DATA SCIENCE METHODS IN EVALUATING INNOVATIVE POTENTIAL AND INNOVATIVE ACTIVITY OF INDUSTRIAL ENTERPRISE UNDER CONDITIONS OF DIGITAL ECONOMY

A.E. Gorokhova, I.A. Zaitsev, V.D. Sekerin

A.E. Gorokhova
Doctor of Economics, Professor, Department of Economics and Organization
Federal State Budgetary Educational Institution of Higher Education "Moscow Polytechnic University", Moscow

I.A. Zaitsev
postgraduate student of the second year of study at the Department of Management of the State Budget Educational Institution of Higher Education of the Moscow Region Technological University, Korolev, Moscow Region

V.D. Sekerin
Doctor of Economics, Professor, Department of Economics and Organization
Federal State Budgetary Educational Institution of Higher Education
"Moscow Polytechnic University", Moscow

Keywords:
Data science, innovative potential, innovative activity, industrial enterprise, digital economy.

Abstract:
The article examines some examples of mathematical methods used in data science to assess the innovative potential and innovative activity of an industrial enterprise in a digital economy. Since the science of data (data science) has a fairly wide mathematical apparatus, which is designed to search for new information, statistical data, as well as to make forecasts and economic proposals, these same methods can be used to work with data when evaluating innovative indicators of an enterprise.

Introduction

In the digital economy, one of the key factors in the development of an industrial enterprise is the availability of information and skills to work with it. Having information as a resource provides a competitive advantage for the company. Based on this advantage, you can develop a strategy for innovative development of the enterprise, as well as develop innovations themselves. In other words, the information resource provides the innovative potential and innovative activity of the enterprise. (Gorokhova ed al., 2018, pp. 103-108).

Data science is the science that studies the problems of analyzing, processing and presenting data in digital form. Using the methods of data science, you can find almost any necessary information to assess the market and economic prospects of innovation, evaluate the preferences of a potential client, which provides a reliable innovative potential for the enterprise.

To assess the innovative potential and innovative activity of an enterprise, a large amount of data must be collected and analyzed. The collected data must be calculated using mathematical methods to obtain the final value of the innovative potential and innovative activity.

1. Goal and method

This article examines similar methods and tools used in data science to assess the innovative potential and innovative activity of an industrial enterprise in the digital economy.

Big Data technology can be used to develop methods for evaluating the innovation activity and innovative potential of an industrial enterprise. The first step in the assessment should be to collect indicators of the characteristics of innovation potential and innovation activity.

2. Results

A total of i indicators are collected, which are divided into j groups. Indicators can be expressed as an integer (the number of employees involved in R & d), a fraction (average salary of employee engaged in R & d in RUB thousand), shares (share of employees with higher education), percent (percentage of R & d expenditure of the total cost of the enterprise).

Big Data technology allows you to automatically upload, store and process indicators every day and without human intervention. The technology is characterized by 3 V (Velocity, Volume, Variety), which indicates that it is able to work quickly with a large volume of heterogeneous data (Morrison et al.,2010)

In this way, an operator who wants to assess the potential or activity of an enterprise at the moment can get information instantly, without making calculations themselves, without understanding the indicators. This is provided by constant work with Big data. When performing an operation, specialists immediately enter data in Big data, which stores data in the desired district There are also calculation programs that read data every day, such as the number of goods sold or units produced, automatically enter and update information in the company's file system and generate a report. The system collects data in such a way that it immediately distributes data into subgroups, which will form a subindex in the future. Information can be collected in these ways. (Schoenberger, Kukier, 2014, 240 p.)

In order to be able to evaluate innovation potential and innovation activity separately, each criterion in the database corresponds to the values (A) or (P), which means innovation activity and innovation potential. In accordance with the fact that we need to evaluate (innovation activity or innovation potential), big data issues only those criteria that we need to calculate. You can implement this using the simplest logical function "if".

For example:
If: x = " innovation activity»
That:
If: zij=...(A)
To: Nij=zij
…
…
Where,

X-value of the query: innovation, innovation activity, innovation potential
Zij – the value of the parameter
Nij - value of the array of parameters to output.

This function can be implemented using a similar logical condition written in a computer language. In other words, initially only the data requested by the operator is received for further processing. This saves time and data processing resources. (Borovikov, 2003, 688 p.)

