Framework for Automated Food Export Gain Forecasting

Dmitry Devyatkin [1] and Yulia Otmakhova [2]

[1] Federal Research Centre "Computer Science and Control" RAS, Moscow, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
[1]devyatkin@isa.ru
[2]otmakhovajs@yandex.ru

Abstract. The food and agriculture could be a driver of the economy in Russia if intensive growth factors were mainly used. In particular, it is necessary to adjust the food export structure to fit reality better. This problem implies long-term forecasting of the commodity combinations and export directions which could provide a persistent export gain in the future. Unfortunately, the existing solutions for food market forecasting tackle mainly with short-term prediction, whereas structural changes in a whole branch of an economy can last during years. Long-term food market forecasting is a tricky one because food markets are quite unstable and export values depend on a variety of different features.

The paper provides a multi-step data-driven framework which uses multimodal data from various databases to detect these commodities and export directions. We propose the quantile nonlinear autoregressive exogenous model together with pre-filtering to tackle with such long-term prediction tasks. The framework also considers textual information from mass-media to assess political risks related to prospective export directions. The experiments show that the proposed framework provides more accurate predictions then widely used ARIMA model. The expert validation of the obtained result confirms that the framework could be useful for export diversification.

Keywords: data-driven market forecasting, international trade, quantile regression, multimodal data

1 Introduction

Sanctions and trade confrontations set difficulties for persistent economic growth. The essential way to overcome them is making the economy more independent and diversified [1]. Due to limited resources, the efforts should be focused on a limited set of development directions. Therefore, the developing of a particular economy field implies discovering a restricted set of the new prospective commodity items and export directions. In this paper, we consider this problem in the case of food and agriculture field. Thus our aim consists in finding the pairs <*Trade partner, Agriculture_OR_FoodCommodity*> with a high probability of the persistent export value growth from a particular country (in our case – from Russia) in several next years.

More precisely, we predict summary export value gain in the following two years based on information about current and two past years. We believe that modern data-driven approaches could be useful to tackle this problem.

This goal is not trivial because of the following issues:

1. Unstable character of many trade flows.

2. Too many features influence on trade flows. If we used them all, it would lead to over-complex prediction models, which aren't trainable with the dataset. Long-term forecasting requires the using of complex models that consider a large number of features and parameters, but the size of the training dataset is strictly limited. Therefore, complex models can be easily overfitted and in some cases give incorrect results on unseen data.

3. Political decisions, economic sanctions strongly affect trade flows, but they hardly ever can be predicted using only statistic databases.

4. Existing regression and classification metrics such as MSE or F1-score poorly reflect the accuracy of the solution to the highlighted problem, since even a small ranking error can lead to the omission of a very profitable direction.

In this paper, we propose a data-driven framework which can mitigate the highlighted problems.

At first, we apply a quantile regression loss since it allows estimating the distribution parameter for the predicted value so that we can process unstable trade flows more accurately.

Secondly, we believe it is possible to mitigate the overfitting problem and instability problems both if one pre-filters pairs with high probabilities of a decline in the future. This can be done with training a binary classifier, which is much simpler than regression and can be performed using simpler models which are not overfitted. Then the "large" errors of the regression model will have less impact on the final result. We also propose compositional features which can describe the market demand for a commodity item compactly to simplify the regression model.

Thirdly, we extract sentiment features from texts, more precisely, from news to assess political risks.

Finally, we suggest calculating ratios between the export value of the top predicted pairs and the export value of the actual top pairs with the highest export gain to assess the usefulness of the prediction.

The rest of the paper is organized as follows: in Section 2 we review related studies; in Sections 3 we present the proposed framework; in Section 4 we describe the results of the experimental evaluation; Section 5 contains conclusion and directions of the future work.

