Ukrainian Journal of Ecology

*Ukrainian Journal of Ecology*, 2017, 7(2), 96–103, doi: 10.15421/2017_25

*ORIGINAL ARTICLE* *UDC 636.5:637.5:637.04*

# Cluster analysis, fuzzy sets, and fuzzy logic models in bird identification

**V.V. Osadchyi[1], V.S. Yeremeev[1], A.V. Matsyura[2]**

*[1]Bogdan Chmelnitskiy Melitopol State Pedagogical University, Melitopol, Ukraine, E-mail:* poliform55@gmail.com

*[2]Altai State University, Barnaul, Russia, E-mail:* amatsyura@gmail.com

*Submitted: 18.01.2017. Accepted: 25.04.2017*

In our resent research (Osadchiy at al., 2016) we considered the mathematical model for the identifying of bird species according to the results of inaccurate field measurements. We used the total length of the bird, the wingspan, the wingbeat frequency, and the flight as the input factors of the model. Testing the model on a hypothetical case of identifying some target species, like Rook, Common raven, Mallard, White Stork, and Lapwing revealed that this model can be used for bird species identification with definite limitations. However, in previous model we applied the recognition algorithm that was based on the classical sections of mathematical statistics. The limitations of those model are obvious - it does not take into account many characteristics and behavioral features of birds that cannot be represented in numerical form, like diurnal activity pattern and flocking behavior. In this case the possibility of using the traditional sections of mathematical statistics is quite limited. The present study is devoted to the development of a mathematical method for the identifying of the bird species that based on cluster analysis with fuzzy logic and fuzzy sets which extends the possibilities of the algorithm that was previously proposed in our research.

**Kew words**: bird species, identification, cluster analysis, fuzzy sets, fuzzy logic.

Bird identification is key point of field ornithological research. By now, the ornithologist is guided by his experience or field guide with information on the body mass, geometric dimensions and plumage color of the birds (Ryabitsev, 2001; Opredelitel', 2017). Thus, researchers from the California University of Technology and Cornell University developed an online service for identifying bird species of the United States and Canada by photographs (Identifikatsiya ptits, 2016).

Nowadays, the use of technical methods and computer data processing are extremely important for the analysis and processing of field bird observations (Bosak, 1990; Potapov, 1990b; Il'ichev et al., 1975; Ganya et al., 1991). The sound spectroscopy makes it possible to recognize the bird species and to study vocalization patters, including even the research of local dialects. Ornithologists at the Cornell Laboratory started to use sonograms to record songs of night birds and observe the exchange of information between them using alarms calls (Zhambyu, 1988).

In addition to scientific interest, ornithological research is of great practical importance. One of the problems relates to the risk of airstrike with the birds. At present, around 2.5-3 thousand of aircraft collisions with the birds are recorded annually in the world. Nowadays, every large airport has its own ornithological service that studies the migration routes of birds and conducts measures to control their abundance.

Technical means, automation of the observation process and application of mathematical methods for processing field data allow to perform ornithological research at more higher level. In last decade, the applied ornithological database was created by Osadchiy et al. (2015) for accumulation of the results of observations over various bird species in the southeastern region of Ukraine. Its content and data reliability depends on large number of factors - weather conditions, the technical limitations of filed surveys, terrain type and so on. Bird species determination is closely connected to the accuracy of using visual or technical means. The use of technical means allows expanding the scope of observation and more precisely organizing the bird counts. Modern technical facilities significantly expand the possibilities of obtaining more correct data, but they still have some limitations. In Osadchiy et al. (2016) the mathematical model for recognition of bird species was presented concerning the results of errors in field measurements. Some four parameters were considered as input factors of the model: the total length of the bird, the wingspan, the wingbeat frequency, and the flight speed. Testing the model on a hypothetical case of recognition

of rooks, crows, ducks, storks, and lapwings suggested that this model could be used at some extent. The model recognition algorithm was based on classic mathematical statistics while the limitations of this model are obvious - it does not consider many characteristics and behavioral features of birds that cannot be represented in numerical form. For example, one species has high activity in the morning time and others - in the afternoon, some species prefer flight in flocks whereas the others are not. In such situations, the possibility of using traditional sections of mathematical statistics is excluded. The present study is devoted to the development of a mathematical method for identifying the bird species by cluster analysis (Zhambyu, 1988) using fuzzy logic (Konysheva, Nazarov, 2011) and fuzzy sets (Eremeev, Baryshevskiy, 2011), which extends the possibilities of the algorithm proposed in (Osadchiy et al., 2016).

