

## **The time series extrapolation model based on maximum likeness set**

I. Chuchueva  
[chuchueva@yandex.ru](mailto:chuchueva@yandex.ru)

### **Introduction**

A time series forecast task is the basis for 1) planning in economics and trading, 2) manufacturing planning, 3) storehouse control, 4) control and optimization of industrial processes etc. Information system progress brings considerable development of extrapolation models.

All extrapolation methods can be divided into two groups: formalized and intuitive ones. According to the article [1] the formalized methods can be mainly divided into two categories: statistical and artificial intelligence ones. In the statistical methods the equations can be obtained showing the relationship between the time series values and the external factors after examining historical data, while the artificial intelligence methods try to imitate the way of human beings' thinking and reasoning to get knowledge from the past experience and to forecast the future value.

Time series methods are based on the assumption that the data have an internal structure, such as autocorrelation, trend or seasonal variation. The methods detect and explore such a structure. ARMA (autoregressive moving average), ARIMA (autoregressive integrated moving average) and ARIMAX (autoregressive integrated moving average with exogenous variables) are the most often used classical regression time series methods.

In this research a new type of autocorrelation extrapolation model is offered. The model is based on the assumption that there are a lot of factors which influence the time series values. But contrary to the classical regression models the influence level of each factor can not be estimated due to the information volume and secure reasons. The main idea of offered model contains the following assumption. In the past period of time the overall external factors influence brought a process into some state which we call initial. It's compulsory that in the future there will be a period of time when the process will have a state which is very similar to the initial one. That assumption is based on Dirichlet's box principle for a pseudorandom sequence with finite sets. In the offered model we deal with the response time series only.

The extrapolation is the process of constructing new data points outside the discrete set of known data points. In the paper a new time series extrapolation model based on maximum likeness set (EMMLS) is offered.

The efficiency of EMMLS is evaluated on time series of day a-head prices and energy consumption of the Russian Wholesale Electricity Market (RWEM). The extrapolated values of the mentioned time series are necessary to control technical and economic parameters of the RWEM. Besides the extrapolated values of the day a-head prices help market participants to increase their financial results and hedge risks.

Note that the day a-head prices extrapolation task is a new task for the RWEM. The peculiarity of the day-ahead prices extrapolation is based on changing market rules: prices up to 01.01.2008 were calculated on one algorithm, later on the algorithm has been changed. The significance of price extrapolation task is described in a list of the relevant articles. There are a lot of specialized extrapolation models for energy consumption time series. But the research in this field is going on together with a new level of accuracy requirements.

In the paper there is the description of EMMLS in chapter 1, identification of EMMLS is reviewed in chapter 2 and numerical results for prices and consumption time series are given in chapter 3.

## Chapter 1. The extrapolation model description

In the research we deal with discrete time series values which are obtained in moments  $t_1, t_2, t_3, \dots, t_N$ . The time moments can be non-equidistant. Let's denote the time series  $Z = z(t_1), z(t_2), z(t_3), \dots, z(t_N)$  with  $Z_1^N = z_1, z_2, z_3, \dots, z_N$ . The collection of consecutive values  $Z_t^M = z_t, z_{t+1}, z_{t+2}, \dots, z_{t+M-1}$  which is located inside time series  $Z_1^N$  we name **set** length  $M$  and start time  $t$ ,  $M \in \overline{[1, N-1]}$ ,  $t \in \overline{[1, N-M]}$ . The time difference between start time  $Z_t^M$  and  $Z_{t-k}^M$  is named **delay**  $k$ ,  $k \in \overline{[1, t-1]}$ .

