GMDH Helps to Build Models Based on Queries to Yandex for Forecast of Economic Crimes

Anna Boldyreva[1], Olexiy Koshulko[2]

[1]*Moscow Institute of Physics and Technology, Russian Academy of National Economy, Moscow, Russia*
[2]*Institute of Cybernetics, National Academy of Science of Ukraine, Kyev, Ukraine*

anna.boldyreva@phystech.edu, koshulko@gmail.com

Abstract. *In the paper we build models for forecast of economic crimes using Group Method of Data Handling (GMDH). Input variables of the model are frequencies of queries with monthly step. They are taken from data bases of the search machine of Yandex company. Output variable is the number of economic crimes. We shortly describe data sources and the process of modeling. The experiments show the average relative error of the forecast 3%-6%. Such an error allows to recommend the proposed technique for departments of Police related to economic crimes.*

Keywords

GMDH, economic crimes, Internet queries

1 Problem setting

Collections of user's queries to the Internet open new possibilities for forecasting dynamics of various processes and events related to human activity. One of such applications is modeling economic crimes. The basic hypothesis consists in the following: a person having some plan to complete an economic crime tries to find any similar cases in the Internet. He (she) looks for information concerning both punishment for the crime and any ways to avoid it. So, it is naturally to reveal a dependence between dynamics of crimes and dynamics of queries, and then to use this dependence for forecasting.

Speaking economic crimes we mean bribe, fraud, extortion, and other crimes, which are registered in the General Prosecutor Office of Russia, and which are included to the academic vocabulary of business terms. Speaking queries we mean words (word collocations), which are registered in data bases of the Yandex search machine, and which are included to the bases of terms. Yandex mentioned here is the largest Russian Internet company. Frequencies (intensities) of various queries are considered as input variables, and amount of crimes is considered as output variable.

The queries to Internet search machines have been already used for forecasting dynamics of various macroeconomic parameters. But the authors do not know any publications related to forecasting economic crimes. The only open publication we could find was our paper [1]. In this paper we continue the mentioned research.

2 Data

2.1 Dynamics of economic crimes

Economic crimes are such crimes as bribe, fraud, extortion, tax evasion, currency crime, illegal enterprise, smuggling, illegal trafficking of metals, sale of counterfeit money, crimes related to drugs, etc. Our final goal is to build models for different type of economic crimes. But by the moment we have data only for the sum of all economic crimes. Just for this reason we consider this sum as the one variable. We downloaded these data from the portal of legal statistics of the General Prosecutor Office[1]. The last available month was February 2015.

[1] The portal of legal statistics of the Prosecutor General's Office of Russia, http:// crimestat.ru/offenses only

2.2 Dynamics of queries

Dynamics of user's queries to Yandex search machine can be downloaded from Yandex data bases[2]. This moment the mentioned data bases contain about 5 millions of queries concerning different topics. The dynamics of these queries covers the period March 2013 – February 2015, that is 2 years. For the preliminary selection of queries we do the following two time-consuming procedures:

- On the first step we search queries from Yandex data bases that simultaneously belong to the academic vocabularies of business and juridical terms. The first vocabulary contains 6470 terms, and the second vocabulary contains 4460 terms.
- On the second step we calculate correlation between queries and crimes. If the correlation exceeds a given threshold then the query with its dynamics is selected. In our research we fix the threshold on the level 0.7.

The resulting list of the preliminary selected queries contained 102 terms. Table 1 presents indicator and some queries from the total list of queries with their frequencies. Speaking descriptors we mean queries.

Tab.1. Indicator and examples of descriptors.

Indicator/descriptors	01.03.2013	01.09.2013	01.03.2014	01.09.2014	01.02.2015	Correlation
Economic crimes	20583	9350	14398	10396	12575	1
Single camera	4685	1103	1298	1344	1561	0,726
Convertible stocks	596	443	528	567	610	0,719
Pension agreement	6887	8690	5118	4759	5885	0,713
Income tax	159756	85489	161599	83293	158537	0,706
Financial statements	289328	110098	377602	98046	166240	0,691

3 Modeling

To build the model for forecasting crimes one should test a huge number of functions reflecting possible dependence between the input and output variables. The Group Method of Data Handling (GMDH) just allows to complete this enumeration: it tests functions from a given class and selects the most significant combination of input variables in the framework of this procedure. [2,3]. In our paper we consider polynomial class of models.

In our experiments we used the specialized tool GMDH Shell including a) procedures of preprocessing b) several algorithms based on GMDH approach. Our previous experience showed the advantage of preliminary data transformation with square root operation for input variables. For this reason we tested algorithms under two options of preprocessing: with and without queries transformation. From the other hand we limited the number of algorithms using only the classical combinatorial algorithm and the neural network like algorithm. To reduce the number of variants we fixed maximum value of lag for input variables. It was equal 3 months. Therefore we elaborated 4 models: 2 of them related to algorithms and the other 2 related to preprocessing. The results are presented in the Table 2. Here: MAPE stands for the mean absolute percentage error.

