Conference Paper

Problems of data collection for the application of the Data Mining methods in analyzing threshold levels of indicators of economic security

Zhukov A. N. and Leonov P. Y.

National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Kashirskoe shosse 31, Moscow, 115409, Russia

Abstract

Determining threshold values of key economic security indices describing an economic situation in any country is an important stage in the assessment of the country's economic stability. A research was undertaken to determine how this problem can be effectively solved with such Data Mining algorithm as decision trees. According to the results of the research the effectiveness of the method was proved, but only with sufficient amount of data available. However, such data collection has a number of significant problems. These problems can be attributed to the following factors: the data for the analysis are presented with varying frequency, there is no possibility to use data over longer time intervals, the lack of a common list of indicators of economic security, which are used in different countries. The purpose of this paper is to analyze the existing problems of the data collection and to submit proposals of solving them.

Keywords: Economic security, Issues of data collection, Data Mining, thresholds.

Corresponding Author:
Zhukov A. N.
bug95@mail.ru

Received: 11 December 2017
Accepted: 20 January 2018
Published: 13 February 2018

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the FinTech and RegTech: Possibilities, Threats and Risks of Financial Technologies Conference Committee.

1. Introduction

Within the framework of the Fifth International Conference W-2017 FiCloud held in Prague at the end of August, a research was conducted [1] on the theme: «the application of the Data Mining algorithms in determining threshold levels of indicators of economic security». It included the following stages: data collection, processing of collected information, building models and analysis of the results obtained. The research allowed to get the results, with the algorithm of building decision trees, which were quite accurate from logical and economical points of view.

However, the collection of data for analysis has caused serious problems associated with both the amount of available data and the nature area of research. These issues and their possible solutions will be discussed in this article.

2. Analysis of the subject area and identification of data collection problems

However, before talking about the collection of data we should define the subject area and the key terms. Firstly, Data mining according to the definition given by Oracle is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Another term required to characterize the domain of the present research is the economic security. According to the Decree of the President of the Russian Federation from 13.05.2017 N 208 economic security strategy "of the Russian Federation for the period up to the year 2030" [2], "Economic security is protection of the national economy from external and internal while the economic sovereignty of the country, the unity of its economic space and conditions for realization of national strategic interests of the Russian Federation are preserved". And finally, it is necessary to define the term "economic security indicators", according to "Economic safety of Russia" [3] these are indicators that reflect quantity of threats to economic security, that have high sensitivity and variability and as a result they are able to warn society, State and market actors about possible dangers associated with the changes in the macroeconomic situation or in economic policy. [3]

It was hypothesized that threshold values of indices can be determined based on decision tree nodes. During the partition of a sample by the "auspicious year" criterion, the value of an explanatory variable in a tree node must comply with or be close to, a threshold value of a given index.

To validate the hypotheses the following indicators were selected: GDP (USD); Unemployment Rate; Population Share Below Poverty Line; Direct Investments in the RF; Coefficient of Fixed Assets Renewal; Coefficient of Fixed Assets Retirement; Degree of Fixed Assets Depreciation; Economic Activity; Level of Population; Fixed Capital Expenditure; Fertility Rate; Life Expectancy; Migration: Arrived/Departed; Money Supply M2 (% of GDP); Inflation Rate; Reserves Including Gold; Import of Goods and Services; National Currency/USD Exchange Rate; Grain Crop Production. Data were taken from the website of the Federal State statistics service, the official statistical agency in the Russian Federation [4].

Values were obtained for the period from 1993 to 2014 years. This period was selected because of completeness of data, and also because it covers important social, economic and political events in the history of the Russian Federation. These events include both periods of crisis and periods of growth and development of the economy.

At this stage, the following problem was revealed: statistical information on indicators is collected with varying frequency, very few indicators are calculated for intervals less than the calendar year. This can be linked to the fact that the collection of statistics on them is connected to financial costs and to the lack of sufficient staff of statistical services. As usual the collection of more frequent data is automated. Another reason is that collection of data on many indicators doesn't have high variability on short periods of time, so expended funds are not comparable with obtained results. However, this statement is true only when we talk about the current approaches to analysis.

Due to the influence of the above factors, the quality of the data collected for analysis was questioned, as well as the results of the research, despite the fact that many of the calculated thresholds accurately comply with stated thresholds.

