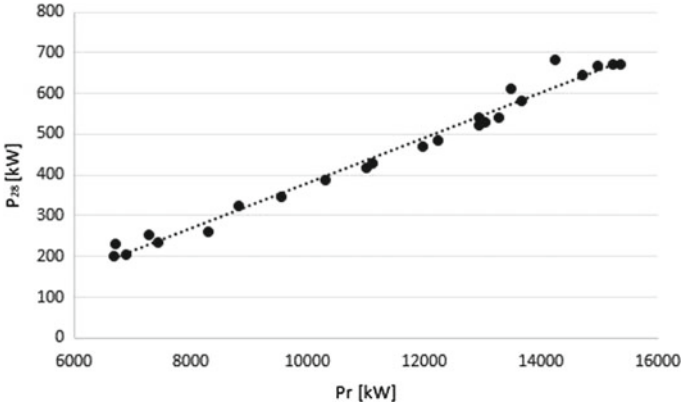

Fig. 3 Linear regression model $P_{28} = 0.0556 \cdot P_r - 175.5$ (Time Frame L_{24})

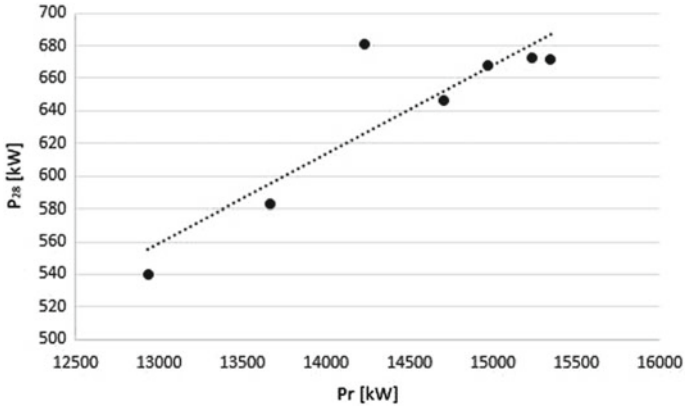

Fig. 4 Linear regression model $P_{28} = 0.0543 \cdot P_r - 146.1$ (Time Frame L_7 ($h_{PL} \pm 3\ h$)

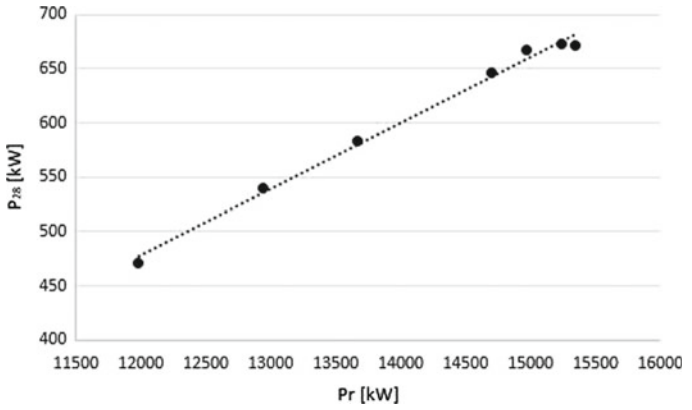

Fig. 5 Linear regression model $P_{28} = 0.0609 \cdot P_r - 251.7$ (Time Frame L_7 ($h_{PL} - 4\ h$; $h_{PL} + 2\ h$))

3.3 Residential Load Estimation Using a Regression—Correlation-Based Method

3.3.1 Solution Description

Load estimation has seen in the latest decades an increase in importance, complexity and need of accuracy. Before 1970, the electricity demand was relatively predictable, and a good forecast required simple mathematical models, limited to trend extrapolation. Also, the "7% rule" was used, which stated the doubling of electricity demand in each 10 years [19]. The load estimation studies are influenced by more factors, which can be grouped as follows [1]:

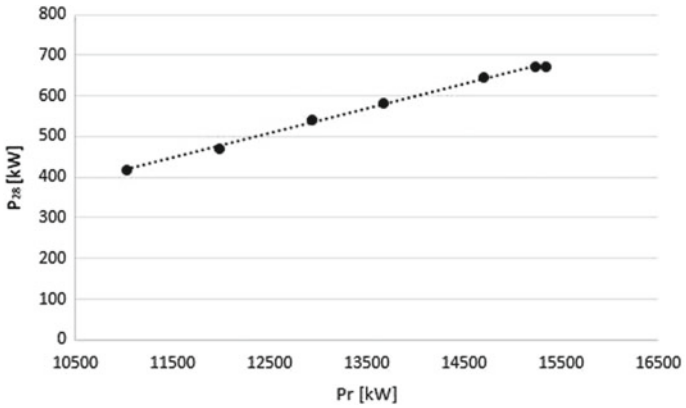

Fig. 6 Linear regression model $P_{28} = 0.0602 \cdot P_r - 244.03$ (Time Frame L_7 ($h_{PL} - 5\ h$; $h_{PL} + 1\ h$))

- Economic: for long and medium time period. These factors aren't responsible for hourly load variations and aren't considered in short term forecasts.
- Temporal—seasons, daily and weekly cycles, holidays, daylight intervals.
- Weather: temperature, humidity, wind speed and direction, clouds, rain.
- Casual: holidays, worker strikes, public events.

Practical studies have shown that the demand variation in time or according to other considered parameters has four main components [17]: season $S(t)$; cyclic $C(t)$; trend $T(t)$, and random $R(t)$. The demand can be written as the sum of the four factors, using the following equation:

$$W(t) = S(t) + C(t) + T(t) + R(t) \tag{24}$$

The mathematical function used in the estimation process is determined by successive steps, taking into account the consumption history and a qualitative and quantitative analysis of the technical and economic factors which influence in time over the consumer demand. In order to obtain a model for the demand of a consumer group or a geographical area requires the testing of several approximation approaches. For electrical load estimation, the optimal approximation functions are obtained using specialized software tools, which choose the best variant among a wide range of options.

The accuracy of the selected estimation model is assessed by computing indices which give the spread of the initial data (earlier demand values) with regard to the considered trend. Usually, a low spread indicates a good approximation which can be expressed by the quality indices (I_k). The mathematical expressions of these indices are given below:

- the mean absolute values of deviations:

$$I_1 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| \hat{y}_i - y_i \right| \tag{25}$$

- the mean absolute percentage values of deviations:

$$I_2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \cdot 100 \tag{26}$$

- the mean absolute deviation:

$$I_3 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| \hat{y}_i - \bar{y} \right| \tag{27}$$

- the dispersion:

$$I_4 = \sigma^2 = \frac{1}{n - m - 1} \cdot \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2. \tag{28}$$

but the value is different with the total variance of y:

$$\sigma_t^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 \tag{29}$$

- the mean square deviation of the selection

$$I_5 = \sigma \tag{30}$$

- the variation coefficient:

$$I_6 = v = \frac{\sigma}{x} \tag{31}$$

- the correlation coefficient from (8) in a particular form:

$$I_7 = \rho = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\pm\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{32}$$

- the particular form of (10) of correlation ratio will be:

$$I_8 = \eta = \sqrt{\frac{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{33}$$

where $\hat{y}_i$—the estimated value, y_i—the real demand, $\bar{y}$—the mean value of the historical consumption, m—the degree of the polynomial used for trend approximation.

In order to compute the trend, as recommended in the literature, continuous functions were used, which can be represented as continuous growth curves and limited growth curves. Their coefficient was determined using time series regression, with normal and modified methods using the sum squared error criterion. This approach is frequently used for residential load estimation.

The load estimation in the MV/LV electric substations is more difficult, because of the lack of historical demand data from consumers. Moreover, load estimation at the level of each DNO is possible with much better accuracy, using load data recorded through the continuous monitoring in the HV/MV electric substations and applying the global estimation methods [17].

