PLISKA

STUDIA MATHEMATICA BULGARICA

ПЛИСКА

БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

Pliska Stud. Math. Bulgar. **16** (2004), 279-290

PLISKA
STUDIA MATHEMATICA
BULGARICA

CLASSIFICATION OF CHENOPODIUM GENUS POPULATIONS AND SPECIES BASED ON CONTINUOUS AND CATEGORICAL VARIABLES

Yanka Tsvetanova, Neli Grozeva

The estimation of statistical distance between populations arises in many multivariate analysis techniques. Whereas distance measures for continuous data are well developed, those for mixed discrete and continuous data are less so because of the lack of a standard model for such data. Such mixture of variables arise frequently in the field of medicine, biometry, psychology, econometrics and only comparatively few models have been developed for evaluating distance between populations. The subject of our study were data in the field of botany. The aim of the presented investigation was to apply methods for analysis of dissimilarity between 44 populations of 13 species of Ghenopodium genus,presented by 15 variables - 10 continuous and 5 categorical. The previously developed by another authors distance measures between populations presented by mixed attributes turned out not appropriate for the available data of Chenopodium genus. For that reason a specific distance measures were applied. The matrices with distances between populations and species were used as input for Hierarchical Cluster Analysis to explore the taxonomic structure of the Chenopodium genus.

2000 *Mathematics Subject Classification:* 62P10, 62H30

Key words: Distance between populations based on continuous and discrete variables, Genus Chenopodium, Cluster analysis

1. Introduction

The effective use of classification methods requires an understanding of the properties of the forms and types of data collected as well as of the measures of association. Data form consists of two-way table of n individuals and p attributes (variables) and the type of attributes can be continuous or categorical (binary, nominal or ordinal).

One of the most popular methods for classification of a set of experimental units, presented by multiply attributes is cluster analysis. The estimation of dissimilarity between populations presented by categorical and continuous data is important step in classification methods and have been an object of many studies. Classification based on all available information on the individuals is much more trustworthy than that based on only continuous or discrete attributes.

Cluster analysis is a partitioning of a heterogeneous set of objects into homogeneous subsets using hierarchical or nonhierarchical methods. The objects for clustering might be individuals or populations. Agglomerative hierarchical clustering techniques use distance (proximity) matrices for finding groups of objects and are basically exploratory methods and could be used as first stage of the study of relationships between observed objects. The agglomerative hierarchical clustering methods start with an original dissimilarity (distance) matrix between all observed objects and fuses the two closest object in a cluster. Next, the cluster-individual dissimilarity between this new group and the remaining objects is calculated. This set of dissimilarities is added to the matrix of dissimilarities among the remaining units to form a new dissimilarity matrix that is one row and column smaller than the original. A new fusion procedure is carried out, and two or more groups are presented, group-group dissimilarities must be computed. The procedure ends when all of the objects are united in one group. The method used for calculating the group-object and group-group dissimilarity is called clustering strategy and various agglomerative clustering strategies have been proposed so far [7].

Genus Chenopodium L. is the largest one in the family Chenopodiaceae of the Bulgarian flora, comprising until now 17 species [8]. Its representatives are nutritive, ruderal and weed species. In the last decades of the 20th century profound studies of species from that genus were carried out by: Crawford [5]; Reynolds and Crawford [22]; Murin et all [19]; Pasnik [21] etc. In Bulgaria genus Chenopodium and the entire family Chenopodiaceae have never been a subject of special studies. A number of scientists think that in spite of the positive efforts genus Chenopodium has not been perfectly studied. The various species and populations of each species are highly variable. In many cases significant

experience and a great number of plants are required to take into consideration the entire diversity when determining them. Some similar and hard to be told apart species are often united in Chenopodium album agg. The use of statistical methods when solving taxonomic problems in highly variable taxones is common in botany. In genus Chenopodium extra difficulties in their implementation arise when the characteristic of their generative organs (flower, seed,fruit) are recorded. In order to differentiate the various species both quantitative and qualitative features during blossoming and fruit-yielding are equally important.

The aim of the presented study was to apply appropriate methods for analysis of dissimilarity between 44 populations of 13 species of Chenopodium genus, presented by 15 variables - 10 continuous and 5 categorical. For calculation of distances between populations and species specific measures have been applied. The matrices with distances between populations and species were used as input for Hierarchical Cluster Analysis.

