More information available at:
www.iba.muni.cz/summer-school2011

Publication was supported by the ESF project no. CZ.1.07/2.2.00/07.0318
"Multidisciplinary Inovation of Study in Computational Biology"
and national budget of the Czech Republic

INVESTMENTS IN EDUCATION DEVELOPMENT

ISBN 978-80-7204-756-7

LEDNICE 2011

Institute of Biostatistics and Analyses
Masaryk University

Proceedings of the 7th Summer School on Computational Biology

Biodiversity: from Genetics to Geography, from Mathematics to Management

15–17 September 2011
Lednice, Czech Republic

Editor:
Jiří Jarkovský

INVESTMENTS IN EDUCATION DEVELOPMENT

Proceedings of the 7th Summer School on Computational Biology

Institute of Biostatistics and Analyses
Masaryk University

Proceedings of the 7th Summer School on Computational Biology

Biodiversity: from Genetics to Geography, from Mathematics to Management

15–17 September 2011
Lednice, Czech Republic

Editor:
Jiří Jarkovský

esf european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitivness
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTMENTS IN EDUCATION DEVELOPMENT

Proceedings of the 7th Summer School on Computational Biology
Biodiversity: from Genetics to Geography, from Mathematics to Management

Editor: Jiří Jarkovský
Cover: Radim Šustr

Published by AKADEMICKÉ NAKLADATELSTVÍ CERM, s.r.o. Brno
Purkyňova 95a, 612 00 Brno
www.cerm.cz

Printed by FINAL TISK s.r.o. Olomučany
1st edition, 2011

ISBN 978-80-7204-756-7

Contents

Foreword

Computational Biology is a modern field of study at the Faculty of Science of Masaryk University (MU). The study programme is guaranteed by the Institute of Biostatistics and Analyses (IBA), which provides computationally oriented courses within the educational concept of the Faculty of Medicine and the Faculty of Science.

The summer schools on Computational Biology are expected to encourage the collaboration among professors, young scientists as well as students of computational biology. Students can participate in informal discussions about novel methods in their field of study and some of them seize this ideal opportunity to the presentation of their own results to the audience. An active contribution of advanced students makes a substantial part of the summer school´s programme.

IBA has initiated a yearly tradition of informal summer schools focused on various aspects of computational science in biology and biomedicine:

2005 – Computational Biology
2006 – Predictive Modelling and ICT in Environmental Epidemiology
2007 – Processing and Analysis of Biodiversity Data: from Genomic Diversity to Ecosystem Structure
2008 – Statistical Methods for Genetic and Molecular Data
2009 – Analysis of Clinical and Biomedical Data in an Interdisciplinary Approach (in Czech]
2010 – Deterministic and Stochastic Modelling in Biology and Medicine

Main objective of the 7th Summer School on Computational Biology is introduction of diversity assessment methodology and its use in various fields of biology from genetics to geographical distribution of organisms, as well as connection to other disciplines from mathematical background to environmental protection management. We hope this specific field of application of computational biology will bring some new viewpoints and experience for all participants.

We are also very grateful for the financial support of the Ministry of Education, Youth and Sports of the Czech Republic, project CZ.1.07/2.2.00/07.0318, Multidisciplinary Innovation of Study in Computational Biology, where this summer school is organized.

On behalf of the programme and organizing committee,

Brno, August 19, 2011

Jiří Jarkovský

Biodiversity: from Genetics to Geography, from Mathematics to Management

Lectures

Evolution of Parasite Diversity: the Importance of Host Genetic Variability

Andrea Šimková, Mária Seifertová

Department of Botany and Zoology, Faculty of Science, Masaryk University, Kotlářská 2, 61137 Brno

Abstract. The first part of this contribution is aimed to introduce three biogeographical hypotheses of species diversity and to present the study analyzing parasite diversity in chub (Squalius cephalus), the cyprinid fish species, based on these hypotheses. The importance of host genetic distance was shown when testing distance decay hypothesis. The second part of this contribution is aimed to introduce coevolutionary hypotheses explaining the associations between major histocompatibility complex (MHC) representing an important component of host immune system and parasite diversity. The models of selection were introduced and the selective pressure acting on MHC diversity in chub was analyzed. Finally, the potential associations between MHC and parasites were studied based on the prediction of heterozygote advantage and rare allele advantage hypotheses.

Keywords: parasite diversity, biogeography, genetic distance, host-parasite coevolution, selection, immune genes

1 Biogeographical patterns of parasite diversity

1.1 Introduction

The diversity and similarity of parasite communities is a result of many determinants widely considered in parasite ecology. The present-day composition and biological diversity of parasite communities are the result of losses and acquisitions of parasite species during the evolutionary history of their hosts [1]. Environmental factors, host ecological traits such as diet and body size, and geographical range are also important determinants of parasite communities [2]. Recently, several ecological studies have emphasized the role of geographical distance between host populations in determining the similarity of parasite assemblages (e.g. [3,4,5]).

Three hypotheses are applicable to the analyses of *biogeographical gradients of parasite biodiversity*: (1) latitudinal gradient, (2) a 'favourable centre' model versus 'local oasis' model, and (3) distance decay, i.e. the role of geographical distance between host populations in the structure (similarity and dissimilarity) of parasite communities.

Following general ecological theory regarding increased biodiversity in the tropics, the analyses of parasite assemblages of fish in these areas suggest that parasite communities exhibit higher diversity in tropical latitudes due to higher evolutionary rates [6]. Temperature, that shows a consistent pattern with the *latitudinal gradient* of parasite diversity, is considered a major biogeographical factor influencing parasite diversity [1,6].

Differences in parasite biodiversity across the host's geographical range may be explained by the '*favourable centre'* model, which is based on the assumption that species abundance is greatest at the centre of the geographical range of the host and declines toward the margins. The optimal conditions for survival and reproduction of a species are supposed at the centre of species range. This hypothesis was also reformulated as '*abundance optimum*' model based on the assumption that species abundance peaks in the locality with the most favourable conditions. When increasing distance from this optimal site the environmental conditions for survival and reproduction are less favourable and thus, population size declines. The 'favourable centre' model and 'abundance optimum' model predict the unimodal distribution of species abundance in space (Fig. 1), whist the multimodal distribution is predicted by the '*local oasis*' model [7] due to changing environmental conditions over space or time.

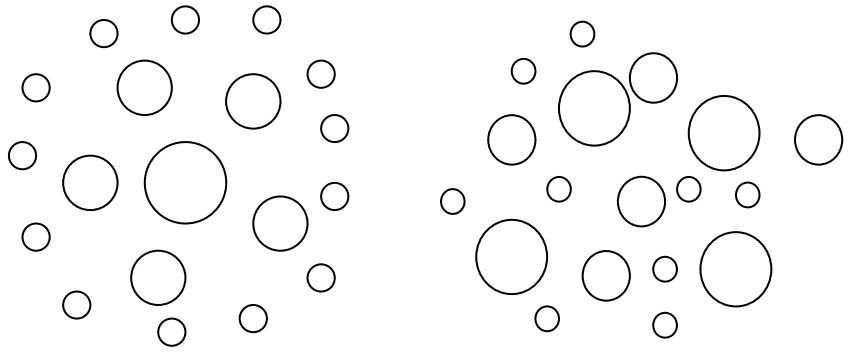


Fig. 1. Abundance optimum model based on unimodal distribution of species abundance (A) and local oasis model based on multimodal distribution of species abundance (B)

The third hypothesis represents decay of similarity in species composition with distance ('*distance decay*'). This decay results from the general prediction that biological similarity decreases with increased geographical distance but also other processes such as landscape topography and spatial configuration, different dispersal of organisms (for instance limited by geographical barriers), and different ability of organisms to survive along climatic and environmental gradient (Fig. 2). Finally, the decay in similarity in species composition may be the result of ecological drift, random dispersal and random speciation according to a neutral theory of biodiversity and biogeography, rather than by environmental heterogeneity. There are many empirical studies demonstrating the negative relationships between parasite community similarity and geographic distance between host populations (e.g. [3,4,8]).

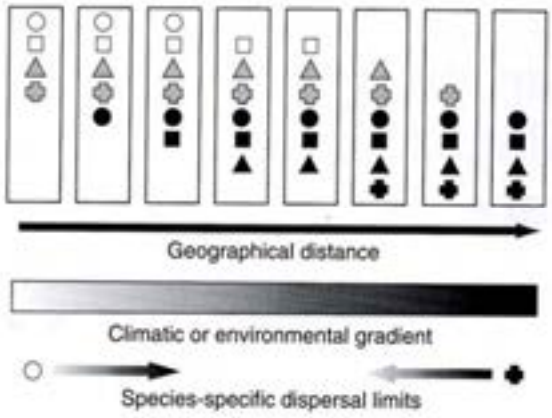


Fig. 2. Schematic representation of eight communities arranged in order of increasing distance. The species present in each community are indicated by different symbols. The communities with small distance share many species, distant communities share few species.

1.2 Empirical study of metazoan parasite diversity in freshwater fish to test the biogeographical hypotheses of parasite diversity

The metazoan parasite communities were studied in 15 populations of chub (*Squalius cephalus*) across much of its geographical distribution in Europe [9]. The number of parasite species per population varied from 5 to 22. Similarity in parasite communities between fish populations was determined using the qualitative Jaccard index on the presence/absence matrix or the quantitative Morisita index on abundance data. The genetic i.e. phylogenetic distance between host populations were calculated using data on cytochrome *b*. Chub as a common European cyprinid species infected by a wide range of metazoan parasites was used as a suitable model species to test three biogeographical hypotheses of parasite diversity.

Concerning the hypothesis of latitudinal gradient hypothesis, it was tested separately for each parasite species. Using a meta-analytical approach, only the abundance of metarcercariae of *Diplostomum* sp. (larval stages of Digenea parasitizing fish eyes) was significantly correlated with latitude ($p=0.007$) but no parasite abundance was significantly correlated with water temperature ($p>0.05$). The absence of a latitudinal gradient in ectoparasitic gill Monogenea may be explained by the fact that higher monogenean species diversity and abundance was recorded in the central regions than in the marginal zones of chub distribution documented by this study.

