# Effect of field sampling design on variation partitioning in a dendritic stream network

Péter Sály[a,*], Tibor Erős[a]

[a]MTA Centre for Ecological Research, Balaton Limnological Institute, H-8237 Tihany, Klebelsberg Kuno u. 3., Hungary

e-mail addresses: saly.peter@okologia.mta.hu (PS), eros.tibor@okologia.mta.hu (TE)

*corresponding author

Abstract

Variation partitioning is one of the most frequently used method to infer the importance of environmental (niche based) and spatial (dispersal) processes in metacommunity structuring. However, the reliability of the method in predicting the role of the major structuring forces is less known. We studied the effect of field sampling design on the result of variation partitioning of fish assemblages in a stream network. Along with four different sample sizes, a simple random sampling from a total of 115 stream segments (sampling objects) was applied in 400 iterations, and community variation of each random sample was partitioned into four fractions: pure environmentally (landscape variables) explained, pure spatially (MEM eigenvectors) explained, jointly explained by environment and space, and unexplained variance. Results were highly sensitive to sample size. Even at a given sample size, estimated variance fractions had remarkable random fluctuation, which can lead to inconsistent results on the relative importance of environmental and spatial variables on the structuring of metacommunities. Interestingly, all the four variance fractions correlated better with the number of the selected spatial variables than with any design properties. Sampling interval proved to be a fundamentally influential sampling design property because it affected the number of the selected spatial variables. Our findings suggest that the effect of sampling design on variation partitioning is related to the ability of the eigenvectors to model complex spatial patterns. Hence, properties of the sampling design should be more intensively considered in metacommunity studies.

Key words: metacommunity; fish assemblage; species distribution modelling; network topology; Moran's eigenvector maps (MEM); relative importance of space and environment

## 1. INTRODUCTION

### *1.1. Properties of field sampling design*

Properties of field sampling design set the window through which ecologists study the spatial and temporal distribution of organisms and the determining factors affecting distribution patterns. The frame of this window is the spatio-temporal scale of the study, which has three elements in ecological sampling theory. Focusing only on the spatial aspect of the scale, the grain size is the size of the sampling units (e.g., quadrates); the sampling interval is the average distance between the neighbouring sampling units; and the extent is the total area included in the investigation (Wiens 1989; Legendre & Legendre 2012 p786). Sample size, another property of sampling design, is the total number of sampling units in the sample, and it is a simple measure of the sampling effort. An additional property is the topology of the sampling units. Topology describes the geometry by which the sampling units are ecologically connected to each other. When sampling units considered being connected, researchers assume that material and individuals can move from one sampling unit to the other one (e.g., Peterson et al. 2013).

### *1.2. Variation partitioning*

Ecologists try to reveal the mechanisms controlling the distribution of organisms by investigating their spatial distributional patterns. One of the most frequently used statistical methods for quantifying different sources of variation of communities is variation partitioning (or variance partitioning), which was introduced into the ecological methodology by Borcard et al. (1992). In a classical approach, this method uses a sites-by-species community matrix as response data, and a sites-by-environmental variables matrix and a sites-by-spatial variables

matrix as explanatory data to decompose additively the total variation of the response data into four variance fractions/proportions by fitting canonical ordination models (canonical correspondence analysis [CCA] or redundancy analysis [RDA]) on the data. One of the variance fractions is the variation explained exclusively by the studied environmental variables, denoted by [a] in the original paper of Borcard et al. (1992). This fraction is usually considered to reflect the importance of environmental effects which could not be associated to spatial co-variation. Another variance fraction ([c]) is explained purely by the spatial variables, and gives estimation on community variation that has no relationship with the environmental variables included into the environmental data matrix. However, depending on the elaboration of the study, there is a possibility that this fraction incorporates some variation that would be explainable by a latent, unmeasured environmental variable. A third variance fraction ([b]) is explained jointly by the studied environmental and spatial variables. In this case the effects of environmental and spatial factors on community structure cannot be disentangled. The last fourth variance fraction is the unexplained residual variation [d].

Peres-Neto et al. (2006) improved variation partitioning by introducing the adjusted redundancy statistic or adjusted coefficient of multiple determination ($R^2_{adj}$). The adjusted redundancy statistic expresses the unbiased form of the variance fractions/proportions which is controlled for the number of explanatory variables in the model and the sample size.

Since its introduction, variation partitioning has become a fundamental method to infer the measure and importance of environment- and space-related mechanisms structuring communities, especially in the field of metacommunity researches. Results mirror that this measure and importance tend to vary according to the studied group of organism (e.g.,

88 Cottenie 2005; Beisner et al. 2006; Marzin et al. 2013), ecological data type (e.g., Cushman &
89 McGarigal 2004; Hoeinghaus et al. 2007; Sály et al. 2011 ), ecosystem type (e.g., Cottenie
90 2005; Heino et al. 2015; Soininen & Weckström 2009), spatial scale of the study (e.g.,
91 Cushman & McGarigal 2004; Declerck et al. 2011; Heino et al. 2015; Mykrä et al. 2007),
92 study region (e.g., Cottenie 2005) and study years (e.g., Mesquita et al. 2006).
93
94 *1.3. Relationship of sampling design and variation partitioning*
95 Differences in the study design are among the most important factors that could lead to
96 apparently inconsistent results of variation partitioning studies. In fact, Dray et al. (2012
97 p262–263) explicitly warned that sampling design introduces an artificial spatial structure into
98 the data in any field study. Despite this casual relevancy, only a little interest has been taken in
99 studying systematically how sample design influences the detected spatial variation of
100 assemblages, although many papers have highlighted the importance of certain spatial scale
101 elements in describing the spatial structure of beta diversity (e.g., Barton et al. 2013; Heino et
102 al. 2015; Mykrä et al. 2007; Soininen 2015).
103
104 In two simulation studies, Smith & Lundholm (2010) and Gilbert & Bennett (2010) found that
105 spatial configuration and sampling strategies affect the results of variation partitioning.
106 Further, they also found that variation partitioning did not model the simulated spatial
107 structures of the data correctly. Migration rates (i.e., dispersal), as a spatial pattern-generating
108 mechanism, influenced both the environment- and space-related variation (Smith &
109 Lundholm 2010); and significant spatially explained variations were found even when the
110 simulated data did not contain spatial component (Gilbert & Bennett 2010).
111

112 Spatial extent, sample size and the topology of the sampling units could obviously affect the
113 environmental and spatial variables that researches consider relevant to describe the spatial
114 variation of assemblages. In many researches, these explanatory variables are identified via a
115 forward selection procedure (Blanchet et al. 2008) prior to variation partitioning. Although,
116 the adjusted form of the variation proportions (Peres-Neto et al. 2006) takes the number of the
117 explanatory variables into account which helps to compare the results of different studies, the
118 effect of the sampling design properties on the number of the relevant (i.e., selected)
119 explanatory variables has not been examined yet.
120
121 For stream-dwelling organisms like fish and aquatic molluscs that have no capacity for
122 terrestrial movement, dispersal connectivity among habitats is completely determined by the
123 physical dendritic structure of the stream network (Fagan et al. 2009), hence topology, beside
124 the dispersal ability of the animals, can be supposed to play a prominent role in their spatial
125 dynamics. The importance of topology of dendritic stream networks has been studied in
126 connection with, for example, fish dispersal (Hitt & Angermeier 2008, 2011) and in the
127 context of the distance-decay similarity relationship for aquatic invertebrates (e.g., Brown &
128 Swan 2010; Cañedo-Argüelles et al. 2015), but the relationship between the topology of the
129 effectively sampled locations of a dendritic network and the space-related community
130 variation is still little known. In fact, the behaviour of variation partitioning as a response of
131 changes in sampling design is still uncovered; therefore we do not know which sampling
132 design properties and variance fractions may be statistically associated to each other.
133
134 In spite of the warning results mentioned above and the lack of a solid understanding of the
135 relationship between sampling design properties and variation partitioning, the latter has been

136 frequently used to study the metacommunity organizations of a wide variety of taxa (e.g.,
137 Alahuhta & Heino 2013; Baldissera et al. 2012; Buschke et al. 2015; Campbell et al. 2015;
138 Erős et al. 2012; Fernandes et al. 2014; Göthe et al. 2013; Grönroos et al. 2013).
139
140 *1.4. Aims*
141 In this paper, we present how sampling design can affect the result of variation partitioning,
142 and how properties of sampling design can influence the number of the selected explanatory
143 variables and the change of the individual variance fractions in a dendritic stream network
144 using presence-absence data of fish species. Applying simple random sampling, we focused
145 on the specific questions as follows. (1) How does sample size (sampling effort) impact the
146 expected value of the estimated variance fractions? Assuming a fix sample size, (2) how does
147 the change of sample configuration influence the relative importance (i.e., rank order) of the
148 estimated variance fractions? (3) Does the change in the sample similarity cause a
149 proportional change in the result of variation partitioning? (4) In what extent can the change
150 of properties of sampling design other than sample size (spatial extent, sampling interval, and
151 topology) explain the change of the individual variance fractions and the number of
152 explanatory variables used for partitioning? Finally, (5) How strong is the association between
153 the amount of the unique variance fractions and the number of the selected explanatory
154 variables used for partitioning?
155
156
157
158
159

## 2. METHODS

Analyses of this study progressed through three main phases. First, environmental data were gathered and fish data were predicted by a statistical model using field survey data. Second, variation partitioning of fish data was done iteratively using simple random sampling with different sample size. Last, results of the variation partitioning were analysed statistically.

### *2.1. Studied stream system, environmental variables, and fish data*

The studied stream system is located in Hungary (Fig. 1), and contains two small rivers, the Zagyva (179 rkm) and the Tarna (105 rkm), and their tributaries (hereafter ZT system). The catchment area of the ZT system is 5676 $km^2$, and it has partly hilly (500 m > altitude ≥ 200 m a.s.l.), partly lowland (altitude < 200 m a.s.l.) geomorphology.

The GIS model of the ZT system used for this study consisted of 115 stream segments (sensu Frissell et al. 1986), that were considered as sampling units (see Erős et al. 2011). Stream segments were characterized with 20 abiotic environmental variables (see Table 1). We used variables which could be relatively easily collated in a GIS environment for each segment, and were widely and successfully used for the predictive modelling of stream fish in former studies (e.g., Park et al. 2006; Hermoso et al. 2011, 2013, 2015).These GIS based data were used from the following data bases: WorldClim (Hijmans et al. 2005), BioClim (Hijmans et al. 2005), Global Human Footprint (Sanderson et al. 2002), Corine Land Cover (Steenmans & Büttner 2006). Note, that instream variables (e.g., substrate composition) could not be used in this case, because these data were not available for all segments. Although this may influence the predictive power of the models, most fish based models use GIS based data exclusively

for predictive modelling (e.g., Leathwick et al. 2005; Hermoso et al. 2011, 2013, 2015; Filipe et al. 2013). Since we used the same variables for each stream segment, which were determined by the same analytical procedure, it is likely that our modelling approach did not influence the final outcome of our simulations, and the main conclusions.