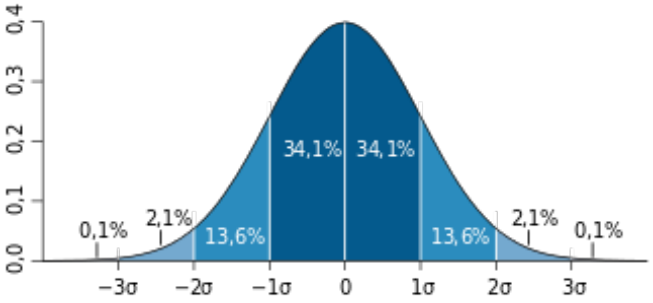
Thus, Big data technology greatly facilitates working with data, and also makes it much more efficient and accessible. Data storage becomes more reliable and long-term. Data does not require physical storage, which reduces storage costs and reduces costs. The most important advantage is that the result can be seen at any time. Indicators can be automatically processed as they are received in the storage, which will speed up the operator's work and allow you to display the final result faster on the screen.

The values of indicators, i.e. big data, are written in different formats: absolute, such as, for example, the number of Universities, relative, such as, for example, the percentage of people with higher education among employees. In order to compare them, the data should be reduced to a single form using mathematical methods.

In Economics, the three Sigma rule is used to set the data sampling frame. All values of indicators of innovation potential and innovation activity should be in the range of three Sigma (formula 1). The probability distribution of finding a number within three Sigma is shown in the figure (Fig. 1), where the mark zero corresponds to the mathematical expectation (Ivchenko, Medvedev, 2010).

Figure 1: *probability Distribution of finding a random variable within 3.*

$$x_i \in (\bar{x} - 3\sigma; \bar{x} + 3\sigma) \quad (1)$$



Source: Compiled by the authors.

If the values go beyond 3 σ, the method of taking the root of degree s is used to reduce them. The degree of the root depends on the coefficient of asymmetry.
The coefficient of asymmetry is equal to the partial of the third Central moment and the standard deviation, raised to a cube (formula 2).

$$\gamma_1 = \frac{\mu_3}{\sigma^3} \quad (2)$$

Where,
γ_1 – Skewness
μ_3 – the third Central moment
σ - the standard deviation

If even after the transformation, the value goes beyond three Sigma, then it does not participate in the calculations, we get rid of it.
The disadvantage of this method is a large level of filtered values, which negatively affects the objectivity of the final value. One of the obvious drawbacks of this method is the case when the

values are equal to one. A root of any degree will give one, when extracted from one. That is, in this case, it is impossible to bring the indicator to the frame and it remains only to throw it out.
Correlation analysis can be used to describe the relationship between the characteristics of innovative potential and innovative activity of an industrial enterprise. To take into account the degree of influence of each indicator, determine their relationship to each other, and filter out indicators that do not affect the number of innovations implemented, you can use correlation analysis.
Correlation analysis is a method for processing statistical data based on the study of correlation coefficients. To determine the correlation, we need several observations of the same variable (Kolemaev, 2006).

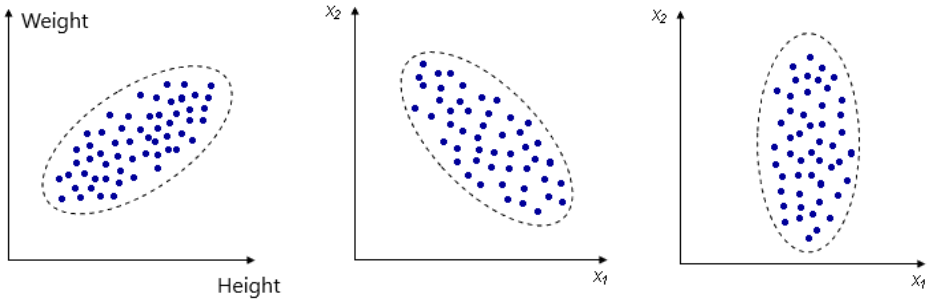
The correlation method allows you to determine: a) whether there is a relationship between the parameters b) what kind and strength of this relationship C) allows you to predict the behavior of one parameter based on data about the second parameter d) helps to classify objects by feature.

The advantages of the method are that the coefficients are quite simple to calculate, and no special mathematical training is required. The interpretation is also quite simple.

The disadvantage is that the method shows only a statistical relationship, not a cause-and-effect one.