2. Previous investigations concerning distances between populations

Mahalanobis squared distance [17] has become the standard measure of distance between two populations when all observed characteristics are quantitative

$\Delta^2 = (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)$,

where μ_i is the mean vector of the $i - th$ population and Σ^{-1} is the inverse of the covariance matrix. The distance measures between two populations when the observed characteristics are qualitative have been proposed by Bhattacharyya [4], Balakrishnan and Sanghvi [2] and Kurczinski [14] who attempt to create Mahalanobis like distance measure for qualitative data.

No universal guidelines have been yet given for evaluating the distance between populations with mixed qualitative and quantitative data. One approach would be to compute one distance between each pair of populations from the quantitative variables and a second distance from the qualitative variables, and then to combine these two distances as a weighted average. This option involve some element of subjectivity with possible loss of information and do not appear very satisfactory in general.

2.1. Mixture models

A joint distribution functions of categorical and continuous variables should be applied in evaluating distance between populations. Olkin and Tate [20] introduced a model where joint distribution of continuous and categorical variables is the marginal distribution of the categorical variables multiplied by conditional

distribution of continuous variables. This model is known as the location model (LM) and it has subsequently applied for discriminant analysis with mixed binary and continuous variables - Krzanowski [10] and more generally for discriminant analysis with mixture of categorical and continuous variables Krzanowski [11], [12]. In it the categorical variables are arranged in a contingency table where the table categories follow multinomial distribution and the continuous variables are assumed to follow a multivariate normal distribution. However the parameters of these multivariate normal distributions depend on their location in the contingency table of categorical variables. Lawrence and Krzanowski [15] proposed the Homogeneous conditional Gaussian mixture model which is based on the original location model of Olkin and Tate . The method combines all levels of the categorical variables into one multinomial variable with m multinomial levels (or cells).

Franko et al. [6] proposed a model where the means, variances and covariances depend not on the specific cell but rather on the corresponding subpopulation. This model is called Independent Mixture model. The difference between Independent model and Homogeneous model is that the vector of means and dispersion matrix of the IM are assumed to be equal for all multinomial cells within a subpopulation, where as for the HCM the vector of means and the dispersion matrix are assumed to be different in each multinomial cell within a subpopulation. A General mixed model for joint distribution density and a general distance measure for mixed nominal, ordinal and continuous data are developed by Leon and Carriere [16].

2.2. Distance between populations with mixed data

Krzanowski [13] was the first who consider the development of mixed data distances based on Matusita's distance [18] using Location model. Krusinska [9] proposed a weighted Mahalanobis distance for mixed data as a weighted sum of the Mahalanobis distances for continuous variables and Mahalanobis -type distance for discrete variables introduced by Kurczinski [14]. Another distance was obtained by Bar-Hen and Daudin [1]. More recently, Bedric et al. [3] derived a Mahalanobis distance for mixed ordinal and continuous data using grouped continuous model. Leon and Carriere [16] propose a generalized Mahalanobis-type distance measure for mixed data with nominal, ordinal and continuous variables.

To describe the common idea of estimating distance measures between populations with mixed data we will introduce some notations. Let c continuous or quantitative variables $Y^T = (Y_1, Y_2, \ldots, Y_c)$ and q discrete or qualitative variables $X^T = (X_1, X_2, \ldots, X_q)$ are measured on each individual and all individuals

are drawn from k populations. The q discrete variables are assumed to define a multinomial vector Z containing s possible states, and the probability of observing state m in population $\mathbf{P_i}$ is assumed to be π_{im}, $(i = 1, \ldots, k; m = 1, \ldots, s)$. Then the joint density of Z at a state m and the vector of $\mathbf{c}$ continuous variables Y is given by a product of the marginal and conditional densities:

$P[Z = m, Y] = P[Z = m].P[Y|Z = m]$. The conditional distribution of the continuous variables vector for the state m of Z is assumed to be multivariate normal with mean vector μ_{im} and dispersion matrix Σ_{im} in population $\mathbf{P_i}$

$(i = 1, \ldots, k; m = 1, \ldots, s)$, thus joint density of Z at the state m and the vector Y in the population $\mathbf{P_i}$ is the product $\pi_{im}.f_{im}(y)$, where $f_{im}(y)$ is the corresponding probability density. There are three special cases of interest in the model:

C1: the conditional dispersion matrix is constant for all states of Z in each population, that is $\Sigma_{im} = \Sigma_i$. $(i = 1, \ldots, k; m = 1, \ldots, s)$.

