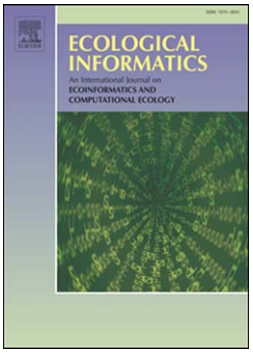
Accepted Manuscript



Use of Self Organising Maps in modelling the distribution patterns of gammarids (Crustacea: Amphipoda)

Eszter Á. Krasznai, Pál Boda, András Csercsa, Márk Ficsór, Gábor Várbíró

PII: S1574-9541(15)00188-0
DOI: doi: 10.1016/j.ecoinf.2015.11.007
Reference: ECOINF 634

To appear in: *Ecological Informatics*

Received date: 18 March 2015
Revised date: 10 September 2015
Accepted date: 9 November 2015

Please cite this article as: Krasznai, Eszter Á., Boda, Pál, Csercsa, András, Ficsór, Márk, Várbíró, Gábor, Use of Self Organising Maps in modelling the distribution patterns of gammarids (Crustacea: Amphipoda), *Ecological Informatics* (2015), doi: 10.1016/j.ecoinf.2015.11.007

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Use of Self Organising Maps in modelling the distribution patterns of gammarids (Crustacea: Amphipoda)

Eszter Á. Krasznai[1], Pál Boda[2], András Csercsa[3], Márk Ficsór[4], Gábor Várbíró[5]

[1] Doctoral School of Chemistry and Environmental Sciences, University of Pannonia, H-8200, Egyetem u. 10. ,Veszprém , Hungary and MTA Centre for Ecological Research, Department of Tisza River Research, H-4026 Debrecen, Bem tér 18/C, Hungary E-mail: krasznai.eszter@okologia.mta.hu

[2] MTA Centre for Ecological Research, Department of Tisza River Research, H-4026 Debrecen, Bem tér 18/C, Hungary E-mail: boda.pal@okologia.mta.hu

[3] Doctoral School of Chemistry and Environmental Sciences, University of Pannonia, H-8200, Egyetem u. 10. ,Veszprém, Hungary and MTA Centre for Ecological Research, Department of Tisza River Research, H-4026 Debrecen, Bem tér 18/C, Hungary E-mail: csercsa.andras@okologia.mta.hu

[4] Laboratory of North Hungarian Environmental Protection, Nature Conservation and Water Inspectorate, H-3530 Miskolc, Mindszent tér 4., Hungary E-mail: ficsor.mark@emikofe.kvvm.hu

[5] MTA Centre for Ecological Research, Department of Tisza River Research, H-4026 Debrecen, Bem tér 18/C, Hungary E-mail: varbiro.gabor@okologia.mta.hu

Corresponding author: Eszter Á. Krasznai, e-mail: krasznai.eszter@okologia.mta.hu

Abstract

Self Organizing Maps (SOMs) are increasingly popular methods in processing high-dimensional ecological data, however, their potentials are not yet fully utilized. It was our objective to prove evidence on an unknown advantage of the SOMs which we aimed to test using data on the spatial distributional patterns of gammarids. Quantitative samples and a wide spectrum of environmental data were obtained from the catchment area of two of the largest side tributaries of the Tisza River. Distributional patterns and habitat preference of three Gammarus species were described by Self Organizing Map methods and regression tree analysis (CART) on spatial and temporal scale. Using SOMs helped us to bring out distinctions in our data and enhance the differences, thus making them easier to recognize and

ACCEPTED MANUSCRIPT

also, with their help, we were able to model the relations of the species to habitat types non-existent among our samples. According to the analysis *Gammarus roeselii* preferred low altitudes, high conductivity, fine substrate, deep actual mean depth and dense plant coverage; *G. fossarum* preferred rocky stream beds, high altitude, lower temperatures and little actual mean depth; while *G. balcanicus* preferred coarse substrate, little or no plant coverage and low temperature. SOMs improved the correlations that proved to be highly useful: besides their use to display complex data in a perspicuous way they have other advantages in bringing out existing relationships in data otherwise difficult to detect.

Keywords

Habitat structure; Niche preference; Interspecific Competition; Regression tree

1. Introduction

Self Organising Maps (SOMs) are a novel yet increasingly popular method in ecology. The SOMs have been used on data of aquatic macroinvertebrates mainly for clustering and visualisation like describing the distributions of the communities in France (Cereghino et al. 2001), in South Korea (Park et al. 2007a) or in China (Li et al. 2012). Besides this they were also successfully, though rarely used in patterning habitat preferences of community assemblages in case of birds (Lee et al. 2010), fish (Dukowska et al. 2013) and even macroinvertebrates (Goethals et al. 2013). They are not only suitable for clustering and to visualize high-dimensional data (Kohonen 1998) but proved an equivalent to conventional statistical methods for ecological patterning (Chon et al. 1996, Giraudel and Lek 2001).

The flowchart of the SOM analysis shows that, using SOM, the input raw dataset, that usually contains a matrix of samples with species as and environmental factors as variables, has three analytical outputs (Fig. 1.):

ACCEPTED MANUSCRIPT

i) The first approach was using SOM as an explorative analysis method to detect and explain sample clusters and sample groupings (Chon et al. 1996). Using SOMs it is possible to convert complex statistical relationships between high-dimensional data into simple geometric relationships on a low-dimensional display preserving the most important topological and metric relationships of the primary data (Fig. 1.a.) (Chon 2011, Várbíró et al. 2007).

ii) By visualising the SOM component planes it is possible to display the groupings of sites (Fig. 1.b.), species abundances and abiotic variables together; therefore, each variable can be evaluated by its influence (Várbíró et al. 2012).

iii) Using the SOM dataset as a model to analyse its data further for correlation and other statistical tools (Fig. 1.c.) is not yet common in ecology. We would like to demonstrate the strength of this approach as it leads to valuable ecological relevancies. To highlight this approach we used a gammarid dataset of a river catchment as a case study.

Gammarids, being easy to collect and abundant in most freshwater ecosystems, proved to be an ideal choice as model organisms. Gammarids belong to the most successful organisms invading aquatic habitats, they often exist in high densities and occasionally can dominate the macroinvertebrate fauna in streams and rivers (Giller and Malmquist 1998, Wesenberg-Lund and Storch 1939). In running waters they can even account for 80–90 % of the macroinvertebrate numbers as well as biomass (van Riel et al. 2006). Their success can be explained by their high tolerance for a wide range of environmental conditions (Bruijs et al. 2001, Devin and Beisel 2007, Wijnhoven et al. 2003), by their high reproductive capacity (Devin et al. 2004) and by their superior competitiveness as predators (Dick et al. 1990, MacNeil and Platvoet 2005). However, spatial as well as temporal isolation of resources and habitats may lead to the coexistence of gammarids (MacNeil et al. 2001). In Hungary, *G. fossarum*, *G. roeselii* and *G. balcanicus* are also considered as part of the native fauna, all of

ACCEPTED MANUSCRIPT

which belong to the most common amphipod species found in Europe (Grabowski and Mamos 2011).

According to our present knowledge, the spatial distribution of these species is mainly influenced by altitude – with *G. balcanicus* living at the highest altitude, followed by *G. fossarum* in the lower regions and *G. roeselii* and *Asellus aquaticus* in the lowermost regions (Pârvulescu 2009). However, besides the altitude, numerous other factors like dissolved oxygen, substrate and plant coverage can also influence the distribution of species (Mauchart et al. 2014).

Our paper had two objectives; the ecological goal was to reveal coexistence patterns of gammarid species, and the statistical/modelling aim was to prove evidence on the superiority of the SOM method against classical tools of statistical methods. We hypothesised that the use of Self Organizing Maps in processing the data considerably reduces the noise and enhances the descriptive value of the model.

2. Materials and Methods

2.1. Study sites and sampling procedure

The sampling locations can be found on the watershed of the Hernád and Sajó rivers, two of the largest side stream tributaries of the Tisza River (Table 1.). The catchment areas of the two rivers cover 18.144 km^2 (Sajó: 12.708 km^2 and Hernád: 5.436 km^2) (Vogt et al. 2007). The Sajó river is 223 km long and it receives 10 streams and a channel (Dobsina, Csermosnya, Csetnek, Murány, Turóc in Slovakia and Keleméri, Hangony, Bán, Tardona, Nyögő, Szuha, Szinva and the Takta channel in Hungary), 3 rivers (Rima in Slovakia and Bódva and Hernád in Hungary) and the Kis-Sajó anabranch, also in Hungary. The Hernád river is 286 km long (118km in Hungary) and it receives 7 streams (Perényi, Gönci, Kis-Hernád, Vasonca, Vadász, Bársonyos, Szartos, all in Hungary) and 3 rivers (Gölnic, Tarca,

ACCEPTED MANUSCRIPT

Olsava, all in Slovakia). The samples were taken at 42 sites on a single occasion during the summer (7-12 of June) of 2012 (Fig. 2.).

The representative units were localized during the field sampling and at each one of them a 20-50 m long section was selected to be the sampling site. We made sure there were no hydromorphological changes (e.g. bridge, bank saving pitching) near them. The sampling was carried out according to the AQEM protocol, thus a sample consisted of 20 sampling units taken from all habitat types according to their share (Hering et al. 2004). The proportions of the different habitat types were mapped at the sampling site and types of less than 5% coverage were not included in the sampling. The following physical factors were measured and registered on every sampling site: temperature, pH, dissolved oxygen and conductivity. The types of the habitats and the substrate found were also recorded along with hydromorphological factors (altitude, width of flood plain, average width and maximum width of stream, actual depth of stream, average depth and maximum depth at high-water). The samples were pre-sorted for vulnerable specimens in the field, their volumes were reduced and they were conserved in 70% ethanol. The sorting and the identification was carried out in the laboratory. The sorted amphipods were identified using relevant taxonomical keys (Gruner 1966, Karaman and Pinkster 1977, Kontschán 2001b, Kontschán et al. 2002).

2.2. Statistical analysis

The SOM method was used to answer our ecological question: we tried to find out which environmental factors had major influence on the presence of gammarids. The raw matrix that was the basis of the SOM analysis was constructed from the relative abundance of the three species and the 17 environmental variables. Sampling sites containing no or less than 30 individuals of gammarid species were removed from our analysed data set as these sites were dominated by *Asellus aquaticus*. The relative abundance values of the species at the different

ACCEPTED MANUSCRIPT

sampling events were plotted on a simplex diagram. Here closeness to the vertices refers to the domination of the given species as its relative abundance is one at the vertex.

