Relations between composition of fishes and hydromorphological variables in a very large river

Stockholm Junior Water Prize – Hungarian Competition 2023

Author:

Benedek Jandó

Veres Pálné Gimnázium

Mentors:

Sándor Baranya

Vivien Füstös

Alexander Anatol Ermilov

Department of Hydraulic and Water Resources Engineering

Budapest University of Technology and Economics

Budapest, 2022.

CONTENTS

ABSTRACT

Today, in any field of science, we can observe interdisciplinary directions, which are born from the fusion of related disciplines. By combining scientific fields and at the same time comparing the methods of different fields, we can get answers to new questions that go beyond a single subject. Understanding the niche model of ecology gives a new dimension to the complex study of the composition of living communities. The habitat of each population is determined by biotic and abiotic factors. The examination of biotic variables is the task of ecology, while abiotic variables cannot be examined with ecological methods, as the scales often used for their evaluation are too robust for detailed analyses. The measurements carried out by hydrologists and hydraulic engineers can provide a much more accurate description of these abiotic variables, so by combining the two, we can discover new relationships. In this study, we assigned the data of the 20 most common fish species in the Hungarian section of the Danube River from 2004 to 2022 to the data of hydrological datasets and hydrodynamic simulation models, and looked for patterns among them using Machine Learning (ML). Among the nine abiotic factors used as independent variables in the analysis, the average depth velocity, water depth and bed material composition were the most decisive variables, which aligns with the results of previous research. In addition, with our Random Forest model, we were able to predict the number of individuals of the 20 most common fish species in the given conditions in the entire Hungarian section of the Danube. These estimates refer to optimal habitat for fish species according to abiotic variables. The model gives accurate values only in a narrow range, the so-called hydromorphological optimum, where our variables determine the abundance of fish. The results of the studies showed that in most cases biotic factors are more dominant than abiotic variables. In addition to the ML analysis, we showed the possibility of using the Danube fish faunistic database, which covers a large area and time, to investigate the relationships of the population (for example, the relationship between invasive and native species) using classical statistical methods. The results found here are in many cases consistent with the Random Forest model, but give reason to extend the model with additional independent variables in order to better understand the ecology of the Danube fish species.

Keywords: *Danube, hydromorphology, fish habitat, Machine Learning, modelling, Random Forest*

1 INTRODUCTION

The hydrography of Hungary is centred on the Danube, the main axis of the country's river network is formed by this river, and it is also organically connected to the country's groundwater system (Implementing the Water Framework Directive in Hungary, 2009). Consequently, it played a major role in the formation of the current inhomogeneous landscape structure of the country, since a significant part of the national landscape typology is made up of river-related landscape types, most of which are located along the Danube (Tózsa, 1998). In addition, as a central river, it has played an important role in trade, water management, residential and agricultural issues for centuries, as a result of which river regulation and flood control were established on the entire Hungarian section (Implementing the Water Framework Directive in Hungary, 2009). This significantly transformed the functioning of wetlands connected to the Danube (Farkas-Iványi & Trájer, 2015).

A total of 105 fish species occur in Hungary, of which about 45% (n=47) are non-native, this includes species that have been introduced in recent years and now form self-sustaining populations (Erős & Vörös, 2017). The majority of these species entered Hungarian waters directly (with human intervention), while in the case of others, it is difficult to determine the exact effect or combination of effects that contributed to their introduction (e.g., Ponto-Caspian gobies) (Tavares, *et al.,* 2020).

The primary reasons for their breeding were food production, recreational (fishing) and aesthetic (aquaristic) uses. However, in many cases, the ecological impact of alien fish species was not explored prior to introduction. Many of these species became invasive after introduction and threaten the native fish and aquatic life (Erős & Vörös, 2017).

Artificial intelligence (AI) has burst into science. Since it brought a lot of new opportunities in almost every field, it is not surprising that its use has quickly spread in ecology. Nowadays, practically all approaches to AI (neural networks, genetic algorithms, random forest systems, deep learning algorithms) are utilized in ecological research (Sylvain, *et al.,* 2019).

Machine Learning (ML) algorithms are generally capable of discovering patterns in data sets. For ecologists, this is extremely beneficial, because it supports the analysis of complex, non-linear data, which is often the case regarding ecological questions (Olden, *et al.,* 2008). The Random Forest (RF) method is a type of ML approach that is well suited for predicting variables in ecological applications (Bergström, *et al.*, 2011; Elmahdy, *et al.*, 2020; Liu, *et al.*, 2018), because compared to classical statistical methods, it can easily investigate complex, non-linear

relationships.

The version of the RF method we use, regression RF, is based on so-called regression trees (RT), which are types of decision trees (DT) built from numerical variables. These trees allow the estimation or clustering of a given variable with logical decisions along pre-calculated threshold values of the variables (Olden, *et al.,* 2008). Based on unique data sets, RTs follow the analogy of rational human decisions and, broken down into linear segments, help predict more complex questions. The downside is that they are not flexible at all. They work well on the dataset they were created on, but cannot be used to interpolate other datasets, because they cannot accommodate different thresholds or different relationships between categories. This is where Random Forest (RF) comes in, combining the efficiency of RTs and the flexibility of ML, although it cannot exceed the range of the dataset used for training. The essence of RF is to generate a number of RTs using randomly selected variables and thresholds, based on a given training data set. To predict the dependent variable(s) based on the new data sets, it runs the value of each independent variable in the new database over all previously constructed RTs. In the case of regression RF, the estimated value will be the average of the values voted by the trees. An RF is thus a set of hundreds or even thousands of RTs, each of which is used to produce new estimates, and is thus able to adapt to unknown cases in the new dataset (Olden, *et al.,* 2008).

2 MATERIALS AND METHODS

2.1 Danube fish faunistic database (DFFD)

2.1.1 General characterisation and sampling methods

The Danube fish faunistic database (DFFD) used for the analyses was compiled from surveys carried out by researchers of the Balaton Limnological Research Institute and the ELRN Centre for Ecological Research between 2004 and 2022 in connection with various projects. With 1668 records it covers the entire section of the Hungarian Danube. Sampling, although using standard methods, was not carried out at standard time intervals and the spatial distribution of the samples is not uniform.