2 Related Work

The vastest branch of studies is devoted to short-term food market forecasting with basic regression and autoregression models. For example, Mor with colleagues propose linear regression and Holt–Winters' models to predict short-term demand for dairy products [2]. The more complex autoregressive integrated moving average

model (ARIMA) allows dealing with non-stationarity time series. This model also widely used for food market forecasting, for example, in [3] to forecast harvest prices based on past monthly modal prices of maize in particular states.

Ahumada et al. [4] proposed an equilibrium correction model for corn price. Firstly they use an independent model for each corn. Then they also observe whether the forecasting precision of individual price models can be improved by considering their cross-dependence. The results show that prediction quality can be improved using models that include price interactions. The multi-step approach is proposed in [5]. The researchers consider the balance between production and market capacity to be the key factor for trade flows forecasting.

For forecasting in volatile markets, it is necessary to reveal detail information about the distribution of the predicted variables, not their mean values only. Quantile regression is a common solution in this case. [6], [7]. For example, researchers [8] apply linear quantile loss to train Support Vector Regressor and use it to assess confidence intervals for predicted values. The paper [9] combines hybrid ARIMA and Quantile Regression (ARIMA-QR) approaches to construct high and low quantile predictions for non-stationary data. The obtained results show that the model yield better forecasts at out-sample data compared to baseline forecasting models.

Let us briefly highlight some studies related to features for food market forecasting which can better explain trade flow dynamic than trade and production values themselves. Paper [10] provides a conclusion that finance matters for export performance, as commodities with higher export-related financial needs disproportionately benefit from better economic development. Jaud with colleagues uses level of outstanding short-term credit and trade credit insurance, reported in the Global Development Finance and Getting Credit Index (EGCc) from the World Bank Doing Business Survey as features related to the level of financial development.

Political factors also influence on the food market. Makombe with colleagues studies the relationship between export bans and food market [11]. The researchers conclude that the prohibitions cause market uncertainty which may have long-run implications for future food security and trade flows. The critical problem here lies in uncertainty in the way how to formalize and consider these factors in models. It is well-known fact, that political decisions often follow by outbursts in mass media, so one can easily predict possible political decisions if he or she analyses the new sentiment. This idea is widely used for short-term analysis in financial markets [12], thus we believe it could be helpful in the proposed framework.

Long-term export prediction assumes considering arbitrary dependencies between the model outcome and the lots of factors in the past. Duration of these dependencies can vary from single days for price movements to dozens of years for political decisions or climate changing. The mentioned approaches cannot model linear, non-linear dependencies and consider a broad set of sophisticated features at the same time though. A natural way to model such complex features and dependencies is to use neural network framework. Pannakkong with colleagues [13] uses a dense multilayer feed-forward network and ARIMA to forecast cassava starch export value. The results show that feed-forward neural network models overcome the ARIMA models in all datasets. Hence, the neural network models can predict the cassava starch exports

with higher accuracy than the baseline statistical forecasting method such as the ARIMA. However, such a simple architecture cannot model long-term interaction.

There are particular network architectures for long-term prediction. In [14], researchers suggested the nonlinear autoregressive exogenous model (NARX) artificial network architecture for market forecasting. They proposed a feed-forward Time Delay Neural Network, i.e. the network without the feedback loop of delayed outputs, which could reduce its predictive performance. The main benefit of the model compared to model compositions is the ability of joint training of linear and non-linear parts of models. Similarly, in [15] authors proved that the generalized regression neural network with fruit fly optimization algorithm (FOA) is effective for forecasting of the non-linear processes.

Unfortunately, neural network approach often leads to inadequately complex models which are needed large datasets to be reliably fitted. We have relatively small dataset thus it is required to find the most straightforward network architecture and tightest feature set which however could achieve satisfactory forecast accuracy.

Because food market is volatile, it would be helpful if the forecasting model provided more information about predicted variables as quantile regression does. Although there are few works in which quantile regression-like loss function was used for training neural networks [16].

