Construction of Multi-step Price Forecasts in Commodity Markets Based on Qualitative and Quantitative Data Analysis Methods

Zinaida K. Avdeeva, Elena A. Grebenyuk, Svetlana V. Kovriga

Construction of multi-step price forecasts in commodity markets based on qualitative and quantitative data analysis methods

Zinaida K. Avdeeva [0000-0002-4517-6750], Elena A. Grebenyuk [0000-0002-5153-8578] and Svetlana V. Kovriga [0000-0001-7675-5192]

[1] Trapeznikov V. A. Institute of Control Sciences of RAS, Moscow, Russia
zinaida.avdeeva@gmail.com, lngrebenuk12@yandex.ru, kovriga@ipu.ru

Abstract. The article proposes a method for constructing and correcting a multi-step forecast for the year ahead (with a monthly breakdown) of prices for raw materials and products of industrial enterprises. The proposed approach consists in the formation of a price forecast taking into account 1) the price of the predicted indicator, the prices of goods participating in the product value chain, and macro indicators (time series); 2) information about the strength and direction of environmental factors affecting the market**.** Structured information about the effects of the external environment is the result of processing expert knowledge and hypotheses from heterogeneous information sources, through analysis and modeling on a cognitive map of the situation (CCS). We form a forecast by constructing an ensemble of time series models, each of which reflects the dependence of the target indicator on its past values and the prices of related products, the composition of which is determined by the results of cognitive modeling and time series analysis. Based on the results of monitoring on the cognitive map of the situation, conducting in order to analyze possible changes in the external environment and digital monitoring of prices, to identify changes in prices modes, we perform a forecast correction. The results obtained in this study show that the use of cognitive modeling and monitoring of changes improve the accuracy of predictions.

Keywords: manufacturing system, multi-step forecasting quantitative and qualitative forecasting, cognitive map, monitoring, time series.

1 Introduction

Forecasting of prices for raw materials and final products of industrial enterprises is an important element of planning of their activity. Since planning is carried out for a year ahead, long-term forecasts (broken down by months) are necessary, accounting for which will increase the profit of the enterprise and reduce unproductive costs in case the forecasts have the necessary accuracy. The task of forecasting the values of time series for several steps ahead is one of the most difficult tasks of forecasting.

Prices on commodity markets, determining the costs of enterprises for raw ma-terials and products have the following features:

- price change processes are nonstationary, with most of them being processes with a stochastic trend, with changing properties and interrelationships,
- there are price interactions between commodities linked together by a product value chain,
- the structure of the interrelations depends on the time scale in which we consider it and may change with changes in scale,
- there is a dependence on macro indicators such as the real interest rate, ex-change rates, oil prices, global demand for commodities, etc. [1],[2].

Future price values for the forecast period depend on quantifiable parameters such as current and past prices for raw materials and products produced from them, export and world prices for goods in the production sector under consideration, transportation tariffs, and macroeconomic indicators.

Under the influence of external environment events (political, economic, natural and man-made disasters, natural disasters, etc.), the set, the strength of the direction of the impact of qualitative factors acting on these prices, change. The current model for describing the dynamics and forecasting market parameters becomes unusable. It is necessary to detect these changes in a timely manner and adjust the model.

Many researchers note and use the influence of extreme events on the prices of commodity markets for forecasting. To identify extreme events, assess their impact on prices, and correct forecasts, [3-5] used such approaches as empirical mode decomposition (EMD), wavelet analysis, and search engine data analysis. The results of the analysis showed that taking into account the influence of ex-ternal influences significantly increases the accuracy of the forecast.