Thus, for a year j from the estimation interval P_{m+j}, the load estimation can be obtained based on a mathematical model which uses historical load data:

$$P_{n+j} = \frac{\sum_{k=0}^{m-1} \sum_{i=1}^{n-j-k} \frac{P_{i+j+k}}{P_i}}{\sum_{k=0}^{m-1} (n - j - k)} \tag{34}$$

where: n—the previous years for which recordings exist; m—the previous years used as forecast base; j—forecast year; k—base year.

Previous studies have shown that ambient temperature has a significant influence on demand [4]. The load estimation with the temperature (computed for several consecutive years) can be:

$$P_{pr} = \frac{P_r}{1 + \frac{a}{b}\Delta\theta} \tag{35}$$

where P_r—the real load, measured in a given year; $\Delta\theta$—the difference between the real and average temperature recorded for several years, over a given time interval; a—regression coefficient with the temperature θ; b—the average load ratio for years j and $j-1$.

Accounting for the (load-temperature) correlation, which differs monthly, and sometimes is greater at the night hours than at the day hours, if temperature forecasts are known for the next year, then the load estimation for the next year can be computed:

$$P_{(n+1),\theta} = P_{n,\theta_n}(b + a \cdot \Delta\theta) \tag{36}$$

where $\Delta\theta$—the difference between the next year temperature forecast and the multi-year temperature; P_{n,θ_n}—the load from the last year; a—regression coefficient with the temperature θ; b—the average load ratio for years j and $j-1$.

Using statistical methods [2, 3, 20] the peak load level growth for individual residential consumers can be computed with:

$$S_{\max} = \overline{S}_{\max} + \lambda \cdot \sigma \tag{37}$$

where $\overline{S}_{\max}$—the mean value of the peak load for the residential consumer:

$$\overline{S}_{\max} = \sum_{i=1}^{n} S_{\max_i} \tag{38}$$

σ—mean square deviation, computed as a particular form:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n}\left(S_{\max_i} - \overline{S}_{\max}\right)^2} \tag{39}$$

n—number of residential consumers with the available measurements;

λ—rated deviation of the normal distribution.

For the estimation of the monthly load, the profile of the warm season (December—month 12) and the profile of the cold season (June, month 6) can be used in any month l:

$$P_{t,l} = \frac{P_{t,12} + P_{t,6}}{2} + \frac{P_{t,12} - P_{t,6}}{2} \cos\frac{\pi \cdot l}{2} \tag{40}$$

where: $P_{t,l}$—the active power at hour $t = 1, \ldots, 24$, in month l; $P_{t,6}$, $P_{i,12}$—the active power at hour $t = 1, \ldots, 24$, in month 6 (June) and month 12 (December).

If the yearly load growth is considered, (40) can be rewritten as:

$$P_{t,l} = \frac{\alpha \cdot P_{t,12} + \frac{1+\alpha}{2} \cdot P_{t,6}}{2} + \frac{\alpha \cdot P_{t,12} - \frac{1+\alpha}{2} \cdot P_{t,6}}{2} \cdot \cos \frac{\pi \cdot l}{2} \tag{41}$$

where α is the yearly load growth coefficient.

The estimation model or function is chosen according to the least squares' criterion, which seeks the minimization of the sum S of the squared differences between the computed and the real energy consumption values, written as:

$$S = \sum_{k=1}^{n} d_k^2 = \sum_{k=1}^{n} [y_k - f(x_k, a_0, a_1, \ldots, a_n)]^2 \tag{42}$$

If the obtained values have different variances, then the measured values were obtained with measurement devices having different precision classes (42) can be rewritten as:

$$S = \sum_{k=1}^{n} d_k^2 = \sum_{k=1}^{n} \left\{ [y_k - f(x_k, a_0, a_1, \ldots, a_n)]^2 \cdot \omega_k \right\} \tag{43}$$

where ω_k are weights inversely proportional with the variance of the measured values, respectively:

$$\omega_1 = \frac{1}{\sigma_1^2}; \quad \omega_2 = \frac{1}{\sigma_2^2}; \quad \ldots \quad \omega_n = \frac{1}{\sigma_n^2} \tag{44}$$