C2: the conditional dispersion matrix is constant for all populations in each state of Z, that is $\Sigma_{im} = \Sigma_m$. $(i = 1, \ldots, k; m = 1, \ldots, s)$.

C3: the conditional dispersion matrix is constant for all states of Z and all populations, that is $\Sigma_{im} = \Sigma$. $(i = 1, \ldots, k; m = 1, \ldots, s)$.

Distance measure between two populations $\mathbf{P_i}$ and $\mathbf{P_j}$ of Krzanowski [13] is the Matusita's distance [18] defined by $\Delta_{ij}^2 = 2(1 - \rho_{ij})$ where the affinity between populations $\mathbf{P_i}$ and $\mathbf{P_j}$ is given by

$$\rho_{ij} = \sum_{m=1}^{s} (\pi_{im}\pi_{jm})^{1/2} I_{ij}^{(m)} \tag{1}$$

where $I_{ij}^{(m)}$ is the affinity between $N(\mu_{im}, \Sigma_{im})$ and $N(\mu_{jm}, \Sigma_{jm})$. In case C3 $I_{ij}^{(m)} = \exp(-\frac{1}{8}\Delta^2)$, where Δ^2 is the squared Mahalanobis distance. The distance Δ_{ij} between populations $\mathbf{P_i}$ and $\mathbf{P_j}$ can now be obtained from ρ_{ij} by using the expression $\Delta_{ij} = \{2(1 - \rho_{ij})\}^{1/2}$.

The generalized Mahalanobis distance proposed by Leon and Carriere [16] use the Kullback-Leiber divergence to the general mixed data model and is given by the formula:

$$\begin{aligned} \Delta_{ij} &= \sum_{m=1}^{s} (\pi_{im} - \pi_{jm}) log(\pi_{im}/\pi_{jm}) \\ &+ \sum_{m=1}^{s} \{(\pi_{im} + \pi_{jm})/2\}(\mu_{im} - \mu_{jm})^T \Sigma^{-1} (\mu_{im} - \mu_{jm}) \end{aligned}$$

They discuss the asymptotic results regarding maximum likelihood estimation of this distance.

In practice we require to evaluate the distance between k groups of sample data. The simplest way to approach this is to treat the data in each group as a sample from the corresponding population $\mathbf{P_i}$ and to replace all parameter values in any Δ_{ij} by their sample estimates. Statistical packages for multivariate analysis of variances or canonical variate analysis are adaptable to produce the estimates of within-category means and variance matrices necessary for the distance calculations of the mixed models.

3. The data of the Chenopodium populations

For our study we had data for 44 populations from 13 species of genus Chenopodium: Chenopodium album L.; C. ambrossoides L.; C. bonus henricus L.; C. botrys L.; C. ficifolium Sm.; C. hybridum L.; C. multifidum L.; C. murale L.; C. opulifolium Schr. ex Koch. et Ziz.; C. polyspermum L.; C. rubrum L.; C. vigatum L.; C. vulvaria L.

Thirty specimens of each studied population have been gathered. The morphometric studies have been carried out in laboratory conditions. For each of the available 1320 individuals 21 attributes - 14 quantitative and 7 qualitative have been recorded. As for some of the attributes there was no variation in populations they were not included in distance estimation. Thus for our study were selected 15 of the attributes - 5 categorical and 10 continuous: number of flowers in a flower group;rate of perianth accretion; presence of dorsal keel on the perianth leaves;perianth tint;seed tint; petal leaf length; petal leaf width; flower diameter; raceme length; seed length; seed width; fruit length; fruit width; seed stickness; fruit stickness.

The previously presented distance measures between populations with mixed categorical and continuous data was not appropriate for the available data because of some peculiarities. They are:

1) All individuals of each population sample belong to the same level of the categorical variables. Thus in the above notations the probability π_{im} of observing state m of Z in population $\mathbf{P_i}$ is 1 or 0.

2) The observed categorical variables have too many levels - $X1(1-4), X2(1-5), X3(1-4), X4(1-7), X5(1-5)$. The observed states of Z are too many and there are only few number of Z states with two species in them.This can be seen from the Tabl.1. Most of the states of the categorical variables are specific for only one species. The reason for this is that some of the species are presented with only one or two populations while other have more than 5 populations (Tabl.2).

