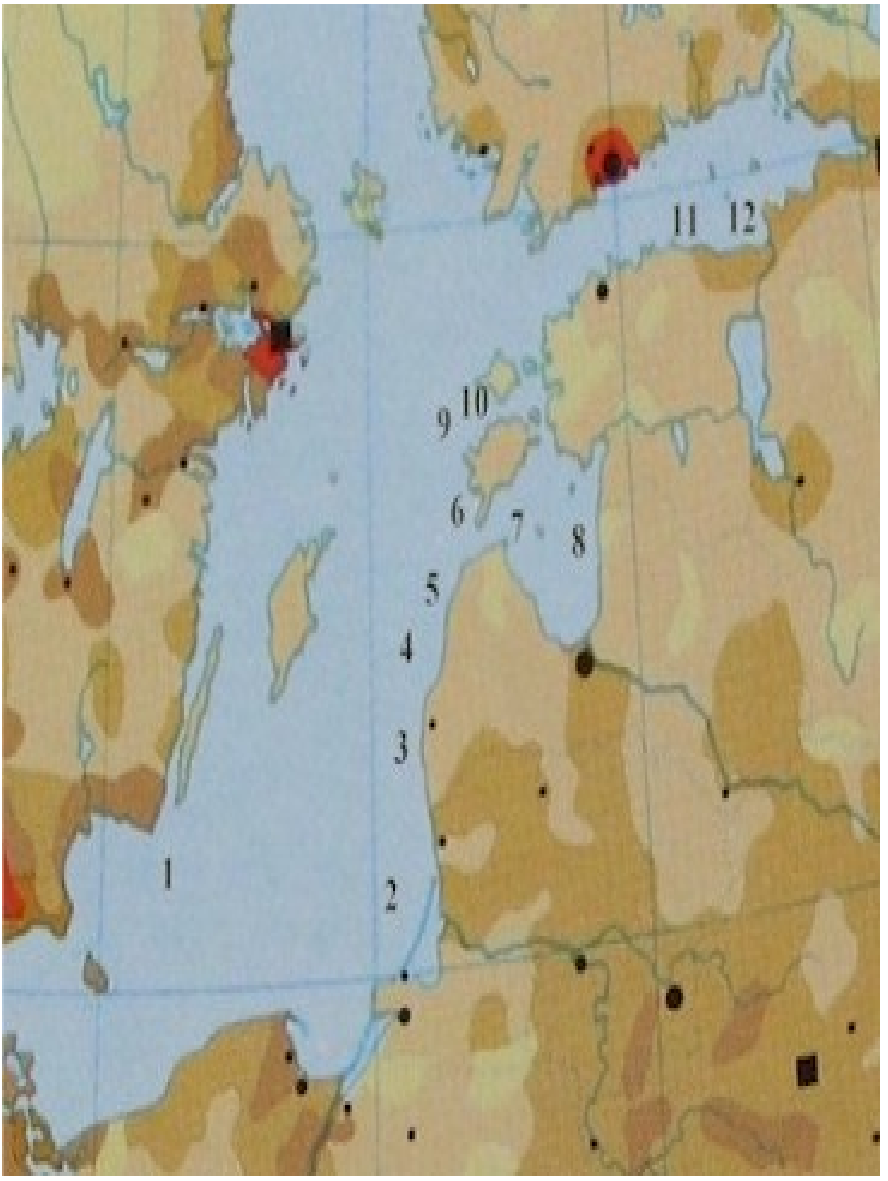
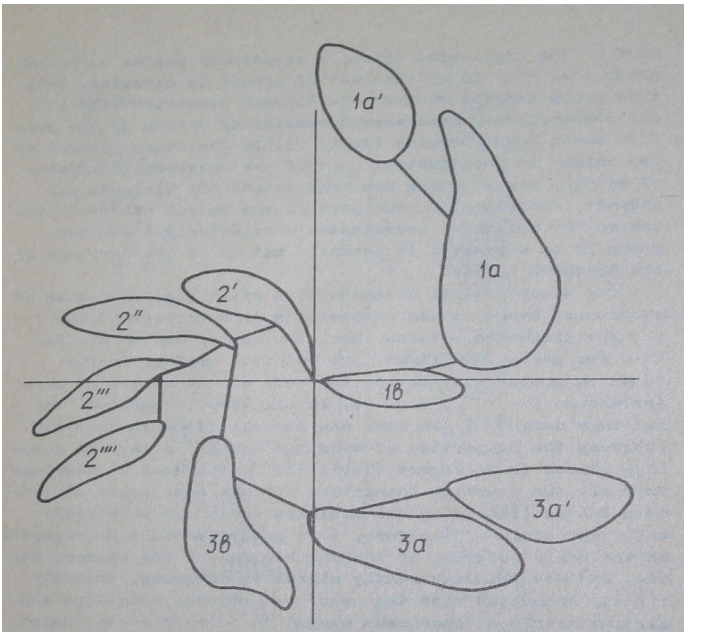