The values $a_0, a_1, \ldots, a_n$, are obtained by minimizing S $(a_0, a_1, \ldots, a_n)$:

$$\frac{\partial S}{\partial a_0} = 0; \quad \frac{\partial S}{\partial a_1} = 0; \quad \ldots \quad \frac{\partial S}{\partial a_n} = 0 \tag{45}$$

By solving (45), the best regression coefficients are determined for a function family $y = f(x)$. The direct extrapolation procedure used for determination the best regression coefficients for the load estimation is illustrated in the following for the logistic and power functions. The logistic function used for the estimation of the trend term in time series has the following expression:

$$y = \frac{a}{1 + b \cdot e^{-c \cdot x}} \tag{46}$$

where a is the limit value of y in time, and can be frequently assessed with non-statistical means.

In order to find a, b and c in (46), a possible approach is to empirically choose three values (y_1, y_2, y_3) which correspond to the (x_1, x_2, x_3) equidistant points illustrated in

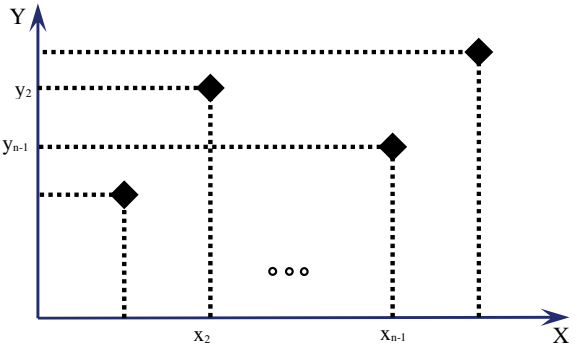

Fig. 7 Representing Y values using equidistant X values

Fig. 7. For simplifying the computation effort, the following notations can be used:

$$x_1 = 0; \quad x_2 = \theta; \quad x_3 = 2\theta \tag{47}$$

Thus, the logistic function (47) can be written:

$$\frac{a - y}{y} = b \cdot e^{-c \cdot x} \tag{48}$$

If $x = x_1 = 0$, then b can be computed with:

$$b = \frac{a - y_1}{y_1} \tag{49}$$

Using the natural logarithm transformation, (48) becomes

$$\ln b - c \cdot x = \ln\left(\frac{a - y_1}{y_1}\right) \tag{50}$$

Similarly, if $x = x_2 = \theta$ and $x = x_3 = 2\,\theta$,

$$\ln b - c\theta = \ln\left(\frac{a - y_2}{y_2}\right); \ \ln b - 2c\theta = \ln\left(\frac{a - y_3}{y_3}\right) \tag{51}$$

By using (50), multiplying the first equation by (-2) and adding it with the second equation from (51), we obtain

$$\frac{a - y_1}{y_1} = \left(\frac{a - y_2}{y_2}\right)^2 \cdot \frac{y_3}{a - y_3} \tag{52}$$

Using (52), a can be written as:

$$a = \frac{2y_1 \cdot y_2 \cdot y_3 - y_2^2(y_1 + y_3)}{y_1 \cdot y_3 - y_2^2} \tag{53}$$

Once a from the logistic function (46) is computed using (53), b can be determined with (49), and c with (50), follows using:

$$c\theta = \ln b - \ln\left(\frac{a - y_2}{y_2}\right) = \frac{a - y_1}{y_1} - \ln\left(\frac{a - y_2}{y_2}\right) \tag{54}$$

or

$$c\theta = \ln \frac{a - y_1 y_2}{a - y_2 y_1}; \quad c = \frac{1}{\theta} \cdot 2.3026 \cdot \log \frac{y_2(a - y_1)}{y_1(a - y_2)} \tag{55}$$

Knowing a, b and c, the logistic function can be computed for any each value of the variable x.

