UDC 631.526.2:582.623:311.12

Using cluster analysis as a method of classification of the genus *Salix* L. representatives

M. V. Roik, *Doctor of Agriculture*
V. V. Balykina, *postgraduate student*
Institute of bioenergy crops and sugar beet NAAS of Ukraine
victoria.mamaisur@gmail.com

Purpose. To study interactions among the representatives of the genus Salix L. through the cluster analysis, form groups of closely related species and hybrid forms basing on differences of morphological parameters of leaves. **Methods.** Field, cluster analysis and tree graphics. **Results.** Willow species were grouped according to absolute parameters of leaf, and three groups of clusters were identified. The degree of affinity between species were assessed using values of an Euclidean distance. Distinctive features of leaf parameters were defined: length of a leaf blade *(Ll)*, distance (cm) between the leaf tip and its maximum width *(SDmxT)* and the distance between the leaf tip (cm) and the line of its width that corresponds to the length of petiole *(SLpT)*. **Conclusions.** Using the willow species collection as an example, diagnostically valuable quantitative parameters of leaves were revealed, the use of which allows to identify willow species and hybrid forms through PC applications.

Keywords: cluster analysis, hierarchical method, K-means method, species identification, leaf parameters.

Introduction. The effectiveness of the whole process of plant breeding depends on high-quality selection of parent material. Cluster analysis can be an Example of searching of advanced techniques of selection – one of the methods of mathematical processing of experimental results [1].

The formation of natural hybrids is typical for willow, so in order to improve the quality of directed hybridization it is advisable to classify samples collection, dividing them into several different groups. Some representatives of groups can be selected for further breeding work.

But not always differences between species and hybrid forms are distinct. An important diagnostic feature is the shape of different kinds of willow leaf, reflecting the complex of quantitative characteristics and varies from round to linear-lanceolate. But the parameter is characterized by considerable intraspecific variability, which complicates its use to identify the species. Mathematical methods allow for complex quantitative parameters to determine samples belonging to certain groups [2, 3].

The purpose of research – by using cluster analysis to explore the relationship between the genus *Salix* L., based on morphological differences in the parameters of leaves and to form groups of species and hybrid forms, closely related to each other.

Materials and methods. Analysis was performed on the basis of research conducted at the Institute of bioenergy crops and sugar beet NAAS for 2011–2014, according to the NDP 22 «Bioenergy resources» sub-programme № 5 «Solid fuels». The collection consisted of 21 species and hybrid forms of willow. For research willow leaves, picked in the spring and summer (50 leaves of five plants each sample) were used.

Differences in form and size of leaves helped to identify nine parameters to measure: the length of leaf blade (*Ll*), width of leaf blade (*Dmx*), length of petiole (*Lp*), distance (cm) from the top of the leaf to its maximum width (*SDmxT*), the distance from the base of the leaf (cm) to its maximum width (*SDmxB*), the width of the leaf at a distance of 0,1*Dmx* from the top (*DmnT*), the width of the leaf at a distance of 0,1*Dmx* from the base (*DmnB*), location (cm) of width of leaf corresponding to the length of the petiole from the top of the leaf (*SLpT*), location (cm) of width of leaf corresponding to the length of the petiole from the base of the leaf (*SLpB*) [4].

Samples were classified by leaf morphological characteristics using methods of cluster analysis – hierarchical cluster analysis and the method of K-means using Cluster Analysis module of STATISTISA program. Hierarchical cluster analysis method was used for making a preliminary decision about the number of clusters (groups) which should be divided amount of source material.

Results and discussion. To set the required number of clusters was used an indicator, that is entitled «metrics» – conditional distance between two clusters selected on the basis of distance measures adopted taking into account conversion values. In this case – square Euclidean distances, that was determined by using standardized values. At the point where a measure of the distance between two clusters is changed abruptly, the process of integration into a new cluster stops to avoid association clusters that are relatively large distance from each other.

Definition of a sufficient number of clusters can be based on analysis of incremental diagram of changes intercluster distance. As a sufficient number of clusters is accepted that the difference is the number of steps, after which intercluster distance increased abruptly [5, 6].