The algorithm and structure of the SOM can be achieved through a neural network that uses self-organizing processes. The SOM is a linear array of artificial neurons with each neuron being represented and arranged in a two dimensional hexagonal lattice in the final presented form (Chon et al. 1996) (Fig. 1.). In our case, raw data matrix for a community containing "n" species and environmental factors (i.e., n dimensions), the abundance of a species, *i*, is expressed as a vector, x_i. Vector x_i is therefore considered to be an input layer for the SOM. Each node, *j*, is connected to each input node, *i*. The connection weights (initially, the weights are randomly assigned), $w_{ij}(t)$, change adaptively at each iteration of the calculation, *t*, until convergence is reached by minimization of the difference, $d_{ij}(t)$, between input data x_i and the weight $w_{ij}(t)$:

$$\text{Eq 1.} \quad d_{ij}(t) = \sum_{i=0}^{N-1} (x_i - w_{ij}(t))^2$$

In the selection phase, the neuron of which weighted vector is in the shortest distance (Eq.1) (minimum $d_{ij}(t)$) to the input vector is chosen as the winner, the best matching unit (BMU) and this neuron is going to have the strongest respond in the next phase. In the learning phase, the chosen neuron and its neighbouring neurons are allowed to adapt by changing weights to further reduce the distance between the weighted vector and the input vector as (Eq. 2.):

$$\text{Eq. 2.} \quad w_{ij}(t+1) = w_{ij}(t) + \eta(t)(x_i - w_{ij}(t))Z_j$$

where Z_j is assigned 1 for the chosen (and its neighbouring) neuron(s) and is assigned 0 for the remaining neurons. The expression $\eta(t)$ denotes the learning factor. The radius-defining neighbourhood is usually set to a larger value early in the training process and is gradually reduced as convergence is reached. The weights of the BMUs and neurons close to

ACCEPTED MANUSCRIPT

it are adjusted toward the input vector through interactive calculation. This process separates the training phases into two steps: the rough (where the large range of neighbouring cells are affected by the learning) and the fine learning phase (usually only one neuron affected by the learning). In our case, the rough tuning phase included 500, while the fine tuning phase lasted 2000 iterations. In order to bring out relationships between environmental and biological variables, Compin and Cereghino (2007) introduced a mask function to give a null weight to the variables which should be excluded in the selection process, whereas variables with a weight of 1 were included in the selection process. Thus, setting mask value to zero for a given component removes the effect of that variable on the organization of the map but the neurons still learn this variable. Therefore the values for these variables could also be visualised on the SOM. By this method it is possible to display the relative abundance of the species and to reveal the effect of environmental factors on the distribution of the species observed. The number of the output neurons is an important factor in the analysis because it influences the final quality and topography of the SOM. We used the suggestion of Vesanto which determines the output map size as 5 times square root of the number of samples (Vesanto 2000).

SOM modelling was conducted using Matlab 8.2 and the SOM Toolbox for Matlab (Alhoniemi et al. 2000). The results of the SOM analysis can be visualized by individual component plains, where each of the species or variables has its own plain. In the plot, darker areas mean higher values of the given variables in the SOM's virtual unit. The resulted SOM matrix, that was analysed later, was built from 32 virtual units and 20 variables (3 biotic, 27 environmental). For correlation analysis between environmental factors we used Pearson product moment correlation. The assumption for normal distribution where needed were achieved by logarithmic transformation of the original raw dataset.

ACCEPTED MANUSCRIPT

To test the applicability of using the SOM method, all the following statistical methods were done both on the raw matrix and on the SOM matrix. Principal components analysis (PCA) was used for highlighting the environmental variables accounting for most of the variance in the distribution of the species (Davis 1987, Harper 1999). The PCA analysis was carried out using PAST (Hammer et al. 2001), the Convex hull delineation was based on the dominant species abundance. We used an analysis of similarities (ANOSIM) to test whether there were significant differences between the dominant species groups (Clarke and Ainsworth 1993). Discriminant Analysis were used to test whether the environmental factor determine the gammarids species distribution (Rohlf and Corti, 2000). Regression tree analysis (CART) was used to identify interrelations among different abiotic variables and the species dominance and to validate whether the results of the SOM model are easier to interpret.

CART methods are well established in ecology for the identification of relations between environmental and density variations (Clapcott et al. 2010). Detailed descriptions of the CART analysis can be found at Breiman et al. (1984). The categorical response dependent variable of the model was the dominant species for the given sample or SOM virtual unit.

3. Results and Discussion

3.1. Ecological question

Gammarus roeselii could be found in only half of the samples collected while both *G. balcanicus* and *G. fossarum* were represented in more than 70% of the samples. Most of the sampled sites (53%) contained two species (75% of these samples contained *G. balcanicus and G. fossarum*), the third of them (30%) all three and in only 17% of the samples was only one of the species present. In the simplex diagram the sites with only one species were

ACCEPTED MANUSCRIPT

displayed at the vertices, while samples containing two species dispersed along the edges. Sites with all three species present were placed inside of the large triangle, but it is also clearly noted that there were no sites where *G. balcanicus* and *G. roeselii* were present together. There is a clear gradient between *G. roeselii* to *G. fossarum* and between *G. fossarum* and *G. balcanicus* but not between *G. roeselii* and *G. balcanicus* (Fig. 3.).

The three species distributed separately in the SOM map (Fig. 4.). Using the mask function it is possible to force the SOM to learn the abiotic composition pattern of the sampling events therefore we could visualize the component planes of the given variable (Fig. 4.). In the case of *G. roeselii* there was a clear negative correlation between the altitude and the abundance of the species. Positive correlation with actual mean depth and the conductivity could refer to the fact that the species is more abundant on the lower reaches of the streams. The species also showed preference for fine (psammal-argylal) substrate and dense plant coverage (Fig. 4.). These results seemed to contradict our present knowledge of the *G. roeselii* preferring macrolithal stream beds and high levels of dissolved oxygen (Henry and Danielopol 1998, Kley et al. 2009). While preferring habitats with coarse bed substrate (mesolithal: 6-20 cm and microlithal: 2-6 cm in diameter), a strong negative correlation was observed between the distributions of *G. balcanicus* and habitats with a dense plant coverage, higher temperature and fine (psammal-argylal) bed substrate (Fig. 4.). *G. fossarum* clearly preferred rocky stream beds at high altitudes, while a strong negative correlation was proven between the distribution of the species and gravel (akal, 2 mm-2 cm in diameter) as bed substrate, temperature and the actual mean depth (Fig. 4.).

The habitat preferences of both G. balcanicus and G. fossarum seemed to be in consort with the expected distributional patterns (Pârvulescu 2009) although not in perfect accordance thus increasing the necessity of testing the possibility of a competition. This suggests that the separation is made by the level of degradation and not the difference in grain size, as opposed

ACCEPTED MANUSCRIPT

to literature data, according to which G. roeselii prefers rocky habitats. Hence, in a degraded mountainous stream the macrophyte and psammal-argylal substrate increases. This feature of hydromorphological degradation has been identified to be one of the most important stressors affecting the in-stream biota in many Central European stream types (Ofenböck et al. 2004). Besides the degradation, both native and non-native biological invasions appeared to be the major driving force in shaping of the diversity of biotic communities (Bollache et al. 2004, Borics et al. 2013, Kinzelbach 1995, Sala et al. 2000). Climatic changes, human induced stress or disturbance events do also often coincide with invading native species which results in altering the previous community pattern (Borics et al. 2013). The above is also true for the gammarid communities of European and Hungarian rivers, as two non-native Mysids appeared or extended their range of distribution at the last five years in Hungarian waters (Borza et al. 2011, Borza and Boda 2013).

3.2. SOM modelling

Using SOM method as an explorative statistical method, basic correlations were revealed between the relative abundance of the species and the environmental variables. Through the SOM's ability of highlighting correlations and bringing out existing relationships in data otherwise difficult to detect, we demonstrated the superiority of the method using both datasets (raw matrix (Appendix 1.) and SOM matrix dataset(Appendix 2.)).

The correlation of the species relative abundances of the raw dataset and species relative abundances of the SOM dataset with the abiotic factors can be found in Table 2. The SOM analysis increased the correlation among environmental factors due to the learning process of the neural network. The SOM dataset-based correlation calculation increases the R value in every case and much more significant correlations can be found in the SOM dataset than in the raw dataset (Table 2.).

ACCEPTED MANUSCRIPT

The CART analysis revealed the thresholds for each species habitat preferences. According to the CART based on the raw dataset, the first factor to influence proved to be the altitude, separating most of the populations of *G. roeselii* from the other species. At altitudes higher than 151 m, the rest of the species split based on the morphology of the bed substrate: most of the populations of *G. fossarum* were separated by the rate of microlithal substrate type. This species preferred habitats with a bed containing 7% or less microlithal substrate. On the next level, the CART suggests that some populations of *G. balcanicus* preferred to have the ratio of the macrolithal substrate type above 4.5%. The remaining group of samples was separated by the altitude again (Fig. 5.a.). Thus, based on the original raw dataset, altitude and the morphology of bed substrate proved to be the most important factors influencing the distribution of the species. The graphs based on the SOM dataset rest on higher correlations and are therefore more clear-cut. Here, too, altitude is the first factor clearly separating *G. roeselii* (below 239.09 m) from the other two species. These species did then separate based on the type of the bed substrate. A remarkably high significance showed *G. balcanicus* to prefer habitats with a share of 21.52% or less Psammal-argylal type of substrate and 7.22% or less plant coverage, while *G. fossarum* preferred habitats with a share higher than 21.52% PSARG (Psammal-argylal) substrate and more than 7.22% plant coverage (Fig. 5.b.).

Physical characteristics of the water based on the raw dataset also seemed to operate the choices of the species: two large groups could be separated based on conductivity, both of which contained more than one species. These could be then further divided into two groups. On one side conductivity remained the driving factor (with *G. balcanicus* preferring above 400.00 µS and some of the *G. fossarum* populations preferring below 400.00 µS), while pH influenced the other side (with *G. roeselii* preferring above pH 7.91 and *G. fossarum* below pH 7.91) (Fig. 6.a.).

ACCEPTED MANUSCRIPT

Physical characteristics of the water based on the dataset of the SOM work in a way very similar to that based on raw data. Here, too, conductivity is the first factor influencing the species. *G. roeselii* is detached from the other two species based on its preference for waters with their conductivity higher than 474.45 µS. The other two species separate based on pH (with most of the *G. balcanicus* populations preferring pH to be higher than 8.22) and conductivity (with the rest of the populations divided by the value of 353.21 µS, *G. balcanicus* preferring the conductivity below and *G. fossarum* above it) (Fig. 5.b.). On the whole, the differences are driven by the same factors but the separation of the sampling sites by their dominant gammarid species is clearer in case of the SOM dataset (Fig. 5.b.; 6.b.).

