Development of a methodology for evaluating the efficiency level of monitoring agroecosystems using big data technologies

Aleksey Petrov[1,*], *Leonid* Andreev[2] and *Sergey* Grigorishin[3]

[1]Norilsk State Industrial Institute, 663310, 50 years of October str. 7, Norilsk, Russian Federation
[2]Northern Trans-Ural State Agricultural Universtity, 625003, 7, Republic st. Tyumen, Russian Federation
[3]Tyumen State University, 625003, 6, Volodarskogo st., Tyumen, Russian Federation

Abstract. The article discusses the theory of monitoring agroecosystems for the effectiveness of using Big Data technologies. The relationship between the agricultural areas of the Tyumen region, the Big Data sources available in them, and the technologies for working with Big Data obtained from sources are described. The article also developed a methodology that makes it possible to assess the level of effectiveness of monitoring agroecosystems using Big Data technologies, based on the result of which a strategy for development of the region as a whole and its agroecosystems, in particular, is formed in terms of equipment with information technologies. The methodology presented in the article is formed on the basis of an engineering ontology, which in the future is able to lower the degree of the human factor in global and local monitoring of agroecosystems for the effectiveness of using Big Data technologies.

1 Introduction

The monitoring and assessment of natural ecosystems primarily implies development and operation of such observation systems, on the basis of which it is possible to assess their condition. The assessment, in turn, is primarily the result of an analysis of the quantitative characteristics of ecosystems, which are dynamic, since they have the property of variability, relative to the impact of external factors (the name of the factors depends on the type of ecosystems). Agricultural ecosystems are commonly referred to as agroecosystems.

Such monitoring of agroecosystems [1], on the one hand, is of great economic importance for the region [2], since, based on its result, the agroecosystem manager can develop a strategy and take measures (crop conservation, pest control, and etc.), on the other hand, it is of great ecological importance, as agroecosystems are part of larger natural ecosystems.

* Corresponding author: darker2012@yandex.ru

2 Methods

In the existing realities, monitoring of agroecosystems is not possible without the use of technologies for processing large amounts of data [3], since the agroecosystem is, first of all, dynamic, which means that the manager of the agroecosystem is interested in obtaining information about its state as quickly as possible.

This, in turn, means that it is possible to develop a methodology for assessing the readiness of an agroecosystem to analyze the data obtained about its state using Big Data technologies, or in other words, a methodology for assessing the level of effectiveness of monitoring agroecosystems using Big Data technologies. This implies that if ten sensors of soil acidity are located on some hypothetical agricultural field (agroecosystem), then using Big Data technologies will be ineffective, since the data obtained can be processed in simple Microsoft Excel and there is no point in creating software for analyzing such data based on Python or another programming language.

This means that there is a certain minimum required amount of data that must come from the agricultural field (agroeosystem) for use of Big Data technologies to be expedient. Moreover, even if the required amount of data is available, it is necessary to determine which of the many currently developed Big Data technologies for agriculture [4] are suitable for analysis.

In the article [5], the authors presented a table responsible for the distribution of technologies for analyzing big data, depending on the field of agriculture. For clarity, we present Table 1.

Table 1. Distribution of Big Data technologies, depending on the field of agriculture.

Agriculture area	Sources of Big Data	Big Data technologies
Weather and climate data	Geospatial data; Meteorological stations; Freely available historical information Other data obtained remotely.	Statistical analysis; Machine notation (K-means clustering algorithm, random/deep tree construction algorithm); GIS analysis; Distributed computing model.
Livestock breeding	Ground sensors; Thermal data; Optical sensors; Incoming feed sensors; Data on meat and dairy products.	Neural networks; Scalable vector machines; Decision trees.
Crops	Freely available historical information Satellite data; Ground sensors.	K-means clustering algorithm; Support Vector Machine; Fourier transform; Wavelet analysis;
Land resources;	Geospatial data; Freely available historical information Other data obtained remotely; Aerial survey data;	K-means clustering algorithm; algorithm for constructing random/deep trees; Image processing; Vegetation index ndvi.
Weeds	Freely available historical information Data received from drones; Aerial survey data; Sensor placed in the fields; Digital Web Libraries.	Neural networks; Logistic regression; Image processing;
Soil condition;	Freely available historical	K-means clustering algorithm;