No general support for the 'abundance optimum' model was found because the majority of correlations were not significant. However, all parasite species exhibited a trend of negative correlation between prevalence and/or abundance and geographical distance from the locality with maximum prevalence. A significant decrease of abundance with an increase of distance from the locality from the most favourable locality was found for ectoparasitic monogeneans of chub (monogeneans were the

most abundant parasite group, see Fig. 3), but not for any endoparasites. This suggests that the pattern of 'abundance optimum' may be associated with level of host specificity i.e. the pattern may hold for highly specific monogeneans (i.e. parasitizing a single host species) because their abundance is not affected by the distribution of intermediate and definitive hosts like in endoparasite species.

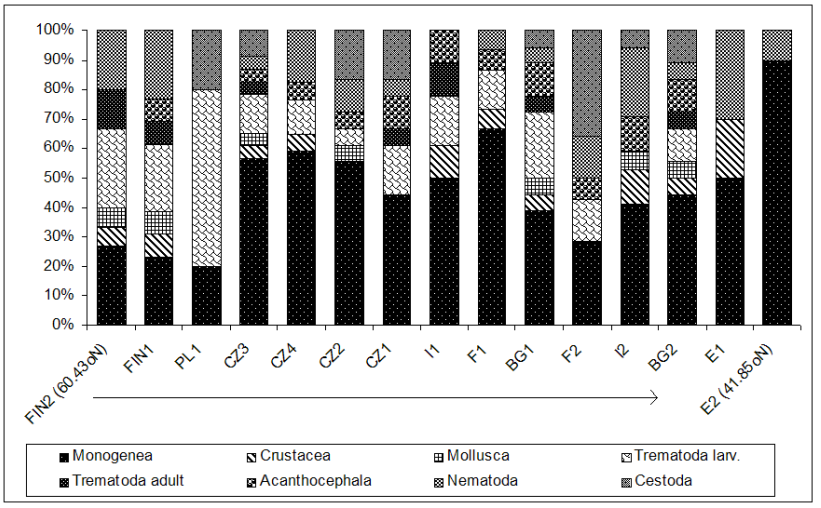


Fig. 3. The proportion of different metazoan groups in the 15 localities studied. Populations are presented according to latitude.

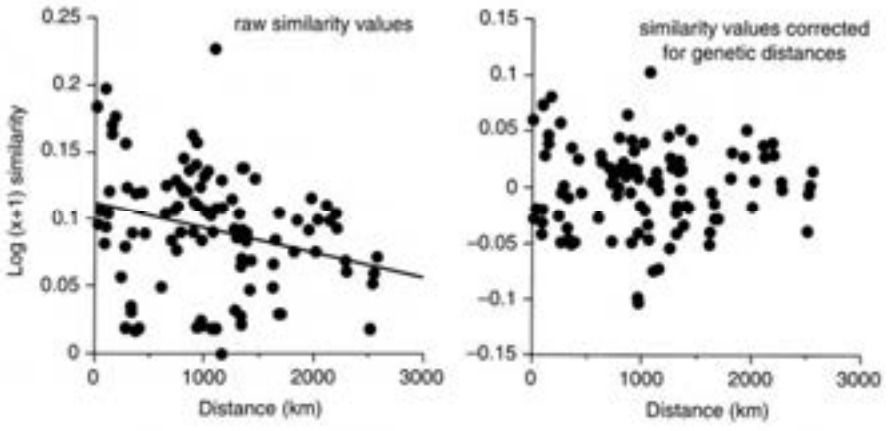


Fig. 4. Distance decay of similarity in metazoan parasite communities of chub. When raw similarity values are plotter against geographic distance, the negative relationship is observed. When similarity values are corrected for genetic distance estimated from cytochrome b sequences, the relationship is not observed.

Using geographical distance is a convenient ecological measure to test ‘distance decay’ hypothesis. However, the presence of parasite species is affected by the current and historical host movements and therefore the *genetic distance between host populations* could provide a more accurate estimate of host movements than inferred from geographical distance. Thus, we looked at both phylogenetic and geographic distance between host populations, and found that in a multivariate analysis, host phylogenetic distances were the only significant determinant of similarity in parasite species composition. After controlling for the influence of phylogenetic distance, no relationship between geographic distance and similarity in parasite communities were found (Fig. 4).

2 Coevolutionary relationships between parasite diversity and host immune system

2.1 Introduction

The capacity of host to control parasite infections is mainly dependent on particular immune genes. Among them, the major histocompatibility complex (MHC) is a multi-gene family controlling immunological self and non-self recognition in vertebrates. MHC genes encode cell surface glycoproteins that present foreign peptides and self-peptides to T lymphocytes, thereby controlling all specific immune response, both cell and antibody mediated [10]. MHC genes are under selective constraints that contribute to maintaining the remarkable high polymorphism at MHC loci. The extensive polymorphism in the MHC genes is especially pronounced in the codons encoding the peptide binding regions (PBR) of the MHC molecule. Gene duplication and inter- and intra-locus recombination, spatially heterogeneous selective pressures and reproductive mechanisms including MHC-based mating preferences, selective abortion and 'allele counting' strategy are considered as the possible mechanisms maintaining MHC diversity. Nevertheless, '*parasite-driven balancing selection*' based on the effects of host-parasite co-evolution leading to '*arms races*' between the immune defense of hosts and the virulence of parasites has been considered as one of the main evolutionary mechanisms maintaining a high MHC polymorphism in wild populations. Parasite-driven balancing selection is explained by two hypotheses [11-14]. First one represents the '*frequency-dependent selection*' hypothesis or rare-allele advantage hypothesis based on the prediction that host genotypes (i.e. frequencies of MHC alleles) constantly change with the frequency of adapted and non-adapted pathogens. It means that host genotypes with a rare allele have a stronger selective advantage and responds better to a new pathogen and therefore they become more frequent; but this is followed by a decrease in its fitness as pathogens adapt to infect the most common host genotype. Second hypothesis is termed as the '*overdominance hypothesis*' or heterozygote advantage hypothesis, which is based on the advantage of being heterozygotes at MHC genes, which permits to recognize a wider range of antigens than in the case of homozygotes.

Chub (*Squalius cephalus*) was used (1) to analyze the role of *selection in forming the diversity of MHC genes* and (2) to investigate the potential *coevolution between MHC and parasite species* at population level. The study was focused on the highly polymorphic exon 2 of MHC class II*B* which includes PBR sites. Two groups of closely linked exon 2 sequences, *DAB1*-like and *DAB3*-like (belonging to MHC IIB genes), were identified.

2.2 Testing the positive selection on MHC genes

The detection of species-specific positively selected sites (PSS) (i.e. sites under positive selection) by means of maximum likelihood methods has became the common approach in recent MHC studies of natural populations of wild living animals. The presence of selection in *DAB* genes of chub was analyzed using the maximum likelihood method in the program CODEML implemented in PAML, version 4.3 [15]. The different models with and without selection incorporated were used to test for the presence of sites under selection and to identify them. The models used the non-synonymous/synonymous rate ratio ($\omega = d_N / d_S$) as an indicator of selective pressure on the protein. Simplistically, values of $\omega < 1$, $= 1$, and > 1 means negative purifying selection, neutral evolution, and positive selection (balancing selection). However, the ratio averaged over all sites is almost never > 1, since positive selection is unlikely to affect all sites over prolonged time. Thus, the interest has been focused on detecting positive selection that affects only some sites. We compared the following models: M0 (one ratio) and M3 (discrete model involving eight classes for ω) - this test is considered as a test of variable ω among sites rather than a test of positive selection, M1a (nearly neutral) and M2a (positive selection), and M7 (β model which uses beta distribution) and M8 (β and ω). Parameters in the site models are shown in *Table 1*. The comparison of M1a and M2a models have limitations in the presence of recombination, while the comparison of M7 and M8 models is robust against the effect of recombination [16]. If alternative models M3, M2a and M8 suggest the presence of sites with $\omega > 1$, all three tests can be considered a test of positive selection [17]. A likelihood ratio test (LRT) statistic (twice the log-likelihood difference between the two compared models ($2\Delta l = 2(l_b - l_a)$) compared with a $\chi 2$ distribution with P_b-P_a degrees of freedom) was used to assess the significance of the differences between models (l_a and l_b are log-likelihood values and P_a and P_b are the number of parameters for each of the models being compared). When the LRT indicated that there was a significant difference, the Bayes empirical Bayes (BEB) method was used to calculate the posterior probabilities (pP) for site classes and to identify sites under selection (the posterior means of ω for positively selected sites are > 1). BEB is implemented under models M2a and M8 only.

Using CODEML, maximum likelihood parameters under different codon models of variable ω across sites in the *DAB1* and *DAB3* datasets were estimated [18]. The LRT statistic comparing the two models shows that the alternative models (M2a, M3 and M8) fit the data significantly better than simpler models M1a, M0 and M7 ($p < 0.001$), which indicates the action of positive selection at specific sites in *DAB* sequences. The variability in selective pressure among sites in the exon 2 and the presence of a number of sites under balancing selection (i.e. positively selected sites)

is shown in Fig. 5. The comparison of positively selected sites in exon 2 of *DAB1*-like and *DAB3*-like sequences showed the differences in evolutionary patterns between *DAB1*-like and *DAB3*-like genes despite their close linkage. This finding suggests potential structural and functional differences between *DAB1*-like and *DAB3*-like genes.