The SOM also enhances the difference among habitat preferences analysed by PCA. Here the main variable loading remains the same (microlithal, PSARG, mesolithal) but there are significant differences among them. The ANOSIM test shows no differences between *G. balcanicus* and *G. fossarum* groups in the raw dataset, however in the SOM model the differences among the groups were significant (Fig. 7.). This was also true for the Discriminant analysis, where the performance of the discrimination on the SOM was 100%, while on the raw dataset it was only 80.23%. In addition, when the group assignment was cross-validated by a leave-one-out cross-validation (jack-knifing) procedure, it resulted 90.62% for the SOM dataset and only 38.3 % for the raw dataset. (Appendix 3., 4.)

Besides habitat differences, species competition also plays a role in the presence and dominance of the species at a site. Temporal data gathered (from 2005 to 2012) by the North Hungarian Inspectorate for Environment, Nature and Water shows that through competition species may replace each other. Long term temporal changes in the species composition of the Sajó - Hernád watershed indicate an increase in the abundance of G. fossarum in Bódva lower section, Jósva stream and Telekes stream, while in the upper Bódva the increase in the abundance of *G. roeselii* occurs (Fig. 8.). While this turnover is a slow process, it can be sped

ACCEPTED MANUSCRIPT

up or triggered altogether by hydromorphological changes, point sources of pollution, or any other human-induced changes.

Reviewing the species assemblages, the temporal changes going on suggested a competition among the native gammarids. This competition (Fig. 8.) suggests that *G. fossarum* appears as a competitor against both *G. balcanicus* and *G. roeselii*. Also, since previous experiments stated that *G. roeselii*, when alone, prefers coarse substrate beds (Kley et al. 2009), one can assume that among the studied species, *G. roeselii* is the weakest competitor being crowded out to occupy streams and rivers with fine substrate beds and lower altitude.

4. Conclusion

The distributional patterns of three gammarid species (*G. fossarum, G. balcanicus, G. roeselii*) were studied on two of the largest side stream tributaries of the Tisza River.

G. roeselii showed negative correlation to the altitude while its relation to the average depth, conductivity and temperature proved to be positive. The species also prefer fine bed substrate and thick plant coverage. *G. balcanicus* showed a strong negative correlation to fine bed substrate, temperature and plant coverage, while it prefers coarse substrate with its size ranging 2-20 cm. The *G. fossarum* in contrast showed a clear positive correlation to the altitude and coarse, rocky bed substrate, while there proved to be a negative correlation between temperature, average depth and gravel as bed substrate.

Using SOMs helped us to bring out distinctions in our data and enhance the differences making them easier to recognize and also, with their help we were able to model the relations of the species to habitat types non-existent among our samples. SOMs improved the correlations in our data that proved to be highly useful: besides their use to display

ACCEPTED MANUSCRIPT

complex data in a perspicuous way, they have other advantages in bringing out existing relationships in data otherwise difficult to detect.

SOMs are useful methods due to their ability of processing and displaying high dimensional data. However, through their facility of highlighting correlations, their utilization can be highly beneficial in case of deficiency, high level of noise or non-linear correlations in the data.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work was funded by the Bolyai János fellowship of the Hungarian Academy of Sciences. Authors would like to thank to Máté Bolbás for the extensive field and laboratory work.

References

Alhoniemi, E., Himberg, J., Parhankangas, J., Vesanto J., 2000. SOM Toolbox for Matlab.

Bollache, L., Devin, S., Wattier, R., Chovet, M., Beisel, J.N., Moreteau, J.C., Rigaud T., 2004. Rapid range extension of the Ponto-Caspian amphipod *Dikerogammarus villosus* in France: potential consequences. Arc.Hydro., 160(1), 57-66.

Borics, G., Várbíró, G., Padisák, J., 2013. Disturbance and stress - different meanings in ecological dynamics? Hydrobiologia, 711, 1-7.

Borza P., Czirok A., Deák Cs., Ficsór M., Horvai V., Horváth Zs., Juhász P., Kovács K., Szabó T., Vad Cs. F., 2011. Invasive mysids (Crustacea: Malacostraca: Mysida) in Hungary: distributions and dispersal mechanisms. North-West J. Zool., 7, 222-228.

Borza, P., Boda P., 2013. Range expansion of Ponto-Caspian mysids (Mysida, Mysidae) in the River Tisza: first record of *Paramysis lacustris* (Czerniavsky, 1882) for Hungary. Crustaceana, 86, 1316-1327.

Breiman, L., Friedman, J., Stone, C. J., Olshen, R. A., 1984. Classification and regression trees. CRC press.

ACCEPTED MANUSCRIPT

Bruijs, M.C.M., Kelleher, B., Van der Velde, G., De Vaate, A.B., 2001. Oxygen consumption, temperature and salinity tolerance of the invasive amphipod *Dikerogammarus villosus*: indicators of further dispersal via ballast water transport. Arc.Hydro., 152(4), 633-646.

Cereghino, R., Giraudel, J.L., Compin, A., 2001. Spatial analysis of stream invertebrates distribution in the Adour-Garonne drainage basin (France), using Kohonen self organizing maps. Ecol. Modell., 146, 167-180.

Chon, T.S., 2011. Self-Organizing Maps applied to ecological sciences. Ecol. Inform., 6, 50-61.

Chon, T.S., Park, Y.S., Moon, K.H., Cha, E.Z., 1996. Patternizing communities by using an artificial neural network. Ecol. Modell., 90, 69-78.

Clapcott, J.E., Young, R.G., Goodwin,E.O., Leathwick, J.R., 2010. Exploring the response of functional indicators of stream health to land-use gradients. Freshwater Biol., 55, 2181-2199.

Clarke, K. R., Ainsworth, A., 1993. A method of linking multivariate community structure to environmental variables. Marine Ecology-Progress, 92, 205-205.

Compin, A., Cereghino, R., 2007. Spatial patterns of macroinvertebrate functional feeding groups in streams in relation to physical variables and land-cover in southwestern France. Landsc. Ecol., 22, 1215-1225.

Davis, Steven J., 1987. Fluctuations in the pace of labor reallocation. Carnegie-Rochester Conference Series on Public Policy, North-Holland, 27, 335-402.

Devin, S., Beisel, J.N., 2007. Biological and ecological characteristics of invasive species: a gammarid study. Biol. Invasions, 9.1, 13-24.

Devin, S., Piscart, C., Beisel, J.N., Moreteau, J.C., 2004. Life history traits of the invader *Dikerogammarus villosus* (Crustacea: Amphipoda) in the Moselle River, France. Int. Rev. Hydro., 89(1), 21-34.

Dick, J.T.A., Elwood, R.W., Irvine, D.E., 1990. Displacement of the native Irish freshwater amphipod *Gammarus duebeni* by the introduced *Gammarus pulex*. Ir.Nat.J., 23, 313-316.

Dukowska, M., Grzybkowska, M., Kruk, A., Szczerkowska-Majchrzak, E., 2013. Food niche partitioning between perch and ruffe: Combined use of a self-organising map and the IndVal index for analysing fish diet. Ecol. Modell., 265, 221-229.

Giller, P.S., Malmqvist, B., 1998. The biology of streams and rivers. Oxford University Press.

Giraudel, J.L., Lek, S., 2001. A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. Ecol. Modell., 146, 329-339.

Goethals, P., Chon, T.S., Kim, D.H., Cho, W.S., 2013. Self-organizing map and species abundance distribution of stream benthic macroinvertebrates in revealing community patterns in different seasons. Ecol. Inform., 17, 14-29.

Grabowski, M., Mamos, T., 2011. Contact zones, range boundaries, and vertical distribution of three epigean gammarids (Amphipoda) in the Sudeten and Carpathian Mountains (Poland). Crustac. Int. J. Crustac. Res., 84.2, 153-168.

Gruner, H.E., 1966. Die Tierwelt Deutschlands und der angrenzenden Meeresteile nach ihren Merkmalen und nach ihrer Lebensweise. 53. Teil. Krebstiere oder Crustacea, V. Isopoda.-2. Lieferung. Jena: Fischer Verl.

Hammer, Ø., Harper, D.A.T., Ryan, P.D., 2001. PAST: Paleontological Statistics Software Package for Education and Data Analysis. Palaeontologia Electronica 4(1), 9pp.

Harper, D. A., 1999. Numerical palaeobiology: computer-based modelling and analysis of fossils and their distributions. John Wiley & Sons Inc.

ACCEPTED MANUSCRIPT

Henry, K.S., Danielopol, D.L., 1998. Oxygen dependent habitat selection in surface and hyporheic environments by *Gammarus roeseli* Gervais (Crustacea, Amphipoda): experimental evidence. Hydrobiologia, 390.1-3, 51-60.

Hering, D., Moog, O., Sandin, L., Verdonschot, P.F.M., 2004. Overview and application of the AQEM assessment system. Hydrobiologia, 516, 1-20.

Karaman, G., Pinkster, S., 1977. Freshwater Gammarus species from Europe, North Africa and adjacent regions of Asia (Crustacea-Amphipoda). Commissie voor de artis bibliotheek.

Kinzelbach, R., 1995. Neozoans in European waters—exemplifying the worldwide process of invasion and species mixing. Experientia, 51.5, 526-538.

Kley, A., Kinzler, W., Schank, Y., Mayer, G., Waloszek, D., Maier, G., 2009. Influence of substrate preference and complexity on co-existence of two non-native gammarideans (Crustacea: Amphipoda). Aquat.Ecol., 43(4), 1047-1059.

Kohonen, T,. 1998. The self-organizing map. Neurocomputing, 21(1), 1-6.

Kontschán, J., 2001b. *Proasellus pribenicensis* Flasarova, 1977. (Crustacea: Iopoda, Asellota), a magyar faunára új víziaszka a Cserehátból. Folia Entomol. Hung., 62, 319-320, Budapest (The first Hungarian record of *Proasellus pribenicensis* Flasarova, 1977.)

Kontschán, J., Muskó, I.B., Murányi, D., 2002. A felszíni vizekben elõforduló felemáslábú rákok (Crustacea: Amphipoda) rövid határozója és elõfordulásuk Magyarországon. Folia Hist.Mus.Matra. 26, 151-157. (Identification and checklist of amphipods (Crustacea: Amphipoda) of the surface waters of Hungary)

Lee, C. W., Jang, J. D., Jeong, K. S., Kim, D. K., & Joo, G. J. (2010). Patterning habitat preference of avifaunal assemblage on the Nakdong River estuary (South Korea) using self-organizing map. Ecological Informatics, 5(2), 89-96.

Li, F., Cai, Q., Qu, X., Tang, T., Wu, N., Fu, X., Duan, S., Jähnig, S.C., 2012. Characterizing macroinvertebrate communities across China: Large-scale implementation of a self-organizing map. Ecol. Indic., 23, 394-401.