	information Ground sensors; Information posted in the database of government agencies; Humidity sensors; Optical sensors;	Neural networks.
Biological resistance	Geospatial data; Freely available historical information Information posted in the database of government agencies;	Statistical modeling; Bayesian functions;
Food safety;	Geospatial data; Freely available historical information Other data obtained remotely; Survey data; Growth depth sensors;	Neural networks; Geospatial modeling; Statistical modeling; Image processing;
Farms;	Freely available historical information Optical sensors; Information posted in the database of government agencies; Meteorological stations; Social networks;	Big Data Benchmarking; Web services; Mobile applications;
Remote sensing;	Geospatial data; Satellite data; Geospatial data; Meteorological stations; Data received from drones; Digital Web Libraries.	Cloud computing with a distributed computing model; Geospatial modeling; Computer vision; Artificial intelligence;
Insurance and finance;	Optical sensors; Information posted in the database of government agencies; Digital web libraries; Information from private banks; Survey data;	Statistical modeling; Predictive analytics; Cloud technologies.

As you can see, for each area of agriculture, a source of information was selected for analysis using Big data technology and directly, the methods of analysis themselves. Of course, the areas of agriculture in Ireland, in any case, their categorization is different from the categorization of similar areas of the Tyumen region. However, the "general" idea can still be traced and "transferred" it to the realities of the Tyumen region. It would be advisable to check the adequacy of the selected methods of analysis based on the applied problem.

3 Results

Currently, in the Tyumen region, there is an acute issue of Big Data study in agriculture. Therefore, using the column "Sources of Big Data", one can assess the level of preparedness of a particular area of agriculture in the region for such studies. Also, based on the direct availability of "Sources of Big Data", it is possible to build an ontological graph of interaction with Big Data technologies, which in turn will lead to a methodology for assessing agricultural areas on the effectiveness of using Big Data technologies and

recommendations in which particular area of agriculture which sources to develop. For this, we present Table 2.

Table 2. Availability of Big Data sources regarding agriculture in the Tyumen region.

Agriculture area	Sources of Big Data	Source availability
Weather and climate data	Geospatial data	Absent
	Meteorological stations	Present
	Freely available historical information	Present
	Other data obtained remotely	Absent
Livestock breeding	Ground sensors	Present
	Thermal data	Absent
	Optical sensors	Absent
	Incoming feed sensors	Present
	Data on meat and dairy products	Present
Crops	Freely available historical information	Present
	Satellite data	Present
	Ground sensors	Present
Land resources	Geospatial data	Absent
	Freely available historical information	Present
	Other data obtained remotely	Present
	Aerial survey data	Absent
Weeds	Freely available historical information	Absent
	Data received from drones	Absent
	Aerial survey data;	Absent
	Sensor placed in the fields	Absent
	Digital Web Libraries	Absent
Soil condition	Freely available historical information;	Present
	Ground sensors;	Present
	Information posted in the database of government agencies;	Present
	Humidity sensors;	Present
	Optical sensors;	Absent
Biological resistance	Geospatial data;	Present
	Freely available historical information	Absent
	Information posted in the database of government agencies;	Present
Food safety;	Geospatial data;	Absent
	Freely available historical information	Absent
	Other data obtained remotely	Present
	Survey data	Absent
	Growth depth sensors	Absent
Farms	Freely available historical information	Present
	Optical sensors	Absent
	Information posted in the database of government agencies	Present
	Meteorological stations	Present
	Social networks	Present
Remote sensing	Geospatial data	Absent
	Satellite data	Absent
	Geospatial data	Absent

	Meteorological stations	Absent
	Data received from drones	Absent
	Digital Web Libraries	Absent
Insurance and finance	Optical sensors	Absent
	Information posted in the database of government agencies;	Present
	Digital web libraries	Absent
	Information from private banks	Absent
	Survey data	Absent

By adding the column "presence of a source", you can quantitatively visualize the filling of areas and display in Table 3.

Table 3. Quantitative indicators of the availability of Big Data sources in relation to agriculture in the Tyumen region.

