Towards Framework for Discovery of Export Growth Points

© Dmitry Devyatkin[1] © Roman Suvorov[1] © Ilya Tikhomitov [1] © Yulia Otmakhova [2]

[1]Federal Research Center Computer Science and Control of the Russian Academy of Sciences,
Moscow, Russia
[2]Novosibirsk State University,
Novosibirsk, Russia

devyatkin@isa.ru rsuvorov@isa.ru tih@isa.ru otmakhovajs@yandex.ru

Abstract. Export value of the Russian Federation has been reducing in the latest years, as well as the corresponding relative yield. Most probably, this trend is caused by Russia total export decline together with growth of food export. Thus, it is very important to not only increase export volumes, but also adjust export structure to fit nowadays reality better. The paper presents a computer-aided framework for export growth points discovery. While the full framework is described briefly, more attention is paid to the first sub-task: growth point candidates ranking. The objective of this sub-task is to reveal combinations of commodities and partner countries with high probability of successful export. The method uses open data about international trade flows and production from United Nations databases and modern machine learning methods. The experimental evaluation shows that taking into account retrospective data allows ranking growth point candidates significantly better. Finally, the limitations and the possible directions of future research are discussed.

Keywords: export growth potential, data mining, international trade, customs statistics, open data, machine learning.

1 Introduction

Sanctions pose both difficulties and opportunities for the Russian economy. On the one hand, traditional foreign markets may be restricted or their growth potential may be exhausted. On another hand, exploring new markets may become a fruitful workaround. We believe that modern big data and machine learning technologies should be useful to discover new foreign markets with high probability of growth in the nearest future. We will refer to the pairs of countries and commodities as potential *growth points*. This paper aims on making a step towards finding new growth points using machine learning and open data analysis.

Authors of [1] consider export growth potential as an opportunity to meet the primary demand for a certain product or service. At the same time, the possibility to satisfy the demand arises locally and has a specific territorial, and, therefore, national binding.

There are two possible ways to satisfy growing demands: extensive and intensive. Intensive way implies improving technologies, scientific and engineering solutions and increasing the resource potential and efficiency of management. Therefore, a product may have high export growth potential if it has high added value, robust interbranch relations and stable external demand. In this paper, we propose a framework for discovery of “export growth points”. High-level procedure of this framework consists of two main steps: (1) finding candidates for “growth points”; (2) assessing each candidate and discovering difficulties with its implementation. The first step consists in ranking pairs <commodity, foreign market> in such a way so most likely growing pairs appear in the beginning of the list.

In this paper we propose a machine-learning-based method that ranks the “growth point” candidates using features, extracted from historical data from FAOSTAT and UN Comtrade databases [2, 3]. The presented evaluation is preliminary, because it is based on retrospective data. We understand such a weakness and we are going to address it in the future work.

The rest of the paper is organized as follows: in the Section 2 we review the most related works published so far; in Sections 3 and 4 we briefly describe our framework and the task of export growth point candidates ranking; in Section 5 we describe our dataset and present the results of experimental evaluation; in Section 5 we conclude and discuss future work.

2 Related work

Most commonly used approaches to foreign trade modeling include: gravity models, computable general equilibrium models, heuristic ranking models, Markovian models, common statistical approaches (regressions, histograms) for manual analysis of a situation.

The paper [4] presents the empirical evaluation of spatial gravity model of Russian trade. The authors concluded that the spatial variables such as the location of the state border checkpoints have a significant effect on the volume and routes of Russian imports. In [5] authors study factors of export and import value-added trade and suggest some recommendations for management of industrial and trade policy. The techniques proposed in this paper allow to determine main directions of economic policy to expand exports

Proceedings of the XIX International Conference “Data Analytics and Management in Data Intensive Domains” (DAMDID/RCDL’2017), Moscow, Russia, October 10–13, 2017

and improve Russian production structure. Duenas and Fagiolo in their paper [6] concluded that gravity models are poorly suited to predicting the presence of trade relations between some two countries. However such models allow us to accurately estimate and forecast the volume, given the knowledge that such trade relation exists. In [7] researchers use gravity models to investigate the export destinations that could be effectively developed with internal financial support. Experimental work was carried out on the data of food export at the firm-level.