Sampling was done using two methods between April and November. One was the shoreline electrofishing (SE), which was carried out both day and night. In this case, fish were taken from a boat drifting downstream directly from the riparian zone along the shore, most often a length of 500 meters on the main channel, but generally between 45 and 800 meters *(Figure 1a)*. The other sampling method was offshore, this is the electrified benthic frame trawl (EBFT) sampling, which was only carried out during the day. In this case, a fishing machine was used for sampling (most often the Hans-Grassl EL65 IIGI), which in most cases fished a 500 meters section, while in general 300-870 meters *(Figure 1b)* (Szalóky, *et al.,* 2014). After identification and measurements,

the sampled fish were returned to the water for both methods. Together, the two methods cover the river basin adequately, making them suitable for combined assessment of the fish fauna of rivers (Zajicek & Wolter, 2018).

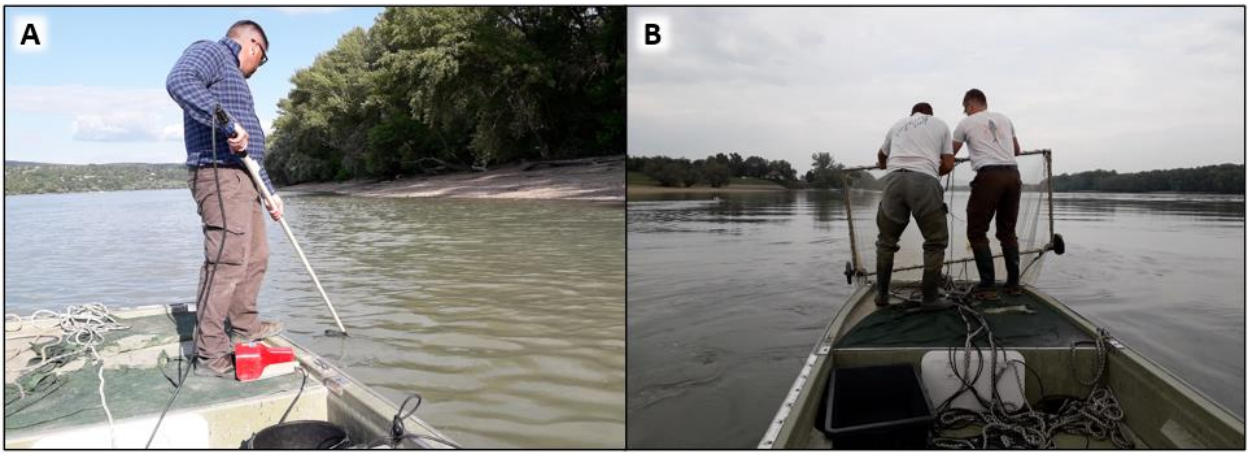

Figure 1. *Shoreline electrofishing (SE) (A) and Electric benthic frame trawl (EBFT) (B) in practice (source: V. Füstös)*

2.1.2 Methods of database cleaning

First, we deleted data that were inaccurate, incomplete or beyond our study area, i.e., not within the Hungarian section of the Danube. Samples were considered incomplete if it was not possible to determine where the sampling took place (in the main channel or a side arm) and if the date, time of day or method of sampling was not indicated. Based on the coordinates, if the distance between the start and end points of the sampling was greater than the possible length of the sampling, the whole data or at least one pair of coordinates was deleted. Redundant variables or variables irrelevant for further analysis (e.g., estimates of riparian vegetation) were deleted. Finally, the annotations were processed too to filter out additional records from the database.

The coordinate pairs used for the analyses were applied as follows: in cases where both coordinate pairs were available (70% of the samples), the average of the two was used to calculate the intersection point. In the case where only the starting point was available (30% of the samples), the starting coordinate pair was used. The projection of the coordinates used for the analyses was HD72/EOV (ESPG:23700).

2.1.3 Analysis of relationships between pairs of species

In addition to RF, we used classical statistical methods to analyse the DFFD, showing some examples of investigating population relationships through large-scale faunistic databases and compared the results with those of RF. For this analysis, we looked at the effect of one species on the number of individuals of the other species for two pairs, when both are present at a given sampling site.

In the case of the two species pairs, we have filtered the samples in which both species are present, because in cases where one species is absent, the ecological complexity of the problem makes it

difficult to determine whether this is due to the other species or to some other variable we have not measured. We then plotted the data pairs on a common scatter plot. We also fitted a reference function f(x)=x to each scatter plot, illustrating the hypothetical case where the number of individuals of the two species is in direct proportion to each other. Based on the pattern of the points and the side of the line on which they clustered, we set up the alternative hypothesis and calculated the Pearson correlation coefficient to evaluate the symmetric relationship between the two variables. Subsequently, we calculated the proportion of each species per record, by dividing the number of individuals of each species with the total number of individuals of the two species. The means of the resulting paired samples were compared using a two-sample paired t-test with one-sided counterhypothesis (in R, t.test(paired=T)). Since the pairs of individuals were in all cases dense at small values and thinned strongly in all directions as they moved towards larger values, their differences did not follow a normal distribution, but the empirical distribution of their differences was mostly uniform and had several peaks rather than skewed. However, due to the large samples (30<=n), we were able to apply the two-sample paired t-test approximately (Reiczigel, *et al.*, 2010).

2.2 Random Forest modelling

2.2.1 20 most common fish species

In the RF model, we used data of the 20 most abundant species, the number of species and the number of individuals in the samples as dependent variables, and 4 sampling conditions and 5 abiotic variables as independent variables. The 20 most abundant fish species were as follows: bleak *(Alburnus alburnus)*, white bream *(Blicca bjoerkna)*, common nase *(Chondrostoma nasus)*, schraetzer *(Gymnocephalus schraester)*, asp *(Leuciscus aspius)*, ide (*Leuciscus idus)*, burbot *(Lota lota)*, monkey goby *(Neogobius fluviatilis)*, round goby *(Neogobius melanostomus)*, European perch *(Perca fluviatilis)*, bighead goby *(Ponticola kessleri)*, Danube whitefin gudgeon *(Romanogobio vladykovi)*, common roach *(Rutilus rutilus)*, cactus roach (*Rutilus virgo)*, pikeperch *(Sander lucioperca)*, Volga pikeperch *(Sander volgensis)*, European chub (*Squalius cephalus)*, vimba bream *(Vimba vimba)*, Danube streber *(Zingel streber)*, zingel *(Zingel zingel).* The sampling condition variables were taken from the DFFD.

2.2.2 Abiotic variables in the model

Of the five abiotic variables, one, water depth, was derived from the DFFD, the others were assigned to the data using two-dimensional (2D) hydrodynamical models or from hydrological databases. The variables derived from hydrological datasets are water stage and flow regime, which have taken one of the values of small-medium-large or flooding-receding-stagnant. The flow velocity (m/s) was estimated for each sampling point using the water stage, flow regime and the slope for that section based on retrospective measurements from measuring stations

placed along the Danube.