## Methods

The main filed characteristics of bird species are: geometric dimensions (length of the bird, wingspan, etc.); wingbeat frequency; flight speed; body weight. The accuracy of the measured parameters depends on the distance to the observer, the technical parameters of observation instruments, the terrain pattern and weather conditions, among which we selected: air transparency (rain, snow, nebula); wind speed and direction; atmospheric pressure; air humidity and temperature.
The mathematical model for recognizing the bird species observed in real time is represented in the form
$W = W(U_1, U_2, \dots U_n, V)$, (1)

where $W = 1,2 \dots i \dots m$ is the set of species, $U_1, U_2, \dots U_k \dots U_n$ - bird parameters measured during the observation, V - noise interference due to measurement errors.
Each bird species, W = i is characterized by a set of properties $U^{0i}_1, U^{0i}_2, \dots U^{0i}_k, \dots U^{0i}_n$, which we unite into the set $U^{0i} = U^{0i}\{U^{0i}_k\}$, where $i = 1,2 \dots m$.
The sets $U^{0i}$ for all the bird species, W, form the reference set $U^0 = U^0\{U^{0i}\}$. The measurement results $U_1, U_2, \dots U_n$ of the unknown species form set $U = U\{U_k\}$ k = 1, ..., n. The set of measurements, U in general does not coincide with any of the sets $U^{0i}$. Therefore, the question whether the observation results belong to one of the reference species should be considered from probability point of view.
Let's suppose that in the process of field observations, data on the parameters $U_1, U_2, \dots U_n$ of an unknown bird species were obtained and we need to select a species W from set of reference objects for which the reference values $U^{0i}_1, U^{0i}_2, \dots U^{0i}_n$ are in highest coincidence with measurements. Possible solution of a similar problem for the parameters $U_1, U_2, \dots U_n$, characterized by continuous random variables, was obtained in (Osadchiy et al., 2016). In our case this restriction is removed.
Some parameters can be set at a qualitative level like the bird species *i* is a nocturnal species while species *i+1* – diurnal species; one bird has black color, another - white, etc. Similar problems can be solved with the help of cluster analysis (Konysheva, Nazarov, 2011) using methods of fuzzy logic (Zhambyu, 1988).
**Conceptual tool for cluster analysis.**
Cluster analysis is widely used in the classification of information, depending on many factors (Hartigan, Wong, 1979). The initial data for solving the problem are the measurements $U_1, U_2, \dots U_n$ and the set of reference values of these parameters $U^{0i}_1, U^{0i}_2, \dots U^{0i}_k, \dots U^{0i}_n$, i = 1, 2, n.
This mathematical model is presented on Fig. 1 as "black box" with reference parameters, where the input factors are the results of measurements, and the output factor is the identified species of observed bird.

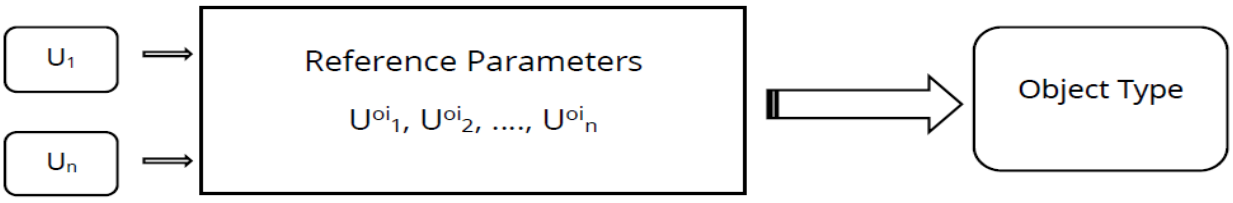


**Fig. 1**. Mathematical model of bird species identification.

The algorithm for solving the problem should ensure the allocation of a set i of the set of known bird species, which is most likely to correspond to the measured values for its reference indicators $U^{0i}_1, U^{0i}_2, \dots U^{0i}_k, \dots U^{0i}_n$. The set of reference parameters $U^{0i}_1, U^{0i}_2, \dots U^{0i}_k, \dots U^{0i}_n$, for each of the form W = i will be associated with the reference cluster under the number i. We assign a set of parameters $U = U\{U_k\}$ k = 1, ..., n to a cluster named "Measurements", characterizing the results of observations. The simplest formula for determining the distance between the *i*-th and *k*-th clusters is (Obzor, 2017):

$$r_{ij} = \sum_{k=1}^{k=n} f_k \, | U_k^{0i} - U_k^{0j} |, i, j = 1,2...m, \quad (2)$$

Where f*k* is the coefficient by which the significance of the individual parameters can be adjusted. The choice of the formula for the calculation of the measure depends on problem statement; to obtain the best result it is necessary to experiment with various formulas. The geometric distance in a multidimensional Euclidean space

$$r_{ij} = \{\sum_{k=1}^{k=n} f_k (U_k^{0i} - U_k^{0j})^2\}^{1/2}, i, j = 1,2...m, \quad (3)$$

or its square is often used:

$$r_{ij} = \sum_{k=1}^{k=n} f_k (U_k^{0i} - U_k^{0j})^2, i, j = 1,2...m, \quad (4)$$