In [8] authors consider Markov models for forecasting the variability of the network of foreign trade financial flows. In [9] an approach for detecting promising areas of export in the sector of both service and goods is proposed. The approach is based on the sequential filtering of potential markets via a number of heuristics, including estimation of the market volume, a level of demand, market openness, etc. In [10] authors studied the relationships between migration flows and foreign trade. They concluded that the trade flows for some products are positively and significantly correlated with migration flows. That feature can be taken into account during analyzing and evaluating the prospects of an export.

In [11] Lall et al. investigated relationship between exports volume and the "complexity" of goods and introduced a metric of "complexity" or "manufacturability" of goods. They mentioned the dependence between the rate of growth of prices on a product and the degree of it manufacturability. This dependence can be used as one of the features for detecting and assessing the export growth potential. Bernard et al. [12] proposed a method for estimating the feasibility of entering the international market for a particular company. They used indicators of the company past activity, including participation in exports, a competitive environment, etc. It is worth noting the weak influence of sectoral state support for exports on the actual volume of exports. In [13] authors considered the relationship of the topology of the international trade network between countries in general with network topologies within each product group. They proposed a methodology for studying the dynamics of changing the structure of several heterogeneous networks that represent trade flows between countries for individual commodity groups. As a result, the most active exporters and importers were detected for separate groups.

In [14] authors try to model the structure and dynamics of the international trade network using the classical methods for solving selecting balls from urns problem. The analysis is carried out at the level of countries and the principle of preferential attachment is implemented ("the rich get richer, the poor get poorer").

In [15] authors propose to model the structure and dynamics of the International Trade Network via the Hamiltonian system. The authors describe the dynamics of the International Trade Network in terms of Hamiltonian, and also make the assumption that the main provisions from the field of statistical physics will also be applicable to modeling the International Trade Network.

Shen et al [16] considered the international trade network at the level of countries and goods. They used flow analysis in graphs and statistics on tops to study the network. The authors draw a number of conclusions related to the specialization of countries, as well as the dominance of developed countries in terms of the diversity of exported products (the principle of preferential accession).

They empirically confirm the fact that food products are mostly traded between the most closely located countries, while high-tech goods are distributed virtually all over the world. Also, the authors detect countries with an anomalous profile of imports, which can talk about a number of economic problems. In [17] authors presented the analysis of export in the service sector on the example of Germany companies. The main goal of the analysis is to determine the dependence of directions and the mode of export on the various features of exported services. They used a non-open dataset from Deutsche Bank. Among other things, the authors detected such heuristics as "exports are more preferable to countries with higher incomes (for countries with lower incomes, an international partnership is more preferable)"; "When selling in more remote countries, international partnership is more profitable."

In [18; 19] researchers developed machine learning models to forecast export dynamics of agricultural products. They compare Support Vector Machines (SVM) and Autoregressive Integrate Moving Average (ARIMA). The experiments showed that SVM achieves significantly smaller error rates.

To sum the review up, we can say that quite extensive efforts have been committed to analyze and predict international trade flows. However, most papers describe fragmentary studies, which are focused on a limited set of factors. Thus, a goal-oriented and comprehensive approach is in high demand.

3 Framework for discovering export growth points

In this section we will try to formalize the problem of export growth points discovery. The objective is to find combinations <*Product$_i$*, *Country$_j$*>, which have the highest unrealized potential for export growth. Also, production and export management of these combinations has to be feasible in the Russian Federation. *Product$_i$* is a product or product category to export and *Country$_j$* is a country or a group of countries to export to.