There are many relatively successfully attempts to take into account information about events affecting a price dynamic fundamental change for its inclusion in predictive models. For this, several types of mainly text information are used: news, analytical reports and expert statements, posts of market participants in professional communities. Especially actively carried out expert events with the digitization of information signals when forecasting on the stock markets, then on the raw materials, energy and commodity markets. The basis for the formation of digitized indices is the adjustment of algorithms for identifying information on a given thematic area, digitization of this information into activity indices with an assessment of emotional coloring, inclusion in the model in different ways. One of the proven methods for assessing the significance of observed changes in their possible impact on predicted parameters are cognitive maps. The CMs, apart from their predictive strength, explicitly structure the knowledge and information. This advantage of the CMs improves the inter-pretation of the links between the factors and links them with real changes in the situation [6-10]. Research area also includes research on the development of integrated methods, where the use of CMs is focused on identifying events (informational causes) that influence the formation of predicted values of the time series [6-10]. An important distinctive feature of such models is that they are constructed using heterogeneous data (quantitative and qualitative, expert) to identify significant factors and parameters of events that may affect the forecast generated. In this case, the formation and/or correction of the predictive model rely on event data extracted from heterogeneous information sources according to the results of CM analysis.

The results of the analysis of the situation on the cognitive map not only indicate and record the presence of factors that caused certain obvious changes in the market, but also allow you to form various scenarios for the development of the current situation for the forecast period. The proposed approach includes the following elements:

- analysis of the history of the predicted market fragment by cognitive modeling methods in order to identify homogeneous periods;
- building models of multi-step forecast of the target indicator values for the year ahead with a monthly breakdown in each of the considered periods;
- monitoring the situation by cognitive modeling methods;
- monitoring the time series of market parameters associated with the predicted target indicator in order to detect disorders in the on-line mode;
- correction of the forecast based on the results of monitoring the situation and time series monitoring during the forecast period.

2 Market analysis based on historical data

2.1 Situation cognitive map construction and estimation of the significance of the causal factors

As a result of cognitive modeling application, at the time of the forecast formation, the main system-forming factors that influence the dynamics of the target indicator and, possibly, the rows associated with it in the database, are determined and ranked according to the strength of their influence on the target indicator. The factors with high ranks will be called significant in what follows.

Let there be a cognitive map of the commodity market situation, including the factors of the value chain in the considered commodity market, $K_f(X, A, f)$, in which $X = (x_1, \ldots, x_n)$ is a set of factors of the situation S; $A = [a_{ij}]$ is the $N \times N$ matrix of factors mutual influence, where $a_{ij} \in [-1; 1]$ is the weight of influence of factor x_i on the factor x_j ([-1; 1] – a discrete scale); f is a function that defines the rule of factor value change at any discrete-time $t \geq 0$.

The state of the situation at any discrete point of time $t \geq 0$ expressed as follows

$$X(t+1) = Q(t)X(0) + Q(t)G(0), \tag{1}$$

where $Q(t) = E_N + A + A^2 + \ldots + A^t = (E_N - A)^{-1}$. When solving the problem of forecasting a target indicator, y, analysis on the matrix of integral influences, Q, allows us to identify the causal-factors, to assess the degree of their impact on the structure, to assess the significance of the influence of any group of factors $\{C\}$. Thus, the set of factors X are divided into classes according to the belonging of factors to the groups $\{C\}$. All causal-factors are divided into X_y^+ - factors and X_y^--factors: $X_y^+ = (x_k: q_{ky} > 0)$ and $X_y^- = (x_k: q_{ky} < 0)$. Submatrices Q_y^+ and Q_y^- are formed from the matrix Q.

The weight of the positive (or negative) influences of the causal-factors from X_y^+ (X_y^-) on the target indicator y is equal to the modulus of the sum of the weights of the cumulative influences $q_y^+ = \sum_{k=1}^{l} \left|q_{ky}^+\right|$ and $q_y^- = \sum_{k=1}^{l} \left|q_{ky}^-\right|$, where q_{ky}^+ (q_{ky}^-) is the weight of the cumulative influence of the causal-factor $x_k \in X_y^+$ ($x_k \in X_y^-$) on the target indicator y. Then the total weight of the causal-factors is $q_y^{total} = q_y^- + q_y^+$ $q_y^{total} = q_y^- + q_y^+$.

Then the significance of any group $\{Ci\}$ is estimated at K_f by the cumulative influence of causal-factors X^{C_i} related to the corresponding parameters of the model M_i (taking into account the identification of time series by the numbers of causal-factors).

