DOI: 10.17323/2587-814X.2022.1.42.55

Information-logical model of express analysis of the state of the enterprise that meets the requirements of standards and regulations, based on publicly available data

Tatiana K. Bogdanova
E-mail: tanbog@hse.ru

Liudmila V. Zhukova
E-mail: lvzhukova@hse.ru

HSE University
Address: 20, Myasnitskaya Street, Moscow 101000, Russia

Abstract

The last 10 years have witnessed an explosive growth in the volume of information posted on the Internet and the digital economy, as well as the formation of official databases of various public authorities. The availability of a large information base open for research has facilitated the development of new methods and approaches to solving analytical problems. Building management and decision-making support systems based on the use of united disparate open data sources allows end users to make the most effective decisions. This is the approach that underpins business growth and managerial maturity at all levels – there is no alternative. Such an approach ultimately creates the conditions for further growth of the economy as a whole. This paper proposes the information and logical model of express analysis of compliance of socio-economic condition of the enterprise with the regulatory requirements of the control and supervisory authorities on the basis of open, publicly available information. The conclusions drawn on the basis of express analysis serve as a basis for deciding on the need for a more detailed, in-depth analysis of the state of individual enterprises.

Keywords: express analysis, information-logical model, enterprise status, regulatory requirements, control and supervisory bodies, open sources of information, publicly available data, structured and unstructured data

Citation: Bogdanova T.K., Zhukova L.V. (2022) Information-logical model of express analysis of the state of the enterprise that meets the requirements of standards and regulations, based on publicly available data. *Business Informatics*, vol. 16, no. 1, pp. 42–55. DOI: 10.17323/2587-814X.2022.1.42.55

Introduction

This paper proposes an information and logical model of express analysis of the compliance of the socio-economic condition of the enterprise with regulatory requirements on the part of the control and supervisory authorities on the basis of publicly available information. An information-logical model is built on the basis of the proposed concept, one of the important features of which is that the concept takes into account any requirements of different regulators, both quantitative and qualitative, imposed on economic objects of different types (enterprises, organizations, educational institutions, etc.) [1]. For different types of enterprises and the requirements imposed on them by the regulator, it is necessary to form different sets of components based on publicly available information, the aggregation of which, using the developed search table, results in the calculation of the value of the integral indicator, which is the basis of the express-analysis. Each component characterizes different aspects of the company's activities: economic, social, financial, technical, etc., and is evaluated in accordance with the methods of machine learning, mathematical statistics and econometrics [2, 3].

Both structured and unstructured information is used to carry out a rapid analysis of the state of an enterprise. Unstructured information is pre-structured using various methods of textual information processing [4–6].

The dynamics of the environment are increasing, the stability of the external environment is decreasing, and the requirements for a rapid response to crises are increasing. The amount of information that needs to be processed to make this or that decision is consistently increasing, at the same time the requirements for the quality, security and relevance of this information are becoming stricter.

The simultaneous use of structured and unstructured statistical data makes it possible to obtain a more accurate qualitative assessment of the object of study, taking into account changes not yet reflected in official statistical reports, which are provided with a certain periodicity and an inevitable time lag.

The result of the express-analysis is an assessment of compliance of the socio-economic condition of the enterprise with the regulatory requirements of the regulator. The conclusions drawn on the basis of the express analysis serve as a justification for deciding on the need for a more detailed, in-depth analysis of individual enterprises.

In recent years, research papers on various economic and mathematical studies have increasingly focused on the use of modern digital technologies for processing large volumes of structured, weakly structured and unstructured data from open Internet sources, machine learning and artificial intelligence methods in decision support models [7–10].

The use of innovative digital capabilities to collect and analyze publicly available information from the Internet allows us to perform additional analysis of the quality characteristics of various enterprises and other research objects. Such open data analysis can be carried out with the help of an auxiliary independent research object evaluation tool created on the basis of analysis of large volumes of structured, weakly structured and unstructured data from open internet sources, and to compare the results with the official research methodology on internal or official statistical data.

Control measures taken on the basis of official statistical information may come with a long delay, because between the end of the reporting period and the transfer of official statistical data on the state of the object to the public authorities may take from 3 to 8 months, which makes it difficult to respond promptly in force majeure situations.

In scientific research, many authors propose various economic and mathematical models based on official statistical information [11]. Most of them are econometric models or models that use machine learning techniques. As a rule, the available statistical data are divided into groups (demographic, social, financial, etc.), ranked, or somehow combined into a single integral indicator, and the factors are assigned weights. Often the result of such a study is an integral indicator (coefficient), which is useful for comparing objects. Such tools rely heavily on internal data or on an existing statistical base [12].