For the classification method of Ward was used, aimed at uniting closely located clusters. The results are shown in Figure 1.

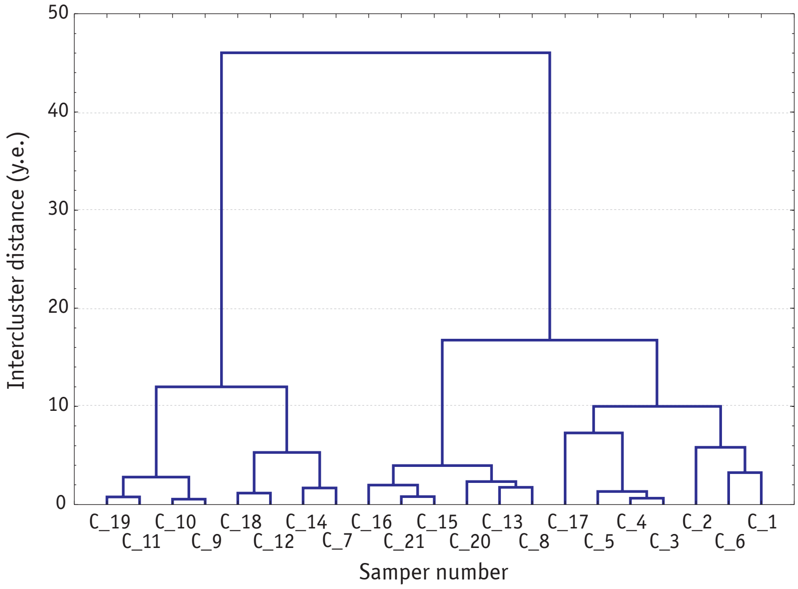

Figure 1. Dendrohrama of clustering of willow samples by morphological parameters of leaf (method of Ward)

Figure 2 shows the graph of incremental changes intercluster distance, through his analysis was determined the required number of clusters. As can be seen from the graph, an abrupt increase intercluster distance held within 19 step. Because the research is subject of 21 sample, it was only three clusters.

