UDC 004.89
DOI: 10.15587/1729-4061.2023.290129

DETERMINATION OF THE NUMBER OF CLUSTERS OF NORMALIZED VEGETATION INDICES USING THE K-MEANS ALGORITHM

Aigul Mimenbayeva
Corresponding author
Master of Sciences, Senior Lecturer
Department of Computational and Data Sciences
Astana IT University
Mangilik ave., 55/11, C1, Astana, Republic of Kazakhstan, 010017
E-mail: aigulka79_79@mail.ru

Samat Artykbayev
Student*

Raya Suleimenova
Candidate of Technical Sciences, Senior Lecturer*

Gulnar Abdygalikova
Candidate of Pedagogical Sciences, Senior Lecturer*

Akgul Naizagarayeva
Master of Engineering, Senior Lecturer*

Aisulu Ismailova
PhD, Associate Professor*
*Department of Information Systems
S.Seifullin Kazakh Agro Technical Research University
Zhenis ave., 62, Astana, Republic of Kazakhstan, 010011

The process of clustering of normalized vegetation indices in five regions with a total area of 2565 hectares of the North Kazakhstan region was studied. A methodological approach to organizing the clustering process is proposed using the vegetation indices NDVI, MSAVI, ReCI, NDWI and NDRE, taking into account individual characteristics in the three main phases of spring wheat development

As a result of the research, vegetation indices were grouped into 3 classes using the k-means clustering method. The first cluster contained vegetation indices whose maximum values occupied about 33.98 % of the total area of the study area. It was found that NDVImax located in the first cluster was positively correlated with soil-corrected vegetation indices MSAVI and crop moisture indicators NDMI (R2=0.92). The second cluster is characterized by minimum values of NDVImax coefficients at the germination, tillering and ripening phases (from 0.53 to 0.55). The lowest values of vegetation indices occupied 35.9 % in the germination phase, 37.9 % in the tillering phase, and 40.1 % of the field from the total area. The third cluster is characterized by average values of vegetation indices in all three phases. A correlation matrix was also constructed to assess the closeness of the relationship between actual yield and NDVI vegetation indices. The maximum coefficient was obtained at the germination phase, R=0.94 with a minimum significance coefficient p=0.018.

The approach used in this study can be useful in the analysis of satellite data, as it can improve the sensitivity of the constellation procedure. From a practical point of view, the results obtained make it possible to assess the condition of agricultural crops in the early stages of the growing season, which makes it possible to improve their productivity based on the results of cluster analysis

Keywords: NVDI, vegetation index, cluster analysis, k-means algorithm, remote sensing data

Received date 14.08.2023
Accepted date 23.10.2023
Published date 30.10.2023

How to Cite: *Mimenbayeva, A., Artykbayev, S., Suleimenova, R., Abdygalikova, G., Naizagarayeva, A., Ismailova, A. (2023). Determination of the number of clusters of normalized vegetation indices using the k-means algorithm. Eastern-European Journal of Enterprise Technologies, 5 (2 (125)), 42–55. doi: https://doi.org/10.15587/1729-4061.2023.290129*

1. Introduction

The use of information technology to monitor crop yields provides enhanced opportunities for solving modern precision farming problems. Crop yields can be effectively planned using digital agriculture technology to monitor field productivity based on satellite imagery. The experience of the leading countries of the world shows the high efficiency of the use of information technologies in digitalization and automation of production, which make it possible to optimally use satellite data in agriculture [1].

One of the promising areas in the field of precision agriculture is the use of geographic information systems, which make it possible to analyze satellite data and create vegetation maps that display the level of vegetation in different areas. Such maps allow to track the condition of crops at various levels, from regional to national levels. This makes it possible for agricultural organizations, government agencies and scientific researchers to monitor the condition and yield of crops, identify problem areas and take appropriate measures. In addition, geographic information systems allow the calculation of various vegetation indices before harvest, which can be used to assess plant health, factors affecting growth and fertility, and to predict crop yields. There are about 160 variants of vegetation indices [2]. Of these, the most common and optimally suitable index for tracking the dynamics of vegetation development is the normalized vegetation index NDVI. Using this index, the active photosynthetic biomass of a plant can be measured. The following formula is used to calculate this index:

$$\mathrm{NDVI}=\frac{\mathrm{NIR}-\mathrm{RED}}{\mathrm{NIR}+\mathrm{RED}}, \qquad (1)$$

where NIR – near infrared light, RED – red light. The values of these coefficients vary from –1 to +1. In Fig. 1 shows a scale of NDVI value ranges and the state of the plant [3].

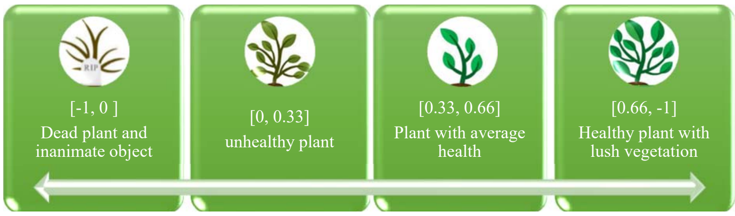

RIP
[-1, 0]
Dead plant and inanimate object
[0, 0.33]
unhealthy plant
[0.33, 0.66]
Plant with average health
[0.66, -1]
Healthy plant with lush vegetation

Fig. 1. Scale of ranges of normalized vegetation NDVI indices

Calculation of normalized NDVI coefficients in geosystems is performed using a raster calculator using pre-loaded data from the field under study. Analysis and processing of vegetation indices over several years provides the necessary information for the formation of a protection system and fertilizer standards, yield forecast in combination with other factors [3].

Based on the above, it can be emphasized that the search and optimization of methods for analyzing the state of agricultural crops using new information technologies is a scientifically relevant area that allows to achieve qualitatively new results in the field of precision agriculture in a number of Asian countries, including Kazakhstan.

In this regard, it is necessary to note one of the most popular and promising areas in the IT industry – data analytics and the use of machine learning algorithms. By analyzing remote sensing data, machine learning algorithms can identify patterns and make predictions that can help farmers optimize their operations and improve yields. The use of machine learning algorithms makes it possible to optimize the process of monitoring and analyzing weather, soil and crop data to obtain accurate forecasts and recommendations.

It is also necessary to emphasize that at present, general methods and methods for assessing the productivity of agricultural crops using remote sensing data have not yet been fully developed [4].

Thus, the analysis and processing of the states of agricultural crops using machine learning methods is a current scientific and applied direction that makes it possible to achieve qualitatively new results in the field of precision agriculture in a number of Asian countries, including Kazakhstan.

2. Literature review and problem statement

The use of effective methods and technologies in agriculture, such as data analytics and the use of machine learning techniques, opens up new opportunities in precision agriculture. Increasing yields and reducing crop losses through the use of effective machine learning methods helps determine the optimal time and place of sowing, watering, fertilizing and harvesting, which in turn allows agricultural producers to obtain more products and increase their income.

Various vegetation indices are used to analyze and process remote sensing data obtained from satellites.

[5] studied various vegetation indices, such as the Atmospheric Resistant Vegetation Index (ARVI), Soil Adjusted Vegetation Index (SAVI), Green Difference Vegetation Index (GDVI), Green Normalized Difference Vegetation Index (GNDVI), Enhanced Vegetation Index (EVI). As a result of studies conducted in different agrometeorological conditions, the authors came to the conclusion that before practically applying the vegetation index, it is necessary to comprehensively analyze the advantages and disadvantages in a particular environment.

The authors of [6] conducted a cluster analysis of the state of plants according to the time of their development according to different vegetation indices. Having studied the above indices in the conditions of the Ladoga lowland plain, the authors came to the conclusion that the most informative vegetation index for assessing the condition of vegetating meadow grasses is NDVI. Also, in work [7] various vegetation indices are considered in detail; the authors emphasized that when choosing vegetation indices for satellite data studies, the seasonal ecological and climatic indicators of the study area should be taken into account.

Taking into account the results of works [5–7], it can be argued that the most informative vegetation index is the NDVI coefficient, which most accurately reflects the state of vegetation in different phenological phases.

Of interest are studies using remote sensing data of the earth in the form of satellite images, in which cluster analysis methods are applied, including the k-means method.

A study conducted by the authors of [8] provided a detailed review of big data clustering methods, including the K-Means algorithm, FCM algorithm, Expectation maximization algorithm, and BIRCH. The authors noted that the most rational approach may be to combine a number of these methods, which makes it possible to take advantage of the benefits that a particular method offers. Despite the theoretical review, the work is not supported by experimental data, which in turn would make it possible to draw a logical conclusion based on actual data.

In work [9], cluster analysis of NDVI vegetation index profiles over several years was used to assess the condition of arable fields in the East Kazakhstan region. The authors have developed their own software module in Python, which calculates and groups NDVI coefficients into several fields using the k-means method. The research results were visually presented and analyzed by specific clusters in the years studied. In accordance with the results obtained, basic descriptive statistics were calculated for each cluster. The advantage of this study is the individuality of the solution, but the disadvantage is the calculation of descriptive NDVI statistics after the clustering stage, which reduces the optimization of the use of the given software module.

The work [10] presents a review of classical image clustering methods from the point of view of statistical pattern recognition. A taxonomy of clustering methods is presented and cross-cutting themes and recent advances in the field are identified. The disadvantage of work [10] is that it does not take into account the features of image clustering, since such a task cannot be classified as an uncontrolled classification of patterns into groups.