The methodological aspects of creating models for export forecasting require further study. Existing models consider some important indicators, but they can be based on erroneous assumptions that cast doubt on the obtained results. For example, the predictive model for assessing the country's diversification of exports provided in the Atlas of Economic Complexity (Feasibility charts) [17]. This model predicts a very curious output, namely that tropical palm oil could be one of the products for diversifying Russia's exports. This is due to the neglecting country's climatic and infrastructure capabilities. That is why the feature set is still not obvious for this problem.

The review shows that the most applicable solution for the food export gain forecasting is to combine long-term prediction models, such as NARX and quantile loss functions. In addition to basic features such as trade flows and production levels, these models should consider heterogeneous macroeconomic and climate indicators. Since the addition of political factors would complicate the regression model, it makes sense to consider them separately. That is, after the regression, we filter obtained export directions if they are related to high political risks.

3 Framework for Export Gain Forecasting

As a test dataset for the framework, we use annual information about trade flows (from UN Comtrade [18]), production values (UN FAOSTAT [19]) and macroeconomic indicators (International Monetary Foundation [20]). We consider the following macroeconomic features: state GDP, inflation level, population level etc. Due to heuristic reasons the dataset includes only the items which are produced in Russia and presented in its trade flows, so the final dataset contains 70 export directions and 50 commodities. We also do not consider records earlier than 2009, because the interna-

tional financial crisis could lead to changes, which we cannot model adequately. Daily climate (temperature, wind speed, humidity, pressure) features were downloaded from RP-5 weather database [21]. The highest, the lowest and average values for each season were calculated, because the time step of the framework is one year. We also used open-available Russian news corpus from Kaggle [22].

It is no doubt to say that trade flows between particular country and its partners depends on trade flows between these partners and the other countries. Unfortunately, if we added all these features directly to the regression model, the model would become too complex and would tend to overfit in the dataset. We propose the *SPR* (Substantial PRoduct) composite features to resolve this problem. The *SPR* shows contribution of an arbitrary exported commodity item from Russia on the global demand satisfying (expression (1)):

$$SPR_i = \frac{X_i}{\sum_{i \in I, j \in D} C_{ij}}. \quad (1)$$

Here I is a set of leading export commodities, D is a set of export directions, X_i is total Russian export value for commodity i, C_{ij} is export value from Russia to country j for commodity i. We consider all trade flows between Russia and its' partners directly and encode the rest flows with the *SPR* features. The comprehensive feature list is presented in Table 1.

Table 1. Feature set for export gain forecasting

Group of features	**Frequency**	**For**	**Features**
Trade flows	Annual	Country, Commodity	Export value
			Import value
			Re-export value
			Re-import value
SPR	Annual	Commodity	SPR
Production	Annual	Country, Commodity	Production value
Macro-economic indicators	Annual	Country	Trade balance
			GDP
			Inflation (CPI)
			Inflation (PPI)
			Population
			Purchasing power parity (PPP)
			Unemployment rate
Climate indicators	Per season: max, min, median	City, town	Temperature
			Humidity
			Wind speed
			Precipitation
			Pressure
			Cloudiness

We propose the following framework for prediction of the promising pairs <*export item,direction*> (Fig. 1). The framework contains regression step and several filtering steps. Pre- and post-filtering steps are proposed to deal with the trade flows instability.