As presented in the literature [21–23], the logistic function can be used for yearly estimations only for longer intervals (8–10 years), especially for consumer categories with similar appliances and demand profiles. As for the use of the power function in load extrapolation, its initial expression is

$$y = a \cdot x^b \tag{56}$$

By using the transformation of natural logarithm, we get

$$y = \ln(A) + B \ln(x) \tag{57}$$

and by substituting $Y = \ln y$; $a = \ln A$; $X = \ln x$; $B = b$, a linear function is obtained:

$$Y = b \cdot X + a \tag{58}$$

The best regression curve fulfils the least mean square criterion:

$$S = \sum_{k=1}^{l} (Y_k - b \cdot X_k - a)^2 \rightarrow \min \tag{59}$$

To find the minimum value of S, it's the first order derivatives in report with a and b must be set to zero ($\partial S/\partial a = 0$; $\partial S/\partial b = 0$), which gives the following equations system:

$$\begin{cases} b \cdot \sum\limits_{i=1}^{l} \ln x_i + m \cdot a = \sum\limits_{i=1}^{l} \ln y_i \\ b \cdot \sum\limits_{i=1}^{l} (\ln x_i)^2 + m \cdot a \cdot \sum\limits_{i=1}^{l} \ln x_i = \sum\limits_{i=1}^{l} \ln x_i \cdot \ln y_i \end{cases} \tag{60}$$

By solving the linear equations system (60), the power function coefficients are obtained:

$$a = \frac{\sum_{i=1}^{l} \ln y_i \cdot \sum_{i=1}^{l} (\ln x_i)^2 - \sum_{i=1}^{l} \ln x_i \cdot \sum_{i=1}^{l} \ln x_i \cdot \ln y_i}{m \cdot \sum_{i=1}^{l} (\ln x_i)^2 - \left(\sum_{i=1}^{l} \ln x_i\right)^2} \tag{61}$$

$$b = \frac{m \cdot \sum_{i=1}^{l} \ln x_i \cdot \ln y_i - \sum_{i=1}^{l} \ln x_i \cdot \sum_{i=1}^{l} \ln y_i}{m \cdot \sum_{i=1}^{l} (\ln x_i)^2 - \left(\sum_{i=1}^{l} \ln x_i\right)^2} \tag{62}$$

The Romanian standards recommends the use a power function for residential load estimation:

$$P(t) = A \cdot t^b = P(t) \cdot t^b \tag{63}$$

If it is considered the 2000–2030 interval, the signification of terms from (63) is the following: *P(t)*—the estimated load for year *t*; *t*—a year from the range [2000, 2020], ($t = 1$ for year 2000); $A = P(t)$—the demand in the first year (2000), used as base value; *b*—regression coefficient, based on historical data, whose value differs for each consumer category.

The estimation functions for the demand evolution in urban areas, considered as power required by MV/LV electric substations, maximum and minimum value, are given in [24] for the 2000–2035 interval. It should also be noted that for the estimation of the demand for the apartments found in crowded areas or in individual buildings more than 4 levels, the following supplemental values should be added: for staircase lighting—0.2 kW/store (4/6 apartments); elevators—10 kW/drive; fire hose enclosure lighting:—2 kW/entrance.

The choice between the maximum and the minimum value should be made in the design stage, taking into account the geographical area, the economic environment, consumer density etc. [5, 6].

3.3.2 Testing the Solution

Using the capabilities of the Smart Meters, which can record consumption values, data was recorded for seven consecutive years (2012–2018) on the LV side of four MV/LV electric substations located in an electric distribution network belonging of a DNO from Romania. The monitored substations supply 390 apartments with 2 and 3 room apartments.

A first category (Group I) contains 205 apartments which use natural gas for cooking and receive hot water and heating from the central thermal power plant. The

second category (Group II) contains 185 apartments which use natural gas for cooking and individual thermal plants for hot water and heating. Table 3 and Fig. 8 show the electricity demand evolution measured in the four monitored MV/LV substations, as measured by the smart meters.