In [11], a semantic image clustering algorithm was proposed to obtain a dataset with grouped images. The disadvantage of this method is the loss of information, since any other image belonging to the cluster can be used as a representative image.

The authors of [12], using the k-means algorithm, experimentally demonstrated the determination of the number of clusters in images of space optical-electronic observation systems. The disadvantage of the work is that the assessment of the state of crops using the clustering procedure is based only on the applied aspect of the k-means method; the work does not consider the features of the space optical-electronic observation system, unlike other objects.

Research conducted by the authors of [13] showed the effectiveness of the hybrid k-means method in the field of precision agriculture for assessing the productivity of agricultural crops. These studies provide a qualitative analysis of modern clustering methods and note that to identify outliers and define clusters, it is preferable to use cross-clustering methods that combine the best aspects of several clustering algorithms. Despite this, the study [13] has some shortcomings related to the choice of dataset for clustering. In this work, for the practical application of clustering, 2 completely different sets of data in different areas were selected. The first data set describes statistics on population, urban areas, and arrests per 100,000 residents for various types of crimes in USA in 1973. The second data set consists of wheat production statistics in various districts of Karnataka from 2009 to 2016. Clustering does not take into account the individual characteristics of the data sets under consideration, which does not justify the result obtained using hybrid clustering methods.

The authors of [14] applied the clustering procedure using the k-means method to general data on the yield of various agricultural crops from 1991-2002. on the territory of India. The disadvantage of the work is that clustering was carried out not only for agricultural crops, but also different types of plants were considered, such as vegetable crops, flowers and aromatic crops, plantation crops. It is worth noting that the yield depends on many factors: the types of crops, individual development phases, etc. In this regard, it is not advisable to compare the yield of certain crops over the years, because in different years, the predecessors can be different crops, and this directly affects the quality of the soil and, ultimately, the yield.

Having reviewed existing clustering methods, it can be noted that many studies in this area emphasize the effectiveness and ease of application of the k-means method from unsupervised learning methods. However, in these studies, clustering was carried out under general conditions, mostly the clustering results were not confirmed with actual data on crop productivity. It is worth noting that each type of individual crop has its own characteristics depending on the phenological phase of development. Therefore, when using machine learning methods, the effectiveness of the result obtained lies not only in the advantage of the selected algorithm, but in the choice of key parameters for cluster analysis.

In this regard, it is possible to note the work of the authors [15], who applied k-means machine learning methods and Ward's method, choosing three vegetation indices NDVI, OSAVI and SAVI and their descriptive statistics at different stages of the growing season as key clustering parameters. The authors collected Marche Region satellite data from seven durum wheat varieties to determine the dynamics of yield components using cluster analysis. Let's also examine in detail the problem of clustering at the stages of germination, tillering and flowering of durum wheat varieties and determined the areas of clusters with homogeneous values of vegetation indices. However, this work does not take into account the individual characteristics of vegetation indices regarding their use in the development phases of winter wheat. It is worth emphasizing that the soil-corrected vegetation index SAVI and the optimized soil vegetation index OSAVI – created to minimize soil brightness – are used preferably for the analysis of crops in the early stages, as well as for monitoring arid areas with low vegetation density and in open areas of land [5].

Therefore, using the vegetation indices NDVI, OSAVI and SAVI as the main factors to assess yields in all three phases of crop development may produce some incorrect results when applying the algorithm.

To summarize, it should be noted that the effectiveness of applying machine learning methods, including the k-means clustering method to remote data, requires a detailed approach, and in the case of choosing vegetation indices as key factors, it is necessary to take into account the individual characteristics of each vegetation index and development phase agricultural crops. It should also be noted that the literature review highlights the effectiveness and simplicity of the clustering algorithm for use in precision agriculture, including the analysis of NDVI coefficients for increasing crop yields.

3. The aim and objectives of the study

The aim of this study is to determine the number of clusters of spring wheat vegetation indices based on the k-means method in the conditions of the North Kazakhstan region during the growing season of 2022.

To achieve the aim, the following objectives were set:

– prepare the appropriate data array from 25.05.2022 to 25.07.2022 in the form of vegetation indices NDVI, OSAVI, NDMI, ReCI and NDRE on the territory of the North Kazakhstan region in the context of five districts;

– perform clustering by vegetation indices NDVI, OSAVI, NDMI, ReCI and NDRE, taking into account the individual characteristics of each index using the k-means method in the phases of tillering, flowering and ripening;

– evaluate the results obtained using the correlation matrix and the significance level of the selected parameters.

4. Materials and methods of research

To monitor and analyze the productivity of spring wheat in the North Kazakhstan region, the vegetation indices NDVI, OSAVI, NDMI, ReCI and NDRE, obtained by interpreting high-resolution multispectral satellite images of Landsat 8 for the growing season of 2022, were selected as research objects. To process satellite images, the EOS Data analytics geographic information system with the Time Series Analysis, Spatial Analysis module was used.

The research hypothesis is that the use of vegetation indices taking into account individual characteristics in the phenological phases of crop development will help improve the efficiency of clustering results for monitoring and assessing the productivity of agricultural crops.

In order to structure the obtained remote data, a cluster analysis multivariate statistical procedure was used. Cluster analysis was performed in the Cluster Analysis module in the STATISTICA software package. When constructing dendrograms, the standardization procedure, the hierarchical tree method and the k-means method were used.

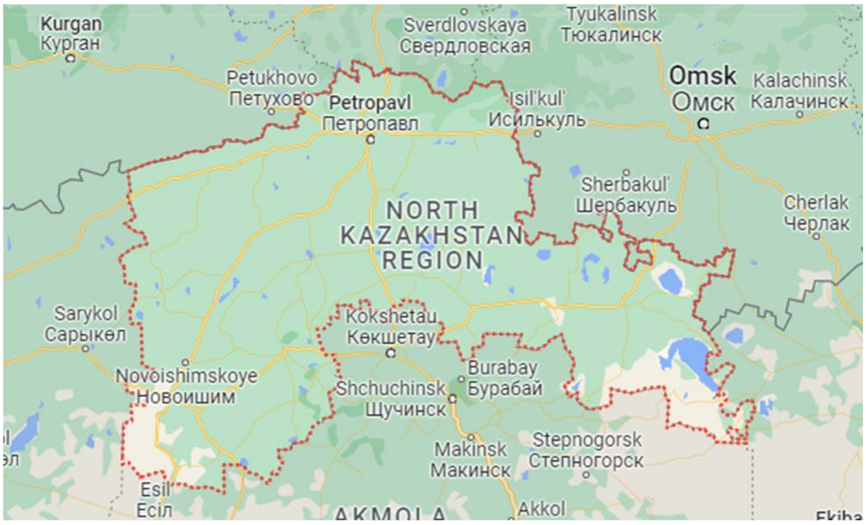


Fig. 2. Territory of the North Kazakhstan region

K-means method. The k-means method is one of the methods of statistical analysis, the purpose of which is to partition given observations from the vector space R^n into a finite number of clusters. The distance between clusters is calculated by the Euclidean formula:

$$p(x,y)=\|x-y\|=\sqrt{\sum_{p=1}^{n}(x_p-y_p)^2}, \tag{2}$$

where x, $y \in R^n$.

Let $(x^{(1)},x^{(2)},x^{(3)},...,x^{(m)}) \in R^n$ be a series of given observations. The k-means method splits a given m observations into $(x^{(1)},x^{(2)},x^{(3)},...,x^{(m)}) \in R^n$ clusters, minimizing the sum of squared distances from each cluster point to its cluster center $\mu_i \in R^n$:

$$\min\left[\sum\nolimits_{i=1}^{k}\sum\nolimits_{x^{(j)}}x^{(j)}-\mu_i^{\ 2}\right], \tag{3}$$

where $i=1,...,k$; $j=1,..., m$.

The following diagram illustrates the algorithm for implementing the k-means method (Fig. 3).

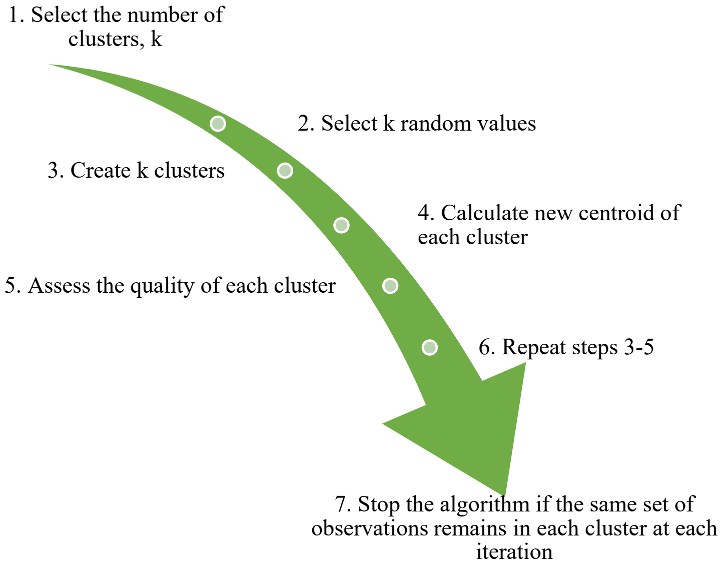

1. Select the number of clusters, k
2. Select k random values
3. Create k clusters
4. Calculate new centroid of each cluster
5. Assess the quality of each cluster
6. Repeat steps 3-5
7. Stop the algorithm if the same set of observations remains in each cluster at each iteration

Fig. 3. Algorithm of the cluster analysis method – k-means

Hierarchical Clustering Algorithm. This procedure allows to identify a set of objects characterized by a certain degree of connectivity. At the initial stage, each object is considered as a separate cluster, in the process of clustering in a hierarchical way, relatively homogeneous groups of objects are combined into a separate cluster, the procedure continues until only one remains. The distance between clusters U and V can be determined using one of the formulas given in Table 1.

