Annals of Parasitology 2016, 62(4), 285–293
doi: 10.17420/ap6204.63

Review articles

The use of selected statistical methods and Kohonen networks in the revision and redescription of parasites

Daniel Zaborski[1], Katarzyna M. Kavetska[2], Wilhelm Grzesiak[1], Katarzyna Królaczyk[2], Emil Dzierzba[2]

[1]Laboratory of Biostatistics, West Pomeranian University of Technology, ul. Doktora Judyma 10, 71-466 Szczecin, Poland
[2]Laboratory of Biology and Ecology of Parasites, West Pomeranian University of Technology, ul. Doktora Judyma 20, 71-466 Szczecin, Poland

Corresponding Author: Katarzyna M. Kavetska; e-mail: katarzyna.kavetska@zut.edu.pl

ABSTRACT. Revisions and redescriptions of species and higher taxa have been known in parasitology since the first description of a parasite. Usually, they are based on standard morphometric methods or more modern genetic analysis. The former are not always sufficiently reliable, while the latter often require expensive equipment, pre-defined genetic markers, and appropriately prepared research material. They may be replaced by multivariate statistical methods, in particular discriminant analysis and cluster analysis, and Kohonen artificial neural networks included in data mining. This paper presents the examples of specific applications of these methods for the verification of the affinity of nematodes. The discriminant analysis showed that it was possible to statistically significantly discriminate individual nematode species, both for males and females, based on morphometric variables. This confirmed the previously assumed division of the species complex *Amidostomum acutum* into three distinct species. Similarly, hierarchical cluster analysis, used for the determination of coherent groups of nematode parasites, allowed the identification of relatively homogeneous clusters of nematode species depending on their circle of hosts, and groups of hosts.

Key words: revision, redescription, discriminant analysis, Kohonen networks, cluster analysis

Introduction

Revisions and redescriptions of species and higher taxa are frequently encountered research problems in parasitology. New reports bring new data, often conflicting with the existing knowledge. Many species have several or even ten or so synonyms. Their systematic positions change, and the existence of species complexes or cryptic species is often demonstrated.

Redescriptions and the resulting revisions usually use standard morphometric methods. However, some new approach to use these data would be proposing the model of body form, for example of leeches [1,2]. The new genetic methods are also an increasingly popular alternative but they do have serious drawbacks. They require expensive equipment and reagents, pre-defined genetic markers and appropriately prepared research material (i.e., parasites), which excludes permanent preparations from research, for example flatworms mounted in Canada balsam.

Multivariate statistical methods (in particular discriminant analysis, cluster analysis), as well as artificial neural networks, are a simple and inexpensive alternative in this type of parasitological research [3]. Their hardware requirements are low (a computer with appropriate software), and analysis may use both fresh and previously collected material (even many years earlier).

Methods

Neural networks

Artificial neural network is an information processing system modeled on the nervous systems of living organisms. Capable of parallel processing of large portions of information, it is resistant to data

errors and omissions [4]. Preparation of such a network is based on supervised or unsupervised learning. Supervised learning uses a set of training cases consisting of input data and desired output values [5]. In unsupervised learning, the training case consists of the input vector only, and the network creates an output structure corresponding to the ordering of the input set.

Kohonen networks are an example of networks using unsupervised learning [6]. They belong to the category of self-organizing maps (SOMs) which transform complex, multi-dimensional input signals into much simpler low-dimensional discretized representation of the input space of the training cases [7]. SOMs apply competitive learning, i.e. nodes compete for the right to respond to a subset of the input data. At any given time, only one output neuron (or only one neuron per group), is active (i.e. „on") at a time. The neuron that wins the competition is called a winning neuron [8].

In our research, the network used for the first time in the revision and redescription of the species complex *Amidostomum acutum* (Lundahl, 1848), a nematode commonly recorded in wild ducks, consisted of the input and output (radial) layers of neurons [3]. The number of neurons in the input layer corresponded to the number of input variables (morphometric traits of nematodes), while the output layer was organized in a square of 3×3 neurons. The neurons of the output layer were strictly ordered [4]. Each neuron of the input layer was connected to all of the output layer neurons. Associated with these connections were weight vectors (corresponding to the synaptic potentials in biological neurons), which are free parameters of the network and the adaptation of which is part of the learning process [9]. Weight vector dimensionality corresponded to the dimensionality of the input vector (13 morphometric variables for males and 18 for females), and their relative lengths were equal 1.

In general, learning in the Kohonen network can be divided into competition, cooperation, and adaptation [10]. In the competition step, the network compares the output values with the input vector according to the adopted discriminant function. Among the output neurons, only one particular neuron is selected; one that is most closely related to the input vector. When this winning neuron is established, the next step is to choose neurons in its neighborhood. Only the weights of these neurons are adjusted towards the input vectors, while the synaptic weights of neurons outside the neighborhood remain unchanged. As the winning neuron is best matched to the input vector in the sense of the Euclidean distance, this winner-take-all principle enables the approaching of the neuron weight vectors of the output layer to the input vectors.