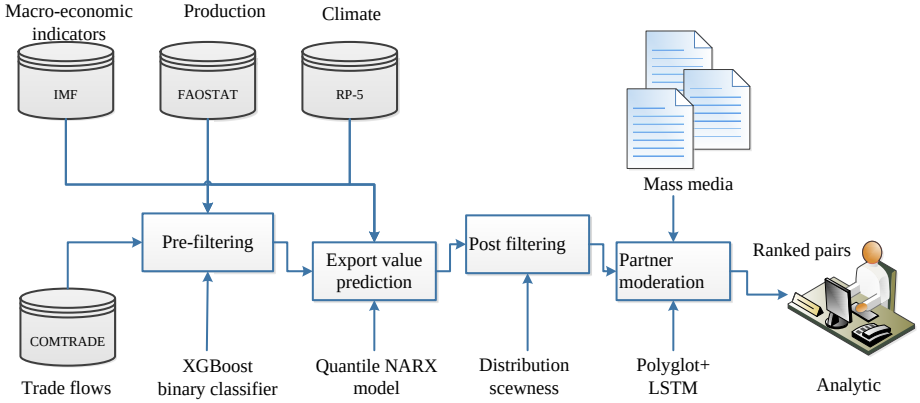

Fig. 1. Framework for prospective export pairs prediction

At first step of the framework we detect pairs which likely tend to decrease. On the one hand the filtering model should be much simpler, than the regression model, but on the other hand it should learn complex non-linear dependencies. The most appropriate approach in this case is decision tree ensembles. We tested several methods such as Random Forest [23], Gradient Boosting [24] and XGBoost [25] to fit these ensembles. The next two steps we realize with the modified NARX quantile regression model (Fig. 2). A single model is used for all directions and commodities since the use of individual models can lead to the loss of information about the interaction between the export value for commodities. We used the following loss function instead of mean squared error to obtain quantile NARX model:

$$L(\omega,\theta) = \sum_{t \vee y_{t \geq f(x_t,\omega)}} \theta(y_t - f(x_t,\omega)) + \sum_{t \vee y_{t < f(x_t,\omega)}} (1-\theta)(y_t - f(x_t,\omega)), \quad (2)$$

here θ is quantile level, x_t is features for time t, $f(x_t,\omega)$ is network output for time t and ω is parameters of the network.

This function was firstly introduced in [16]; it is a direct application of quantile regression [6] for networks training. Thanks to error-backpropagation framework the network architecture does not have any affection on the function (2). The modified NARX model allows predicting values for different quantile levels. We predict export flows with quantile levels 0.25, 05 and 0.75 and assessed skewness of the results. Than pairs with positive distribution skew are filtered. We also applied the Autoregressive Integrated Moving Average (ARIMA) model as a baseline.

In the last step, we filter unreliable trade partners with various models for sentiment analysis. We tested two neural network models, namely Attention-based Long-Short Term Memory (LSTM) [26] and Contextual LSTM [27]. Polyglot sentiment analyzer was also tested as a baseline. We used the Kaggle corpus with more than

10K news reports in Russian to train these models. Post-filtering itself consists of two parts. At first, we apply the Polyglot library [28] together with country name dictionary to extract news, mentioned Russia and some other country together. Then we apply a sentiment analyzer and filter trade partners with highly negative sentiment scores from the results.

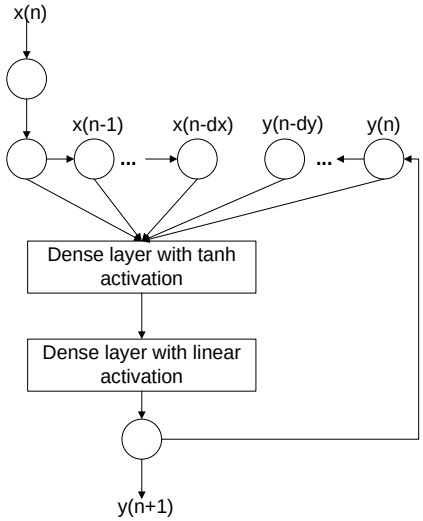

Fig. 2. Nonlinear autoregressive exogenous model for export value forecasting

4 Results and Discussion

We evaluated the proposed framework as well as its crucial parts. At first, the classification performance for the filtering step was evaluated. We tested several decision tree ensembles (Random Forest, Gradient Boosting, XGBoost) and Linear Support Vector Machine (SVM) classifier as a baseline. The XGBoost method revealed the best accuracy in cross-validation, so we add it to the framework. It is worth to note that the filtering itself can be done with relatively high quality (about 73% F1 on 5-fold cross-validation, see Table 2, “filtering” column) using a simple feature-set because this task is much simpler than the whole regression task.