The use of structured and unstructured data analysis from open internet sources is the most comprehensive and versatile way to fully analyze the state of the economic object of study in a comprehensive way. It provides objective information on the current situation without intermediate processing based on the analysis of a wide variety of relevant data stored in the public domain in all Internet sources. If necessary, the results of external data analysis can be correlated with the results of similar analytical activities carried out using internal data. In addition, the results of analysis from open publicly available sources can complement official or internal data in some aspects of the subject's activities.

The advantage of using open data is the ability to obtain information at any periodicity (without reference to the regularity of updating, officially published statistical reporting), to expand and check compliance of the actual socio-economic condition of the object of research with official data.

1. Classification of publicly available information sources

All publicly available information can be represented in the form of different types of data. Currently, all existing data can be divided into:

1) structured;
2) poorly structured;
3) quasi-structured;
4) unstructured.

Structured data refers to data that is organized in a certain way, has a given structure, and describes a specific subject area. Taken together, this allows for reliable and in-depth analysis of this data. This information is most often presented in the form of tables.

Loosely structured data is data that does not follow a clear structure of tables and relationships in the database, but contains special delimiters (tags) that allow us to do semantic separation of the entire data set. Examples include XML documents.

Quasi-structured data is data in an unstructured format, which requires a lot of time to be processed by special tools. An example of such data is a website page.

Unstructured data is data that does not have a specific form and is not strictly fixed. At the moment this is the predominant data format due to the development of the information society. Approximately 80% of all currently available information is unstructured. Examples of such data are images, video, audio and textual information from social media.

Depending on the type of data, it requires its own preprocessing and processing methods. Most methods in mathematical statistics and econometrics are based on the analysis of structured information. Machine learning methods, neural networks allow us to analyze weakly structured, quasi-structured and unstructured data, identifying patterns in them. Furthermore, with various preprocessing procedures, these data can be reduced to structured data and incorporated into classical mathematical models.

If unstructured data is represented by text, pre-processing it using vectorization and classification methods allows us to bring it to a structured form.

The information to form the research base for the express analysis can be obtained from different sources, differing in status, frequency of updating and the degree of reliability of the information provided. *Table 1* presents the classification of publicly available information sources according to the reliability of the source.

Table 1.

Classification of sources of publicly available information

Source of information	Characteristics of information source	Example of information source	Type of information	Update on source
Official data generators and aggregators	Websites of federal and regional statistical bodies, websites of ministries and agencies that publish thematic data under the information disclosure regulations, the reliability of which is confirmed by the relevant public authority.	rosstat.gov.ru zakupki.gov.ru fssp.gov.ru cbr.ru wciom.ru	Structured data.	As a rule, the frequency of updates is once a quarter, or less frequently.
The websites and social media pages of the research subjects	Websites of enterprises, organizations of all forms of ownership, websites of platforms on which they are obliged to post information about their activities. The credibility of the information is usually confirmed only by the object of the research itself.	technomoscow.ru uniconf.ru tinkoff.ru 57.mskobr.ru	All data types.	Constant updating.
Unofficial data generators	Websites of organizations engaged in activities related to the research subjects and publishing data about them in open sources. Credibility is ensured by internal monitoring and control of information.	cian.ru hse.ru/rlms	Predominantly structured data.	According to the approved methodology, updates can be carried out either at set intervals or on an ongoing basis.
Unofficial data aggregators	Russian and international data aggregators, usually providing data for scientific and other studies. Credibility is ensured by internal monitoring.	bankodrom.ru banki.ru avtostat.ru data.worldbank.org.	Predominantly structured data.	Updates are usually carried out at intervals that correspond to the frequency with which official data are updated.
Unofficial internet sources of expert studies	Russian and international websites of expert organizations, rating agencies, personal pages of recognized experts. The reliability of the data is ensured by the reputation of the expert.	raexpert.ru ra–national.ru	All types of data.	The update is carried out in accordance with the source's internal rules.
Unofficial publicly available internet sources	Social media pages, blogs, comments on content, informal community pages. The validity of the data is usually not subject to verification.	moneyzz.ru pedsovet.su	Predominantly unstructured or weakly structured data.	Constant updating.