There are two main types of hierarchical clustering approaches: agglomerative and divisive clustering.

In agglomerative methods, new clusters are created by sequentially combining individual objects or their groups into larger subsets, thus creating a hierarchical tree from leaves to trunk.

Table 1

Formulas for determining the distance between clusters

1	Single linkage method	$R_{\min}(U,V)=\min\{d(u,v):u\in U,v\in V\}$
2	Complete linkage method	$R_{\max}(U,V)=\max\{d(u,v):u\in U,v\in V\}$
3	Unweighted pair group method with arithmetic mean	$R_{avg}(U,V)=\frac{1}{\lvert U\rvert\cdot\lvert V\rvert}\sum_{u\in U}\sum_{v\in V}d(u,v)$
4	Unweighted pair group method with centroid arithmetic average	$R_c(U,V)=p^2\left(\sum_{u\in U}\frac{u}{\lvert U\rvert}\cdot,\sum_{u\in U}\frac{v}{\lvert V\rvert}\right)$
5	Ward's method	$R_{ward}(U,V)=\frac{\lvert U\rvert\cdot\lvert V\rvert}{\lvert U\rvert+\lvert V\rvert}p^2\left(\sum_{u\in U}\frac{u}{\lvert U\rvert}\cdot,\sum_{u\in U}\frac{v}{\lvert V\rvert}\right)$

Note: d(u,v) – distance between objects u, v belonging to clusters U and V, respectively

Divisional methods use the reverse process in which new clusters are created by splitting larger clusters into smaller ones, thus creating a tree from the root to the leaves. The resulting tree structure of objects of a given array, constructed from a matrix of proximity measures, is called a dendogram. Fig. 4 demonstrates the main steps of the groups of clustering types described above [17].

Correlation analysis. Correlation allows to determine how much the considered system parameters are interrelated with each other. If there is a correlation between them, then it is possible to predict the values of one quantity based on the values of the other. The correlation may be positive, negative, or absent. A positive correlation means that as the values of one quantity increase, the values of the other quantity also increase. A negative correlation, on the contrary, indicates that as the values of one quantity increase, the values of another quantity decrease. In the absence of correlation, changing the values of one quantity does not in any way affect the values of another quantity.

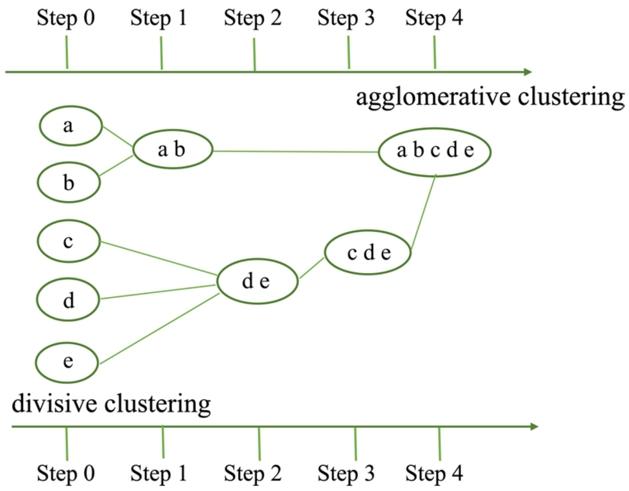


Fig. 4. Dendogram of agglomerative and separative clustering

One way to quantify this relationship is to use the Pearson correlation coefficient (R), which is a measure of the linear relationship between two variables. The value of the correlation coefficient can range from –1 to 1. A value close to 1 indicates a strong positive correlation, one close to –1 indicates a strong negative correlation, and a value close to 0 indicates no correlation.

In correlation analysis, P-value is used as a measure of the statistical significance of the results and to assess the reliability of statistical conclusions. It shows the probability of obtaining the observed data or more extreme results, given that the null hypothesis (usually that there is no effect or no difference between groups) is true.

If the P-value is very small (usually less than a specified significance level, such as 0.05), it indicates that the observed data are unlikely to be true given the null hypothesis. In this case, it is possible to reject the null hypothesis in favor of the alternative hypothesis, which suggests there is an effect or difference between groups.

Vegetation indices. The vegetation index allows to assess the health of vegetation, the degree of its development, and physiological state.

Normalized vegetation index NDVI. This index allows to track changes in vegetation cover, identify zones of soil degradation and assess the state of ecosystems. NDVI can also be used to analyze agricultural land to determine its potential for growth and yield. For example, low NDVI values may indicate soil problems or pests in the fields. Using NDVI in combination with other data, such as weather or soil data, can help agronomists and scientists make decisions about vegetation management and agricultural production. NDVI is widely used in various fields including ecology, forestry, agriculture, hydrology and geology. Its use continues to evolve with the development of new technologies for surveying and processing satellite data.

Modified Soil Vegetation Index (MSAVI). The MSAVI vegetation index calculation method is based on the spectral features of two image bands: red and near-infrared. It uses information about alkaline soil cover, which contains less chlorophyll, which helps reduce the influence of soil noise and improves the accuracy of land cover data. The basic principle of MSAVI is to account for variation in the ratio of plant chlorophyll to other components of the plant canopy. This allows to more accurately determine the presence and intensity of vegetation, even with a low level of vegetation cover. Thus, the MSAVI index is a useful tool for monitoring and studying vegetation, especially under conditions where NDVI cannot provide sufficiently accurate data, for example, in the case of low chlorophyll content or insufficient vegetation. It can also be used to monitor vegetation in the early stages of development, when the vegetation cover is not yet sufficiently developed, i.e. during the period of seedling formation.

Normalized Differential Red Edge Index (NDRE). The NDRE vegetation index is used to determine the maturity of crop plants, as well as to assess the overall health of vegetation and determine the level of plant stress. NDRE can also be useful in studying the relationship between crop yield levels and land cover. This vegetation index is often used in agronomy and environmental studies at the stage of crop maturation. An NDRE value closer to 1.0 indicates higher foliar nitrogen content, which may indicate the need to reduce fertilizer applications. An NDRE value closer to -1.0 indicates lower foliar nitrogen content, which may indicate the need for increased fertilization. In this way, NDRE can help the user make more informed decisions regarding fertilizers and optimize their use in the field.

Vegetation photosynthetic activity index (ReCI). This index is an indicator that is used to assess the amount of chlorophyll in plant leaves. It can also be used to assess the active growth phase of plants. ReCI is based on the spectral characteristics of land cover, which are reflected in spectral data obtained from Earth remote sensing. This index uses information about the absorption and reflection of light in different wavelength ranges to determine the amount of chlorophyll in plant leaves. The higher the ReCI value, the greener and more actively growing the plants. Low ReCI values may indicate a lack of chlorophyll or the presence of stressful conditions such as drought or disease. ReCI is a useful tool for monitoring vegetation conditions and assessing vegetation health. It can be used in various fields such as agriculture, ecology and forestry to assess crop yields, determine the degree of greenness and detect plant diseases.

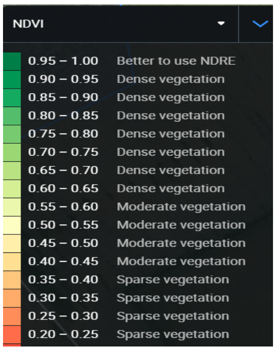
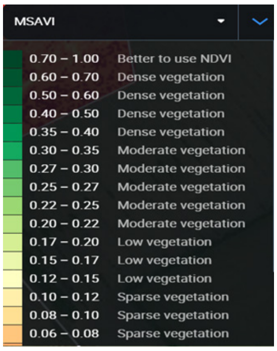
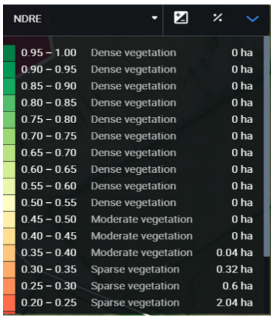
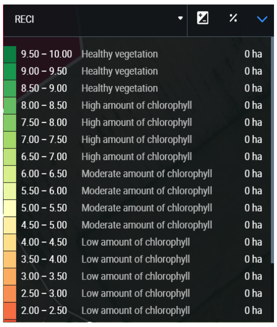
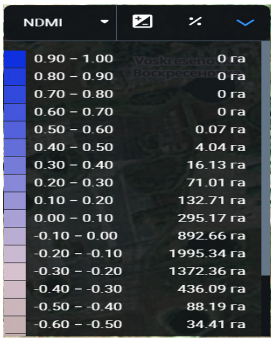
The Normalized Difference Moisture Index (NDMI) is an index that is used to determine the moisture content of plants based on the difference in spectral ranges in the near-infrared (NIR) and short-wave infrared (SWIR) bands. This index is a reliable indicator of moisture deficiency in agricultural crops. The more negative the NDMI value, the more pronounced the moisture deficit in plants. High NDMI values indicate good moisture levels in plants.