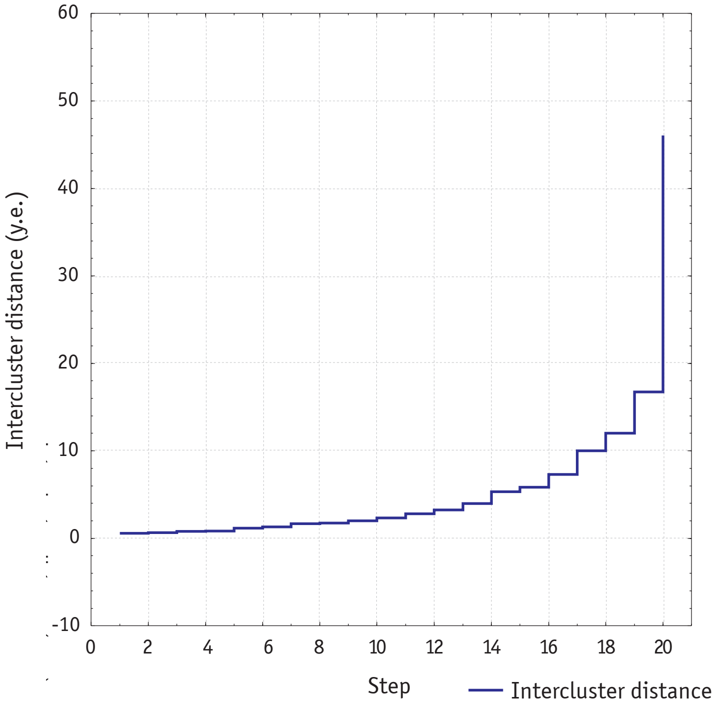

Figure 2. Graph of incremental changes of intercluster distance

Using this method made it possible to decide about the number of clusters, required for further analysis. The results were used during clustering by the method of K-means. As a result of cluster analysis were grouped species in absolute terms of willow leaf and were defined the following groups and clusters: cluster I – *S. caspica* Pall., *S. uralensis* Hort. ex K. Koch., *S. alba* L. *f. splendens*, *Salix rosmarinifolia* L., *S. purpurea* L. × *S. viminalis* L., *S. caspica* Pall. × *S. purpurea* L., *S. integra* Thunb L. × *S. acutifolia* Mild., *S. viminalis* L. × *S. caprea* L.; cluster II – *S. acutifolia* Mild., *S. alba* L., *S. alba* L. local form, *S. cangensis* Nakai, *S. viminalis* L. × *S. acutifolia* Mild., *S. caprea* L. × *S. purpurea* L.; cluster III – *S. viminalis* L., *S. triandra* L. local form, *S. triandra* L., *S. matsudana* Vill., *S. cinerea* L., *S. repens* L., [(*S. viminalis* L. × *S. purpurea* L.) × (*S. caspica* Pall. × *S. caprea* L.)].

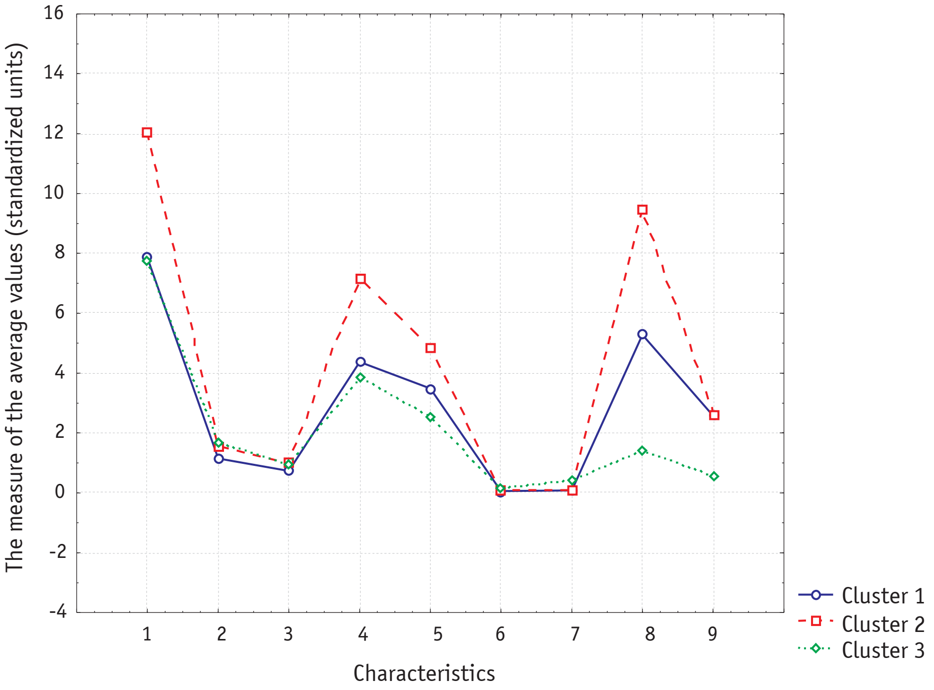

Figure 3. The average values for each cluster

In line graph (Figure 3) are shown the average value for each cluster which clearly differ from each other. This suggests about qualitative division of samples into groups. According to the graph distinctive parameters are the length of leaf blade (*Ll*), distance (cm) from the top of the leaf to its maximum width (*SDmxT*) and the location (cm) of width of leaf corresponding to the length of the petiole from the top of the leaf (*SLpT*).

To assess the degree of relationship between species was used value of the Euclidean distance (Table 1).

Table 1

The Euclidean distance between clusters

	Cluster I	Cluster II	Cluster III
Cluster I	0,00000	4,95265	2,28535
Cluster II	2,225455	0,0000	11,48029
Cluster III	1,511736	3,388257	0,00000

The Euclidean distance is the most common measure distances between objects, which is the geometric distance between objects in multidimensional space. In our case it is the distance between morphological parameters of leaf for each species; it is equivalent to the distance between species according to selected parameters. The more smaller the distance between objects, the more they are similar. The square Euclidean distance used when it is required to increase in order the value of the distances between each other very distant objects [7].