MacNeil, C., Platvoet, D., 2005. The predatory impact of the freshwater invader *Dikerogammarus villosus* on native *Gammarus pulex* (Crustacea: Amphipoda); influences of differential microdistribution and food resources. J. Zool., 267(1), 31-38.

MacNeil, C., Dick, J.T., Elwood, R.W., Montgomery, W.I., 2001. Coexistence among native and introduced freshwater amphipods (Crustacea); habitat utilization patterns in littoral habitats. Arc.Hydro., 151(4), 591-607.

Mauchart, P., Bereczki, Cs., Ortmann-Ajkai, A., Csabai, Z., Szivák, I., 2014: Niche segregation between two closely related gammarids (Peracarida, Amphipoda) – Native vs. Naturalised non-native species. – Crustaceana 87(11–12): 1296-1314.

Ofenböck, T., Moog O., Gerritsen J., & Barbour M., 2004. A stressor-specific multi-metric approach for monitoring running waters in Austria using benthic-macroinvertebrates. Hydrobiologia, 516, 251-268.

Park, Y.S., Chon, T.S., Song, M.Y., Hwang, H.J., Kwak, I.S., Ji, C.W., Oh, Y.N., Youn, B.J., 2007. Self-organizing mapping of benthic macroinvertebrate communities implemented to community assessment and water quality evaluation. Ecol. Modell., 203, 18-25.

Pârvulescu, L., 2009. The epigean freshwater Malacostracans (Crustacea: Malacostraca) of the rivers in the Anina Mountains (SW Romania). Studia Universitatis Babes-Bolyai, Biologia, 54(2), 3-17.

Rohlf, F.J., Corti, M., 2000. Use of two-block partial least squares to study covariation in shape. Systematic Biology, 49, 740-753.

Sala, O.E., Chapin, F.S., Armesto, J.J., Berlow, E., Bloomfield, J., Dirzo, R., Huber-Sanwald, E., Huenneke, L.F., Jackson, R.B., Kinzig, A., Leemans, R., Lodge, D.M., Mooney,

ACCEPTED MANUSCRIPT

H.A., Oesterheld, M., Poff, N.L., Sykes, M.T., Walker, B.H., Walker, M., Wall, D.H., 2000. Global biodiversity scenarios for the year 2100. Science, 287(5459), 1770-1774.

Van Riel, M.C., Van der Velde, G., Rajagopal, S., Marguillier, S., Dehairs, F., Bij de Vaate, A., 2006. Trophic relationships in the Rhine food web during invasion and after establishment of the Ponto-Caspian invader *Dikerogammarus villosus*. In Living Rivers: Trends and Challenges in Science and Management. Springer Netherlands, 39-58.

Várbíró, G., Borics, G., Kiss, T.K., Szabó, K.E., Plenković-Moraj, A., Ács, É., 2007. Use of Kohonen Self Organizing Maps (SOM) for the characterization of benthic diatom associations of the River Danube and its tributaries. Arch. Hydrobiol., Supplementband. Large rivers, 17.3-4, 395-403.

Várbíró, G., Borics, G., Csányi, B., Fehér, G., Grigorszky, I., Kiss, K. T., Tóth, A., Ács, É., 2012. Improvement of the ecological water qualification system of rivers based on the first results of the Hungarian phytobenthos surveillance monitoring. Hydrobiologia, 695(1), 125-135.

Vesanto, J., 2000. Neural network tool for data mining: SOM toolbox. In Proceedings of symposium on tool environments and development methods for intelligent systems (TOOLMET2000), 184-196.

Vogt, J., Soille, P., Jager, A. D., 2007. A pan-European river and catchment database. JRC Reference Reports, EUR 229220 EN. European Commission.

Wesenberg-Lund, C., Storch, O., 1939.Biologie der süsswassertiere (Wirbellose Tiere). Springer; Wien: 1-817.

Wijnhoven, S., Van Riel, M.C., Van der Velde, G., 2003. Exotic and indigenous freshwater gammarid species: physiological tolerance to water temperature in relation to ionic content of the water. Aquat. Ecol., 37.2, 151-158.

Captions for tables and figures

Table 1. List of the sampling sites with their codes, names, exact locations and altitude

Table 2. The Pearson's correlation coefficients of the species abundance with abiotic parameters (numbers in bold indicate significant correlations at $p<0.05$).

Figure 1. The Flowchart of the SOM analysis.

Figure 2. The locations of the sampling sites (• sites with *Gammarus* sp. dominance; ▲ sites with *Asellus sp*. dominance; ■ monitoring sites of temporal investigations).

Figure 3. Ternary plots of the sampling sites based on *Gammarus* sp. abundance. Vertices of the triangle represent perfect dominance (100%) of the given *Gammarus* sp.

ACCEPTED MANUSCRIPT

Figure 4. The results of the Self Organizing Map (SOM) analysis. The darker shading of the component planes means higher abundance of the given species and higher variable values in the given SOM unit. Component planes arranged in a table by their positive (solid line) and negative (dotted line) correlation with the given species presented in the first column.

Figure 5. Optimum CART tree for the distribution of sites with *Gammarus* sp. dominances based on habitat composition factors. a. raw dataset, b. SOM dataset

Figure 6. Optimum CART tree for the distribution of site with *Gammarus* sp. dominances based on chemical factors. a. Raw dataset, b. SOM dataset

Figure 7. PCA score plot for sampling sites classified according to habitat composition. The convex hulls represent a morphospace constrained *Gammarus* sp. dominance. (□ *G. rosellii,* ● *G. balcanicus* , + - *G. fossarum*)

Figure 8. Temporal changes in the distribution of *Gammarus* sp. abundance in selected streams from 2005-2012.

Appendix 1. Raw dataset used for PCA and CART analysis

Appendix 2. SOM dataset used for PCA and CART analysis

Appendix 3. Confusion matrix, A table with the numbers of points in each given group (rows) that are assigned to the different groups (columns) by the classifier.

Appendix 4. Result of the discriminant analysis : Loading scores of the Raw and Som dataset's environmental variables to the first two canonical axes produces maximal and second to maximal separation between all groups. The axes are linear combinations of the original variables as in PCA, and eigenvalues indicate amount of variation explained by these axes.