Table 2. Evaluation results for filtering methods

Method for filtering	F_1-binary
Random Forest	0.64±0.04
Gradient Boosting	0.59±0.02
XGBoost	0.73±0.05
LinearSVM	0.63±0.02

We also assessed the importance of different types of features for the filtering. The classifier was trained and tested on modified feature sets, in which distinct group of features had been omitted (see Table 3, "filtering" column). This column contains difference between binary F_1 score obtained on the full feature set and the score obtained on clipped feature set. The higher the F_1 score drop is, the more important related subset of features is. Results show that the most important features are SPR, climate and macro-economic indicators.

Table 3. Importance of the particular feature groups for the pre-filtering and for regression results

	Filtering	Regression
Group of features	**ΔF_1**	**ΔPredicted export value gain, in %**
Trade flows	0.05	7.1
SPR	0.17	6.8
Production	0.12	1.2
Macro-economic indicators	0.29	24.7
Climate indicators	0.05	17.5

Then we studied the importance of the different types of features for the regression. As for filtering step, we separated features into distinct subsets and trained the regressor with them (see Table 3, "regression" column). The "Predicted export value gain" here and in the next tables means ratios between the export value of the top-10 predicted pairs and export value of the top-10 actual pairs with the highest export gain. The "ΔPredicted export value gain" column contains difference between the gain obtained on the full feature set and the gain obtained on clipped feature set. The results showed that the most significant features are macro-economic indicators, climate and past export flows. This confirms the limitation of models which do not consider these features, for example [17].

Table 4. Importance of the particular filtering steps

Filtering	Predicted export value gain, %
Without	12.5
Pre-filtering (only)	38.9
Post-filtering (only)	22.7
All (combined)	61.6

We also evaluated the contribution of the filtering steps to the results. The filtering steps together help significantly improve the obtained results, as one can conclude from Table 4. These steps allow removing the pairs with the highest decline risk.

Table 5 contains results for considered sentiment analysis methods. We evaluated these methods on the test subset of the Kaggle sentiment dataset. Attention LSTM model shows slightly better result on this task, so we added it to the proposed framework.

Table 5. Results for the sentiment analysis

Method	**F_1-macro**
Polyglot sentiment	0.62
CLSTM	0.79
Attention LSTM	0.82

Fig. 3 shows average sentiment for mass-media news related to main trade partners of Russia, which are often mentioned in Russian news. The news was gathered from Lenta.ru dataset [29], because the timeline of the Kaggle dataset is not appropriate. We filtered news, contained both "Russia" and other country names and evaluated average sentiment for them with the Attention LSTM model. It's easy to note that unreliable partners get lower marks.

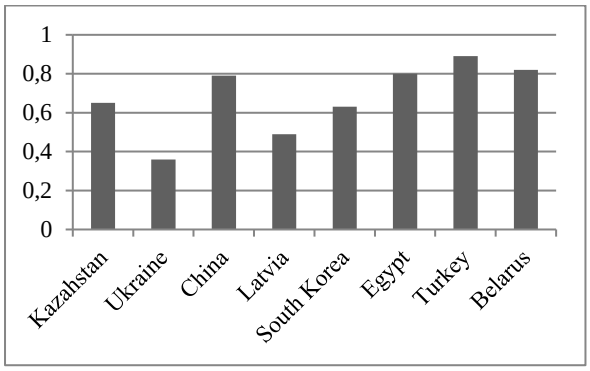

Fig. 3. Partner assessment with sentiment analysis

We tested the overall framework on retrospective data, more precisely records from 2009 to 2014 were used to train and other data (2015–2016) we left for the evaluation. The detailed results for the whole framework are presented in Table 6. The "Actual" column contains ranked pairs with the highest average export gain in 2015–2016 for Russian Federation. Summary average gain for the top 10 pairs amounted to 1.5 billion USD. The "Predicted" columns contain results of the forecasting.

The economic analysis of the detailed results showed that the list of the partners with the highest summary export gain did not match with the list of top importers for the study period.