The detailed learning algorithm involved the following steps:

1. In the first step of learning ($t = 0$) the components of the weight vector $\mathbf{w}_j(t)$ of the j-th neuron of the output layer had small random values $\{w_1, w_2, \ldots, w_m\}$ [11]:

$$\mathbf{w}_j(t) = [w_1, w_2, ..., w_m] \qquad (1),$$

where: m – the number of network inputs (morphometric traits).

2. Then, after feeding the network with input variables for a given training case (previously scaled to the range [0, 1]), the Euclidean distance between each neuron's weight vector and the input vector was determined. The neuron which produced the smallest distance became the winning neuron j^* [12]:

$$\mathbf{w}_{j^*}(t) = \mathbf{w}_j(t) \leftarrow \min_{j=1\ldots r}\left(\left\|\mathbf{x}(t) - \mathbf{w}_j(t)\right\|\right), \qquad (2)$$

where $\mathbf{w}_{j^*}(t)$ – weight vector of the winning neuron j^* in step t, $\mathbf{w}_j(t)$ – weight vector of the j-th neuron of the output layer in step t, $\mathbf{x}(t)$ – the input vector in step t, r – the number of neurons in the output layer ($r = 9$), $\|\cdot\|$ – the Euclidean norm.

3. The weights of the winning neuron and the neurons in its neighborhood were adjusted according to the formula [13]:

$$\mathbf{w}_j(t+1) = \mathbf{w}_j(t) + \eta(t)[\mathbf{x}(t) - \mathbf{w}_j(t)],\ 0 < \eta(t) < 1,$$
$$j \in N_g(t), \quad (3)$$

where: $\eta(t)$ is the learning coefficient in step t ensuring the convergence of the process, while $N_g(t)$ is the radius of the neighborhood in step t. The weights of other neurons (outside the neighborhood) did not change in this step of the algorithm [14].

4. The network was fed with another input (training) vector, and the Euclidean distance was determined between this vector and weight vectors of output layer neurons. The total number of training cases was 61 for males, 63 for females and 10 validation cases for both sexes (to monitor network error during the learning process). This procedure was repeated until all cases of the training

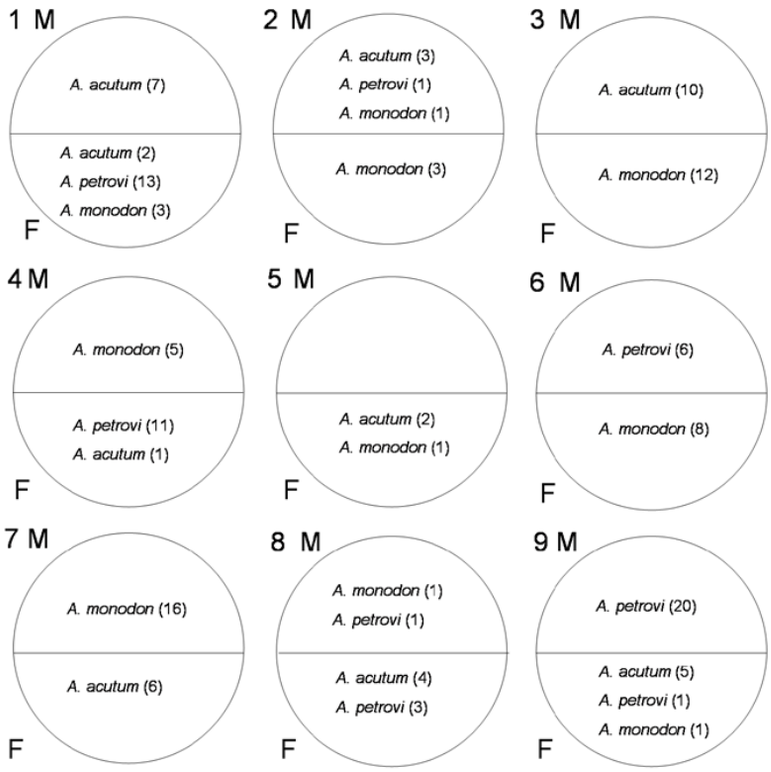

Fig. 1. A topological map with neuron numbers and the labels of *Amidostomoides* species for males (M) and females (F) (the number of cases activating a neuron is shown in brackets)

data set were given to the network (one training epoch) [15].

5. This step involved the reduction in learning rate $\eta(t)$ and the radius of the neighborhood $N_g(t)$. Usually, $\eta(t)$ decreases rapidly, e.g. during the first 100 epochs, from 0.9 to 0.1 (period of ordering). It is when that decrease may be linear. After this initial stage, $\eta(t)$ should be small (e.g. less than or equal to 0.01) for a long time. It is adopted that $\eta(t) \propto 1/t$ Similarly, $N_g(t)$ initially has relatively high values which gradually decrease during the step of ordering, e.g. down to 1. In the final step of learning, the radius of the neighborhood may be zero (only the winning neuron is activated) [16]. In this paper, these parameters decreased linearly from initial 0.3 and 1.0 to the final 0.01 and 0.

6. The last step consisted in moving to the next iteration ($t=t+1$) and repeating the steps 2 - 5 until the maximum number of iterations T ($t=T$) [14]. This number should be at least 500 times greater than the number of neurons in the output layer to achieve the desired level of statistical accuracy [11]. In the present study, the number of training epochs was 3500 for males and 5000 for females.