The information and logical model of building an integral indicator for express analysis of compliance of the socio-economic condition of the enterprise with the regulatory requirements of the control and supervisory authorities proposed in this article is based on the conceptual model of express analysis set out in [1]. A distinctive feature of the proposed conceptual model is that the authors propose to take into account the requirements of regulatory authorities as a starting point, while most Russian and foreign studies assess the state of the research object based on the requirements imposed on the object by its owners or investors. Another advantage of the conceptual model is the use of publicly available data, i.e. the possibility to obtain information at any time without being bound to the periods of updating the officially published statistical reports, and the possibility to check the compliance of the actual state of the research object with the official data. The proposed information-logical model is a combination of an algorithm for calculating the individual components of the integral index by using mathematical, econometric and statistical methods, the characteristics of input and output information at each stage, and, actually, the algorithm of calculating the values of the integral index by using a logical function based on a search table.

2. Components of the integral index and methods of their estimation

The integral index is a flexible express-analysis tool based on publicly available structured and unstructured data. The construction algorithm for the Integral Indicator is based on the aggregation of the individual values of each component in the set using a look-up table. Each component is estimated using mathematical, econometric and statistical methods, such as: logistic regression model, clustering and grouping methods, thematic modelling methods, etc.

The flexible toolkit of express analysis for management decision-making developed on the basis of the conceptual model is a sequence of five stages, starting from the requirements on the part of control and supervisory authorities, development and evaluation of a set of components characterizing the research object, their aggregation into a single integral indicator based on the search table, and ending with the monitoring and ranking of research objects according to the results of calculations [1].

Depending on the type of research object (industrial enterprise, banking organization, educational institution, etc.), based on the requirements of various regulators, a list of data sources for rapid analysis is formed: websites of research objects, news sources, electronic platforms or information aggregators, websites of state authorities, etc. The research database is created on the basis of the information from these sources. The flexibility of the tools proposed in the article is due to the fact that the list of components necessary for rapid analysis can be supplemented depending on the type of research subject, the frequently changing requirements of regulatory and supervisory authorities and an increasing number of publicly available information sources.

Table 2 provides a list of the possible components identified by the authors relating to the four blocks of types of input information for component calculation, types of variables of the calculated value of each component (according to the metrics proposed by Robert S. Kaplan and David P. Norton), and methods for estimating component values [13].

Various estimation methods are used to estimate the components of the integral indicator based on information about the survey objects from the database.

Table 2.

Components of the integral indicator and methods of their estimation

№	Components	Type of input information	Type of variable component calculated value	Method of estimation
Characterization of the financial condition of the object of study				
1	Probability of financial disadvantage	structured	categorical, ordinal	logistic regression model
The status identity of the object of study				
2	Status of the object of study in terms of scale	structured	categorical	cluster analysis
3	Status of the object of study as belonging to an abnormal group	structured	categorical	cluster analysis
Characteristics of the external information environment				
4	Media activity in relation to the object of study	weakly structured, quasi–structured and unstructured data	quantitative	semantic analysis
5	Positive tone of references to the subject of the study in online sources		quantitative	semantic analysis
6	Negative tone of references to the subject of the study in online sources		quantitative	semantic analysis
Regulatory requirements for the condition of the object of study				
7	Compliance with the requirements of public authorities	structured	binary or categorical	statistical and index analyses

2.1. Component 1. Probability of financial distress

Represents the probability of an unfavorable financial condition of the research object (bankruptcy, revocation of a license for financial reasons). In order to estimate this probability, a logistic regression model is applied based on financial statements data and their volatility indicators: standard deviation and variance, data on macroeconomic variables, data on public procurement as a supplier or buyer, In general, a logistic regression model takes the form [1]:

$$P\left(Y=1 \mid x, m, v\right)=\frac{1}{1+e^{-z}},$$

$$z=\beta_0+\sum \beta_i x_i+\sum \gamma_j m_j+\sum \varphi_k v_k,$$

where:

$P\left(Y=1 \mid x, m, v\right)$ – the conditional probability of the financial condition of the object under investigation being adverse;

β_0 – constant;

x_i – the variables that characterize the financial condition of the subject of the study;

m_j – variables characterizing the environment external to the object of study (macroeconomic factors);

v_k – non-quantitative indicators of the subject's performance;

β_i, γ_j, φ_k – regression coefficients to be estimated.

2.2. Components 2 and 3. Study object status by scale and abnormal group membership

Represents the clustering results to determine whether the survey object belongs to one of the classes. These components allow us to take into account specific features of all objects of the study type in terms of location, scale, type of activity, etc. The specifics of the obtained cluster of objects are taken into account, all of which allows us to assess more objectively the state of the enterprise in relation to objects from its class.

Clustering algorithms are divided into two types:

1. Hierarchical methods.
2. Non-hierarchical methods.

Hierarchical clustering methods are of two types [14, 15]:

1. Agglomerative (combining).

In this category of methods the initial objects are combined and the number of clusters is reduced [16]. This approach is carried out "bottom-up": creating small clusters and combining them into larger ones.

