International Journal of Botany Studies
ISSN: 2455-541X; Impact Factor: RJIF 5.12
Received: 20-11-2020; Accepted: 05-12-2020: Published: 12-12-2020
www.botanyjournals.com
Volume 5; Issue 6; 2020; Page No. 516-522

The use of grouping morphological characteristics of *Lettuce varieties* L. var. *capitata* for the difference test in Ukraine

NV Leschuk[1*], NS Orlenko[2], OV Khareba[3], OJ Dydiv[4]
[1, 2] Ukraine Institute for Plant Variety Examination, Kyiv, Ukraine
[3] National University of Life and Environmental Sciences of Ukraine, Kyiv, Ukraine
[4] Lviv National Agrarian University, Lviv, Ukraine

Abstract
The purpose of our research is to substantiate the use sequence of clustering methods for grouping the morphological characteristics of lettuce varieties var. capitata for the test for the difference. For analytical research, the results of the qualification examination of lettuce varieties for distinctness, uniformity, and stability were used. Analyzed data for the period from 2009 to 2019 year, which obtained by international requirements Guidelines for the conduct of tests for distinctness, uniformity, and stability (TG/13/11). For more precisions, identification varieties for distinctness were used hierarchical cluster analysis and Machine Learning Method in sequence. It was experimentally revealed that the most adequate model of similar varieties groups of lettuce is formed when the attribute "head: size" is used as the target variable and the attribute "head density" as the focal variable. Findings indicate the prospects of using the algorithm of the nearest neighbors in the identification of similar varieties *o Lactuca sativa* L. var. *capitata*.

Keywords: plant variety, lettuce, characteristics, cluster analysis, grouping, difference test

1. Introduction

Identification of new varieties by the method of morphological description of vegetative and generative organs of plants enables to reveal the morphological code formula of the phenotype of the corresponding lettuce variety and to determine the criteria for distinctness, uniformity, and stability for preparing proposals for state registration of the variety (Kalloo and Krug, 1980; Helm, 1954) [4, 3]. It is known that the variety is the main link in the technology of all lettuce types growing. For efficiency and ease of description of new varieties of morphological characteristics, when identifying candidate varieties, they are grouped according to the corresponding characteristics. The grouping of similar varieties within a type (leaf, head, romaine, and stem) is based on the correct choice of the mathematical and statistical apparatus for the classification of plant varieties by morphological characteristics (Orlenko *et al*., 2019) [12]. We should notice that testing for distinctness uniformity and stability is a way to determine if a variety (for which an application is submitted for inclusion in the State Register of Plant Varieties suitable for distribution in Ukraine) differs from the well-known varieties of *Lactuca sativa* var. *capitata* L. (partial distinctness) or pronounced characteristics used to reveal the distinctness, are uniformly the same (partial uniformity) and these characteristics do not change during reproduction over subsequent generations (partial stability). Varieties of head lettuce *Lactuca sativa* var. *capitata* L. are identified by 42 morphological features during the examination to determine the criteria of distinctness, uniformity, and stability (DUS-test). To date, there is not a single statistical method for establishing an unambiguous distinctness between one variety and another. Distinctness, if any, are revealed by observation and measurement through internationally agreed protocols of countries within The International Union for the Protection of New Varieties of Plants (UPOV), based in Geneva, and the Community Plant Variety Office (CPVO), an agency of the European Union, located in Anger, France. According to the authors, the automated grouping of the most similar varieties with subsequent visual identification of distinctive varieties within the group should help to simplify the technological procedure for determining the distinctness of the variety. The classification methods suitable for processing field and laboratory data of plant varieties range from very simple to very complex. Therefore, one should be careful when choosing statistical methods, in particular, cluster. Usually, the classification is carried out according to the following stages: the data of clustering objects is selected (in our case, this is a dataset of head lettuce), many variables are determined to evaluate the objects in the sample and when necessary, the values of variables are normalized (in the context of this study, this is the length of the outer part of the perianth, the width of the outer part of the perianth, the length of the inner part of the perianth and the width of the inner part of the perianth, calculation of the values of the similarity measures between objects). Closest similarity analysis is a method for classifying observations gained from an examination for distinctness, uniformity, and stability (DUS) based on the similarity of the results obtained in the study of morphological traits of plant varieties (observations on morphological traits at various phenological stages of plant growth and laboratory analysis of the manifestation of plant properties). Thus, the distance between two observations is a criterion for their difference (Leshchuk *et al*., 2019) [6]. For computer processing of morphological signs data (codes of sign manifestation), a nominal and ordinal scale was used. The nominal scale was used to group similar varieties according to the following morphological characteristics: "seed: color", "seedling: anthocyanin coloration", "seedling:

cotyledons by size (at full development)", "leaf: position at the stage of 10-12 leaves", "leaf: position according to harvest ripeness (outer leaves in the head lettuce)", "leaf: shape", and the ordinal scale for the features: "plant: diameter", "plant: head formation", "head: by density", "head: size ", "leaf: by thickness ", "leaf: the color intensity of the outer leaves ". It is proposed to carry out such a grouping with the consistent use of hierarchical cluster analysis and machine learning, which will allow classifying similar varieties most accurately. The manifestation of morphological traits by which they differ should be revealed among similar varieties. This will significantly reduce the processing time of the results of the DUS examination and make this process less laborious and more reliable because intuitive, introspective estimation can be applied only for small sets of objects.

2. Materials and Methods

2.1. Description of study data and methodology: The results of identification of head lettuce varieties were entered into the source databases of the Ukrainian Institute for Plant Variety Examination, which is the basis for an expert opinion on the compliance of the new variety with the protection criteria. For an analytical study, the results of the qualification examination of head lettuce varieties for distinctness, uniformity, and stability (DUS-test) were used. During the period 2009-2019 we analyzed data obtained by international requirements of "Guidelines for the conduct of tests for distinctness, uniformity, and stability (TG / 13/11 lettuce UPOV code (s): lactu_sat, 2017) [12] and "The method of lettuce varieties (*Lactuca sativa* L.) examination on Distinctness, Uniformity, and Stability" (UPOV, 2017; Leshchuk, 2007) [7]. The object of our research was the data of the morphological description of 30 cultivars *Lactuca sativa* L. var. *capitata* of domestic and foreign selection (Ukraine, Netherlands, Germany, France, Czech Republic, Poland), included in the State Register of Plant Varieties, suitable for distribution in Ukraine (State register of plant varieties suitable for dissemination in Ukraine, 2019) [2].

2.2. Description of study area: The varieties Lactuca sativa var. capitata were studied on four experimental plots in each of the two Plant division region research farm owned by the Ukrainian Institute for Plant Variety Examination (UIPVE). The first study plots were located in the Yakymivs`kyi division at the Zaporizhia region in south-east Ukraine (approximately 46.4′ N. 35, 2′ E). The other plates were in the Gostomel division of UIPVE at the Kiyv region (approximately 50, 33′ N, 30, 2′ E) central Ukraine.

2.3. Description of statistical analysis stages: According to the recommendations provided in the publications, at the first stage of data processing, the clustering results were compared using various methods and metrics of hierarchical cluster analysis and identifying the most suitable morphological characteristics for analysis, namely classification (Nearest Neighbor, Furthest Neighbor, Median Clustering, Centroid Clustering, and Ward's Method) using Euclidean and non-Euclidean metrics. It was found that the most suitable for clustering is Ward's method using the Euclidean metric (Orlenko and Mazhuha, 2019) [11].

Ward's method: minimizes the sum of the squares of the criterion. The distance between two clusters is defined by the formula:

$$D_{kj} = \frac{\| \bar{x}_k - \bar{x}_j \|^2}{\frac{1}{N_k} - \frac{1}{N_j}}$$

The classical Euclidean Distance Metric was used. The formula for this distance is:

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

The following variables were selected for clustering: Head density, Head: size, Head: shape in longitudinal section, Plant: diameter.