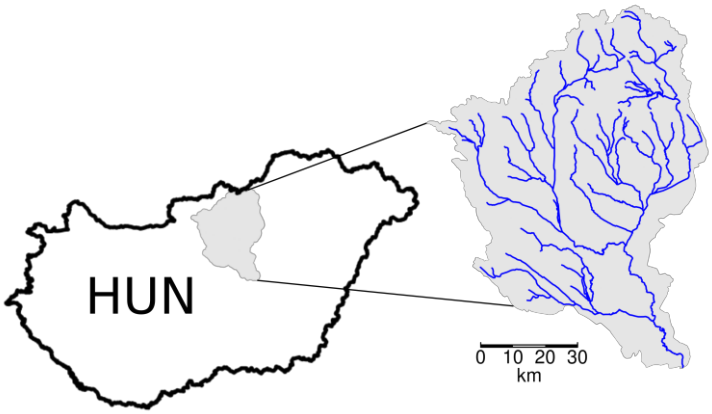



*Fig. 1. Location of the Zagyva-Tarna stream system in Hungary. Stream segments (stream reaches between two confluences) were considered as sampling units of the study.*

*Table 1. Abiotic environmental variables used in this study. All the listed variables acted as a potential predictor in the MARS modelling. However, only variables marked with an asterisk (*) were included in the variation partitioning procedure, because of strong linear associations among the variables.*

| Variable | Description | Min. | Max. | Mean ± SD |
|---|---|---|---|---|
| *Distance from source | Stream distance of the midpoint of the segment from the flow origin. (rkm) | 0.68 | 163.22 | 20.35 ± 28.17 |
| *Sinuosity index | Sinuosity index of the segment. Calculated as (l-d)/d, where l is the channel length, d is the Euclidean distance between the upstream and downstream endpoints of the segment. 0 means straight flow. | 0.00 | 0.72 | 0.16 ± 0.13 |

| Variable | Description | Min. | Max. | Mean ± SD |
|---|---|---|---|---|
| Altitude | Average altitude above sea of the raster cells touched by the segment. Derived from the Alt16 raster of the WorldClim database. (m) | 83.00 | 582.14 | 178.23 ± 84.76 |
| *Annual mean temperature | Annual mean temperature averaged across the raster cells touched by the segment. Derived from the BIO1 raster of the BioClim database. (°C) | 7.69 | 10.70 | 10.03 ± 0.59 |
| Maximum temperature of the warmest month | Maximum temperature of the warmest moth averaged across the raster cells touched by the segment. Derived from the BIO5 raster of the BioClim database. (°C) | 23.47 | 27.28 | 26.48 ± 0.63 |
| Minimum temperature of the coldest month | Minimum temperature of the coldest moth averaged across the raster cells touched by the segment. Derived from the BIO6 raster of the BioClim database. (°C) | -7.10 | -4.28 | -5.31 ± 0.67 |
| Isothermality | The proportion of the mean diurnal temperature range to the annual temperature range averaged across the raster cells touched by the segment. Derived from the BIO3 raster of the BioClim database. (%) | 29.00 | 31.00 | 30.35 ± 0.54 |
| Temperature seasonality | Averaged value of the raster cells touched by the segment. Derived from the BIO4 raster of the BioClim database. (Standard deviation × 100) | 7523.71 | 7937.44 | 7828.59 ± 67.03 |
| Annual precipitation | Annual precipitation averaged across the raster cells touched by the segment. Derived from the BIO12 raster of the BioClim database. (mm) | 518.00 | 648.86 | 546.38 ± 23.45 |
| Precipitation of the wettest month | Precipitation of the wettest month averaged across the raster cells touched by the segments. Derived from the BIO13 raster of the BioClim database. (°C) | 67.00 | 90.29 | 71.88 ± 4.30 |

| Variable | Description | Min. | Max. | Mean ± SD |
|---|---|---|---|---|
| Precipitation of the driest month | Precipitation of the driest month averaged across the raster cells touched by the segments. Derived from the BIO14 raster of the BioClim database. (°C) | 27.00 | 36.43 | 29.26 ± 1.65 |
| *Precipitation seasonality | Averaged value of the raster cells touched by the segment. Derived from the BIO15 raster of the BioClim database. (Coefficient of variation) | 24.86 | 32.86 | 28.97 ± 2.25 |
| *Human footprint | Human Footprint score averaged across the raster cells touched by the segment. Derived from the Global Human Footprint (Geographic) v2 (1995–2004) database. A value of 0 means no human influence, whereas a value of 100 means maximum human influence. | 21.00 | 76.00 | 45.18 ± 11.36 |
| *Artificial surfaces (CLC) | Relative area of the artificial surfaces within a 60 m width buffer zone around the segment. Derived by unifying the area of the land cover patches coded by 111, 112, 121, 122, 123, 124, 131, 132, 133, 141 and 142 in CORINE 2006 database. | 0 | 0.98 | 0.12 ± 0.17 |
| Agricultural surfaces (CLC) | Relative area of the agricultural surfaces within a 60 m width buffer zone around the segment. Derived by unifying the area of the land cover patches coded by 211, 213, 221, 222, 231, 242 and 243 in CORINE 2006 database. | 0 | 1 | 0.63 ± 0.29 |
| *Forested vegetation (CLC) | Relative area of the forested vegetation surfaces within a 60 m width buffer zone around the segment. Derived by unifying the area of the land cover patches coded by 311, 312 and 313 in CORINE 2006 database. | 0 | 1 | 0.15 ± 0.23 |
| *Scrub and herbaceous | Relative area of the scrub and herbaceous vegetation surfaces within a 60 m width buffer zone around the | 0 | 0.65 | 0.05 ± 0.10 |

| Variable | Description | Min. | Max. | Mean ± SD |
|---|---|---|---|---|
| vegetation (CLC) | segment. Derived by unifying the area of the land cover patches coded by 321, 322, 323 and 324 in CORINE 2006 database. | | | |
| *Wetlands (CLC) | Relative area of inland wetlands within a 60 m width buffer zone around the segment. Derived by unifying the area of the land cover patches coded by 411 and 412 in CORINE 2006 database. | 0 | 0.47 | 0.02 ± 0.06 |
| Water bodies (CLC) | Relative area of inland water bodies within a 60 m width buffer zone around the segment. Derived by unifying the area of the land cover patches coded by 511 and 512 in CORINE 2006 database. | 0 | 0.86 | 0.03 ± 0.11 |
| *Ponds | Relative area of ponds within a 60 m width buffer zone around the segment. Derived from a national Water Framework Directive GIS layer. | 0 | 0.32 | 0.02 ± 0.06 |

Fish occurrence (presence-absence) data associated to each stream segment was obtained from predictive species distribution modelling. It was necessary, because fish data from field surveys (altogether 251 surveys conducted at 132 sites between 2003 and 2014) were only available for 68 segments (literature and own data on a total of 42 species). For building the species distribution models we used actual field data. The standardized sampling protocol consisted of the single pass electrofishing of representative habitats of the segments, with the total length examined depending on the type of the waterbody (for details see Erős, 2007). For streams, a battery-powered electrofishing device was used (Hans-Grassl IG 200/2B, PDC). The crew sampled a 150 m long reach, slowly walking upstream and with single-pass fishing of the whole stream width. For non-wadeable rivers, boat electrofishing was applied with a

207 generator driven device (Hans-Grassl EL64 II GI, SDC), slowly moving downstream and
208 electrofishing 500 m long reaches in near shore areas. This division in sampling length
209 between streams and rivers was necessary to optimize sampling effort and to sample fish
210 assemblages representatively and proportionally to the size of the water body (see e.g.,
211 Oberdorff et al. 2001; Pont et al. 2006). Species richness estimators showed that such an effort
212 catches most fish species (> 85%) in a single occasion in both streams and rivers in this
213 ecoregion (see Erős, 2007; Sály et al. 2009 for details). After identification and counting, fish
214 were released into the water at the site of capture. Note, that segments where former faunistic
215 studies did not justify the existence of fish were considered unrepresentatively surveyed.
216
217 As a first step of the predictive modelling, fish data of the surveys were pooled within the
218 stream segments. Species occurring at less than four segments (~5%) were excluded from the
219 analysis. Data of the remaining species were used as a training data set in a multiresponse
220 multivariate regression splines (MARS) model (Leathwick et al 2005). In the model, the 20
221 abiotic environmental variables were used as potential predictors. MARS was fitted with a
222 generalised linear model with binomial error distribution option on the training data.
223 Predictive performance of the model was evaluated by a mean AUC value (area under a
224 receiver operating characteristic curve) computed from ten 4-fold cross validations for each
225 species separately. Species with a mean AUC value less than 0.7 (an arbitrary threshold) were
226 excluded (see Appendix), and the model was refitted on the data of the retained species.
227 Consequently, weakly predictable species, e.g., ubiquitous ones, did not influence the general
228 predictive performance of the model. In the second step, the trained MARS model was fitted
229 on all the stream segments to get occurrence probability of the species. As a last step,
230 occurrence probabilities were converted into binary presence-absence data using a threshold

231 criterion that maximizes the sum of sensitivity and specificity (Jiménez-Valverde & Lobo
232 2007), which resulted in a complete fish data set for the entire ZT system.
233
234 *2.2. Reducing the number of environmental variables*
235 Collinearity among explanatory variables can lead to unreliable parameter estimations and to
236 inflation of the coefficient of (multiple) determination of statistical models. In variation
237 partitioning, strongly correlated explanatory variables can cause negative estimated variance
238 fractions (Peres-Neto et al. 2006). Therefore, during preliminary data analyses, the 20
239 environmental variables were screened on the basis of pairwise Pearson correlations (its
240 absolute value would not be greater than 0.7) and expert judgement to find a subset of them in
241 which there was no strong collinearity among the variables. As a result of this screening
242 process 10 out of the initial 20 environmental variables were selected for further analysis
243 (marked with an asterisk in Table 1), and used as input variables in forward selection
244 procedures before variation partitioning.
245
246 *2.3. Iterative randomization procedure: sampling, forward selection, variation partitioning*
247 *and sampling design characterization*
248 The statistical sampling distributions of the variance fractions were generated using an
249 iterative randomization procedure (Monte Carlo simulation). This procedure was conducted
250 with four sample sizes, choosing 23, 46, 69, and 92 stream segments randomly from the 115
251 ZT stream segments (statistical population). These sample sizes corresponded to 20%, 40%,
252 60% and 80% information coverage of the statistical population.
253

Each random sample was analysed as if it had been a single field sample, correspondingly, the steps of its analysis followed a scenario that is commonly used in variation partitioning by field ecologists. When it was necessary, the geographic localization of the unique stream segments was modelled by latitude and longitude coordinates of the midpoint of the segments during the analysis process. Segment midpoint is the point that is halfway stream distance from both endpoints of the stream segment.

The iteration process was initiated by choosing a random sample of the ZT segments. Then, the sample was subjected to a Moran's eigenvector maps (MEM) analysis (Dray et al. 2006) to get the potential spatial explanatory variables of the particular sample. To start this analysis, the pairwise stream distance matrix of the midpoint of the sample stream segments was transformed into a matrix of normalized distances:

$$d'_{ij} = 1 - (d_{ij}/d_{max})$$

where $d'_{ij}$ is the normalized distance for the distance of segment $i$ and segment $j$; $d_{ij}$ is the original distance (rkm) of segment $i$ and segment $j$; $d_{max}$ is the maximum of the pairwise distances (rkm) of the sample segments.

Two stream segments were considered neighbours (i.e., connected) only if there was a direct path (i.e., a path that did not go through a third stream segment included in the given sample) between them along the stream network. Otherwise they were considered unconnected. Connectivity relationships were summarized in a symmetric binary matrix (CM) in which 1s coded the connected and 0s the unconnected pairs of segments.

278

279 In order to get a spatially weighted connectivity matrix, CM was weighted with the matrix of
280 the normalized distances. Then, the result matrix (Hadamard product) was eigen-analysed.
281 Eigenvectors with positive eigenvalue were retained as potential spatial explanatory variables
282 of the given sample.