ACCEPTED MANUSCRIPT

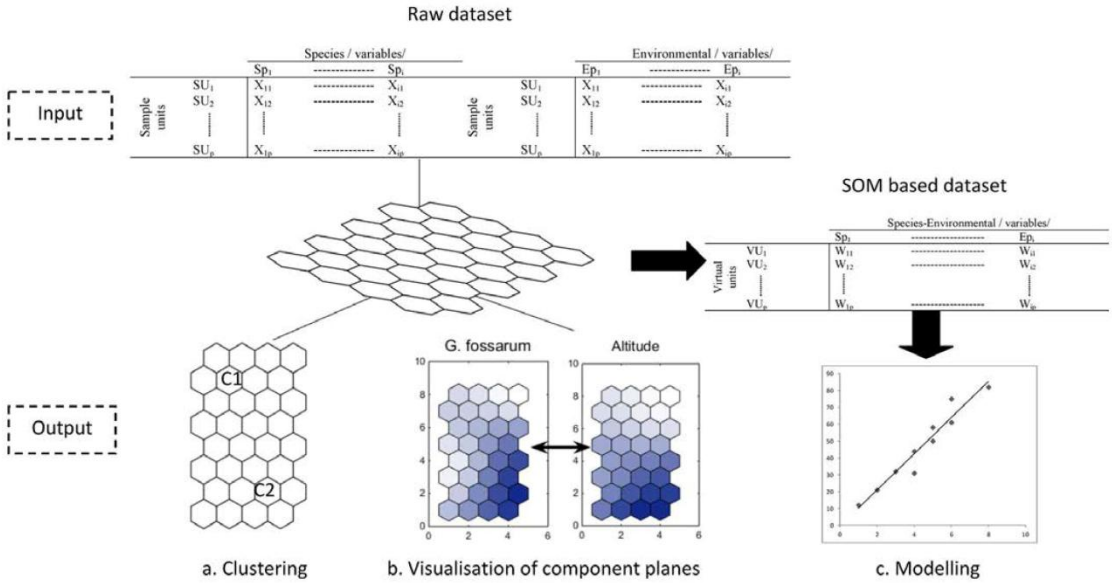


Figure 1

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

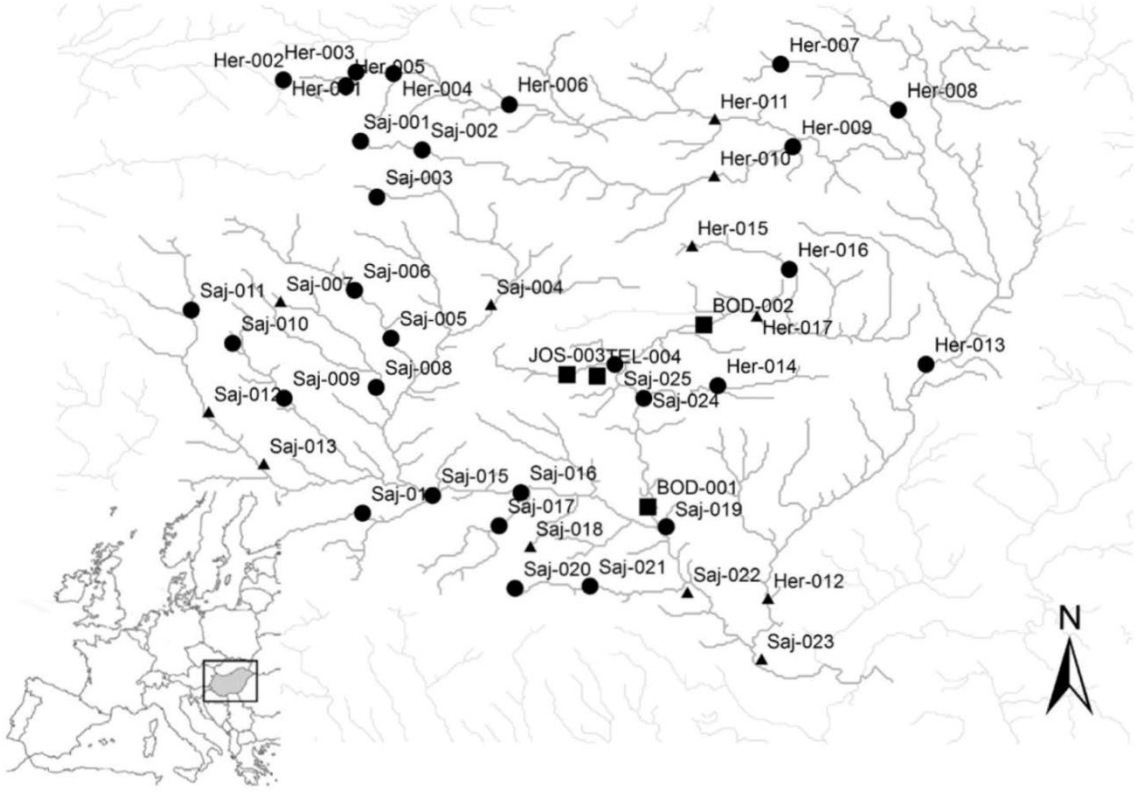


Figure 2

ACCEPTED

ACCEPTED MANUSCRIPT

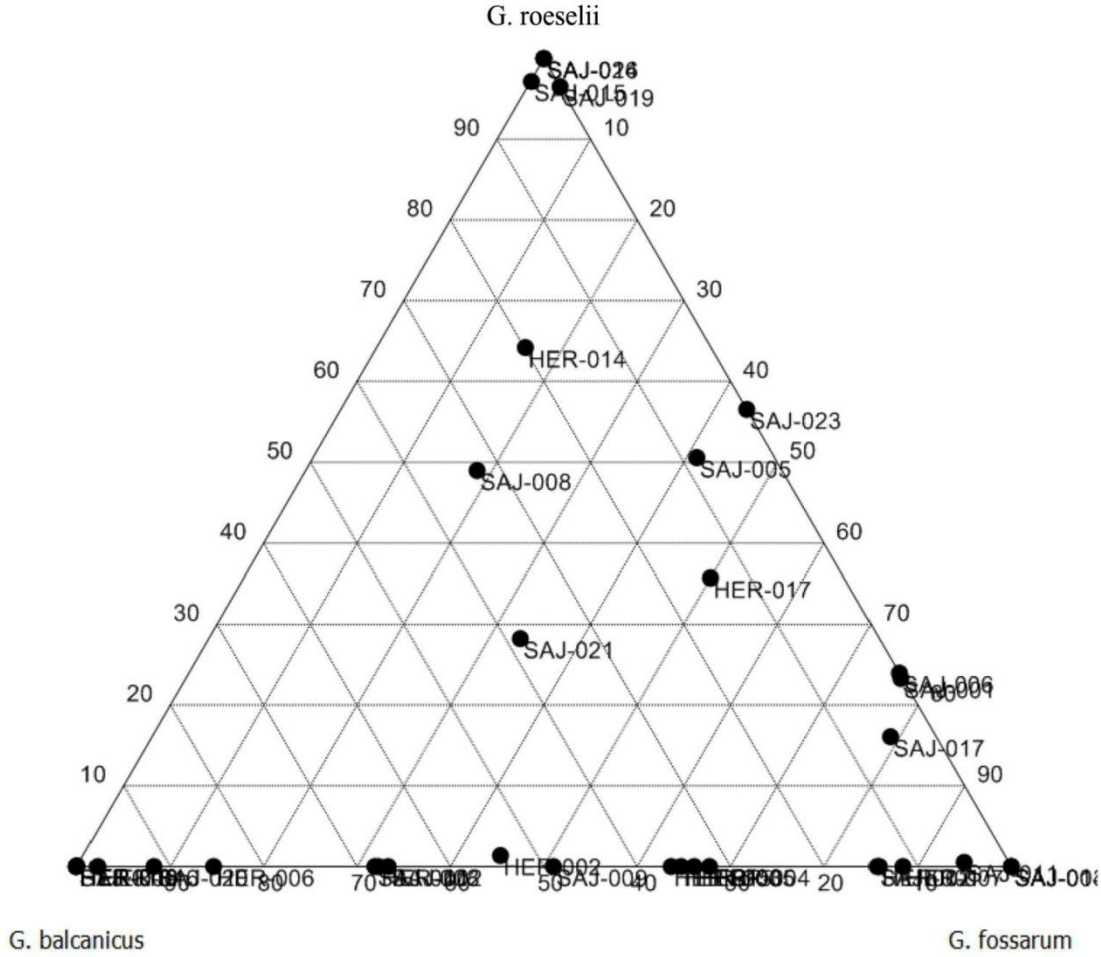


Figure 3

ACCEPTED MANUSCRIPT

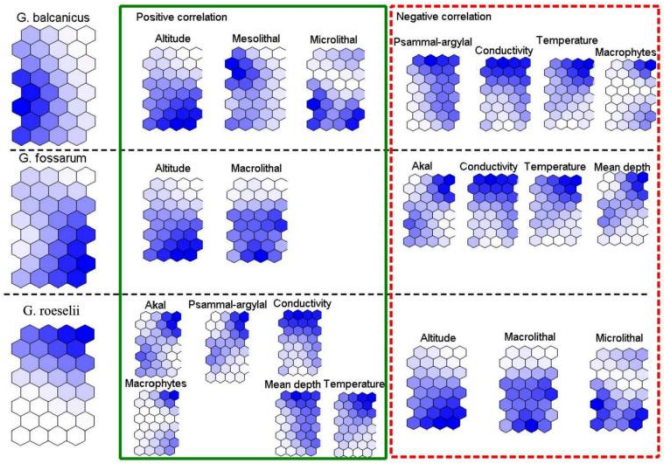


Figure 4

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

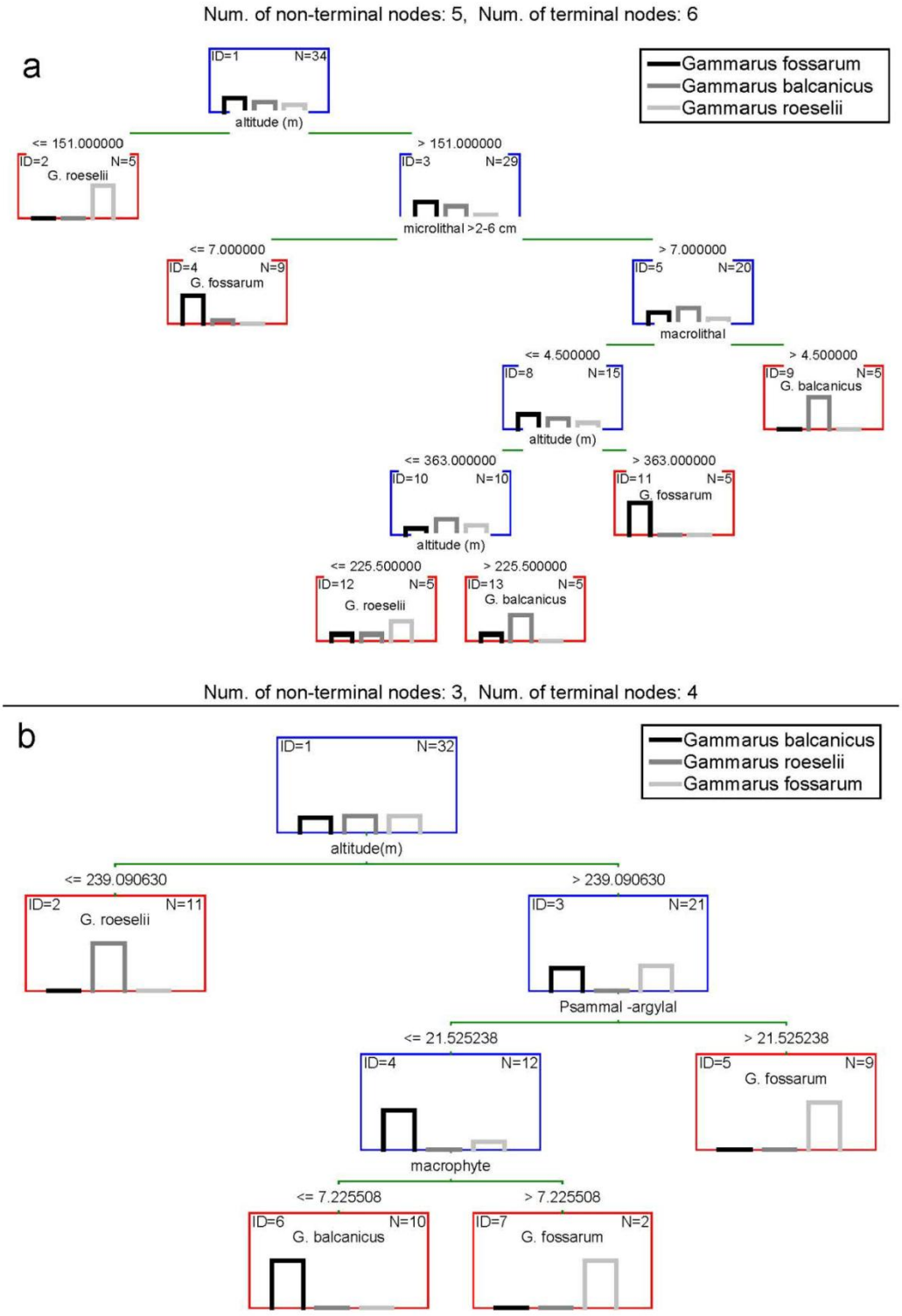


Figure 5

ACCEPTED MANUSCRIPT

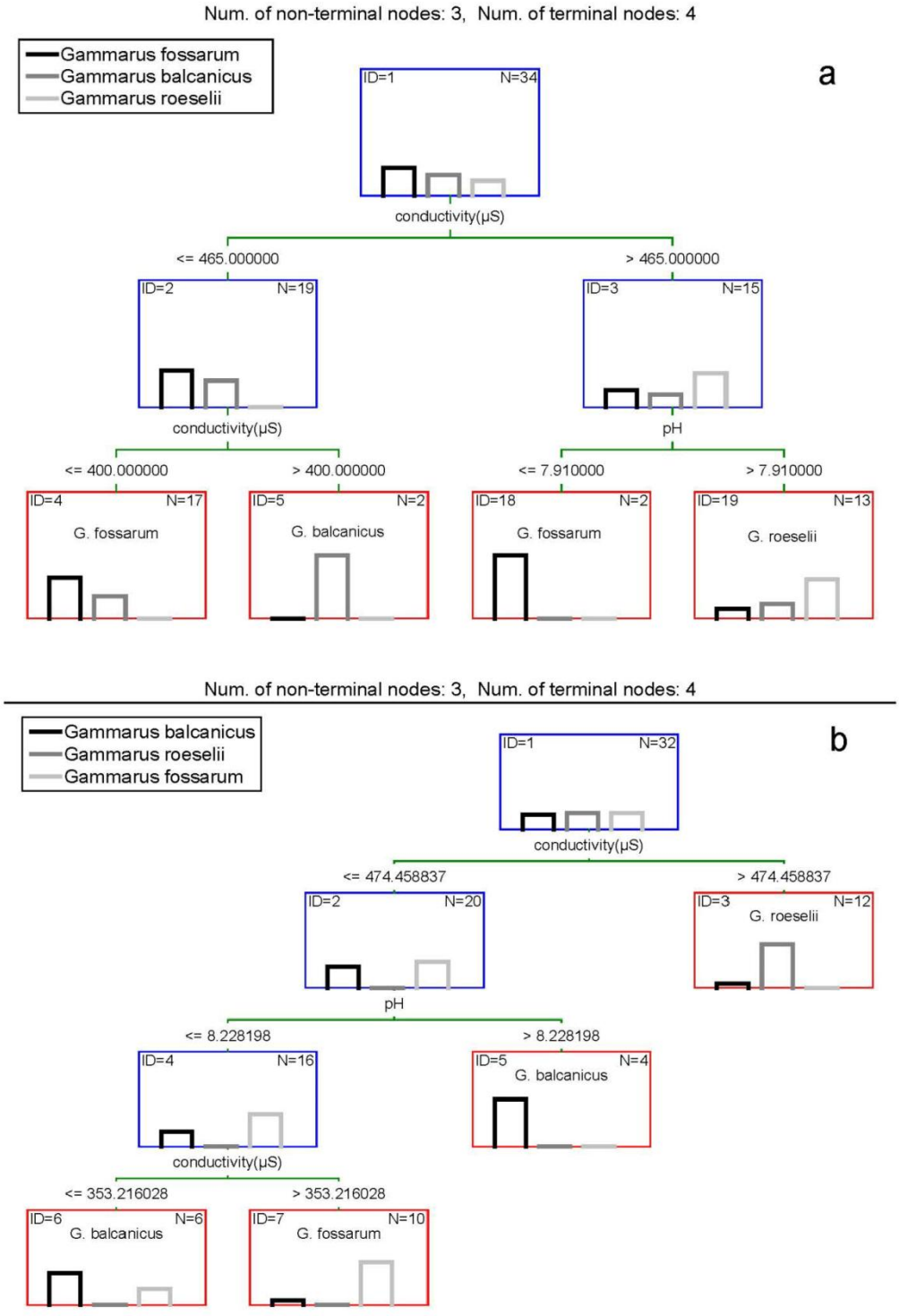


Figure 6

ACCEPTED MANUSCRIPT

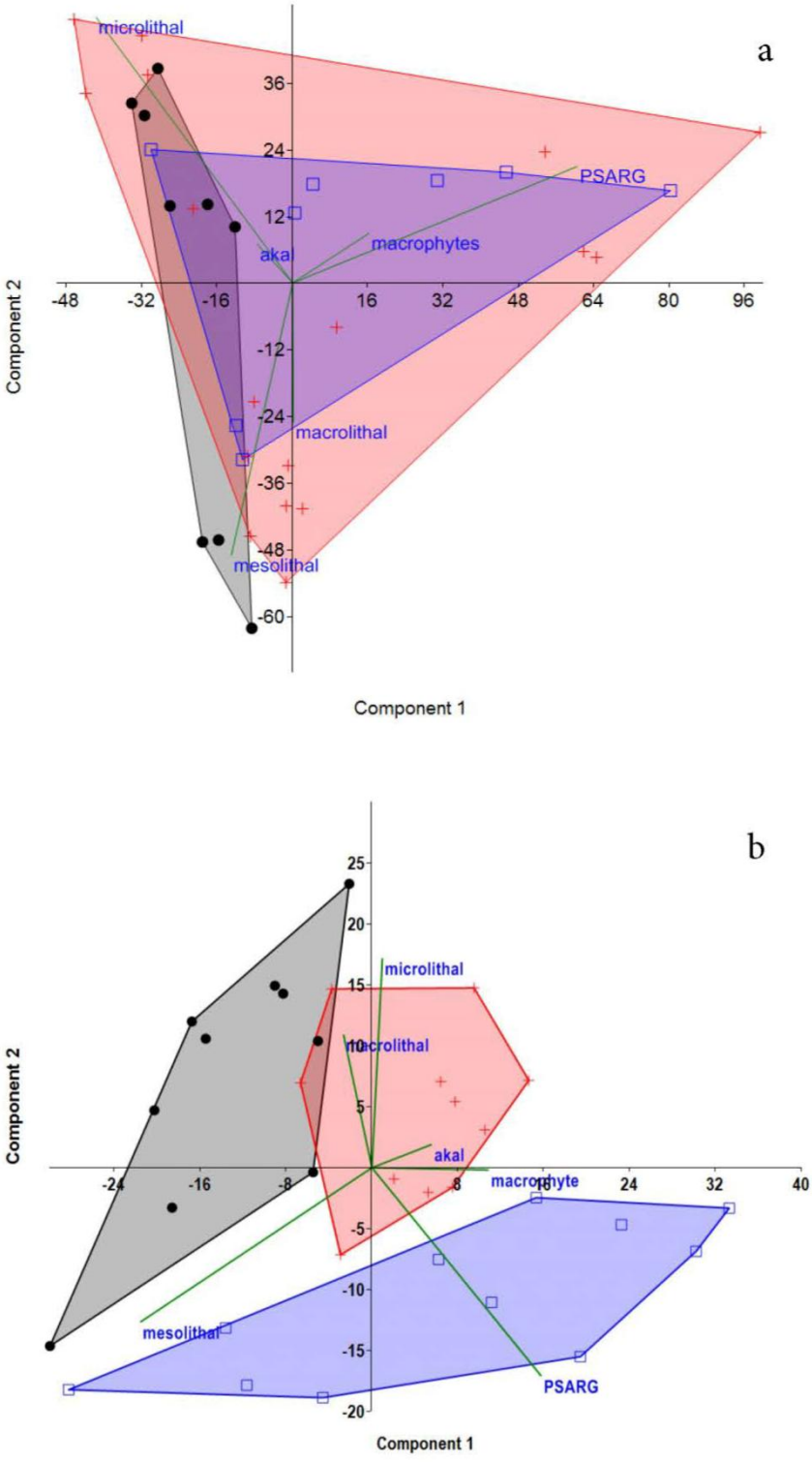


Figure 7

ACCEPTED MANUSCRIPT

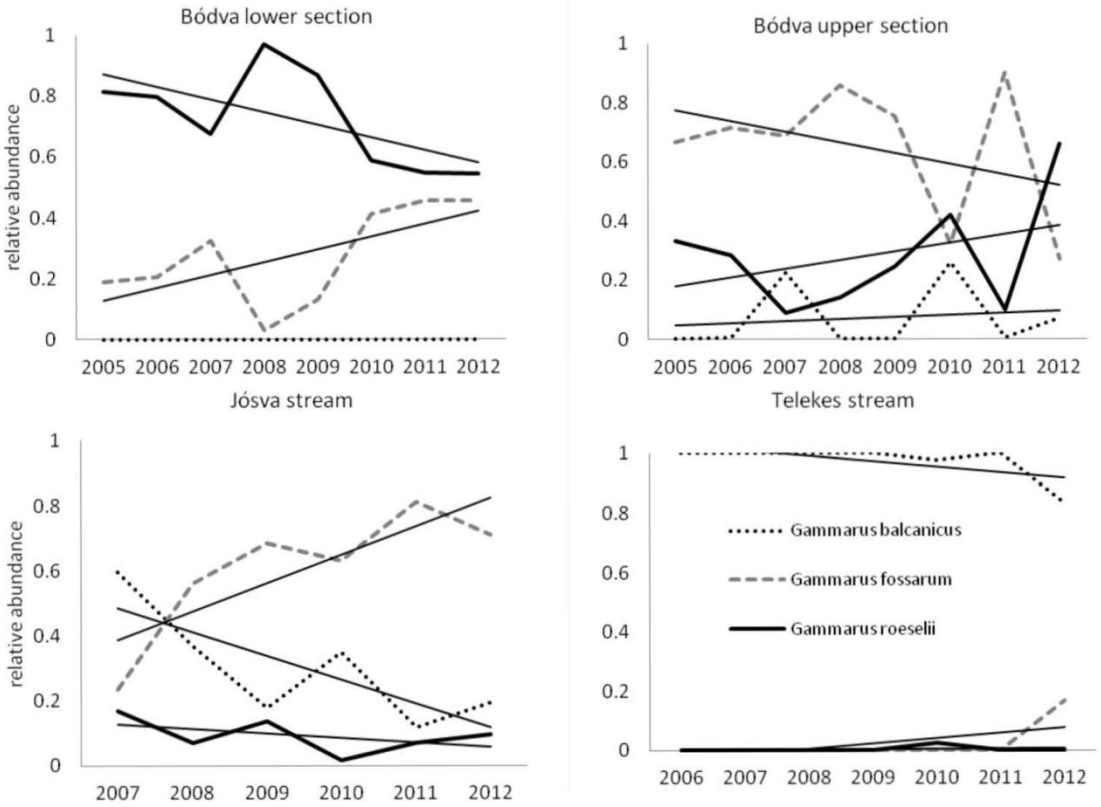


Figure 8

ACCEPTED MANUSCRIPT

Table 1.

Code	Stream name	Country	Settlement	Latitude (N)	Longitude (E)	Altitude (m)
HER-001	Hernád	SK	Hernádfő/Vikartovce	48°59'56.16	20°60'48.72"	903
HER-002	Hernád	SK	Hernádfő/Vikartovce	48°59'07.66"	20°07'01.68"	869
HER-003	Hernád	SK	Hernádfalu/Spišské Bystré	49°00'12.85"	20°14'35.30"	648
HER-004	Hernád	SK	Szepesvéghely/Hranovnica	49°00'00.62"	20°18'28.63"	601
HER-005	Bystra	SK	Hernádfalu/Spišské Bystré	48°59'42.08"	20°14'05.92"	767
HER-006	Hernád	SK	Szepessümeg/Smižany	48°56'55.88"	20°30'26.79"	483