We propose to use open data analysis and modern machine learning techniques to find such growth points. The high-level algorithm of our framework consists of the following steps:

1. Construct a list of growth point candidates<*Product$_i$*, *Country$_j$*>. Reorder this list so the candidates with higher likelihood of becoming successful export direction appear earlier.
2. Analyze supply chains which contain commodities from our candidate list. Products

with higher added value should be reviewed first. Consider the product lifecycle (including production, storage, transportation and processing for the selected products) in order to detect the most probable difficulties for each stage of the lifecycle in the context of the Russian Federation. Propose intensive or extensive ways of overcoming them. Products with too many difficulties are removed from the list.

Novelty of our approach consists in maximum possible automation. We can automate step 1 (candidates ranking) and aid step 2. Ranking in Step 1 can be carried out with a predictive machine-learning based model. Step 2 can be highly facilitated by developing a specialized information retrieval system which uses big collections of scientific and engineering documents, such as patents, scientific papers, grant reports. Step 1 is discussed in detail later in this paper. We are going to consider step 2 in future.

4 Data Driven Candidates Ranking

Formally, the problem of candidates ranking is a Learning-To-Rank (LTR) problem. Traditionally, each LTR problem is specified by three components: a set of possible queries, a set of objects and a target metric to optimize. In this work each query is formulated as "Which products to which countries should we try to export to increase budget income, in the context of current macroeconomic situation and our state of industry?". In other words, a query is specified by current economic context (wide or narrow, depends on implementation). Objects that are ranked relative to that query are export growth point candidates or pairs <*Product$_i$*, *Country$_j$*> (what and where to export).

The main difficulty with LTR problem statement is target metric construction. This metric must reflect the likelihood of success if export of *Product$_i$* to *Country$_j$* from the Russian Federation will be established. Such a metric cannot be constructed in purely data-driven way, because no database of such cases exists. To overcome this issue, we propose to base on two sources of knowledge: (1) opinion of experts in the field of food market and international trade; (2) retrospective data about dynamics of international trade. On the one hand, retrospective data alone cannot be used to predict future, because the world context is changing and it will almost never become same again. On another hand, experts base on a limited number of factors and limited knowledge (it may be very deep but still limited). Thus, we propose to use experts to take into account factors which are hard to formalize; and retrospective data - to measure prior likelihood of trade flow of *Product$_i$* to *Country$_j$* to grow.

Taking into account expert opinion requires labeling a training dataset. In this paper we conduct preliminary studies only using retrospective data, due to limitations of time and resources. Experiments with manually annotated datasets will be considered in future.

In other words, in this paper we study only export dynamics prediction. One can dispute that LTR is a reasonable approach to this problem and claim that traditional regression is a better fit. We chose LTR due to three main reasons. The first one is that information about order is more abstract than information about exact increase of trade value or volume (and thus the corresponding predictive model should generalize better). The second reason is that we plan to use LTR in more general case and thus we want to conduct experiments as close to the proposed framework as possible. And the third reason is that we can generate more data to train LTR model and thus try to reduce overfitting.

To facilitate solution of the described LTR problem, we treat it as pairwise ranking problem: we build a regression model, which is given a pair of two export growth point candidates <*Product$_1$*, *Country$_1$*> and <*Product$_2$*, *Country$_2$*> returns a difference between export flows for the first and second pair. Generally, such a model operates on a feature set consisting of three major parts: description of global macroeconomic situation; description of trade flows for the first candidate; description of trade flows for the second candidate. Ideally, information about both candidates should also somehow describe prices, competitiveness, quality etc.

The objective of the experimental evaluation in this paper is to verify that retrospective data is useful to compare trade flow dynamics for different commodities and foreign markets. To achieve this goal, we applied ARIMA model as a baseline and also built two machine learning models: "baseline" and "advanced".