The bed material composition of the sampling point was derived from a 2D model using the methods and models of the following papers: (Baranya, *et al.*, 2018; Füstös, *et al.*, 2019; Füstös, *et al.*, 2021).

The essence of the local bed material estimation is the bed shear stress calculated based on the model results for each grid point of a grid based on the water discharge, from which the bed material can be inferred. The water discharge in each area is obtained by a maximum likelihood estimation method, which, because of the relationship between the two, allows a value for the shear stress to be estimated probabilistically. And based on previously taken bed material samples and measurements, bed material types (grouped by grain size) can be assigned to the estimated shear stress at each grid point, yielding a very fine-scale mapping of the bed material composition.

2.2.3 Model parameters

In this paper, a regression RF model was used, which was trained on 80% of the available data, while the remaining 20% was used for validation. The RF model contained 150 trees and could get down a maximum of five levels.

2.2.4 Testing the model

The accuracy and effectiveness of the model was tested on data from a validating dataset of a fish species. Testing was performed by comparing selected samples with a randomly thinned 2D grid query. The median of the estimated values within a maximum distance of 150 m from each sampling site was compared to the measured values.

3 RESULTS AND EVALUATION

3.1 RF analysis results

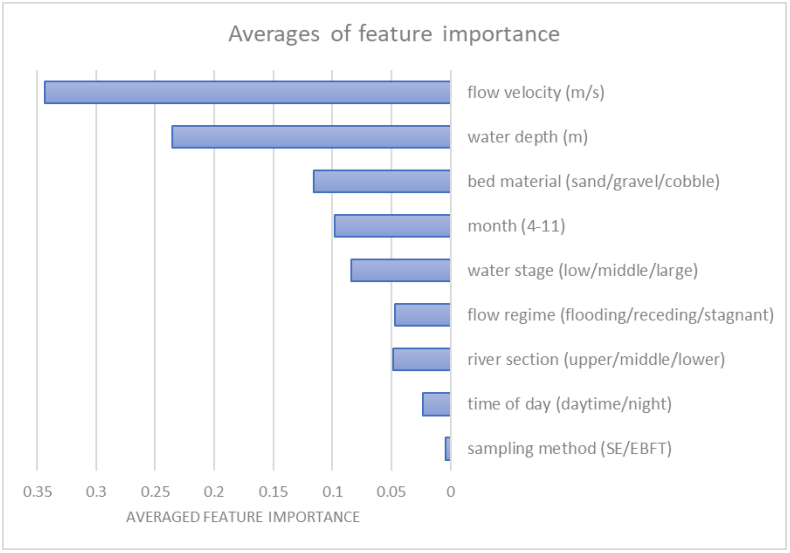

Figure 2. *Average weight (Feature importance) in the presence of the most common fish species in the Danube of the nine biotic variables used in the model.*

The RF analysis showed that from the abiotic variables selected for the study, flow velocity and water depth were the particularly important ones, however, bed material quality was ranked as the third most significant *(Figure 2 2)*. The high priority given to the first two parameters is consistent with the fact that, in the early days of habitat suitability research, US researchers began to rate habitats on these same two parameters (U.S. Fish and Wildlife Service, 1985), and the quality of bed material was added to the set later (Baranya, *et al.,* 2018). Some additional parameters should be treated with caution. From our own experience, we know that time of day and sampling method also influence the species composition of the sampled fish assemblage, which is not reflected in the weights of these parameters. It is important to stress, however, that the dependent variables in our study were numbers of individuals and species, not species classification. This approach does not therefore allow us to show the change in composition, but only the change in species number and number of individuals (which, according to the model results, was not significant along the two sampling methods and the two time periods). It is also important to note that we omitted several rarer species from the model for which the sampling method would certainly have had a higher weight. A good example is the sterlet *(Acipenser ruthenus)*, which can only be fished successfully with the EBFT method, as it prefers open water habitats (Szalóky, *et al.*, 2014).

When the weights of the nine abiotic parameters are considered by species, it's clearly visible, that in case of some species the flow velocity is multi-determinative compared to the other variables *(Figure 3)*.

It is known that the European perch (*Perca fluviatilis*) typically prefers stagnant waterbodies, while the vimba bream (*Vimba vimba*), on the contrary, prefers faster flowing rivers with faster current (Harka & Sallai, 2004). Flow velocity is therefore an equally important limiting factor in the occurrence of both species, as shown by our model. Another observed phenomenon is related to the round goby (*Neogobius melanostomus*), in the occurrence of which RF has homogeneously defined a low weight for the three most important parameters *(Figure 3)*. This leads to the conclusion that the species has no particular requirements in terms of hydrological and morphological variables. This is supported by the fact that we are talking about a generalist alien species that has been able to spread in the Hungarian Danube in a short period of time (Harka & Sallai, 2004); the invasion success of this species maybe due to this kind of 'undemandingness'.

