Use of big data technologies in animal husbandry

Alexander Sokolov[1*], *Vera* Batova[2], and *Andrey* Volkov[2]

[1]Saratov Branch of the Institute of State and Law of the Russian Academy of Sciences, Chernyshevskogo str., 135, Saratov, 410028, Russia
[2]Penza State Technological University, Baidukov's passage/ Gagarin street, 1a / 11, Penza, 440039, Russia

Abstract. The growing interest in big data is driven by several identified factors. Nowadays, humanity uses a large number of equipment that generates multi-format signals and the amount of the generated information is growing exponentially, but the biggest part of this data is unstructured information. In this regard, choosing relevant information and achieving correct interpretations of its flow are becoming more relevant and complex issues. The technologies of Big data allow processing huge volumes and diverse compositions of information that can be regularly updated and located in different sources. The use of theses technologies leads to an increased work efficiency and competitiveness, and develops new knowledge. In this paper, the purpose of the study is to investigate and identify the opportunities of business processes' digitization in agricultural production.

1 Introduction

In our modern world, many sources of big data can be identified:
• Incoming data from measuring devices,
• Events from radio frequency identifiers,
• Streams of messages from social media,
• Meteorological data,
• Earth remote sensing data,
• Data flow indicating the subscribers' location in a cellular network,
• Information from audio and video recording devices.

According to an IDC study, investments in IoT technologies will reach 1 trillion dollars by the end of 2020, which is a clear indicator that the number of "smart" and connected devices will grow. At the same time, the number of devices and sensors that collect and transmit data will grow exponentially. As a result, the flows of information generated by these connected devices will also grow.

Big Data technologies include more than just analyzing big amounts of information. The problem faced is not the quick growth of data volumes, but the format in which they are represented that does not correspond to the traditional structured database format (DB)

* Corresponding author: aysockolov@mail.ru

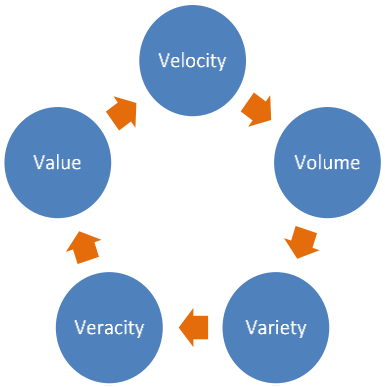
– that includes blogs, video recordings, text documents, machine language, or, for example geospatial data. The information is generally stored in many different repositories or data processing centers (DPCs), sometimes in other countries or external organizations. As a result, users may have big volumes of data without having the necessary tools to organize, classify, group and establish correlations between the different information that lead to meaningful conclusions. Besides, the data is updated more and more often, and as a result, traditional methods of information analysis cannot keep up with the huge volumes of constantly updated data, which ultimately require Big Data technologies (Fig. 1).

Nowadays, it is important not only to be able to accumulate information, but also to extract

business benefit from it. The industries working directly with the consumers (telecommunications, banking, retail) were the first to come to this conclusion. At this point, the interaction processes reached a new level, allowing the establishment of communication between different devices using tools of augmented reality, which opened up new possibilities to optimize the companies' business processes.

A scientific approach was needed to systematize, group and classify information, and developpe algorithms, which led to the emergence of Data Science as a professional activity with an effective and reliable search for patterns, that extracts knowledge from, and presents it in a form suitable for processing by people, software systems and control devices. The results of this technology can lead to optimal and well-founded management decisions.

In fact, data science is a collection of different specific disciplines oriented towards data analysis and search for optimal solutions. Mathematical statistics was the only discipline applied before introducing machine learning and artificial intelligence adding computer science and optimization as analysis methods.



Characteristics	Traditional Database	Big Data
Information volume	From Gigabyte to Terabyte	From Petabyte to Exabyte
Storage method	Centralized	Decentralized
Data structuring	Structured	Multi-structured or not structured
Data storage and processing model	Vertical model	Horizontal model
Data interconnection	Strong	Week

Fig. 1. Visual representation of the Big Data technology.

The main tools of data science are mathematical and algorithmic methods that are optimized to efficiently identify complex models. This means that their use involves building complex analytical models based on an array of data in order to extract new knowledge (Fig.2).