2. Divisive (decoupling).

Divisive type algorithms are characterized by the initial condition of having one cluster. This initial cluster is divided into smaller clusters. Dividing algorithms work top-down.

The disadvantage of these methods is the computational complexity on high dimensional data. A characteristic feature of hierarchical clustering methods is that observations once in a cluster cannot move to another cluster when further combining (disjoining) objects, in contrast to non-hierarchical methods.

The main distinctive idea of non-hierarchical clustering methods is to determine the center of the cluster and group all objects that are at a distance from the cluster center within a given threshold value [14, 15]. The group of non-hierarchical clustering methods includes algorithms of *k*-means family [16].

For high-dimensional data with an unknown number of clusters, the BIRCH (two-step or two-stage clustering) method based on *k*-means method is proposed. Two-step clustering does not require the number of clusters to be specified, since in the first step the optimal number of clusters is determined, and then the partitioning into homogeneous groups already takes place. This method makes it possible to analyze large amounts of both quantitative and qualitative data and works well with small memory sizes.

The quality of the resulting clustering can be evaluated using the silhouette measure *Sil* [17]:

$$Sil = \frac{1}{N}\sum\nolimits_{c_k \in C}\sum\nolimits_{x_i \in c_k}\frac{b(x_i,c_k)-a(x_i,c_k)}{\max\left(a(x_i,c_k),b(x_i,c_k)\right)},$$

where:

Sil – the overall value of the silhouette measure of clustering of all data;

N – total number of objects in the sample;

C – set of all clusters;

c_k – k-th cluster on the set C;

x_i – i-th object, $i \in [1, N]$;

$a(x_i,c_k) = \frac{1}{|c_k|}\sum\nolimits_{x_j \in c_k}\left\|x_i - x_j\right\|$ –

the average distance from object $x_i \in c_k$ to other objects x_j in that cluster ck (compactness);

$|c_k|$ – number of objects in a cluster c_k;

$b(x_i,c_k) = \min_{c_l \in C \setminus c_k}\left\{\frac{1}{|c_l|}\sum\nolimits_{x_j \in c_l}\left\|x_i - x_j\right\|\right\}$ –

average distance from the site to objects x_j from another cluster c_l: $k \neq l$, k, $l \in [1, C]$.

Silhouette measure *Sil* takes values on the interval –1 to +1, where:

1 – all observations are located exactly in the centers of their clusters;

–1 – all observations are located at the centers of some other clusters;

0 – the observations are located at equal distances on average from the center of their cluster and the center of the nearest cluster.

2.3. Components 4, 5 and 6. Media activity in relation to the research object, positive and negative tone of mentions of the research object in Internet sources

The evaluation of these components is an analysis of unstructured or weakly structured data, predominantly textual. The semantic analysis to assess the meanings of the components characterizing media activity and the tone of the references to the object of research requires a preliminary preprocessing of this data, technical and linguistic data cleaning, compilation of the vocabulary of the words used in the texts.

Tone is the author's emotional attitude towards some object expressed in the text [18, 19]. One way to determine the tonality is to search for the emotional component in the text by the previously formed tonal dictionaries using linguistic analysis. The application of ready-made dictionaries to purified textual data allows us to classify textual units (sentences, words) into three categories: ambivalent, positive and negative. Semantic analysis of media activity, text categorization and application of machine learning techniques require text vectorization.

Vectorization is the process of converting textual documents into a numeric vector. The choice of vectorization method usually depends on a specific case, conditions, available hardware and technological tools. New methods and algorithms that improve vectorization quality and processing speed are constantly appearing and make it possible to introduce natural language processing into a model.

Currently the most popular algorithm implemented in many statistical packages, is the Bag-of-Words. Bag-of-words is a vector representation of an unordered set of words into a vector of dimension n [20–23]. Schematically, the algorithm can be represented as follows.

The whole text can be represented as a set of processed words, that is, individual terms (t_j), which with the help of this algorithm are translated into numerical data from the space R^n.

$$B : \text{words} \to R^n,$$

$$B\,(\textit{'some text in the Internet'}) = (w_{i,1}, w_{i,2}, ..., w_{i,n}),$$

where:

t_j – term j;

w_{ij} – the weight of term j in the document; the weight of the documents is rationed so that $0 < w_{ij} < 1$, для $\forall i$;

n – number of terms in space.

The document is then set up as follows:

$$d = (w_1, w_2, ..., w_{|V|}),$$

where:

d – document vector;

$|V|$ – the number of unique terms in the document.