Compared to other classification methods, the use of Kohonen networks enables to retain topology; after learning, similar cases end up in one class or similar classes according to the definition of neighborhood applied during learning [17]. Input values are then replaced by a smaller number of coding vectors. In this way, the set of data is compressed. In this process, noise and outliers are removed, as the map contains only coding vectors and not the original input, and each of these vectors represents the sample of input data. Outliers or noise-containing data are mapped as a coding vector representing the cluster to which the data belong [9]. The results of this classification are presented on the topological map (Fig. 1).

The map shows that in the case of males, neurons 4 and 7 grouped most cases (91%) for the species *Amidostomoides monodon*, while neurons 6 and 9 grouped the majority of cases (93%) representing *Amidostomoides petrovi*. Other cases for these species were grouped into neurons 2 and 8. Cases for *Amidostomoides acutum* activated neurons located in the upper part of the topological maps, wherein approx. 85% of these observations were grouped into neurons 1 and 3. Neuron 5 did not represent any of the species. In turn, the topological map for females (Fig. 1) showed that neurons 2, 3 and 6 grouped the majority of cases (82%) of the

species *A. monodon* and neuron 7 was the winner for approx. 33% of the observations of the species *A. acutum*. The remaining cases were distributed between neurons 4, 5, 8 and 9. *A. petrovi* was represented by neurons located in the left and bottom parts of the topological map.

Discriminant analysis

In order to verify the results of clustering using Kohonen network, and select morphometric variables (traits) with the greatest discriminatory power, we used linear discriminant analysis (LDA). LDA is a method of supervised learning (categories of the grouping variable must be previously defined). It seeks projection axes on which observations from different classes are distanced from each other, while requiring that observations of the same class are close to each other [18].

In this study, LDA applied the linear canonical discriminant functions [19]:

$$D = w_0 + w_1x_1 + w_2x_2 + ... + w_mx_m, \quad (4)$$

where D is the value of the discriminant function, $x_1,..,x_m$ are discriminant variables, $w_1,..,w_m$ are coefficients for discriminant variables, w_0 – constant, m – number of discriminant coefficients (13 for males and 18 for females).

Canonical discriminant function coefficients were determined so as to maximize between-class distance and minimize the within-class distance. The data from the training set were presented in the form of a matrix, where n is the number of observations, m is the number of discriminant variables. We calculated within-class and between-class scatter matrix according to the following formula [20]:

$$\mathbf{S}_w = \sum_{i=1}^{c} \sum_{\mathbf{x}_i \in X_i} (\mathbf{x}_i - \overline{\mathbf{x}}_i)(\mathbf{x}_i - \overline{\mathbf{x}}_i)^T, \quad (5)$$

$$\mathbf{S}_b = \sum_{i=1}^{c} n_i (\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})^T, \quad (6)$$

where: $\mathbf{S}_W$ – within-class scatter matrix, $\mathbf{S}_b$ – between-class scatter matrix, $\boldsymbol{x}_i$ – column vector corresponding to the i-th row of the matrix $\boldsymbol{X}$, $\boldsymbol{X}_i$ – matrix consisting of matrix $\boldsymbol{X}$ rows corresponding to the class i, $\overline{\mathrm{x}}_i$ – the average for the class i, $\overline{\mathrm{x}}$ – the general average, c - the number of classes ($c = 3$), n_i – the number of observations in the class i, and T – transposition.

The following relationships occur [18]:

$$\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w \quad (7),$$

$$\mathbf{S}_t = \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T \quad (8),$$

where: $\mathbf{S}_t$ is the total-scatter matrix.

Class and general means were calculated according to the following formula [21]:

$$\overline{\mathbf{x}}_i = \frac{1}{n_i} \sum_{\mathbf{x}_i \in \mathbf{X}_i} \mathbf{x}_i, \quad (9)$$

$$\overline{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{c} \sum_{\mathbf{x}_i \in \mathbf{X}_i} \mathbf{x}_i. \quad (10)$$

With the invertible matrix $\mathbf{S}_W$, the first LDA vector was determined according to the optimization formula [22]:

$$\max_{w_1} \frac{\mathbf{w}_1^T \mathbf{S}_b \mathbf{w}_1}{\mathbf{w}_1^T \mathbf{S}_w \mathbf{w}_1}, \quad (11)$$

where $\boldsymbol{w}_1$ is the first vector, while the second LDA vector was determined to maximize the scatter between classes and at the same time minimize scatter within classes along all axes perpendicular to the first LDA vector, etc. The optimization of this ratio consisted in solving the eigenvalue problem [23]:

$$\mathbf{S}_b \mathbf{w}_i = \lambda_i \mathbf{S}_w \mathbf{w}_i, \quad (12)$$

where λ_i is the i-th eigenvalue.