283

284 After MEM analysis, the fish data of the sample was checked, and species that did not occur
285 in any sample segments were deleted from the data table. Similarly, environmental data of the
286 sample were checked as well, and environmental variables with zero variance were deleted.

287

288 Before variation partitioning, a forward selection procedure (Blanchet et al. 2008) was applied
289 to identify the relevant environmental and spatial variables that can serve as explanatory
290 variables of the given sample. Forward selection was controlled by three stopping criteria to
291 avoid overfitting: (1) a preselected variable had to explain a significant portion of the
292 explained variance, in other words, significance value of a preselected variable had to be
293 larger than 0.05; (2) a preselected significant variable had to increase the coefficient of
294 multiple determination ($R^2$) by at least 0.01; (3) the adjusted coefficient of multiple
295 determination ($R^2_{adj}$) did not have to be larger than a value of that derived from a global test
296 (i.e., including all the environmental variables or spatial variables). The numbers of the
297 selected environmental and spatial variables (i.e., the numbers of the effective explanatory
298 variables) were recorded.

299

300 Then, an RDA-based variation partitioning with adjusted coefficients of multiple
301 determination was used to get the purely environmentally, the purely spatially, the jointly
302 explained, and the residual variance fractions (Peres-Neto et al. 2006).
303
304 After variance partitioning, sampling design properties of the particular random sample were
305 recorded. Spatial extent was measured as the area of the rectangle expanding between the
306 westernmost and easternmost, and southernmost and northernmost sample segments.
307 Sampling interval was measured as the average Euclidean distance between the neighbouring
308 stream segments. We note here that during preliminary analyses sample interval had been
309 measured by using stream distances instead of Euclidean distances, but this showed weaker
310 relationships with the variance fractions than Euclidean distance did, hence it was omitted.
311 Topology of the sampling units in a certain sampling design was quantified as average
312 eccentricity of the nodes of a graph of the sample segments. This connected graph was made
313 from the symmetric binary connectivity matrix (CM, see above), and its nodes represented the
314 sample segments, whereas its (unweighted) edges represented the connections between them
315 (see Erős et al. 2011 Fig. 1). Eccentricity of a single node is the maximum topological
316 (shortest path) distance between the particular node and any other node of the graph. The
317 greater the mean eccentricity of the graph nodes, the more elongated the topology of the
318 sampling design. In preliminary analyses, we had quantified the topology by other graph
319 theoretic measures (Harary index, degree centrality, betweenness centrality, closeness
320 centrality) (Minor & Urban 2008; Ricotta et al. 2000), but these measures were rather strongly
321 associated (mostly linearly) with each other, therefore we used only the mean eccentricity in
322 the main analysis.
323

Random sampling and the subsequent analysis process described above was iterated 400 times at every sample size level, which resulted in a total of 1600 (4 sample sizes × 400 repetitions) unique sampling designs and variation partitioning analyses.

After the randomization procedure, variation of the statistical population (i.e., data of all the 115 ZT segments) was also decomposed by the same analytical procedure that had been used for the random samples.

*2.4. Statistical analysis of variation partitioning results*

Finishing the random sampling procedure, the sampling distribution of the variance fractions and the number of the selected environmental and spatial explanatory variables was characterised by descriptive statistics.

Variance fractions of all the 1600 partitioning analyses were ranked to quantify their relative importance; and the frequency distribution of the unique rank order vectors was used to assess the robustness of the variance partitioning against sampling design alteration for every sample size.

The strength of the general relationship between sampling design modification and the results of variance partitioning was quantified and tested by Mantel tests with 999 randomizations for each sample size. In these tests, pairwise sample similarity was measured by Kulczynski index, and pairwise difference in variation partitioning results by Euclidean distance using variance fractions [a], [b] and [c].

Specific relationships between the variance fractions, the number of selected environmental and spatial variables, and sampling design properties were explored by generalised least squares regression models (i.e., weighted linear regression) with maximum likelihood estimation (Zuur et al. 2009). Variance fractions and the number of the selected environmental and spatial variables were the response variables, whereas spatial extent, sampling interval, topology measure acted as explanatory variables nested within the sample size (categorical variable) in each regression model. Because variance of the response variables depended on the groups of the sample size, a variance structure that allows different variances for each group was built in the models (Zuur et al. 2009). After model fitting, significance of each explanatory variable at a level of alpha equals 0.05 was judged with a t-test. Non-significant explanatory variables were excluded and the model was refitted on the data in order to get a minimum adequate model that had no any insignificant terms (Crawley 2007).

Relationships between the unique variance fractions and the number of the selected environmental and spatial variables were examined through correlation analyses.

*2.5. Software tools*

GIS data processing was done with QGIS (QGIS Development Team 2014). All the statistical analyses were conducted in R environment (R Core Team 2015). MARS modelling was carried out as it is implemented in the earth package (Milborrow et al. 2014). Thresholds to convert predicted probabilities into presence-absence data were identified with PresenceAbsence package (Freeman & Moisen 2008). MEM analysis was conducted with the spacemakeR package (Dray 2013). The packfor package (Dray et al. 2013) was used for the forward selection procedure. Pairwise stream distance matrix was computed with shp2graph

372 (Lu 2014) and igraph (Csárdi & Nepusz 2006) packages. Variation partitioning was done with
373 the varpart function of the vegan package (Oksanen et al. 2015). Line graph construction and
374 eccentricity computation were also carried out with the igraph package (Csárdi & Nepusz
375 2006). Package vegan (Oksanen et al. 2013) was used for the Mantel tests too. Generalised
376 least squares regressions were conducted with nlme package (Pinheiro et al. 2015).
377
378
379
380
381 3. RESULTS
382
383 *3.1. Species distribution modelling*
384 Out of the 42 fish species of the field data set of the Zagyva-Tarna system, 14 species were
385 excluded owing to rarity, and 11 species because of poor predictability. MARS algorithm
386 selected two environmental predictors (distance from source and precipitation of the wettest
387 month) to model the distribution of the remaining 17 fish species that were included into the
388 main analyses (Table 2). The fit of the MARS model on the training data measured by the
389 coefficient of determination ($R^2$) averaged across the 17 species was 0.30 ± 0.13 (mean ± SD).
390 The same value of the generalized coefficient of determination ($GR^2$, it is corrected for the
391 effective number of model parameters and the number of observations [see earth package
392 vignette 'Notes on the earth package' at http://www.milbo.org/doc/earth-notes.pdf])
393 measuring the generalization performance of the model was 0.20 ± 0.14. The mean AUC
394 value of the ten 4-fold cross validations averaged across the 17 species was 0.80 ± 0.06 (Table
395 2).

396

397

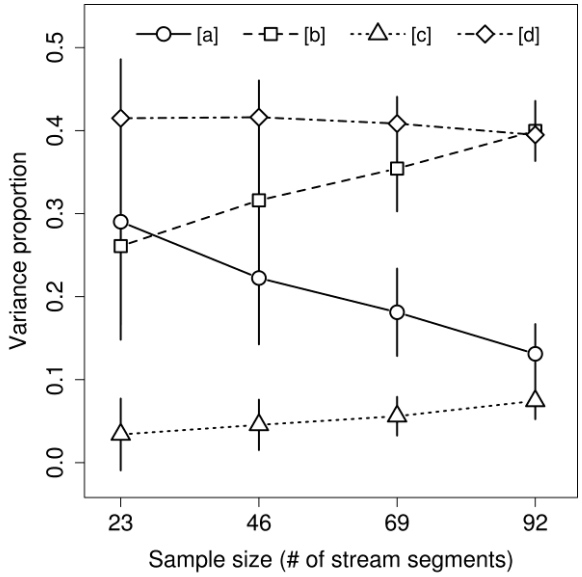
398 *Table 2. Relative occurrence frequency (i.e., prevalence) of the fish species in the training*
399 *data and MARS–GLM performance. $R^2$: coefficient of determination; $GR^2$: generalized*
400 *coefficient of determination; AUC: area under a receiver operating characteristic curve*
401 *averaged across the results of ten 4-fold cross validations.*

| Species | Common name | Rel. occ. fr. (n=68) | $R^2$ | $GR^2$ | AUC (mean ± SD) |
|---|---|---|---|---|---|
| *Alburnoides bipunctatus* | Schneider (spirlin) | 0.176 | 0.177 | 0.069 | 0.733 ± 0.194 |
| *Alburnus alburnus* | bleak | 0.529 | 0.358 | 0.274 | 0.766 ± 0.111 |
| *Barbatula barbatula* | stone loach | 0.544 | 0.273 | 0.178 | 0.739 ± 0.152 |
| *Blicca bjoerkna* | white bream | 0.309 | 0.302 | 0.210 | 0.764 ± 0.145 |
| *Carassius gibelio* | Prussian carp | 0.500 | 0.177 | 0.069 | 0.727 ± 0.124 |
| *Cobitis elongatoides* | spined loach | 0.618 | 0.331 | 0.243 | 0.827 ± 0.095 |
| *Esox lucius* | northern pike | 0.353 | 0.428 | 0.353 | 0.853 ± 0.100 |
| *Gobio gobio* | gudgeon | 0.588 | 0.137 | 0.024 | 0.732 ± 0.142 |
| *Leuciscus aspius* | asp | 0.074 | 0.145 | 0.033 | 0.811 ± 0.171 |
| *Leuciscus leuciscus* | common dace | 0.088 | 0.160 | 0.050 | 0.848 ± 0.155 |
| *Proterorhinus semilunaris* | Western tubenose goby | 0.309 | 0.590 | 0.537 | 0.952 ± 0.041 |
| *Rhodeus sericeus* | bitterling | 0.500 | 0.425 | 0.349 | 0.833 ± 0.123 |
| *Romanogobio vladykovi* | Danube whitefin gudgeon | 0.147 | 0.319 | 0.230 | 0.846 ± 0.146 |
| *Rutilus rutilus* | roach | 0.559 | 0.416 | 0.339 | 0.844 ± 0.100 |
| *Sander lucioperca* | pikeperch | 0.147 | 0.177 | 0.069 | 0.765 ± 0.155 |
| *Scardinius erythrophthalmus* | rudd | 0.279 | 0.300 | 0.208 | 0.810 ± 0.104 |
| *Squalius cephalus* | chub | 0.632 | 0.363 | 0.280 | 0.761 ± 0.142 |

| Species | Common name | Rel. occ. fr. (n=68) | $R^2$ | $GR^2$ | AUC (mean ± SD) |
|---|---|---|---|---|---|
| mean and SD of species | – | 0.374 ± 0.197 | 0.299 ± 0.126 | 0.207 ± 0.143 | 0.801 ± 0.060 |

*3.2. Descriptive statistics of variance fractions and the number of the selected environmental and spatial variables*

Descriptive statistics of the sample distribution of the variance fractions varied as sample size changed (Table 3). Mean value of variance fraction [a] decreased, and that of variance fraction [b] increased considerably with increasing sample size. Although, the mean of variance fraction [c] also increased, its changes were moderate. Interestingly, the mean of variance fraction [d] remained virtually the same at all the four sample sizes (Fig. 2; Table 3). Further, the mean value of the residual variance fraction was reasonably close to the residual variance fraction obtained from variation partitioning of the total statistical population (115 ZT segments) even at the smallest sample size. Whereas the mean value of the other variance fractions approximated the corresponding variance fractions in greater steps with increasing sample size (Table 3).