At the second stage data of first and fourth clusters were processed by Machine Learning Methods with the *K*-Nearest Neighbors Algorithm (Leskovets *et al*., 2016; Marmanis and Babenko, 2011) [8, 9]. According to this algorithm (Nasledov, 2013) [10], it is assumed that there are already a certain number of objects with an accurate classification (in our case, similar varieties of plants of head lettuce), and it is necessary to develop a rule that allows attributing a new variety to one of the possible classes (a set of morphologically similar plant varieties). And the KNN algorithm selects a coefficient that determines the degree of similarity for new varieties of head lettuce, and k is the number of records that will be considered close using the following rules:

1. 1. $d(x,y) \geq 0$, $d(x,y) = 0$ if and only if $x = y$;
2. $d(x,y) = d(y,x)$;
3. $d(x,z) \leq d(x,y) + d(y,z)$, provided that the points x, y, z do not lie on one straight line (Lantz, 2013; Compton**,** 1994) [5, 1]. Where x, y, z are feature vectors of the objects that are being compared. The ordering of the attribute values was carried out using the Euclidean distance.

As the target variable "Time of harvest maturity" was chosen, the focal case identifier was "Plant: head formation", the case label was "Name of the variety". The following signs make up a list of features: Leaf: attitude at 10-12 leaf stage: Leaf blade: division: Plant: diameter: Plant: head formation: Varieties with closed head formation only: Head: the degree of overlapping of the upper part of leaves: Head: density: Head: size: Head: shape in longitudinal section: Leaf: thickness: Leaf: shape: Leaf: the shape of tip: Leaf: the hue of green color of outer leaves: Leaf: the intensity of the color of outer leaves: Leaf: glossiness of upper side: Leaf: blistering: Leaf: the size of blisters: Leaf blade: a degree of undulation of margin: Leaf blade: incisions of margin on the apical part: Leaf blade: venation: Axillary sprouting; Time of beginning of bolting under long-day conditions.

3. Results and Discussion

The results of the frequency analysis of the manifestation of head lettuce morphological traits are presented in Figure 1.

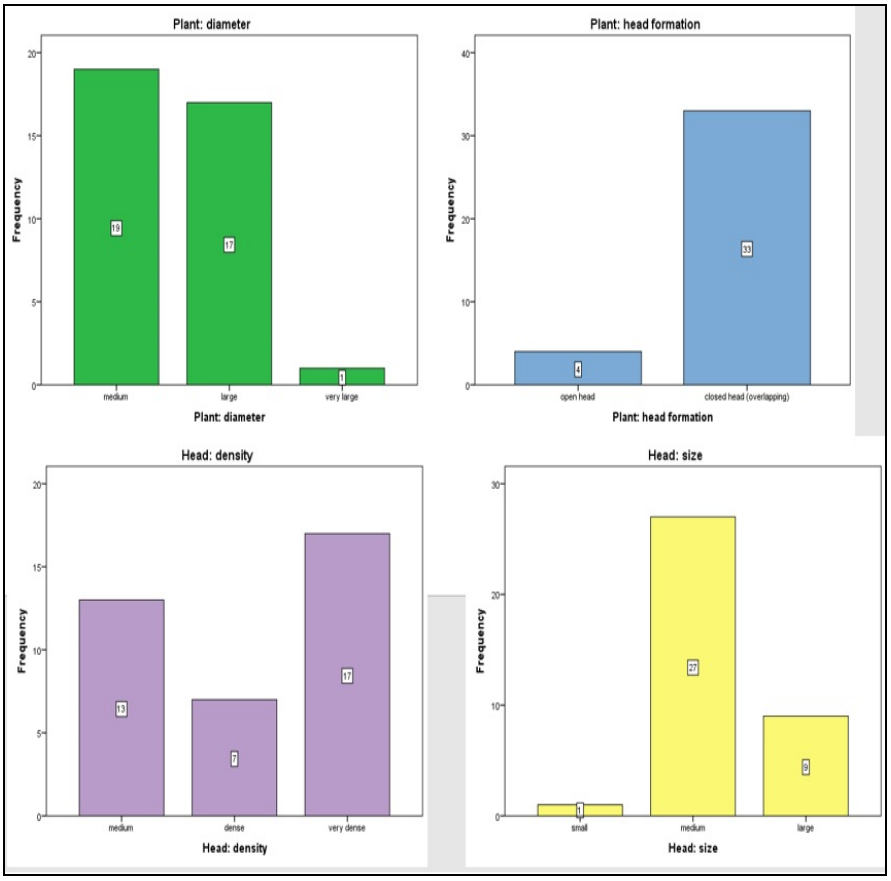

Fig 1 : Grouping of common varieties of *Lactuca sativa* var. *capitata* L. by morphological traits.