When determining the contribution of each variable to the determination of the class we used Wilks' lambda coefficient Λ, according to the formula [24]:

$$\Lambda = \prod_{i=1}^{q} \frac{1}{1 + \lambda_i} \quad (13),$$

where q is the maximum number of discriminant functions. This statistic takes into account both the differences between the classes, as well as homogeneity within the groups. Because it takes into account the inverse values $[1/(1+\lambda)]$, variables with the lowest ratio have the highest discriminant power [25].

We also calculated the value of the partial Wilks' lambda, denoted as $\Lambda(u|\boldsymbol{x})$, which determines the change in the value of Wilks' lambda of the model $\Lambda(\boldsymbol{x})$ after the addition of the variable u to the subset $\boldsymbol{x}=[x_1,x_2,\ldots,x_m]$ [26]:

$$\Lambda(u \mid \mathbf{x}) = \frac{\Lambda(\mathbf{x} \mid u)}{\Lambda(\mathbf{x})} \quad (14).$$

This statistic was used directly to calculate the value of F statistic and the level of significance for each discriminant variable.

In addition, we determined tolerance coefficient T, a measure of redundancy in the model of a given discriminant variable [27]:

$$T = 1 - R^2 \qquad (15),$$

where R^2 is the coefficient of determination.

We also verified LDA assumptions: normality of variable distribution and equality of variance and covariance matrices for each class.

The last step was to determine the classification function coefficients used for assignment of objects (individuals) to the distinguished classes [28]:

$$V_i = c_i + w_{i1}x_1 + w_{i2}x_2 + ... + w_{im}x_m \qquad (16)$$

where: V_i – resultant classification value, i – the number of the class (i = 1,2,3), c_i – constant for the i-th class, w_{ij} – the weight of the j-th variable (morphometric trait) when calculating the classification for the i-th class, x_j – value observed for a given case for the j-th variable, $j=1,2, ..., m$, m – the number of discriminant variables (13 for males and 18 females).

The object was assigned to a class for which the value of the classification function was the biggest, establishing *a priori* probabilities proportional to the size of classes. The results of discriminant analysis showed a high capacity of discrimination between species of nematodes under the applied set of discriminant variables. Only one case was incorrectly classified in the case of males and 4 cases in females, giving an overall percentage of correct classification of 98.59% and 94.52%, respectively.

It should be noted that discrimination of nematode species was statistically significant (Λ=0.0205; F=26.11; p≤0.05 for males and Λ=0.0280; F=34.11; p≤0.05 for females). Fig. 2 presents a scatterplot of the canonical values for male and female nematodes, which confirms a good separation of the three species by the obtained discriminant functions. In the case of males, 6 of the 13 morphometric variables turned out to be statistically significant (p≤0.05). Among females, a significant contribution to the class discrimination was observed for 8 out of 18 analyzed morphometric variables (p≤0.05). Wilks' lambda coefficients (0.0206 – 0.0342 for males and 0.0281 – 0.0403 for females) and tolerance coefficients (0.5151 – 0.8151 for males and 0.1951 – 0.9077 for females) indicated relatively high discrimination power of those variables and their low redundancy.