The correlation coefficient for sets  $Z_t^M$  and  $Z_{t-k}^M$  calculated by

$$\rho_k = \frac{\text{cov}(Z_t^M, Z_{t-k}^M)}{\sqrt{D[Z_t^M]} \cdot \sqrt{D[Z_{t-k}^M]}}, \quad (1)$$

where  $\text{cov}(Z_t^M, Z_{t-k}^M)$  is covariance of sets  $Z_t^M$ ,  $Z_{t-k}^M$  and  $D[Z_t^M]$ ,  $D[Z_{t-k}^M]$  are their variances respectively [2]. Based on literature (1) we introduce **likeness measure** for sets  $Z_t^M$  and  $Z_{t-k}^M$

$$\rho_k^M = \frac{|\text{cov}(Z_t^M, Z_{t-k}^M)|}{\sqrt{D[Z_t^M]} \cdot \sqrt{D[Z_{t-k}^M]}} \in [0,1].$$

Value  $\rho_k^M$  depends on the set length M and the delay between  $Z_t^M$  and  $Z_{t-k}^M$ . The likeness measure  $\rho_k^M$  reflects the level of a linear dependence. The closer  $\rho_k^M$  to 1 higher the linear dependence is.

Let's calculate for set  $Z_t^M$  values  $\rho_1^M, \rho_2^M, \dots, \rho_{t-1}^M$  and find maximum  $\rho_{k \max}^M = \max(\rho_1^M, \rho_2^M, \dots, \rho_{t-1}^M)$ . The set which corresponds to delay  $k_{\max}$  we denote  $Z_{t-k \max}^M$  and name **maximum likeness set** for  $Z_t^M$ . It's easy to see that for  $Z_{t-k \max}^M$  exists equation

$$\rho_{k \max}^M = \frac{|\text{cov}(Z_t^M, Z_{t-k \max}^M)|}{\sqrt{D[Z_t^M]} \cdot \sqrt{D[Z_{t-k \max}^M]}}$$

**Likeness hypothesis.** If sets  $Z_t^M$  and  $Z_{t-k \max}^M$  have  $\rho_{k \max}^M$  value close to 1, then for some P and sets  $Z_t^{M+P}$  and  $Z_{t-k \max}^{M+P}$  value  $\rho_{k \max}^{M+P}$  will be also close to 1. The extrapolation results which are given in chapter 3 confirm the hypothesis for the researched time series. For the other time series the hypothesis must be verified.

Let's switch to vector designation and denote set  $\mathbf{Z}_t^M = (z_t, z_{t+1}, \dots, z_{t+M-1})^T$ , and time series  $\mathbf{Z}_1^N = (z_1, z_2, \dots, z_N)^T$ . Set  $\mathbf{Z}_t^M$  can be approximated using set  $\mathbf{Z}_{t-k \max}^M$  as follows:

$$\mathbf{Z}_t^M = a_1 \cdot \mathbf{Z}_{t-k \max}^M + a_0 + \mathbf{E}^M;$$

$$\check{\mathbf{Z}}_t^M = a_1 \cdot \check{\mathbf{Z}}_{t-k \max}^M + a_0. \quad (2)$$

Here  $a_1$  and  $a_0$  are constants,  $\mathbf{E}^M$  is the vector of approximation errors,  $\check{\mathbf{Z}}_t^M$  are the model values for  $\mathbf{Z}_t^M$ . Then set  $\mathbf{Z}_{N+1}^P$  can be expressed using some set  $\mathbf{Z}_\tau^P$ , which is located inside the original time series  $\mathbf{Z}_1^N$  as follows

$$\check{Z}_{N+1}^P = a_1 \cdot Z_{\tau}^P + a_0. \quad (3)$$

The algorithm for defining set  $Z_{\tau}^P$  consists of the next steps.