Initially, in order to identify the most representative mathematical model for the load estimation, as described in the previous sections, continuous growth functions (linear, parabolic, polynomial, exponential) limited growth (power, logarithmic, modified exponential, logistic) and modified combinations functions were used.

In the second stage, the regression coefficients were determined for each function and apartment category, using the time as interest variable and the minimum least square criterion. The results confirmed that the power function has the smallest sum of squared estimate of errors (SSE), confirming the validity of the estimation function type recommended in the standards. However, the regression coefficients differ slightly:

$$\text{Group I}: \; W(t) = 0.420 \cdot t^{0.201} \tag{64}$$

Table 3 The demand evolution for each apartment category, [kW/ap]

Year	Group I	Group II
2012	0.420	0.467
2013	0.482	0.543
2014	0.523	0.594
2015	0.554	0.632
2016	0.58	0.664
2017	0.602	0.691
2018	0.637	0.736

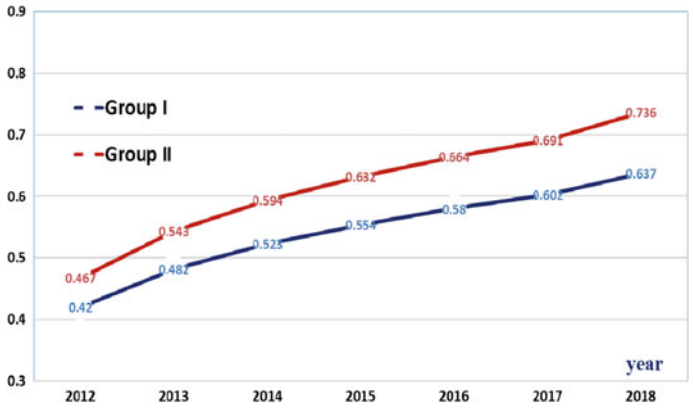

Fig. 8 The demand evolution measured at the LV side of electric substation, for the considered apartment categories and years 2012–2018

Table 4 The demand evolution on the LV side of electric substations, for the considered apartment categories using new and the recommended coefficients

Year	Calculated values		The extreme values of range from the Romanian normative	
	Group I	Group II	Minimum	Maximum
2012	0.420	0.467	0.450	0.520
2013	0.482	0.543	0.500	0.580
2014	0.523	0.594	0.540	0.630
2015	0.554	0.632	0.570	0.670
2016	0.580	0.664	0.600	0.710
2017	0.602	0.691	0.630	0.739
2018	0.637	0.736	0.680	0.800
2019	0.667	0.773	0.727	0.851
2020	0.692	0.804	0.768	0.899
2021	0.703	0.818	0.790	0.920
2022	0.713	0.832	0.804	0.942
2023	0.733	0.857	0.840	0.980
2024	0.750	0.879	0.870	1.020
2025	0.767	0.899	0.899	1.053
2026	0.781	0.918	0.927	1.085
2027	0.788	0.927	0.940	1.100
2028	0.804	0.947	0.969	1.133
2029	0.816	0.963	0.992	1.160
2030	0.829	0.979	1.016	1.187

$$\text{Group II}: \; W(t) = 0.467 \cdot t^{0.219} \tag{65}$$

It should be noted that in the Romanian standard, the power function coefficients have different values according to the number of rooms in the apartment and the heating/cooking type, as described earlier.