The list of each group varieties is given in Figure 2 and Table 1.

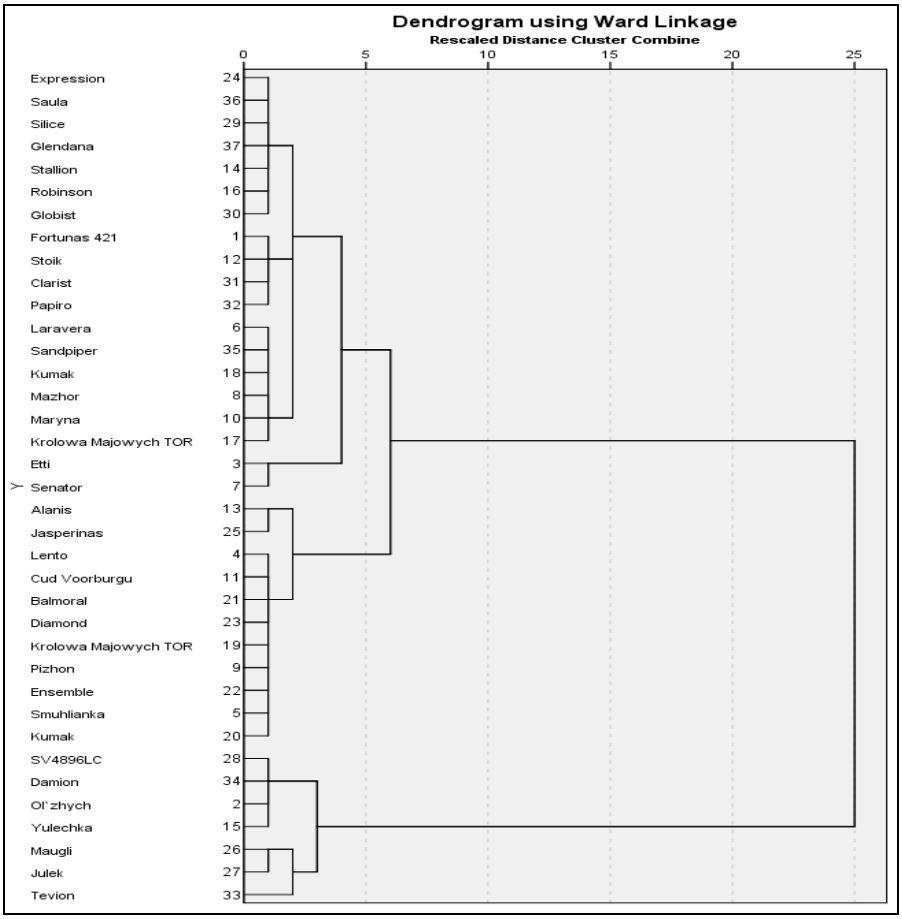

Fig 2: Dendrogram of clustering lettuce varieties using Ward's Method.

The collection of head lettuce varieties, which are included in the State Register of Plant Varieties suitable for distribution in Ukraine, consists mainly of varieties that are characterized by a closed head, medium size with high or very high density. As a result of cluster analysis, groups of lettuce were formed. The first group includes 15 varieties. This group contains varieties with a closed head shape, large or medium size, dense or very dense. The second group includes four varieties with a closed and very dense head. The third cluster consists of only two varieties with a closed, dense, and large head. The fourth cluster was formed by varieties with an open head shape of medium density and medium size. In the fifth cluster, there are three varieties with a closed head of medium density.