4.1 Dataset

We used excerpts from FAOSTAT [2] and UN Comtrade (Comstat) [3] databases from 2011 to 2015 years. The main source of data is Comstat (import, export, re-import, re-export). From FAOSTAT we took information about production volumes. The last year FAOSTAT contains data about is 2014, so 2015 is the last year we could predict for. Full dataset contained 307 million data points.

Due to limited time and computational resources, we conducted experiments only on the 10 most exported from the Russian Federation commodities. Also, we selected 20 countries in the same way. Thus, we got 200 growth points. Surely, in future experiments we should consider much larger set of commodities and countries, not only those well-developed already.

The testbed was set up as follows. All available data were split into two parts: train and test. Train subset contained information about trade from 2013 to 2014. Test subset contained information about only 2015. Each subset consisted of datapoints each representing a pair of export growth point candidates to compare. Features were constructed using "current" and "previous" year. Outcomes were constructed on the base of the "next" year. Thus, in train features were constructed on the base of 2011-2012 (2013 as "next") and 2012-2013 (2014 as "next") and outcomes were constructed on the base of 2013 and 2014 correspondingly. In test subset features

Table 1 Top 5 predicted export growth points and their summary proportion in the total export gain

No	Actual		Predicted					
			ARIMA		Baseline model		Advanced model	
	Partner Country	Commodity	Partner Country	Commodity	Partner Country	Commodity	Partner Country	Commodity
1	Saudi Arabia	Barley	Libya	Barley	Azerbaijan	Potatoes	Italy	Maize
2	China	Soybeans	Spain	Soybeans	Georgia	Maize	Spain	Maize
3	Turkey	Maize	Ukraine	Wheat	Uzbekistan	Wheat	Libya	Maize
4	Azerbaijan	Wheat	Ukraine	Molasses	Ukraine	Potatoes	Spain	Rye
5	Italy	Maize	Kazakhstan	Soybeans	China	Wheat	Ukraine	Molasses
Export gain $	360059k		11710k		13830k		19197k	
%	**76.2**		**2.4**		**2.9**		**4.0**	

were constructed using 2013-2014 and outcomes represented difference in dynamics in 2015. Each subset was symmetric: for each pair <A, B> there was also pair <B, A>. Samples with outcome of 0 were excluded from both subsets.

4.2 Baseline model

The objective of baseline model is to estimate, how accurate candidates can be compared using only knowledge about titles of these candidates. Baseline is implemented as Bernoully Naive Bayes classifier with feature set, consisting only of <*Product*$_1$*, Country*$_1$> *(only elements of left hand part of comparison).* Etalon oucomes for training the baseline model were constructed as $sign(dEV_1 - dEV_2)$, where dEV_i is the first difference of export value of *Product*$_i$ from the Russian Federation to *Country*$_i$.

Thus, this classifier estimates prior marginal probability of each candidate to grow faster than each other candidate. This model is very naive and measures skewness of our dataset and most frequent patterns of the Russian Federation international trade.

4.3 «Advanced» model

The objective of this model is to estimate, how much simple context information can improve comparison accuracy. There are several differences from the baseline: the feature set, the machine learning method used and the loss function.

The feature set consists of two parts: historical information about trade of the Russian Federation with *Product*$_i$ and *Country*$_i$; and the same information about the second candidate. "Historical information about trade" includes the following basic values from UN Comtrade database: export amount (in tonnes), export value (in USD), export prices (as ratio of value to amount), export monopolization; the same corresponding parameters for re-export, import, re-import. The feature set also contains information about production (from FAOSTAT database). Prior dynamics is taken into account using first order differences and ratios. First order difference (or ratio) is the difference (or ratio) of the value for the current year and that for the previous one. Monopolization (or competitiveness, or concentration) is estimated using Herfindahl index (sum training data was split so that data for each year was used solely either for the train or for evaluation. After best hyperparameters were chosen, the model was refitted using all training data. Finally, we decided to use LightGBM to train that model, because it showed the most promising results. All the results presented for "advanced" model were constructed using LightGBM.