11
12
9
10
6
7
8
5
4
3
1
2

A Study of Infraspecific Groups of the Baltic East Coast Autumn Herring by new Method based on Cluster Analysis



Positions of the autumn herring subgroups differentiated by the method described.

Figure 1

E. Ojaveer, Estonian Laboratory of Marine Ichthyology (1975)

"In the Baltic Sea the autumn spawning herring forms a smaller number of groups than the spring herring does. This is probably connected with the different location of their spawning grounds. Spawning grounds of the spring herring are concentrated in favorable sites near the coast (in gulf, estuaries, etc.) while between such spawning centers gaps occur usually. Contrary to it, in most parts of the Baltic spawning places of the autumn herring form a continuous chain situated in the open sea. Therefore, differences in environment conditions between the autumn spawning grounds of neighboring areas are small and in large districts the characters of the autumn herring do not reveal essential differences. For instance, there is no significant difference between the autumns herrings caught on various grounds off the Polish coasts. The autumn herring of the Swedish Baltic coasts can be divided into four groups (that of the Gulf of Bothnia, that of the Bothnia Sea, the herring of the Swedish east coast and that of the Swedish south coast), between which a gradual transition occurs."

Appendix 1, J. Mullat (1975), Tallinn Technical University

While cluster is a concept in common usage, there is currently no consensus on its exact definition. There are many intuitive, often contradicting, ideas on the meaning of cluster. Consequently, it is difficult to develop exact mathematical formulation of the cluster separation task. Yet, several authors are of view that clustering techniques are already well established, suggesting that the focus should be on increasing the accuracy of data analysis. The available examples of data clustering tend to be rather badly structured, whereas application of the formal techniques on such data fails to yield results when the classification is known *a priori*. These issues are indicative of the fundamental deficiencies inherent in many numerical taxonomy techniques.

Following the standard nomenclature, a vector of measurements can describe every object $\langle x_1, x_2, ..., x_k \rangle$. Thus, for every pair of objects E_i and E_j a distance d_{ij} between those objects can be defined as

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{ik} - x_{jk})^2} \qquad (1)$$

However, it should be noted that all measurements are usually standardized beforehand.

Applying Eq. (1) on N objects yields a full matrix of distances

$$D = \begin{Vmatrix} 0 & d_{12} & d_{13} & . & . & d_{1k} \\ d_{21} & 0 & d_{23} & . & . & d_{2k} \\ . & . & . & . & . & . \\ d_{k1} & d_{k2} & . & . & . & d_{kk} \end{Vmatrix} \qquad (2)$$

Authors of many empirical studies employ methods utilizing the full matrix of distances as a means of identifying clusters on the set $\{E_1, ..., E_i, ..., E_k\}$.

In this section, we describe a new and highly effective clustering method, underpinned by some ideas offered by the graph theory. As the first step in our novel approach, we emphasize that, for elucidating the structure of the system of objects, knowledge of all elements of the matrix of distances given above is rarely needed. We further posit that, for every object, it is sufficient to consider no more than M of its nearest neighbors.