Table 1: Cluster distribution of head lettuce varieties

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Fortunas 421	Ol`zhych	Etti	Lento	Maugli
Laravera	Yulechka	Senator	Smuhlianka	Julek
Mazhor	SV4896LC		Pizhon	Tevion
Maryna	Damion		Cud Voorburgu	
Stoik			Alanis	
Stallion			Balmoral	
Robinson			Ensemble	
Krolowa Majowych TOR			Diamond	
Kumak			Jasperinas	
Expression			Balmoral	
Silice			Ensemble	
Globist				
Clarist				
Papiro				
Sandpiper				

The first cluster included 15 varieties, and the fourth - 10 ones. There was a need to more accurately identify similar varieties separately for each group using the Mechine learning method. The simulation results using this method are visualized on the "Predictor Space" (Fig. 3 and 5) and Peers Chart (Fig. 4 and 6). The Metric Space Chart is interactive. Each axis represents a measure in the model, and the location of the points on the chart shows the values of those measures for observations in the training and control groups. The diagram shows the relationship between the data in the training sample and the resulting sample. The circle indicates the varieties that are part of the training sample, and the triangle indicates the varieties of the control sample. The dots in the diagram represent varieties *Lactuca sativa* L. var. *capitata* selected as focal values. The location of the points on the diagram shows the values of these indicators for observations in the training and control groups. A bold outline indicates that the observation is focal. Focal observations are connected to their k Nearest Neighbors. The varieties of the first group, in the process of clustering using the KNN algorithm, were distributed as follows: 85.7% Training and 14.3% Holdout.

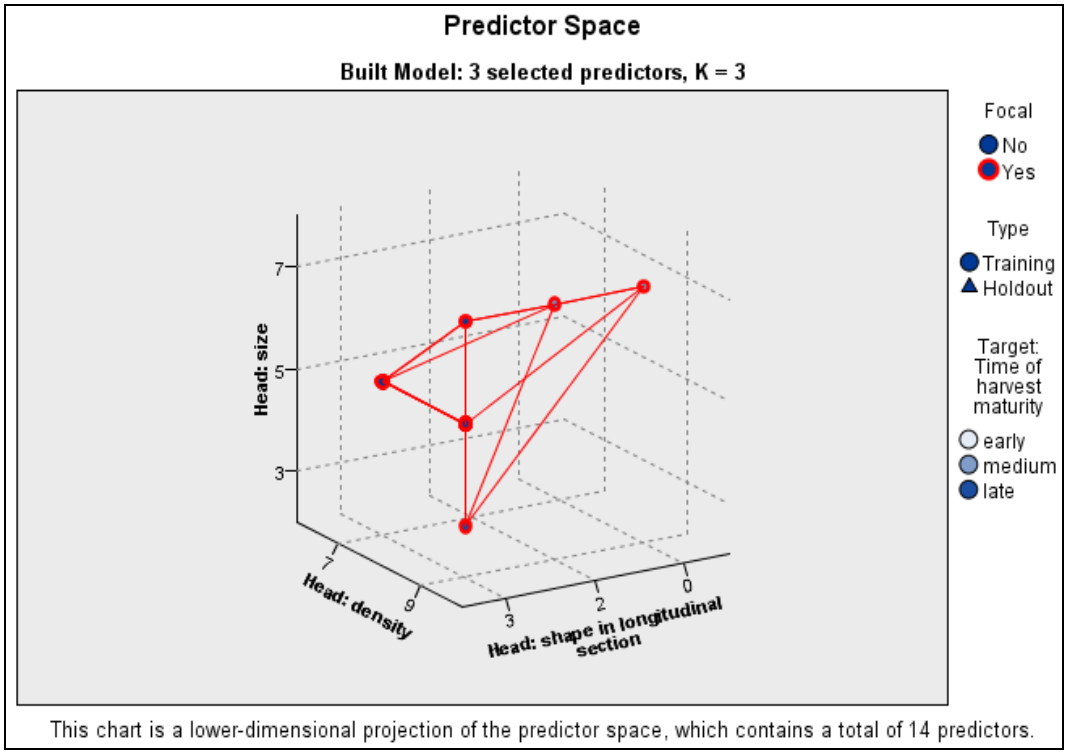

Fig 3: Diagram Predictor Space of the first cluster group of varieties

According to Table 2, the smallest distance is 2.00.