418

419 *Fig. 2. Mean value and standard deviation of the variance fractions at different sample sizes.*
420 *Values were computed from the results of RDA-based variation partitioning analyses of 364*
421 *(for sample size 23) or 400 (for sample size 46, 69, 92) random samples. Circles stand for the*
422 *pure environmentally explained ([a]), squares for the jointly explained by environment and*
423 *space ([b]), triangles for the pure spatially explained, and diamonds for the unexplained ([d])*
424 *variance fractions.*

425

426 All the dispersion indices (SD, CV%, IQR and range) decreased monotonically as sample size
427 increased. Despite of this trend, the range of stochastic fluctuation of each variance fraction
428 exceeded 0.10 (i.e., 10%) even at the largest sample size that is when dispersion was the
429 smallest for every variance fraction. Considering a given sample size, the residual variance
430 fraction ([d]) showed the smallest, and the pure spatial variance fraction ([c]) the largest
431 relative variability measured by the coefficient of variation (Table 3).

432

433 Mean value of the number of the selected environmental and spatial variables also showed a
434 positive relationship with sample size. Further, increasing sample size had a greater effect on

435 the number of the selected MEM variables, than on the number of the selected environmental
436 ones. Similarly to the case of variance fractions, mean values of these two variables computed
437 at the largest simple size were the closest to the number of the selected environmental and
438 MEM variables obtained from the forward selection of the total statistical population (115 ZT
439 segments) (Table 3).
440
441 Standard deviation and range of the number of the selected MEM variables depended on the
442 sample size too, but those of the number of the selected environmental variables did not so
443 (Table 3).
444

*Table 3. Descriptive statistics of the variance fractions and the number of the selected environmental and spatial variables derived from an iterative randomization procedure. [a] purely environmentally explained variance fraction. [b] variance fraction jointly explained by environmental and spatial variables. [c] purely spatially explained variance fraction. [d] unexplained residual variance fraction. Sample size refers the number of stream segments of the random samples. n: the number of random samples drawn during the iterative randomization procedure; SD: standard deviation; CV (%): coefficient of variation (SD/mean × 100); Q1: the first quartile; Q3: the third quartile; IQR: interquartile range. Note that variation partitioning was not done in 36 cases out of the 400 random samples at the level of sample size 23. Note also that the last row shows the result of variation partitioning of the entire data set (i.e., all the 115 ZT segments).*

| Sample size (relative sample size) | Statistics | [a] | [b] | [c] | [d] | # of selected env. vars | # of selected spatial vars |
|---|---|---|---|---|---|---|---|
| 23 (0.20) | n | 364 | 364 | 364 | 364 | 364 | 364 |
| | min | 0.017 | 0.018 | -0.047 | 0.222 | 1 | 1 |
| | Q1 | 0.201 | 0.173 | 0.002 | 0.367 | 2 | 1 |
| | median | 0.286 | 0.254 | 0.026 | 0.412 | 2 | 2 |
| | Q3 | 0.379 | 0.344 | 0.055 | 0.465 | 3 | 3 |
| | max | 0.605 | 0.582 | 0.247 | 0.643 | 6 | 9 |
| | mean | 0.290 | 0.260 | 0.034 | 0.415 | 2.511 | 2.44 |
| | SD | 0.124 | 0.113 | 0.043 | 0.071 | 0.759 | 1.338 |
| | CV (%) | 42.73 | 43.23 | 127.20 | 17.12 | 30.22 | 54.83 |
| | IQR | 0.177 | 0.172 | 0.053 | 0.098 | 1 | 2 |
| | range | 0.588 | 0.564 | 0.294 | 0.422 | 5 | 8 |
| 46 (0.40) | n | 400 | 400 | 400 | 400 | 400 | 400 |

| Sample size (relative sample size) | Statistics | [a] | [b] | [c] | [d] | # of selected env. vars | # of selected spatial vars |
|---|---|---|---|---|---|---|---|
| | min | 0.042 | 0.061 | -0.014 | 0.290 | 2 | 1 |
| | Q1 | 0.166 | 0.262 | 0.024 | 0.386 | 3 | 4 |
| | median | 0.212 | 0.324 | 0.040 | 0.416 | 3 | 5 |
| | Q3 | 0.280 | 0.372 | 0.061 | 0.443 | 4 | 7 |
| | max | 0.488 | 0.529 | 0.170 | 0.550 | 6 | 13 |
| | mean | 0.222 | 0.316 | 0.046 | 0.416 | 3.298 | 5.495 |
| | SD | 0.080 | 0.080 | 0.030 | 0.044 | 0.846 | 2.122 |
| | CV (%) | 35.93 | 25.36 | 66.65 | 10.64 | 25.65 | 38.61 |
| | IQR | 0.113 | 0.110 | 0.037 | 0.058 | 1 | 3 |
| | range | 0.446 | 0.469 | 0.184 | 0.260 | 4 | 12 |
| 69 (0.60) | n | 400 | 400 | 400 | 400 | 400 | 400 |
| | min | 0.061 | 0.205 | 0.006 | 0.311 | 2 | 2 |
| | Q1 | 0.147 | 0.318 | 0.040 | 0.387 | 4 | 7 |
| | median | 0.178 | 0.355 | 0.054 | 0.411 | 4 | 9 |
| | Q3 | 0.214 | 0.39 | 0.070 | 0.432 | 4 | 11 |
| | max | 0.343 | 0.498 | 0.145 | 0.499 | 7 | 17 |
| | mean | 0.181 | 0.354 | 0.056 | 0.409 | 4.065 | 8.918 |
| | SD | 0.053 | 0.052 | 0.023 | 0.032 | 0.776 | 2.532138 |
| | CV (%) | 29.07 | 14.56 | 41.63 | 7.89 | 19.10 | 28.39 |
| | IQR | 0.067 | 0.073 | 0.030 | 0.045 | 0 | 4 |
| | range | 0.282 | 0.292 | 0.139 | 0.188 | 5 | 15 |
| 92 (0.80) | n | 400 | 400 | 400 | 400 | 400 | 400 |
| | min | 0.031 | 0.282 | 0.012 | 0.311 | 3 | 6 |

| Sample size (relative sample size) | Statistics | [a] | [b] | [c] | [d] | # of selected env. vars | # of selected spatial vars |
|---|---|---|---|---|---|---|---|
| | Q1 | 0.105 | 0.379 | 0.058 | 0.380 | 4 | 12 |
| | median | 0.127 | 0.402 | 0.072 | 0.395 | 4 | 14 |
| | Q3 | 0.152 | 0.424 | 0.090 | 0.411 | 5 | 16 |
| | max | 0.241 | 0.486 | 0.156 | 0.455 | 7 | 23 |
| | mean | 0.131 | 0.400 | 0.074 | 0.395 | 4.518 | 13.97 |
| | SD | 0.036 | 0.036 | 0.022 | 0.023 | 0.718 | 2.771 |
| | CV (%) | 27.31 | 9.045 | 29.59 | 5.89 | 15.90 | 19.84 |
| | IQR | 0.047 | 0.045 | 0.031 | 0.031 | 1 | 4 |
| | range | 0.210 | 0.203 | 0.143 | 0.144 | 4 | 17 |
| 115 (total statistical population) | – | 0.103 | 0.429 | 0.084 | 0.384 | 5 | 18 |

456

457

458 *3.3. Rank order of variance fractions*

459 Stochastic fluctuation of the variance fractions affected strongly their rank order. Considering
460 all the four variance fractions, frequency distribution of the rank orders consisted 10, 6, 5 and
461 4 different rank order vectors for the sample size 23, 46, 69 and 92, respectively (Table 4). If
462 we considered only the variance fractions [a], [b] and [c], the numbers of the unique rank
463 order vectors were 5, 3, 3 and 2 for the sample size 23, 46, 69 and 92, respectively (Table 5).

464

*Table 4. Frequency distribution of the unique rank orders considering all the four variance fractions ([a] pure environmentally explained, [b] jointly explained by environment and space, [c] pure spatially explained, [d] unexplained). Rank 1 denotes the smallest of the variance fractions. At every sample size, the frequency distribution was made from the result of 400 variation partitioning analyses. NAs mean that variation partitioning was not done because there were not any significant spatial variable for 36 random sample configuration. Therefore, in these cases all the explained variance can be interpreted as pure environmentally explained variance.*

| Sample size (relative sample size) | type of rank order vector | [a] | [b] | [c] | [d] | frequency | rel. freq. |
|---|---|---|---|---|---|---|---|
| 23 (0.20) | 1 | 3 | 2 | 1 | 4 | 127 | 0.3175 |
| | 2 | 2 | 3 | 1 | 4 | 92 | 0.2300 |
| | 3 | 4 | 2 | 1 | 3 | 69 | 0.1725 |
| | 4 | 2 | 4 | 1 | 3 | 48 | 0.1200 |
| | 5 | 1 | 4 | 2 | 3 | 10 | 0.0250 |
| | 6 | 1 | 3 | 2 | 4 | 9 | 0.0225 |
| | 7 | 4 | 3 | 1 | 2 | 4 | 0.0100 |
| | 8 | 4 | 1 | 2 | 3 | 4 | 0.0100 |
| | 9 | 3 | 4 | 1 | 2 | 1 | 0.0025 |
| | 10 | NA | NA | NA | NA | 36 | 0.0900 |
| 46 (0.40) | 1 | 2 | 3 | 1 | 4 | 212 | 0.5300 |
| | 2 | 3 | 2 | 1 | 4 | 102 | 0.2550 |
| | 3 | 2 | 4 | 1 | 3 | 62 | 0.1550 |
| | 4 | 4 | 2 | 1 | 3 | 10 | 0.0250 |
| | 5 | 1 | 4 | 2 | 3 | 10 | 0.0250 |

| Sample size (relative sample size) | type of rank order vector | [a] | [b] | [c] | [d] | frequency | rel. freq. |
|---|---|---|---|---|---|---|---|
| | 6 | 1 | 3 | 2 | 4 | 4 | 0.0100 |
| 69 (0.60) | 1 | 2 | 3 | 1 | 4 | 300 | 0.7500 |
| | 2 | 2 | 4 | 1 | 3 | 68 | 0.1700 |
| | 3 | 3 | 2 | 1 | 4 | 21 | 0.0525 |
| | 4 | 1 | 4 | 2 | 3 | 10 | 0.0250 |
| | 5 | 1 | 3 | 2 | 4 | 1 | 0.0025 |
| 92 (0.80) | 1 | 2 | 4 | 1 | 3 | 174 | 0.4350 |
| | 2 | 2 | 3 | 1 | 4 | 174 | 0.4350 |
| | 3 | 1 | 4 | 2 | 3 | 48 | 0.1200 |
| | 4 | 1 | 3 | 2 | 4 | 4 | 0.0100 |
| 115 (total statistical population) | true rank order | 2 | 4 | 1 | 3 | – | – |

*Table 5. Frequency distribution of the unique rank orders considering all the pure environmentally explained ([a]), the jointly explained by environment and space ([b]), and the pure spatially explained ([c]) variance fractions. Rank 1 denotes the smallest of the variance fractions. At every sample size, the frequency distribution was made from the result of 400 variation partitioning analyses. NAs mean that variation partitioning was not done because there were not any significant spatial variable for 36 random sample configuration. Therefore, in these cases all the explained variance can be interpreted as pure environmentally explained variance.*

| Sample size (relative sample size) | type of rank order vector | [a] | [b] | [c] | frequency | rel. freq. |
|---|---|---|---|---|---|---|
| 23 (0.20) | 1 | 3 | 2 | 1 | 200 | 0.5000 |

| Sample size (relative sample size) | type of rank order vector | [a] | [b] | [c] | frequency | rel. freq. |
|---|---|---|---|---|---|---|
| | 2 | 2 | 3 | 1 | 141 | 0.3525 |
| | 3 | 1 | 3 | 2 | 19 | 0.0475 |
| | 4 | 3 | 1 | 2 | 4 | 0.0100 |
| | 5 | NA | NA | NA | 36 | 0.0900 |
| 46 (0.40) | 1 | 2 | 3 | 1 | 274 | 0.6850 |
| | 2 | 3 | 2 | 1 | 112 | 0.2800 |
| | 3 | 1 | 3 | 2 | 14 | 0.0350 |
| 69 (0.60) | 1 | 2 | 3 | 1 | 368 | 0.9200 |
| | 2 | 3 | 2 | 1 | 21 | 0.05250 |
| | 3 | 1 | 3 | 2 | 11 | 0.02750 |
| 92 (0.80) | 1 | 2 | 3 | 1 | 348 | 0.8700 |
| | 2 | 1 | 3 | 2 | 52 | 0.1300 |
| 115 (total statistical population) | true rank order | 2 | 3 | 1 | – | – |

### *3.4. General relationship between sampling design modification and results of variation partitioning*

Although the mean of the pairwise Euclidean distances of the variation partitioning results of the random samples crashed, and the mean of the pairwise sample similarities (Kulczynski index) increased sharply as the sample size increased, there was not any kind of association between them at any levels of a single sample size (Table 6.).