The numerical order of parameters $U^{0i}_k$ of the same species can be significantly different. For example, the wingbeat frequency can be measured by 1-2 Hz, and the length of the bird is like 100-130 cm, which leads to a disproportionate contribution of the corresponding parameters to calculated value/distance (3) or (4). With the aim of leveling such a shortcoming, a relative difference was used in (Osadchiy et al., 2016):

$$(U_k^{0i} - U_k^{0j}) / U_k^{0j}, \quad (5)$$

The transformation (5) gives good results at a relatively high measurement correctness. We denote the variance of the measurement error $U_k^{0i}$ by $(\sigma^i_k)^2$. As the accuracy decreases, when the root-mean-square deviation $\sigma^i_k$ approaches the difference $U_k^{0i} - U_k^{0j}$, it is advisable to use the value

$$Z_k^{0ij} = (U_k^{0i} - U_k^{0j}) / \sigma_k^{ij}, \quad (6)$$

instead of the value (5), where $\sigma^{ij}_k$ is the root-mean-square deviation equal to the square root of the variance, $(\sigma_k^{ij})^2$ and determines the errors of the *k*-th parameter measurement for bird species with indices *i* and *j* in accordance with the formula:

$$(\sigma_k^{ij})^2 = (\sigma_k^i)^2 + (\sigma_k^j)^2, \quad (7)$$

## Results

In the present paper, we suggested to calculate the distance between clusters due to:

$$r_{ij} = \{\sum_{k=1}^{k=n} [f_k (Z_k^{0ij})^2]\}^{1/2}, i, j = 1,2...m, \quad (8)$$

instead of formula (3), where the relative difference between two values for the same parameter is determined by the formula (6). As was mentioned earlier, some parameters have discrete characteristics. This excludes the possibility of their estimation with the help of a continuous set of real numbers and, therefore, excludes the possibility of calculating distances by formulas (2, 3, 4, 8). In this case, we use fuzzy definitions. Let us consider three examples.

**The first example.**

Let one of the parameters of the observed species determine the degree of blackness of bird plumage. The values of the parameter "Black color" will be given by numbers from 0 to 2.86 (the second column on the left in Table 1), namely: the value "Absolutely black" is 2.86, the value "Rather black than other color" is 1.66, etc. The value of U from 0 to 2.86 in Table 1 should be considered as reference points for an approximate estimation of plumage color. The corresponding probabilities are given in the third column of Table 1: the value "Absolutely black" corresponds to 1, the value "Rather black than other color" - 0.90, etc. The numerical characteristics for the parameter U in Table 1, 2, and 3 were chosen so that the random variable (6) obeys a normal distribution with zero mathematical expectation and unit variance.

**Table 1**. Quantitative characteristics of plumage blackness in observed bird species.

| Parameter for bird plumage color | Value, *U* | Probability, *P(U)* |
|---|---|---|
| “Absolutely black” | 2.88 | ≈1.0 |
| “Rather black” | 1.66 | 0.90 |
| “Undefined” | 0.97 | 0.67 |
| “Rather non-black” | 0.46 | 0.35 |
| “Non-black” | 0.0 | 0.0 |

**The second example.**

Let one of the parameters characterized the daily activity of bird species towards three periods like: "Morning” (until 10.00), "Daytime” (from 10.00 to 15.00), "Evening” (from 15.00).

**Table 2**. Daytime bird activity

| Daily activity parameter | Value, *U* | Probability, *P(U)* |
|---|---|---|
| Before 10.00 AM | 2.88 | ≈1.0 |
| From 10.00 till 3.00 PM | 0.97 | 0.67 |
| From 3.00 PM | 0.0 | 0.0 |

**Example three.**

It is known that some bird species demonstrate flocking behavior. The individual, being in the flock, spends less time tracking the danger and more time for feeding. On the other hand, in this case part of the energy is spent on social conflicts (fights, demonstrative behavior) (Matematicheskie modeli, 2017). This could be presented by fuzzy logic and illustrated by Table 3.

**Table 3**. Estimation of flocking behavior tendency

| Spatial parameter | Value, *U* | Probability, *P(U)* |
|---|---|---|
| Flock pattern | 2.88 | ≈1.0 |
| Individual pattern | 0.0 | 0.0 |

If the relative difference (6) is used in the formula (4) instead of the difference $U_k^{0i} - U_k^{0j}$, then the square of the distance between the measures is determined by the formula:

$$R_{ij} = \sum_{k=1}^{k=n} [f_k (Z_k^{0ij})^2], i, j = 1,2...m, \quad (9)$$

In literature, a more general power-law measure of this distance is also used:

$$S_{ij} = \{\sum_{k=1}^{k=n} [f_k (Z_k^{0ij})^p]\}^{1/r}, i, j = 1,2...m, \quad (10)$$

For p = r = 2 expression (10) coincides with (9) and for p = 2 and r = 1 with the formula (8). It is not possible to give preference to any one expression from (8, 9, 10). It all depends on the properties of the clusters. The optimal choice of a metric is made on a specific material and requires additional research.