It was decided to test a hypothesis by taking similar indicators for different countries to increase the sample size. A total of 19 countries were chosen. They included both mature economies (U.S.A., England, Germany, etc.) and countries with developing economy (Ukraine, Belarus, etc.). Asian countries were represented in our sample collection by China, Japan, South Korea and Singapore

Such approach to solving the problem increasing sample was used due to the nature of the subject area, particularly because of the lack of opportunities to increase the time interval for collecting data. The absence of such a possibility is the second key problem in data collection. This problem is related to the variability of the world economy, and its globalization. For example, the best time interval for the Russian Federation is about 20-25 years, because previously country had a different economic model, and, after the USSR collapse a 10-year period of time was marked by a significant decline in all economic and social indicators. Surely 10 years ago, such periods could not be excluded from the analysis, but at the moment the Russian economy has gone far ahead of 90-ies level. A similar situation applies to the world economy as a whole, in the context of globalization and Informatization and high variability of the current political and economic environment.

Unfortunately, with the increase of the number of countries we permanently lost the ability to use many indicators due to the absence of information in open access in selected States. As a result, the number of indicators was reduced from 19 to 11 Fertility Rate; Life Expectancy; Money Supply M2; Inflation rate; Reserves Including

Gold; Import of Goods and Services; National Currency/USD Exchange Rate; Grain Crop Production; Country; GDP (USD); Net Foreign Investments; Unemployment Rate.

Absence of opportunity to increase the time interval is compounded by gaps in the available data for the analysis period. If we consider the situation in the context of the several States, it becomes even more difficult. However, within a few years, it will be possible to completely get rid of gaps in the data because the interval will move ahead for another few years and absent data will not be used for further research. A very similar and more significant problem is the lack of a common format of presentation of statistical information on a number of indicators in many countries, e.g. GDP and other economic indicators as usually measured in the national currency and the question is, in what currency and at what exchange rate it is best to measure such economic indicators, but a far more significant problem is the problem of absence of many indicators.Usually such specific indicators as the number of crimes per 1,000 people, the level of confidence in the Government and other social indicators do not relate directly to the work of statistical services and are gathered by them on the basis of the data provided by other ministries and departments When collecting data for analysis it is very difficult to find the original sources of such information about foreign States. A creation of a single independent statistics service with similar standards and requirements to data could serve as a method of solving this problem. Public sources of statistical information on various countries are certainly available, but their reliability and validity of the information should be proved.

3. Results analysis

Despite the fact that when analyzing data the suggested hypothesis was confirmed and current thresholds indicators of economic security were adjusted, the identified problems of data collection remain an important obstacle to conducting of such studies.

Summing up the above, these problems include: different frequency of collecting data for analysis, absence of ability to use data over longer time intervals, absence of a list of economic security indicators, which would be used by different countries.

4. Possible solutions to identified problems

There are several possible approaches to address these issues.

Firstly, it is an increase of frequency of collection of statistical data. However, this decision implies an increase in financial costs of statistical research and requires time for implementation.

The second solution is a critical revision of the statistical data collected in favor of those that possess greater variability and are more instructive on smaller time intervals. The disadvantage of this approach is that you will need time to form a sample of such indicators and gain their value over a sufficient period of history.

The third decision could be a creation of a single independent international statistics service with similar standards and requirements to data for all its participants. As with the previous cases, it will take time and considerable financial and human resources.

5. Conclusion

In conclusion it could be stated that there is no high-priority problems of statistical data collection. They all are interrelated, and their solving can only be achieved with comprehensive measures, such as improvement of communication between statistical services, increase of the frequency of the collection of statistical information, automation of collection of statistical data, implementation of international statistical standards and statistical indicators.

Acknowledgements

This work was supported by Competitiveness Growth Program of the Federal Autonomous Educational Institution of Higher Education National Research Nuclear University MEPhI (Moscow Engineering Physics Institute).

References

[1] A.N. Zhukov, K.V. Evsikov, Leonov P.Y., Morozov N.V., Domashova J.V. Use of Data Mining Algorithms While Determining Threshold Values of Economic Security Indices. -2017 5^{th} International Conference on Future Internet of Things and Cloud Workshops. -p. 20-24.

[2] the Decree of the President of the Russian Federation from 13.05.2017 N 208 "economic security strategy of the Russian Federation for the period up to the year 2030"

[3] Senchagov, v.k. Economic safety of Russia [text]/V. Senchagov. -Moscow: BEAN, 2015[3] -815

[4] the Federal State statistics service [electronic resource]: access mode: http://www.gks. ru /