Table 1. Parameters in the site models (according to Yang [15])

Model	NSsites	#p	Parameters
M0 (one ratio)	0	1	ω
M1a (neutral)	1	2	*p0 (p1 = 1 – p0),* ω0< 1, ω1= 1
M2a (selection)	2	4	*p0, p1 (p2 = 1 – p0 – p1),* ω0< 1, ω1= 1, ω2> 1
M3 (discrete)	3	5	*p0, p1 (p2 = 1 – p0 – p1)* ω0, ω1, ω2
M7 (beta)	7	2	*p, q*
M8 (beta&ω)	8	4	*p0 (p1 = 1 – p0), p, q, ωs> 1*

#p is the number of free parameters in the ω distribution. Parameters in parentheses are not free and should not be counted: for example, in M1a, $p1$ is not a free parameter as $p1 = 1 - p0$. In both likelihood ratio tests comparing M1a against M2a and M7 against M8, df = 2. The site models are specified using NSsites.

2.3 Similarity between host populations based on the genetic variability and parasites

The geographically distant populations were more variable in their microsatellites and more dissimilar in their parasite composition (*Table 2*). Concerning MHC diversity, populations with dissimilar MHC allelic profiles were geographically distant populations with significantly different variability in microsatellites and a dissimilar composition of parasite communities. Significant positive correlations were found between MHC distance and metazoan parasite similarity based on abundance data (Morisita index). However, no significant correlation was found between MHC variability measured by amino acid distance and microsatellite distance. Multiple regression analysis with backward elimination calculated using a permutation method, showed that only geographic distance (b = -0.535, $P < 0.001$) and microsatellite variability (b = -0.241, $P = 0.004$) had a statistically significant contribution to MHC similarity (N = 105, $R^2 = 0.465$, $P < 0.001$).

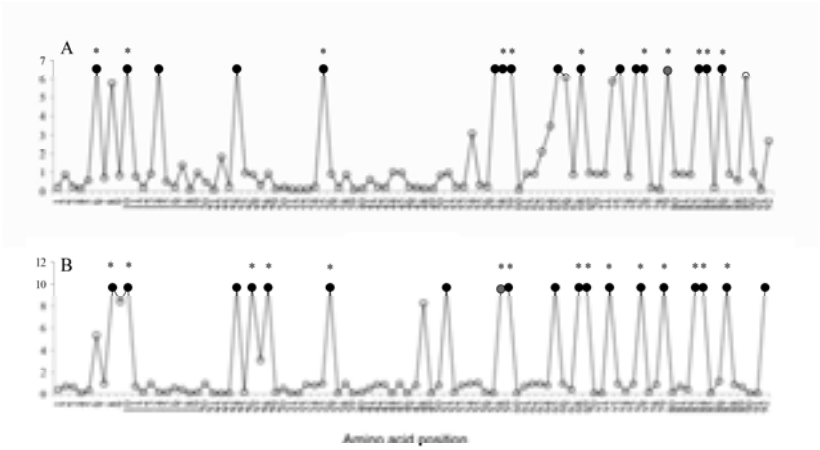


Fig. 5. Representation of DAB1-like (A) and DAB3-like (B) sequence variants. Approximate posterior means of ω, calculated as the weighted average of ω over the 11 site classes, weighted by the posterior probabilities under the random sites model M8 (β and ω) are shown. Sites picked, representing a target of positive selection at the 99% level, are indicated by black circles, and those at the 95% level by gray circles.

Table 2. Summary of the Mantel test performed in R v. 2.9.1 using 15 chub populations. The values of the Spearman correlation coefficient are shown below the diagonal and P - values (10,000 permutations) for each pair-wise comparison are shown above the diagonal. Significant correlations are shown in italic.

	A	**B**	**C**	**D**	**E**	**F**
A: Genetic distance (microsatellites)		0.005	0.032	0.001	0.001	0.148
B: Parasite presence (Jaccard index)	*-0.419*		0.009	0.034	0.039	0.077
C: Parasite abundance (Morisita index)	*-0.273*	*0.397*		0.012	0.043	0.039
D: Geographic distance	*0.578*	*-0.292*	*-0.304*		0.009	0.409
E: Presence of MHC alleles	*-0.465*	*0.288*	*0.236*	*-0.353*		0.381
F: Amino acid distance (MHC variability)	-0.132	0.201	*0.204*	0.021	0.045	

2.4 MHC diversity: parasite-mediated or neutral selection?

The ANOVA revealed significant differences in metazoan parasite load between fish individuals expressing a different number of *DAB* alleles. Fish with a higher number of *DAB* alleles (more than 2 alleles) have a significantly higher abundance and species richness of metazoan parasites (Fig. 6). GLM analysis using univariate models

revealed the significant influence of population effect, ectoparasite abundance and species richness, endoparasite abundance and species richness, and microsatellite variability on MHC diversity expressed by amino acid distance. However, using multivariate models, only the population effect had a significant influence on MHC variability (*Table 3*). Thus, MHC diversity was confounded by the population effect, potentially related to specific habitat character and/or linked to population phylogeny. Even initially a link between individual MHC diversity and parasitism was found; both parasitism and microsatellite variability explained the very low (i.e. non-significant) proportion of MHC diversity at population level.

Table 3. Variability sources explaining the individual MHC amino-acid distance (only non zero values were included)

Source	Variability explained (%)	p
Univariate models		
Population	57.00%	*<0.001*
Brillouin index diversity	0.20%	0.577
Total parasite abundance (ln)	0.90%	0.210
Ectoparasite abundance (ln)	5.70%	*0.002*
Endoparasite abundance (ln)	8.40%	*<0.001*
Total parasite species (ln)	0.60%	0.309
Ectoparasite species (ln)	8.60%	*<0.001*
Endoparasite species (ln)	10.70%	*<0.001*
Microsatellite variability	7.90%	*<0.001*
Multivariate models		
Population	57.61%	*<0.001*
Microsatellite variability	1.20%	0.077
Total parasite species (ln)	0.59%	0.216
Total parasite abundance (ln)	0.34%	0.348
Brillouin index diversity	0.17%	0.507
Whole model	59.90%	
Population	56.35%	*<0.001*
Microsatellite variability	1.66%	0.053
Ectoparasite abundance (ln)	0.67%	0.218
Ectoparasite species (ln)	0.02%	0.821
Whole model	58.70%	
Population	57.31%	*<0.001*
Microsatellite variability	1.12%	0.077
Endoparasite abundance (ln)	0.75%	0.149
Endoparasite species (ln)	0.13%	0.550
Whole model	59.30%	

2.5 Specific associations between MHC alleles and parasite species

The multivariate co-inertia analysis, COIA [19], is a statistical method applied to investigate the relationships between genetic matrix including the presence/absence of each MHC allele and parasite matrix including the abundance of each metazoan parasite species [20]. The first step of COIA involved the separate analyses of each matrix, i.e. the analysis of the genetic matrix using a principal components analysis

(PCA) and the analysis of parasite matrix using a correspondence analysis (CA). Thereafter, the co-inertia (COIA) analysis of the two matrices was performed. The PCA based on the MHC class II*B* allele matrix (presence/absence data) revealed that the first two axes explained 25.6% of the variance in the data (F1: 14.2%, F2: 11.4%). The CA based on the abundance of metazoan parasite species revealed that the first two axes explained 31.5% of the total variance in the data (F1: 17.1%, F2: 14.4%). MHC alleles and metazoan parasite species exhibited significant covariance in the COIA model (global co-inertia = 0.725, $P < 0.05$). The first two axes of the COIA model explained 39% of the variance shared between the MHC and metazoan parasite matrices (F1: 21%, F2: 18%). The co-structure of MHC and metazoan parasite variables on the COIA factor maps (Fig. 7) indicated four groups of population-specific alleles separated by the first two axes, i.e. two groups of alleles specific to Czech populations, the group of alleles specific to Italian populations and the group of alleles specific to Finnish populations (Fig. 7A), and the same four groups of parasite species-specific populations (Fig. 7B). Several associations between population specific alleles and population specific parasite species were identified along the first and second axes (Fig. 7).

Following COIA analysis, *DAB3* alleles were more involved in the parasite-MHC allele associations compared to *DAB1* alleles. This could suggest that the rare-allele advantage as a mechanism of parasite-mediated selection drives the diversity of *DAB3*-like genes in European chub populations. The analyses of structure and selection patterns in *DAB1*-like and *DAB3*-like genes in chub showed that *DAB3*-like genes are under stronger positive selection compared to *DAB1*-like genes which suggests potential structural and functional differences between two groups of closely linked genes (see above). However, the results suggesting the potential associations between specific *DAB* alleles and parasites species should be interpreted carefully. These associations could either indicate negative frequency dependent selection acting on MHC diversity within four groups of chub populations or alternatively imply that the presence of parasites results from specific characteristics of those regions (for instance the presence of intermediate hosts or the specific character of the habitat) or the biogeographical gradient of parasite distribution.

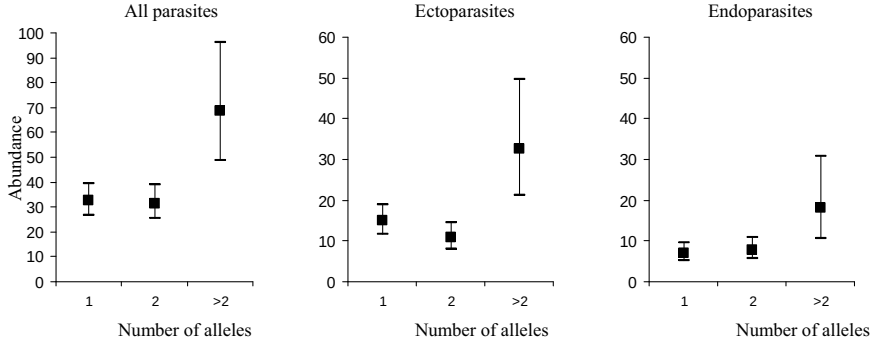


Fig. 6. The relationships between parasite species richness (A – total parasites, B - ectoparasites, C- endoparasites) and number of DAB alleles. p - statistical significance is based on ANOVA, a,b - homogeneous groups based on Tukey post hoc test