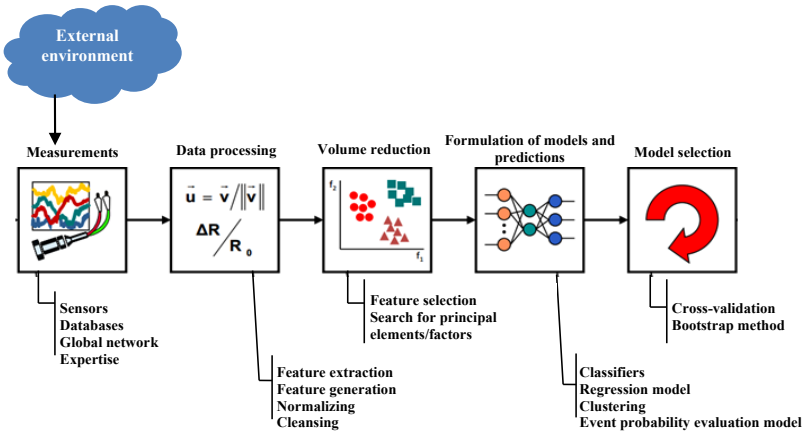


Fig. 2. Model building scheme based on Big Data technology.

There are many different methods for analyzing datasets using statistics and computing tools like machine learning for example. Furthermore, some of these methods are not necessarily applicable exclusively to Big Data, and can be used for smaller data arrays (A/B testing, regression analysis). We should note that using multiple analysis methods on a data array leads to more precise and relevant conclusions at the output.

2 Materials and methods

A/B testing. A method in which a control sample is compared to other samples in order to identify the optimal combinations of indicators. This can be used, for example, to identify marketing proposals that induce the best consumer reactions. The large number of iterations allowed by Big Data can guarantee a statistically reliable result.

Association rule learning. A set of methods that aim to identify correlations, between variables in large data sets. This method is used in data mining.

Classification. A set of methods that can predict consumer behavior in a segment of a particular market (purchase decisions, outflow, consumption volumes…). This method is used in data mining.

Cluster analysis. A statistical method that used to classify objects in groups, based on previously unknown common characteristics. This method is used in data mining.

Crowdsourcing. A method to collect data from a large number of sources.

Data fusion and data integration. A set of methods that analyze user comments on social media, comparing them to sale's results in real time.

Data mining. A set of methods used to identify the most receptive consumers to a certain promoted product or service, or the characteristics of the most successful employees. it also helps when it comes to consumer behavior prediction.

Ensemble learning. This method uses a set of prediction models, thereby improving the quality of forecasts.

Genetic algorithms. In this method, the results are presented as "chromosomes", which can combine and mutate as in the process of natural evolution based on the concept “survival of the fittest”.

Machine learning. A discipline in computer science (formerly called “artificial intelligence”), aiming to create self-learning algorithms by exploiting experimental data analysis.

Natural language processing (NLP). A set of methods borrowed from computer science and linguistics to recognizing human natural language.

Network analysis. A set of methods applied to analyze the links in a social network. The application of this approach on social media, allows to identify the links between individual users, companies and communities.

Optimization. A set of digital methods used to redesign complex systems and processes in order to improve one or more elements. This method is used to make strategic decisions, such as investment analysis, or selecting a product range composition to be introduced on the market.

Pattern recognition. A set of methods using self-learning elements to predict consumer behavior patterns.

Predictive modeling. A set of methods used to create mathematical models in order to forecast predetermined probable scenarios and their outcomes. For example, it can be used to analyze a CRM system's database to identify possible factors that could eventually push subscribers to change providers.

Regression. A set of statistical methods used to identify patterns between the alterations of a dependent variable and one or more independent variables. This method is often used to forecast and predict outcomes, and also in data mining.

Sentiment analysis. This method uses natural human language recognition technologies to assess consumer feelings. Using this method, we can isolate specific messages related to a predefined subject of interest (a product for example). Then assess whether the subject is positively or negatively judged, as well the degree of the expressed emotions, etc.

Signal processing. A set of methods used in radio engineering to recognize and analyze a determined signal in a noisy background.

Spatial analysis. A method used to analyze the spatial data, partially based on statistics - topology, geographic coordinates and geometry of objects. In this case, the source of Big Data is the geographic information systems (GIS).

Statistics. The science of collecting, organizing and interpreting data by designing surveys and conducting experiments. Statistical methods are often used to evaluate connections between separate events.

Supervised learning. A set of methods based on self-learning technologies that can identify functional connectivity in an analyzed data set.

Simulation. Modeling the behavior of a complex system is often used to predict and develop various planning scenarios.