*Table 6. Results of Mantel tests of variation partitioning (Euclidean distances) vs. sample similarities (Kulczynski index). Euclidean distances were computed from the three variance fractions as follows: pure environmentally explained ([a]), jointly explained by environment and space ([b]), and pure spatially explained ([c]). p-values were computed from 999 randomizations.*

| Sample size (relative sample size) | Mantel statistics (Spearman correlation) | p-value |
|---|---|---|
| 23 (0.20) | -0.018 | 1 |
| 46 (0.40) | -0.022 | 1 |
| 69 (0.60) | -0.033 | 1 |
| 92 (0.80) | -0.036 | 1 |

*3.5. Relationships between properties of sampling design and unique variance fractions*

Number of explanatory variables contained by the minimum adequate regression models varied across the models of the different response variables (i.e., variance fractions). In general, the strength of the linear relationships of the properties of the sampling design with the unique variance fractions were moderate (see pseudo-$R^2$s in Table 7) and sample size dependent.

Pure environmentally explained variance fraction ([a]) was affected negatively by spatial extent although its effect was only marginally significant ($0.05 < p \leq 0.10$) at sample size 69, and significant ($p < 0.05$) at sample sizes 46 and 92. Estimated effect of sampling interval on variance fraction [a] was positive at all the sample sizes, but it was marginally significant at

509 sample size 69 and significant at sample size 92. Interestingly, the effect size (regression
510 coefficient b) and its statistical significance (p-value) of sampling interval increased
511 consistently as sample size increased. Mean eccentricity (topology) showed significant
512 positive effect on [a] at sample size 92, and marginally significant positive effects at sample
513 sizes 46 and 69.
514
515 Variance fraction explained jointly by environment and space ([b]) was significantly
516 associated only with sampling interval in a negative way at each sample size. Similarly to the
517 case of variance fraction [a], the effect size and significance of this association also increased
518 consistently with increasing sample size.
519
520 Pure spatially explained variance fraction ([c]) was negatively influenced by sampling interval
521 and mean eccentricity, but only at the largest sample size. The effect of these two explanatory
522 variables was highly insignificant at other sample sizes.
523
524 Residual variance fraction ([d]) was affected by spatial extent positively at sample sizes 46,
525 69, 92, by sampling interval also positively at sample sizes 23, 69, 92, and by mean
526 eccentricity negatively at sample size 69.
527
528 *3.6. Relationships between properties of sampling design and the number of the selected*
529 *spatial and environmental variables*
530 Variation of the number of the selected environmental and spatial variables was better
531 explainable by sample design properties than that of the variance fractions (see pseudo $R^2$
532 values at Table 7). The number of the selected environmental variables was positively related

533 to spatial extent at larger sample sizes (69, 92). On the other hand, the number of the selected
534 spatial variables was influenced only by sampling interval and in a negative way. Apart from
535 sample size 46, this relationship was significant at all the other sample sizes (Table 7).
536
537 *3.7. Correlations between variance fractions and number of the selected environmental and*
538 *spatial variables*
539 Pairwise Pearson correlation coefficients showed that each unique variance fraction covaried
540 much stronger with the number of the selected spatial variables than with spatial extent,
541 sampling interval or mean eccentricity independently of sample size. The direction of the
542 covariation was consistent across sample sizes for every variance fraction. On the contrary,
543 strength and direction of covariation between unique variance fractions and the number of the
544 selected environmental variables depended on sample size and type of variance fraction
545 (Table 8).

*Table 7. Results of the generalised least squares models. Estimated partial regression coefficients (b), their standard error (SE), significance value, and the standardized partial regression coefficients (i.e., beta coefficients [Quinn & Keough, 2002]) (beta). Pseudo-$R^2$ means the proportion of explained variation; it was computed as 1 - RSS/TSS where RSS is the residual sum of squares and TSS is the total sum of squares. Note that the spatial extent, sampling interval and mean eccentricity was nested within sample size, but models did not contain sample size as a main effect. Consequently, the estimation of the intercept parameter is meaningless and is not shown in the table.*

| | | Sample size 23 (0.20) | | | Sample size 46 (0.40) | | | Sample size 69 (0.60) | | | Sample size 92 (0.80) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Explanatory variables | | | Explanatory variables | | | Explanatory variables | | | Explanatory variables | | |
| Response variable (pseudo-$R^2$) | | spa. ext. | sampl. int. | mean ecc. | spa. ext. | sampl. int. | mean ecc. | spa. ext. | sampl. int. | mean ecc. | spa. ext. | sampl. int. | mean ecc. |
| [a] (0.352) | b | $-2.6\times10^{-06}$ | $3.0\times10^{-4}$ | -0.004 | $-1.6\times10^{-05}$ | 0.004 | 0.008 | $-9.2\times10^{-06}$ | 0.008 | 0.005 | $-1.2\times10^{-05}$ | 0.020 | 0.006 |
| | SE | $9.9\times10^{-06}$ | 0.003 | 0.008 | $7.2\times10^{-06}$ | 0.004 | 0.004 | $5.2\times10^{-06}$ | 0.004 | 0.003 | $4.8\times10^{-06}$ | 0.005 | 0.002 |
| | t statistics | -0.261 | 0.092 | -0.558 | -2.245 | 1.105 | 1.925 | -1.757 | 1.941 | 1.718 | -2.472 | 4.038 | 2.700 |
| | p-value | 0.793 | 0.927 | 0.577 | 0.025 | 0.269 | 0.054 | 0.079 | 0.052 | 0.086 | 0.014 | $5.7\times10^{-05}$ | 0.007 |
| | beta | -0.015 | 0.006 | -0.032 | -0.121 | 0.063 | 0.104 | -0.091 | 0.105 | 0.089 | -0.126 | 0.213 | 0.137 |
| [b] (0.324) | b | | -0.005 | | | -0.007 | | | -0.016 | | | -0.024 | |

| Response variable (pseudo-$R^2$) | | Sample size 23 (0.20) Explanatory variables | | | Sample size 46 (0.40) Explanatory variables | | | Sample size 69 (0.60) Explanatory variables | | | Sample size 92 (0.80) Explanatory variables | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | spa. ext. | sampl. int. | mean ecc. | spa. ext. | sampl. int. | mean ecc. | spa. ext. | sampl. int. | mean ecc. | spa. ext. | sampl. int. | mean ecc. |
| | SE | | 0.002 | | | 0.003 | | | 0.003 | | | 0.005 | |
| | t statistics | | -2.051 | | | -2.264 | | | -4.519 | | | -5.159 | |
| | p-value | | 0.040 | | | 0.024 | | | $6.7\times10^{-06}$ | | | $2.8\times10^{-07}$ | |
| | beta | | -0.107 | | | -0.113 | | | -0.221 | | | -0.250 | |
| [c] (0.197) | b | | -0.001 | 0.002 | | 0.002 | 0.001 | | 0.001 | $3.2\times10^{-04}$ | | -0.012 | -0.004 |
| | SE | | 0.001 | 0.003 | | 0.001 | 0.002 | | 0.002 | 0.001 | | 0.003 | 0.001 |
| | t statistics | | -0.863 | 0.855 | | 1.399 | 0.891 | | 0.450 | 0.242 | | -4.065 | -3.143 |
| | p-value | | 0.388 | 0.393 | | 0.162 | 0.373 | | 0.653 | 0.809 | | $5.1\times10^{-05}$ | 0.002 |
| | beta | | -0.049 | 0.049 | | 0.074 | 0.047 | | 0.024 | 0.013 | | -0.206 | -0.160 |
| [d] (0.086) | b | $4.5\times10^{-06}$ | 0.005 | 0.002 | $1.1\times10^{-05}$ | 0.004 | -0.004 | $8.4\times10^{-06}$ | 0.008 | -0.005 | $8.2\times10^{-06}$ | 0.017 | -0.002 |
| | SE | $5.5\times10^{-06}$ | 0.002 | 0.005 | $3.9\times10^{-06}$ | 0.002 | 0.002 | $3.1\times10^{-06}$ | 0.002 | 0.002 | $3.0\times10^{-06}$ | 0.003 | 0.001 |

| Response variable (pseudo-$R^2$) | | Sample size 23 (0.20) Explanatory variables | | | Sample size 46 (0.40) Explanatory variables | | | Sample size 69 (0.60) Explanatory variables | | | Sample size 92 (0.80) Explanatory variables | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | spa. ext. | sampl. int. | mean ecc. | spa. ext. | sampl. int. | mean ecc. | spa. ext. | sampl. int. | mean ecc. | spa. ext. | sampl. int. | mean ecc. |
| | t statistics | 0.818 | 2.957 | 0.480 | 2.904 | 1.801 | -1.920 | 2.723 | 3.381 | -2.662 | 2.719 | 5.405 | -1.410 |
| | p-value | 0.414 | 0.003 | 0.631 | 0.004 | 0.072 | 0.055 | 0.006 | 0.001 | 0.008 | 0.007 | $7.5\times10^{-08}$ | 0.159 |
| | beta | 0.046 | 0.181 | 0.027 | 0.153 | 0.100 | -0.102 | 0.135 | 0.175 | -0.133 | 0.133 | 0.273 | -0.069 |
| # of selected env. vars (0.518) | b | $-7.7\times10^{-06}$ | | -0.002 | $-4.4\times10^{-05}$ | | 0.030 | $2.4\times10^{-04}$ | | 0.127 | $4.4\times10^{-04}$ | | 0.068 |
| | SE | $5.5\times10^{-05}$ | | 0.045 | $7.1\times10^{-05}$ | | 0.041 | $7.3\times10^{-05}$ | | 0.041 | $9.3\times10^{-05}$ | | 0.040 |
| | t statistics | -0.141 | | -0.035 | -0.624 | | 0.723 | 3.289 | | 3.128 | 4.660 | | 1.690 |
| | p-value | 0.888 | | 0.972 | 0.533 | | 0.470 | 0.001 | | 0.002 | $3.4\times10^{-06}$ | | 0.091 |
| | beta | -0.007 | | -0.002 | -0.031 | | 0.036 | 0.162 | | 0.154 | 0.229 | | 0.083 |
| # of selected spatial vars (0.783) | b | | -0.074 | | | -0.081 | | | -0.623 | | | -1.229 | |
| | SE | | 0.029 | | | 0.084 | | | 0.171 | | | 0.357 | |

| Response variable (pseudo-$R^2$) | | Sample size 23 (0.20) | | | Sample size 46 (0.40) | | | Sample size 69 (0.60) | | | Sample size 92 (0.80) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Explanatory variables | | | Explanatory variables | | | Explanatory variables | | | Explanatory variables | | |
| | | spa. ext. | sampl. int. | mean ecc. | spa. ext. | sampl. int. | mean ecc. | spa. ext. | sampl. int. | mean ecc. | spa. ext. | sampl. int. | mean ecc. |
| | t statistics | | -2.564 | | | -0.967 | | | -3.640 | | | -3.442 | |
| | p-value | | 0.010 | | | 0.334 | | | $2.8\times10^{-04}$ | | | $5.9\times10^{-04}$ | |
| | beta | | -0.133 | | | -0.048 | | | -0.180 | | | -0.170 | |