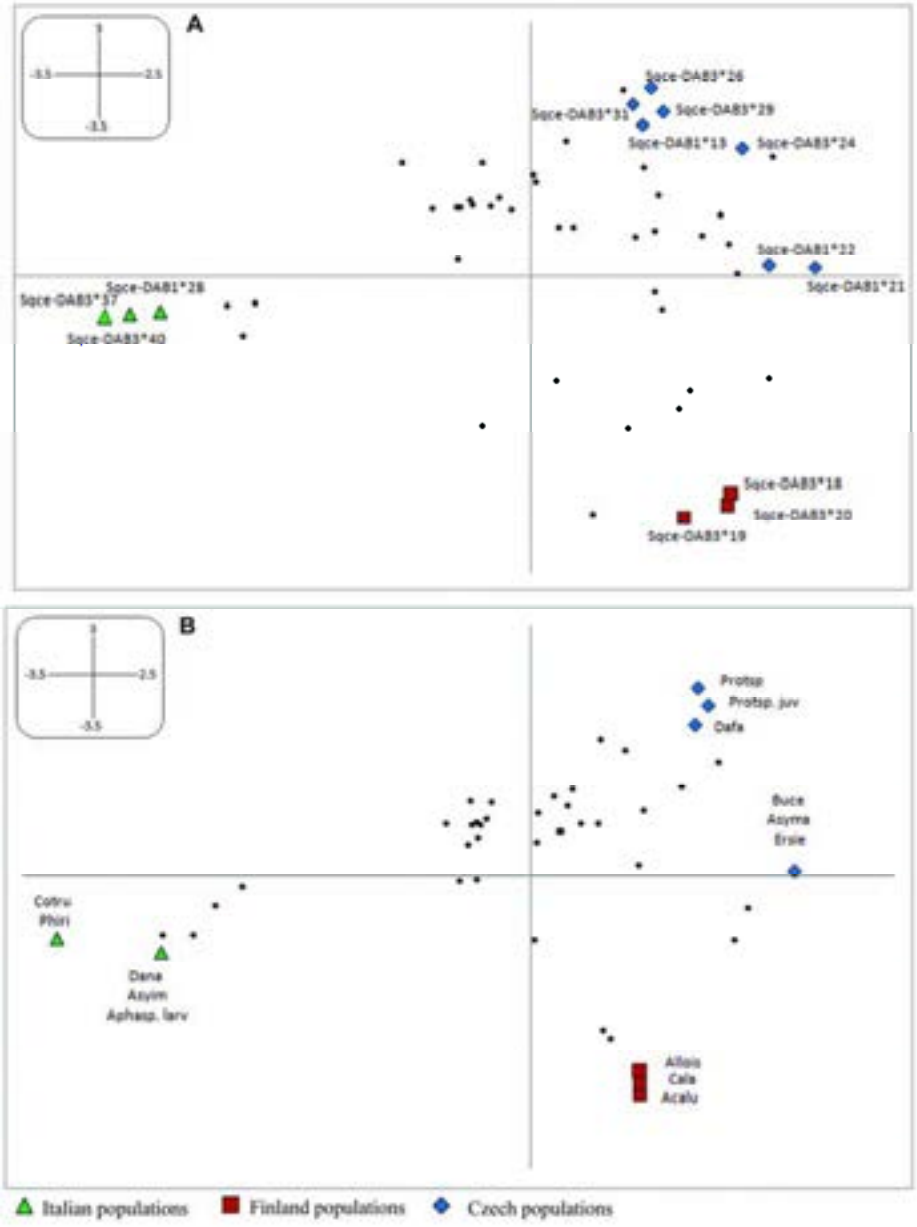


Fig. 7. Co-inertia analyses of MHC alleles (A) and metazoan parasite species (B) in chub populations. Only positively associated variables (i.e. MHC alleles susceptible to metazoan parasites) are labelled; population specificity is indicated by different symbols.

Acknowledgments. This study was funded by the Grant Agency of the Czech Republic, project No. 524/07/0188. MS was supported by the Ichthyoparasitology Research Centre of the Ministry of Education, Youth and Sports of the Czech

Republic LC 522 and partially by the Rector's Programme in Support of MU Students' Creative Activities. AŠ was supported by the Research Project of Masaryk University (No. MSM0021622416).

References

1. Poulin, R., Rohde, K.: Comparing the richness of metazoan ectoparasite communities of marine fishes: controlling for host phylogeny. Oecologia 110, 278--283 (1997)
2. Poulin, R.: Species richness of parasite assemblages: evolution and patterns. Ann. Rev. Ecol. Syst. 28, 341--358 (1997)
3. Poulin, R., Morand, S.: Geographical distances and the similarity among parasite communities of conspecific host populations. Parasitology 119, 369--374 (1999)
4. Poulin, R.: The decay of similarity with geographical distance in parasite communities of vertebrate hosts. J. Biogeogr. 30, 1609--1615 (2003)
5. Poulin, R.: The structure of parasite communities in fish hosts: ecology meets geography and climate. Parassitologia 49, 169--172 (2007)
6. Rohde, K.: Latitudinal gradients in species diversity: the search for the primary cause. Oikos 65, 514--527 (1992)
7. Poulin, R., Dick, T. A.: Spatial variation in population density across the geographical range in helminth parasites of yellow perch Perca flavescens. Ecography 30, 629--636 (2007)
8. Oliva, M. E., González, M. T.: The decay of similarity over geographical distance in parasite communities of marine fishes. J. Biogeogr. 32, 1327--1332 (2005)
9. Seifertová, M., Vyskočilová, M., Morand, S. and Šimková, A.: Metazoan parasites of freshwater cyprinid fish (Leuciscus cephalus): testing biogeographical hypotheses of species diversity. Parasitology 135, 1417--1435 (2008)
10. Klein, J.: Origin of major histocompatibility complex polymorphism: The trans-species hypothesis. Hum. Immunol. 19, 155--162 (1987)
11. Klein, J., Figueroa, F.: The evolution of class I MHC genes. Immunol. Today 7, 41--44 (1986)
12. Klein, J.: Of HLA, tryps, and selection: an essay on coevolution of MHC and parasites. Hum. Immunol. 30, 247--58 (1991)
13. Hughes, A.L., Nei, M.: Maintenance of MHC polymorphism. Nature 355, 402--403 (1992)
14. Hedrick, P.W.: Pathogen resistance and genetic variation at MHC loci. Evolution 56, 1902--8 (2002)
15. Yang, Z.H.: PAML4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586--1591(2007)
16. Anisimova, M., Nielsen, R., Yang, Z.H.: Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164 (3),1229--1236 (2003)
17. Yang, Z.H., Nielsen, R., Goldman, N., Pedersen, A.M.K.: Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155 (1), 431--449 (2000)
18. Seifertová, M., Šimková, A.: Structure, diversity and evolutionary patterns of expressed MHC class IIB genes in chub (Squalius cephalus), a cyprinid fish species from Europe. Immunogenetics 63(3), 167--181 (2011)
19. Doledec, S., Chessel, D.: Co-inertia analysis - an alternative method for studying species environment relationships. Fresh. Biol. 31, 277--294 (1994)
20. Tollenaere, C., Bryja, J., Galan, M., Cadet, P., Deter, et al. Multiple parasites mediate balancing selection at two MHC class II genes in the fossorial water vole: Insights from multivariate analyses and population genetics. J. Evol. Biol. 21: 1307--1320 (2008)

Genetic Diversity in Populations

Natália Martínková[1,2], Barbora Zemanová[2]

[1] Institute of Biostatistics and Analyses, Masaryk University, Brno, Czech Republic
[2] Institute of Vertebrate Biology, Academy of Sciences, v.v.i., Brno, Czech Republic

Abstract. Genetic diversity accumulates over time on the level of DNA sequence with accumulation of mutations. It is additionally increased with population admixture, and the decrease in genetic diversity is often the first indication of detrimental processes affecting populations, such as reduction in the number of breeding individuals or breeding of close relatives. Nucleotide diversity shows how different the sequences of a given gene are in a population. Gene diversity estimates how likely two individuals are to share the same sequence of a gene. Number of alleles is the count of different versions of a gene with differing sequences, and this is corrected for sample size in estimation of allelic richness. Heterozygosity is the average frequency that an individual will have two different copies of a gene.

Keywords: nucleotide diversity, gene diversity, haplotype diversity, allelic richness, heterozygosity.

1 Introduction

Genetic diversity expresses the amount of genetic variation, which is a background for evolutionary processes. It accumulates in time. On the level of DNA sequence, the diversity increases with occurrence of mutations and fluctuates based on the size of the population.

Molecule of the DNA is composed of four nucleotide bases, where their composition and order define information encoded in the molecule. This information changes with time when changes in sequence of nucleotides are introduced with mutations. Point mutations affect a single nucleotide base, which might be replaced by a different base, or a base might be deleted or inserted. Base replacements are called substitutions. Deletions and insertions are referred to by a common term – indel. Mutations occur as errors in copying the DNA molecule during the process of cell division. They occur in all cells, but in multicellular organisms, only mutations that affect DNA in gametes or spores will be transferred to the next generation. Those will contribute to the observed genetic diversity of the next generation and will enable the reconstruction of the evolutionary process.

Organisms that share the same sequence of a specific genomic region, a gene or a locus, are referred to as sharing the same allele. In case of a haploid genome, in gametes, mitochondria or chloroplasts, identical sequences are referred to as the same haplotype.

Measuring genetic diversity is always a comparison between more than one sequences. Changes detected within a group of sequences represent diversity, and comparisons between the groups or even between individual sequences are the measurements of divergence.

2 Diversity Indices

Genetic diversity indices that help describe population structure on DNA sequence level include number of differences between sequences, nucleotide diversity, gene diversity, and on the level of whole genes or other markers, number of alleles, allelic richness and heterozygosity. These indices are then used to model evolutionary processes that shape population structure, such as reconstruction of colonization pathways, historical demographic changes or signal for selection.

The number of differences between sequences represents the basic measure of genetic distance. As the DNA molecule consists of only four nucleotide bases, multiple mutations might accumulate at any given site. The occurrence of multiple mutations would then underestimate genetic distance and needs to be corrected using an appropriate substitution model. This was addressed in detail at the 4th Summer School [1] and it affects predominantly sequence datasets that are expected to contain old divergence events. Multiple mutations are very rare in recent history and are usually ignored in studying population processes during human history.