Time series analysis. A set of methods used to analyze repetitive data sequences using statistics and digital signal processing. One of its most common uses is to track the stock market or patient’s morbidity.

Unsupervised learning. A set of methods based on machine learning technologies, used to identify the functional connectivity hidden in the analyzed data sets.

Data Visualization. A method that aims to graphically present the results of Big Data analysis in the form of diagrams or animated illustrations to facilitate their interpretation. (Fig. 3).

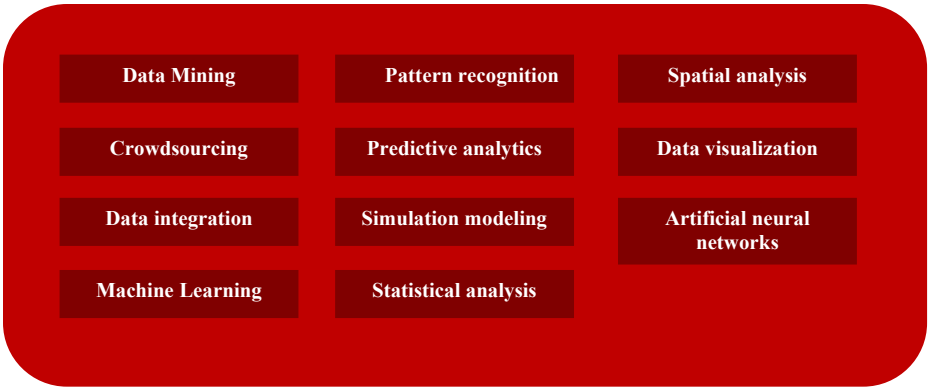


Fig. 3. Big Data Analysis Methods.

In general, Big Data can be defined as:

- Storage technologies for large volumes of structured and unstructured data;
- Data processing technologies;
- Data quality management;
- Technologies of data delivery to supply consumers.

When working with Big Data, the results are usually obtained in the cleansing process using sequence modeling: first, a hypothesis is made and a model is built (statistical, visual or semantic model). Using the built model, we verify the accuracy of the advanced hypothesis, before moving to the next one. To conduct this process, we need to interpret visual values, or to formulate interactive knowledge-based queries, otherwise it's necessary to developpe an adaptive "machine learning" algorithm to achieve the desired result. We note that the lifetime of such an algorithm could be quite short.

An effective big data analysis, can help companies gain an important competitive advantage and achieve their fundamental goals. Nowadays, these companies are using various tools and technologies such as Python to analyze their sets of data. More and more companies focus their efforts on identifying the causes of certain occurring events, in this case, predictive analytics can be used to identify trends and predict probable future scenarios.

As a matter of fact, the massive spread of the above-mentioned technologies and the new ways of using the variety of Internet devices and services, have served as a starting point for the introduction of big data into almost all spheres of human activity, including scientific research, commercial and governmental activities.

As a result of the significant computing power's growth and the improvement of storage technologies, the use of big data analysis is gradually becoming accessible for small and medium-sized businesses and is no longer exclusively reserved for large companies and research centers. This accessibility is also facilitated by the development of the cloud computing.

3 Results

We must mention, that for a long-time agriculture was not an activity that attracts investors, because of its long production cycle which is subject to natural risks and high yield losses; it is also very difficult to automate the biological processes occurring during the cultivation of a specific product; it is practically impossible to predict an increase in productivity; innovation can also be a big risk to take. Until lately, the use of information technologies in

agriculture was mainly restricted to financial management and business transactions' monitoring.

The quick evolution of the technologies used in agriculture, was mainly affected by the global scientific and technological advancements, as well as the growth of institutional investors and the size of financial resources. On the other hand, a big leap was made when tech companies learned how to control the full cycle of crop or animal production using smart devices, which drew attention to agriculture. These devices transmit and process the current parameters of each object and its environment (using equipment and sensors that measure the parameters of the soil, plants, microclimate, animal characteristics, etc.), using wireless communication channels between them and external partners.

It is now possible to combine these objects into a single network using the appropriate digital platforms. This facilitates the exchange and management of data, using the "Internet of Things" and the high computing capacity. Software and cloud platforms also developed rapidly, allowing the automatization of a large number of agricultural processes by creating a virtual (digital) model of the entire production cycle interconnected with the production line. This digitization of the agricultural production has made it possible to plan work schedules with mathematical precision, and avoid losses by taking emergency measures against occurring threats, it also allows to calculate the possible yield, the cost of production and the profit. The production line of agricultural products can be represented as follows (Fig. 4).