**Distances between reference clusters**

Reference data for some bird species are given in Table 4 and contain parameters determined by a continuous set of real numbers (the length of the bird, the wingspan, the wingbeat frequency, the flight speed), and the qualitative parameters (the plumage blackness, species activity at different daily times, flocking behavior tendency) in accordance with the data of Table 1 and 3. The first four parameters for rook, raven, duck mallard, stork and lapwing are taken from (Osadchiy et al., 2016), whereas the last three were suggested by experts. Let's consider the mallard duck as an example of plumage color estimation. According to experts, the degree of blackness regards "black-not black" scale rather could be accepted as "difficult to determine". Therefore, this value, U, considered to be 0.97 according to Table 1. The errors of the set value can be determined by the nearest upper and lower values of the blackness parameter from Table 1. The nearest upper value is 1.66. Adding half the difference between 1.66 and 0.97 to 0.97, we get the maximum estimated value equal to 1.31. The closest lower value for this parameter is 0.46. By subtracting from 0.97 half the difference between 0.97 and 0.46, we get the minimum estimated value equal to 0.72. All cells of Table 4 were filled in similar pattern. The parameters of individuals can vary considerably depending on the habitat and time of observation, therefore the authors do not pretend to exact values set in Table 4.

**Table 4.** The reference values $U_i$ for some bird species.

| Species | Bird species parameter | | | | | | |
|---|---|---|---|---|---|---|---|
| | Bird length, $U_1$, cm | Wingspan, $U_2$, cm | Wingbeat frequency, $U_3$, Hz | Flight speed, $U_4$, km/h | Plumage darkness, $U_5$ | Daily activity, $U_6$ | Spatial pattern, $U_7$ |
| Rook, $W_1$ | 60-70 | 130-140 | 3-4 | 50-60 | 2.27-2.88 | 0.48-1.90 | 1.43-2.86 |
| Raven, $W_2$ | 60-70 | 120-130 | 3-4 | 40-50 | 1.32-2.26 | 0.48-1.90 | 0.0-1.42 |
| Mallard, $W_3$ | 57-62 | 85-95 | 5-7 | 72-97 | 0.72-1.31 | 1.91-2.86 | 1.43-2.86 |
| Stork, $W_4$ | 100-115 | 155-165 | 1,5-2,5 | 35-45 | 0.0-0.23 | 0.0-0.47 | 0.0-1.42 |
| Lapwing, $W_5$ | 27-33 | 78-88 | 40-45 | 95-105 | 0.24-0.71 | 0.0-0.47 | 0.0-1.42 |

At the first stage, we calculated the distances between the reference clusters $r_{ij}$ in formula (8) with the coefficients $f_k = 1$. It was assumed that root-mean-square deviations $\sigma_k^i$ and $\sigma_k^j$ in formula (7) are equal to half the difference between the maximum and minimum values in the reference parameters of Table 4. The results of the calculations are presented in Table 5.

**Table 5**. Distances between the reference clusters $r_{ij}$ calculated by formula (8).

| | $W_1$,Rook | $W_2$,Raven | $W_3$,Mallard | $W_4$,Stork | $W_5$,Lapwing |
|---|---|---|---|---|---|
| $W_1$,Rook | 0 | 1.005 | 3.09 | 3.83 | 7.95 |
| $W_2$,Raven | 1.005 | 0 | 2.44 | 2.98 | 7.75 |
| $W_3$,Mallard | 3.09 | 2.44 | 0 | 25.15 | 6.12 |
| $W_4$,Stork | 3.83 | 2.98 | 5.01 | 0 | 9.13 |
| $W_5$,Lapwing | 7.95 | 7.75 | 6.12 | 9.13 | 0 |

The distance between clusters *i* and *j* is equal to the distance between clusters *j* and *i*, so Table 5 is symmetrical with respect to the diagonal elements.

The calculation of the diagonal elements themselves requires additional explanation. Formally, according to expression (6), the value $Z_k^{0ij}$ for i = j is 0. Therefore, the diagonal elements $r_{ii}$ calculated by formula (8) are also equal to 0, which is reflected in Table 5.

Such an approach for computing $r_{ii}$ is justified when the reference values of the parameters $U_k^{0i}$ are specified with absolute accuracy. In fact, each of these parameters is defined in a certain probability interval $[U_k^{0i\min}, U_k^{0i\max}]$.

Suppose there are several reference databases. We denote the values of the same parameter for species i in different databases as $U_k^{0i1}$ and $U_k^{0i2}$. Since $U_k^{0i1}$ and $U_k^{0i2}$ are within the interval $[U_k^{0i\min}, U_k^{0i\max}]$ their values differ by not more than $U_k^{0i\max} - U_k^{0i\min}$, i.e. by a quantity of the order $(U_k^{0i\max} - U_k^{0i\min})/2$ on average. Earlier it was indicated that the standard deviation $\sigma^i_k$, which characterizes the errors of parameter k for species i, was assumed to be equal to $\sigma_k^i = (U_k^{0i\max} - U_k^{0i\min})/2$.