The weight of a term can be set in several ways:

1. In a binary way:

$$w_i = \begin{cases} 1, t_i \in d \\ 0, t_i \notin d \end{cases}.$$

2. According to the number of occurrences of the term:

$$w_i = n_i,$$

where n_i – the number of occurrences of the term in the document.

3. Term Frequency – TF.

$$w_i = tf(t_i, d) = \frac{n_i}{\sum_{k=1}^{|V|} n_k},$$

where:

tf – thermal frequency;

n_i – the number of occurrences of the term in the document;

$\sum_{k=1}^{|V|} n_k$ – number of terms in the document.

4. Term Frequency – Inverse Document Frequency (TF–IDF).

Representation in the form of two parameters:$w_{ij} = tf_i \cdot idf_i$, where tf_{ij} – is the ratio of the number of terms t_i on paper d_j to the total number of terms in this document, $iidf_i$ – the number inverse of the number of documents in which the term occurs t_i. Thus, the more often a word occurs in this document, but less often in all documents in general, the greater the weight of that term in the document:

$$tf(t_i, d) = \frac{n_i}{\sum_{k=1}^{|V|} n_k},$$

$$idf(t, d) = \log \frac{|D|}{|d_i \supset t_i|},$$

where:

$d_i \supset t_i$ – the number of documents in which it occurs t_i;

$|D|$ – the number of documents in the enclosure.

The weight is then calculated as follows:

$$w_i = tf - idf(t_i, d, D) = tf(t_i, d) \cdot idf(t, d).$$

After vectorization, semantic text analysis algorithms are applied to determine tone, main themes, media activity, etc.

To calculate the values of the components, statistical methods are used to summarize the information about the object of study, e.g. by directly counting the occurrence of positive and negative words, the overall tone of the text is determined.

2.4. Component 7. Compliance with government requirements

This component is defined as a binary or ordinal indicator calculated using indices and statistical indicators. It represents an estimate of the number of irregularities in the activity of the object of study, in case normative and threshold values are given by the control or oversight state authorities.

A consolidated representation of the above is the information-logical model of express analysis of the compliance of the socio-economic state of the object of research with the requirements of control and supervisory authorities (*Fig. 1*). In *Fig. 1*, stage 3, which is key in the algorithm for calculating the integral indicator, is shown in general form. Detailed elaboration of stage 3 of the information-logical model is presented in *Fig. 2.* In this stage, the components of the integral index are evaluated and the values of the integral index itself are calculated depending on the values of each component in the set.

The interquartile range of IQR for the sample size n is proposed to transform the values of the component which characterize the media activity (component 4), the tone of the reference about the research object in Internet sources (components 5 and 6) and compliance with the requirements of public authorities (component 7). Here:

$F_n(x)$ – selective distribution function;

$IQR = Q_3 - Q_1$, where $Q_3 = 0.75$; $Q_1 = 0.25$.

The proposed information and logical model was tested on the basis of data from a group of industrial enterprises and financial sector enterprises.

A rapid analysis was conducted to match the need for financial assistance for 506 industrial enterprises registered in Moscow and the feasibility of its provision to federal and regional authorities. The express analysis was based on