Since the Euclidean distances are greater than one, it can be confirmed that clusters are located at great distances from each other, so species and hybrid forms of willow, that form these clusters, have a low degree of relationship. The biggest is the distance between clusters II and III, meaning they are least similar.

As a result of analysis of variance (Table 2) were defined intergroup and intragroup variance, the difference between which the parameter shows the sample belongs to a cluster and clustering quality. Parameters with values of p > 0.05 can be removed from cluster-

Table 2

Results of analysis of variance

№	Parameters	Intergroup SS	df	Intragroup SS	df	Fisher's ratio test, F	Significance level, p
1	*Ll*	76,3648	2	64,23703	18	10,69917	0,000867
2	*Dmx*	8,1818	2	6,47554	18	1,64256	0,221192
3	*Lp*	0,4551	2	0,32265	18	7,11545	0,005284
4	*SDmxT*	40,5284	2	17,17839	18	21,23339	0,000018
5	*SDmxB*	17,1633	2	14,72036	18	10,49359	0,000953
6	*DmnT*	0,3531	2	0,25548	18	1,87137	0,182651
7	*DmnB*	0,4749	2	0,25742	18	16,6032	0,000082
8	*SLpT*	208,651	2	20,51737	18	91,52531	0,000000
9	*SLpB*	18,7042	2	7,88811	18	21,34073	0,000018

ing procedure. In our case *Dmx* and *DmnT* indicators will be removed.

The best specimens belonging to the cluster characterize parameters *SLpT*, *SDmxT* and *Ll* as they correspond to the biggest difference between inter- and intragroup variances, the worst (that correspond to the smallest difference of variances) – parameters *Dmx*, *DmnT* and *DmnB*. Characteristics F and p also characterize the contribution of the parameter in the division of samples into groups. Better clustering correspond to higher values of the first and the lower – of the other parameters. According to the table above best parameter corresponds to biggest the difference (F–d) [8].

The way to determine the nature of the clusters is to check the average values for each cluster and for each measurement. Table 3 shows the parameters of the analysis of leaf.

Table 3

Characteristics of groups (clusters) on morphological parameters of leaves

№	Parameters	Clusters					
		I		II		III	
		$\bar{x}$	y	$\bar{x}$	y	$\bar{x}$	y
1	*Ll*	7,86±0,11	1,00	12,03±0,08	0,77	7,74±1,08	3,01
2	*Dmx*	1,15±0,03	0,49	1,52±0,03	0,52	1,70±0,07	0,76
3	*Lp*	0,73±0,02	0,11	0,99±0,01	0,06	0,92±0,02	0,19
4	*SDmxT*	4,38±0,10	0,97	7,17±0,03	0,51	3,85±0,19	1,25
5	*SDmxB*	3,49±0,02	0,43	4,86±0,01	0,29	2,56±0,26	1,47
6	*DmnT*	0,06±0,01	0,02	0,10±0,01	0,08	0,17±0,02	0,19
7	*DmnB*	0,08±0,01	0,02	0,10±0,02	0,05	0,41±0,01	0,20
8	*SLpT*	5,28±0,16	1,18	9,45±0,24	1,36	1,42±0,03	0,50
9	*SLpB*	2,58±0,07	0,81	2,57±0,09	0,82	0,58±0,01	0,05

Based on the assumption that the most distinctive characteristics to identify willow plants are the length of the leaf blade (*Ll*), location (cm) of width of leaf corresponding to the length of the petiole from the top of the leaf (*SLpT*) and distance (cm) from the top of the leaf to its maximum width (*SDmxT*), determine the following: to the cluster I belong plants with length of the leaf blade 7,86 ± 0,11 cm, with parameters *SDmxT* = 4,38 ± 0,10 cm and *SLpT* = 5,28 ± 0,16 cm; for plants that belong to the cluster II, identical parameters have the following meanings: *Ll* = 12,03 ± 0,08 cm, *SDmxT* = 7,17 ± 0,03 cm, *SLpT* = 9,45 ± 0,24 cm, respectively; for plants, that put together the cluster III, parameters are: *Ll* = 7,74 ± 1,08 cm, *SDmxT* = 3,85 ± 0,19 cm, *SLpT* = 1,42 ± 0,03 cm respectively.