One can notice that we do not explicitly use information about global economic situation. We omitted it from the feature set due to two main reasons: (1) it is very difficult to represent in such a way so a machine learning-based model can take full advantage of it (unclear how to prepare features); (2) some global information is implicitly encoded into difference between production, import and export, and also in monopolization estimates. Surely, explicitly taking into account the global economic situation is very important. We will consider it in next papers.

5 Experimental evaluation

As written before in the paper, the main objective of experimental evaluation is to estimate how much the detailed retrospective data about international trade is useful for the problem of growth point candidate ranking. Because of the nature of the problem, the standard classification or regression scores are not well applicable to measure the prediction quality, i.e. miscomparison of different pairs may have very different significance. Therefore, we used a proportion of the predicted export growth points in the total export gain as the score. In other words, the bigger part of export growth the model detects (the list "%" row in tables), the better the model works. These percent values may be treated as quantitative prediction quality measures.

Table 1 contains the scores for the top 5 actual growth points and for the predicted alternatives. Sum absolute export value growth for the predicted pairs is presented. The last row (%) contains the portion of total growth of export from Russia in 2015, calculated for all growth point candidates (as specified above). From this table one can see that it is nearly impossible to predict short one-year trade flow dynamics without additional information about global economic situation.

A notable difficulty here is high volatility of the product market, while the creation or development of a food manufacture is a long-term process. Therefore, we think that prediction of averaged, long-term trends would yield a more meaningful ranking.

Advanced model achieved slightly better results than baseline and ARIMA models. From that we conclude that retrospective data is useful to predict flow dynamics. This in turn means that combining open retrospective data about international trade with expert opinions makes much sense in order to maximize both likelihood and novelty.

Table 2 Top 5 predicted commodities and their proportion in the total export gain

No	Actual	ARIMA	Baseline model	Advanced model
1	Barley	Barley	Potatoes	Maize
2	Soybeans	Soybeans	Maize	Rye
3	Maize	Wheat	Wheat	Molasses
4	Wheat	Molasses	Linseed	Soybeans
5	Potatoes	Maize	Rye	Wheat
$	446903k	440272k	137694k	225233k
%	**94.6**	**93.2**	**29.1**	**47.6**

Table 2 presents five commodities with the highest expected growth. The last row (%) contains the portion of total growth. One can see how much Russian food export is non-diversified: 5 commodities occupy more than 90% of total export value growth. Also, we can see that ARIMA predicts commodity dynamic much better than both baseline and advanced model. We think that this is mostly due to inertia of flows: if something grows today, it will most probably grow tomorrow. Again, "advanced" model performed better than baseline. This means that prior information is not very useful to predict commodity dynamics.

Table 3 Top 5 predicted directions and their proportion in the total export gain

No	Actual	ARIMA	Baseline model	Advanced model
1	Saudi Arabia	Libya	Azerbaijan	Italy
2	China	Spain	Georgia	Spain
3	Turkey	Ukraine	Uzbekistan	Libya
4	Azerbaijan	Kazakhstan	Ukraine	Ukraine
5	Italy	Georgia	China	Armenia
$	374755k	49666k	145263k	47982k
%	**79.3**	**13.6**	**31.8**	**13.1**

Table 3 presents five countries with the highest expected import growth from the Russian Federation. From this table we conclude that Russia export is not only commodity-non-diversified, but also partner-non-diversified. From this table we can see that purely prior-based "baseline" model performed best: it predicted more than 30% of actual export growth. ARIMA and "advanced" model performed approximately equally. So, we conclude that almost no new markets are explored: we will trade tomorrow with those, who we trade today. Additional unaccounted factors may include politics, wars, sanctions, etc.

7 Conclusion and future work

In this paper we have reviewed and discussed the problem of export growth points discovery. The main contribution of this paper is an automated data-driven framework that addresses the problem. The framework uses open data from many data sources and modern machine learning techniques. We also conducted preliminary experiments to evaluate the possibility to use retrospective data to rank growth point candidates. The experiments were based on open data from FAOSTAT and UN Comtrade.