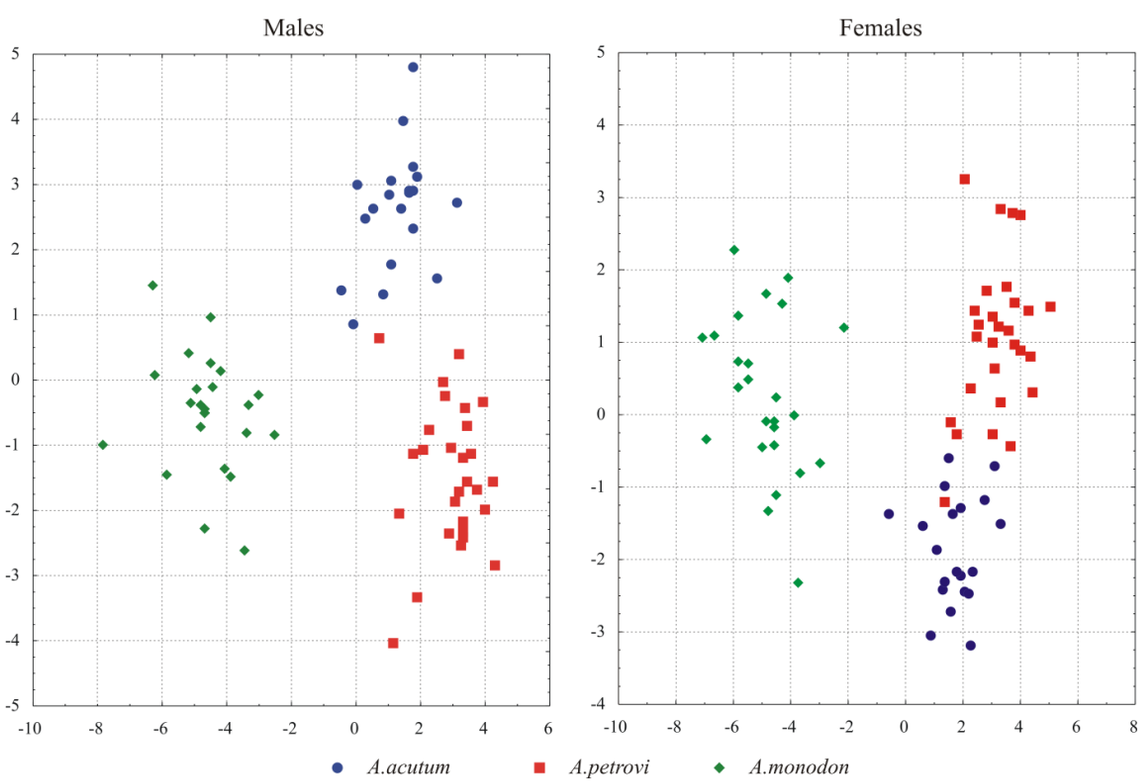

Fig. 2. A scatterplot of the canonical values for male and female nematodes [reproduced from: Kavetska KM et al. 2011. Revision of the species complex *Amidostomum acutum* (Lundahl, 1848) (Nematoda: Amidostomatidae). *Parasitology Research* 109: 105-117]

Cluster analysis

Cluster analysis is aimed at the discovery of natural groups (clusters) of objects existing in a multidimensional dataset. The groups should be as internally homogeneous as possible, and also as different from one another as possible [29]. In the analysis, all variables are treated as interdependent, and its purpose is to identify the structure of the tested set of variables or objects. Among the various methods used in cluster analysis, hierarchical agglomerative clustering is a general procedure, which sequentially connects pairs of most similar clusters in accordance with the arbitrary function called the measure of similarity, thus generating a nested set, also called a hierarchy [30]. The result of the agglomerative method can be presented as a dendrogram in which clusters of similar objects occur as separate branches. Selection of a suitable similarity measure indirectly influences the shape of the cluster, and thus plays a key role in the process of clustering. Although dependent on the problem being solved, this selection is usually confined to a small number of the most commonly used measures, such as the Euclidean distance or Pearson's correlation coefficient [31].

In our study, ordering of similarities was determined based on the cluster analysis using the Ward algorithm, with the Euclidean distance adopted as a measure of affinity [32]. In the Ward's method two clusters are connected with each other on the basis of the information loss criterion, the sum of squared errors (*SSE*). For each cluster *i*, its average (or centroid) and the sum of squared errors (SEE_i) are calculated. SSE_i is the sum of the squared deviations of each sample in a cluster from its average. For *k* clusters, there are *k* values of *SSE* (SSE_1, SSE_2, ..., SSE_k), the sum of which gives the total value of *SSE*:

$$SSE = \sum_{i=1}^{k} SSE_i \qquad (17).$$

For each pair of clusters *m* and *n*, we first calculated the mean (centroid) for the created *mn* cluster. Then we calculated the sum of squared errors for the *mn* cluster (SSE_{mn}) according to the formula (18):