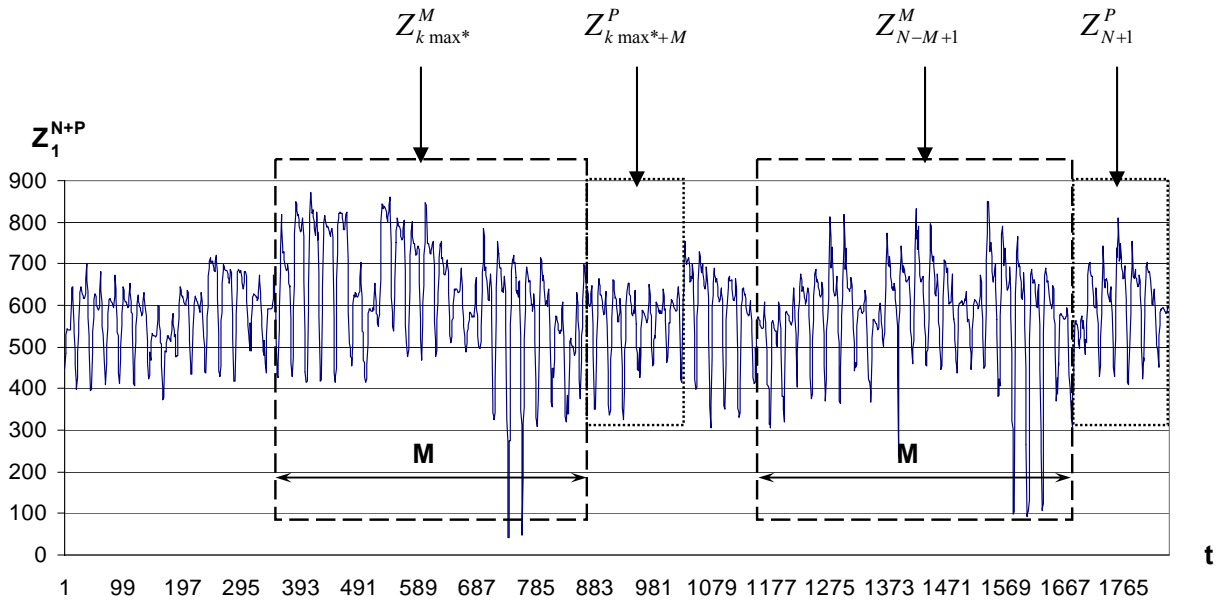
- For set  $Z_{N-M+1}^M$  find maximum likeness set  $Z_{k \max^*}^M$ , where  $k \max^* = N - M + 1 - k \max$ .
- According to literature (2), there is an approximation for  $Z_{N-M+1}^M$  using set  $Z_{k \max^*}^M$ :

$$\check{Z}_{N-M+1}^M = a_1 \cdot Z_{k \max^*}^M + a_0.$$

Constants  $a_1$  and  $a_0$  can be calculated by solving the following task using least square method which is widely described in literature

$$S = \sum_{i=1}^M (\check{Z}_{N-M+1-i} - Z_{N-M+1-i})^2 \rightarrow \min.$$

- According to **likeness hypothesis** in place set  $Z_{\tau}^P$  must be used set  $Z_{\tau}^P = Z_{k \max^*+M}^P$ , i.e. set which is located right after the maximum likeness set (see pic 1).



Pic 1. Location of sets  $Z_{k \max^*}^M$ ,  $Z_{k \max^*+M}^P$ ,  $Z_{N-M+1}^M$  and  $Z_{N+1}^P$

The extrapolated values  $\check{Z}_{N+1}^P$  for time series  $Z_1^N$  can be calculated by following formula

$$\check{Z}_{N+1}^P = a_1 \cdot Z_{k \max^*+M}^P + a_0 = \text{EMMLS}(M). \quad (4)$$

The formula (4) describes the extrapolation model based on the maximum likeness set, EMMLS.