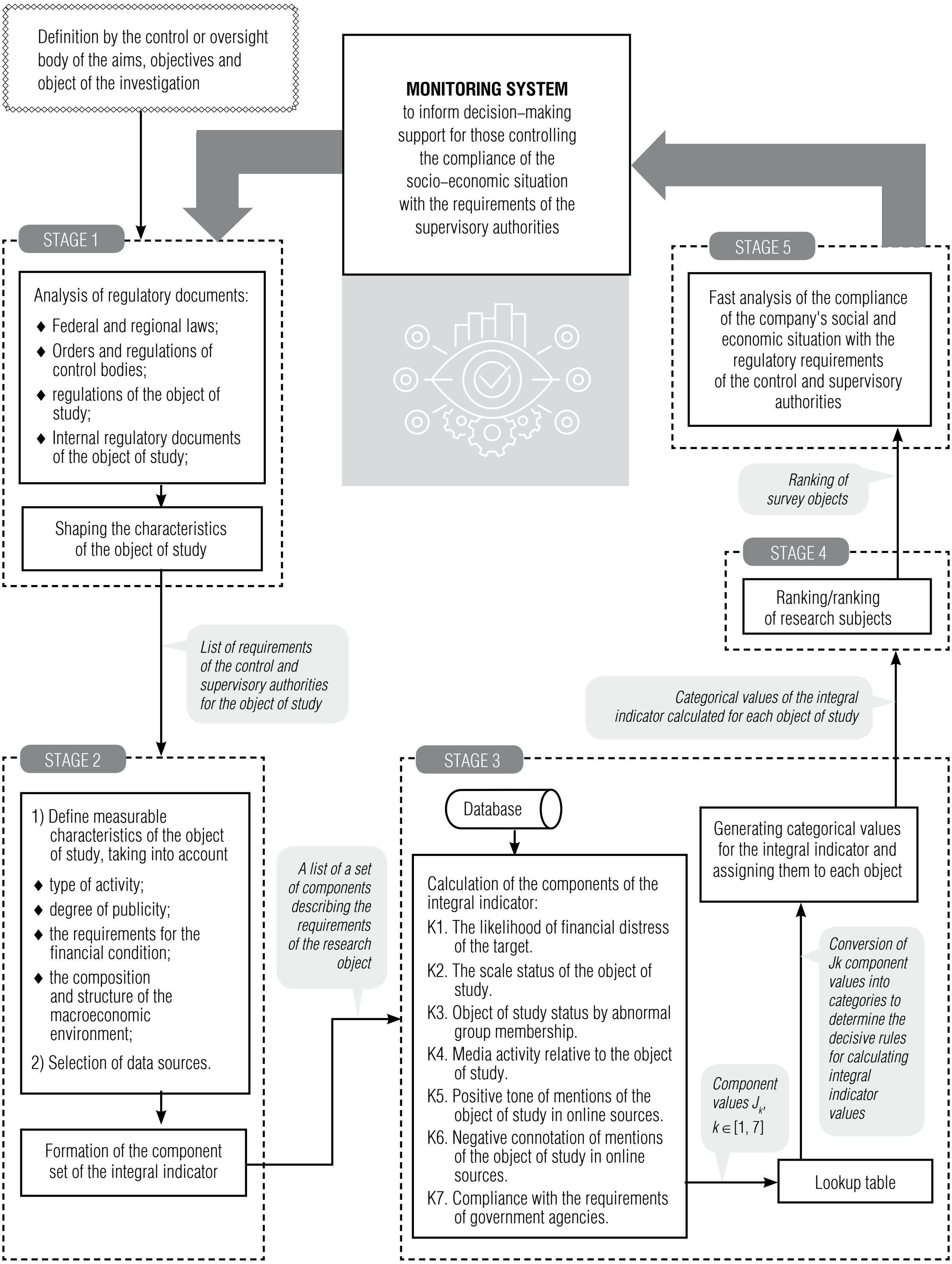

Fig. 1. Information–logical model of the algorithm for calculating the components of the integral indicator.