$$SSE = SSE_1 + SSE_2 + ... + SSE_k - SSE_m - SSE_n + SEE_{mn}$$

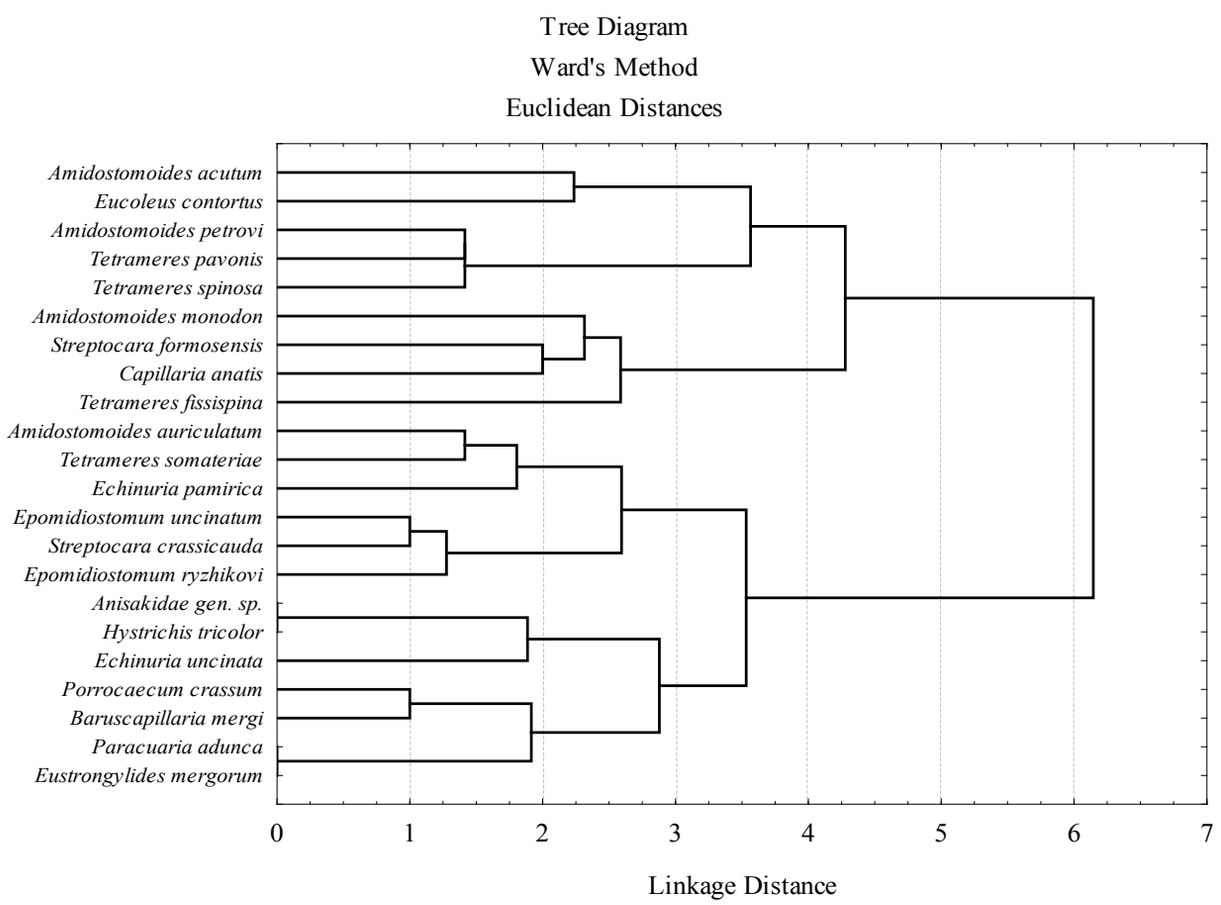

Fig. 3. Coexistence of nematode species taking into account host species

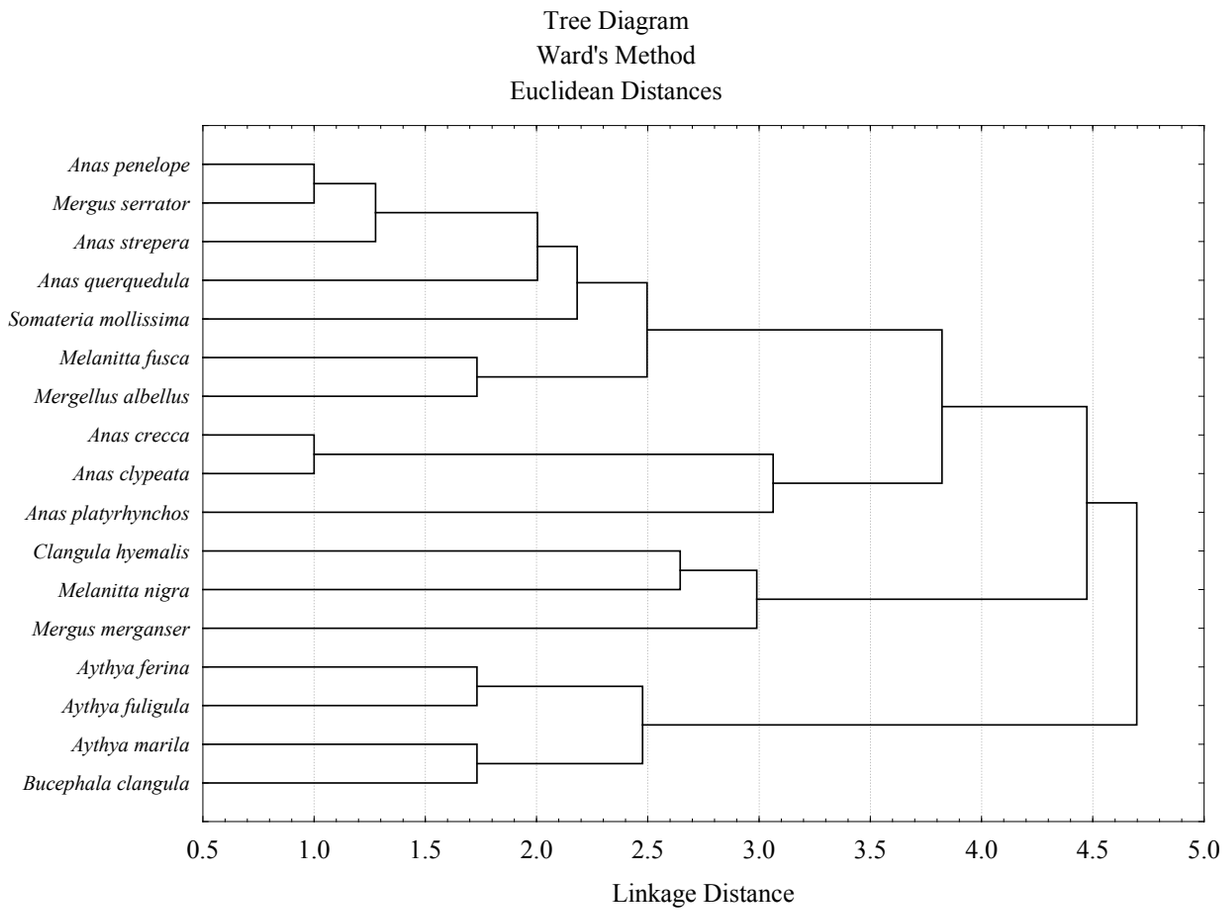

Fig. 4. Similarity of ducks in terms of their nematofauna

Clusters *m* and *n* which showed the smallest increase in *SSE* (the smallest information loss) were then combined [33].