To explicate this strategy, let us consider a system of 9 objects (Fig. 2) with their interconnections — edges. The matrix of nearest neighbors for such a graph is given by:

$$\boldsymbol{MND} = \begin{Vmatrix} 5(1) & 6(1) & 3(2) & 0 & 0 & 0 \\ 4(1) & 3(2) & 7(3) & 0 & 0 & 0 \\ 4(1) & 5(1) & 1(2) & 2(2) & 0 & 0 \\ 2(1) & 3(1) & 5(1) & 7(3) & 0 & 0 \\ 1(1) & 3(1) & 4(1) & 6(1) & 7(3) & 0 \\ 1(1) & 5(1) & 7(3) & 0 & 0 & 0 \\ 2(3) & 4(3) & 5(3) & 6(3) & 8(3) & 9(3) \\ 7(3) & 9(3) & 0 & 0 & 0 & 0 \\ 7(3) & 8(3) & 0 & 0 & 0 & 0 \end{Vmatrix}$$

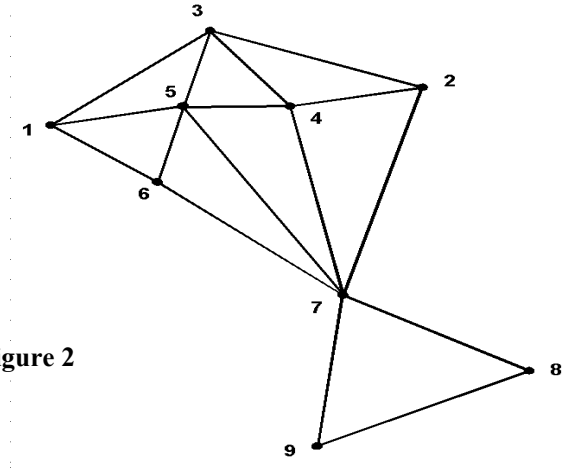


Figure 2

It can be easily verified that each row i of that matrix contains a list of objects j directly connected with a given object E_i, with the distances d_{ij} given in parentheses. Based on this argument, henceforth, we will denote the matrix of nearest neighbor distances by ***MND***.

In most cases, having data pertaining to about 8-10 nearest neighbors is sufficient. This is highly important for computation, where the goal is to minimize the required memory space. By applying this method on, e.g., the case of 1,000 objects, only 10,000 memory locations would be needed, which is a significant saving relative to the 500,000 required when the full matrix is processed.

We will use the ***MND*** defined above as a starting point to create some useful mathematical constructs.

Let W be the list of edges (pairs of objects) in the ***MND***. For every edge $e=[a,b]$, a subset W_b^a of the list W can be defined as follows.

Definition 1. Subset W_b^a of W represents a proximity space of edge $[a,b]$ if

a) for every pair of objects x and y, which are connected with at least one edge in W_b^a, there exists a path joining x and y, and

b) every edge that is a member of that path belongs to the subset W_b^a.

According to the graph theory postulates, proximity space is a sub-graph connected with the edge $[a,b]$.

Example. Let us consider the edge $[4,5]$ shown in Fig. 1. According to the aforementioned rules, its proximity space, denoted as W_5^4, is the sub-graph $W_5^4=\{[3,4],[3,5],[4,7],[5,7],[2,4],[1,5],[5,6],[4,5]\}$.

Definition 2. The system of proximity spaces is referred to, as the proximity structure if for each edge $w=[a,b]$ there exists a nonempty proximity space W_b^a in the system.

Sometimes it is useful to exclude the edge $[a,b]$ from the proximity space W_b^a. In line with the Venn diagram annotation, this exclusion is denoted as $W_b^a \setminus [a,b]$, whereby the resulting subset can be referred to as a reduced proximity space.