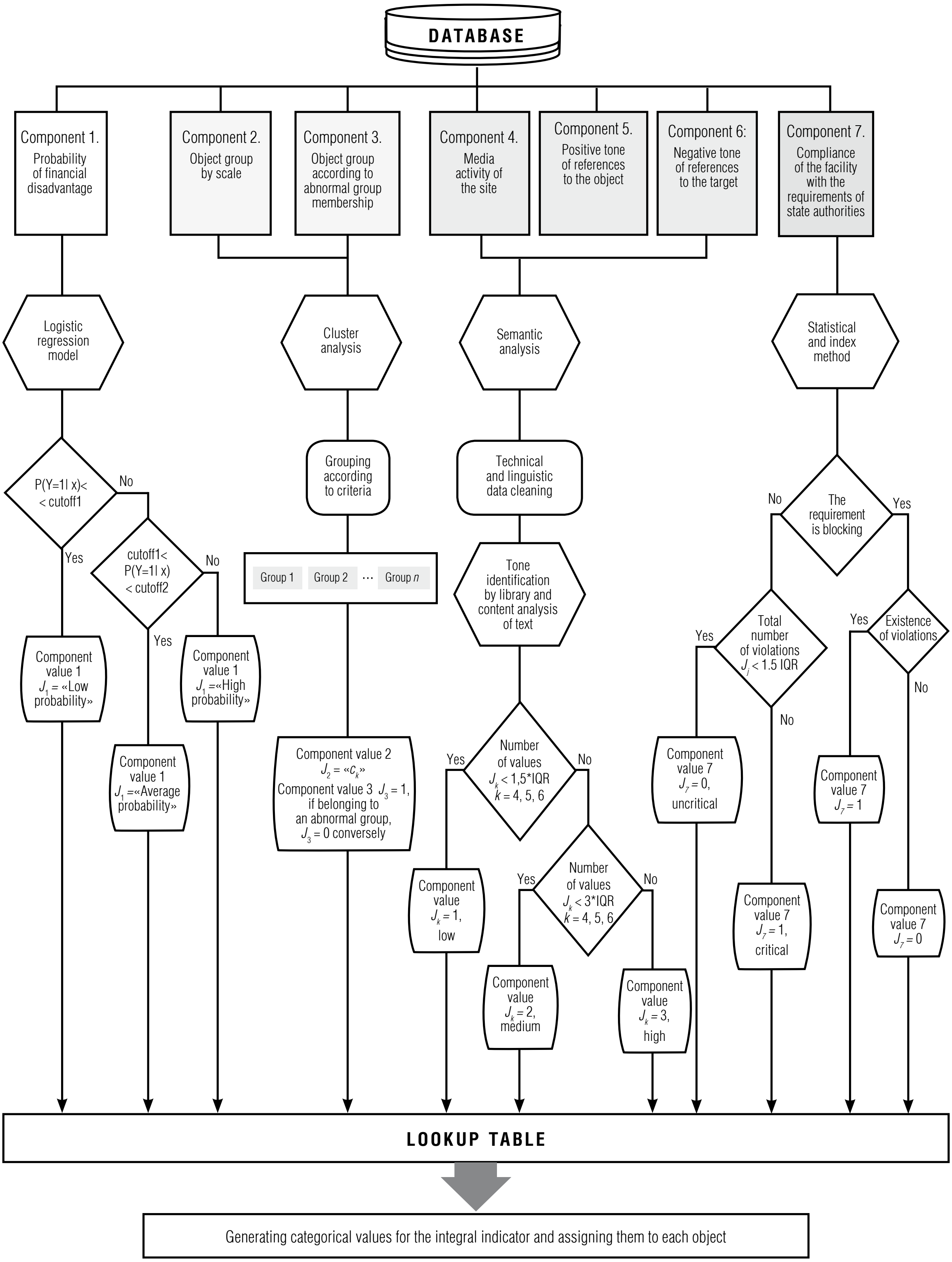

Fig. 2. Detailed step 3 of the information–logical model of the algorithm for calculating the components of the integral indicator.

open data for 2016, 2017 and 2018. The results obtained were in line with the actual data for the following year on the assignment of subsidies and benefits by the Moscow City Government [24].

The proposed conceptual model of express analysis of the compliance of the socio-economic condition of the object of research with the stated requirements on the part of control and supervisory authorities has been tested to assess the socio-economic condition of a commercial bank. The controlling body in this case is the Central Bank of Russia — the supervisory authority in the banking sphere. In accordance with the CBR requirements for bank reliability, the values of the four components of the integral index were obtained and its value for each bank was calculated. The predictive ability of the constructed model was confirmed by their actual state as of March 2020 [1].

Conclusion

This article suggests an information and logical model for express analysis of compliance of the social and economic condition of the object with the requirements of the control and supervisory bodies with the use of open public data. The proposed information-logical model is based on the concept of using an integral indicator for rapid analysis of compliance of the socio-economic condition of the object, regardless of its type of requirements imposed on it by control and supervisory authorities.

The classification of information sources and methods of processing them depending on the type of data is given.

An algorithm is proposed for calculating the possible components allocated by the authors relating to the four blocks of input information types, types of variables of the calculated value of each component (in accordance with the metrics proposed by Robert S. Kaplan and David P. Norton), and methods for estimating component values.

The developed conceptual model has been tested to carry out a rapid analysis of the compliance of the socio-economic condition of two different types of facilities with the requirements imposed on them by the supervisory authorities on samples of 506 industrial enterprises [24] and 111 banks [1]. ■

References

1. Bogdanova T.K., Zhukova L.V. (2021) The concept for valuation the position of the control object based on a universal complex indicator using structured and unstructured data. *Business Informatics*, vol. 15, no. 2, pp. 21–33 (in Russian). http://doi.org/10.17323/2587-814X.2021.2.21.33

2. Krichevskiy M.L. (2019) Methods of machine learning in choosing a strategy of an enterprise. *Russian Journal of Innovation Economics*, vol. 9, no. 1, pp. 251–266 (in Russian). https://doi.org/10.18334/vinec.9.1.40093

3. Opekunov A.N., Kuzmina M.G. (2019) Principles of forming models for forecasting the probability of bankruptcy of enterprises using machining elements. *Models, Systems, Networks in Economics, Technology, Nature and Society*, no. 4, pp. 24–31 (in Russian).