Let's consider the correlation on the example of indicators of innovative activity and innovative potential of an industrial enterprise. Figure 2a shows a positive correlation, where each point is a single dimension. Figure 2b shows a negative correlation. Figure 2c shows values without correlation (Basovsky, 2002).

Figure 2: *a) positive correlation b) negative correlation c) lack of correlation.*



2a) 2b) 2c)

Source: Compiled by the authors.

To characterize the degree of correlation, there is a value describing the correlation process – the correlation coefficient. The correlation coefficient is a measure of the linear relationship between two variables.

If the correlation is linear, the correlation coefficient is 1 or -1. If there is no correlation between the two objects, the correlation coefficient will be 0. The correlation scale is shown in the table (tab. 1).

Table 1: *correlation scale of values.*

Values of the coefficient r	The degree of connection
0,75 - 1	Very high positive
0,50 – 0,74	High positive
0,25 – 0,49	Average positive
0 – 0,24	Weak positive
-0,24 - 0	Weak negative
-0,49 - -0,24	Average negative
-0,74 - -0,50	High negative
-1 - -0,75	Very high negative

Source: Compiled by the authors.

Correlation analysis is suitable for calculating the innovative potential and innovative activity of an industrial enterprise. It is able to identify the relationship, clearly show it in the form of a correlation coefficient. However, it is better to use the regression analysis method in the calculation, since it gives an accurate characteristic and degree of influence. Similarly, correlation analysis shows the relationship of two related events, and regression analysis can describe the relationship of two poorly related events.

In addition to correlation analysis, regression analysis is often used. The methodology for assessing the innovative potential and innovative activity of an enterprise is based on the assessment of indicators of a set of criteria that characterize these concepts. To take into account the degree of influence of each indicator, as well as to filter out indicators that do not affect the number of implemented innovations, you should use mathematical methods to determine the relationship of indicators and the degree of their influence on the main indicator. One of these methods is regression analysis, which results in a coefficient of correlation between the indicator and the number of implemented innovations. The method describes the nature of the relationship between two variables. The goals of regression analysis are:

- Determining the degree of determinism of dependent variable variants by independent variables
- Calculating the value of a dependent variable using independent variables
- Determining the contribution of private independent variables to the variation of the dependent variable

Regression analysis shows the degree of influence of characteristics on the main factor. The method is based on regression equations, mathematical formulas that are applied to variables in order to predict the dependent variable (formula 3,4) (Berezhnaya, Berezhnoy, 2005)

$$E(Y \mid X) = F(X, \beta) \quad (3)$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \quad (4)$$

Where,

Y - is a dependent variable that we need to determine the effect on. The variable describes the process that is most important to us. In our case, this is the number of implemented innovations.

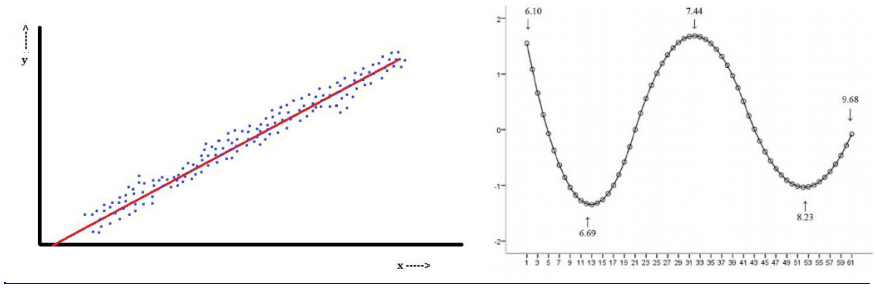
β - is the coefficient of influence of variable x on the dependent, main variable Y. the coefficient describes the strength and type of relationship between two variables.

ε - is a random error, it is the set of all random variables that occur and cause the error.

x - is a variable whose degree of influence and relationships on the dependent variable we find by regression method.

Creating a regression model is an iterative process that seeks to find effective variables x that explain variables Y. the Process is repeated many times, adding or removing variables. Depending on the nature of the process and the relationship, the process can be linear or nonlinear (Fig. 3) (Gladilin, 2006)

Figure 3: *a) linear regression, b) - not linear regression.*



a) b)