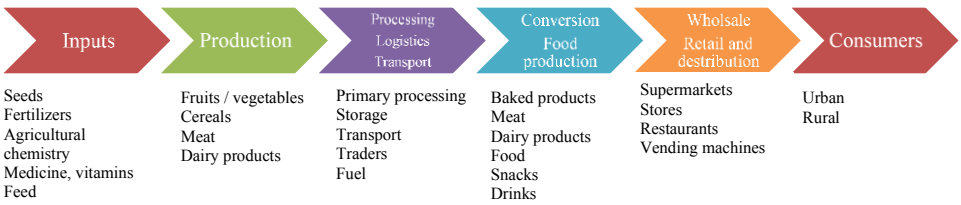


Fig. 4. The production line of plus-value in the agricultural industry.

It can be difficult to represent all the business processes in this chain in the form of a mathematical model or computer-recognizable data. For this reason, the business processes must be precisely described and formulated. Only formulated processes can be used to identify causal interconnections between them. This means, only formulated business processes can be algorithmized and presented in the form of an economic and mathematical model.

It is important to note that traditional agricultural production is still influenced by a large number of poorly predictable factors like weather making this process very difficult to algorithmize.

The flow of digital information coming from measuring devices to characterize each of the business processes, will contain an unstructured multi-format data presented by several method. This makes the data difficult to process in a comparable form using one single digital platform.

The first step taken by the authors to describe and model the business processes illustrated in this diagram, was the pull method. This method focuses on the value that a business process creates for its consumers.

While the organizational structure of a system defines the subsystems (e.g., departments, positions), and also contains information about the tasks in each subsystem [9], the organization of a business process is a tool to track tasks, as well as the temporal

and spatial aspects of their execution (who does what, when and how). The small elements contributing to the execution of each individual task are at the same time the main components of the process itself [10]. Pull modeling is about detailing this individual tasks.

Therefore, while modeling the processes, it is necessary to record every single significant input and output. At the same time, it is recommended to classify process inputs into:

- Transforming inputs,
- Convertible inputs,
- Control inputs.

Fig. 5 illustrates the inputs and outputs of the process, arranged according to the SADT methodology and the IDEF0 notation.

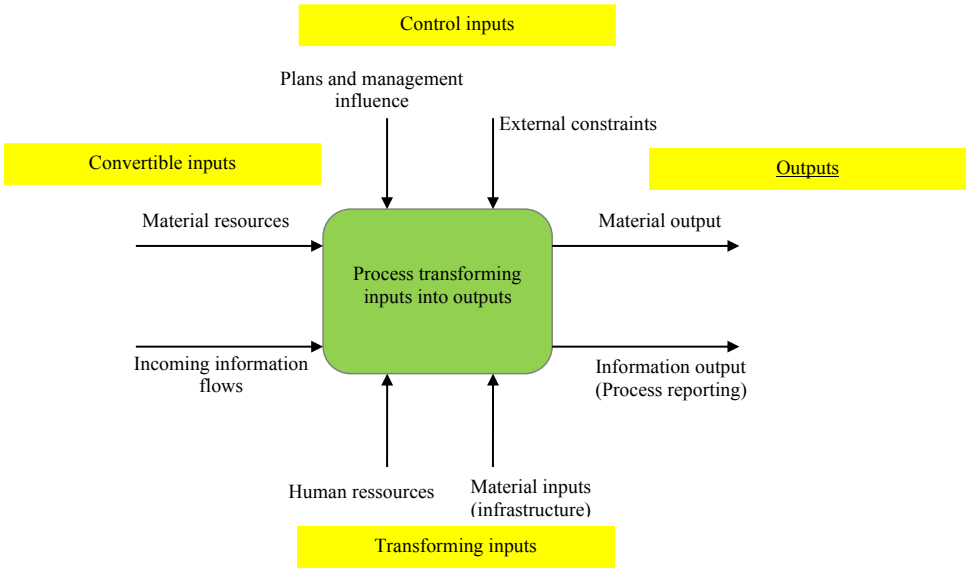


Fig. 5. Business process modeling.

The authors decided that it was essential to present the business processes' information flows in a tabular form to achieve a complete presentation. Meanwhile, it is recommended to start from the final operations, associated to the final results of the process. According to the authors, the information flow from the main operations of each business process in the production line of plus-value, include the following types (see Annex 1).