2.1 Nucleotide Diversity - π

Nucleotide diversity is the average proportion of substitutions observed between any two randomly chosen sequences within a group of sequences. It is often expressed as percent, and for any given population, nucleotide diversity is different for different loci.

It is calculated as the sum of all per site numbers of differences between unique pairs of haplotypes in a group of sequences.

$$\pi = n/(n-1) \sum_{ij} p_i p_j \pi_{ij} \quad \textbf{(1)}$$

where n is the number of sequences, p_i and p_j are frequencies of haplotype i and j in the dataset and π_{ij} is the number of nucleotide differences per site between i and j.

It shows how different sequences belonging to a group are expected to be (Fig. 1).

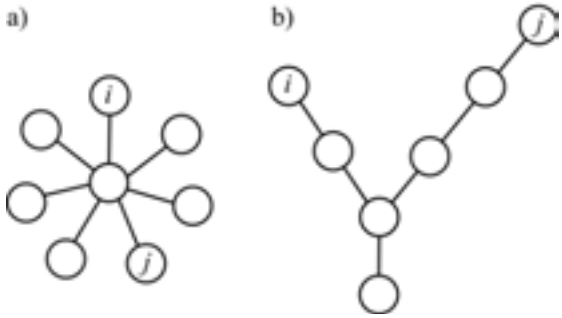


Fig. 1. Haplotype networks indicating alternative relationships of DNA sequences that would attain different nucleotide diversity. Both datasets contain eight unique sequences (circles), but the number of differences between *i* and *j*, represented by edges is smaller in (a) than in (b). The specific values would depend on sequence length, as π_{ij} in formula 1 is a per site number of differences between specific sequences.

2.2 Gene or Haplotype Diversity - *h*

Gene diversity represents a measure of what proportion of individuals in a group or a population share the same haplotype or allele. The higher the gene diversity, the greater the chance that two individuals randomly sampled from a population would have different sequence of a target gene.

$$h = 1 - \sum_i p_i^2 \qquad (2)$$

where p_i is the frequency of haplotype *i* in the dataset.

Datasets displayed in Fig. 1 would both have $h = 1$, because each sequence differs from all others. The number of differences between sequences is not reflected in this index. The gene diversity is dependent on the number of different haplotypes in a population, but also on frequency with which the haplotypes occur. Populations where a single haplotypes dominates are probably rapidly growing or selective pressure on the gene limits genetic variability of the gene in the population, whereas a population where haplotypes occur with more evenly distributed frequency were most probably stable in numbers in recent history and the gene in question evolves neutrally (Fig. 2).

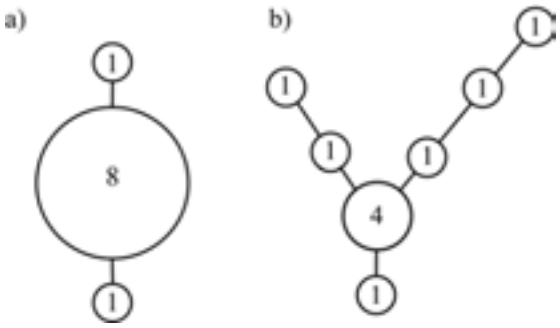


Fig. 2. Haplotype networks indicating alternative population structure with varying frequency of specific haplotypes (indicated by numbers in circles). Haplotype diversity in (a) would be $h = 1 - [(1/10)^2 + (8/10)^2 + (1/10)^2] = 0.34$, whereas in (b) $h = 1 - [6 * (1/10)^2 + (4/10)^2] = 0.78$. Note that π would be smaller in (a) than in (b).

Nucleotide and haplotype diversity can be calculated from a sequence alignment in program DnaSP (http://www.ub.edu/dnasp/).

2.3 Number of Alleles and Allelic Richness

In diploid datasets, one individual might inherit a copy of each gene from each parent and these copies might differ. The studies that utilize this fact analyze occurrence of different alleles with different specific DNA sequence. The data are differentiated on the gene or locus level, not the nucleotide level. This means that the number of substitutions or indels in the DNA sequence distinguishing each allele is irrelevant.

Number of alleles is a simple count of different sequences. In laboratory analyses, allele sequences are often estimated on the level of an individual. The gene sequence would contain ambiguous bases in positions, where the alleles differ. To obtain a haplotype, the exact gametic phase of the allele, the ambiguities need to be resolved. Experimentally, the gametic phases – or haplotypes, as the gametes are haploid – are sequenced from molecular clones of the alleles. From population datasets, gametic phases can be reconstructed using a computational algorithm.

In Fig. 1, both panels contain the same number of different haplotypes. In Fig. 2, panel (a) contains 3 haplotypes, whereas panel (b) contains 7. But in these situations, all samples, represented by the figure panels, have comparable size. If we would draw a random subsample of 3 sequences from each of the datasets, we would underestimate the number of alleles in the dataset. In all cases, the true number of alleles will be higher than our sample size with the exception of that from Fig. 2a. If the sampling was random, we might miss the rare haplotypes even in Fig. 2a. The number of discovered alleles would grow with increasing sample size, although not indefinitely. The number of alleles needs to be corrected for sample size. The number of alleles expected to be found in n individuals, $\tau(n)$, is modeled by allelic accumulation function.

$$\tau(n) = \sum_i [1 - (1 - \psi_i)^n] \quad \textbf{(3)}$$

where n is the number of individuals in the dataset and ψ_i is the probability that an allele is present in an individual i.

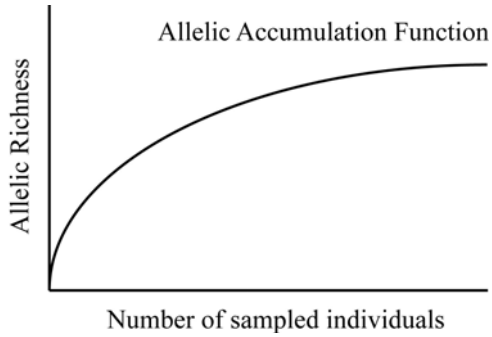


Fig. 3. The relationship between number of individuals sampled from a population and allelic richness. The number of alleles discovered in a population would be higher for larger samples, allelic richness then corrects for differences in sample size.

Allelic richness can be calculated for a dataset that contains a list of individuals sampled in a given population for which alleles of a selected marker are known. Software FSTAT (http://www2.unil.ch/popgen/softwares/fstat.htm) can be used to calculate allelic richness.

2.4 Heterozygosity – H_O, H_E

Heterozygosity is a state when an individual has two different alleles of a gene or locus. The individual is heterozygous. If an individual has both copies of a gene that are identical, they are the same allele, the individual is homozygous. In a population where individuals mate randomly, the heterozygots would occur proportionally to the frequency of different alleles based on Mendelian inheritance. The observed heterozygosity (H_O) within a population is then the proportion of individuals that are found to be heterozygous for the given locus. Over multiple loci, average observed heterozygosity is computed.

When frequency of observed heterozygots in a population matches that expected (H_E) from the frequency of alleles, the population is in Hardy-Weinberg equilibrium (HWE). For the simplest case, where in total two alleles occur in a population, the proportion of heterozygotes expected under HWE would be calculated from the frequency of each allele.

$$p^2 + 2pq + q^2 = 1 \,, \qquad (4)$$

where p is the frequency of the first allele and q is the frequency of the second allele present in the population. HWE for two alleles in a population is a binomial expansion of $(p + q)^2$. For three alleles, HWE would be a trinomial expansion of $(p + q + r)^2$ and so on.

Deviations from HWE is estimated from Pearson's χ^2 test:

$$\chi^2 = \sum_i [(O_i - E_i)^2 / E_i] \quad (5)$$

where O_i is the observed number of individuals with genotype i, E_i is the expected number of individuals with genotype i, and degrees of freedom is the difference between number of genotypes and number of alleles. The null hypothesis is that the population is in HWE.

Observed and expected heterozygosity are calculated in FSTAT listed above or in Genepop software (http://genepop.curtin.edu.au/).

3 Case Study

Genetic diversity naturally increases with time as genetic variability is increased by accumulation of novel mutations and populations admix and exchange genetic information. Changes in evolutionary history of populations might exacerbate or reduce changes in genetic diversity.

Martínková et al. [2] tested genetic diversity of populations of a small terrestrial carnivore, the stoat, on the British Isles. They found that both the nucleotide and haplotype diversity of partial sequences of the mitochondrial genome was higher in stoats from Ireland than in those from Great Britain. In fact, the Irish stoats exhibited genetic diversity comparable to that found in continental Europe. That is not in agreement with expectations where both the Irish and British stoats should have similar genetic diversity that would be smaller than the diversity found on the continent. The authors applied additional analyses and explained the deviation from expectation by differences in evolutionary history between different populations. The stoats from Ireland represent remnants of an ancient population that was able to survive the last glaciation in Ireland, very close to the ice-sheet. Contrary to that, the British stoat population is younger and originated from direct immigrants from continental Europe.

Genetic diversity reflects also changes in population structure caused by humans. In chamois, a mountain ungulate that lives in large mountain ranges in Europe, genetic structure of populations bears signature of historical translocations and introductions [3]. Several populations of chamois were introduced to new areas or translocated to repopulate regions where the species was overexploited over the last century. Those populations have lower allelic richness, heterozygosity and nucleotide diversity. Such populations originated from a small number of founders that harbored limited genetic variation. It is now reflected in similarly limited genetic diversity. The opposite, populations with very high genetic diversity are characterized by consisting of individuals of different geographic origin. Distant populations are expected to be divergent, and if such individuals are introduced to a single population, they would bring divergent haplotypes, increasing the overall genetic diversity of a population.

4 Conclusion

Genetic diversity indices reflect basic genetic structure of a population. Their combination provides valuable information about possible heterogeneity of origin of a population, its age or a signal that first indicates possible problems associated with genetic information such as inbreeding.