The aim of cluster analysis in the present study was to identify internally coherent groups of nematode species depending on the circle of their hosts, as well as similar groups of host species depending on the nematodes living in them. The results of these analyses are shown in a dendrogram in Fig. 3 and 4.

The first dendrogram (Fig. 3) highlighted all the analyzed elementary coinvasive systems: *Amidostomoides acutum* and *Eucoleus contortus*; *A. petrovi*, *Tetrameres pavonis* and *T. spinosa*; *A. monodon, Streptocara formosensis* and *Capilaria anatis*; etc., which at the highest level form two main clusters. The first cluster (upper) consisted mainly of nematodes reported in a wide range of hosts (generalists). These included all polyxenic species and *T. pavonis* which matched this system because of its specificity regarding environmental hosts. It is also worth noting that each of the three above mentioned elementary systems was created with the participation of one of the three species of the genus *Amidostomoides*, dominant in each of the three tribes. The second cluster (bottom) consisted of species with narrow specificity (specialists), found most often in one or two species of ducks (including *Eustrongylides mergorum*, *Paracuaria adunca*, *Tetrameres somateriae*, *Epomidiostomum ryzhikovi*, *Hystrichis tricolor*). It is worth noting that the nematodes from the lower cluster were not observed in 9 out of 17 of the analyzed duck species, including none of the Aythyini tribe. At this stage, it is very difficult to form conclusions about a cluster containing nematode species found in single individuals.

Similarly, a diagram showing the similarities between host Anatinae species revealed two major groups of birds (Fig. 4). A clearer bottom cluster was formed by the Aythyini tribe individuals and very closely related *Bucephala clangula*. The cluster of nematofauna in this species was definitely closer to Aythyini, even though *B. clangula* belongs to Merginae. The group of these four species of ducks (*B. clangula*, *A. marila*, *A. fuligula* and *A.*

ferina) was clearly separated from all Anatinae. The upper cluster, containing as many as 13 species of ducks, was not as consolidated, and formed specific and hierarchically structured sub-sets. The first of these sub-sets (*Clangula hyemalis*, *Melanitta nigra* and *Mergus merganser*) was a relatively compact group of Merginae, with a relatively high numbers of individuals. A cluster with a similar distance was the one containing other two sub-sets: first of all a species-rich group within Anatinae (*Anas platyrhynchos*, *Anas clypeata* and *Anas crecca*) and a group of other Anatinae ducks and Merginae, less frequently studied. Perhaps further parasitological research, especially in the latter group, will allow to propose a clearer structure, showing more interrelationships between nematodes and between their hosts. However, it is most likely that it will correspond to the system of three tribes.

Conclusions

Morphometric traits of nematodes, used in the first part of the paper, significantly discriminated individual species, both for males and females. This confirmed the previously established division of the species complex *Amidostomum acutum* into three distinct species (*Amidostomoides acutum*, *A. petrovi* and *A. monodon*). Similarly, hierarchical cluster analysis, used in the second part of the study, allowed the isolation of relatively homogeneous clusters of nematode species depending of their circle of hosts, and groups of hosts depending on the nematodes living in them. Our analyses showed that the more traditional multivariate methods, i.e. linear discriminant analysis and hierarchical cluster analysis, and also a more innovative data mining method of the Kohonen artificial neural network, can be very useful tools in the revision and redescription of parasites, especially with limited access to alternative techniques such as sequence analysis.

References

[1] Bielecki A. 1997. Fish leeches of Poland in relation to the Palearctic piscicolines (Hirudinea: Piscicolidae: Piscicolinae). *Genus* 8: 223-375.

[2] Cichocka J.M., Bielecki A. 2015. Phylogenetic utility of the geometric model of the body form in leeches (Clitellata: Hirudinida). *Biologia* 70/8: 1078-1092. doi:10.1515/biolog-2015-0121

[3] Kavetska K.M., Królaczyk K., Stapf A., Grzesiak W., Kalisińska E., Pilarczyk B. 2011. Revision of the species complex *Amidostomum acutum* (Lundahl, 1848) (Nematoda: Amidostomatidae). *Parasitology Research* 109: 105-117.

[4] Tadeusiewicz R., Gąciarz T., Borowik B., Leper B. 2007. Discovering neural network properties using C# programs. PAU, Cracow.

[5] Higgs P.G., Attwood T.K. 2005. Bioinformatics and molecular evolution. Wiley-Blackwell, Oxford.

[6] Kohonen T. 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics* 43: 59-69.

[7] Larose D.T. 2005. Discovering knowledge in data: an introduction to data mining. John Wiley & Sons, Hoboken, N.J.

[8] Haykin S.S. 2009. Neural networks and learning machines. 3rd ed. Prentice Hall, New York.

[9] Samarasinghe S. 2007. Neural networks for applied sciences and engineering: from fundamentals to complex pattern recognition. Auerbach, Boca Raton, FL.

[10] Yang B.S., Han T., An J.L. 2004. ART–KOHONEN neural network for fault diagnosis of rotating machinery. *Mechanical Systems and Signal Processing* 18: 645-657.