Therefore, the value $Z_k^{0ij} = \left|(U_k^{0i1} - U_k^{0j2})\right| / \sigma_k^{ij}$ calculated regarding formula (7) is 1/√2. In this case, the diagonal elements $U_k^{0i1}$ and $U_k^{0i2}$ equal to the distance between and calculated by formula (8), will be equal to √3.5 ≈ 1.87. Replacing zero values by 1.87 in the diagonal elements of Table 5 we rewrite it (Table 6).

**Table 6**. Distances between the reference clusters rij, calculated by formula (8) regards the errors of the diagonal elements.

| | *W1*,Rook | *W2*,Raven | *W3*,Mallard | *W4*,Stork | *W5*,Lapwng |
|---|---|---|---|---|---|
| *W1*,Rook | 0 | 1.005 | 3.09 | **3.83** | 7.95 |
| *W2*,Raven | 1.005 | 1.87 | 2.44 | 2.98 | **7.75** |
| *W3*,Mallard | 3.09 | 2.44 | 1.87 | **5.01** | **6.12** |
| *W4*,Stork | **3.83** | 2.98 | **5.01** | 1.87 | **9.13** |
| *W5*,Lapwing | **7.95** | **7.75** | **6.12** | **9.13** | 1.87 |

All distances in Table 6 are calculated by the same formulas for the normalized values of the parameters (6). Since the diagonal elements of this table, which are equal to 1.87, characterize the errors of setting reference parameters, then all the values of distances $r_{ij}$ between clusters, less than 1.87, can be attributed to insignificant, and the corresponding individuals are considered indistinguishable. Since $r_{12}$ = 1.005 <1.87, then the identification of hooded crow from rook is impossible in considered measurement correctness.

The distances $r_{14}$ = 3.83, $r_{15}$ = 7.95, $r_{25}$ = 7.75, $r_{34}$ = 5.01, $r_{35}$ = 6.12, and $r_{45}$ = 9.13 are by several times larger than the diagonal element 1.87, so the pairs "rook" - "stork", "rook" - "lapwing" "raven" - "lapwing", "duck" - "stork", "duck" - "lapwing", "stork" - "lapwing" can be attributed to well-distinguishable. Distances $r_{23}$ = 2.44, $r_{24}$ = 2.98, $r_{13}$ = 3.09 are slightly larger than the diagonal element. Therefore, with some confidence, we can assume that the pairs "raven" - "duck", "raven" - "stork", and "rook" - "duck" are rather distinct than indistinguishable. To obtain more correct conclusions, the distances between the reference clusters $R^{ij}$ were calculated by the formula (9). The results of the calculations are given in Table 7.

**Table 7**. Distances between the reference clusters $R_{ij}$, calculated by the formula (9).

| | *W1*,Rook | *W2*,Raven | *W3*,Mallard | *W4*,Stork | *W5*,Lapwing |
|---|---|---|---|---|---|
| *W1*,Rook | 0 | 1.01 | **9.56** | **14.66** | **63.15** |
| *W2*,Raven | 1.01 | 0 | 5.93 | **8.88** | **60.13** |
| *W3*,Mallard | **9.56** | 5.93 | 0 | **25.15** | **37.51** |
| *W4*,Stork | **14.66** | **8.88** | **25.15** | 0 | **83.27** |
| *W5*,Lapwing | **63.15** | **60.13** | **37.51** | **83.27** | 0 |

The deviation of the reference parameter $U^{0j}_k$ from the reference parameter $U^{0i}_k$ is a random variable that obeys normal distribution with a variance approximately equal to $(\sigma_k^{ij})^2$. The normalized parameter (6) is also a random variable that obeys normal distribution with a mathematical expectation equal to zero and a variance equal to unity. Therefore, the sum of deviation squares (9) for two random species with indices i and j obeys the Pearson distribution $\chi^2$ (chi-square) with *k* degrees of freedom (Kremer, 2004). The density of this distribution could be:

$$\chi^2 = \sum_{i=1}^{i=k} Z_i^2 , \qquad (11)$$

Where $Z_i$ is a random variable that obeys the normal distribution law with zero mathematical expectation and unit variance. The density of the distribution $\chi^2$ is:

$$\varphi(x) = \frac{x^{\frac{k}{2}-1} e^{\frac{-x}{2}}}{2^{\frac{k}{2}} \Gamma(\frac{k}{2})}, x \geq 0$$

$$0, x \leq 0$$

Where $\Gamma(y) = \int_0^{\infty} e^{-t} t^{y-1} dt$ is the Euler gamma function, which is (y-1) for positive integers!

Since the sum of the distances between the clusters $R_{ij}$ obeys the distribution (11) in Table 7, then by using the Pearson criterion, we can test the $H_0$ hypothesis that the distances between two reference species are equal to zero. The critical values of $\chi^2_{cr}$ for k degrees of freedom are given in Table 8.