Acknowledgment. The authors thank the organizers of the Summer School for their invitation to present in this forum.

References

1. Vallo, P.: Basic Phylogenetic Methods. In: Dušek, L., Haruštiaková, D., Martínková, N. (eds.) Proceedings of the 4th International Summer School on Computational Biology: Statistical Methods for Genetic and Molecular Data, 29-31 May, 2008, pp. 72--82. Institute of Biostatistics and Analyses, Masaryk University, Brno (2008)
2. Martínková, N., McDonald, R.A., Searle, J.B.: Stoats (*Mustela erminea*) Provide Evidence of Natural Overland colonization of Ireland. Proc. Roy. Soc. B. 274, 1387--1393 (2007)
3. Crestanello, B., Pecchioli, E., Vernesi, C., Mona, S., Martínková, N., Janiga, M., Hauffe, H.C., Bertorelle, G.: The Genetic Impact of Translocations and Habitat Fragmentation in Chamois (*Rupicapra*) spp. J. Hered. 100, 691--708 (2009)

Biodiversity: a Principle of Life in the Hands of Computational Science

Jiří Jarkovský[1,2], Ladislav Dušek[1,2], Jana Koptíková [1], Danka Haruštiaková [2]

[1] Institute of Biostatistics and Analyses, Faculty of Science and Faculty of Medicine, Masaryk University, Kamenice 126/3, 625 00 Brno, Czech Republic
dusek@iba.muni.cz
[2] Research Center for Environmental Chemistry and Ecotoxicology, Faculty of Science, Masaryk University, Kamenice 126/3, 634 00 Brno, Czech Republic

Abstract. This paper has been prepared to provide a brief educational overview of biodiversity data as a subject of different types of studies. The biodiversity is defined in all levels of organization of biological systems, from molecular and genomic level to ecosystem scale. A special attention is given to the methodology of different types of analyses, including the widely-used modeling of species-abundance relationships. The analysis of biodiversity is widely available in many software packages and hundreds of measures can be used; however, it must always be done with a very careful and correct interpretation. The so-called dual concept of biodiversity is discussed: (1) the component measuring number of forms (species) in the system and (2) component identifying the quantity (population size) of these forms. The majority of biodiversity measures combine both these components in various kinds of ratio indices. Several key numerical principles underlying large families of biodiversity measures are explained. These are so-called Shannon's concept of biodiversity, principle of dominance, principle of evenness, species-abundance cumulative profiles and niche-oriented modeling. The paper concludes that the remarkable specifics of biodiversity data and biodiversity itself make this field extremely challenging for computational science, including computer-assisted simulations and modern data mining techniques. Furthermore, the estimation of uncertainty that is associated with different biodiversity measures is still not adequately addressed in computationally-oriented literature. This might be extremely important for the application of biodiversity monitoring as a standardized input in ecological risk assessment studies.

Keywords: biodiversity, diversity indices, species-abundance models, niche-oriented models

1 Introduction: biodiversity, its forms and interpretation

The diversity simply means the variety of forms in some examined system, regardless of its biotic or abiotic character. The biodiversity then logically means the variety of biological forms in a biological system. Of course, no one can imagine our world without its biological (natural) variability and that is why we can call the biodiversity "a principle of life". Everything starts from the very origin of the life. The variability

is intrinsically associated with the genome structure, starting even from the elementary codon sequence. The so-called gene polymorphism is defined as a genetic variant that appears in at least 1% of a population. By setting the cut-off at 1%, we exclude the spontaneous mutations that may have occurred in a single family (and spread through the descendants). The gene polymorphism and many other molecular mechanisms involved in genome replication form a base that generates variability or changeability in all subsequent levels of organization of biological systems. It means the variability of forms from the taxonomic point of view, as well as the morphologic, physiologic, metabolic or even behavioral ones. We can conclude that the biodiversity is diverse itself. The modern biology distinctly recognizes many mutations of this phenomenon, each with its very specific consequences and interpretation. To mention the most frequent attributes of the word biodiversity, we must mention the following adjectives: genetic, physiological and eco-physiological, structural, taxonomic, and behavioral.

This paper is focused mainly on the explanation of general, widely accepted, let's say standard, principles of biodiversity. Our primary topic is the taxonomic diversity measuring the variety of species in biological communities or in ecosystems. This field of biology initiated the research of biodiversity in the past and stimulated also first computational attempts to mathematically standardize the evaluation (Pielou, 1969, 1975; Krebs, 1989). However, the world is changing and nowadays we can even read the human genome and map DNA varieties with less than 1 % incidence (Barnhart, 1989; DeLisi, 2001). Using this genome diversity research, we can identify the causes of different characteristics that relate to specific segments of human population as well as of other biological populations. It can strongly contribute to our understanding of human evolution. Another benefit could be the research of diseases. The diversity research could help us to explain why certain groups are vulnerable to certain diseases and how populations have adapted to these vulnerabilities. All these modern features of biodiversity research are beyond the scope of this introduction and they are addressed in the other papers of the proceedings.

This methodical introduction is somewhat simplified because there is no straightforward link between genotype and phenotype variability, especially in diploid or aneuploid cells or organisms. In the real world, the final variety of forms, e.g. really expressed gene alleles, translated genes or the assemblage of species, inevitably depends on factors of surrounding environment. In some sense, we must accept the environment as a factor constituting the variability of biological world; at the same time, however, the environment is constituted or at least influenced by the processes mediated by living organisms. To conclude, although we primarily examine the biodiversity, we must also study the diversity of environmental conditions. This principle is applied in any level of organization, including genetic or molecular systems.

And vice-versa, the remarkable changes in biodiversity indicates probable changes in life, nutritional or environmental conditions in the examined system. The changes in structure of biological communities due to multiple anthropogenic stresses have recently received an increasing attention as a perspective indicator system of ecosystem integrity. Scientific data concerning this topic – and especially statistically derived outputs – play an indispensable role in the identification or the retrospective evaluation of risks associated with environmental disturbances.

If we need to characterize the biodiversity as an environmental indicator system in one word, we should use the word complexity. Yes, biodiversity is really very complex, as it comprises the status of many forms, very distinct species, with their own evolutionary strategy, habitat and niche preferences. The complexity is surely positive if we are able to measure it representatively – in such case, it brings a really complex and valid interpretation. The ecological risk assessment, based on the diversity of biological communities, provides a very relevant and straightforward interpretation for examined ecosystems.

However, the complexity is inevitably coupled with a relatively low specificity. The more complex end-point is used, the less specific can be the interpretation of changes. That is why the biodiversity is frequently listed among the so-called integrating parameters. The explanation of this term is simple. To register significant changes in biodiversity, especially at ecosystem level, we need a long-term time series of measurements of many biological communities. Changes in such system naturally integrate numerous stimuli, both natural (nutritional status) and anthropogenic (toxic pollution). Complex communities comprise a variety of species with different susceptibility to the environmental stress which could make biodiversity patterns rather difficult to identify. The bio-indication of detrimental changes at this level should always be coupled with a well-optimized and powerful statistical treatment.

In view of these features that could mask real mechanisms of stress influence, the comparative evaluation of biodiversity in different stressed communities appears to be a fertile area of research. Apart from an increasing scientific interest, there is still lack of standard statistical methodology in this field (Washington, 1984; Hughes, 1986; Krebs, 1989; Fausch *et al.*, 1990; Tokeshi, 1993). Statistical methods are often used only as a tool for indication of patterns, rather than for explanation of mechanisms or quantity.

Last but not least, we must mention the communication and presentation power of biodiversity. The biodiversity in the colloquial and aesthetic sense represents a widely accepted and required attribute of nature and countryside. The drop in biodiversity is clearly recognized as a risk by general public. So although biodiversity is a relatively complex endpoint, which requires high-volume and long-term data, it is a favorable endpoint for environmental studies. The biodiversity has the power to demonstrate the value of ecosystem to be protected. Monitoring with incorporated biodiversity measures can help to communicate and to manage the risk.

2 Biodiversity data as a unique challenge for the computational science

This chapter brings a brief methodical overview of approaches currently available for the analysis of biodiversity data. A more detailed description of individual techniques can be found in the other chapters of the proceedings.

2.1 Measures of biodiversity: why and how to analyze them

The relatively complex definition of biodiversity presented in the previous chapter necessarily opens the question how to analyze it. If we wish to work with the biodiversity as an indicator system, we must be able

- to quantify it
- to define the reference status (at least probabilistically)
- to mutually compare different systems
- to detect changes in time

We strongly need some numerical measures that can indicate the increasing/decreasing variability of forms in the analyzed system. And not only blind indication: we need to interpret what has happened to the biological system and "inside" the biological system. It means to detect what components disappeared or increased in incidence, complex structural changes, changed proportions among components, etc. The so-called biodiversity measures are therefore assessed also from the viewpoint of their computational background and relevant interpretation. The current literature offers hundreds of such measures, but not all represent the proper and best choice for all types of data. That is why we try to explain here the main types of biodiversity metrics in mutual comparison. A more detailed insight into the concept of biodiversity measures can be found in the chapter written by Jarkovsky et al. in the proceedings.

2.2 Biodiversity data and its specifics

The biodiversity at any level of organization is described through two basic components (we speak about the "dual concept" in biodiversity measurement, see also Fig. 1):

(1) The count of forms that comprise the biological system. A typical example is the number of species (species richness) as a simple count of species in the community. The community can thus be species-rich or poor.

(2) The quantification of occurrence (incidence) of the forms. This is typically the abundance specifying the number of individuals per species, or the more frequently used biomass quantification, respiration or another metabolic activity, etc. This component is often called the heterogeneity measure. The role of this component is to quantify the relative size of populations of

species that is valuable to give them relevant weights, e.g. to identify dominant species, rare species, etc.