Currently, it is very difficult to say for sure, which method is more useful for the final task – growth point discovery. Different methods compared to each other differently, depending on how to compare (top5 growth points, top5 commodities or top5 directions). This fact gives some clues on what a better model should look like. Another thing that has to be changes is the objective function: predicting short-term export value changes is very difficult and useless, because developing a new manufacture needs much more than one year. Thus, it makes much more sense to predict long-term trends.

Main directions of future work include (a) repeating experiments with adjusted methodology; (b) creating a manually-annotated dataset of growth points; (c) incorporating information about global economic situation and substitutes.

Acknowledgment

The research is supported by Russian Foundation for Basic Research, project 16-29-12877.

References

[1] Rodrik D. Institutions, integration, and geography: In search of the deep determinants of economic growth //In Search for Prosperity: Analytic Narratives on Economic Growth. Princeton University Press, Princeton. – 2003

[2] Food and Agriculture Organization of the United Nations. URL: http://www.fao.org/faostat/en/

[3] UN Comtrade: International Trade Statistics. URL: https://comtrade.un.org/data/

[4] Kaukin A., Idrisov G. The gravity model of Russia's international trade: the case of a large country with a long border. Working paper. – 2014

[5] Gordeev D. et al. Analysis of Global Supply Chains in International Trade Patterns. – 2016. – №. 765

[6] Duenas M., Fagiolo G. Modeling the International-Trade Network: a gravity approach //Journal of Economic Interaction and Coordination. – 2013. – Vol. 8(1). – pp. 155-178

[7] Jaud M., Kukenova M., Strieborny M. Financial Development and Sustainable Exports: Evidence from Firm product Data //The World Economy. – 2015. – Vol. 38(7). – pp. 1090-1114

[8] Snijders T. A. B. Models for longitudinal network data //Models and methods in social network analysis. – 2005. – Vol. 1. – pp. 215-247

[9] Grater S. et al. Linking export opportunities of products and services: the case of South Africa.

[10] Sgrignoli P. The World Trade Web: A Multiple-Network Perspective //arXiv preprint arXiv:1409.3799. – 2014

[11] Lall S., Weiss J., Zhang J. The “sophistication” of exports: a new trade measure //World Development. – 2006. – Vol. 34(2). – pp. 222-237.

[12] Bernard A. B., Jensen J. B. Why some firms export //Review of Economics and Statistics. – 2004. – Vol. 86(2). – p. 561-569

[13] Barigozzi M., Fagiolo G., Garlaschelli D. Multinetwork of international trade: A commodity-specific analysis //Physical Review E. – 2010. – Vol. 81(4). – p. 46-104

[14] Peluso S. et al. International Trade: a Reinforced Urn Network Model. – 2016. – №. 1601.03067.

[15] Fronczak A. Structural Hamiltonian of the international trade network //No. – 2012. – Vol. 1. – No. arXiv: 1205.4589. – pp. 31-46

[16] Shen B., Zhang J., Zheng Q. Exploring multi-layer flow network of international trade based on flow distances //arXiv preprint arXiv:1504.02361. – 2015

[17] Kelle M. et al. Cross border and Foreign Affiliate Sales of Services: Evidence from German Microdata //The World Economy. – 2013. – Vol. 36(11). – pp. 1373-1392

[18] Sujjaviriyasup T., Pitiruek K. Agricultural Product Fore-casting Using Machine Learning Approach //Int. Journal of Math. Analysis. – 2013. – Vol. 7. – №. 38. – p. 1869-1875

[19] Sujjaviriyasup T., Pitiruek K. Hybrid ARIMA-support vector machine model for agricultural production planning //Applied Mathematical Sciences. – 2013. – Vol. 7. – №. 57. – p. 2833-2840

[20] Microsoft. https://github.com/microsoft/lightgbm