4. Kasevich V.B. (1977) *Elements of general linguistics*. Moscow: Nauka (in Russian).

5. Savenkov P.A. (2019) Using methods and algorithms of machine learning in management decision support systems. *Bulletin of Science and Education*, nos. 1–2 (55), pp. 23–25 (in Russian). https://doi.org/10.24411/2071-6168-2019-10207

6. Popova S.V., Khodyrev I.A. (2012) Keyword extraction. *Scientific and Technical Bulletin of Information Technologies, Mechanics and Optics*, no. 1 (77), pp. 67–71 (in Russian).

7. Eliseeva E.N. (2019) Financial instruments for assessing the insolvency of industrial enterprises. *Region: systems, economy, management*, no. 3 (46), pp. 132–140 (in Russian).

8. Medvedev D.A. (2019) Big data: the reasons for their emergence and how they can be used. *Science and Education Today*, no. 4 (39), pp. 14–16 (in Russian).

9. Pang B., Lee L. (2008) Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, vol. 2, nos. 1–2, pp. 1–135. https://doi.org/10.1561/1500000011

10. Morozov A.N. (2018) Alternative sources of statistical information as the basis for political decision making. *Problems of State and Municipal Management*, no. 2, pp. 50–70 (in Russian).

11. Puzanov A.S., Trutnev E.K., Markvart E., Popov R.A., Safarova M.D. (2017) *Strategic planning and urban regulation at the municipal level*. Moscow: Delo (in Russian).

12. Andreeva N.A., Ugrimova S.N. (2019) To the question of the application of statistical methods of integral estimation of effectiveness of system of management of industrial enterprises. *Accounting and Statistics*, no. 1 (53), pp. 42–49 (in Russian).

13. Norton D. P., Kaplan R.S. (2008). *Balanced scorecard*. Moscow: Olymp-Business.

14. Chugunov V.R., Zhukova L.V., Kovalchuk I.M., Kovaleva A.S. (2017) Mathematical methods of data grouping for making management decisions in planning tasks. *Actual Problems of System and Software Engineering 2017. Proceedings of the 5th International Conference on Actual Problems of System and Software Engineering Supported by Russian Foundation for Basic Research. Project #17-07-20565*, pp. 333–341 (in Russian).

15. Baresyagin A.A., Kupriyanov M.S., Stepanenko V.V., Kholod I.I. (2007*) Data analysis technologies: Data Mining, Visual Mining, Text Mining, OLAP*. 2nd edition. St. Petersburg: BXV-Petersburg (in Russian).

16. Shalymov D.S. (2008) Stable clustering algorithms based on index functions and stability functions. S*tochastic optimization in computer science*, vol. 4, pp. 236–248 (in Russian).

17. Kaufman L., Rousseeuw P. (2005) *Finding groups in data: An introduction to cluster analysis*. Wiley-Interscience.

18. Semina T.A. (2020) Sentiment analysis: modern approaches and existing problems. *Social sciences and humanities. Domestic and foreign literature. Ser. 6, Linguistics*, no. 4, pp. 47–64 (in Russian).

19. Liu B. (2010) Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition* (eds. N. Indurkhya, F.J. Damerau). London: Chapman and Hall/CRC.

20. Bolshakova E.I., Vorontsov K.V., Efremova N.E., Klyshinsky E.S., Lukashevich N.V., Sapin A.S. (2017) *Automatic text processing in natural language and data analysis. Tutorial*. Moscow: HSE (in Russian).

21. Demidova L.A., Stepanov M.A. (2019) An approach to solving problem of the structural transformations detection in the time series' groups. *Cloud of science*, no. 2. pp. 201–226 (in Russian).

22. Popova S.V., Khodyrev I.A. (2012) Extraction of keyword combinations. *Scientific and Technical Bulletin of Information Technologies, Mechanics and Optics*, no. 1 (77), pp. 67–71 (in Russian).

23. Krasnyansky M.N., Obukhov A.D., Solomatina E.M., Voyakina A.A. (2018) Comparative analysis of machine learning methods for solving the problem of classifying documents of scientific and educational institution. *Bulletin of Voronezh State University*, no. 3, pp. 173–182 (in Russian).

24. Zhukova L.V. (2021) Express-analysis of the state of industrial enterprises of Moscow using the universal comprehensive indicator. *Economic Science of Modern Russia*, vol. 4 (95), pp. 89–96 (in Russian).

About the authors

Tatiana K. Bogdanova

Cand. Sci. (Econ.);

Assistant Professor, Department of Business Informatics, Graduate School of Business, National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: tanbog@hse.ru

ORCID: 0000-0002-0018-2946

Liudmila V. Zhukova

Assistant Professor, Department of Applied Economics, Faculty of Economic Sciences, National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: lvzhukova@hse.ru

ORCID: 0000-0003-1647-5337