Table 2: Distances of the "Nearest Neighbors" varieties of the first group

KNN_ Focal Case_ Case Number	KNN_ Nearest Neighbor_ Case Number_1	KNN_ Nearest Neighbor_ Case Number_2	KNN_ Nearest Neighbor_ Case Number_3	KNN_ Nearest Neighbor Distance_1	KNN_ Nearest Neighbor Distance_2	KNN_ Nearest Neighbor Distance_3
Fortunas 421	Stoik	Laravera	Robinson	3.16	3.46	3.74
Laravera	Maryna	Mazhor	Papiro	3.16	3.16	3.46
Mazhor	Laravera	Kumak	Papiro	3.16	3.46	3.46
Maryna	Robinson	Laravera	Krolowa Majowych Tor	2.83	3.16	3.16
Stoik	Fortunas 421	Silice	Robinson	3.16	3.16	3.46
Stallion	Krolowa MajowychTor	Globist	Expression	2.45	2.83	2.83
Robinson	Expression	Maryna	Krolowa Majowych Tor	2.83	2.83	2.83
Krolowa Majowych	Globist	Expression	Stallion	2.00	2.00	2.45
Kumak	Clarist	Glendana	Silice	2.83	3.46	3.46
Expression	Silice	Globist	Krolowa Majowych Tor	2.00	2.00	2.00
Silice	Expression	Globist	Krolowa Majowych Tor	2.00	2.00	2.83
Globist	Krolowa Majowych Tor	Silice	Expression	2.00	2.00	2.00
Clarist	Papiro	Kumak	Globist	2.45	2.83	3.16
Papiro	Clarist	Globist	Silice	2.45	2.83	2.83
Silice	Glendana	Robinson	Expression	2.00	2.45	2.83
Glendana	Robinson	Maryna	Kumak	3.16	3.16	3.46

This indicates that the varieties 'Expression', 'Silice', 'Globist', 'Krolowa Majowych Tor'are the most similar. Common manifestations of the traits of these varieties are a closed head of medium size with weak blistering of small leaves with an obvious dissection of the tip edge and fan-shaped venation. The onset of peduncle emergence in long day conditions is very late. Cultivars 'Glendana' and 'Stallion' have a distance of 2.45 to these cultivars.

The Peers Chart diagram allows us to estimate manifestation of what signs distinguishes varieties to a greater extent. Common manifestations of the traits of these varieties are a closed head of medium size with weak blistering of small leaves with an obvious dissection of the tip edge and fan-shaped venation. The onset of peduncle emergence in long day conditions is very late. Cultivars 'Glendana' and 'Stallion' have a distance of 2.45 to these cultivars.

As Figure 4 shows, the variety 'Krolowa Majowych TOR' is distinguished from all varieties by the Time of harvest maturity attribute, which is very late; the varieties 'Fortunas 421','Stoik', 'Stallion' with dense heads differ according to the Head: density trait; the 'Papiro' variety, which has a small head size, differs according to the Head: size attribute; the 'Laravera' variety, which has a wide-elliptical head shape, differs according to the Head: shape in longitudinal section.; according to the Plant: diameter traits the variety 'Stallion' with a very large diameter differs from other varieties; the variety 'Fortunas 421' with anthocyanin coloration of seedlings differs according to the trait Seedling: anthocyanin coloration

Fig 4: Diagram Peers Chart for the first variety group

Figure 5 displays the model for the second cluster group, the varieties of which, in the clustering process using the KNN algorithm, were distributed as follows: 70% Training and 30% Holdout.

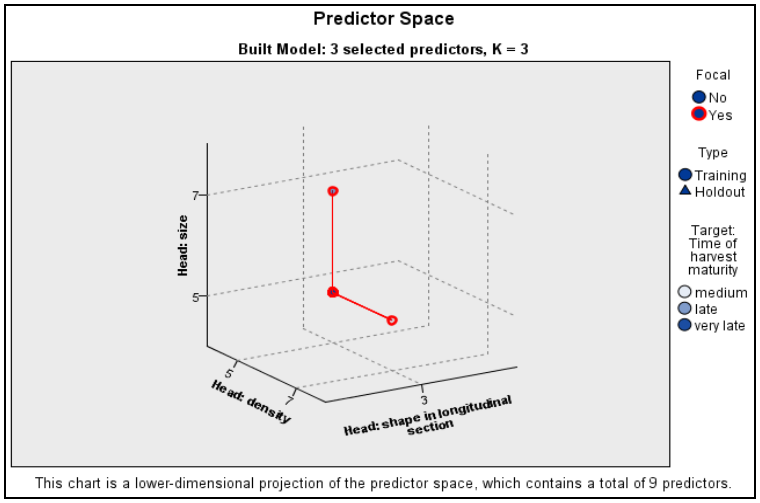

Fig 5: Diagram Predictor Space of the second cluster group of varieties

According to Table 3, the smallest distance is 1.41, which indicates that varieties 'Balmoral', 'Diamond' and 'Cud Voorburgu' are the most similar.