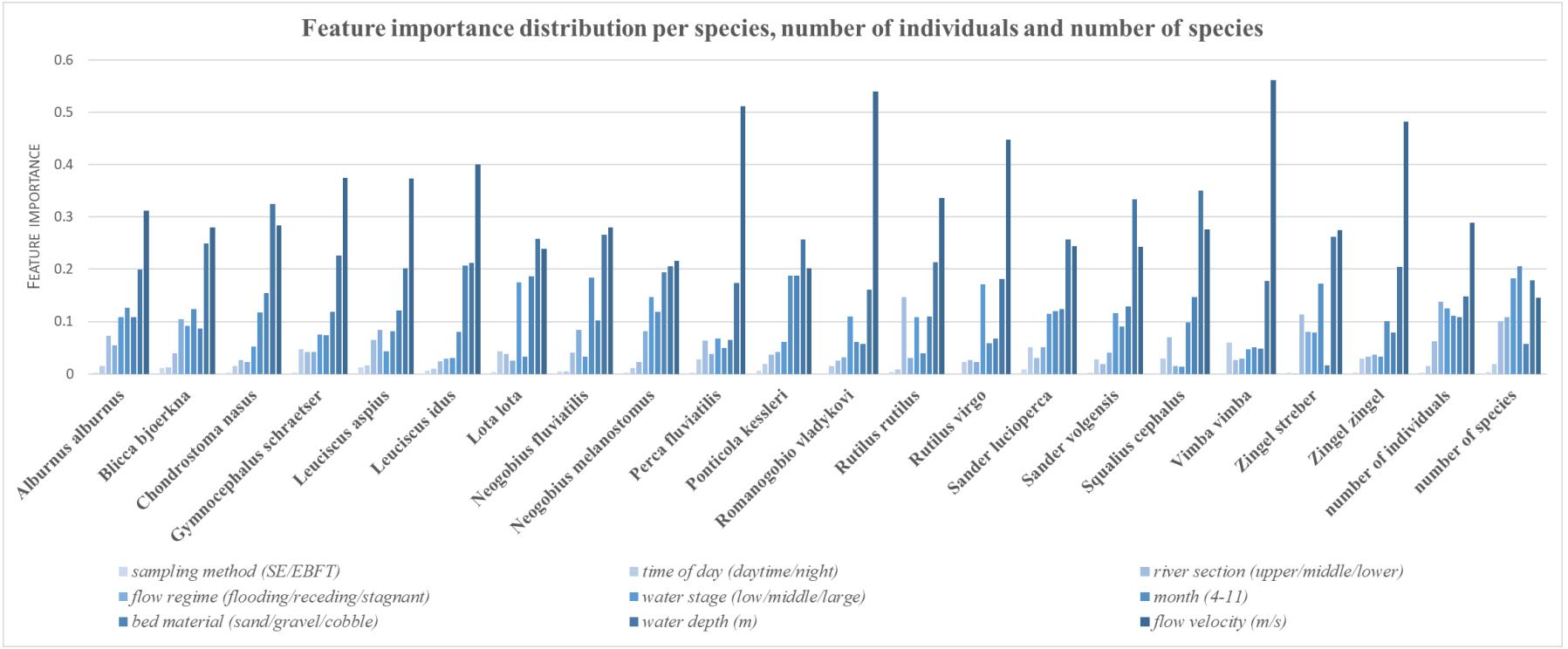

Figure 3. *Weights of the 9 abiotic parameters (feature importance) involved in the occurrence of the 20 most abundant fish species for each species and for the number of individuals and number of species (RF).*

3.2 Using the model

With the help of the RF model, we can create fictitious situations, queries, in which we set the conditions (e.g., season, water stage) and get an estimate of the individual number of each fish species at a given location in the Danube based on the learned relationships. This allows us to model the spatial distribution of fish according to abiotic variables, i.e., the suitable habitats to each species.

In a query, the model provides estimates of the individual numbers of the 20 most abundant fish species at each point on a small-scale 2D grid over the Hungarian section of the Danube (Füstös, *et al.*, 2021). The value obtained is often not a whole number, as the model averages the results of the regression trees, but it can be used to classify river habitats in terms of the species based on the relationships between the variables.

3.2.1 Large-scale query: abundance map of the common roach (Rutilus rutilus) in spring, in a receding and flooding flow regime

The roach is a common fish species found throughout Europe. It lives mainly in large rivers and lakes. It is favoured by the construction of canals and dams, as it prefers areas with slower current (Kottelat & Freyhof, 2007). In Hungary, it occurs everywhere except in shallow streams and faster flowing river sections. Juveniles are found in the kelp vegetation, while the larger ones are found near the shores (Harka & Sallai, 2004).

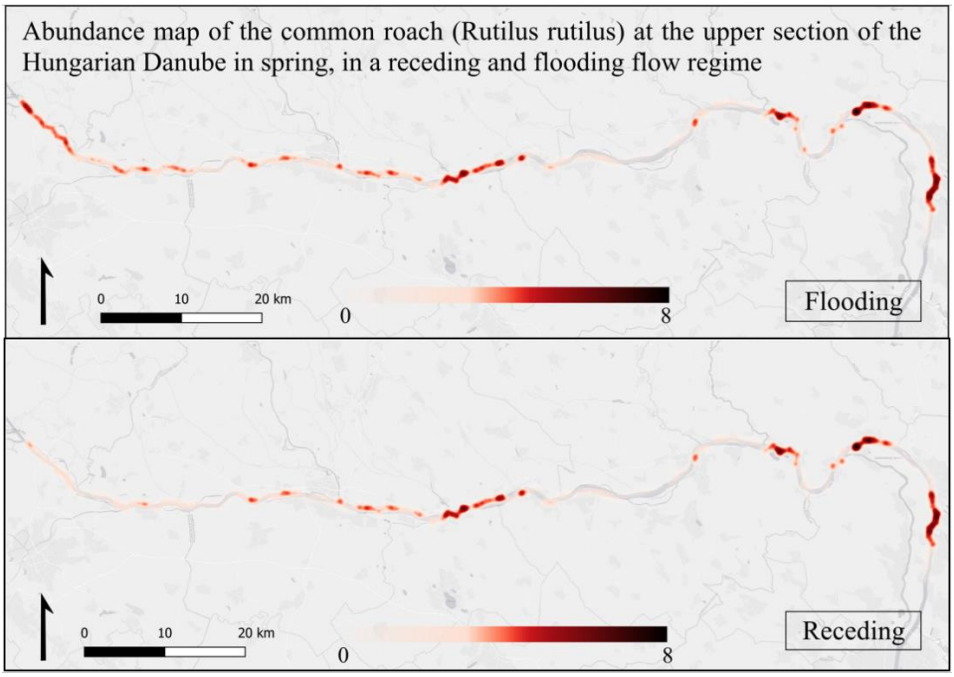

Figure 4. *Predicted habitats of the roach (Rutilus rutilus) in the upper section of the Danube on spring during flooding and receding water (RF).*

The RF model weights suggest that for this species, in addition to flow velocity and water depth, the bed material, river stage and water stage also have a significant influence. This is clearly

shown in Figure 4, which shows an overall increase in abundance during flooding, which is particularly evident in the Szigetköz area (a braided section of the Danube, where the river enters the country, on the left side of the figure). This is related to the widening of the coastal region due to the flood, which is a suitable habitat for the roach.

It also shows *(Figure 4)* that the greatest densification areas are constant regardless of the flow regime. These are areas, where islands or bays provide pretty suitable habitats for the roach.

3.2.2 Small-scale queries: abundance map of the bleak (Alburnus alburnus) in summer, in receding flow regime on the Radványi Island

The bleak *(Alburnus alburnus)* is a common cyprinid species, which occurs in almost all of Europe, except the Pyrenees and Apennines. (Kottelat & Freyhof, 2007). It is a social species found in all, average to large, stagnant and flowing lakes and rivers. Its largest populations are found, among other places, in the bream zone of rivers, as it prefers slower-flowing regions (Harka & Sallai, 2004).