**Table 8**. Critical values of $\chi^2_{cr}$ for significance level q = 0.3 (Obzor, 2017).

| k | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|
| $\chi^2_{кр}$ | 3.66 | 4.88 | 6.06 | 7.23 | 8.38 |

The number of parameters k, in our case is equal to 7. According to Table 7 the critical value of χ2cr is equal to 8.38. The distances between clusters for all pairs of birds, except for the pairs "rook" - "duck" and "raven" - "duck" are much larger than the critical value of 8.38, so it can be argued that despite the errors of specifying the parameters of birds, most species can be assumed as distinguishable (Table 7).

We noted that the measure (9) is more sensitive in comparison to (8) when species are identified. So, the measure (9) in Table 7, in contrast to measure (8) in Table 6 allows us to distinguish clusters in pairs "rook" - "duck" and "crow" - "duck".

**Testing "Measurements" cluster vs one from reference clusters**

The results of measurements $U_1$, $U_2$, ... $U_n$ will be merged into a cluster called "Measurements". Let's provide it the number m+1. The distance between the "Measurement" cluster and any reference cluster with the number i in Euclidean space is determined by formulas (8) or (9):

$$r_{i,m+1} = \{\sum_{k=1}^{k=n} [f_k (Z_k^{0i,m+1})^2]\}^{1/2}, i = 1,2...m, \quad (12)$$

Or

$$R_{i,m+1} = \sum_{k=1}^{k=n} [f_k (Z_k^{0i,m+1})^2], i = 1,2...m, \quad (13)$$

Where $Z_k^{0i,m+1} = (U_k^{0i} - U_k^{m+1}) / \sigma_k^{i,m+1}$. (14)

The power function for calculating the distance can be written similarly to (10) in the form:

$$S_{ij} = \{\sum_{k=1}^{k=n} [f_k (Z_k^{0i,m+1})^p]\}^{1/r}, i, j = 1,2...m, \quad (15)$$

We can consider an algorithm for identifying an unknown species using the reference parameters presented in Table 4. Suppose we could obtain the data that based on observations of an unknown species and present them in Table 9.

**Table 9**. Parameter values for identification of an unknown bird species (cluster with number m=6 "Measurements").

| Species | Unknown bird species parameter | | | | | | |
|---|---|---|---|---|---|---|---|
| | Bird length, cm | Wing span, cm | Wingbeat frequency, Hz | Flight speed, km/'h | Plumage darkness | Daily activity | Distribution pattern |
| Unknown | 50-60 | 80-90 | 4-6 | 70-90 | 0.24-0.71 | 1.91-2.86 | 0.0-1.42 |

The distance $r_{i6}$ between the "Measurement" cluster and the reference clusters was computed using formula (12). It was assumed that the root-mean-square deviation $\sigma_k^6$ in formula (14) is half the difference between the maximum and minimum values for the cluster parameter "Measurements" (Table 9), and the standard deviation $\sigma_k^i$ in the formula is half the difference between the maximum and minimum values for the reference parameters in Table 4. The results of the calculations are presented in ranking series:

$r_{mallard}$= 0.86< $r_{raven6}$=2.64<$r_{rook6}$=3.42<$r_{stork6}$=4.89<$r_{lapwing6}$=5.82, (16)

it is clear from (16) that the duck has the greatest probability of identifying the observed object, followed by the raven, rook, stork, and lapwing (in descending order

In some cases, when one of the criteria is an order of magnitude more than alternative variants, such mathematical processing of data can fully satisfy the researcher. The very probability of identification remains unknown, although some preliminary conclusions can be obtained from the following considerations.

The root-mean-square deviations $\sigma_k^6$ and $\sigma_k^i$ characterize the variances of the *k*-th parameter for the "Measurements" cluster and the reference parameter of the *i*-th species. The calculated distances $r_{i6}$, formula (12) are equal to the square root of the sum of squares of the factor $Z_k^{0i,m+1}$, due to formula (14). The numerator $Z_k^{0i,m+1}$ is equal to the difference between the reference value of the reference cluster parameter $U_k^{0i}$ and the cluster parameter "Dimensions" $U_k^{m+1}$. The denominator is equal to the root-mean-square deviation $\sigma_k^{i,m+1}$, which characterizes the measurement error. For $Z_k^{0i,m+1} \leq 1$, the value $\sigma_k^{i,m+1}$ is not less than the difference $|U_k^{0i} - U_k^{m+1}|$, which indicates the coincidence of these parameters with a probability of at least 0.5-0.6.

Since the number of parameters in our case is 7, the coincidence of the clusters "Measurements" and "Ducks" has the probability of approximately equal to (0.5-0.6) / 7≈0.1. Similar reasoning allows concluding that there is a low probability of coincidence of the unknown bird with the stork and lapwing. The question of identifying an unknown species with a crow and rook requires additional analysis, which we could perform by formula (13). The results of the calculations are presented below in ranking series:

$R_{mallard} = 0.74 < R_{raven}6 = 6.99 < R_{rook}6 = 11.73 < R_{stork} = 23.89 < R_{lapwing}6 = 33.89$, (17).