At the stage of building forecasting models, the factors of the cognitive map are a source of keywords for finding sources of qualitative and quantitative information about factors, candidates for organizing regular monitoring; ranking factors or groups of factors by significance allows you to set a monitoring model based on the signs of expected changes by groups of factors with the greatest significance. Such an assessment allows you to determine an explanatory scheme for the conditions of serious qualitative changes in the analyzed factor (change in direction and change in the strength of changes). The ranking of groups of causal factors can also serve as the basis for recommending the construction of quantitative forecasting models based on certain quantitative indicators corresponding to the factors of the cognitive map.

Thus control factors according to the degree of significance of their integral influence on the achievement of the vector of goals are ranked.

2.2 Construction of digital models of multistep forecast for the year ahead

To build a digital projections, we use ensembles of linear models, each of which computes the one step forecasts $\hat{Y} = \{\hat{y}_{1|1-h}, \hat{y}_{2|2-h}, \dots, \hat{y}_{t|t-h}, \dots\}$ values of the target $Y = \{y_1, y_2, \dots, y_t, \dots\}$ as a combination of past values $y_{t-h}, y_{t-2h}, \dots$, and values of the factors $X_1, X_2, \dots X_k$, where $X_i = \{x_{i1}, x_{i2}, \dots, x_{it}, \dots\}$, $i = 1,2, \dots, k$. Since the intervals on which the model is built are short, we have limited the number of regressors included in the model to two.

For the forecast, we build ensembles of VAR and VECM models. These models are successfully used to form long-term multi-step forecasts both separately and as part of hybrid forecasting algorithms that include methods for analyzing the state of the external environment [11-13]. The build process consists of the following steps.

1. For each forecast horizon (month, 2 months, quarter, half year, year), we search in the database for time series that satisfy the conditions:

- the order of integration of the series is equal to the order of integration of the target indicator;
- according to the results of cognitive modeling, the dynamics of the series is influenced by one or several significant system-forming factors.

2. From the series selected at step 1, we form groups, including the target indicator and the series (no more than two), reflecting the relationship between the values of the target indicator and the significant system-forming factors (prices for raw materials,

products, macro indicators, etc.), and test each group for cointegration. If cointegration exists and the series X_{i1}, X_{i2} are Granger causal for the target Y, then we build VEC models for these groups. If there is no cointegration and the series of factor differences $\Delta X_{i1}, \Delta X_{i2}$ are Granger causal for the target, ΔY (i.e., the series of factor differences help to predict the values of ΔY, but ΔY does not help to predict the values of the series of differences), then we build VAR models by series differences. We repeat this procedure for each forecast horizon.

3. We calculate the forecasts in each group as the average values of the forecasts of the models included in it.

4. The results of the forecasts of each group are averaged with weights obtained using cognitive modeling, which determine the significance of each group based on the significance of the system-forming factors:

$$\mathrm{C}_1, \dots, \mathrm{C}_k\colon \sum_{i=1}^{k} \mathrm{C}_i = 1,$$

where k is the number of groups, $\mathrm{C}_1, \dots, \mathrm{C}_k$ are their values, estimated in K_f by the cumulative influence of causal-factors X^{C_i} related to the corresponding parameters of the model M_i (section 2.1). The forecast formula for each horizon h has the form:

$$\hat{Y}_{t|t-h} = \sum_{i=1}^{k} C_j \hat{y}^i_{t|t-h} \tag{2}$$

where $\bar{y}^{\mathrm{i}}_{\mathrm{t|t-h}}$ is the average value of the forecasts at time t, calculated for the models of the i-th group on the horizon h (step 3).

5. To form an integral multi-step forecast on the basis of a direct strategy, we build monthly, 2-month, quarterly, semiannual, annual forecasts for one step forward and determine the upper and lower limits in the intervals between the forecasts.

3 Forecast correction in the forecast interval

During the forecasting interval the changes caused both by global environmental events: catastrophes, accidents, crises, etc., and changes in the structure of supply and demand as a result of events related to a particular industry are possible. These changes require correction of the forecast and, accordingly, of the models forming it.