HER-007	Svinka	SK	Frics/Fričovce	49°11'07.06"	20°58'54.15"	491
HER-008	Svinka	SK	Janó/Janov	48°56'23.96"	21°10'59.74"	293
HER-009	Gölnic/Hnilec	SK	Jekelfalva/Jaklovce	48°52'48.01"	21°00'02.95"	331
HER-010	Gölnic/Hnilec	SK	Nagykuncfalva/Helcmanovce	48°49'46.64"	20°51'31.03"	403
HER-011	Hernád	SK	Korompa/Krompachy	48°55'34.62"	20°51'06.54"	366
HER-012	Hernád	HU	Hernádkak	48°03'56.51"	20°57'58.18"	109
HER-013	Hernád	HU	Hidasnémeti	48°29'52.60"	21°13'49.10"	155
HER-014	Gölnic/Hnilec		Rakaca	48°27'37.10"	20°52'07.30"	187
HER-015	Gölnic/Hnilec	SK	Stósz/Štós	48°42'13.80"	20°49'23.40"	395
HER-016	Sajó	SK	Jászó/Jasov	48°39'50.00"	20°59'40.30"	246
HER-017	Sajó	SK	Zsarnó/Žarnov	48°34'54.30"	20°56'09.40"	181
SAJ-001	Murán	SK	Vernár	48°53'07.40"	20°14'42.00"	910
SAJ-002	Murán	SK	Sztracena/Stratená	48°52'04.07"	20°20'51.67"	809
SAJ-003	Turiec	SK	Sajóréde/Rejdová	48°47'41.23"	20°17'27.85"	620
SAJ-004	Turiec	SK	Szalóc/Slavec	48°36'04.15"	20°28'34.09"	245
SAJ-005	Kalosa	SK	Lice/Licince	48°32'30.63"	20°18'09.48"	208
SAJ-006	Kalosa	SK	Jolsva/Jelšava	48°37'07.86"	20°14'30.33"	253
SAJ-007	Rimava	SK	Szirk/Sirk	48°36'09.57"	20°70'28.06"	320