NDMI is a useful tool for assessing moisture deficiency and identifying low moisture zones in crops. It can be used to monitor crops and make watering decisions to prevent potential plant damage due to moisture deficiency.

Table 2 shows the mathematical description and scale of values of the vegetation indices NDVI, MSAVI, NDRE, ReCI and NDWI.

Table 2

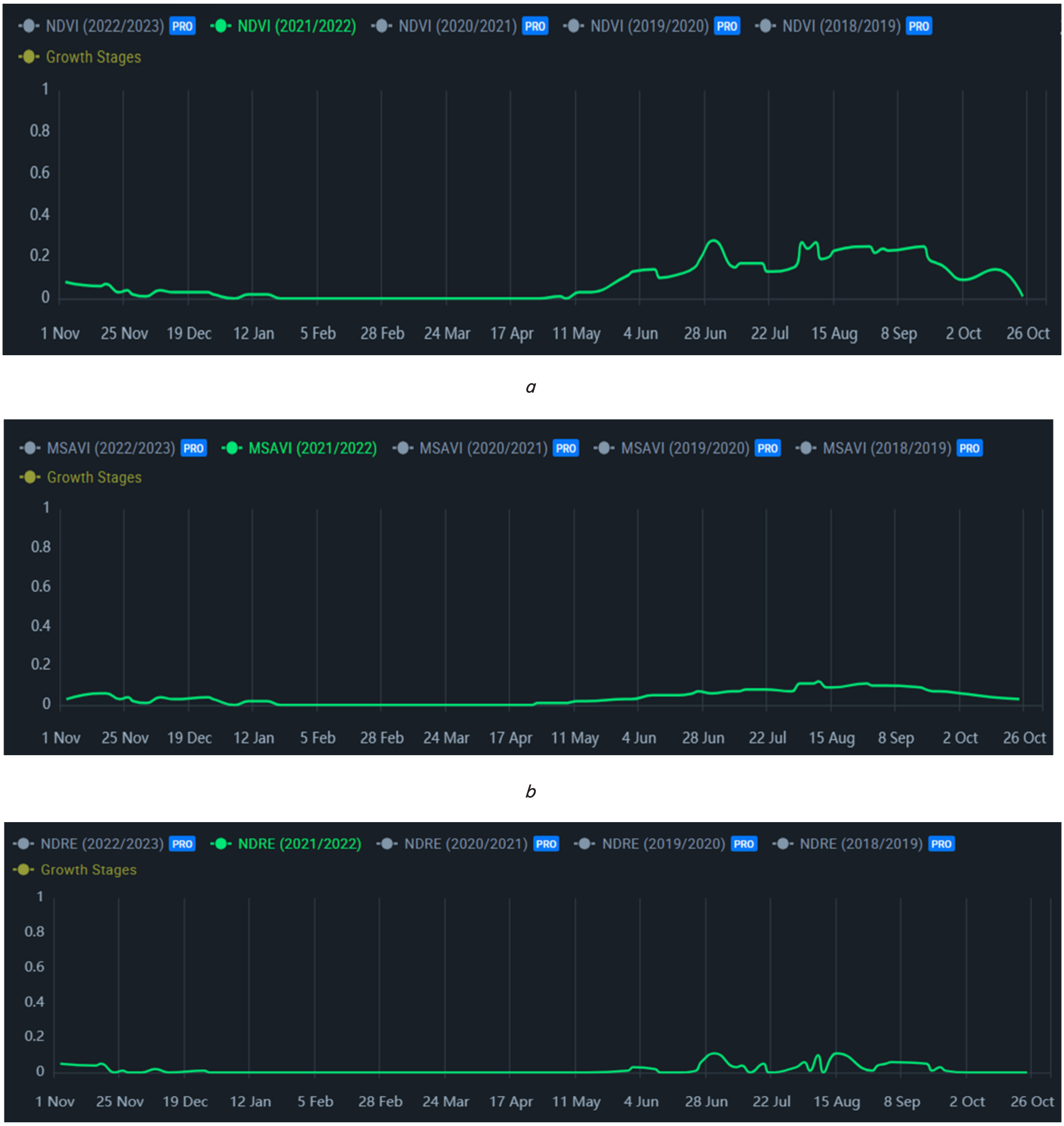
Formula and brief description of vegetation indices NDVI, MSAVI, NDRE , ReCI, NDMI

VI	Mathematical formula	Value scale	Source
NDVI	$$\text{NDVI}=\frac{(P_{\text{NIR}}-P_{\text{RED}})}{(P_{\text{NIR}}+P_{\text{RED}})},$$ where PNIR – NIR band, PRED – RED band	 NDVI 0.95 – 1.00 Better to use NDRE 0.90 – 0.95 Dense vegetation 0.85 – 0.90 Dense vegetation 0.80 – 0.85 Dense vegetation 0.75 – 0.80 Dense vegetation 0.70 – 0.75 Dense vegetation 0.65 – 0.70 Dense vegetation 0.60 – 0.65 Dense vegetation 0.55 – 0.60 Moderate vegetation 0.50 – 0.55 Moderate vegetation 0.45 – 0.50 Moderate vegetation 0.40 – 0.45 Moderate vegetation 0.35 – 0.40 Sparse vegetation 0.30 – 0.35 Sparse vegetation 0.25 – 0.30 Sparse vegetation 0.20 – 0.25 Sparse vegetation	[2]
MSAVI	$$\text{MSAVI}=\frac{(P_{\text{NIR}}-P_{\text{RED}})}{(P_{\text{NIR}}+P_{\text{RED}}+L)(1+L)},$$ where PNIR – NIR band, PRED – RED band, L=1–2·S·NDVI·WDVI	 MSAVI 0.70 – 1.00 Better to use NDVI 0.60 – 0.70 Dense vegetation 0.50 – 0.60 Dense vegetation 0.40 – 0.50 Dense vegetation 0.35 – 0.40 Dense vegetation 0.30 – 0.35 Moderate vegetation 0.27 – 0.30 Moderate vegetation 0.25 – 0.27 Moderate vegetation 0.22 – 0.25 Moderate vegetation 0.20 – 0.22 Moderate vegetation 0.17 – 0.20 Low vegetation 0.15 – 0.17 Low vegetation 0.12 – 0.15 Low vegetation 0.10 – 0.12 Sparse vegetation 0.08 – 0.10 Sparse vegetation 0.06 – 0.08 Sparse vegetation	[3]
NDRE	$$\text{MSAVI}=\frac{(P_{\text{NIR}}-P_{\text{RE}})}{(P_{\text{NIR}}+P_{\text{RE}})},$$ where PNIR – NIR band, PRE – RED EDGE band	 NDRE 0.95 – 1.00 Dense vegetation 0 ha 0.90 – 0.95 Dense vegetation 0 ha 0.85 – 0.90 Dense vegetation 0 ha 0.80 – 0.85 Dense vegetation 0 ha 0.75 – 0.80 Dense vegetation 0 ha 0.70 – 0.75 Dense vegetation 0 ha 0.65 – 0.70 Dense vegetation 0 ha 0.60 – 0.65 Dense vegetation 0 ha 0.55 – 0.60 Dense vegetation 0 ha 0.50 – 0.55 Dense vegetation 0 ha 0.45 – 0.50 Moderate vegetation 0 ha 0.40 – 0.45 Moderate vegetation 0 ha 0.35 – 0.40 Moderate vegetation 0.04 ha 0.30 – 0.35 Sparse vegetation 0.32 ha 0.25 – 0.30 Sparse vegetation 0.6 ha 0.20 – 0.25 Sparse vegetation 2.04 ha	[2, 3]
ReCI	$$\text{ReCl}=\frac{P_{\text{NIR}}}{P_{\text{RE}}}-1$$	 RECI 9.50 – 10.00 Healthy vegetation 0 ha 9.00 – 9.50 Healthy vegetation 0 ha 8.50 – 9.00 Healthy vegetation 0 ha 8.00 – 8.50 High amount of chlorophyll 0 ha 7.50 – 8.00 High amount of chlorophyll 0 ha 7.00 – 7.50 High amount of chlorophyll 0 ha 6.50 – 7.00 High amount of chlorophyll 0 ha 6.00 – 6.50 Moderate amount of chlorophyll 0 ha 5.50 – 6.00 Moderate amount of chlorophyll 0 ha 5.00 – 5.50 Moderate amount of chlorophyll 0 ha 4.50 – 5.00 Moderate amount of chlorophyll 0 ha 4.00 – 4.50 Low amount of chlorophyll 0 ha 3.50 – 4.00 Low amount of chlorophyll 0 ha 3.00 – 3.50 Low amount of chlorophyll 0 ha 2.50 – 3.00 Low amount of chlorophyll 0 ha 2.00 – 2.50 Low amount of chlorophyll 0 ha	[2, 3]
NDMI	$$\text{NDMI}=\frac{(P_{\text{NIR}}-P_{\text{SWIR1}})}{(P_{\text{NIR}}+P_{\text{SWIR1}})}$$	 NDMI 0.90 – 1.00 0 га 0.80 – 0.90 0 га 0.70 – 0.80 0 га 0.60 – 0.70 0 га 0.50 – 0.60 0.07 га 0.40 – 0.50 4.04 га 0.30 – 0.40 16.13 га 0.20 – 0.30 71.01 га 0.10 – 0.20 132.71 га 0.00 – 0.10 295.17 га -0.10 – 0.00 892.66 га -0.20 – -0.10 1995.34 га -0.30 – -0.20 1372.36 га -0.40 – -0.30 436.09 га -0.50 – -0.40 88.19 га -0.60 – -0.50 34.41 га	[1]