In the first business process, the final results are information about the quality of the products obtained by growing seeds, or raising young animals. And also, information about the associated raw materials used (animal feed, medicines and vitamins for animals, mineral fertilizers and agrochemicals for plants). The source of his information can be a warehouse or a farm, and can also include the time and the quantity of resources related to the process. The information must be compared to the planned objectives.

In the second business process, the initial agricultural raw materials are converted into finished products or semi-finished products. For example, cultivated seeds will be harvested and transferred (sold) for storage, processing, or sale; the raised animals will be transferred (sold) to a slaughterhouse, the cow milk will be delivered to dairy plants or agricultural producers who will distribute it themselves. Therefore, these processes need to be controlled by various measuring instruments that can generate information about soil

composition, seed quality, plant and animal diseases, periods of fertilization and feeding and the volume of the reached agricultural products.

The third business process involves the delivery of semi-finished products for further processing using means of transport that can be owned by the supplier or the client. The data that characterizes this business model must include information about fuels and lubricants, spare parts, and cargo details.

In the fourth business process, semi-finished products are converted into finished products for wholesale or retail. The data in this case, must contain information (gathered from control sensors on each step of the transformation process) about the finished products' volume and the deviations of the technological process.

In the fifth business process, trading operations take place to sell the finished products (advertising, promotion, etc.). Usually, only stores are affected by the selling fees, but if agricultural producers sell their own finished products, the selling fees must be taken into consideration. Information relating to these operations can be collected from cargo loading / unloading documents, expected losses, unfinished production volumes and price changes.

In the sixth business process, payments are made (in cash, bank transfer, by card, etc.) by the sellers of the finished products. If the products are transferred for sale, then there must be a delay between the product's receiving time and the payment. Which means, the data must contain information about the amount of money, the receiving time, and the total of customer debt.

Such volumes of information require a scientific approach to develop algorithms, collect, organize, store data and ensure rapid access. Each business process must have a mathematical description that precisely characterizes its essence. The continuous flow of information requires systematization in order to have sufficient perception, quick identification of the deviations from plans and standards, and interactive responses that adapt to changes. Such volumes of information can only be processed using appropriate computer programs, methods and technologies.

As an example of the production processes' digitization in Russian companies, we can take the breeding center of the Damate group in the Tyumen region. The flow of information arriving to the Center's server includes satellite field images, fertilization maps, information from movement sensors and animal monitoring. Using this flow of information, the different situations are assessed and the Rapid Response Center operates 24 hours a day.

Streams from the cameras installed in the premises is also transmitted to the situation assessment center, as well as all information about production processes. Furthermore, the Centre's employees record every day all violations or dysfunctions including absenteeism and equipment placement. In addition to that, every employee can report an emergency situation using an internal hot line.

One of the important missions of the Center is to identify systematic non-compliance to planned schedules. The planning of the production processes is compared to the results obtained from calculations, which makes it possible to conclude whether it has been respected or not. This is done online using email, instant messaging and phone calls.

The Center has three main operating scenarios. The first scenario is to process the daily collected information and identify hidden problems. Based on the daily results, reports are generated and sent in electronic form every day before 6 a.m.

The second scenario is when an irregularity happens in a given time and location. "In addition to the report, we also send emails and instant messages to all responsible employees, - explains Andrey Zaitsev. - we have messaging groups, which bring together all the people concerned, and in the case of taking a key decision or a quick reaction, that helps a lot".

The third scenario – “Fire situation”, when everyone in the company is overwhelmed by the emergency. This happens when a decision needs to be made immediately. The center’s employees do not solve the problems themselves, but in an emergency cases there is an alert algorithm covering 90% of the outcomes that may arise. “Our task is to find someone who can solve the problem and do it quickly” notes Zaitsev.

Fig. 6. Modern dairy farm in the "Damate" group.

The software used for herd monitoring:
- TMR Tracker, for feeding control,
- SELEX, for herd control,
- DAIRY PLAN, for milking control,
- SMARTBOW, to monitor cows using movement sensors,
- HYBRIMIN, for ration preparation.

The data is combined in the UNIFORM AGRI framework that makes it possible to monitor all the details in a single system, without the need to switch between different programs. The SMARTBOW program is convenient for vets, inseminators and the herd managers, especially when looking for a specific animal.