Table 3: Distances of the "Nearest Neighbors" varieties of the second group.

KNN_ Focal Case_ Case Number	KNN_ Nearest Neighbor_ Case Number_1	KNN_ Nearest Neighbor_ Case Number_2	KNN_ Nearest Neighbor_ Case Number_3	KNN_ Nearest Neighbor Distance_1	KNN_ Nearest Neighbor Distance_2	KNN_ Nearest Neighbor Distance_3
Lento	Cud Voorburgu	Jasperinas	Ensemble	2.00	2.45	2.45
Smuhlianka	Lento	Kumak	Ensemble	2.45	2.83	2.83
Pizhon	Balmoral	Alanis	Diamond	2.00	2.45	2.45
Cud Voorburgu	Balmoral	Lento	Jasperinas	1.41	2.00	2.00
Alanis	Balmoral	Pizhon	Diamond	2.00	2.45	2.45
Balmoral	Diamond	Cud Voorburgu	Krolowa Majowych TOR	1.41	1.41	2.00
Ensemble	Kumak	Cud Voorburgu	Lento	2.00	2.00	2.45
Jasperina	Cud Voorburgu	Diamond	Lento	2.00	2.00	2.45

Common manifestations of the traits of these varieties are the absence of anthocyanin color of the seedling, undivided leaf blade edge, average diameter of plants, round shape of a longitudinal section of the head of medium density, weak axillary branching, fan-shaped venation of the leaf blade with the existing dissection of the edge of the apex, weak glossiness of the upper side of the leaf and weak puffiness. These varieties have a very late time of the onset of peduncle emergence in conditions of a long day.

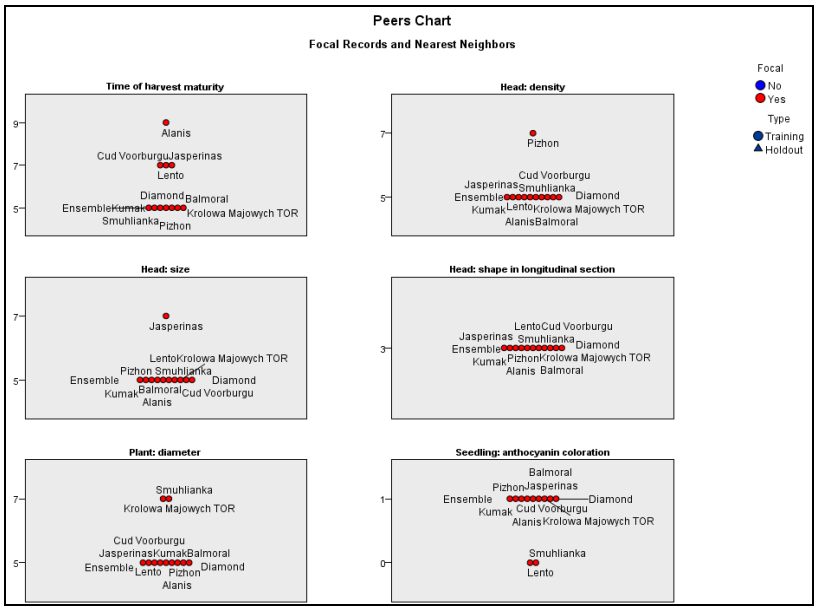

Fig 6: Diagram Peers Chart for fours variety group**.**

As Figure 6 shows, 'Alanis' is distinguished by the Time of harvest maturity feature, 'Pizhon' differs from all the others in terms of head density, 'Jasperinas' differs in head size, and 'Smuhlianka' by plant size and the presence of anthocyanin seedling coloring.