Tab.2. Values of MAPE for different experiments.

Neuronet, no roots	**Neuronet, roots**	**Combi, no roots**	**Combi, roots**
1,3%	2,6%	5,5%	5,6%

[2]The service of the statistics for queries to Yandex. http:// wordstat.yandex.ru

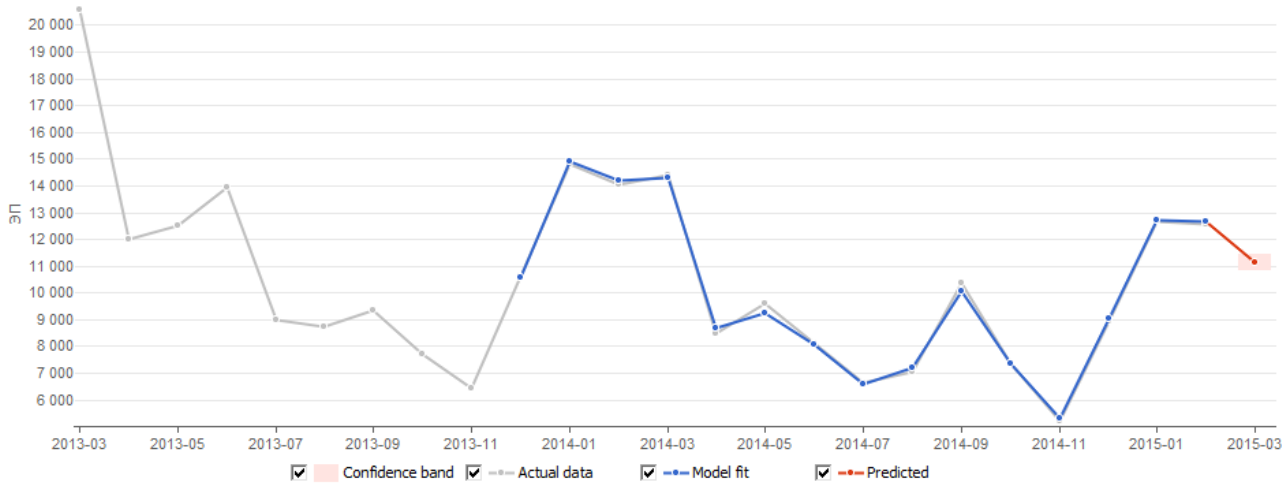

Fig.1. The best model for economic crimes:

neural algorithm without transformation.

As an example we demonstrate the results of one experiment where neural algorithm without preliminmary transformation were used. Here is the equation for this model:

$$Y1\,[t] = 95.5535 - Sequestration\,[t-4] * 0.0613822 + N2 * 1.02263$$

$$N2\,[t] = -196.624 + Sequestration\,[t-3] * 0.0663601 + N3 * 0.983966$$

$$N3\,[t] = -795.589 + Wound\ Ballistics\,[t-1] * 16.678 + N5 * 0.934989$$

$$N5\,[t] = -771.286 - Blizhnyaya\ Duma\,[t-2] * 2.37261 + N6 * 1.17293$$

$$N6\,[t] = -614.82 + N9 * 0.673973 + N12 * 0.387116$$

$$N12\,[t] = 8801.59 - Taxable\ income\,[t-4] * 1.26206 + Active\ electoral\ right\,[t-3] * 2.83552$$

$$N9\,[t] = 19313.1 - Poison\,[t-3] + 0.115881 * Equal\ rights\ of\ citizens\,[t-3] * 7.48979$$

where: *N*xx are neural layers, words in italic are queries

4 Conclusions

In the paper we build the models for forecasting dynamics of economic crimes based on the dynamics of queries to the Yandex search machine. We study several options of modeling to select the best one. The accuracy of models proved to be enough high. So, the proposed technology could be recommended for Police departments related to economic crimes. In the future we intend to build particular models for forecasting economic crimes in Russian regions.

References

[1] Boldyreva A., Koshulko O.: Forecasting models of economic crimes according to queries to Internet: Regression vs GMDH. *In: Mathematical modeling of social processes, Proc. of Sociol. Faculty of MSU, Publ. House MSU (Moscow State Lomonosov Univ.)*, vol. 17, 2015, pp. 35-43 [rus]

[2] Stepashko V.: Ideas of academician O. Ivakhnenko in Inductive Modeling field from historical perspective. *In: Proc. of 4th Intern. Conf. on Induct. Modeling (ICIM-2013), Kyev: NAS of Ukraine, Prague: Prague Tech. Univ.*, 2013, pp. 31-37

[3] Stepashko V.: Method of critical variances as an analytical tool of the inductive modeling theory. *Journ. of Inform. and Automat. Sciences, Begell House Inc.*, 2008, vol. 40, No. 3, pp. 47-58