## Chapter 2. The extrapolation model identification

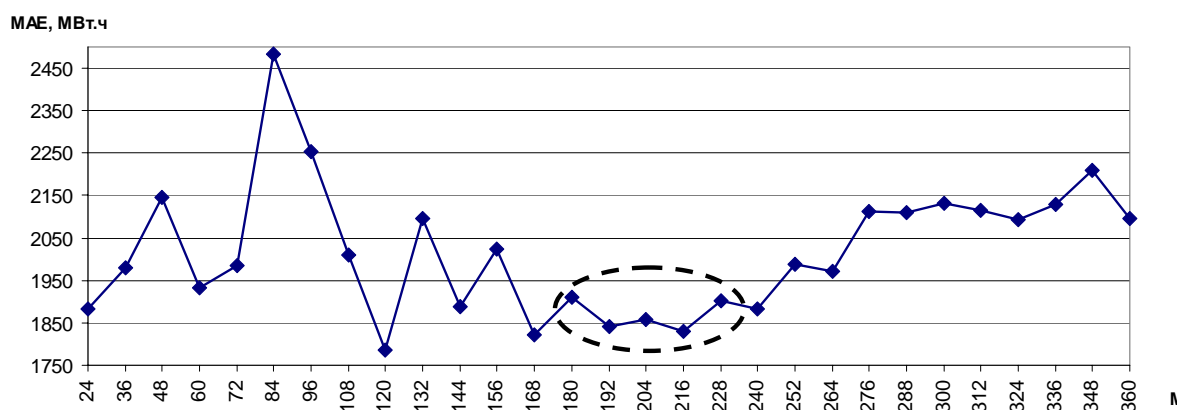
The identification of the offered model should be done by the following algorithm.

- Divide the original time series  $Z_1^N$  into 3 parts in proportion 2:2:1. Each part is named as **base period** (40%), **assessment period** (40%) и **test period** (20%) of the time series respectively.
- Concerning forecast task define  $P$  and a range of possible values  $M$ . Initially we recommend to take into account a wide range for parameter  $M$ , for example,  $M \in [P, 30 \cdot P]$ , and later specify it.
- For each  $M$  value extrapolate the time series inside the **assessment period**.
- The extrapolation results which have been obtained with previous step must be evaluated using mean absolute error,  $MAE$ :

$$MAE = \frac{1}{K} \cdot \sum_{i=1}^K |\tilde{z}_i - z_i|. \quad (6)$$

- After calculating all  $MAE$  values for each  $M$  inside the assessment period choose range of  $M$  values which correspond to the stable minimum.
- As the last step the expert chooses particular  $M$  value from defined stable minimum range.

As an example see the dependence  $MAE$  from  $M$  value for the energy consumption time series European price zone of the RWEM. (see pic 2.)



Pic 2. Dependence *MAE* from *M* value,  $M \in [24,360]$ , step 12

Initially has been chosen *M* value range,  $M \in [24,360]$ . Inside initial range has been specified stable minimum range,  $M \in [180,228]$ . Finally has been estimated  $M = 216$ .

### Chapter 3. The efficiency of extrapolation model

The efficiency research of offered model (4) has been implemented on price and consumption time series, which have been given by ATS organization:

- Energy consumption European price zone Russian energy market (EPZ);
- Energy consumption Siberian price zone Russian energy market (SPZ);
- Day a-head price European price zone Russian energy market (EPZ);
- Day a-head price Siberian price zone Russian energy market (SPZ).

Each time series consists of values in time range from 01.09.2006 to 30.09.2009. Except extrapolation of hourly time series it makes sense to extrapolate daily time series: for the energy consumption time series – integrated value for day; for the price time series – mean value for day. Characteristics of time series are given in Table 1.

Table 1. Time series characteristics

Resolution	Time series	Length	Mean	Standard deviation	Minimum	Maximum
Hourly	Energy consumption EPZ (MWh)	27 023	81 441	10 640	57 847	110 586
	Energy consumption SPZ (MWh)	27 023	22 338	3 023	15 328	30 666
	Dayahead price EPZ (RUB/ MWh)	27 023	610	194	0	1 559
	Dayahead price SPZ (RUB/ MWh)	27 023	368	185	0	1 029
Daily	Energy consumption EPZ (MWh)	1 126	1 954 502	218 255	1 519 197	2 438 008
	Energy consumption SPZ (MWh)	1 126	536 101	68 930	408 408	695 606
	Dayahead price EPZ (RUB/ MWh)	1 126	610	144	221	1 204
	Dayahead price SPZ (RUB/ MWh)	1 126	368	176	0	734

The extrapolation of hourly time series imply calculations of 24 values for the next day, extrapolation of daily time series imply calculations of the single value for the next day. According to [2] such kind of extrapolation is the short-term extrapolation.