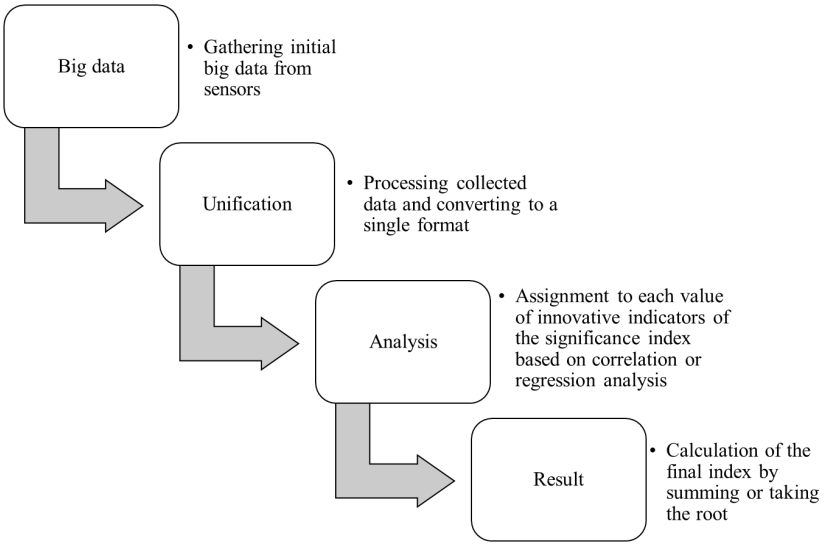
Source: Compiled by the authors.

The advantages of this method are its accuracy, simplicity, and visibility. The method clearly shows the relationship of two variables and the degree of influence of one on the other, which allows you to set the significance coefficients, as well as exclude values that do not affect the main value or those values that have a close relationship with other factors of influence and duplicate them. The method is not complicated and therefore accessible and easy to apply. You can calculate the degree of relationship and influence using the simplest Microsoft Excel program with the Power Query add-in. The method is also visual, since it gives coefficients that explicitly characterize the type of connection.

The method of regression analysis is well applied for evaluating the innovative potential and innovative activity of an industrial enterprise.

Based on these methods, we can make an algorithm for working with data to assess the innovative potential and innovative activity of an industrial enterprise (Fig. 4).

Figure 4: *the sequence of mathematical operations used to evaluate the innovative potential and innovative activity of an enterprise.*



Source: Compiled by the authors.

Conclusion

First, big data is collected from devices and sensors in the enterprise, then it is converted to a common format so that it is easy to read. After that, each value and each group of values is assigned a significance coefficient using the correlation or regression analysis method. Then the values are calculated as the sum or root.

Mathematical methods used for searching and analyzing new information in data science are perfectly applicable for calculating the innovative potential and innovative activity of an industrial enterprise in the digital economy and should be used in industrial enterprises.

References

MORRISON Alan et al. Big Data: how to extract information from it. *Technology forecast. Quarterly magazine*, Russian edition, 2010 issue 3. PricewaterhouseCoopers(December 17, 2010). Retrieved 12 November 2011. Archived from the original on March 11, 2012.

BASOVSKY L. E., *Forecasting and planning in market conditions*, textbook. Moscow: INFRA-M, 2002.

BEREZHNAYA E. V., BEREZHNOY V. I., *Mathematical methods of modeling economic systems*, textbook, 2nd ed.,- Moscow: Finance and statistics, - 2005.

BOROVIKOV V. STATISTICA. *The art of data analysis on a computer: for professionals* / V. Borovikov. — SPb.: Peter, 2003. — 688 p. — ISBN 5-272-00078-1.

MAYER-SCHOENBERGER V. KUKIER K. Big data. A revolution that will change the way we live, work, and think = Big Data. *A Revolution That Will Transform How We Live, Work, and Think Ina Gaiduk*. - Moscow: Mann, Ivanov, Ferber, 2014. - 240 p — ISBN 987-5-91657-936-9.

GLADILIN A. V., *Econometrics: study guide*. Moscow: KNORUS, 2006.

GOROKHOVA A.E., GAISINA L.M., GAREEV E.S., SHUTOV N.V., SHAKIROVA E.V. Application of Coaching Methods at Agricultural and Industrial Enterprises to improve the Quality of Young Specialists' Adaptation In: *Quality-Access to Success*. – Volume: 19, Issue: 164, 2018, Pages: 103-108

IVCHENKO G. I., MEDVEDEV Yu. I. *Introduction to mathematical statistics*. Moscow: LKI Publishing house, 2010. - §2.2. Selective moments: exact and asymptotic theory. — ISBN 978-5-382-01013-7.

KOLEMAEV V. A. *Econometrica*: textbook, - Moscow: INFRA-M, 2006.