3.1 Cognitive-map-driven monitoring of the current situation and scenarios simulation

Cognitive-map driven monitoring is based on monitoring information sources, where analytical materials on the situation are concentrated. In addition to the use of analysis in the monitoring cycle, when the analytical department generates reasonable conclusions about the significance of the observed changes in the external environment, analysis on the cognitive map allows you to create a semantic observation model in information sources on important factors and topics. Such monitoring includes:

1) *Situation development monitoring* by tracking information sources by keywords chosen from CM and determining of factors' initial values in an appropriate linguistic scale. The IT-infrastructure is configured to monitor the information space based on the factors of the model. A special category is created for each factor and the user/analyst receives a notification in case a message regarding a certain factor appears in mass media or any other connected source of information.

Let, $S_y^d(X(0);G(0);K_f;R(r_j = sign(x_j(0)))$ – the d-th qualitative forecast on the map K_f, depending on the values of factors X at the time $t = 0$, is determined by Eq. 1. That is, in each scenario d, we obtain the value of the factor y^* at any discrete point of time $t \geq 0$ and an estimate of the dynamics $r_y = sign\left(y^* - y^0\right)$.

2) Assessment of the significance of information occasions on the situation model

In the block of qualitative forecasting of a target indicator for given input conditions for factors X, in addition to assessing their value, an assessment is made of the significance in the change of certain factors-reasons. For this simulation scenario, a Q^R sub-matrix is formed on the Q, in which the factors of the $X_{R(y)}^{inpS}(x_i : edf_i \neq 0)$ *and those* $x_R^{M_i} : (edf_i \neq 0)$ represented in the rows and the factors $\{ X_R^{M_i} \}$ and the target indicator y in columns; on the intersections of rows and columns – the corresponding integral influence q_{jk}, if there is no influence, then $q_{ij} = 0$.

Then a significance of the M_i for some scenario S_y^d defined as

$$q_{S_y^d}^{M_i} = \sum_{k1=1}^{m1} \sum_{j1=1}^{l1} r_{x_{k1}} \times q_{x_{k1}}^{x_{j1}^{M_i}} + \sum_{j2=1}^{l2} r_{x_{j2}^{M_i}} \times q_{x_{j2}^{M_i}}^{y^{M_i}}, \tag{3}$$

where the first addend is the cumulative influence of the factors from $X^{X^{M_i}}$ on causal factors related to a model M_i on target indicator, *y,* and the second addend is the cumulative influence of the causal factors related to M_i on target indicator *y*.

Accordingly, for each qualitative forecast $S_y^d(X(0);K;EDF(edf_i = sign(x_i(0)))$, we obtain the following estimates for the forecasting models $\{M_i, (S_{x_k}^d; \{q_{S_{x_k}^d}^{M_i}\}; edf_{x_k})\}$. Also, we can obtain an estimate of a forecasting model M_i in the form

$$q{*}_{S_y^d}^{M_i} = \sum_{k2=1}^{m2} r_{x_{k2}^{other_i}} \times q_{x_{k2}^{other}}^{y^{M_i}},$$

which estimates a cumulative effect of the causal factors has not been connected with factors from X^{M_i}.

Systematically monitoring of such situation by means of IT-tools is needed to use a simulation module for dynamic analysis and prediction. The software tools should provide the collection, processing and consolidation of heterogeneous unstructured data - text and rich media - from internal and external sources (databases, the Internet, file systems, corporate information systems, television and radio broadcasting, etc.) in a close to real-time mode. A fundamental feature of the system is that it supports a full cycle of data processing, i.e. transform data into information and extract actionable knowledge from information through in-depth text analysis and situation modeling. [14].

3.2 Monitoring the dynamics of quantitative indicators

If changes occur in the prediction interval, the models used stop describing the incoming data. These changes against the background of random perturbations are little noticeable and without the use of special algorithms they are detected with a large lag. Therefore, for monitoring, we use algorithms aimed at the fastest detection of changes, subject to the limitation of time between false detections, capable of dealing with non-stationary processes [15-17].