For the apartment types used in the study case, two different coefficient sets are provided:

$$\text{Minimal } W(t) = 0.305 \cdot t^{0.35} \tag{66}$$

$$\text{Maximal } W(t) = 0.357 \cdot t^{0.35} \tag{67}$$

For a comparative analysis of the coefficients associated the estimation function obtained in the study case, relations (64) and (65), and given in the standards, relations

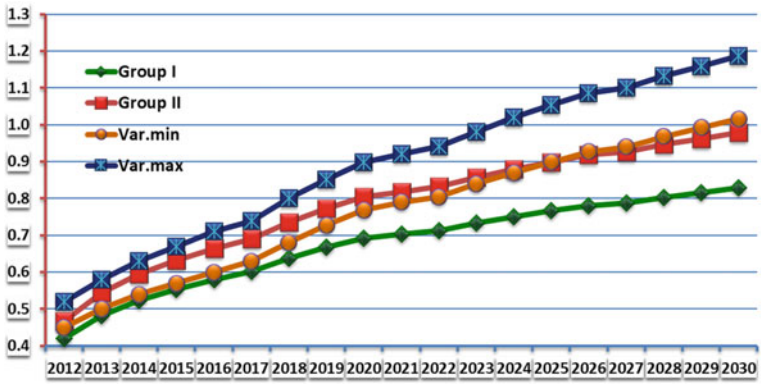

Fig. 9 The demand evolution for the categories of considered apartment

(66) and (67), Table 4 and Fig. 9 present the demand evolution on the LV side of electric substations, between 2012 and 2030 (estimated).

The following conclusions can be highlighted:

- The coefficients of the power function computed in the study case have different values comparative with those from the normative.
- The new estimated values are inside the range given in [24] for the apartments using individual thermal substations.

4 Conclusions

The estimation of the loads in different parts of the distribution system represents a main function of the DNOs. The electricity cannot be efficiently stored on a large scale (relative to the produced amount), which means that for the DNOs, the estimation of the loads is an indispensable factor in the distribution process. The regression models are some of the most commonly used statistical techniques. For the estimation of electricity/power consumption such approaches are used to model the relationship between consumption and other factors such as weather, type of day, nature of consumption, etc. Usually RL model is used using in most cases the temperature. The advantages of this model are related to the relatively simple implementation, the easy understanding of the relationship between the input and output variables and the easy estimation of the performance of the forecasting model. However, due to the complex dependence between electricity consumption and influence factors, inherent problems arise in identifying the correct model.

To solve this problem, the regression analysis-based approaches for the load modelling from the nodes of electricity distribution networks were treated in the chapter. The approaches refer to estimation of the required powers in the supply

points with a mixt structure of the load (i.e. residential, commercial, and industrial) at the hour when the maximum value of the load is recorded and the demand of residential consumers which represent the highest percentage from the load structure fed from the LV/MV electric substations. The proposed approaches were tested in real operation conditions of MV distribution networks from Romania. Thus, the estimation of the loads from the MV/LV electric substations of a test network, at the hour when the maximum value (peak load) was recorded, using the proposed method based on the power correlation, led at an average error for the time frame 7 h (hPL $-$ 5 h; hPL $+$ 1 h) below 1.48% than in the others frames L_{24} frame or L_7 frames ($h_{PL} \pm 3\ h$), ($h_{PL} - 4\ h$; $h_{PL} + 2\ h$).

Regarding the estimation of the demand in the case of residential consumers, the comparative analysis of the coefficients associated the estimation function and to those given in the Romanian standard highlighted that the estimated values are lower than the minimum recommended values for the apartments which use natural gas for cooking and individual thermal plants for hot water and heating. This behaviour could be the result of a modified behaviour of the customers or due to the used database which belonging of a characteristic electric substation from the analysed area, while in used data from the Romanian standard are collected from the whole country.