The proposed NARX model allows predicting the most growing export commodities quite precisely. Linseed is the only mismatched position, but it reflects a new

prospective market. Moreover, linseed production for export has been strongly supported by the Russian government since 2016. Thereby the model detected this potential market with past data and predicted that decision.

Table 6. Detailed results of the export value prediction

		Predicted			
Actual		**ARIMA**		**NARX + pre-filtering + quantile-filtering**	
Partner	Commodity	Partner	Commodity	Partner	Commodity
Bangladesh	Wheat	Egypt	Wheat	Egypt	Wheat
Egypt	Wheat	Bangladesh	Wheat	Saudi Arabia	Barley
China	Soybeans	China	Soybeans	Nigeria	Wheat
Nigeria	Wheat	Turkey	Wheat	Morocco	Wheat
China	Oil of Sunflower Seed	China	Oil of Sunflower Seed	Sudan	Maize
Rep. of Korea	Maize	Algeria	Oil of Soybeans	Turkey	Barley
Lebanon	Wheat	Azerbaijan	Wheat	Turkey	Linseed
Algeria	Oil of Soybeans	Saudi Arabia	Barley	Bangladesh	Wheat
Saudi Arabia	Barley	Lebanon	Wheat	Italy	Wheat
China	Oil of Soybeans	China	Oil of Soybeans	China	Oil of Soybeans
Export value gain, M USD	1474.4; 100%	674.8; 45.7%		908.65; 61.6%	

The most often cause of the NARX model errors is neglecting features, related to technological development. However, the appearance of new technologies leads to dramatic changes in the markets. New deep-processed commodities appear, and prices for existing raw products can decline, which leads to an export value drop for traditional providers. Existing counterparties (Turkey, for example) may switch to other commodities. Therefore, there is a need to add technological features to the model, which would make it possible to predict prospect commodities with an assessment of the related technologies for primary and deep processing.

To sum up, the results of the proposed framework could be useful for export diversification since NARX model provides new prospective commodity items. The variants of the NARX model and ARIMA are also helpful for counterparty countries exploration.

5 Conclusion

In this paper, we propose a data-driven framework for food export gain forecasting. The framework considers multimodal open data from many data sources and corpora. In this research, we tried to mitigate the set of problems, related to machine forecasting of food export gain: large feature set dimension, volatility of markets, factors which are difficult to formalize (political risks).

In the experiments, we used open data from FAOSTAT, UN Comtrade, information about global economic situation from International Monetary Foundation, climate information and reports from news corpora. According to the results, quantile loss function and NARX model is a promising combination for long-term prediction of trade flows for food commodities.

In the future research we plan to consider logistical and infrastructure conditions as well as technological features in the framework. The next steps of our research also include detailed analysis of the obtained commodity items and finding technologies which could help to push the export for these commodities up.

Acknowledgements

The project is supported by the Russian Foundation for Basic Research, project No. 16-29-12877 “ofi_m”.

References

1. Awokuse, T.: Does agriculture really matter for economic growth in developing countries? In: 2009 Annual Meeting. Agricultural and Applied Economics Association, vol. 49762. Milwaukee, Wisconsin (2009).
2. Mor, R. and Bhardwaj, A.: Demand forecasting of the short-lifecycle dairy products. In: Chahal, H., Jyoti, J., Wirtz, J. (eds.) Understanding the Role of Business Analytics, pp. 87–117. Springer, Singapore (2019).
3. Darekar, A. and Reddy, A.: Price forecasting of maize in major states. Maize Journal **6** (1&2), 1–5 (2017).
4. Ahumada, H. and Cornejo, M.: Forecasting food prices: The case of corn, soybeans and wheat. International Journal of Forecasting **32** (3), 838–848 (2016).
5. Burlankov, S., Ananiev, M., Gazhur, A., Sedova, N., and Ananieva, O.: Forecasting the development of agricultural production in the context of food security. Scientific Papers Series-Management, Economic Engineering in Agriculture and Rural Development **18** (3), 45–51 (2018).
6. Koenker, R. and Hallock, K.: Quantile regression. Journal of economic perspectives **15** (4), 143–156 (2001).
7. Maciejowska, K., Nowotarski, J., and Weron, R.: Probabilistic forecasting of electricity spot prices using Factor Quantile Regression Averaging. International Journal of Forecasting **32** (3), 957–965 (2016).
8. Li, G. Xu, S., Li, Z., Sun, Y., and Dong, X.: Using quantile regression approach to analyze price movements of agricultural products in China. Journal of Integrative Agriculture **11** (4), 674–683 (2012).