4. Conclusions

Cluster analysis is an effective tool for analyzing expertise data on DUS since it greatly facilitates the search for patterns among a large data set. Also, the presence of a division of data into training and control allows us to "train" the model on a dataset of common varieties. Information technology tools are a very effective tool for analyzing expert examination data at DUS and greatly facilitate the search for patterns among a large set of data on the actual manifestation of morphological characteristics of plant varieties. It was experimentally revealed that the most adequate model of similar varieties groups of head lettuce is formed when the “head: size” feature is used as the target variable and the “head density” feature as the focal variable. The obtained results indicate that the use of the Nearest Neighbors algorithm is promising for the identification of similar varieties of lettuce *Lactuca sativa* L. var. *capitata*.

5. References

1. Compton ME. Statistical methods are suitable for the analysis of plant tissue culture data. Plant Cell Tiss. Organ Cult. 1994; 37(3):217-242. doi: 10.1007/BF00042336
2. Derzhavnyi reiestr sortiv roslyn, prydatnykh dlia poshyrennia v Ukraini na 2019 rik [State register of plant varieties suitable for dissemination in Ukraine in 2019], 2019. Retrieved from https://sops.gov.ua/reestr-sortiv-roslin. (Ua)
3. Helm J, *Lactuca sativa* L. in morphologisch-systematischer Sicht. Die Kulturpflanze. 1954; 2(1):72-129. doi: 10.1007/BF02095730
4. Kalloo, Krug H. Sortendifferenzierung bei Kopfsalat (*Lactuca sativa var*. capitata) – Vorläufige Mitteilung. Die Gartenbauwissenschaft. 1980; 45(3):116-120.
5. Lantz B. Machine Learning with R. Learn how to use R to apply powerful machine learning methods and gain an insight into real-world applications. Birmingham: Packt Publishing Ltd, 2013.
6. Leshchuk NV, Orlenko NS, Khareba OV. Informatsiino-tekhnichni osoblyvosti testu na vidminnist novykh sortiv *Lactuca sativa* var. capitata L. Plant Varieties Studying and Protection. 2019; 15(3):241-248. doi:10.21498/2518-1017.15.3.2019.181081 (Ua)
7. Leshchuk NV. The technique of examination of varie-ties of lettuce (*Lactuca sativa* L.) for distinctness, uniformi-ty and stability. Okhorona prav na sorty roslyn [Protection of Rights to Plant Varieties]. 2007; 3(2):366-379. (Ua)
8. Leskovets Yu, Radzharaman Yu, Ulman D. Analiz bol’shikh naborov dannykh [Analysis of large data sets]. Moscow: DMK. (Ru), 2016.
9. Marmanis Kh, Babenko D. Algoritmy intellektual’nogo Interneta. Peredovye metodiki sbora, analiza i obrabotki dannykh [Algorithms of the intellectual Internet. Advanced techniques for data collecting, analyzing and processing]. Moscow: Simvol. (Ru), 2011.
10. Nasledov AD. IMB SPSS Statistics 20 i AMOS: professional’nyy statisticheskiy analiz dannykh [IMB SPSS Statistics 20 and AMOS: professional statistical data analysis]. St. Petersburg: Piter. (Ru), 2013.
11. Orlenko NS, Mazhuha KM, Dushar MB, Maslechkin VV. Comparative analysis of clustering methods sui-table for plant varieties morphological characteristics data processing. Visnyk poltavskoi derzhavnoi ahrarnoi akademii [News of Poltava State Agrarian Academy]. 2019; 2: 261-269. doi: 10.31210/visnyk2019.02.35 (Ua)
12. Orlenko NS, Leshchuk NV, Symonenko NV, Tahantsova MM. Osoblyvosti vykorystannia zasobiv Machine learning pid chas identyfikatsii podibnykh sortiv roslyn (na prykladi *Lactuca sativa* L. var. sapitata). Visnyk poltavskoi derzhavnoi ahrarnoi akademii [News of Poltava State Agrarian Academy]. 2019; 4:233-240. doi: 10.31210/visnyk2019.04.30 (Ua)
13. UPOV. *Lettuce Lactuca sativa.* Guidelines for the conduct of tests for distinctness, uniformity and stability (TG/13/11TG/13/11), 2017. Retrieved from https://www.upov.int/edocs/tgdocs/en/tg013.pdf