ACCEPTED MANUSCRIPT

SAJ-008	Rimava	SK	Otrokocs/Otročok	48°28'29.09"	20°16'36.58"	183
SAJ-009	Rimava	SK	Vámosbalog/Veľký Blh	48°26'06.58"	20°70'01.22"	193
SAJ-010	Hangony-	SK	Balogrussó/Hrušovo	48°30'50.92"	20°20'58.56"	293
SAJ-011	Hangony-	SK	Nyustya/Hnúšťa	48°35'27.65"	19°57'14.11"	300
SAJ-012	Sajó	SK	Cserencsény/Čerenčany	48°25'10.77"	19°58'48.93"	215
SAJ-013	Bán-patak	SK	Rimapálfalva/Pavlovce	48°19'37.85"	20°40'40.16"	183
SAJ-014	Tardona-patak	HU	Ózd	48°13'56.19"	20°15'30.48"	169
SAJ-015	Bódva	HU	Sajónémeti	48°16'18.78"	20°22'33.06"	147
SAJ-016	Garadna-patak	HU	Vadna	48°16'26.23"	20°33'33.56"	138
SAJ-017	Szinva-patak	HU	Bánhorváti	48°12'54.49"	20°29'23.86"	174
SAJ-018	Szinva-patak	HU	Tardona	48°12'01.99"	20°34'00.01"	230
SAJ-019	Sajó	HU	Boldva	48°12'31.31"	20°46'29.07"	118
SAJ-020	Rakaca	HU	Lillafüred	48°06'26.00"	20°31'06.10"	590
SAJ-021	Jósva-patak	HU	Alsó-Hámor	48°40'55.91"	20°49'10.69"	243
SAJ-022	Rakaca	HU	Miskolc	48°60'10.72"	20°49'38.03"	114
SAJ-023	Bódva	HU	Köröm	47°59'08.00"	20°56'42.50"	96
SAJ-024	Bódva	HU	Szalonna	48°26'12.98"	20°44'19.54"	142
SAJ-025	Bódva	HU	Perkupa	48°29'48.07"	20°41'16.18"	155

ACCEPTED MANUSCRIPT

Table 2.

	Raw data			SOM		
	G. roeselii	*G. balcanicus*	*G. fossarum*	*G. roeselii*	*G. balcanicus*	*G. fossarum*
altitude (m)	**-0.48**	-0.01	**0.47**	**-0.87**	**0.25**	**0.77**
width of floodplain	0.08	0.29	**-0.36**	0.09	**0.48**	**-0.63**
width at high water	0.31	0.14	**-0.43**	**0.61**	0.01	**-0.74**
depth at high water	0.30	0.12	**-0.40**	**0.64**	-0.07	**-0.69**
width (actual)	**0.36**	0.01	**-0.35**	**0.63**	-0.22	**-0.52**
mean depth (actual)	**0.39**	0.01	**-0.40**	**0.66**	-0.21	**-0.58**
depth (max.)	0.31	-0.00	-0.31	**0.66**	-0.31	**-0.46**
temperature (°C)	**0.54**	-0.15	**-0.37**	**0.94**	**-0.47**	**-0.62**
pH	-0.02	0.09	-0.05	0.22	0.20	**-0.49**
O2 (mg/l)	-0.25	0.08	0.18	**-0.58**	**0.36**	0.31
conductivity (µS)	**0.42**	-0.29	-0.08	**0.93**	**-0.63**	**-0.42**
macrolithal	-0.19	0.05	0.14	**-0.79**	0.32	**0.58**
mesolithal >6-20 cm	-0.16	0.16	-0.04	-0.16	**0.44**	-0.30
microlithal >2-6 cm	-0.18	0.19	0.00	**-0.55**	**0.45**	0.20
akal >2mm-2cm	0.29	0.12	**-0.40**	**0.51**	-0.02	**-0.59**
PSARG	0.27	**-0.34**	0.09	**0.75**	**-0.85**	0.03
macrophyte	0.29	-0.30	0.04	**0.53**	**-0.58**	0.02

ACCEPTED MANUSCRIPT

Electronic appendix 1.

Use of Self Organising Maps in modelling the distribution patterns of gammarids (Crustacea: Amphipoda) Eszter Á. Krasznai, Pál Boda, András Csercsa, Márk Ficsór, Gábor Várbíró

Raw dataset

	G. roesellii	G. balcanicus	G. fossarum	Dominant taxa	altitude (m)	width of floodplain	width at high water	depth at high water	width (actual)	mean depth (actual)	depth (max.)	temperature (°C)	pH	O2 (mg/l)	conductivity (µS)	macrolithal (%)	mesolithal >6-20 cm (%)	microlithal >2-6 cm (%)	akal >2mm-2cm (%)	PSARG (%)	macrophyte cover(%)
HER-001	0.00	0.13	0.82	G. fossarum	903	20	5	0.3	1.0	0.1	0.3	15.4	7.8	7.2	276	0	10	80	10	0	4
HER-002	0.01	0.54	0.44	G. balcanicus	869	2	1	0.1	0.5	0.1	0.2	15.4	7.8	8.1	165	5	70	15	10	0	2
HER-003	0.00	0.28	0.52	G. fossarum	648	20	10	1.0	2.0	0.3	1.0	19.8	8.1	8.2	240	0	10	0	10	80	0
HER-004	0.00	0.32	0.66	G. fossarum	601	5	5	1.0	5.0	0.4	1.0	20.9	8.2	7.8	252	60	10	0	30	0	10
HER-005	0.00	0.30	0.59	G. fossarum	767	5	5	0.5	5.0	0.1	0.5	20.5	8.4	8.1	232	80	10	0	10	0	15
HER-006	0.00	0.84	0.14	G. balcanicus	483	15	12	2.0	8.0	0.7	1.0	20.0	8.7	8.3	453	40	60	0	0	0	0
HER-007	0.00	0.11	0.86	G. fossarum	491	7	7	2.0	4.0	0.3	0.4	18.3	8.4	8.7	797	0	0	20	0	80	0
HER-008	0.00	0.66	0.31	G. balcanicus	293	10	10	1.5	4.0	0.2	0.4	22.0	8.6	8.8	702	0	10	70	10	10	0
HER-009	0.00	1.00	0.00	G. balcanicus	331	20	15	2.0	4.0	0.3	1.0	22.8	8.2	7.8	237	0	0	70	20	10	10
HER-013	0.00	1.00	0.00	G. balcanicus	155	200	80	10.0	30.0	1.2	3.0	23.7	8.3	7.7	494	5	5	40	50	0	0
HER-014	0.64	0.20	0.16	G. roesellii	187	9	9	1.0	2.0	0.2	0.5	24.4	8.2	7.4	701	4	0	20	4	70	0
HER-015	0.00	0.34	0.59	G. fossarum	395	10	8	3.0	3.0	0.1	1.5	17.3	8.0	9.4	146	0	0	70	20	10	0
HER-016	0.00	0.97	0.02	G. balcanicus	246	50	15	2.5	5.0	0.5	1.8	19.6	8.2	9.0	238	10	0	70	15	5	0
HER-017	0.36	0.14	0.50	G. fossarum	181	25	10	3.0	5.0	0.5	1.5	23.3	8.0	7.5	368	20	0	0	0	80	0
SAJ-001	0.23	0.00	0.76	G. fossarum	910	12	8	0.5	4.0	0.2	0.8	16.0	8.0	9.2	295	90	4	4	1	0	0
SAJ-002	0.00	0.13	0.80	G. fossarum	809	15	12	1.0	8.0	0.3	1.0	22.0	8.6	9.5	321	0	0	80	10	10	15
SAJ-003	0.00	0.00	1.00	G. fossarum	620	3	1	0.3	0.5	0.1	0.1	21.2	7.9	8.3	116	0	0	90	10	0	10
SAJ-005	0.48	0.08	0.39	G. roesellii	208	20	12	2.0	7.0	0.5	1.0	24.4	8.4	8.3	477	0	50	20	20	10	0
SAJ-006	0.24	0.00	0.76	G. fossarum	253	10	10	2.5	8.0	1.0	1.2	21.4	8.4	8.3	389	0	0	0	0	10	0

ACCEPTED MANUSCRIPT

SAJ-008	0.46	0.31	0.17	G. roesellii	183	10	10	2.5	5.0	0.8	1.6	22.0	8.1	8.0	510	0	60	20	10	10	10
SAJ-009	0.00	0.45	0.47	G. fossarum	193	8	8	1.5	4.0	0.2	0.4	22.8	7.8	5.5	478	10	70	0	10	10	10
SAJ-010	0.00	0.64	0.30	G. balcanicus	293	8	4	1.0	2.0	0.2	0.4	21.3	8.1	8.6	313	0	70	10	20	0	10
SAJ-011	0.01	0.05	0.92	G. fossarum	300	20	12	0.5	10.0	0.5	1.2	22.8	8.3	8.6	313	0	60	20	10	10	15
SAJ-012	0.00	0.50	0.25	G. balcanicus	215	12	10	2.0	10.0	0.8	1.2	24.6	8.1	7.9	240	10	40	20	20	10	10
SAJ-013	0.00	1.00	0.00	G. balcanicus	183	25	12	0.8	10.0	0.8	1.2	26.4	8.0	7.2	270	0	10	40	40	10	0
SAJ-014	0.00	0.00	1.00	G. fossarum	169	6	2	0.6	1.0	0.2	0.2	25.7	8.4	7.3	866	0	0	0	0	100	90
SAJ-015	0.97	0.03	0.00	G. roesellii	147	10	10	2.0	3.0	0.7	0.7	26.3	8.0	5.0	812	0	0	10	60	30	80
SAJ-016	1.00	0.00	0.00	G. roesellii	138	60	50	7.0	30.0	1.0	4.0	22.8	8.1	7.6	491	0	0	0	0	100	0
SAJ-017	0.14	0.04	0.70	G. fossarum	174	10	10	2.0	5.0	0.5	1.7	21.1	7.8	5.9	755	0	20	50	20	10	10
SAJ-019	0.97	0.00	0.03	G. roesellii	118	50	30	2.0	20.0	1.0	1.2	24.8	8.2	7.9	511	0	20	70	0	10	10
SAJ-020	0.00	0.89	0.08	G. balcanicus	590	1	1	1.0	0.5	0.1	0.1	17.4	8.4	8.7	411	30	0	40	20	10	10
SAJ-021	0.28	0.38	0.33	G. balcanicus	243	8	8	1.5	2.0	0.1	0.4	20.4	8.7	9.0	518	0	70	10	10	10	0
SAJ-023	0.57	0.00	0.43	G. roesellii	96	35	30	6.0	20.0	1.0	3.0	26.2	8.7	10.2	661	0	0	20	60	20	0
SAJ-024	1.00	0.00	0.00	G. roesellii	142	15	15	2.0	4.0	0.2	0.4	30.0	8.1	7.3	496	0	0	20	70	10	60

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Appendix 2.