5. Results of studies on the use of vegetation indices in the development phases of spring wheat for monitoring and assessing crop productivity

5. 1. Preparation of data sets of vegetation indices NDVI, OSAVI, NDMI, ReCI and NDRE and their statistical indicators

As part of this study, a total of 52 observations, 156 images in three phases of the spring wheat growing season were downloaded and digitized using Landsat 8 and Sentitel 2 atellite channels using a raster calculator in the EOS Land Viewer system. For cluster analysis of remote sensing data using the k-means method, digitized satellite images from May 28, 2022 were used. until July 27, 2022 in the form of vegetation indices NDVI, NDVI, OSAVI, NDWI, ReCI and NDRE and their statistical indicators of spring wheat in the North Kazakhstan region. The normalized vegetation index NDVI was chosen as the main indicator for the research. The time series of spring wheat NDVI, MSAVI, NDRE coefficients for the study period is illustrated in Fig. 5.



a

b

c

Fig. 5. Analysis of time series of vegetation coefficients of spring wheat in the North Kazakhstan region: *a* – NDVI; *b* – MSAVI; *c* – NDRE

In the cluster analysis, the following coefficients were included in the basis of groupings throughout the entire array, in parallel with the values of the NDVI vegetation indices: standard deviation (Std) – standard deviation, variance – dispersion, Q1 – lower quartile, Q3 – upper quartile, P10 – low percentile, P90 – higher percentile, median – median, scene_id – the scene's identifier, view_id – view's identifier.

5. 2. Cluster data analysis using the k-means method

At the first stage of the clustering procedure, in order to bring all values of vegetation indices to a single range of values, a standardization procedure was carried out. Next, using the hierarchical tree method, a dendogram was constructed (Fig. 6). Based on the resulting dendogram, it was determined that the data can be divided into three clusters: C1–C6; C7–C37; C38–C53. Therefore, the main hypothesis further was that there were three clusters of data. Cluster analysis was carried out using the k-means method and stable 3 clusters were obtained in one iteration. In order to assess the stability of these clusters, the method of variance analysis was used. As a result, it was revealed that for all considered variables the significance values $p=0.00<0.05$, i.e. the clusters under consideration are well separable and each variable that we entered into the analysis is important for constructing these clusters. Fig. 7 illustrates the clusters for NDVI coefficients and statistical characteristics.

Fig. 6 demonstrates the result of cluster analysis of vegetation indices in the North Kazakhstan region.

Based on the above operations, 3 observational clusters were obtained based on satellite data obtained during the growing season (Table 3).

Using the results of cluster analysis, at each harvest stage, cluster areas were identified in which the values of vegetation indices were homogeneous. This allows for a more detailed study and comparison of the characteristics of different clusters at each stage of the agricultural cycle. This approach to satellite data analysis allows for more accurate determination of changes in vegetation and agricultural activity at different stages of the harvest. This can be useful for planning and optimizing agricultural processes and solving various problems related to yield and vegetation characteristics in certain regions:

1. Cluster analysis in the germination phase of spring wheat

A map of clusters compiled on the basis of the NDVI, MSAVI and NDMI indices taken from seedlings is presented in Fig. 4. 3 clusters were identified, the highest values are represented by cluster 1. The median value, standard deviation and analysis of variance of the data from each cluster are presented in Table 3. Average NDVI values ranged from 0.58 to 0.78 for cluster I, from 0.15 to 0.28 for cluster II, from 0.4 to 0.66 for cluster III.

In the germination phase of the first cluster, the highest $NDVI_{max}=0.78$ is observed in the Mamlyut and Kyzylzhar regions, which is confirmed by the high coefficient of the modified adjusted soil index MSAVI in these regions. Accordingly, in these areas in the germination phase, a high degree of soil moisture is observed, which is explained by the high values of the normalized difference moisture index NDMI (Fig. 8).

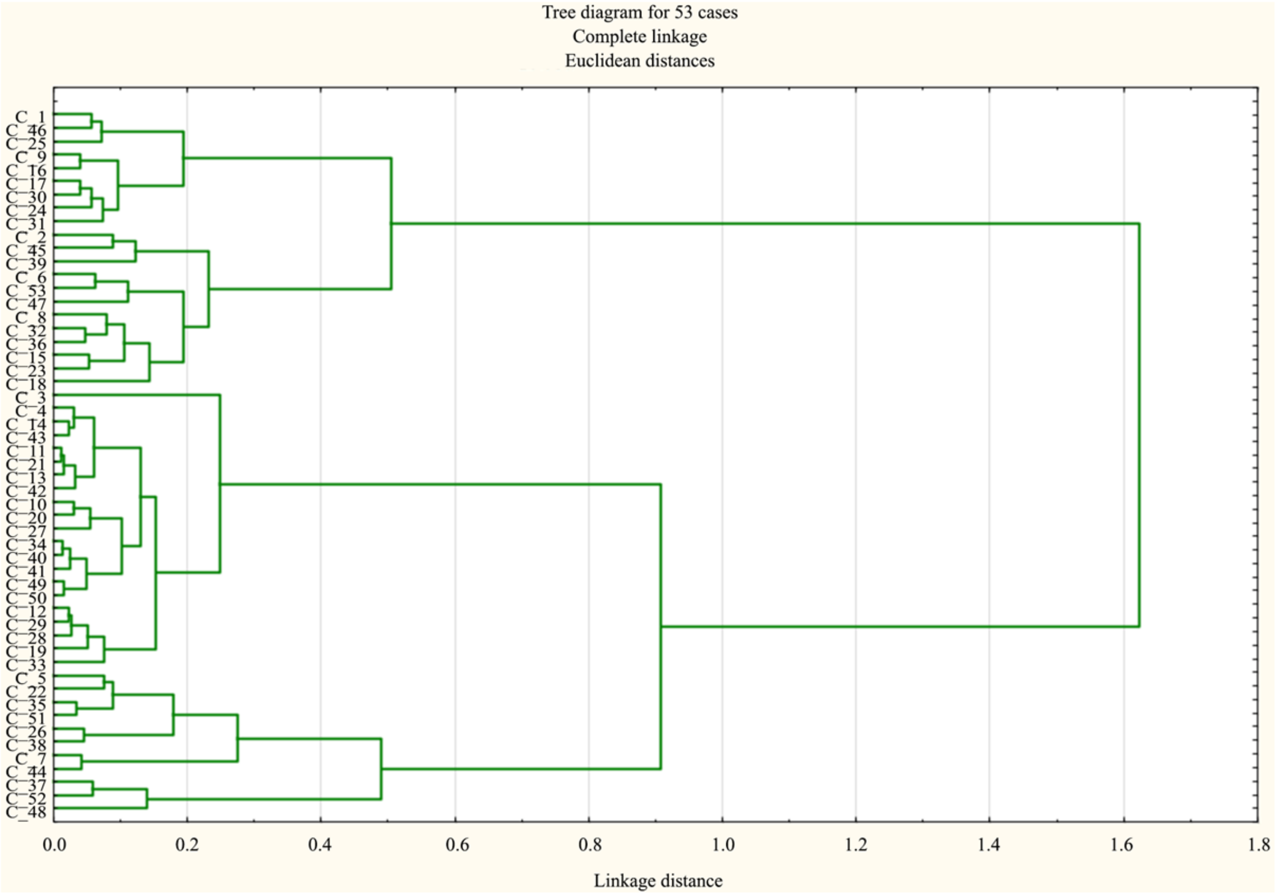


Fig. 6. Result of cluster analysis using Euclidean distance

]

Table 3

Distribution of NDVI coefficients in the germination phase by clusters in the context of regions of the North Kazakhstan region

NDVI germination			
Region	I cluster	II cluster	III cluster
Chaghly	53.92 ha/10.52 %; average NDVI: 0.69	293.33 ha/57.25 %; average NDVI: 0.26	165.15 ha/32.23 %; average NDVI: 0.4
Ayyrtausky	104.35 ha/20.33 %; average NDVI: 0.58	106.76 ha/20.80 %; average NDVI: 0.15	302.13 ha/58.87 %; average NDVI: 0.4
Kyzylzharsky	174.2 ha/33.98 %; average NDVI: 0.78	130.59 ha/25.47 %; average NDVI: 0.5	207.85 ha/40.54 %; average NDVI: 0.66
Akkayinsky	125.58 ha/24.48 %; average NDVI: 0.76	80.42 ha/15.67 %; average NDVI: 0.17	307.07 ha/59.85 %; average NDVI: 0.38
Mamlyutsky	139.88 ha/27.25 %; average NDVI: 0.78	123.28 ha/24.02 %; average NDVI: 0.28	250.08 ha/48.73 %; average NDVI: 0.56

Maps of vegetation indices at the germination phase of spring wheat in five districts of the North Kazakhstan region are shown in Fig. 6.