Examination of cluster analysis made it possible to compare similar types and hybrid forms of willow by morphological characteristics of leaf to determine significant differences between them. By applying the method of K-means was allocated groups of closely related samples to facilitate the identification of phenotypically similar species.

Conclusions. It in effectually to use cluster analysis for diagnostic species of the genus *Salix* L., to determine which sometimes can be quite difficult. As a result of cluster analysis collection samples divided into groups by morphological parameters of leaf based on affinity in terms of the Euclidean distance. Cluster analysis of complex of parameters of leaf allowed to detect diagnostically valuable quantitative characteristics by which it is possible to identify willow species using applications for PC. In addition, the results make it possible to assert the similarity of norm of reaction of the genetic apparatus for collection samples in regularity of detection of quantitative characteristics.

Bibliography

1. Тищенко В. Н. Использование кластерного анализа для идентификации и отбора высокопродуктивных генотипов озимой пшеницы на ранних этапах селекции / В. Н. Тищенко Н. М. Чекалин, М. Е. Зюков // Фактори експериментальної еволюції організмів : зб. наук. пр. – К. : Аграрна наука, 2004. – Т. 2. – С. 270–278.
2. Гашева Н. А. Классификационно-диагностическая шкала рода *Salix* как возможность мониторинговых и таксационных ЭВМ-тестирований / Н. А. Гашева // Вестн. Оренбург. ун-та. – 2006. – № 4. – С. 23–27.
3. Chao N. On the classification and distribution of the family Salicaceae / N. Chao, G. T. Gong J. Liu // Journal of Sichuan Forestry Science and Technology. – 1998. – No 19. – P. 9–20.
4. Гашева Н. А. Опыт применения дискриминантного анализа для различия фенотипически сходных видов ив / Н. А. Гашева // Вестник экологии, лесоведения и ландшафтоведения. – 2005. – № 6. – С. 123–130.
5. Леончик Е. Ю. Кластерный анализ: терминология, методы, задачи : конспект лекций / Е. Ю. Леончик, О. В. Савастру. – Одесса : Одесский нац. ун-т им. И. И. Мечникова, 2007. – 48 с.
6. Комп'ютерні методи в сільському господарстві та біології : навч. посіб. / О. М. Царенко, Ю. А. Злобін, В. Г. Скляр, С. М. Панченко. – Суми : Університетська книга, 2000. – 202 с.
7. Дронов С. В. Многомерный статистический анализ / С. В. Дронов. – Барнаул : Изд-во Алтайского гос. ун-та, 2003. – 213 с.
8. Буреева Н. Н. Многомерный статистический анализ с использованием ППП «STATISTICA» / Н. Н. Буреева. – Нижний Новгород, 2007. – 112 с.