Table 1: The distribution of the 13 species according to the states of the categorical attributes

Levels	X1	X2	X3	X4	X5
1	He	Vu, B, Al, Hy, Po	F, Am, Al, Mur, Po, Pu	Al, Pu, He	F,Al,Mur, Pu, Po
2	Al, B, Hy, Mul, Mur, Po, R	He, Po, Al	F, Am	F, He, B, Hy, Mul, Mur	F, Vi, Am, He, B, Hy, Pu
3	F, Al, Am, Mur, Pu, Po, Vu	F, Vi, He, Mur, R	Vi, He, B	F, Am, B, R	B,Vu
4	Vi, Al	Am, Pul, F, Mur	Hy, R, Vu	F	R
5		Mul	Mul	F	Mul
6				Po	
7				Vu, Vi	

Notations: Al - albim; Am - amrossoides, B - botris; F - ficifolium, He - henricus; Hy - hybridum; Mul - multifidum; Mur - murale; Po - Polyspermum; Pu - Opulifolium; R - rubrum; Vi - vigatum; Vu - vulvaria.
X1 - number of flowers in a flower group; X2 - rate of perianth accretion; X3 - presence of dorsal keel on perianth leaves; X4 - perianth tint; X5 - seed tint.

Table 2: The species of Chenopoduum genus and their corresponding populations

Name of the species	The numbers of populations belonging to the species
1. Ficifolium	f25,f31,f36,f54,f71
2. Vigatum	vi37,vi77,vi78
3. Album	al52,al60,al61,al62,al63,al64,al65,al66,al67,al68,al69,al70
4. Ambrosoides	am32,am33
5. Henricus	he27,he28
6. Botrys	b29,b30,b34,b35,b45,b53
7. Hybridum	hy38,hy80
8. Multifidum	mu39,mu40
9. Murale	mr41,mr75
10. Opulifoliun	pu17,pu42,pu55
11. Polyspernum	po44,po18
12. Rubrum	r43,r79
13. Vulvaria	vu3

3) There are different number of individuals in the species because of the mentioned above reason concerning the number of population in them.

These special features of the data were a reason to apply some specific measures for distance between the observed populations and then between species.

4. The proposed distance measures

4.1. Distance between populations

According to the above notations we assume that the distribution of continuous variables in the population $\mathbf{P_i}$ is multivariate normal with mean vector μ_i and a dispersion matrix Σ_i, $(i = 1, \ldots, k)$.

In the case when the dispersion matrix is constant for all populations, distance between populations $\mathbf{P_i}$ and $\mathbf{P_j}$ is supposed to be calculated by the following method. The distance measure is :

$$\Delta_{ij} = (1 - \frac{\sum_{l=1}^{q} \omega_l \delta_{ij}^{(l)}}{q+1})\{(\mu_i - \mu_j)^T \Sigma^{-1} (\mu_i - \mu_j)\}^{1/2} \tag{2}$$

where $\delta_{ij}^{(l)} = 1$ when the state of the corresponding qualitative variable X_l in the both populations is the same, otherwise $\delta_{ij}^{(l)} = 0$; ω_l is the weight of the corresponding categorical variable X_l, $(l = 1, \ldots, q)$. When all weights are equal to 1, then

$$\Delta_{ij} = (1 - \frac{q_c}{q+1})\{(\mu_i - \mu_j)^T \Sigma^{-1} (\mu_i - \mu_j)\}^{1/2} \tag{3}$$

where q_c is the number of the categorical variables for which both population have the same value.

When populations do not coincide for any of the categorical variables the squared Mahalanobis distance between population considering only qualitative data is taken for Δ_{ij}.

4.2. Distance between species

Let denote by S_I and S_J the two species with k_I and k_J corresponding number of populations in them. Distance between the two species is defined by the formula:

$$D_{IJ} = \frac{1}{k_I k_J} \sum_{i=1}^{k_I} \sum_{j=1}^{k_J} \Delta_{ij} \tag{4}$$

Thus D_{IJ} is the average distance between all pairs of populations - one from each of them.

5. Cluster analysis

The matrix with squared Mahalanobis distances between populations was obtained from Discriminant analysis procedure of STATISTICA 6.0 package and the matrices with proposed distances between populations and species were obtained as results from STATISTICA Visual Basic (SVB) programs composed additionally.