551

552 *Table 8. Pairwise Pearson correlation coefficients (lower triangle) and their p-values (upper triangle) of the sampling design properties, number of the*

553 *selected environmental and spatial variables, and the unique variance fractions.*

| sample size 23 | spatial extent | sampling interval | mean ecc. | # of selected env. vars | # of selected spatial vars | [a] | [b] | [c] | [d] |
|---|---|---|---|---|---|---|---|---|---|
| spatial extent | | < 0.001 | 0.063 | 0.918 | 0.198 | 0.845 | 0.552 | 0.096 | 0.021 |
| sampling interval | 0.427 | | < 0.001 | 0.073 | 0.011 | 0.830 | 0.041 | 0.202 | < 0.001 |
| mean eccentricity | -0.093 | -0.378 | | 0.996 | 0.283 | 0.533 | 0.461 | 0.204 | 0.392 |
| # of selected environmental variables | 0.005 | 0.090 | 0.000 | | 0.615 | < 0.001 | 0.931 | 0.003 | < 0.001 |
| # of selected spatial variables | -0.068 | -0.134 | 0.056 | -0.026 | | < 0.001 | < 0.001 | < 0.001 | < 0.001 |
| [a] | -0.010 | 0.011 | -0.033 | 0.229 | -0.733 | | < 0.001 | < 0.001 | < 0.001 |
| [b] | -0.031 | -0.107 | 0.039 | 0.005 | 0.769 | -0.816 | | < 0.001 | < 0.001 |
| [c] | -0.087 | -0.067 | 0.067 | -0.158 | 0.576 | -0.410 | 0.236 | | < 0.001 |
| [d] | 0.121 | 0.191 | -0.045 | -0.311 | -0.292 | -0.201 | -0.307 | -0.268 | |
| sample size 46 | | | | | | | | | |
| spatial extent | | < 0.001 | 0.121 | 0.569 | 0.813 | 0.065 | 0.515 | 0.155 | < 0.001 |
| sampling interval | 0.321 | | < 0.001 | 0.866 | 0.334 | 0.830 | 0.024 | 0.244 | < 0.001 |

| sample size 23 | spatial extent | sampling interval | mean ecc. | # of selected env. vars | # of selected spatial vars | [a] | [b] | [c] | [d] |
|---|---|---|---|---|---|---|---|---|---|
| mean eccentricity | 0.078 | -0.336 | | 0.498 | 0.957 | 0.142 | 0.785 | 0.655 | 0.013 |
| # of selected environmental variables | -0.029 | 0.008 | 0.034 | | 0.736 | 0.104 | 0.796 | 0.150 | 0.016 |
| # of selected spatial variables | -0.012 | -0.048 | 0.003 | 0.017 | | < 0.001 | < 0.001 | < 0.001 | < 0.001 |
| [a] | -0.092 | -0.011 | 0.074 | 0.081 | -0.713 | | < 0.001 | < 0.001 | 0.404 |
| [b] | -0.033 | -0.113 | -0.014 | 0.013 | 0.733 | -0.858 | | < 0.001 | < 0.001 |
| [c] | 0.071 | 0.058 | 0.022 | -0.072 | 0.572 | -0.428 | 0.267 | | < 0.001 |
| [d] | 0.177 | 0.183 | -0.123 | -0.121 | -0.433 | 0.042 | -0.444 | -0.396 | |
| sample size 69 | | | | | | | | | |
| spatial extent | | < 0.001 | 0.131 | 0.003 | 0.093 | 0.170 | 0.078 | 0.101 | < 0.001 |
| sampling interval | 0.278 | | < 0.001 | 0.271 | < 0.001 | 0.286 | < 0.001 | 0.692 | < 0.001 |
| mean eccentricity | -0.076 | -0.293 | | 0.005 | 0.117 | 0.190 | 0.301 | 0.908 | < 0.001 |
| # of selected environmental variables | 0.150 | -0.055 | 0.142 | | 0.204 | 0.134 | 0.395 | 0.043 | 0.018 |
| # of selected spatial variables | -0.084 | -0.179 | 0.078 | -0.064 | | < 0.001 | < 0.001 | < 0.001 | < 0.001 |
| [a] | -0.069 | 0.053 | 0.066 | 0.075 | -0.668 | | < 0.001 | < 0.001 | 0.203 |

| sample size 23 | spatial extent | sampling interval | mean ecc. | # of selected env. vars | # of selected spatial vars | [a] | [b] | [c] | [d] |
|---|---|---|---|---|---|---|---|---|---|
| [b] | -0.088 | -0.221 | 0.052 | 0.043 | 0.700 | -0.854 | | < 0.001 | < 0.001 |
| [c] | 0.082 | 0.020 | 0.006 | -0.101 | 0.577 | -0.458 | 0.344 | | < 0.001 |
| [d] | 0.194 | 0.252 | -0.195 | -0.118 | -0.446 | 0.064 | -0.453 | -0.525 | |
| sample size 92 | | | | | | | | | |
| spatial extent | | < 0.001 | 0.010 | < 0.001 | 0.344 | 0.099 | 0.066 | 0.306 | < 0.001 |
| sampling interval | 0.290 | | < 0.001 | 0.050 | 0.001 | 0.005 | < 0.001 | 0.001 | < 0.001 |
| mean eccentricity | -0.129 | -0.267 | | 0.285 | 0.097 | 0.053 | 0.163 | 0.037 | 0.001 |
| # of selected environmental variables | 0.219 | 0.098 | 0.054 | | 0.058 | < 0.001 | 0.175 | < 0.001 | 0.996 |
| # of selected spatial variables | -0.047 | -0.170 | 0.083 | -0.095 | | < 0.001 | < 0.001 | < 0.001 | < 0.001 |
| [a] | -0.083 | 0.139 | 0.097 | 0.179 | -0.675 | | < 0.001 | < 0.001 | < 0.001 |
| [b] | -0.092 | -0.250 | 0.070 | -0.068 | 0.683 | -0.882 | | < 0.001 | < 0.001 |
| [c] | 0.051 | -0.164 | -0.104 | -0.178 | 0.610 | -0.500 | 0.339 | | < 0.001 |
| [d] | 0.221 | 0.330 | -0.159 | 0.000 | -0.602 | 0.305 | -0.517 | -0.704 | |

554 4. DISCUSSION

555

556 This methodological investigation provides an insight into the relationship between
557 ordination-based variation partitioning and the properties of sampling design in a dendritic
558 network context. Although a recent prominent study (Gilbert & Bennett 2010) has touched
559 this problem in a lattice grid context, to our knowledge, this study is the first which focused
560 on the effect of sampling design primarily on the relative importance of the environment- and
561 space-related component of assemblage variations, and on the specific relationships between
562 the unique variance fractions and sampling design properties.

563

564 *4.1. Effect of sample size*

565 In general, because our dendritic study system (Zagyva-Tarna stream system) consists of a
566 finite number of sampling units (stream segments), sample size usually interacts with the
567 effects of the other sampling design properties.

568

569 Expected values of the variance fractions estimated by the sample mean behaved in a peculiar
570 way as sample size increased. Interestingly, residual variance fraction [d] changed negligibly
571 as sample size increased. This result suggests that given a certain set of environmental
572 descriptor variables, the total explainable variation of assemblages can be estimated with
573 rather high accuracy independently from the sample size of the study. At the same time, the
574 dispersion statistics of the unique residual variance fraction showed that the precision of this
575 estimation can be low, especially at small or medium sample size.

576

577 Contrary to the residual variation, the mean values of the environment- and space-related
578 variance fractions varied highly and their relative importance changed with changes in sample
579 size. The decreasing of the mean environmentally explained variance with increasing sample
580 size could, on the one hand, be a data set specific phenomenon. Both distance from source and
581 precipitation of the wettest month, the two predictors used to model fish species distributions
582 by MARS, can be associated with the longitudinal profile of a stream system. If species
583 distribution is controlled mainly by the longitudinal profile associated environmental factors,
584 the pure environmentally explained variance is expected to be low at small sample size,
585 because spatially compact (i.e., less eccentric) sampling design with a short environmental
586 gradient is more probable to occur at small sample size than at large sample size. On the other
587 hand, the most fundamental environmental factors that control the spatial distribution of
588 riverine fish assemblages at large scale, such as altitude, channel slope, discharge, are strongly
589 related to the longitudinal aspect of running waters (Matthews 1998). Therefore, this natural
590 character of stream systems can also result in low environmentally explained variance.
591
592 The greater the sample size, the more complex network structures can be combined from the
593 sample segments. This can be the reason why space-related variance increased with sample
594 size. In other words, the number of possible unique topological configurations (i.e., possible
595 spatial patterns) of the sampling units depends on the number of the sample units and on their
596 topological position within the stream network. This assumption is supported by the result
597 that the mean number of the selected MEM variables also increased as sample size increased.
598 On the other hand, least squares regression model showed that at a certain sample size, the
599 number of the selected MEM variables was influenced by sampling interval. Eigenanalyses-
600 based spatial models, like MEM analysis and the analysis of principal coordinates of

601 neighbour matrices (PCNM; Borcard & Legendre 2002), have the ability to model complex
602 spatial patterns at various spatial scales (Dray et al. 2012). Smith & Lundholm (2010) argued
603 for the sophisticated behaviour of the PCNM method about that variation partitioning could
604 not distinguish between environment-related and space-related patterns. Similarly, Gilbert &
605 Bennett (2010) also showed that PCNM predictors inflated the explained variation in spite the
606 use of the adjusted coefficient of determination ($R^2_{adj}$). Therefore, it can be supposed that
607 space-related variances revealed by these eigenanalyses-based techniques primarily reflect the
608 complexity of the design in terms of the number and spatial arrangement of the sampling
609 units. If this is really the case, ecologists should be cautious when they infer the importance of
610 dispersal of the studied organisms from purely the spatially explained variance of
611 assemblages, especially when they have no reasonable knowledge on the movement ability of
612 the studied species.
613
614 *4.2. Effect of sampling configuration*
615 Given a fix sample size, the stochastic fluctuation of the estimated variance fractions induced
616 by the change of sampling configuration seems to be not consistent with each other. As a
617 consequence, rank order of the variance fractions can change randomly as well. Considering
618 the relative frequency of the experienced unique rank order vectors suggest that the
619 uncertainty of the estimation of the true rank order (i.e., the rank order obtained by variation
620 partitioning of the total statistical population [115 segments]) is the greatest at small sample
621 size. However, as our results demonstrate, it is possible that even at 80% information
622 coverage of the statistical population there could be roughly 0.13 probability chance to miss
623 the true rank order vector when researchers aim to assess the relative importance of variance
624 fractions [a], [b] and [c]. Moreover, small sample size could involve such sample