References

1. Tishchenko, V. N., Chekalin, N. M., & Zyukov, M. Ye. (2004). Ispolzovanie klasternogo analiza dlya identifikatsii i otbora vysokoproduktivnykh genotipov ozimoy pshenitsy na rannikh etapakh selektsii [Using cluster analysis for the identification and selection of high-yield genotypes of winter wheat in the early stages of selection]. *Faktory eksperimentalnoy evolyutsii organizmiv – Factors of experimental evolution of organisms, 2*, 270–278 [in Russian].
2. Gasheva, N. A. (2006). Klassifikatsionno-diagnosticheskaya shkala roda *Salix* kak vozmozhnost monitoringovykh i taksatsionnykh EVM-testirovaniy [Classification and diagnostic scale of genus *Salix* as the possibility of monitoring and taxation of computer testing]. *Vestnik Orenburgskogo universiteta – Bulletin of the Orenburg University, 4*, 23–27 [in Russian].
3. Chao, N., Gong, G. T., & Liu, J. (1998). On the classification and distribution of the family Salicaceae. *Journal of Sichuan Forestry Science and Technology, 19*, 9–20.
4. Gasheva, N. A. (2005). Opyt primeneniya diskriminantnogo analiza dlya razlichiya fenotipicheski skhodnykh vidov iv [Experience of using discriminant analysis to differentiate phenotypically similar species willows]. *Vestnik ekologii, lesovedeniya i landshaftovedeniya – Journal of Ecology, Forest and Landscape, 6*, 123–130 [in Russian].
5. Leonchik, Ye. Yu., & Savastru, O. V. (2007). *Klasternyy analiz: terminologiya, metody, zadachi: konspekt lektsiy [Cluster analysis: terminology, methods, objectives: lecture notes]*. Odessa: Odesskiy nats. un-t im. I.I. Mechnikova [in Russian].
6. Tsarenko, O. M., Zlobin, Iu. A., Skliar, V. H., & Panchenko, S. M. (2000). *Kompiuterni metody v silskomu hospodarstvi ta biolohii [Computer methods in agriculture and biology]*. Sumy: Universytetska knyha [in Ukrainian].
7. Dronov, S. V. (2003). *Mnogomerniy statisticheskiy analiz [Multivariate statistical analysis]*. Barnaul: Izd-vo Altayskogo gos. un-ta [in Russian].
8. Bureeva, N. N. (2007). *Mnogomernyy statisticheskiy analiz s ispolzovaniem PPP «STATISTICA» [Multivariate statistical analysis using PPP «STATISTICA»]*. Nizhniy Novgorod [in Russian].

УДК 631.526.2:582.623:311.12

М. В. Роїк, В. В. Баликіна. Використання кластерного аналізу як методу класифікації представників роду *Salix* L. Сортовивчення та охорона прав на сорти рослин. – 2015. – № 1–2 (26–27). – С. 32–36.

Мета. За допомогою кластерного аналізу дослідити взаємозв'язки між представниками роду *Salix* L., сформувати групи видів і гібридних форм, близькоспоріднених між собою, ґрунтуючись на відмінностях морфологічних параметрів листків. **Методи.** Польовий, кластерного аналізу й деревоподібної графіки. **Результати.** Види верби згруповано за абсолютними показниками листка й виділено три групи кластерів. Оцінено ступінь спорідненості видів між собою за значенням евклідової відстані. Визначено відмінні показники листка: довжина листкової пластинки *(Ll)*, відстань (см) від верхівки листка до максимальної його ширини *(SDmxT)* та відстань від верхівки листка (см) до лінії ширини, що відповідає довжині черешка *(SLpT)*. **Висновки.** На прикладі колекції верби виявлено діагностично цінні кількісні показники листків, використання яких дає можливість ідентифікувати види й гібридні форми верби за допомогою прикладних програм для ПК.

УДК 631.526.2:582.623:311.12

Н. В. Роик, В. В. Балыкина. Использование кластерного анализа как метода классификации представителей рода *Salix* L. // Сортовивчення та охорона прав на сорти рослин. – 2015. – № 1–2 (26–27). – С. 32–36.

Цель. С помощью кластерного анализа исследовать взаимосвязи между представителями рода *Salix* L., сформировать группы видов и гибридных форм, близкородственных между собой, основываясь на различиях морфологических параметров листьев. **Методы.** Полевой, кластерного анализа и древовидной графики. **Результаты.** Виды ивы сгруппированы по абсолютным показателям листка и выделены три группы кластеров. Оценена степень родства видов между собой по значению евклидового расстояния. Определены отличительные показатели листка: длина листовой пластинки (*Ll*), расстояние (см) от верхушки листа до максимальной его ширины (*SDmxT*) и расстояние от верхушки листа до линии (см) ширины, соответствующей длине черешка (*SLpT*). **Выводы.** На примере коллекции ивы обнаружены диагностически ценные количественные показатели листьев, использование которых дает возможность идентифицировать виды и гибридные формы ивы с помощью прикладных программ для ПК.

Ключевые слова: кластерный анализ, иерархический метод, метод К-средних, идентификация видов, показатели листа.

Надійшла 7.05.2015