We conduct digital monitoring using weekly data. To perform it, we build on the interval containing no changes in the time series models of the target indicator and significant factors and calculate the errors of the models under stable conditions:

$$Res_t = \Delta y_t - \tilde{\mu} - \sum_{i=1}^{k-1} \tilde{\beta}_i \, \Delta y_{t-i} \quad (4)$$

where Δy_t are the differences of weekly values of the monitored series,, $\tilde{\mu} - \sum_{i=1}^{k-1} \tilde{\beta}_i \, \Delta y_{t-i}$ are the parameters of its model description built on the differences of weekly data.

If the drift $\tilde{\mu}$ has changed in the tracked indicator, then the conditional mathematical expectation of the sequence (4) changes, and the asymptotic properties of the conditional variance remain unchanged, if the volatility of the process has changed, then the variance of sequence (4) changes [15].

For change detection we use the algorithm of abrupt change diagnosis (detection and isolation) in random signals proposed in [16]. This algorithm detects changes in random signals in the presence of multiple alternatives and isolating the change that occurred. Let $\Delta y_1, \Delta y_2, \ldots, \Delta y_{t0}, \Delta y_{t0+1}, \ldots$, be a random independent sequence distributed as

$$F(\Delta y_t, \theta) = \begin{cases} F(\Delta y_t, \theta_0), & if < t0 \\ F(\Delta y_t, \theta_l), & if \geq t0 \end{cases}$$

where $l = 1,2, \ldots, m$, $F(\Delta y_t, \theta_l)$, is a family of distributions differing in their parameters: θ_i, $\theta_i \neq \theta_j$, if $j \neq i$ for all i, j=0,1,2,…,m. The change time t_0 and the distribution index i after changing properties are unknown (but nonrandom). As a result, the algorithm generates a signal t_0, l , where t_0 is the alarm time at which the change was detected, and $l \in \{1,2,\ldots,m\}$ is the type of change in accordance with (4). This algorithm minimizes of the maximum mean delay for detection/isolation [16].

4 Description of the experiment and results

Realization results of the proposed approach have been demonstration on example of building a multi-step forecast of scrap prices in commodity markets. In the interval 2015-2018, we used monthly data to build VEC and VAR models and form a multistep forecast for the year ahead (2019) with a monthly breakdown, taking into account the recommendations of the cognitive map for the composition and assessment of the significance of causal factors and the results of analyzing the properties of temporal series describing these factors (sections 2.1 - 2.2).

Monitoring of the situation revealed a change in demand for metal products in Europe in March 2019 and the introduction of US duties for Russia. Based on these data, the qualitative forecast for the CM (SM scenario) showed that the significance of the quantitative forecast for a group of factors, reflecting the demand for finished products, is the highest. In August, it was discovered: the absence of causality of factors relative to the target in some models and, consequently the violation of cointegration in its.

Based on the results of the detected changes, the group of models that form prices for scrap by factors reflecting the demand for finished products was assigned the maximum weight, and the models with violation of cointegration was replaced with other models of the corresponding groups. According to the data for 2015-2019, the forecast for 2020 was carried out in a similar way.

To check the results of the algorithm, we used, as in [17], two measures of accuracy: the mean absolute percent error (MAPE) and the root mean squared error (RMSE):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_t-\hat{Y}_{t|t-h}}{y_t}\right| \tag{5}$$

and

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_t-\hat{Y}_{t|t-h}\right)^2} \tag{6}$$

where y_t is the value target at the moment t, $\hat{Y}_{t|t-h}$ is the forecast at the moment of time t, calculated with the horizon h at the moment of time $t-h$.

To assess the algorithm performance, we used metrics (5) and (6) obtained by forecasting a number of scrap prices for a year with a monthly breakdown for the interval of 2016-2019 (in dollars). For comparison, we calculated forecasting errors by the following algorithms: naive forecasting, forecasting by ARIMA model, forecasting using the proposed algorithm without taking into account the recommendations of the cognitive map. The results are given in Table 1.

To assess the algorithm performance, we used metrics (5) and (6) obtained by forecasting monthly scrap prices time series for a year with a monthly breakdown in 2019 and 2020 years (in dollars). For comparison, we calculated forecasting errors by the following algorithms: naive forecasting, forecasting by ARIMA model, forecasting using the proposed algorithm without taking into account the recommendations of the cognitive map. The results are given in Table 1.