The distance matrices between populations and species obtained by the proposed formulas were used as input to the hierarchical cluster analysis with Ward's method. From a statistical perspective the Ward's method seems better than the other hierarchical clustering strategies. This is because it has an objective function to minimize the within group sum of variability therefore to maximize the among group variability; thus, it gives natural connection to the analysis of variances. Furthermore, the Ward's method is appropriate for multinormal data distribution. The dendrograms obtained by cluster analysis are shown in Fig.1 and Fig.2

The dendrogram of the populations (Fig.1) shows that the Chenopodium rubrum; C. murale; C. botrys; C. ambrossoides ; C. bonus henricus; C. vigatum populations, belonging to various sections, are correctly united in separate clusters. The representatives of section Chenopodium the Chenopodium opulifolium; C. ficifolium; C. vulvaria; C. hybridum; C. murale; C. polyspermum; C. album populations are also correctly grouped. The regrouping of C. album; C. ficifolium; C. murale and C. polyspermum populations is absolutely admissible since the important feature for distinguishing these 4 species is the correlation between the length and the width of the leaf blade. Vegetative features are not included in this study.

The Cluster analysis of the species (Fig.2) shows absolutely correct grouping of the various types of clusters depending on which section of genus Chenopodium they belong taxonomically.

6. Conclusions

The proposed formula for evaluating dissimilarity between populations and species could be applied for investigation of the proximity between Chenopodium and some other genus. The obtained dendrograms fro cluster analysis show that the proposed distance measures can be successfully used for solving taxonomical problems in complex studies of variable taxones. In the future work when the data for the species will be extended with information for more populations the distances of Krzanowski [13] and Leon and Carriere [16] could be applied to the data. It will be interesting comparison of the results from cluster analysis based

on that distances with those received by the proposed formulas for the distances between populations and species.

Fig1. Dendrogram of 44 populations from Chenopodium genus obtained from the proposed distances between populations through Ward's hierarchical cluster method.

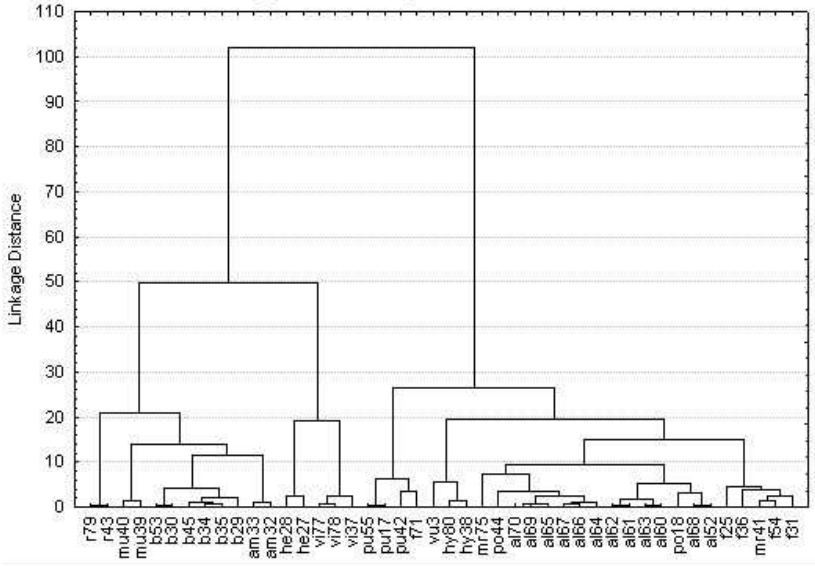
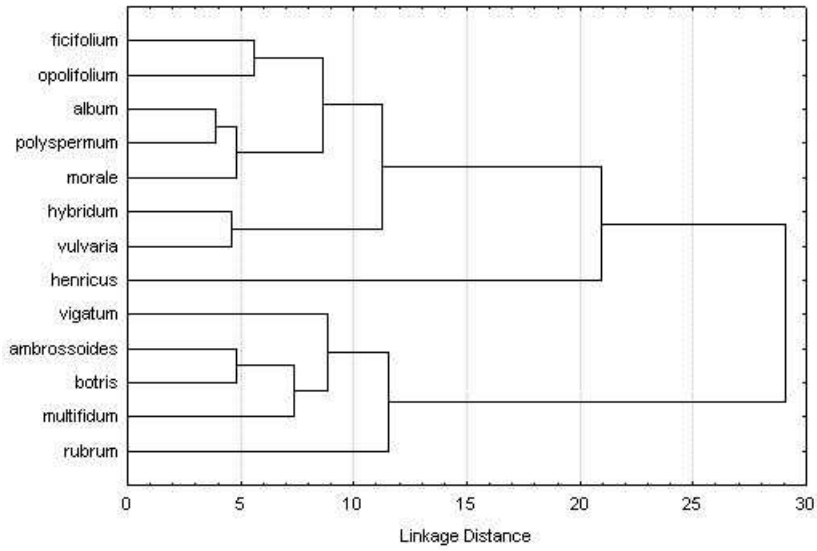