625 configurations from which MEM eigenvectors are not able to cover any significant spatial
626 structures at a significance level of alpha equals 0.05. This result supports Alahuhta & Heino's
627 (2013) conclusion that the relative contribution of environmental and spatial mechanisms to
628 metacommunity structuring varies in a rather unpredictable way.
629
630 As Mantel tests revealed, the change of sample design similarity seems not to cause a
631 proportional modification in the result of variation partitioning. In other words, a small
632 change in sample similarity of two random samples can result in both a great and a small
633 difference between the results of the variance partitioning of the two random samples alike.
634 This surprising result suggests that the effect of sampling design on variation partitioning can
635 be hardly predicted on the basis of sample similarity. The rationale behind this must be related
636 to the identity of the sampling units. Considering a compositional difference between two
637 equal-sized samples caused by only a single pair of randomly selected stream segments, the
638 biological similarity (species pool) can vary according to the topological position of the
639 selected segments. For example, two stream segments with the same Strahler order (e.g., two
640 headwater segments) tend to have much more similar species pool than two segments with
641 different Strahler order (e.g., one headwater and one mainstem segment).
642
643 *4.3. Effect of spatial extent, sampling interval and topology*
644 Results of the GLS models suggest that spatial extent affect mainly the environmentally
645 explainable variation of species assemblages. This involves an indirect influence on the
646 residual variation as well. Interestingly, Grönroos et al. (2013) found that spatial extent was
647 not related to metacommunity structuring. Because they had different number of local sites at
648 the different spatial extents, the modifying role of sample size and/or topology may be the

649 reason for the apparent lack of the effect of spatial extent. Sampling interval appears to
650 modify both the environment- and space-related variation, but its effect on these two unique
651 variance fractions could depend on the sample size. However, the emergent and negative
652 effect of sampling interval as it can be detected in the residual variation seems to be
653 independent on sample size. Although, topology seems to affect both the pure environmental
654 and spatial variance its influence can be powerful only at large sample size. To sum up, results
655 suggest that variation partitioning in a dendritic system (i.e., in a system with a finite number
656 of sampling units) is more sensitive to the properties of the sampling design when the
657 informational coverage of the statistical population is large than when that is small or
658 medium.
659
660 Spatial extent and topology tend to influence the selected number of the abiotic variables,
661 although their effect seems significant only at large sample sizes. On the contrary, sampling
662 interval could reduce the number of the selected spatial explanatory variables. That is
663 sampling interval might influence the complexity of the spatial structure that can be modelled
664 by an eigenanalysis-based spatial method in dendritic networks.
665
666 Probably the most surprising result emerging from our study was that each variance fraction
667 was correlated much stronger with the number of the selected MEM variables than with any
668 of the sampling design properties. Further, pseudo-$R^2$ values of the GLS models indicated that
669 sampling interval tend to explain better the variation of the number of the selected MEM
670 variables than that of any variance fractions. Hence, it is likely that sampling interval
671 primarily affects the number of the selected MEM variables in the forward selection
672 procedure, which in turn influences the estimated variance fractions in variation partitioning.

673 The increased number of the selected MEM variables tend to increase the spatially explained
674 variance, and reduce the environmentally explained and the residual variance fractions (see
675 correlations in Table 8). This finding corresponds to Gilbert & Bennett (2010) who reported a
676 statistical artefact nature of eigenanalysis-based spatial methods, because selection of some
677 eigenvector variables can involve selecting additional ones leading to inflated explained
678 variance. Therefore, spatial patterns behind the increased spatially explained community
679 variation sometimes can be ecologically meaningless.
680
681
682 5. CONCLUSIONS
683
684 The findings of this study clearly indicate that sampling design has a considerable and
685 unpredictable effect on the result of multivariate variation partitioning. Of sampling design
686 properties, it seems that sample size and sampling interval influences notably the results. It is
687 highly probable that this influencing effect is strongly related to the ability of eigenanalysis-
688 based spatial variables to model complex patterns. Apart from other important factors, such as
689 biogeographic regions and anthropogenic modifications, differences in sampling design could
690 have a significant role in the inconsistency of the results of metacommunity studies of stream
691 organisms.
692
693 Acknowledgements
694 This work was supported by the OTKA K104279 grant and the Bolyai János Research
695 Scholarship of the Hungarian Academy of Sciences (TE).
696

697 REFERENCES

698

699 Alahuhta, J. and Heino, J. 2013. Spatial extent, regional specificity and metacommunity
700 structuring in lake macrophytes. – Journal of Biogeography 40: 1572–1582.

701

702 Baldissera, R. et al. 2012. Metacommunity composition of web-spiders in a fragmented
703 neotropical forest: relative importance of environmental and spatial effects. – Plos One 7:
704 e48099.

705

706 Barton, P. S. et al. 2013. The spatial scaling of beta diversity. – Global Ecology and
707 Biogeography 22: 639–647.

708

709 Beisner, B. E. et al. 2006. The role of environmental and spatial processes in structuring lake
710 communities from bacteria to fish. – Ecology 87: 2985–2991.

711

712 Borcard, D. and Legendre, P. 2002. All-scale spatial analysis of ecological data by means of
713 principal coordinates of neighbour matrices. – Ecological Modelling 153: 51–68.

714

715 Brown, B. L. and Swan, C. M. 2010. Dendritic network structure constrains metacommunity
716 properties in riverine ecosystems. – Journal of Animal Ecology 79: 571–580.

717

718 Buschke, F. T. et al. 2015. Partitioning the variation in African vertebrate distributions into
719 environmental and spatial components - exploring the link between ecology and
720 biogeography. – Ecography 38: 450–461.

721

722 Campbell, R. E. et al. 2015. Flow-related disturbance creates a gradient of metacommunity
723 types within stream networks. – Landscape Ecology 30: 667–680.
724

725 Cañedo-Argüelles, M. et al. 2015. Dispersal strength determines meta-community structure in
726 a dendritic riverine network. – Journal of Biogeography 42: 778–790.
727

728 Cottenie, K. 2005. Integrating environmental and spatial processes in ecological community
729 dynamics. – Ecology Letters 8: 1175–1182.
730

731 Crawley, M. J. 2007. The R Book. – Wiley Publishing.
732

733 Csárdi, G. and Nepusz, T. 2006. The igraph software package for complex network research. –
734 InterJournal Complex Systems: 1695.
735

736 Cushman, S. A. and McGarigal, K. 2004. Patterns in the species-environment relationship
737 depend on both scale and choice of response variables. – Oikos 105: 117–124.
738

739 Declerck, S. A. J. et al. 2011. Scale dependency of processes structuring metacommunities of
740 cladocerans in temporary pools of High-Andes wetlands. – Ecography 34: 296–305.
741

742 Dray, S. 2013. spacemakeR: Spatial modelling. – R package version 0.0-5/r113.
743

744 Dray, S. et al. 2012. Community ecology in the age of multivariate multiscale spatial analysis.
745 – Ecological Monographs 82: 257–275.
746
747 Dray, S. et al. 2013. packfor: Forward Selection with permutation (Canoco p.46). – R package
748 version 0.0-8/r109.
749
750 Erős, T. et al. 2012. Temporal variability in the spatial and environmental determinants of
751 functional metacommunity organization - stream fish in a human-modified landscape. –
752 Freshwater Biology 57: 1914–1928.
753
754 Erős, T. et al. 2011. Network thinking in riverscape conservation - A graph-based approach. –
755 Biological Conservation 144: 184–192.
756
757 Fagan, W. F. et al. 2009. Riverine landscapes: Ecology for an alternative geometry. – In:
758 Cantrell, S., Cosner, C. et al. (ed.), Spatial Ecology. Chapman and Hall/CRC, pp. 85-100.
759
760 Filipe, A. et al. 2013. Forecasting fish distribution along stream networks: brown trout (*Salmo*
761 *trutta*) in Europe. – Diversity and Distributions 19: 1059–1071.
762
763 Fernandes, I. M. et al. 2014. Spatiotemporal dynamics in a seasonal metacommunity structure
764 is predictable: the case of floodplain-fish communities. – Ecography 37: 464–475.
765
766 Freeman, E. A. and Moisen, G. 2008. PresenceAbsence: An R Package for Presence Absence
767 Analysis. – Journal of Statistical Software 23: 1–31.

768

769 Frissell, C. A. et al. 1986. A hierarchical framework for stream habitat classification - viewing
770 streams in a watershed context. – Environmental Management 10: 199–214.
771

772 Gilbert, B. and Bennett, J. R. 2010. Partitioning variation in ecological communities: do the
773 numbers add up? – Journal of Applied Ecology 47: 1071–1082.
774

775 Göthe, E. et al. 2013. Metacommunity structure in a small boreal stream network. – Journal of
776 Animal Ecology 82: 449–458.
777

778 Grönroos, M. et al. 2013. Metacommunity structuring in stream networks: roles of dispersal
779 mode, distance type, and regional environmental context. – Ecology and Evolution 3: 4473–
780 4487.
781

782 Heino, J. et al. 2015. Metacommunity organisation, spatial extent and dispersal in aquatic
783 systems: patterns, processes and prospects. – Freshwater Biology 60: 845–869.
784

785 Hermoso, V. et al. 2011. Addressing longitudinal connectivity in the systematic conservation
786 planning for freshwaters. – Freshwater Biology 56, 57–70.
787

788 Hermoso, V. et al. 2013. Data Acquisition for Conservation Assessments: Is the Effort Worth
789 It? – PLoS ONE 8: e59662.
790

791 Hermoso, V. et al. 2015. Evaluating the costs and benefits of systematic data acquisition for
792 conservation assessments. – Ecography 38: 283–292.
793
794 Hijmans, R. et al. 2005. Very high resolution interpolated climate surfaces for global land
795 areas. – International Journal of Climatology 25: 1965–1978.
796
797 Hitt, N. P. and Angermeier, P. L. 2011. Fish community and bioassessment responses to
798 stream network position. – Journal of the North American Benthological Society 30: 296–309.
799
800 Hitt, N. P. and Angermeier, P. L. 2008. Evidence for fish dispersal from spatial analysis of
801 stream network topology. – Journal of the North American Benthological Society 27: 304–
802 320.
803
804 Hoeinghaus, D. J. et al. 2007. Local and regional determinants of stream fish assemblage
805 structure: inferences based on taxonomic vs. functional groups. – Journal of Biogeography 34:
806 324–338.
807
808 Jiménez-Valverde, A. and Lobo, J. M. 2007. Threshold criteria for conversion of probability
809 of species presence to either-or presence-absence. – Acta Oecologica-international Journal of
810 Ecology 31: 361–369.
811
812 Leathwick, J. R. et al. 2005. Using multivariate adaptive regression splines to predict the
813 distributions of New Zealand's freshwater diadromous fish. – Freshwater Biology 50: 2034–
814 2052.

Lu, B. 2014. shp2graph: Convert a SpatialLinesDataFrame object to a "igraph-class" object. – R package version 0-2.

Marzin, A. et al. 2013. The relative influence of catchment, riparian corridor, and reach-scale anthropogenic pressures on fish and macroinvertebrate assemblages in French rivers. – Hydrobiologia 704: 375–388.

Matthews, W. J. 1998. Patterns in freshwater fish ecology. – Chapman and Hall.

Mesquita, N. et al. 2006. Spatial variation in fish assemblages across small mediterranean drainages: Effects of habitat and landscape context. – Environmental Biology of Fishes 77: 105–120.

Minor, E. S. and Urban, D. L. 2008. A graph-theory framework for evaluating landscape connectivity and conservation planning. – Conservation Biology 22: 297–307.