Below numerical extrapolation results are given (see Table 2 and Table 3). The period from 01.03.2009 to 30.09.2009 (7 months, about 5000 values for hourly time series, about 200

for daily time series) has been selected to forecast and validate the performance of the EMMLS (**test period**). There are also  $M$  values for each time series in Table 2 and Table 3. Calculated  $M$  values can be taken as a basis while working with consumption and day-ahead prices time series, i.e. the similar nature time series. The model validation has been obtained using mean absolute percentage error, MAPE

$$MAPE = \frac{1}{Q} \cdot \sum_{i=1}^Q \frac{|\tilde{z}_i - z_i|}{z_i} \cdot 100\%$$

$Q$  is a number of the time series values inside period from 01.03.2009 to 30.09.2009.

### 3.1 The consumption forecast

Table 2. Consumption forecast results

Time series	Resolution	M	MAE, MWh	MAPE, %	Running time, h
Energy consumption	Hourly	216	1 347	<b>1.04</b>	1.80
EPZ	Daily	6	19 736	<b>1.10</b>	0.15
Energy consumption	Hourly	24	378	<b>1.86</b>	4.20
SPZ	Daily	8	7 727	<b>1.44</b>	0.05

A good performance of the extrapolation model can be observed. The MAPE values are around 1.5%, where the lowest mean error is 1.04% and the highest one is 1.86%. Compared to the results which are published in other articles we make a conclusion that achieved level of accuracy is efficient for the consumption time series.

### 3.2 The day-ahead prices forecast

Table 3. Day-ahead prices forecast results

Time series	Resolution	M	MAE, RUB/MWh	MAPE, %	Running time, h
Dayahead price EPZ	Hourly	360	49.54	<b>7.00</b>	2.10
	Daily	16	34.50	<b>4.81</b>	0.04
Dayahead price SPZ	Hourly	84	65.90	<b>39.78</b>	4.93
	Daily	30	60.39	<b>32.93</b>	0.04

In table 3 there are MAPE values 39.78% and 32.93%. Those values can be explained that 0 (zero) values of day-ahead prices SPZ are occurred in the forecast period. During the processing zero values have been replaced with 0.01 RUB/WMh. As was mentioned above the task of prices extrapolation is a new for the RWEM that's why it's not possible to provide a comparison with other results to make conclusion about the efficiency of EMMLS.

Compared to the forecast results for the other markets (for example, California and Spain markets, [2]) we can assume that EMMLS is efficient for the RWEM. Obtained MAE and MAPE values are very reasonable taking into account the complex nature of the price time series and the results previously reported in the technical literature.

It's clear that for the day-ahead price SPZ time series the additional improvement of model is required.

## **Conclusion**

This paper has proposed EMMLS to forecast hourly and daily electricity prices and consumption time series for the RWEM. Consumption forecast MAPE values are between 1.04% and 1.86%. Obtained MAE values for the day-ahead prices EPZ are 34.50 RUB/MWh and 49.54 RUB/MWh for daily and hourly resolution respectively. The day-ahead prices SPZ MAE values are 60.39 RUB/MWh and 65.90 RUB/MWh respectively. The model for the day-ahead prices SPZ requires further improvements. The differences in EPZ and SPZ price results may reflect different technical and financial zone structure.

The high efficiency of EMMLS for the energy consumption time series forecast is estimated.

In the future the extrapolation researches in such fields as FOREX indexes, environment temperature, NYMEX futures prices, blood sugar level and some others are planned.

## **References**

1. M. Sc. Jingfei Yang. Power System Short-term Load Forecasting – Elektrotechnik und Informationstechnik der Technischen Universität Darmstadt. – 2006. – P. 84.
2. Reinaldo C. Garcia A. GARCH Forecasting Model to Predict Day-Ahead Electricity Prices. – German Institute of Economic Research, Germany. – 2003. – P. 12.