[11] Emamian V., Kaveh M., Tewfik A.H. 2000. Robust clustering of acoustic emission signals using the Kohonen network. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 2000 (ICASSP'00).* 5-9 June 2000, Istanbul, Turkey. Vol. 6: 3891-3894.

[12] Wolkenstein M., Hutter H., Mittermayr C., Schiesser W., Grasserbauer M. 1997. Classification of SIMS images using a Kohonen network. *Analytical Chemistry* 69: 777-782.

[13] Hoffmann M., Várady L. 1998. Free-form surfaces for scattered data by neural networks. *Journal for Geometry and Graphics* 2: 1-6.

[14] Chon T.-S., Park Y. S., Moon K. H., Cha E. Y. 1996. Patternizing communities by using an artificial neural network. *Ecological Modelling* 90: 69-78.

[15] Chen Z., Feng T.J., Houkes Z. 1999. Texture segmentation based on wavelet and Kohonen network for remotely sensed images. In: *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, 1999 (IEEE SMC '99).* 12- 15 October 1999, Tokyo, Japan. Vol. 6: 816-821.

[16] Hernandez-Pajares M., Floris J. 1994. Classification of the Hipparcos input catalogue using the Kohonen network. *Monthly Notices of the Royal Astronomical Society* 268: 444-450.

[17] Cottrell M., Rousset P. 1997. The Kohonen algorithm: a powerful tool for analysing and representing multidimensional quantitative and qualitative data. In: *Biological and Artificial Computation: From Neuroscience to Technology: International Work-Conference on Artificial and Natural Neural Networks, IWANN'97.* 27 June 1997,

Lanzarote, Canary Islands, Spain. (Eds. J. Mira, R. Moreno-Diaz, J. Cabestany). Vol. 1240: 861-871.

[18] Cai D., He X., Han J. 2007. Semi-supervised discriminant analysis. In: *Proceedings of the IEEE 11th International Conference on Computer Vision, 2007 (ICCV 2007).* 14-20 October 2007, Rio de Janeiro, Brazil: 1-7.

[19] Chan H., Wei D., Helvie M., Sahiner B., Adler D., Goodsitt M., Petrick M. 1995. Computer-aided classification of mammographic masses and normal tissue: linear discriminant analysis in texture feature space. *Physics in Medicine and Biology* 40: 857-876.

[20] Chiang L.H., Russell E.L., Braatz R.D. 2000. Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis. *Chemometrics and Intelligent Laboratory Systems* 50: 243-252.

[21] Ye J., Janardan R., Li Q. 2005. Two-dimensional linear discriminant analysis. In: *Advances in Neural Information Processing Systems 17 (NIPS 2004),* December 2004. (Eds. L.K. Saul, Y. Weiss, L. Bottou). Vancouver, BC, Canada 17:1569-1576.

[22] Chiang L.H., Kotanchek M.E., Kordon A.K. 2004. Fault diagnosis based on Fisher discriminant analysis and support vector machines. *Computers and Chemical Engineering* 28: 1389-1401.

[23] Zhao W., Chellappa R., Phillips P.J. 1999. Subspace linear discriminant analysis for face recognition. University of Maryland, College Park, MD.

[24] Jung-Senssfelder K. 2007. Equity Financing and Covenants in Venture Capital: An Augmented Contracting Approach to Optimal German Contract Design. Springer Science & Business Media, Berlin.

[25] Klecka W.R. 1980. Discriminant analysis. SAGE, Beverly Hills, CA.

[26] Garcia H.C. 2008. A framework for the Self Reconfiguration of Automated Visual Inspection Systems. ProQuest, Ann Arbor, MI.

[27] Katsanos M. 2010. Intermarket Trading Strategies. John Wiley & Sons, Chichester.

[28] Hill T., Lewicki P. 2007. Statistics: methods and applications. StatSoft, Tulsa, OK.

[29] Jain A.K. 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* 31: 651-666.

[30] Marrelec G., Messé A., Bellec P. 2015. A Bayesian alternative to mutual information for the hierarchical clustering of dependent random variables. *PLOS One* 10: e0137278.

[31] D'haeseleer P. 2005. How does gene expression clustering work? *Nature Biotechnology* 23: 1499-1501.

[32] Kavetska K.M. 2006. Biological and ecological background of nematode fauna structure formation in the alimentary tracts of wild Anatinae ducks in north-western Poland. Dissertation, Agricultural University, Szczecin.

[33] Villwock R., Steiner M.T.A., Siqueira P.H. 2011. Pattern clustering using ants colony, Ward method and Kohonen maps. In: *ECTA and FCTA 2011 - Proceedings of the International Conference on Evolutionary Computation Theory and Applications and the Proceedings of the International Conference on Fuzzy Computation Theory and Applications [parts of the International Joint Conference on Computational Intelligence IJCCI 2011].* 24-26 October 2011, Paris, France. (Eds. A.C. Rosa, J. Kacprzyk, J. Filipe, A.D. Correia). SciTePress, Setubal, Portugal: 137-145.

Received 12 July 2016
Accepted 2 September 2016