9. Arunraj N. and Ahrens D.: A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. International Journal of Production Economics **170**, 321–335 (2015).
10. Jaud, M., Kukenova, M., and Strieborny, M.: Financial Development and Sustainable Exports: Evidence from Firm-product Data. The World Economy **38** (7), 1090–1114 (2015).
11. Makombe, W. and Kropp, J.: The effects of Tanzanian maize export bans on producers' welfare and food security. In: Selected Paper prepared for presentation at the Agricultural & Applied Economics Association, vol. 333-2016-14428. Boston, MA (2016).
12. Nassirtoussi, A., Aghabozorgi, S., Yuing Wah, T., and Chek Ling Ngo, D.: Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. Expert Systems with Applications **42** (1), 306–324 (2015).
13. Pannakkong, W., Huynh, V., and Sriboonchitta S.: ARIMA versus artificial neural network for Thailand's cassava starch export forecasting. Causal Inference in Econometrics, pp. 255–277. Springer, Cham (2016).
14. Menezes, Jr J. M. P. and Barreto, G.: A. Long-term time series prediction with the NARX network: An empirical evaluation. Neurocomputing **71** (16–18), 3335–3343 (2008).
15. Li, H., Guo, S., and Sun, J.: A hybrid annual power load forecasting model based on generalized regression neural network with fruit fly optimization algorithm. Knowledge-Based Systems **37**, 378–387 (2013).
16. Taylor, J.W.: A quantile regression neural network approach to estimating the conditional density of multiperiod returns. Journal of Forecasting **19** (4), 299–311 (2000).
17. The atlas of economic complexity, http://atlas.cid.harvard.edu, last accessed 2019/07/01
18. UN Comtrade: International Trade Statistics, https://comtrade.un.org/data/, last accessed 2019/04/28
19. Food and Agriculture Organization of the United Nations, http://www.fao.org/faostat/en/ last accessed 2019/04/28
20. International monetary foundation, http://www.imf.org/en/Data, last accessed 2019/04/28
21. RP5 weather archive, http://rp5.ru, last accessed 2019/04/28.
22. Kaggle Russain news dataset for sentiment analysis, https://www.kaggle.com/c/sentiment-analysis-in-russian/overview, last accessed 2019/04/28
23. Breiman, L.: Random forests. Machine learning **45** (1), 5–32 (2001).
24. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189–1232 (2001).
25. Chen, T. and Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining, pp. 785–794. ACM (2016).
26. Wang, Y., Huang M., Zhao L., and Zhu X.: Attention-based LSTM for aspect-level sentiment classification. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp. 606–615. Association for Computational Linguistics, Austin, Texas (2016).
27. Ghosh, S., Vinyals, O., and Strope B.: Contextual LSTM (CLSTM) models for large scale NLP tasks. In: arXiv preprint arXiv:1602.06291. ACM (2016).
28. Al-Rfou, R., Kulkarni V., and Perozzi, B.: Polyglot-NER: Massive multilingual named entity recognition. In: Proceedings of the 2015 SIAM International Conference on Data Mining, pp. 586–594. Society for Industrial and Applied Mathematics (2015).
29. Lenta.ru Russian news dataset, https://github.com/yutkin/Lenta.Ru-News-Dataset, last accessed 2019/04/28.