Fig.2. Dendrogram for the species of Chenopodium genus obtained from the matrix of proposed distance between species through Ward's hierarchical cluster method.



REFERENCES

[1] A. Bar-Hen, J. J. Daudin. Generalization of the Mahalanobis distance in the mixed case. *Journal of Multivariate Analysis* **53** (1995), 332–342.

[2] V. Balakrishnan, L. D. Sanghvi. Distance between populations on the basis of attribute data. *Biometrics* **24** (1968), 859–865.

[3] E. J. Bedrick, J. Lapidus, J. F. Powell. Estimating the Mahalanobis distance from mixed continuous and discrete data. *Biometrics* **56** (2000), 394–401.

[4] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Sankhya* **7** (1946), 401–406.

[5] D. J. Crawford. Syntematic relations on the narrow leaved species of Chenopodium of the Western United States. *Brittonia* **27(3)** (1975), 279–288.

[6] J. E. Franco, J. Crossa, J. Villaseor, S. Taba, S. A. Eberhart. Classifying genetic resources by categorical and continuous variables. *Crop Science* **38** (1998), 1688–1696.

[7] J. A. Hartigan. Clustering Algorithms, Wiley, New York, 1975.

[8] S. Kozuharov. Opredelitel na visshite rastenia v Bulgaria, Nauka i izkustvo, Sofia, 1992, 787.

[9] E. Krusińska. A valuation of state of object based on weighted Mahalanobis distance. *Pattern Recognition* **20** (1987), 413–418.

[10] W. J. Krzanovski. Discrimination and classification using both binary and continuous variables. *Journal of the American Statistical Association* **70** (1975), 782–790.

[11] W. J. Krzanowski. Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics* **36** (1980), 493–499.

[12] W. J. Krzanowski. Mixtures of continuous and categorical variables in discriminant analysis: a hypothesis testing approach. *Biometrics* **38** (1982), 991–1002.

[13] W. J. Krzanowski. Distance between populations using mixed continuous and categorical variable. *Biometrika* **70** (1983), 235–243.

[14] T. W. Kurczyǹski. Generalized distance and discrete variables. *Biometrics* **26** (1970), 525–534.

[15] C. J. Lawrence, W. J. Krzanowski. Mixture separation for mixed-mode data. *Statistics and Computing***6** (1996), 85–92.

[16] A. R. Leon, K. C. Carrière. Distance for Mixed Data. Available: http://www.stat.ualberta.ca/ brg/ms/GEN602.pdf (Accessed May 2003)

[17] P. C. Mahalanobis. On the generalized distance in statistics. *Proc. Nat. Inst. Sci.,India* **2** (1936), 49–55.

[18] K. Matusita. Classification based on distance in multivariate Gaussian cases. *Proc. 5th Berkeley Sump.* **1** (1967), 299–304.

[19] A. Murin, I. Haberova, C. Zamsran. Karyological studies of some species of the Mongolian flora, *Folia. Geobot. Phytotax* **15** (1980), 395–405.

[20] I. Olkin, R. F. Tate. Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics* **32** (1961), 448–465 (correction in **36**, 343–344).

[21] A. Pasnik. Notes of Chenopodium pedunculare and Chenopodium striatiforme / Chenopodiaceae/ in Poland: taxonomy and distribution, *Fragm. Flor. Geobot.* **44(1)** (1999), 63–70.

[22] J. Reynolds, D. J. Crawford. A quantitative study of variation in the Chenopodium atrovirens - desicatum - pratericola complex. *Amer. J. Bot.* **67(9)** (1980), 1380–1390.

Yanka Tsvetanova
Faculty of Agriculture
Trakia University, Stara Zagora, Bulgaria
e-mail: yanka@uni-sz.bg