2. Cluster analysis in the spring wheat tillering phase.

The map of clusters, compiled on the basis of NDVI, ReCI and NDWI indices taken from seedlings, is presented in Fig. 7. 3 clusters were identified, the highest values are represented by cluster 1. The average values of NDVI varied from 0.22 to 0.79 for cluster I, from 0.06 to 0, 52 for cluster II, that is 0.07 to 0.67 for cluster III.

In the spring wheat tillering period, the maximum value of the normalized vegetation index $NDVI_{max}$=0.79 was obtained in the Kyzylzhar district. High NDVI values are also observed in the Mamlyutsky and Akkainsky districts.

The maximum value of the normalized vegetation index NDVI characterizes the intensity of vegetation growth, including agricultural crops. The higher the NDVImax value, the "greener" the vegetation. And also, for a more detailed analysis of the state of spring wheat in the tillering phase, it is possible to conduct an additional study on the vegetation indices ReCI and NDMI. The ReCI index shows the amount of chlorophyll in leaves in the active phase of the growing season [2]. On the satellite images of the Kyzylzharsky, Mamlyutsky and Akkainsky districts, a large percentage of vegetation with a high content of chlorophyll is observed (Fig. 9). In Chagly and Ayyrtausky districts, the content of chlorophyll in crops during the tillering period is very low. In the Kyzylzharsky district, the moisture content of agricultural crops is also observed to be high.

Maps of vegetation indices at the tillering phase of spring wheat in five districts of the North Kazakhstan region are shown in Fig. 8.

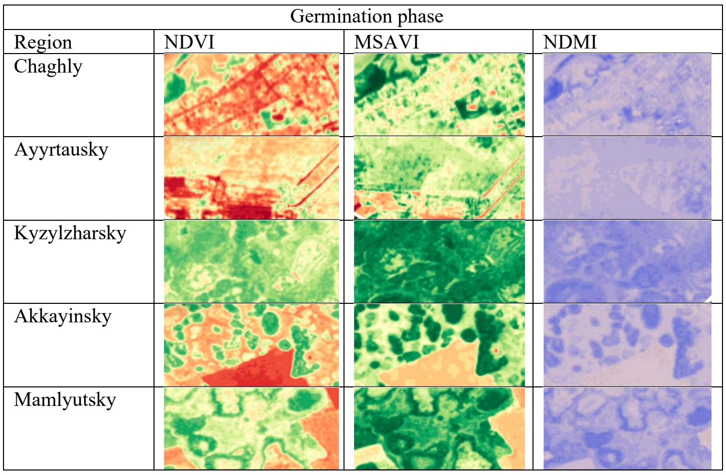


Fig. 7. Maps of vegetation indices NDVI, MSAVI, VDMI during the germination phase of spring wheat

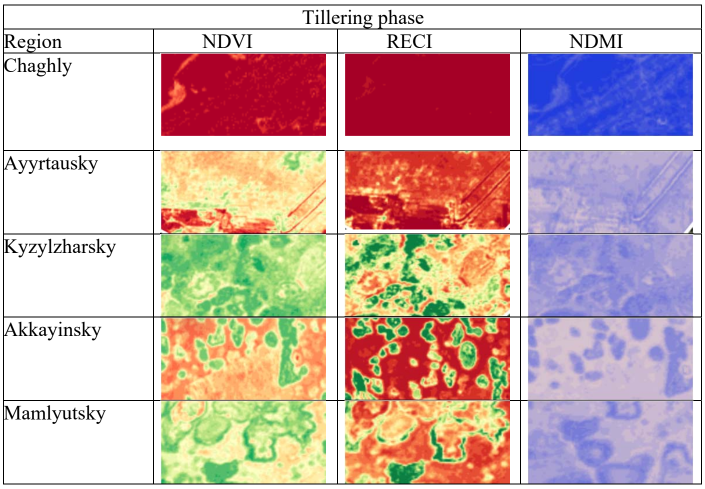


Fig. 8. Maps of vegetation indices in the spring wheat tillering phase

Table 4

Distribution of NDVI coefficients during the tillering phase by clusters across the regions of the North Kazakhstan region

NDVI tillering			
Region	I cluster	II cluster	III cluster 2
Chaghly	13.62 ha/2.66 %; average NDVI: 0.22	47 ha/9.17 %; average NDVI: 0.07	451.77 ha/88.17 %; average NDVI: 0.06
Ayyrtausky	125.56 ha/24.47 %; average NDVI: 0.6	295.53 ha/57.58 %; average NDVI: 0.41	92.14 ha/17.95 %; average NDVI: 0.13
Kyzylzharsky	186.21 ha/36.32 %; average NDVI: 0.79	228.42 ha/44.56 %; average NDVI: 0.67	98.01 ha/19.12 %; average NDVI: 0.52
Akkayinsky	105.62 ha/20.59 %; average NDVI: 0.76	107.57 ha/20.97 %; average NDVI: 0.47	299.87 ha/58.45 %; average NDVI: 0.3
Mamlyutsky	196.63 ha/38.31 %; average NDVI: 0.75	228.87 ha/44.59 %; average NDVI: 0.56	87.73 ha/17.09 %; average NDVI: 0.37

in order to obtain a good harvest in this region, it is recommended to increase the amount of mineral fertilizers during the ripening phase of spring wheat.

Maps of vegetation indices during the ripening phase of spring wheat in five districts of the North Kazakhstan region are shown in Fig. 9.

The normalized humidity index NDMI shows values higher than average in Kyzylzharsky and Mamlyutsky districts. In the Chagly and Ayyrtausky districts, the vegetation moisture indicator shows a moisture deficit in the crops. Accordingly, in the latter, the quantity and quality of agricultural crops is the smallest.

This table demonstrates the values of vegetation indices and their descriptive statistics NDVI in three phenological phases of spring wheat development (Table 6).

Fig. 10 illustrates maps of vegetation indices of five districts of North Kazakhstan after the clustering procedure.

3. Cluster analysis in the ripening phase of spring wheat ripening.

The map of clusters, compiled on the basis of NDVI, NDRE and NDWI indices taken from seedlings, is presented in Fig. 10. 3 clusters were identified, the highest values were represented by cluster 1. Average NDVI values varied from 0.64 to 0.80 for cluster I, from 0.43 to 0.68 for cluster II, from 0.15 to 0.53 for cluster III. The maximum NDVI coefficient was obtained in the Mamlyutsky district $NDVI_{max}$=0.80. Also, high NDVI values were observed in the Akkayinsky and Kyzylzharsky districts. Accordingly, let's investigate the normalized differential Red Edge indices in the ripening phase of spring wheat. This index is used at the stage of ripening of agricultural crops, the maximum value of NDRE shows a high nitrogen content in the leaves of the crop.

In the Mamlyutsky district, coefficient values of $0.65<NDRE_{max}<0.75$ occupied 45 % of the field with dense vegetation and $0.25<NDRE_{min}<0.45$, which revealed 55 % of the field with sparse vegetation. Accordingly,

Table 5

Distribution of NDVI coefficients in the ripening phase by clusters in the section of the regions of the North Kazakhstan region

NDVI ripening			
Region	I cluster	II cluster	III cluster 2
Chaghly	49.42 ha/9.64 %; Average NDVI: 0.68	119.7 ha/23.36 %; Average NDVI: 0.43	343.28 ha/66.99 %; Average NDVI: 0.27
Ayyrtausky	186.59 ha/36.36 %; Average NDVI: 0.64	240.7 ha/46.90 %; Average NDVI: 0.45	85.95 ha/16.75 %; Average NDVI: 0.15
Kyzylzharsky	225.43 ha/43.97 %; Average NDVI: 0.75	223.06 ha/43.51 %; Average NDVI: 0.64	64.15 ha/12.51 %; Average NDVI: 0.5
Akkayinsky	103.29 ha/20.13 %; Average NDVI: 0.76	143.45 ha/27.96 %; Average NDVI: 0.52	266.32 ha/51.91 %; Average NDVI: 0.29
Mamlyutsky	171.56 ha/33.43 %; Average NDVI: 0.8	194.02 ha/37.80 %; Average NDVI: 0.68	147.65 ha/28.77 %; Average NDVI: 0.53