Agriculture area	Source availability
Weather and climate data	2/4
Livestock breeding	2/5
Crops	3/3
Land resources	2/4
Weeds	0/5
Soil condition	4/5
Biological resistance	2/3
Food safety	1/5
Farms	4/5
Remote sensing	0/5
Insurance and finance	1/5

So it can be seen that the most prepared area of agriculture is Crops, Soil Condition, and Farms. Based on them, it will be possible to make initial ontological graphs for creating a methodology based on engineering ontology.

4 Discussion

Consider the process of creating an ontological graph based on the "Crops" area.

Table 4. Analysis of the "Crops" area.

Crops	Freely available historical information	K-means clustering algorithm Support Vector Machine Fourier transform Wavelet analysis
	Satellite data	
	Ground sensors	

To begin with, you need to understand to which subdomain "Crops" which Big Data technology to implement most effectively.

The K-means clustering algorithm technology consists in dividing a certain volume of observations (points) into an arbitrary number of clusters, limited by a certain standard [6]. The key feature is the presence of a centroid for each cluster and the separation of points in

accordance with similarity. In other words, distribution into clusters occurs, depending on the location of the central point (centroid), which is obtained by averaging the positions of all points, or vice versa. After distribution into clusters, through the calculation of the distance, two or more samples are compared with each other.

The use of this technology is correct if there are a number of points associated with a geographic location. That is, from the available "Sources of Big Data" only ground sensors in tandem with satellite or other methods of obtaining spatial information about ground sensors can be used. A source such as “Freely available historical information” is too broad concept; therefore, its rating of use for the analysis of “Crops” by Big Data technologies will be low.

Support vector machine technology [7] consists in learning by precedents that take into account the cases of linear separability, optima and gaps between classes. It is noteworthy that analytics using this technology most often occurs in tasks related to linear regression, which in parallel includes the search for minimization.

The correct application of this technology is possible if there is a time axis, in other words, to trace the change in information over time. Among the offered "Sources of Big Data", this can be "Freely available historical information", but on condition that it can be expressed in quantitative metrics.

The technology of the Fourier transform [8] consists in the comparison of one function into a complex variable, which describes the coefficients in the expansion of the original function into elementary components. In the area we are considering, such a transformation is used to process signals and represent them in the form of time series, with the possibility of displaying them in the form of a frequency spectrum.

Correct application is possible in presence of constantly incoming signals, in other words, in the presence of ground sensors.

Wavelet analysis technology [9] consists in the analysis of different types of frequency components of the data, most often along the "scale - time - level" chain. Most often they are used to obtain more accurate data after the Fourier transform, or to refine this transformation, since they give a more accurate binding of the quantitative features of the parameter, relative to time. Therefore, we can conclude that the wavelet analysis is only an auxiliary analysis and is used in special cases together with the Fourier transform.

Based on the above, an ontological graph of the "Crops" area can be built. For construction, we use the Protégé environment [10], the resulting graph is shown in Figure 1.

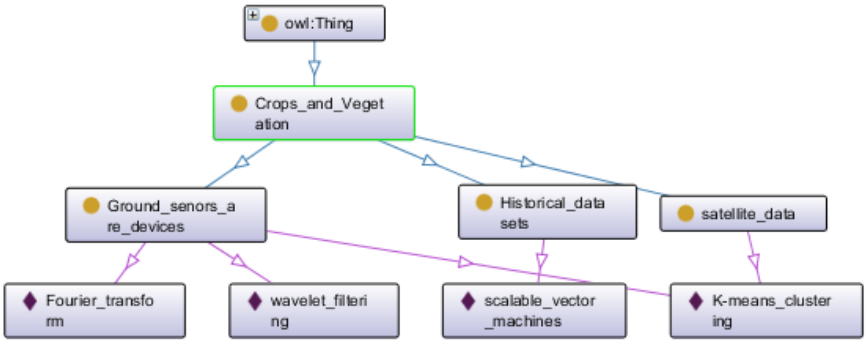

Fig. 1.. Ontological graph of the "Crops" area.

As you can see in the presented column, "Big Data Sources" were represented in the form of "Entities" of the Protégé environment, and "Big Data technologies" in the form of "Individuals", such a solution gives us, first of all, the opportunity to trace "paired" sources for big data analysis. From an applied perspective, this means the presence of the same analysis technologies, the same software for data from different sources, but in the same

agroecosystem. Thus, it will be easier for big data analysts to navigate when finding dependencies, obtaining additional information, making forecasts, and more.