The distances $R_{i6}$ in (17) obey the χ2 distribution. According to Table 8 the critical value of Pearson's criterion for seven degrees of freedom at a significance level of q = 0.3 is equal to 8.38. Therefore, the last three species - rook, stork, and lapwing (17) should be excluded from consideration towards identifying with an unknown bird. The question of the coincidence or difference in clusters "Dimensions" and "crows" remains open. The answer can be obtained by using more precise measuring devices.

**Impact of measurement errors on the validity of unknown bird identification.**

The relevance of unknown bird identification depends on two standard deviations. One of them determines the error in measuring the *k*-th parameter of an unknown bird $\sigma_k^{m+1}$, the other $\sigma_k^{0i}$ is the error of setting the *k*-th parameter for the reference cluster of the *i*-th species. In this paper, the calculation of distances between clusters was carried out using formulas (12) and (13) in a multidimensional Euclidean space. The values calculated from formula (13) allow us to determine the reliability of the conclusions using the statistical distribution χ2. Therefore, when studying the correctness of the identification of an unknown species, we use this formula.

In Table 10 we presented the results of calculations of χ2 for various dispersions $(\sigma_k^{i,m+1})^2$ for fixed values of the average reference parameters, given in Table 4, and average parameters of unknown species for the cluster "Measurements", Table 9. The second column in Table 10 corresponds to the values $(\sigma_k^{i,m+1})^2$ used in the construction of the ranking series (17). The first, third, and fourth columns are obtained for values $(\sigma_k^{i,m+1})^2$ multiplied by 0.5, 2.0, and 4.0, respectively.

**Table 10.** Influence of dispersion $(\sigma_k^{i,m+1})^2$ errors on the distance between the reference clusters and the cluster "Measurements", calculated by the formula (13).

| Species | $0.5(\sigma_k^{i6})^2$ | $1.0(\sigma_k^{i6})^2$ | $2.0(\sigma_k^{i6})^2$ | $4.0(\sigma_k^{i6})^2$ |
|---|---|---|---|---|
| Rook | **23,46** | **11,73** | 5,86 | 2,93 |
| Raven | **13,98** | 6,99 | 3.48 | 1,74 |
| Mallard | 1,48 | 0,74 | 0.37 | 0,18 |
| Stork | **47,78** | **23,89** | **11,94** | 5,98 |
| Lapwing | **67,78** | **33,89** | **16,95** | 8,7 |

The critical value of the Pearson criterion for the seven degrees of freedom with a significance level of q = 0.3 is 8.38. The decrease in the accuracy of measurements is even two times higher than the standard value corresponding to the second column in Table 10, so we can state that the clusters “Measurements” - “rook”, Measurements" - "stork", and "Measurements"- "lapwing” do not coincide. The coincidence between clusters "Measurements" and "Crows" can be considered using more precise methods of observation. Reducing the dispersion $(\sigma_k^{i6})^2$ by two times will increase the distance between these clusters to 13.98 and draw an appropriate conclusion.

## Conclusions

In this paper, we suggested the algorithm for identifying the bird species using cluster analysis. This method is based on the concept of "Cluster" (Konysheva, Nazarov, 2011), which in this case consists of a set of parameters that characterize a certain

bird species. As an example, the clusters related to the rook, raven, mallard duck, white stork, and lapwing were considered. The number of characteristics in a cluster is not limited. Numerical results of the cluster analysis are obtained for a set of seven parameters: the length of the bird, wingspan, wingbeat frequency, the flight speed, the plumage blackness, the activity of behavior at day time, and the tendency to fly in the flock. The distance in the multidimensional Euclidean space was chosen when performing calculations - formulas (8, 9, 12 and 13).

The distances between the reference species, calculated from formula (8), are given in Table 6. The calculated values allow us to determine the cluster differences in the pairs "rook" - "raven", "rook" - "duck", "rook" - "white stork", "rook" - "lapwing", "raven" - "mallard duck", "raven" - "white stork", "raven" - "lapwing", "mallard duck" - "white stork", "mallard duck" - "lapwing", "white stork" - "lapwing". The smallest distance of 1.005, obtained using the formula (8), refers to the pair "rook" - "raven" that indicates a slight difference in these species from the results of the observations presented in Table 4 and due to the systematic closeness of the species. The highest distance 9.13 was calculated for the pair "white stork" - "lapwing".

Application of formula (9) is more informative (see Table 7). In this case, the smallest distance 1.01 was also obtained for the pair "rook" - "raven", and the highest distance 83.27 was calculated for the pair "white stork" - "lapwing"

Current method approbation was carried out by cluster analysis towards the observation parameters of an unknown species with fuzzy data on the wingspan, wingbeat frequency, flight speed, plumage blackness, activity at diurnal time, and ability to fly in the flock. The Pearson criterion value obtained for the discussed study case testified that the unknown observed species could be a mallard duck with high degree of significance.