References

1. Phuangpornpitak N, Prommee W (2016) A study of load demand forecasting model in electric power system operating and planning. Greater Mekong Subreg Acad Res Netw Int J 10:19–24
2. Neagu B, Georgescu G (2010) Load and energy forecast on a proximate, medium and long horizon in public electricity repartition and distribution systems. Buletinul Institutului Politehnic din Iaşi LVI (LX) 3:71–82
3. Li K, Wang B, Wang Z, Wang F, Mi Z, Zhen Z (2017) A baseline load estimation approach for residential customer based on load pattern clustering. Energy Procedia 142:2042–2049
4. Avdakovic S, Ademovic A, Nuhanovic A (2013) Correlation between air temperature and electricity demand by linear regression and wavelet coherence approach: UK, Slovakia and Bosnia and Herzegovina case study. Arch Electr Eng 62(4):521–532
5. Vu D, Muttaqi KM, Agalgaonkar AP (2014) Assessing the influence of climatic variables on electricity demand. In: IEEE power and energy society general meeting, Washington, USA, 2014, pp 1–5
6. Mfonobong Umoren A, Okpura N, Markson I (2017) Rural electrification peak load demand forecast model based on end user demographic data. Math Softw Eng 3(1):87–98
7. Grigoraș G, Cârțină G (2013) The fuzzy correlation approach in operation of electrical distribution systems. Int J Comput Math Electr Electron Eng 32(3):1044–1066
8. Grigoraş G, Scarlatache F, Neagu BC (2010) Clustering in power systems. Applications. Lambert Academic Publishing, Germany
9. Chen C, Zhou JN (2014) Application of regression analysis in power system load forecasting. Adv Mater Res 960(961):1516–1522
10. Zheng R, Zhijian JG, Hongqiao Peng J, Zhu Y (2019) Regression analysis of time series for forecasting the electricity consumption of small consumers in case of an hourly pricing system. Int Trans Electr Energy Syst e12100033
11. Gelman A, Hill J (2007) Data analysis using regression and multi-level/hierarchical models. Cambridge University Press, New York

12. Jaba E (2000) Statistics (in Romanian). Economic Publishing, Bucharest, Romania
13. Nagumo T, Ito H, Sano T (2017) Load current forecasting using statistical analysis. In: 24th international conference & exhibition on electricity distribution (CIRED), Glasgow, Scotland
14. Supapo KRM, Santiago RVM, Pacis MC (2017) Electric load demand forecasting for Aborlan-Narra-Quezon distribution grid in Palawan using multiple linear regression. In: IEEE 9th international conference on humanoid, nanotechnology, information technology, communication and control, environment and management (HNICEM), Manila, Philippines
15. Bai J, Jiang M, Liu L, Sun Y, Wang Y, Zhang J (2019) Correlation analysis and prediction of power network loss based on mutual information and artificial neural network. In: IOP conference series: earth and environmental science [Online]. https://iopscience.iop.org/article/. https://doi.org/10.1088/1755-1315/227/3/032023/pdf
16. Rahman H, Selvarasan I, Begum J (2018) A short-term forecasting of total energy consumption for India-A black box based approach. Energies 11, paper 2442 [Online]. https://www.mdpi.com/1996-1073/11/12/3442
17. Almeshaiei E, Soltan H (2011) A methodology for electric power load forecasting. Alexandria Eng J 50(2):137–144
18. Azad MK, Uddin S, Takruri M (2018) Support vector regression based electricity peak load forecasting. In: IEEE 11th international symposium on mechatronics and its applications (ISMA), Sharjah, United Arab Emirates
19. Abdulkareem A, Okoroafor EJ, Awelewa A, Adekitan A (2019) Pseudo-inverse matrix model for estimating long-term annual peak electricity demand: the Covenant University's experience. Int J Energy Econ Policy 9(4):103–109
20. Friedrich L, Armstrong P, Afshari A (2014) Mid-term forecasting of urban electricity load to isolate air-conditioning impact. Energy Build 80:72–80
21. Rabie AH, Saleh AI, Abo-Al-Ez KM (2015) A new strategy of load forecasting technique for smart grids. Int J Modern Trends Eng Res 2(12):332–341
22. Muller CJ, MacLehose RF (2014) Estimating predicted probabilities from logistic regression: different methods correspond to different target populations. Int J Epidemiol 43(3):962–970
23. Gajowniczek K, Ząbkowski T (2017) Two-stage electricity demand modeling using machine learning algorithms. Energies 10(10):1547
24. PE 132/2003, Normative of design of the public distribution networks, S.C. ELECTRICA S.A., Bucharest (2003)