In addition, in the ontological graph, you can see that such a source as "Freely available historical information" is analyzed by only one technique and does not intersect with other "Sources of Big Data". This, in turn, suggests that with development of Big Data in the region, there are two ways to go:

1. Stop this kind of source, without the need for its further development and analysis of big data.
2. Present data in an accessible form for the analysis of other Big Data technologies.

As you can understand, such ontological graphs can be built for any area of agriculture and matched for any agroecosystem. For clarity, let's expand the existing graph to include areas such as Soil Condition, Farms, and Weather and Climate Data, thereby creating a hypothetical agro-ecosystem that includes the listed areas. That is, the regions are chosen in such a way that, in fact, they are responsible for the analysis of the meta-region and cover the interests in Big Data of farms engaged in crop production.

Thus, we have obtained a search ontological graph. It is difficult to visualize it in full, but Protégé has search functions, by searching for the keyword "data" we see what is shown in Figure 2.

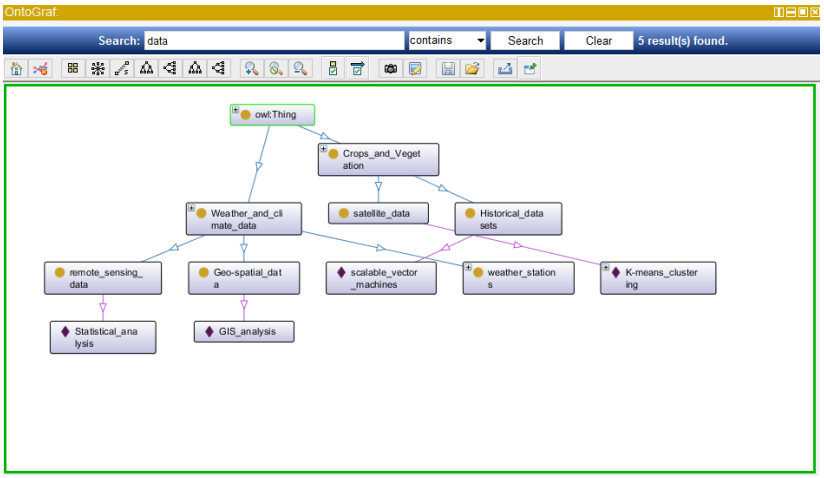

Fig. 2. The result of displaying a search ontological graph by the keyword "data".

On this search ontological graph, we see that all "Entities" related to the column "Sources of Big Data" and one way or another associated with data, as well as the "Individuals" and areas of agriculture attached to them, were displayed. In view of this, a kind of "coupling" is obtained, where, on the one hand, we can observe the fullness of Big Data technologies, and on the other hand, the field of agriculture. Such a visualized dependence and work with the resulting ontology can give a specialist the following information:

1. Understanding of the readiness of the agricultural sector for Big Data technologies.
2. Understanding of the fullness of the agricultural field with Big Data sources.
3. What sources of Big Data in what area of agriculture needs to be brought in to increase the efficiency of using Big Data technologies.
4. What Big Data technologies to use in what sources.
5. What "paired" Big Data sources can be used to capture the largest number of areas of agriculture for analysis.

6. What Big Data technologies can be used, depending on the type of agricultural enterprise.

As a result, it turns out that the very methodology for assessing the level of effectiveness of monitoring agroecosystems using Big Data technologies is, first of all, in the correctness of work with the resulting ontology. In order to translate the methodology directly into a quantitative metric, that is, to be able to "put down" each agroecosystem its own "score" from one to ten points, it is necessary to use such Protege functionality as "Data properties" and "Object properties".

"Data properties" in Protégé are literal properties and can be attached to "Entities", so it becomes possible to represent "Entities" as a list of property names and associated quantitative metrics. As a result, each "Entities" can be represented by a strictly defined numerical value, which can be changed only if additional information is added to the ontology.

"Object properties" in Protégé are a property pointer to "Entities". The pointer, in turn, has two key properties, namely the address allocated for a specific "Entities" and the connection with a specific type of "Entities".

Having filled in "Data properties" and "Object properties" for each "Entities", we will be able to quantify each area of agriculture, in other words, present them in the form of a line histogram, at Figure 3.