Notes: alg1 is the proposed algorithm without taking into account the results of analysis by the control map; alg2 is the algorithm that corrects alg1 according to the results of analysis by the cognitive map.

Table 1. Average error of the annual forecast with a monthly split

Algorithm	MAPE 2019	MAPE 2020	RMSE($)	RMSE($)
alg1	4,2	3,5	10,29	10,2
alg2	6,8	5,8	16,8	15,8
ARIMA	12,95	13,5	38,57	39,5
naïve	12,02	14,84	35,14	37,14d

The experimental results presented here demonstrate the advantage of our approach.

References

1. J. P. Byrne, G. Fazio, and N. Fiess, Primary Commodity Prices: Co-movements, Common Factors and Fundamentals. Journal of Development Economics 101, 16-26 (2013).
2. C. Liu, X. Sun., J. Wang, J. Li, J. Chen, Multiscale information transmission between commodity markets: An EMD. Research in International Business and Finance 55(1), 101318 (2021).
3. L. Tanga, C. Zhanga, L. Lib, S. Wang, A multi-scale method for forecasting oil price with multi-factor search engine data. Applied Energy 257(1), 114033 (2020)
4. X. Zhang, L. Yu a, S. Wang, Kin Keung Lai, Estimating the impact of extreme events on crude oil price: An EMD-based event analysis method. Energy Economics 31, 768-778 (2009)
5. L Lin, Y Jiang, H Xiao, Z Zhou Crude oil price forecasting based on a novel hybrid long memory GARCH-M and wavelet analysis model. Physica A: Statistical Mechanics and its application 543(1), 123532 (2020).
6. E.I., Papageorgiou, K., Poczeta "A two-stage model for time series prediction based on fuzzy cognitive maps and neural networks. Neurocomputing. 232, 113-121 (2017).
7. D., Shan, W., Lu, and J., Yang. The data-driven fuzzy cognitive map model and its application to prediction of time series. International Journal of Innovative Computing, Information and Control 14(5), 1583-1602 (2018).
8. W., Froelich, and J., Salmeron. Evolutionary learning of fuzzy grey cognitive maps for the forecasting of multivariate, interval-valued time series. Int. J. Approx. Reason 55, 1319-1335 . (2014).
9. W., Homendaa, and A., Jastrzebska. Clustering techniques for fuzzy cognitive map design for time series modeling. Neurocomputing 232, 3-15 (2017).
10. T., Hong, and I., Han. Integrated approach of cognitive maps and neural networks using qualitative information on the World Wide Web: the KBNMiner. Expert Systems 21(5), 243-252 (2004).
11. A Sangasoongsong, STS Bukkapatnam, J. Kim, P.S. Iyerc, R.P. Sureshc Multi-step sales forecasting in automotive industry based on structural relationship identification. International Journal of Production Economics 140(2), 875-887 (2012).
12. H Jiang, C Liu Forecasting construction demand: a vector error correction model with dummy variables. Construction Management and Economics 29(9), 969-979 (2011).
13. A. Parot, K. Michell, W.D. Kristjanpoller, Using Artificial Neural Networks to forecast Exchange Rate, including VAR - VECM residual analysis and prediction linear combination. Intelligent Systems 26(1), 3-15 2019.
14. Makarenko D. Analytical and Predictive Monitoring: a Cognitive-Map-Based Approach and Toolkit To Socio-Political Stability Governance Support. IFAC-PapersOnLine 51(30), 522-526 (2018).
15. Grebenyuk E.A. Monitoring and identification of structural shifts in processes with a unit root. Proceedings of the 13th International Conference (MLSD) Moscow, IEEE, https://ieeexplore.ieee.org/document/9247829 (2020).
16. I.V. Nikiforov, A simple recursive algorithm for diagnosis of abrupt changes in random signals. IEEE Transactions on Information Theory 46(7), 2740-2746 (2000).
17. Tang L, Dai W, Yu L. A novel CEEMD-based EELM ensemble learning paradigm for crude oil price forecasting. Int J Inf Technol Decis Mak 14, 141-69 (2015).