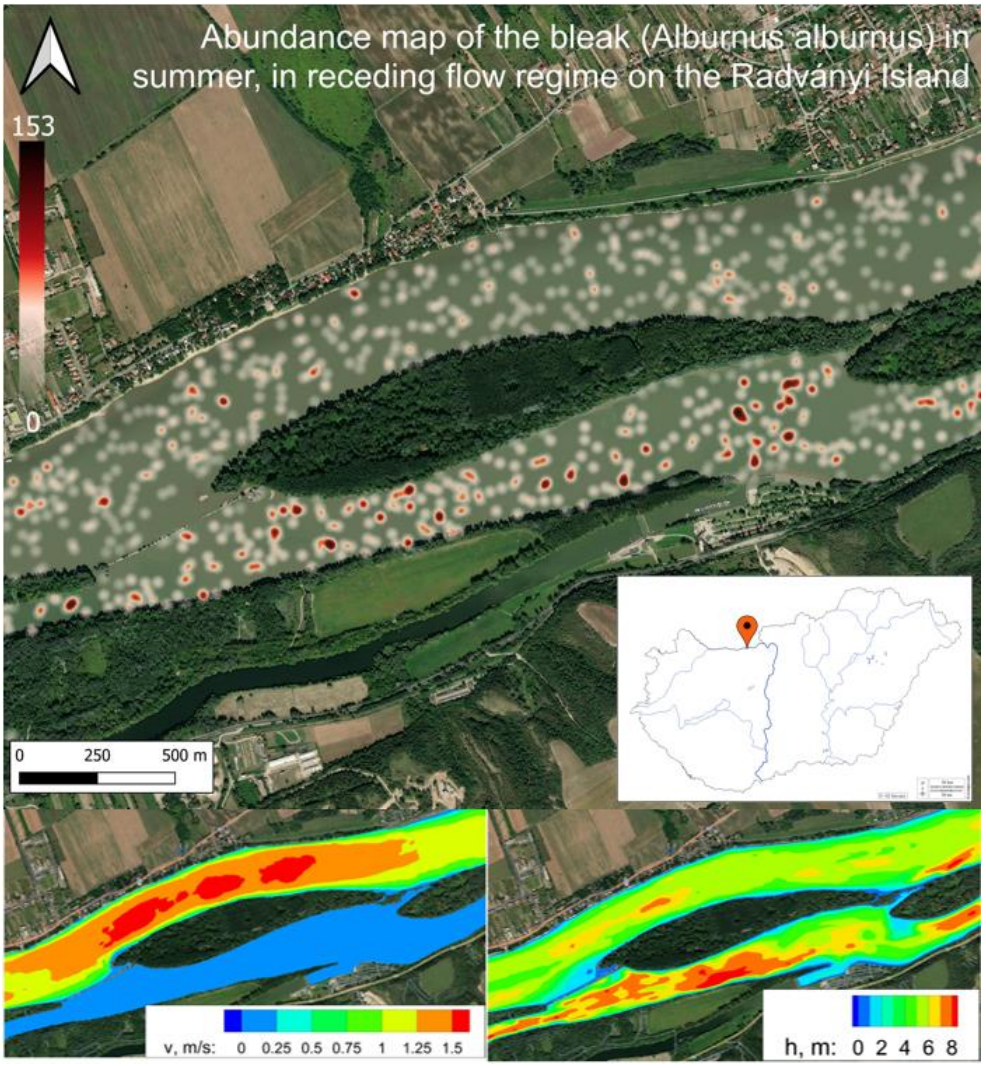

Figure 5. *Predicted map of optimal habitats for the bleak (Alburnus alburnus) near the Radványi island compared with flow velocity (v) and water depth (h).*

Based on the weights given by the RF, the flow velocity is the most limiting factor in its occurrence *(Figure 3)*. However, unlike many other species, it is less sensitive to water depth due to its lifestyle: feeding most often near the surface on planktonic organisms, insects that fall into the water, plant parts or organic debris (Harka & Sallai, 2004). This is can clearly be observed in the proximity of Radványi Island (Neszmély) *(Figure 5)*, the most suitable habitats for the species are found in the stagnant side arm, despite the fact that the water depth value is quite high.

3.3 Testing the model using the data of the bleak *(Alburnus alburnus)*

To test the accuracy and effectiveness of the model, we used data on the number of individuals of bleak *(Alburnus alburnus)*, which was found to be highly sensitive to flow velocity based on the weights given by the model *(Figure 3)* and the validation dataset contained the highest number of non-zero records for this species. The samples were selected according to the variables affecting the velocity. So that 26 samples taken in July (the month was chosen because, although not very important for this species, it is dominant in the spatial distribution) in mid-water stage and receding flow regime, which were represented in the validation data to which the medians of the model estimates were assigned.