In the preceding discussion, for every edge $[a,b]$, only the value of the distance $d[a,b]$ between $[a,b]$ was taken into account. In what follows, it is useful to introduce a new notation. For example, it is beneficial to assign a real number (credential π), which is different from the distance, to every edge on the graph. For example, let us define the credential of every edge in the diagram shown in Fig. 1 as

$$\pi[x,y]=d[x,y]+r[x,y].$$

For example, $\pi[4,7]=3+2$, $\pi[7,8]=3+1$ on the edge $[x,y]$, where $d[x,y]$ is the Euclidean distance (1) between x,y and $r[x,y]$; $r[x,y]$ is the number of triangles that can be built around $[x,y]$.

Let us further assume that a proximity structure $\mathcal{L}$ of a graph W is known and that $f(x)$ is a real function.

Definition 3. The function $f_b^a(\pi)$ defined for all credentials of the edges in W_b^a is called the influence function of the proximity structure $\mathcal{L}$ if the following holds $f_a^b(\pi[x,y]) \le \pi[x,y]$ for each $[x,y] \in W_b^a \setminus [a,b]$, where $\pi[x,y]$ is the credential of the edge $[x,y]$.

In other words, for every edge $[x,y]$, we can find a new credential in the reduced proximity space $W_b^a \setminus [a,b]$

$$\pi'[x,y] = f_b^a(\pi[x,y]).$$

To demonstrate the benefit of introducing the influence function, let us again consider the diagram depicted in Fig. 1. Graphically, the influence function represents the value of the number of triangles after the elimination of the edge $[a,b] \in W_b^a$ from the list W_b^a. Using the set W_5^4 as an example, this corresponds to

$$f_5^4(\pi[3,4]) = f_5^4((d_{34} + r_{34}) = (1+1)) = (d_{34} + r_{34}) = (1+0) = 1;$$

$$f_5^4(\pi[3,4]) = f_5^4((d_{56} + r_{56}) = (1+0)) = (d_{34} + r_{34}) = (1+0) = 1;$$

$$f_5^4(\pi[3,4]) = f_5^4((d_{47} + r_{47}) = (3+1)) = (d_{34} + r_{34}) = (3+0) = 3.$$

$$\textbf{\textit{MNW}} = \begin{Vmatrix} 5(3) & 6(2) & 3(3) & 0 & 0 & 0 \\ 4(3) & 3(3) & 7(4) & 0 & 0 & 0 \\ 4(3) & 5(3) & 1(3) & 2(3) & 0 & 0 \\ 2(3) & 3(3) & 5(3) & 7(5) & 0 & 0 \\ 1(3) & 3(3) & 4(3) & 6(3) & 7(5) & 0 \\ 1(2) & 5(3) & 7(4) & 0 & 0 & 0 \\ 2(4) & 4(5) & 5(5) & 6(4) & 8(4) & 9(4) \\ 7(4) & 9(4) & 0 & 0 & 0 & 0 \\ 7(4) & 8(4) & 0 & 0 & 0 & 0 \end{Vmatrix}$$

It is evident that knowledge of the influence function of an edge allows us to easily find the set of new credentials for an entire subset $H \in W$. Let us consider the set $\overline{H} = W \setminus H$ and arrange its edges in some order $\langle e_1, e_2, ... \rangle$. Applying the steps shown above, we can find the proximity spaces of the edges in $\langle e_1, e_2, ... \rangle$ and apply Eq. (3) recursively.

Using the information delineated thus far, we can now introduce our algorithm, the aim of which is to identify the data structure.

At this point, we can assume that steps pertaining to the selection of the proximity structure and the influence function have been completed. Thus, we can proceed through the algorithm as follows:

A1. Find the edge with the minimum credential and store its value.
A2. Eliminate the edge from the list of all edges and compute the credentials for proximity spaces of the minimal edge using the recursive procedure (3).
A3. Traverse through the list of edges and identify the first edge with the credential less or equal to the stored credential. Return to **A2** to eliminate that edge. If no such edge exists, proceed to **A4**.
A4. Check whether there are any further edges in W. If yes, return to **A1**, otherwise terminate the calculations.