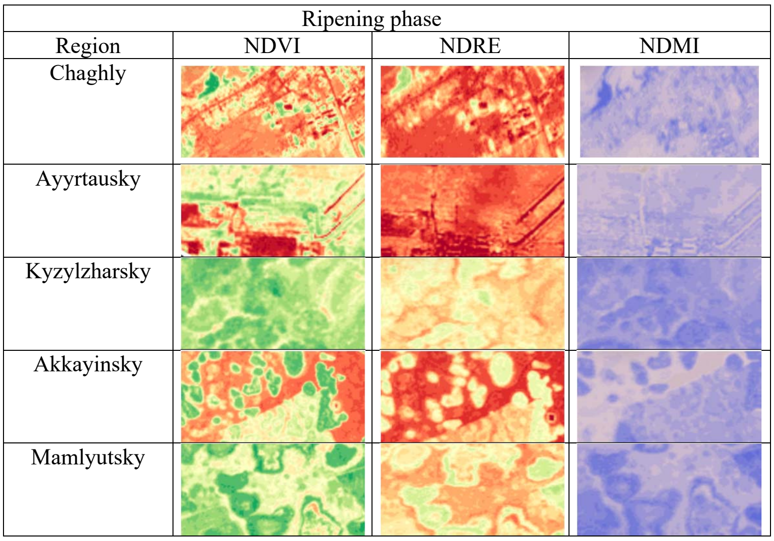


Fig. 9. Maps of vegetation indices during the ripening phase of spring wheat

Table 6

Values of vegetation indices and their descriptive statistics of spring wheat

Region	Date	min	max	aver-age	std	vari-ance
Ayyrtausky	11.07.2022	–0.3237	0.8427	0.5209	0.1549	0.024
Kyzylz-harsky	14.07.2022	–0.3845	1	0.5298	0.2098	0.044
Akkayinsky	14.07.2022	–0.0013	0.824	0.4905	0.14	0.0196
Mamlyutsky	13.07.2022	–0.3866	0.8471	0.5819	0.2018	0.0407
Chaghly	14.07.2022	–0.4954	0.8504	0.3612	0.1169	0.0137
Region	Date	min	max	aver-age	std	vari-ance
Ayyrtausky	14.06.2022	–0.3979	0.8578	0.4835	0.1958	0.0383
Kyzylz-harsky	24.06.2022	–0.3969	1	0.4827	0.2061	0.0425
Akkayinsky	09.06.2022	–0.1532	0.8252	0.3446	0.1394	0.0194
Mamlyutsky	09.06.2022	–0.4501	0.8548	0.2515	0.0864	0.0075
Chaghly	23.06.2022	–0.2753	0.8192	0.2304	0.0411	0.0017
Region	Date	min	max	aver-age	std	vari-ance
Ayyrtausky	06.06.2022	–0.3173	0.8408	0.4452	0.1996	0.0398
Kyzylz-harsky	05.05.2022	–0.4286	0.8568	0.2112	0.1253	0.0157
Akkayinsky	09.06.2022	–0.1532	0.8252	0.3446	0.1394	0.0194
Mamlyutsky	08.05.2022	–0.3952	0.6581	0.1936	0.1592	0.0253
Chaghly	09.06.2022	–0.4501	0.8548	0.2515	0.0864	0.0075

5. 3. Evaluation of the obtained clustering results using the correlation matrix and the level of significance of the parameters

Table 7 shows the values of vegetation coefficients and yield of spring wheat in the germination, tillering and ripening phases. Using these data for 2021–2022 in the regions of the North Kazakhstan region, a correlation matrix was constructed and p-values were calculated (Fig. 11).

Table 7

The yield of spring wheat in the North Kazakhstan region in three vegetation phases by region for 2021–2022

Region	NDVIgerm	NDVItill	NDVImatur	Cwheat_yield, c/ha
Chaghly	0.69	0.22	0.68	17.5
Ayyrtausky	0.58	0.6	0.64	16
Kyzylzharsky	0.78	0.79	0.75	18.5
Akkayinsky	0.76	0.76	0.76	18
Mamlyutsky	0.78	0.75	0.8	17.68

The maximum Pearson correlation coefficient R_{germ}=0.94 is reached in the germination phase with the optimal coefficient of significance p=0.018, which means that the yield of spring wheat in the North Kazakhstan region is 94% dependent on the maximum normalized vegetation index in the tillering phase (Table 8). The lowest correlation coefficient R_{till}=0.33 shows that the values of the normalized vegetation indices in the tillering phase NDVI_{till} weakly correlate with the yield values of the North Kazakhstan region, i.e. the values of these indices in this phase do not determine the amount of spring wheat harvest in this region.

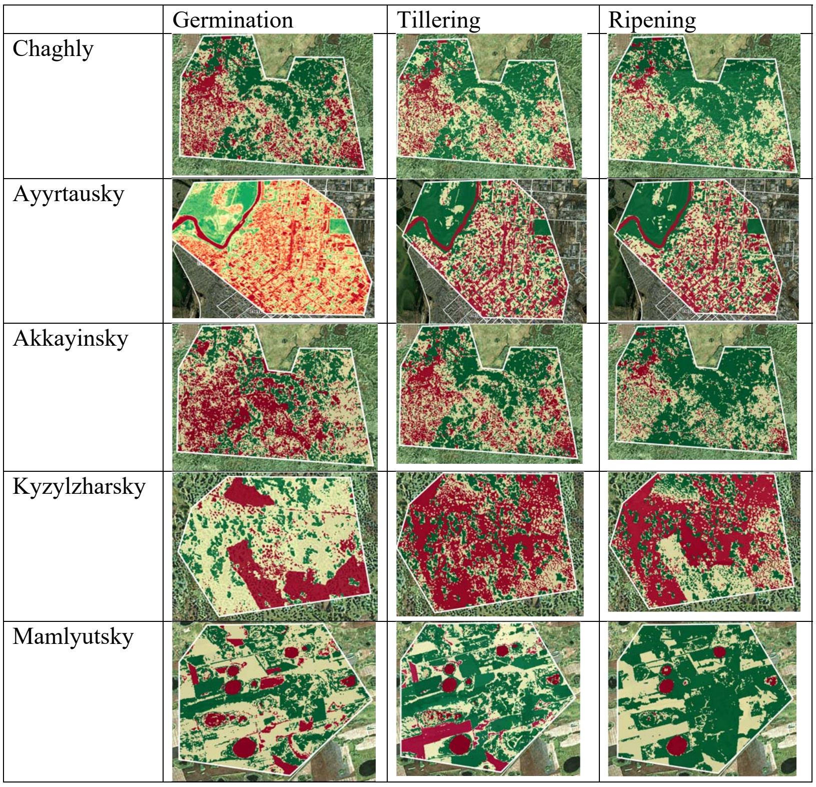


Fig. 10. Maps of vegetation indices after the clustering procedure

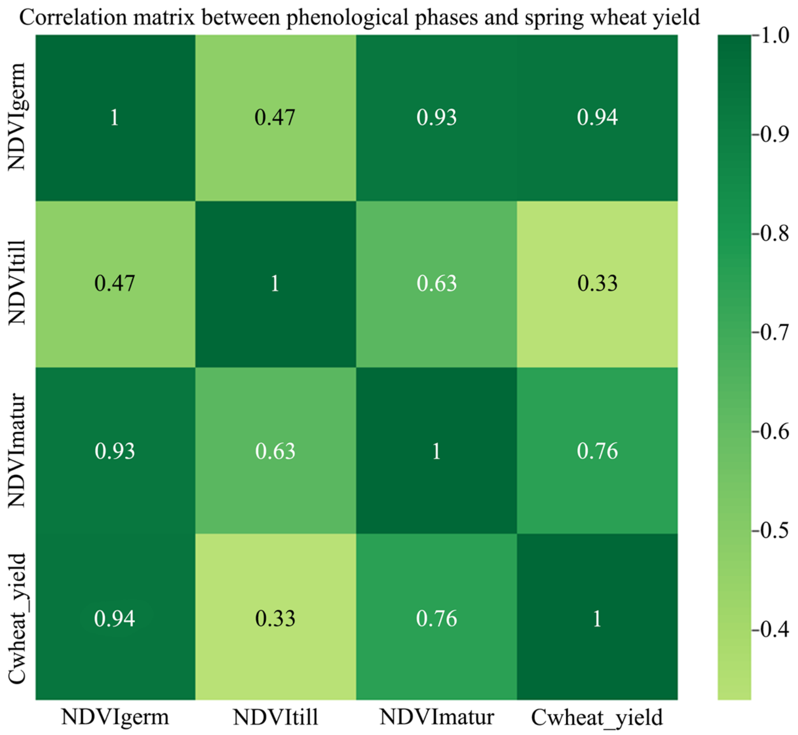


Fig. 11. Correlation matrix between phenological phases of development and yield of spring wheat for 2021–2022

A Table 8 of Pearson correlation coefficients and *p*-value was constructed.

Table 8

Values of Pearson correlation coefficients and coefficient of significance in phenological phases of spring wheat for 2021–2022

Type of assessment	Germination	Tillering	Ripening
Pearson's correlation coefficient	0.94	0.33	0.76
p-value	0.018	0.589	0.140

6. Discussion of the results of the study of the number of clusters of normalized vegetation indices and their descriptive statistics

A feature of the study is the determination of the number of clusters of normalized vegetation indices, taking into account individual characteristics using the k-means algorithm (Fig. 1).

On the basis of our research, a conceptual solution to the analysis of the composition of agricultural crops is proposed using the clustering of digitized remote sensing data of the North-Kazakhstan Agricultural Experiment Station, which are used to determine the quality of photosynthetically active biomass of agricultural crops. Determining the number of clusters by the proposed method will make it possible, in comparison with studies [8–15], to calculate the most favorable cluster in the phenological phase for increasing the productivity of spring wheat in the North Kazakhstan region.

The vegetation indices NDVI, NDRE, OSAVI, NDMI, and ReCI were selected as the main parameters for assessing the condition of spring wheat in three vegetation phases. The above-mentioned vegetation indices taking into account individual characteristics made it possible, in comparison with studies [10–17], to obtain the most effective analysis of the fields of the North Kazakhstan region.

Having information about the coefficients of NDVI and its descriptive statistics with the help of digitized images of agricultural crops, in the phase of flowering or ripening, it is possible to predict the amount of the harvest.