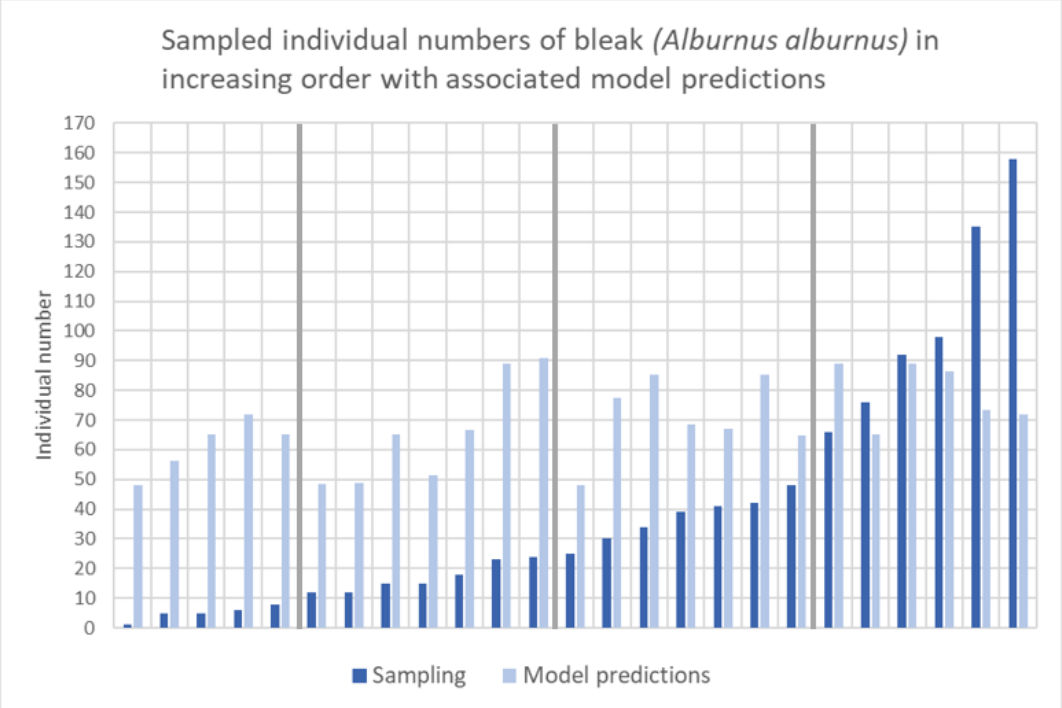

Figure 6. *Sampled individual numbers of bleak (Alburnus alburnus) from the validating dataset (dark blue) in increasing order with quartiles and estimated individual values belonging to them (light blue).*

The model gave similar values for all samples under the given conditions (σ = 14,25; range = 42,81). In contrast, the true individual numbers took values over a much larger range (σ = 41,53; range = 157). In most cases the model significantly overestimates, but the difference between

the true and estimated values decreases from quartile to quartile and is smallest in the 4th quartile. In addition, as shown in the figure *(Figure 6)*, there is a narrow interval where the error is acceptably small. This range is a kind of hydromorphological optimum where these variables determine the number of individuals. In the other cases, other abiotic variables (e.g., dissolved oxygen) or mostly biotic variables cause the large divergence observed, which is negative in most cases, so that biotic factors have a greater limiting power than the hydromorphological variables we used. In addition, of course, there are the large extreme values, which are outside the scope of the models.

3.4 Results of the analysis of species pairs

3.4.1 Neogobius melanostomus – Zingel streber (alien generalist – native specialist)

The round goby *(Neogobius melanostomus)* is a highly generalist species, as confirmed by the weights in the RF model, as there is no prominent limiting variable *(Figure 3)*. It does not prefer deep areas with too strong currents but can be found in very high abundance near the bottom, on practically any kind of bed material type and anywhere. It is the most common of the ponto-caspian goby species in the Danube. Like the other species, it is difficult to assess the impact of human activity in its establishment (Erős & Vörös, 2017; Tavares, *et al.*, 2020). In contrast, the Danube streber *(Zingel streber)* is a protected species native to the Danube drainage, which is primarily a current-favouring species, and thus occurs in higher numbers in deeper parts of the riverbed in sections with stronger currents (Füstös, *et al.*, 2021).

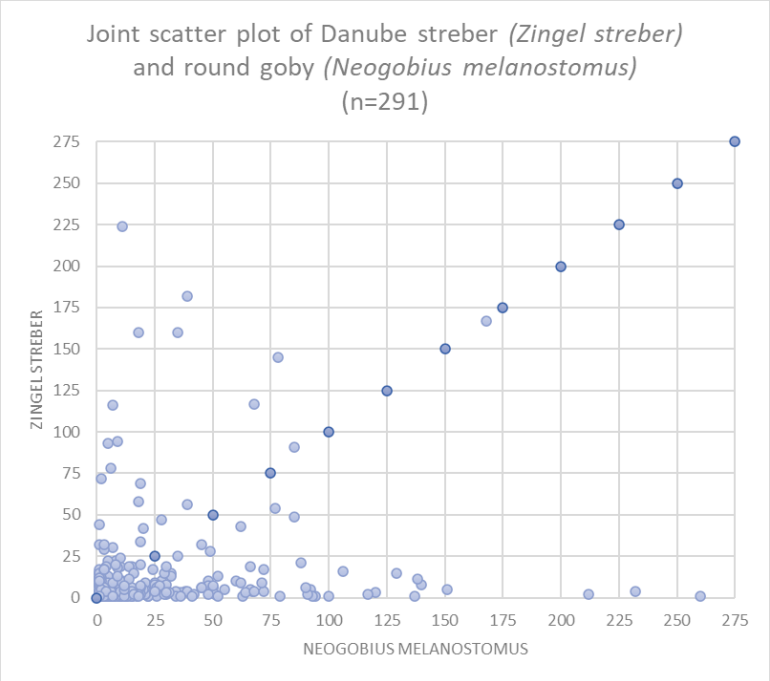

Figure 7. *Joint scatter plot of the native, specialist danube streber (Zingel streber) and the invasive, generalist round goby (Neogobius melanostomus) with a f(x)=x line.*

There were 291 co-occurrences of the two species. The scatter plot of the number of individuals does not show any definite relationship between the two variables *(Figure 7)*. On the other hand, the dots tend to the right of the reference line, i.e., towards the round goby *(Figure 7)*, which is confirmed by the results of the paired t-test, showing that the mean difference between the proportions of the species is 0,20 ($p<0,0001$), i.e., when they occur together, there are on average 20% more round gobies than strebers. This may represent a serious trophic competition between the two species and the values obtained are worrying for the Danube streber population. However, the scatter plot shows that in areas where strebers are abundant, i.e., in deep and strong current areas, the number of round gobies is low *(Figure 7)*, so that the specialist species can 'beat' the aggressively expanding invasive species in its optimal habitat.

3.4.2 Sander lucioperca – Perca fluviatilis (trophic competitors)

The pikeperch *(Sander lucioperca)* and the European perch *(Perca fluviatilis)* can be competitors for food up to a certain age, as both species consume small fish as juveniles. The methods used for sampling on the Danube can effectively capture this size range, so it is interesting to investigate how these two species divide the available habitat between each other.