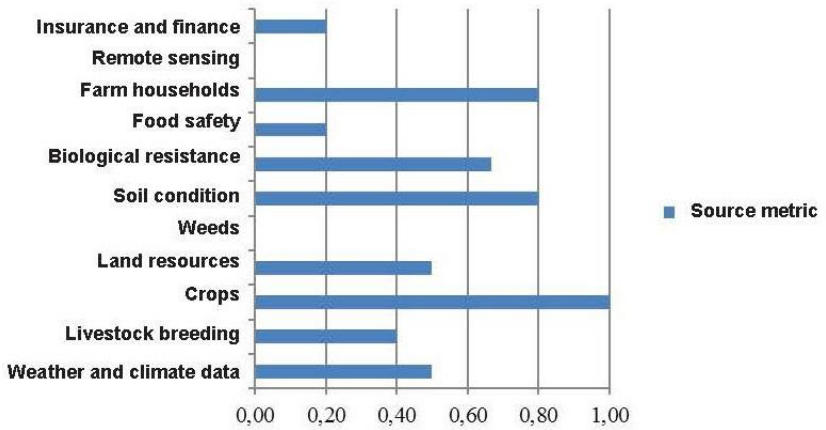

Fig. 3. Linear histogram for assessing the level of effectiveness of monitoring agroecosystems using Big Data technologies.

From this diagram, we can immediately observe the lagging areas of agriculture, namely the areas "Remote sensing" and "Weeds". This is due to the fact that such regions are not considered separately in the Tyumen region. However, it is still logical that in order to increase the efficiency of using Big Data technologies, a quantitative increase is necessary, first of all, “Sources of Big Data”. By itself, such a distribution was seen already at the stage of adding the column "presence of a source" and was displayed in Table 3. However, it is the addition of quantitative metrics for “Data properties” and “Object properties” to the ontology that will allow us to adjust the source metrics for the agroecosystem and see a more accurate picture of “the effectiveness of using Big Data technologies” at a more detailed level for a specialist. Of course, when making the right decisions for the development of Big Data for a particular agroecosystem, the cost of its equipment will be reduced (adding various kinds of data sources), by understanding which sources to put and accelerating decision-making for development of the region.

5 Conclusion

As a summary of this article, a number of the following conclusions can be made:

– The method of monitoring agroecosystems to assess the level of efficiency of using Big Data technologies allows visualizing the lagging and leading areas to form an effective development strategy.

– The agroecosystem manager shall guide not only the selection of Big Data technologies for analyzing information about the agroecosystem, but also focus on the available data sources about it.

– The methodology can be improved through use of engineering ontologies and Protégé software, which can internally evaluate technologies for agro-ecosystems and search for hidden knowledge, excluding the human factor.

To improve the methodology in the future, it is necessary to compare different agroecosystems for the effectiveness of using Big Data technologies and analyze the results.

References

1. P. Srivastava, R. Singh, R. Bhadouria, Pal D. Bahadur, P. Singh, S. Tripathi, ENPS, (2021)
2. S.Cao, G.Xie, L.Zhen, Ecological Economics, **69**, 7 (2010)
3. V.P. Yakushev, V.V. Yakushev, V.L. Badenko, D.A. Matveenko, Y.V. Chesnokov, Sel'skokhozyaistvennaya Biologiya, **55**, 3 (2020)
4. L.N. Hudson, T. Newbold, S.Contu, S.L.L. Hill, I. Lysenko, A. De Palma, H.R.P. Phillips, T.I. Alhusseini, F.E. Bedford, D.J. Bennett, H. Booth, V.J. Burton, Ecology and Evolution, **7**, 1 (2017)
5. D. Bose, Big data analytics in Agriculture (2020)
6. M. Ding, T. Wang, X. Wang, ACM Transactions on Knowledge Discovery from Data **16**, 2 (2022)
7. A. Singh, S. Saha, M. Hasanuzzaman, A. Jangra, Expert Systems with Applications, **186** (2021)
8. M.M. Alam, M.M.R. Howlader, Sensing and Bio-Sensing Research, **34** (2021)
9. F. Xu, S. Tan, Expert Systems with Applications, **186** (2021)
10. A. Petrov, A. Popov, M. Chekardovsky, A. Pushkarev, CEUR Workshop Proceedings, **2843** (2021)