We should note that the program runs on a tablet that can be taken anywhere on the premises. The data collected from all the sensors installed on each cow can be viewed on the map. If for example, an animal has been mentioned in a report and needs the supervision of a specialist (in case of a lack of rumination for example), it will be possible to locate the animal on the map, which makes it easier to find it in the herd. The program also allows to see a graphic illustration of an animal's activity during the last 24 hours, or to perform an analysis based on specific parameters for the last three days in order to evaluate the health condition. If an animal expresses deviation from the fixed norms, the program automatically detects it and reports it immediately. The installed sensors are programmed to read chewing movements.

4 Conclusion

In the new chapter of analytics’ development, “Big Data” has become a two-level processing model. The first level is the traditional Big Data analysis, where large amounts of information are analyzed in non-real time. On the other hand, the new second level offers

the possibility to analyze relatively large amounts of information in real time, primarily using "in-memory" analysis technologies. The second level offers at the same time, both new and traditional methods of data analysis, in a way to achieve an "on-the-fly" analysis and react to events as they occur, which opens up new business opportunities.

Nowadays, Edge computing is gaining more and more popularity thanks to the sensors themselves. The development of this technology will continue mainly as a result of IoT that covers basic computing systems. Edge Computing can offer to companies the ability to store broadcasted data near the source and analyze it in real time. Edge computing is also an alternative to Big Data analytics, which requires high performance storage devices and significantly more network bandwidth. With the growing number of devices and sensors, more businesses are choosing Edge Computing for its ability to solve bandwidth, latency and connectivity issues. Furthermore, combining Edge and Cloud technologies forms a synchronized infrastructure that can minimize the risks associated to data analysis and management.

On April 17, 2020, the state-owned company Rostec announced that it had signed an agreement with the Ministry of Agriculture of Russia, in order to introduce digital technologies in the agro-industrial complexes.

The document mentions the implementation of digitization projects in the agricultural sector, to stimulate the development of public-private partnerships and increase the export potential of the Russian agro-industrial sector.

As announced by the industrial director of Rostec, Sergey Sakhnenko: "The potential of digitization in agriculture is one of the highest in all economic sectors. About 70% of farms in the United States, Canada and Western Europe are already using "smart" technologies in agriculture. In Russia, demand is only increasing. To achieve maximum effect, it is important to introduce complex solutions for process automation in agro-industrial complexes, instead of introducing separate systems. This can have a synergistic effect and lead to an increase in agricultural productivity".

Rostec's candidate technologies that can be implemented in agro-industrial complexes, include Farm Management software systems, Robotic systems, Unmanned agricultural machinery, monitoring agricultural facilities using drones and Precision Agriculture technologies based on IoT.

The reported study was funded by RFBR, project number 20-011-00740 "Development of a model of regulation of legislative imbalance in the field of veterinary medicine"

References

1. A. I. Voronova, A.V. Levenets, Information Technologies of the XXI Century. Collection of Scientific Papers, 448-456 (2019)
2. A. A. Grabar, E. V. Karanina, Economics and Management: Problems, Solutions, **5(2)**, 11-15 (2018)
3. E. Klyukhina, O. Luchina, T. V. Lesina, Bulletin of the Educational Consortium Central Russian University. Information Technology, **1(7)**, 58-62 (2016)
4. T. A. Soldatenko, S. R. Yessimzhanova, Economics: Strategy and Practice, **15(2)**, 107-113 (2020)
5. A. Yu. Zakusilova, International Journal of Applied Sciences and Technologies Integral, **4(1)**, 21 (2019)
6. A. A. Petrov, Trade Policy, **3(11)**, 46-74 (2017)
7. S. S. Chagin, Professional Education. The Capital, **7**, 27-30 (2018)

8. V. D. Churakov, *Big Data and jurisprudence: are we on the same path?* In Proceedings of the VII International Scientific and practical Conference. Ser. "Electronic legislation" Scientific ed. by N. A. Sheveleva, pp. 136-144, 2017
9. S. M. Doguchaeva, Economy. Business. Cans, **1(22)**, 96-104 (2018)
10. D. V. Kalimbet, Central Scientific Bulletin, **3**, **№ 9S(50S)**, 26-27 (2018)
11. V. A. Fedotov, Improving the methodology for assessing the technological properties of grain and predicting the quality of bakery and pasta products from wheat flour, PhD Thesis (Technical Sciences, Orenburg, 2020)
12. D. S. Kurochkin, Innovative mechanism for improving the efficiency of implementing a process approach to enterprise management, PhD Thesis (Moscow, 2009)
13. S. Mitrovich, Business Informatics, **4(42)**, 40-46 (2017), DOI: 10.17323/1998-0663.2017.4.40.46