Use of Self Organising Maps in modelling the distribution patterns of gammarids (Crustacea: Amphipoda) Eszter Á. Krasznai, Pál Boda, András Csercsa, Márk Ficsór, Gábor Várbíró

SOM dataset

	G. roesellii	G. balcanicus	G. fossarum	Dominant taxa	altitude (m)	width of floodplain	width at high water	depth at high water	width (actual)	mean depth (actual)	depth (max.)	temperature (°C)	pH	O2 (mg/l)	conductivity (µS)	macrolithal (%)	mesolithal >6-20 cm (%)	microlithal >2-6 cm (%)	akal >2mm-2cm (%)	PSARG (%)	macrophyte cover(%)
VU1	0.00	0.67	0.29	G. balcanicus	401	21	11	1.6	4.6	0.3	0.7	20.7	8.2	8.3	384	8	33	36	18	5	5
VU2	0.00	0.72	0.24	G. balcanicus	388	31	15	2.1	6.4	0.4	1.0	21.1	8.2	8.0	349	11	25	35	20	8	5
VU3	0.00	0.87	0.10	G. balcanicus	332	46	20	2.8	8.6	0.5	1.2	21.7	8.2	8.1	349	11	16	42	24	8	4
VU4	0.02	0.80	0.13	G. balcanicus	338	39	18	2.6	8.5	0.6	1.2	21.7	8.3	8.1	354	14	24	32	20	10	4
VU5	0.10	0.70	0.16	G. balcanicus	313	30	16	2.4	7.3	0.5	1.1	21.4	8.4	8.3	406	13	37	26	16	8	3
VU6	0.30	0.41	0.25	G. balcanicus	248	14	10	2.0	4.7	0.4	0.9	21.6	8.4	8.4	489	5	54	16	11	14	3
VU7	0.42	0.33	0.22	G. roesellii	208	10	9	2.0	3.9	0.5	1.1	21.8	8.3	8.3	531	1	55	17	10	17	5
VU8	0.52	0.26	0.19	G. roesellii	193	10	10	1.8	3.7	0.4	1.0	22.9	8.3	7.9	588	2	36	19	9	35	5
VU9	0.00	0.51	0.44	G. balcanicus	501	13	8	1.4	4.1	0.2	0.7	20.0	8.1	8.1	318	18	31	27	16	7	7
VU10	0.00	0.46	0.47	G. fossarum	477	17	10	1.5	4.8	0.3	0.8	20.6	8.1	7.9	324	17	26	26	16	15	8
VU11	0.01	0.70	0.24	G. balcanicus	375	33	16	2.3	7.3	0.5	1.1	21.5	8.2	7.9	339	13	23	30	19	14	5
VU12	0.06	0.53	0.32	G. balcanicus	362	25	13	2.1	7.2	0.5	1.1	21.8	8.2	8.0	351	14	27	21	15	21	5
VU13	0.14	0.55	0.24	G. balcanicus	307	23	13	2.2	7.0	0.5	1.1	21.9	8.3	8.2	398	13	39	19	14	14	3
VU14	0.37	0.27	0.32	G. roesellii	231	17	12	2.2	6.4	0.5	1.1	22.9	8.4	8.3	486	6	39	16	15	23	4
VU15	0.48	0.27	0.23	G. roesellii	204	12	11	1.9	4.4	0.4	0.9	22.7	8.3	8.1	558	2	41	17	11	28	5
VU16	0.67	0.16	0.15	G. roesellii	177	17	14	1.9	6.1	0.4	1.0	24.3	8.2	7.6	619	2	16	21	14	46	11
VU17	0.00	0.32	0.62	G. fossarum	563	10	7	1.2	4.1	0.2	0.8	20.1	8.2	8.1	302	23	20	29	14	13	11

ACCEPTED MANUSCRIPT

VU18	0.01	0.22	0.71	G. fossarum	527	12	7	1.1	4.2	0.3	0.7	20.5	8.2	8.0	366	14	17	31	11	25	14
VU19	0.03	0.35	0.53	G. fossarum	449	18	10	1.6	5.2	0.4	0.9	21.0	8.1	7.9	355	13	22	23	13	26	9
VU20	0.15	0.19	0.58	G. fossarum	393	16	10	1.8	5.9	0.5	1.1	21.3	8.1	7.9	405	16	17	18	11	28	6
VU21	0.24	0.26	0.43	G. fossarum	293	19	12	2.3	7.2	0.6	1.3	22.5	8.2	8.1	416	13	26	15	14	26	3
VU22	0.49	0.10	0.38	G. roesellii	208	23	16	2.9	9.4	0.6	1.5	24.0	8.3	8.2	503	7	20	17	22	30	6
VU23	0.64	0.12	0.22	G. roesellii	177	22	17	2.5	8.6	0.5	1.2	24.5	8.3	7.8	566	2	21	20	21	35	12
VU24	0.86	0.05	0.09	G. roesellii	150	28	22	2.9	11.6	0.6	1.4	25.5	8.2	7.3	593	1	8	23	27	41	26
VU25	0.00	0.18	0.77	G. fossarum	568	10	7	1.0	4.3	0.2	0.7	20.5	8.2	8.3	364	14	13	39	11	22	16
VU26	0.02	0.10	0.84	G. fossarum	523	11	7	1.0	4.0	0.3	0.6	20.6	8.2	8.1	428	7	13	39	8	30	19
VU27	0.07	0.13	0.75	G. fossarum	480	13	8	1.2	4.5	0.4	0.8	20.6	8.1	8.0	422	12	14	28	8	29	13
VU28	0.19	0.07	0.70	G. fossarum	411	14	10	1.8	5.7	0.5	1.2	20.7	8.1	7.9	455	19	10	21	9	25	7
VU29	0.31	0.09	0.55	G. fossarum	302	19	13	2.6	7.8	0.6	1.4	22.3	8.2	8.1	463	16	13	15	15	28	3
VU30	0.55	0.05	0.39	G. roesellii	202	26	19	3.4	11.4	0.7	1.8	24.2	8.3	8.3	530	8	11	18	28	30	9
VU31	0.77	0.04	0.18	G. roesellii	154	29	22	3.3	12.6	0.7	1.6	25.4	8.2	7.7	567	2	11	22	31	34	22
VU32	0.93	0.02	0.05	G. roesellii	139	33	25	3.3	13.9	0.7	1.6	25.9	8.1	7.2	581	0	6	24	32	37	33

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Appendix 3.

Use of Self Organising Maps in modelling the distribution patterns of gammarids (Crustacea: Amphipoda) Eszter Á. Krasznai, Pál Boda, András Csercsa, Márk Ficsór, Gábor Várbíró

Confusion matrix
A table with the numbers of points in each given group (rows) that are assigned to the different groups (columns) by the classifier.

Original

	Raw dataset 80.24 % of correctly classified				SOM dataset 100 % of correctly classified			
	G. fossarum	G. balcanicus	G. roesellii	Total	G. balcanicus	G. roesellii	G. fossarum	Total
G. fossarum	13	2	0	15	10	0	0	10
G. balcanicus	1	10	0	11	0	11	0	11
G. roesellii	0	1	7	8	0	0	11	11
Total	14	13	7	34	10	11	11	32

Jackknifed

	Raw dataset 38.24 % of correctly classified				SOM dataset 90.63 % of correctly classified			
	G. fossarum	G. balcanicus	G. roesellii	Total	G. balcanicus	G. roesellii	G. fossarum	Total
G. fossarum	7	4	4	15	8	1	1	10
G. balcanicus	2	4	5	11	0	11	0	11
G. roesellii	2	4	2	8	1	0	10	11
Total	11	12	11	34	9	12	11	32

ACCEPTED MANUSCRIPT

Appendix 4.

Use of Self Organising Maps in modelling the distribution patterns of gammarids (Crustacea: Amphipoda) Eszter Á. Krasznai, Pál Boda, András Csercsa, Márk Ficsór, Gábor Várbíró

Result of the discriminant analysis : Loading scores of the Raw and Som dataset's environmental variables to the first two canonical axes produces maximal and second to maximal separation between all groups. The axes are linear combinations of the original variables as in PCA, and eigenvalues indicate amount of variation explained by these axes.

	Raw dataset		SOM dataset	
	Axis 1	Axis 2	Axis 1	Axis 2
altitude (m)	99.04	-2.44	21.42	-28.94
width of floodplain	-4.62	-6.32	0.52	1.87
width at high water	-3.90	-0.96	-0.34	1.00
depth at high water	-0.51	-0.07	-0.06	0.13
width (actual)	-2.02	0.19	-0.23	0.36
mean depth (actual)	-0.10	0.01	-0.01	0.02
depth (max.)	-0.20	0.08	-0.03	0.03
temperature (°C)	-1.28	0.57	-0.31	0.25
pH	-0.03	-0.05	0.00	0.02
O2 (mg/l)	0.06	-0.19	0.03	0.00
conductivity (µS)	-52.67	46.75	-20.47	9.37
macrolithal (%)	4.92	0.49	1.08	-1.13
mesolithal >6-20 cm (%)	-1.47	-7.24	0.51	1.95
microlithal >2-6 cm (%)	1.18	-4.30	0.97	-0.15
akal >2mm-2cm (%)	-5.47	-1.01	-0.42	1.16
PSARG (%)	-0.96	10.44	-2.17	-0.93
macrophyte cover(%)	-1.98	5.18	-0.82	-0.43

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Highlights

- We present the spatial distributional patterns of three Gammarus species.
- We test the use of Self Organizing Map(SOM) in distribution modelling.
- We compare the raw and SOM-based datasets in describing distributional patterns.

ACCEPTED MANUSCRIPT