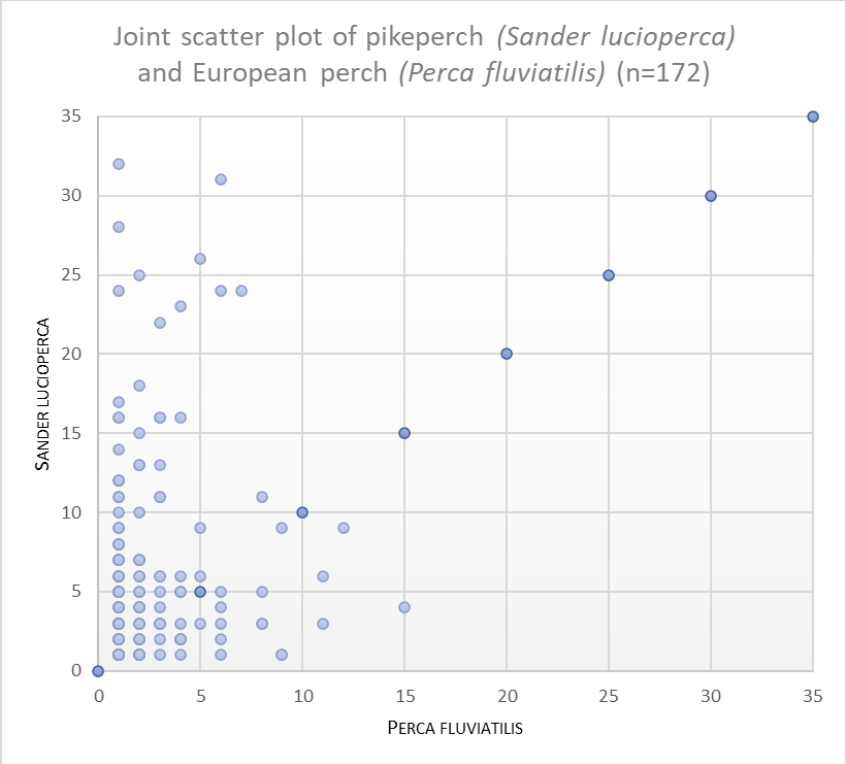

Figure 8. *Joint scatter plot of pike perch (Sander lucioperca) and European perch (Perca fluviatilis) with f(x)=x line.*

The combined scatter plot shows that the points are almost exclusively concentrated on the left side, i.e., at the pikeperch *(Figure 8).* This is supported by the results obtained from the proportion variables, as the estimated value of the mean difference of the two species was 0,29 ($p<0,0001$ (paired t-test)). When the two species occur together, there are therefore on average 29% more pikeperch than European perch. In addition, the scatter plot shows that with small numbers of individuals, the number of individuals often equal each other *(Figure 8)*, however, as the results show, the pikeperch is more dominant in the Danube than its relative. This is supported by the results of the RF model, as the presence of the European perch is very strongly limited by the flow velocity *(Figure 3)*, suggesting some kind of drift-based partitioning between the two species, with the pikeperch being more abundant in areas with stronger current and the European perch preferring slower flowing or stagnant water. Although the two species are not directly related (Sloss, *et al.,* 2004), but it is possible that a form of vicariance underlies the phenomenon, similarly to the evolution of other species in the family *Percidae* (Near, *et al.,* 2001).

4 SUMMARY AND CONCLUSIONS

In our study, fish faunistic surveys of the Danube over the last cca. 15 years (2004-2022) were processed and used to train an RF model. The nine parameters used as independent variables cover the most important hydromorphological variables and a significant part of the abiotic factors. The 20 fish species used as dependent variables cover the major species groups in the Danube (Kottelat & Freyhof, 2007). The model is unique among fish distribution prediction models in that it is based on a Machine Learning approach and has a large number of dependent and independent variables. It also covers a very large continuous area, unlike previous habitat modeling on the Danube, which furthermore used other methods and fewer fish species as dependent variables, but with similar results (Baranya, *et al.*, 2018; Füstös, *et al.*, 2019). The method above is generally applicable, given the existence of a pre-existing, detailed database of sampling history.

Based on the weights of the RF model, we were able to determine the role of each abiotic variable in the abundance of the 20 most common fish species. The distribution of weights per species provides information on the generalist or specialist nature of some species. In summary, we found that flow velocity, water depth and bed material quality are the three most important variables, which aligns well with the literature (U.S. Fish and Wildlife Service, 1985; Baranya, *et al.,* 2018). However, it is important to emphasize that the low weights obtained for each

variable do not necessarily imply a lower importance of the variable, it simply did not prove to be a determining factor for the given parameters (e.g., sampling method, time of day).

The model is able to predict the distribution of individual fish species in the entire section of the Danube under given conditions using queries based on 2D hydrodynamic model results. The values it gives also follow the longitudinal and transverse relationships very well. Hence, our model is an excellent tool to investigate the spatial distribution of fish habitats and to predict changes in the conditions at both small and large scales.

Our model is not aiming to accurately predict the actual number of fish, but rather to infer the appropriate habitats based on the values it provides. The tests have shown very well that biotic variables are much more limiting the fish distribution and abundance in some cases.

There was a narrow interval, the hydromorphological optimum, a range where the model estimates were very close to reality. It sets the course of further investigation raising questions about variability in extent and location of this hydromorphological optimum between fish species. So, are there species of fish for which biotic variables are less or differently dominant?

The analysis of species pairs using classical statistical methods showed that the weights given by the model can explain population relationships and determine the habitat of each species, and conversely, population relationships as biotic variables also can be determinant in the number of individuals.

Our research has raised many more questions. It may be important to add more well-measured or assessable abiotic variables to the model for more accurate predictions of the suitable riverine habitats corresponding to each fish species. The independent variables in the model could be extended to include, for example, pH, which is a very important determinant in lakes (Matuszek & Beggs, 1988) or dissolved oxygen, which is an important abiotic factor in general. In this way we can indirectly investigate the influence of biotic factors.

There are further opportunities to increase the spatial applicability of the model, which could cover additional stretches of the Danube or its side rivers. An obstacle to this, especially for side rivers, is the different hydrodynamic and hydromorphological conditions, which would require the introduction of new models to predict certain abiotic variables. Beyond overcoming these methodological obstacles, however, there are already some technical solutions to increase the spatial usability of ML models (Meyer & Pebesma, 2021).

For the model to be able to estimate the number of individuals accurately, a large number of biotic variables would have to be added. Many of these, such as trophic relationships, are

extremely difficult to measure or assess (e.g., a species may be at different trophic levels at different ages). It is also extremely difficult to determine the impact of a species on the ecosystem, so it is difficult to know which biotic variables need to be included to obtain an accurate estimate of abundance for example (Franco, *et al.*, 2020). Adding biotic variables would therefore require a deeper understanding of the population relationships of each species.

Our research took another step towards AI-supported ecology, in presenting the applicability of a Random Forest model in assessing fish-habitat relationships in a very large river. The results on one hand showed that Machine Learning methods are promisingly able to distinguish the different weights of parameters affecting habitat selection of fish, and thus may help in increasing the efficiency of future samplings. On the other hand, it once again emphasized the importance of detailed, long-term databases, being the core training material for Machine Learning approaches.