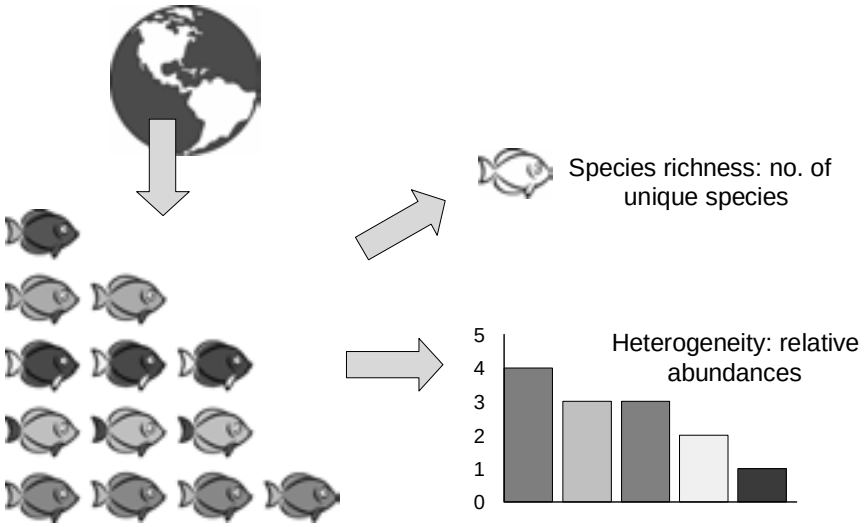


Fig. 1. Dual concept associated with the biodiversity data: the number of species and the relative size of their population

In a relevant biodiversity analysis, both components mentioned above should be incorporated. We cannot rely only on simple number of species (forms). Furthermore, various measures can be simultaneously employed as quantifying component 2 and such multivariate investigation can indicate different underlying mechanisms. It must be however emphasized here that the two components might not be necessarily coupled or correlated. The increasing number of species is not always associated with an increasing abundance without any change in the community structure. This fact forms the real substance of biodiversity measurement where we examine relationships between the number of forms and their quantitative presence in the system. Indeed, the majority of the biodiversity measures are based on ratio or relative weighting of the species richness and heterogeneity measures.

There is one very important aspect that completely separates processing of biodiversity data from the other fields of biostatistics. It refers to the sample size phenomenon. The standard statistics tends to optimize the sample size in order to reach a sufficient power of applied statistical test(s), e.g. for the comparison of two different experimental groups. Such an approach simply cannot be adopted for biodiversity analysis: here, the number of species (forms) is strictly given by the type of the system. Some biological communities are species-rich, some are not, and this fact itself contributes to the biodiversity estimation. That is why it is so important to weight number of forms by their relative quantity (size) in the system. The heterogeneity measures generate patterns that are at least partially independent on the number of species present. These distributional patterns are called species-abundance profiles.

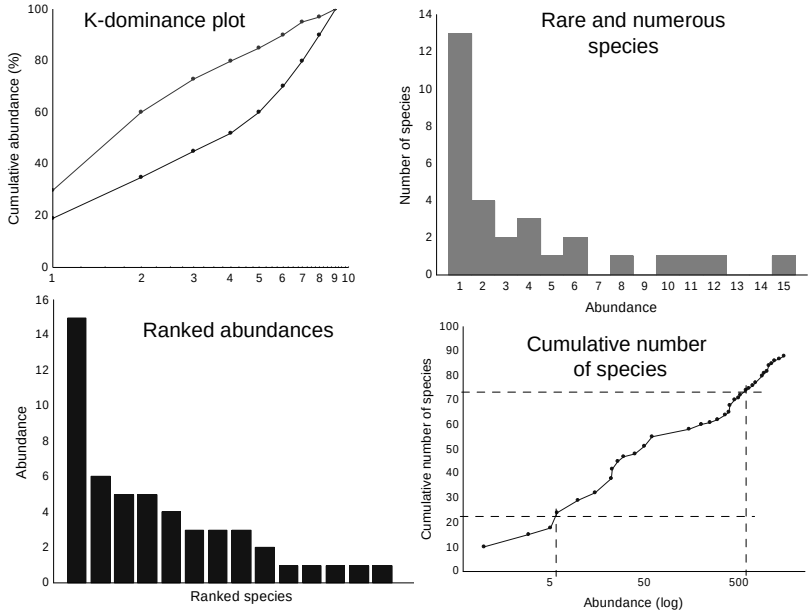


Fig. 2. Important types of plots used to document species-abundance profiles

Figure 2 summarizes different types of the so-called species-abundance plots that are commonly used to show species-abundance profiles in different communities. The most important "species rank plot" arranges species rank in the X axis according to their quantity plotted in the Y axis. The plot can easily indicate very important changes in the community structure (see also Fig. 3):

- changes in proportion of different species (Fig. 3A)
- loss of rare species (Fig. 3B)
- loss of some remarkable part of the community or some part sensitive to the influential factors (Fig. 3C, D)

The principle of biodiversity measures thus cannot stem only from the standard stochastic methodology. Instead, we use several families of measures with the interpretation more or less focused on the number of forms and their relative quantity.

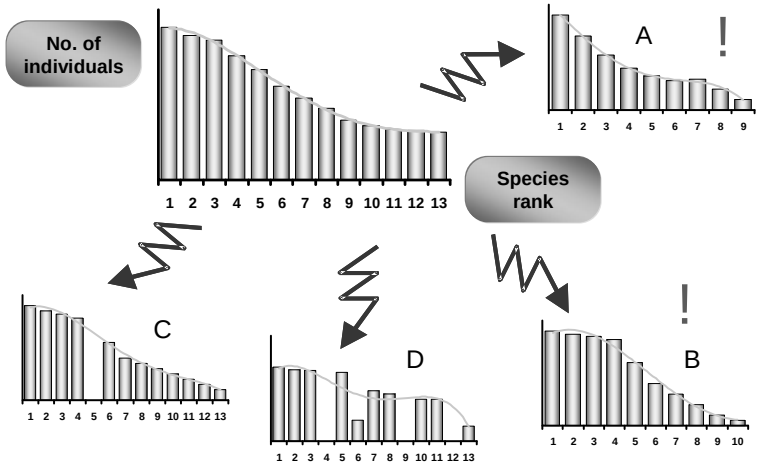


Fig. 3. Species-abundance profiles and indication of impact of influential factor (marked as arrow: stress factor, changes in nutritional resources, etc.)

2.3 Biodiversity indices

Biodiversity indices are a frequently used measure that became nearly synonymous to the term biodiversity. Very typically, these measures were developed on the basis of some empiric experience or requirements and they got specific names that correspond to different fields of biology or to the authors' names. The interpretation of many indices is associated with a given biological discipline. Table 1 brings a short overview of the most important measures. We should mention here the following principles that are intrinsically associated with biodiversity indices:

- The *species richness measure* is the simplest dimension of biodiversity derived from counts of unique species (or other examined forms). This dimension is commonly marked as S in formulas and is called the alpha-diversity. Typical representatives of this family are Margalef or Menhinick index (Table 1). Usability of species richness as separated measure might be disputable: it always depends on given assumptions and the investigator must control the risk of bias. The species richness is recommended for the comparison of the whole ecosystems, where it simply comprises tens-hundreds of species and serves as an overall "health" indicator.
- The *concept of dominance* can be easily derived from species richness. It measures the relative population size of the most frequent species. Again, this is a very simplified measure, however with a serious interpretation, because it can be attributed to the type of biological community. It also reflects nutritional, seasonal and other determining environmental conditions. Basically, it can be easily understood: for example, if we have the community with dominant species representing 90 % of the whole size (measured as individuals, biomass, etc.) or a community with 1-2 dominant species altogether only with 20 % of the whole size. For the formula, see for example the Berger-Parker index in Table 1.

Table 1. Overview of biodiversity indices (S=number of species, N=number of individuals, n_i=number of individuals of the i-th species)

Index	Equation
Margalef index (Cliphord & Stephenson, 1975)	$D_{Mg} = \frac{(S-1)}{\ln N}$
Menhinick index (Whittaker, 1977)	$D_{Mn} = \frac{S}{\sqrt{N}}$
Shannon index (Pielou, 1975).	$H' = -\sum p_i \ln p_i$, where $p_i = \frac{n_i}{N}$
Brillouin index (Pielou, 1969, 1975)	$HB = \frac{\ln N! - \sum \ln n_i!}{N}$
Simpson index (May, 1975)	$D = \sum \left(\frac{n_i(n_i-1)}{N(N-1)} \right)$
Berger-Parker index (Berger & Parker, 1970, May, 1975)	$d = \frac{N_{\max}}{N}$ where N_{max} - abundance of the most abundant species

- The *Shannon's concept* underlying the Shannon's diversity index (H'; see Table 1) measures the *information entropy*, considering species as symbols and their relative population sizes as probability. The advantage of this concept is that it simultaneously takes into account the number of species and their relative distribution, and thus it introduces the measure of the so-called *evenness*. The index is increased either by having more unique species, or by having a greater evenness. The maximum evenness corresponds to the ideal model situation when all the species reach equal population size. This concept was widely adopted in many biological disciplines as it offers a valuable interpretation also to the nutritional and environmental conditions. More species with a more equal distribution imply more favorable conditions with less frequent competitive interactions. In the case of a limited number of species (typically in biodiversity monitoring in urban areas), the Brillouin's formula is recommended as less biased than original Shannon's measure (Table 1).
- The *concept of probabilistic similarity* is based on probability calculations, e.g. the probability to catch two individuals of the same species from the pool of N species. As an example of such approach, we used the Simpson index (Table 1).

Of course, the estimation of biodiversity using any type of indices can be complicated with some probability of error. *The level of uncertainty* related to the diversity measures is not simple and it is not even theoretically completely solved. In fact, some diversity indices are applied without an adequate theoretical reasoning. To estimate the stochastic variability in terms of standard error and/or confidence limits, we often use bootstraping or jackknifing as iterative methods to estimate the variability of different indices. However, there are additional risks of bias that cannot

be simply detected by formulas, namely the risk of non-representative (biased) sampling. The analysis of this problem is beyond the scope of this overview; for more information, we recommend an excellent monograph written by Magurran (1983).