Performance of the algorithm will be demonstrated by applying the aforementioned steps to the graph shown in Fig. 1.

First, the credentials for all edges should be defined using the following expression:

$$\pi[x,y] = d[x,y] + r[x,y].$$

To do so, we must compute the matrix of credentials using the matrix of distances (2).

We will demonstrate all steps of the algorithm described above.

A1. Minimal edge is $[1,6]$ and the associated credential is $\pi[1,6] = 2$. To store its value, let $u = 2$.

A2. We eliminate the edge $[1,6]$ from the list W and therefore have to change the credentials of $\pi'[6,7] = 4$:

$$W_6^1 \setminus [1,6]:\ \pi'[1,3] = 3;\ \pi'[1,5] = 2;\ \pi'[5,6] = 2.$$

A3. Proceeding through the list, we encounter the edge $[1,5]$ as the first edge with the credential less or equal to u. Now, we return to step A2. After 9 steps with $u = 2$, we have the following sequence of edges:

$$\langle [1,6], [1,5], [1,3], [3,5], [3,4], [2,4], [2,3], [4,5], [5,6] \rangle.$$

Now, we consider the case $u = 3$, and after applying the preceding steps, we obtain $\langle [2,7], [4,7], [5,7], [6,7] \rangle$. Finally, using $u = 4$ yields $\langle [7,8], [7,9], [8,9] \rangle$.

It can be easily verified that those ordered lists of edges provide accurate representation of our graph's structure.

For graphical output, we can utilize the ordered edges to construct a connected tree (a tree is a graph without circles).

For the example given above, we can construct the tree using the ordered lists of edges, while excluding all edges $[a,b]$ if both their end points, a and b, are already members of the list. This approach results in the sequence

$$\langle [1,6],[1,5],[1,3],[3,4],[2,4],[2,7],[7,8],[7,9] \rangle$$

based on which the tree in Fig. 3 can be constructed.

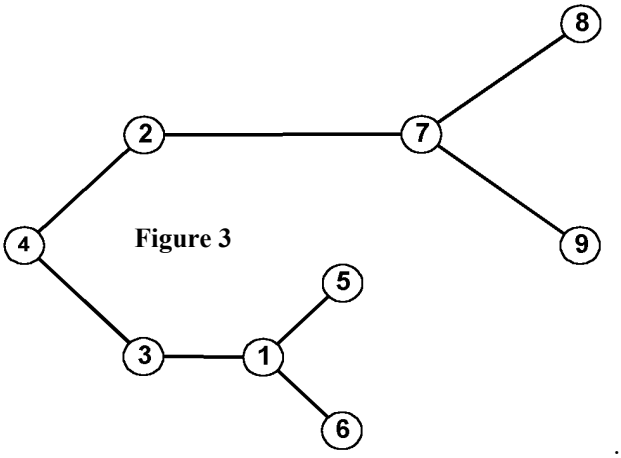


Figure 3

Using this simplified diagram, relative position of any object in the tree can be established by considering the number $S(x,y)$ of steps needed to reach the point y from the point x on the tree (e.g., $S(1,2)=3$, $S(1,8)=5$). Hence, for every object x, we can identify another object from which the maximum number of steps is required to reach x. For example, to identify the object at the top of the tree, we will take the object for which that maximum is minimum. Using real data, and applying these rules, we obtain the tree shown in Fig. 1.

LITERATURE

Ojaveer E., Mullat J.E. (Appendix I) and L.K. Võhandu (Appendix II). (1975)
A Study of Infraspecific Groups of the Baltic East Coast Autumn Herring
by two new Methods based on Cluster Analysis, Appendix I,
pp. 42-47 have been written by Mullat, main body by Ojaveer E.,
Estonian Laboratory of Marine Ichthyology. Estonian Contribution to the International Biological Program, VI, TARTU, pp. 22-50.