5 BIBLIOGRAPHY

Baranya, S. *et al.*, 2018. Habitat mapping of riverine fish by means of hydromorphological tools. *Ecohydrology,* 11(7).

Bergström, P., Gonzalez-Mirelis, G. & Lindegarth, M., 2011. Interaction between classification detail and prediction of community types: implications for predictive modelling of benthic biotopes. *Marine Ecology Progress Series,* Volume 432., pp. 31-44.

Elmahdy, S. I. *et al.*, 2020. Spatiotemporal Mapping and Monitoring of Mangrove Forests Changes From 1990 to 2019 in the Northern Emirates, UAE Using Random Forest, Kernel Logistic Regression and Naive Bayes Tree Models. *Frontiers in Environmental Science,* Volume 8., pp. 102-125.

Erős, T. & Vörös, J., 2017. Áttekintés a hazai idegenhonos inváziós halak, kétéltűek és hüllők jelenlegi helyzetéről (in Hungarian). *Magyar Tudomány,* Volume 4, pp. 426-428.

Farkas-Iványi, K. & Trájer, A., 2015. The influence of the river regulations on the aquatic habitats in river Danube, at the Bodak branch-system, Hungary and Slovakia. *Carpathian Journal of Earth and Environmental Sciences,* 10.(3.), pp. 235-245.

Franco, A. C. S., Garcia-Berthou, E. & dos Santos, L. N., 2020. Ecological impacts of an invasive top predator fish across South America. *Science of The Total Environment,* Volume 761.

Füstös, V. *et al.*, 2019. Habitat based hydrodynamic investigation of the Upper-Hungarian Danube River (in Hungarian). *Pisces Hungarici,* Volume 13., pp. 81-90.

Füstös, V., Erős, T. & Józsa, J., 2021. 2D vs. 3D Numerical Approaches for Fish Habitat Evaluation of a Large River - Is 2D Modeling Sufficient?. *Periodica Polytechnica Civil Engineering,* 65(4), pp. 1114-1125.

Harka, Á. & Sallai, Z., 2004. *Fish fauna of Hungary (in Hungarian).* Szarvas: Nimfea Természetvédelmi Egyesület.

Implementing the Water Framework Directive in Hungary, 2009. *River Basin Management.* Budapest: Central Directorate of Water and Environment.

Kottelat, M. & Freyhof, J., 2007. *Handbook of European Freshwater Fishes.* Cornol, Switzerland and Berlin, Germany: Maurice Kottelat and Jörg Freyhof.

Liu, Z. *et al.*, 2018. Application of machine learning methods in forest ecology: recent progress and

future challenges. *Environmental Reviews,* 26(10), pp. 339-350.

Matuszek, J. E. & Beggs, G. L., 1988. Fish Species Richness in Relation to Lake Area, pH, and Other Abiotic Factors in Ontario Lakes. *Canadian Journal of Fisheries and Aquatic Sciences,* 45(11), pp. 1931-1941.

Meyer, H. & Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution,* 12.(9), pp. 1620-1633.

Near, T. J., Page, L. M. & Mayden, R. L., 2001. Intraspecific phylogeography of Percina evides (Percidae: Etheostomatinae): an additional test of the Central Highlands pre-Pleistocene vicariance hypothesis. *Molecular Ecology,* Volume 10., pp. 2235-2240.

Olden, J. D., Lawler, J. J. & Poff, N. L., 2008. Machine learning methods without tears: A primer for ecologists. *The Quarterly Review of Biology,* 83(2), pp. 171-193.

Reiczigel, J., Harnos, A. & Solymosi, N., 2010. *Biostatisztika nem statisztikusoknak (In Hungarian).* Nagykovácsi: Pars Kft.

Sloss, B. L., Billington, N. & Burr, B. M., 2004. A molecular phylogeny of the Percidae (Teleostei, Perciformes) based on mitochondrial DNA sequence. *Molecular Phylogenetics and Evolution,* Volume 32., pp. 545-562.

Sylvain, C., Hervet, É. & Lecomte, N., 2019. Applications for deep learning in ecology. *Methods in Ecology and Evolution,* Volume 10, pp. 1632-1644.

Szalóky, Z. *et al.,* 2014. Application of an electrified benthic frame trawl for sampling fish in a very large European river (the Danube River) – Is offshore monitoring necessary?. *Fisheries Research,* Volume 151, pp. 12-19.

Tavares, C. N. *et al.,* 2020. Tracing the colonization process of non-native gobies into a large river: the relevance of different dispersal modes. *Biological Invasions,* Volume 22., pp. 2421-2429.

Tózsa, I., 1998.. Tájképi homogenitás Magyarországon (In Hungarian). *Földrajzi Értesítő,* 47(3), pp. 432-445.

U.S. Fish and Wildlife Service, 1985. Habitat suitability models and instream flow suitability curves: pink salmon. *Biological report,* p. 10.109.

Zajicek, P. & Wolter, C., 2018. The gain of additional sampling methods for the fish-based assessment of large rivers. *Fisheries Research,* Volume 197., pp. 15-24.

Benedek Jandó is currently a student at the Veres Pálné Gimnázium, but besides his studies he works as a volunteer for several organisations, such as the Birdlife Hungary and the Hungarian Biodiversity Research Society. His main interests are ecology, evolutionary biology, ornithology, biostatistics, and hydrobiology. He took part in many bird monitoring and research projects as a bird ringer. Since his early childhood he was interested in nature and regularly asked questions about its functioning. Thanks to this curiosity, he conducted several self-made projects, which was dealt with the ecology of birds and other living organisms. He would like to be a researcher, who studies a wide range of topics, such as the ecological modelling of fish habitats.

Summary

In this study, we assigned the data of the 20 most common fish species in the Hungarian section of the Danube River from 2004 to 2022 to the data of nine abiotic parameters, and looked for patterns among them using ML. Among the abiotic factors used as independent variables in the analysis, the average depth velocity, water depth and bed material composition were the most decisive variables, which aligns with the results of previous research. In addition, with our RF model, we can predict the suitable habitats of each fish species under given conditions and examining the impacts of biotic factors. Our methods can be transplanted to other large rivers, to investigating the ecology of fishes.

Statement on the student's own work

The student worked under our guidance and supervision, but he did the following workflows alone: doing bibliographical research; cleaning the raw Danube fish faunistic database and preparing it for the analyses; figuring out the methods and carrying out the classic statistical analysis of the database; drawing conclusions from the RF results; making maps from the RF queries and interpreting them; testing the model with the validating dataset and interpreting it; making every single figure in the article.