Mykrä, H. et al. 2007. Scale-related patterns in the spatial and environmental components of stream macroinvertebrate assemblage variation. – Global Ecology and Biogeography 16: 149–159.

Oberdorff, T. et al. 2001. A probabilistic model characterizing fish assemblages of French rivers: a framework for environmental assessment. – Freshwater Biology 46: 399–415.

Oksanen, J. et al. 2015. vegan: Community Ecology Package. – R package version 2.3-0.

840

841 Park, Y. et al. 2006. Stream fish assemblages and basin land cover in a river network. –
842 Science of The Total Environment 365: 140–153.

843

844 Peterson, E. E. et al. 2013. Modelling dendritic ecological networks in space: an integrated
845 network perspective. – Ecology Letters 16: 707–719.

846

847 Pinheiro, J. et al. 2015. nlme: Linear and Nonlinear Mixed Effects Models. – R package
848 version 3.1-122.

849

850 Pont D., et al. 2006. Assessing river biotic condition at a continental scale: a European
851 approach using functional metrics and fish assemblages. – Journal of Applied Ecology 43:
852 70–80.

853

854 QGIS Development Team 2014. QGIS Geographic Information System. – Open Source
855 Geospatial Foundation.

856

857 R Core Team 2015. R: A Language and Environment for Statistical Computing. – R
858 Foundation for Statistical Computing. Vienna, Austria.

859

860 Ricotta, C. et al. 2000. Quantifying the network connectivity of landscape mosaics: a graph-
861 theoretical approach. – Community Ecology 1: 89–94.

862

863 Sály, P. et al. 2009. Assemblage level monitoring of stream fishes: the relative efficiency of
864 single-pass *vs*. double-pass electrofishing. – Fisheries Research 99: 226–233.

865

866 Sály, P. et al. 2011. The relative influence of spatial context and catchment- and site-scale
867 environmental factors on stream fish assemblages in a human-modified landscape. – Ecology
868 of Freshwater Fish 20: 251–262.

869

870 Sanderson, E. W. et al. 2002. The human footprint and the last of the wild. – Bioscience 52:
871 891–904.

872

873 Smith, T. W. and Lundholm, J. T. 2010. Variation partitioning as a tool to distinguish between
874 niche and neutral processes. – Ecography 33: 648–655.

875

876 Soininen, J. 2015. Spatial structure in ecological communities – a quantitative analysis. –
877 Oikos DOI: 10.1111/oik.02241

878

879 Soininen, J. and Weckström, J. 2009. Diatom community structure along environmental and
880 spatial gradients in lakes and streams. – Fundamental and Applied Limnology 174: 205–213.

881

882 Steenmans, C. and Büttner, G. 2006. Mapping land cover of Europe for 2006 under GMES. –
883 In: Matthias, B. (ed.), Proceedings of the 2nd Workshop of the EARSeL SIG on Land Use and
884 Land Cover. The European Association of Remote Sensing Laboratories and the Center for
885 Remote Sensing of Land Surfaces at the Rheinische Friedrich-Wilhelms-Universität Bonn, pp.
886 202–207.

887

888

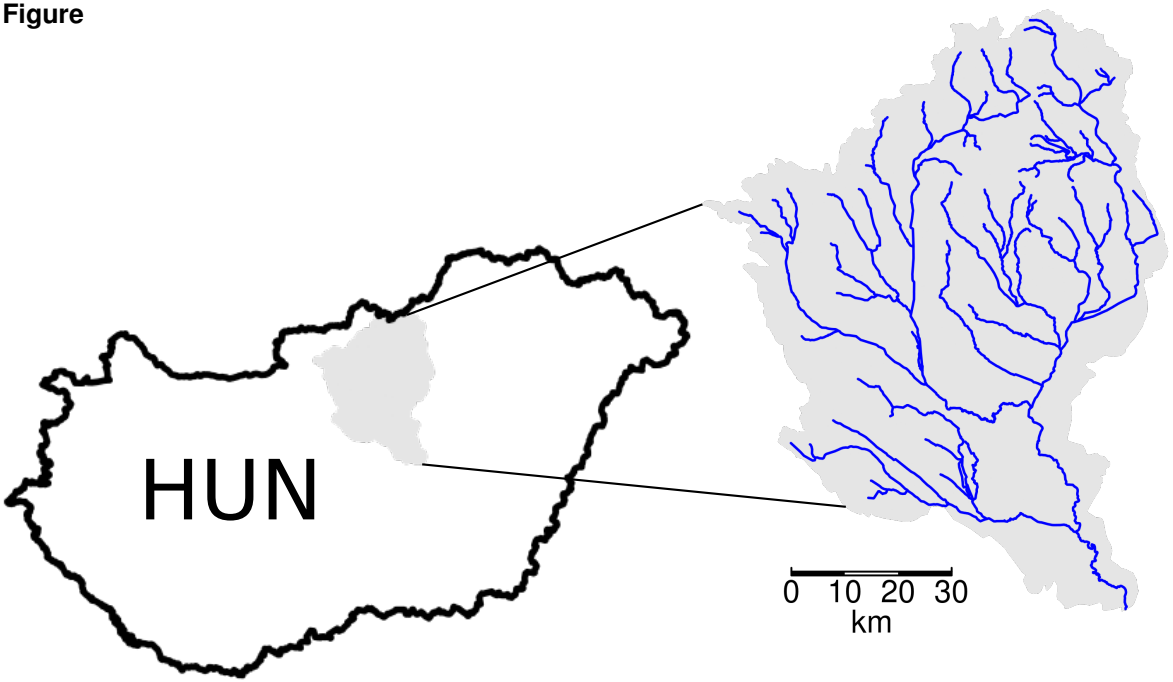
889 APPENDIX

890 *Species excluded from the MARS–GLM modelling because of low predictability (i.e., with a*
891 *mean AUC value less than 0.7). Rel. occ. fr.: relative occurrence frequency; $R^2$: coefficient of*
892 *determination; $GR^2$: generalized coefficient of determination; AUC: area under a receiver*
893 *operating characteristic curve averaged across the results of ten 4-fold cross validations.*

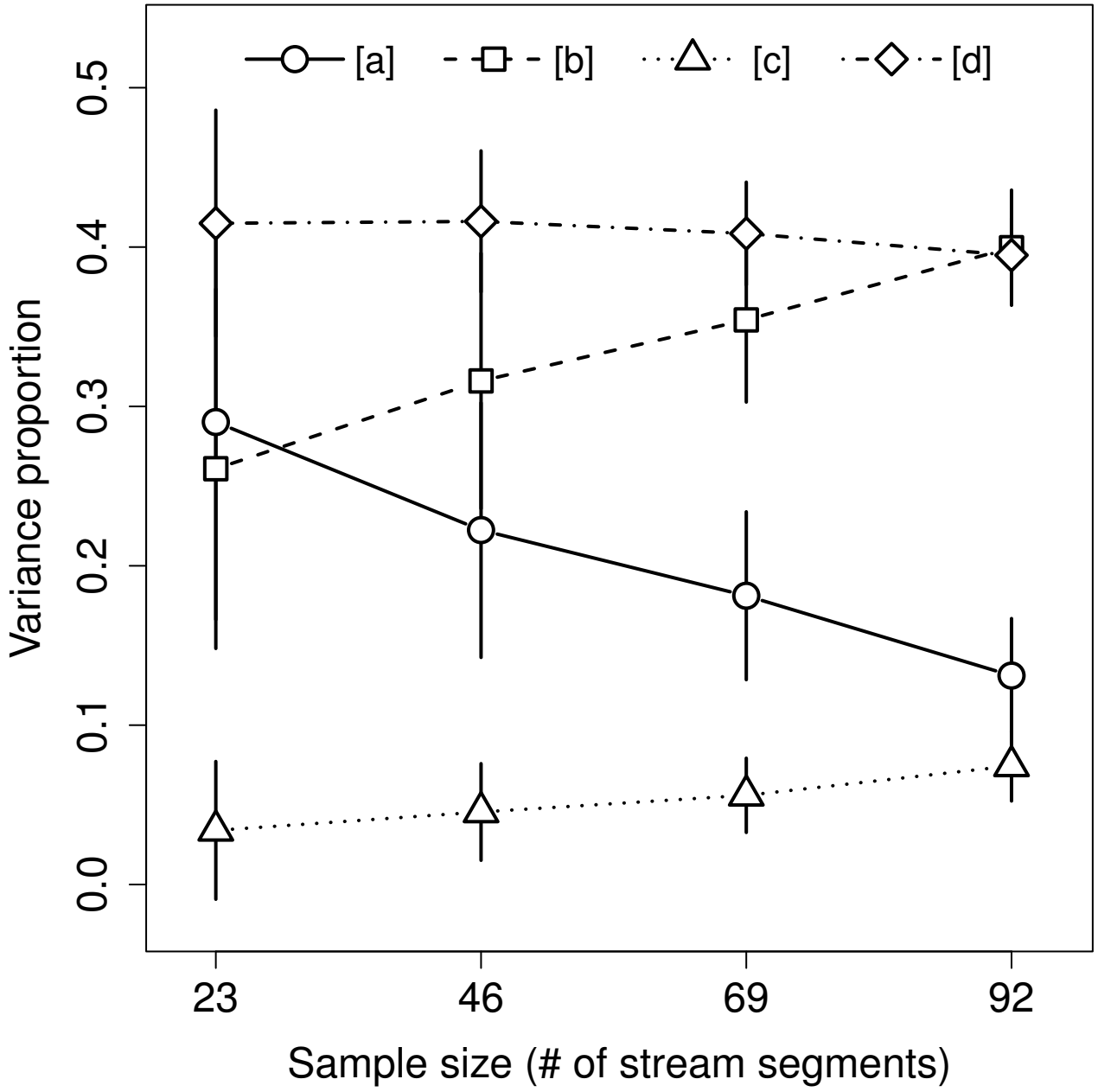
| Species | Common name | Rel. occ. fr. (n=68) | $R^2$ | $GR^2$ | AUC (mean ± SD) |
|---|---|---|---|---|---|
| *Abramis brama* | common bream | 0.191 | 0.109 | -0.007 | 0.691 ± 0.114 |
| *Ameiurus melas* | black bullhead | 0.118 | 0.069 | -0.053 | 0.656 ± 0.148 |
| *Carassius carassius* | Crucian carp | 0.088 | 0.078 | -0.042 | 0.629 ± 0.252 |
| *Gymnocephalus cernua* | ruffe | 0.103 | 0.104 | -0.014 | 0.664 ± 0.220 |
| *Lepomis gibbosus* | pumpkinseed | 0.221 | 0.060 | -0.063 | 0.629 ± 0.148 |
| *Leucaspius delineatus* | belica | 0.059 | 0.015 | -0.114 | 0.485 ± 0.247 |
| *Leuciscus idus* | ide | 0.118 | 0.070 | -0.052 | 0.675 ± 0.194 |
| *Misgurnus fossilis* | weatherfish | 0.103 | 0.073 | -0.048 | 0.681 ± 0.191 |
| *Neogobius fluviatilis* | monkey goby | 0.059 | 0.025 | -0.102 | 0.636 ± 0.237 |
| *Perca fluviatilis* | European perch | 0.309 | 0.146 | 0.034 | 0.693 ± 0.112 |
| *Pseudorasbora parva* | stone moroko | 0.265 | 0.112 | -0.004 | 0.677 ± 0.138 |

894

**Figure**



**Figure**


[a]
[b]
[c]
[d]
Variance proportion
0.0
0.1
0.2
0.3
0.4
0.5
23
46
69
92
Sample size (# of stream segments)