Usually, along with the value of the normalized vegetation index, natural factors (natural soil fertility, weather), biological (organic fertilizers, seeds, hybrids), as well as organizational and technological factors (tillage, mineral fertilizers, meliorants, plant protection products). For the quantitative accounting of each influencing factor, it is necessary to periodically perform direct or indirect measurements, which in some cases is a time-consuming process that is not easy to implement on the scale of industrial crop production. This explains the effectiveness of using cluster analysis with digitized satellite data in the form of normalized vegetation indices and their descriptive statistics, which can be easily obtained through open access geosystems.

The maximum correlation coefficient obtained during the experiment shows a positive, fairly close relationship between the vegetation index in the budding phase NDVIgerm and the yield value in the studied period.

The proposed solution for the application of the k-means clustering method to the normalized difference vegetation indices NDVI, NDRE, MSAVI, NDMI, ReCI can be an example of the analysis of various vegetation indices, such as the weather-resistant vegetation index (ARVI), the vegetation index with soil correction (SAVI), Green Difference Vegetation Index (GDVI), Green Normalized Differential Vegetation Index (GNDVI), Enhanced Vegetation Index (EVI), without resorting to downloading satellite images. Digitized values of satellite images in the form of various vegetation indices and their descriptive statistics can be obtained in the open access of such geoinformation services as EOS Land Viewer, Sentinel Hub, EO Browser.

It should be noted that the effectiveness of remote sensing varies depending on the type of plants. Some varieties react more sensitively to changes in growing conditions and differ better in terms of vegetation indices, while other varieties may be less sensitive to such changes.

The most productive areas of clustering were clearly identified on the basis of remote sensing data. The result of this study shows that vegetation indices, taking into account individual characteristics, can be used to identify areas with high yield potential and more effectively allocate resources and manage cultivation. However, at the harvest stage, the relationship between vegetation indices and agronomic components of yield becomes less significant due to a small difference between agronomic parameters. This may be due to the fact that at this stage other factors, such as mechanical harvesting and other types of tillage, play a more important role in the formation of the final harvest.

However, it is necessary to take into account that the stage of growing plants and their varieties can significantly affect the effectiveness of the application of vegetation indices.

On the basis of the conducted research, the expediency of using cluster analysis for the values of vegetation indices and its descriptive statistics for further use in forecasting productivity in the conditions of the North Kazakhstan region is substantiated.

In general, this study confirms the importance of using vegetation indices taking into account individual characteristics regarding the phases of development of agricultural crops for quality and yield management. However, additional research and consideration of differences between varieties of agricultural crops are required to reduce the influence of territorial restrictions in the phases of development for more accurate forecasting of yield and optimization of cultivation processes.

The disadvantage of the study is that when applying the k-means algorithm, a fixed number of clusters is determined at the initial stage, so the calculated calculation time was exceeded in the process. Thus, in order to reduce this shortcoming, the initial k clusters may first be chosen randomly, and the next k clusters were chosen in proportion to the square of the distance from the nearest center using k-means.

The proposed solution of this research is applicable to different types of vegetation and vegetation indices, when determining the phenological phases of development. However, it is necessary to take into account the climate of the studied region, because the phases of plant development directly depend on agrometeorological conditions. Despite this, determining the number of clusters in this way makes it possible to determine the most favorable growing season for improving the productivity of plants and will strengthen biological and agro-technological measures that increase productivity in this phase of plants.

Further research can be directed to the development of clustering methods that allow to reduce the dependence of cluster definition on territorial conditions.

7. Conclusions

1. To implement continuous monitoring of the state of agricultural crops on the territory of the North Kazakhstan Agricultural Research Station, a database of digitized high-resolution multispectral space images of Sentitel 2 and Landsat 8 for the growing season of 2022 has been collected in the form of vegetation indices NDVI and their descriptive indices, as well as indices NDRE, OSAVI, ReCI, NDMI.

The graph of the time series analysis of the normalized vegetation indices NDVI of spring wheat in the North Kazakhstan region is demonstrated. The values of the coefficients of the linear trend and visual analysis show the phenological phases of the vegetative development of agricultural crops. According to the graph, it is possible to trace the growth trend of spring wheat from the growing season to mid-summer, and then the trend of decline until the earing period.

2. Using the hierarchical tree method, a dendogram has been constructed, with the help of which the main clusters C1–C6 were determined; C7–C37; C38–C53. As a result of the application of the k-means method, all variables were divided into 3 clusters, the stability of which was assessed by the method of variance analysis. The coefficient of significance of each variable in clusters is $p=0.00<0.05$.

3. The values of the correlation matrix between the vegetation index NDVIgerm and the yield of spring wheat in the regions of the North Kazakhstan region have been calculated. The obtained results have been shown that 94 % of the maximum normalized vegetation indices positively correlate with spring wheat yield in the germination phase. Vegetation indices in the tillering phase showed the lowest correlation (33 %).

Clusters of vegetation indices in three phenological phases of spring wheat development in the regions of the North Kazakhstan region were also analyzed.

Conflict of interest

The authors declare that there is no conflict of interest regarding this research, including financial, personal, authorship, or any other nature that could affect the research and its results presented in this article.

Financing

The research was conducted without financial support.

Data availability

The manuscript contains data included as additional electronic material.

References

1. Mutanga, O., Masenyama, A., Sibanda, M. (2023). Spectral saturation in the remote sensing of high-density vegetation traits: A systematic review of progress, challenges, and prospects. ISPRS Journal of Photogrammetry and Remote Sensing, 198, 297–309. doi: https://doi.org/10.1016/j.isprsjprs.2023.03.010
2. Xue, J., Su, B. (2017). Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications. Journal of Sensors, 2017, 1–17. doi: https://doi.org/10.1155/2017/1353691
3. Xiao, X., Braswell, B., Zhang, Q., Boles, S., Frolking, S., Moore, B. (2003). Sensitivity of vegetation indices to atmospheric aerosols: continental-scale observations in Northern Asia. Remote Sensing of Environment, 84 (3), 385–392. doi: https://doi.org/10.1016/s0034-4257(02)00129-3
4. Mimenbayeva, A., Zhukabayeva, T. (2020). A review of free resources for processing and analyzing geospatial data. Proceedings of the 6th International Conference on Engineering & MIS 2020. doi: https://doi.org/10.1145/3410352.3410800

5. Liu, D., Yang, F., Liu, S. (2021). Estimating wheat fractional vegetation cover using a density peak k-means algorithm based on hyperspectral image data. Journal of Integrative Agriculture, 20 (11), 2880–2891. doi: https://doi.org/10.1016/s2095-3119(20)63556-0
6. Komarov, A. A., Kirsanov, A. D., Malashin, S. N. (2021). Comparative characteristics of various vegetation indices (vi) when the vegetation cover state of forage grasses assessing. Izvestya of Saint-Petersburg State Agrarian University, 63 (2), 18–29. doi: https://doi.org/10.24412/2078-1318-2021-2-18-29
7. Somvanshi, S. S., Kumari, M. (2020). Comparative analysis of different vegetation indices with respect to atmospheric particulate pollution using sentinel data. Applied Computing and Geosciences, 7, 100032. doi: https://doi.org/10.1016/j.acags.2020.100032
8. Ntayagabiri, J. P., Ndikumagenge, J., Ndayisaba, L., Philippe, B. K. (2023). Study on the Development and Implementation of Different Big Data Clustering Methods. Open Journal of Applied Sciences, 13 (07), 1163–1177. doi: https://doi.org/10.4236/ojapps.2023.137092
9. Tlebaldinova, A. S., Ponkina, Ye. V., Mansurova, M. Ye., Ixanov, S. Sh. (2021). Using satellite images to assess the state of arable fields on the example of the East Kazakhstan region. Bulletin of the National Engineering Academy of the Republic of Kazakhstan, 82 (4), 179–186. doi: https://doi.org/10.47533/2020.1606-146x.128
10. Pandey, S., Khanna, P. (2014). A hierarchical clustering approach for image datasets. 2014 9th International Conference on Industrial and Information Systems (ICIIS). doi: https://doi.org/10.1109/iciinfs.2014.7036504
11. Prasad, M., Thota, S. (2023). Buddy System Based Alpha Numeric Weight Based Clustering Algorithm with User Threshold. doi: https://doi.org/10.20944/preprints202308.1676.v1
12. Khudov, H., Makoveichuk, O., Komarov, V., Khudov, V., Khizhnyak, I., Bashynskyi, V. et al. (2023). Determination of the number of clusters on images from space optic-electronic observation systems using the k-means algorithm. Eastern-European Journal of Enterprise Technologies, 3 (9 (123)), 60–69. doi: https://doi.org/10.15587/1729-4061.2023.282374
13. Vandana, B., Kumar, S. S. (2019). Hybrid K Mean Clustering Algorithm for Crop Production Analysis in Agriculture. Special Issue, 9 (2S), 9–13. doi: https://doi.org/10.35940/ijitee.b1002.1292s19
14. Umarani, R., Tamilarasi, P. (2019). Data analysis of crop yield prediction using k-means clustering algorithm. Journal of Emerging Technologies and Innovative Research, 6 (4), 535–538. Available at: https://www.jetir.org/papers/JETIR1904582.pdf
15. Marino, S., Alvino, A. (2021). Vegetation Indices Data Clustering for Dynamic Monitoring and Classification of Wheat Yield Crop Traits. Remote Sensing, 13 (4), 541. doi: https://doi.org/10.3390/rs13040541