## References


Eremeev, V.S., Baryshevskiy, S.O. (2011). Graficheskiy metod resheniya zadach nechetkogo lineynogo programmirovaniya s chetko postavlennoy tsel'yu pri nechetkikh ogranicheniyakh. Geometrichne modelyuvannya i informatsiyni tekhnologiï proektuvannya: Tavrian State Agrotechnial Academy, 4(49), 27-32 (in Russian).

Ganya, I.M., Zubkov, N.I., Kotyatsy, M.I. (1991). Radiolokatsionnaya ornitologiya. Kishinev: Shtiintsa (in Russian).

Hartigan, J. A., Wong, M.A. (1979). Algorithm AS 136: A k-means clustering algorithm. Applied Statistics, 28(1), 100–108.

Identifikatsiya ptits po peniyu. Available from: http://muz4in.net/news/instrument_vtoroj_mirovoj_vojny_kotoryj_izmenil_nashi_sposoby_izuchenija_penija_ptic/2015-12-11-39847/ (Accessed on 15.05.2017).

Identifikatsiya ptits. Available from: https://nplus1.ru/news/2015/06/09/hitchcock-knowed-about-birds-better/ (Accessed on 15.05.2017).

Il'ichev, V.D., Vasil'ev, B.D., Zhantiev, R.D. (1975). Bioakustika. Moscow: Vysshaya shkola (in Russian).

Konysheva, L.K., Nazarov, D.M. (2011). Osnovy teorii nechetkikh mnozhestv. Saint Petersburg: BKhV-Peterburg Press (in Russian).

Kremer, N.Sh. (2004). Teoriya veroyatnostey i matematicheskaya statistika. Moscow: YuNITI- DANA (in Russian).

Kto i kak spasaet samolety ot ptits. Available from: http://www.yaplakal.com/forum3/topic1404151.html/ (Accessed on 15.05.2017).

Matematicheskie modeli staynogo povedeniya kulikov. [Elektronnyy resurs] Rezhim dostupa: http://dom-i-zveri.ru/povadki-ptic/matematicheskie-modeli-stajnogo-povedeniya-kulikov.html/ (Accessed on 15.05.2017)

Obzor algoritmov klasterizatsii dannykh. Available from: http://habrahabr.ru/post/101338/ (Accessed on 15.05.2017).

Opredelitel' ptits stran SNG. Available from: http://onbird.ru/opredelitel-ptic/p8/ (Accessed on 15.05.2017).

Osadchiy, V.V., Siokhin, V.D., Gorlov, P.I., Vasil'ev, V.M., Pechers'ki, P.I. (2015). Komp'yuterna programa "Web portal formuvannya informatsiynoï bazi danikh z migratsiï ptakhiv v Azovo-Chornomors'komu regioni Ukraïni". Ukrainian Patent 62480 from 12.11.2015 (in Ukrainian).

Osadchyi, V.V., Matsyura, A.V., Eremeev, V.S. (2016). Mathematical model of bird species identifying: implication of radar data processing. Biological Bulletin of Bogdan Chmelnitskiy Melitopol State Pedagogical University, 6(3), 463-471. Doi: http://dx.doi.org/10.15421/2016119

Potapov, E.R. (1990a). Uchet khishchnykh ptits v ravninnykh tundrakh (pp. 12-16). In Metody izucheniya i okhrany khishchnykh ptits. Metodicheskie rekomendatsii. E.P. Kryukova (Ed.). Tver': Oblastnaya tipografiya (in Russian).

Potapov, E.R. (1990b). Bioradiotelemetriya v izuchenii khishchnykh ptits: sredstva i vozmozhnosti (pp. 1138-164). In Metody izucheniya i okhrany khishchnykh ptits. Metodicheskie rekomendatsii. E.P. Kryukova (Ed.). Tver': Oblastnaya tipografiya (in Russian).

Ryabitsev, V.K. (2001). Ptitsy Urala, Priural'ya i Zapadnoy Sibiri. Spravochnik-opredelitel'. Ekaterinburg: Ural University Press (in Russian).

V N'yu-Yorke ubili 70 tysyach ptits dlya bezopasnykh poletov. Available from: http://comments.ua/world/571583-v-nyu-yorke-ubili-70-tisyach-ptits.html/ (Accessed on 15.05.2017).

Zhambyu, M. (1988). Ierarkhicheskiy klaster-analiz i sootvetstviya. Moscow: Finansy i statistika (in Russian).


*Citation:*
Osadchyi, V.V., Yeremeev, V.S., Matsyura, A.V. (2017). Cluster analysis, fuzzy sets, and fuzzy logic models in bird identification. *Ukrainian Journal of Ecology, 7*(2), 96–103.