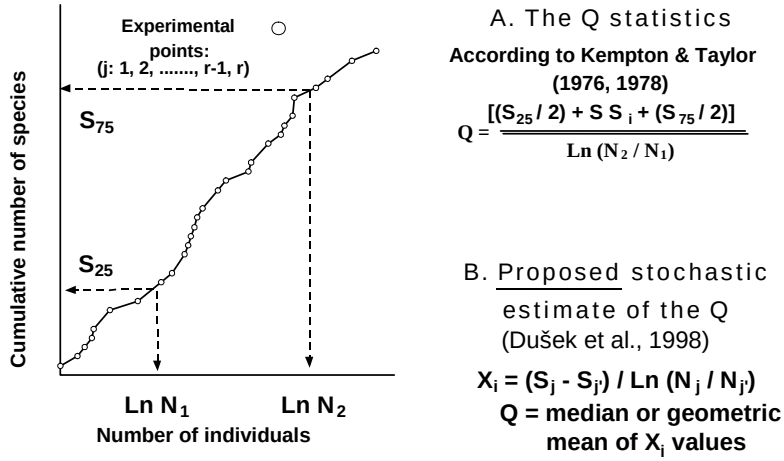


Fig. 4. Cumulative species-abundance curve and its stochastic outcome: the Q statistics

2.4 Stochastic analysis of cumulative species–abundance curves

This is one of the most frequently used graphs for biodiversity data in cumulative species-abundance curve (see Fig. 2 and Fig. 4). It can be simply drawn by plotting the species rank against the cumulative sum of their population size. The figure can be plotted for biological communities of any size and multiple lines can be easily compared directly in just one figure! We can hardly find another type of graph with such an user-friendly layout. Moreover, the cumulative curves provide a very important computational added value – we can estimate the slope as a measure of diversity. It is called the Q statistics and should be mentioned here for its robustness and flexibility.

The analyses of the cumulative species abundance curves were applied as a rather robust alternative to common biodiversity indices (Kempton and Taylor 1976, 1978) or as a stochastic estimate of the Q using consecutive computational steps (Fig. 4, Tab. 2). If we choose the median estimate of the slope of species cumulative frequency, it is important to note that its flexibility generally allows the estimation of the Q for a variety of abundance values from the whole range to very small intervals. The estimates can be based on the whole range or the inter-quartile range of abundance values (marked as Q_{total}, Q_{intq}). Dusek et al. (1998) proved a highly significant correlation between Q_{intq} and Q according to Kempton and Taylor (1978) and justified the proposed stochastic approach as being compatible with the previous algorithm.

Table 2. Q statistic as stochastic outcome from cumulative species–abundance curves

Index	Equation
Q statistics inter-quartile (Kempton and Tailor 1976, 1978)	$Q = \frac{\frac{1}{2}n_{R1} + \sum_{R1+1}^{R2-1} n_r + \frac{1}{2}n_{R2}}{\log\left(\frac{R2}{R1}\right)}$, where Σn_r – total number of interquartile species, R1 a R2 – 25% a 75% percentile, n_{R1} – number of species in lower quartile class, n_{R2} – number of species in upper quartile class.
Q statistics stochastic (Dušek et al,1998)	$X_i = \frac{S_j - S_{j'}}{\log(\frac{N_j}{N_{j'}})}$, for all combinations of S_j, $S_{j'}$ and N_j, $N_{j'}$ ($j > j'$, j=1,2, …….r) where S – cumulative number of species, N – number of individuals in given class, r – number of classes (i=1,2,……r(r-1)/2). Final Q is computed as median or geometric mean of X_i.

2.5 Correlation of different biodiversity measures

It is evident that the correct application of biodiversity measures is strongly associated with their information potential and interpretation. However, there is an universal tendency, leading to the mutual correlation of biodiversity indices and other related statistics. The biodiversity in general should grow with the increasing number of unique species (concept of species richness) and with a more equal proportional size of their populations (growing heterogeneity). A very important factor is our ability to define positive correlation with some widely used biodiversity metric with an already accepted interpretation, for example with Shannon's index. In such way, the correlation between stochastically estimated Q statistics (cumulative species-abundance curves, see 2.4) and Shannon's concept was documented in large communities of *Ephemeroptera* and *Plecoptera* (Zahrádková et al., 1998). Such findings represent a key point in the interpretation of the Q statistics. As expected, a more profound relationship was found between H´and Q_{total} than between H´ and Q_{intq} (see also text in 2.4). However, a non-random distribution of species with an extremely low or high abundance could bias the estimate, as it is known for Shannon's H´ (Pielou, 1975).

We can conclude that many diversity indices should principally positively correlate, of course when measured in similar types of communities. Similarly, we can use directly the primary abundance values to quantify the position of species within the community and to perform multivariate analyses. Such analyses are

typically used to cluster species or sampling sites according to similarities in the incidence of certain species. Biodiversity measures can also enter such analyses. More information on multivariate biodiversity analyses can be found in the paper of Jarkovsky et al. in the proceedings.

2.6 Species-abundance models

In addition to the species richness and heterogeneity measures, which have been described above, we can evaluate the biodiversity using the so-called species-abundance models. We can select from many types of model templates with an already formalized mathematical background. One may ask now, what is the real added value of such models, if we have so many easily calculated indices? We will try to answer this expectable question in the following points:

a) The *species-abundance models present a more sophisticated alternative* to simple indices. Once we fit a proper model to the species-abundance profile of the community, we can monitor substantial changes (i.e. changes that really affect the structure of the community).
b) *The models contribute to the typology of different biological systems* with the most remarkable outcome in the standard ecology. Some communities follow certain (typical) species-abundance pattern. Such typology cannot be done only on the basis of diversity indices because these measures cannot describe the profile characteristic for the position of different community components.
c) *Species-abundance models* represent a very powerful tool for the studies relating biodiversity patterns to environmental conditions or available nutrients. Of course, the structure of biological community reflects these conditions and some types of models can be typically found under stressed situation. For example, log-normal model (see also Table 3 and Figure 5) had been frequently attributed to undisturbed communities with a tendency to be replaced by log series or geometric series in stressed communities (Gray & Pearson, 1982).

The literature distinctly recognizes the so-called stochastic species-abundance models that can be fit according to clearly given templates and equations, using standard stochastic methodology. It means goodness-of-fit test (for small assemblages of species: Kolmogorov-Smirnov one-sample test) that can assess a satisfactory fit to primary data. Basic types of the models are listed in Table 3 and displayed in Figure 5.

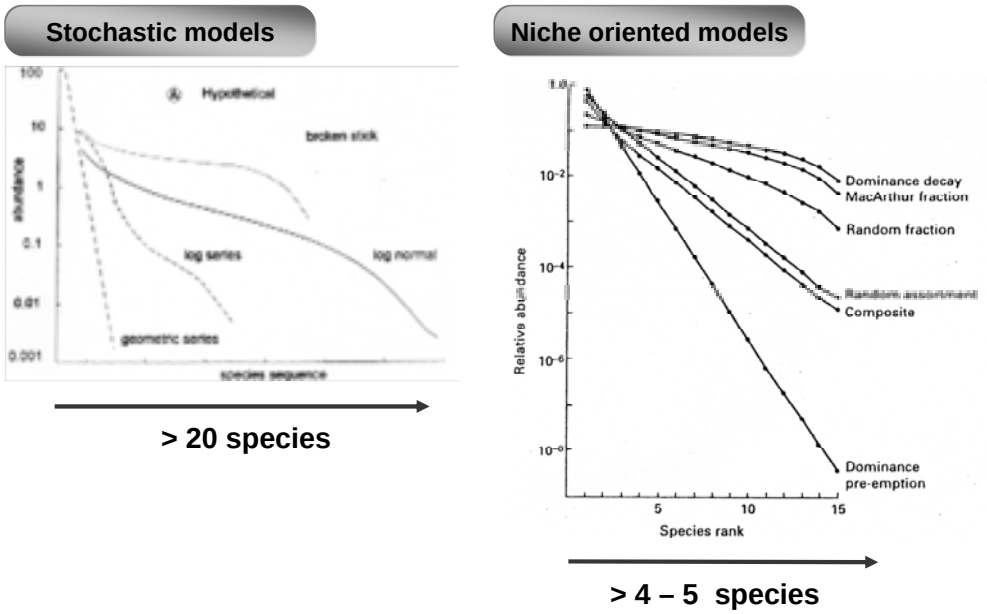


Fig. 5. Stochastic and niche-oriented species-abundance models

Table 3. List of species-abundance models with relevant references

Type of model	Model	Author
Stochastic models	Logarithmic series	Fisher et al. (1943)
	Log normal	Preston (1948, 1962)
	Negative binomial	Anscombe (1950), Bliss and Fisher (1953)
	Zipf-Mandelbrot	Zipf (1949, 1965), Mandelbrot (1977, 1982)
Niche oriented models	Geometric series	Motomura (1932)
	Particulate niche	MacArthur (1957)
	Overlapping niche	MacArthur (1957)
	Broken stick	MacArthur (1957)
	MacArthur fraction	Tokeshi (1990)
	Dominance pre-emption	Tokeshi (1990)
	Random fraction	Tokeshi (1990)
	Sugihara's sequential breakage	Sugihara (1980)
	Dominance decay	Tokeshi (1990)
	Random assortment	Tokeshi (1990)
	Composite	Tokeshi (1990)

However, the stochastic techniques suffer from sample size that is mostly relatively small in biological communities (we cannot optimize species richness; it is definitely given for a certain biological community). This fact can strongly limits the

discrimination power of these models, the verification becomes inconclusive and the models cannot be taken as statistically proved at reasonable significance level. Similar conclusion can be made in environmental or ecotoxicological studies applying stochastic species-abundance models. Stochastic foundation of the models determines their excellent properties as a generalized comparative tool, sufficiently flexible even for the evaluation of quite different heterogeneous communities (Routledge, 1980). Although this robustness might be desirable in comparative ecotoxicological studies, the rather complex models cannot directly reflect the mechanisms underlying the observed, environmentally induced changes.

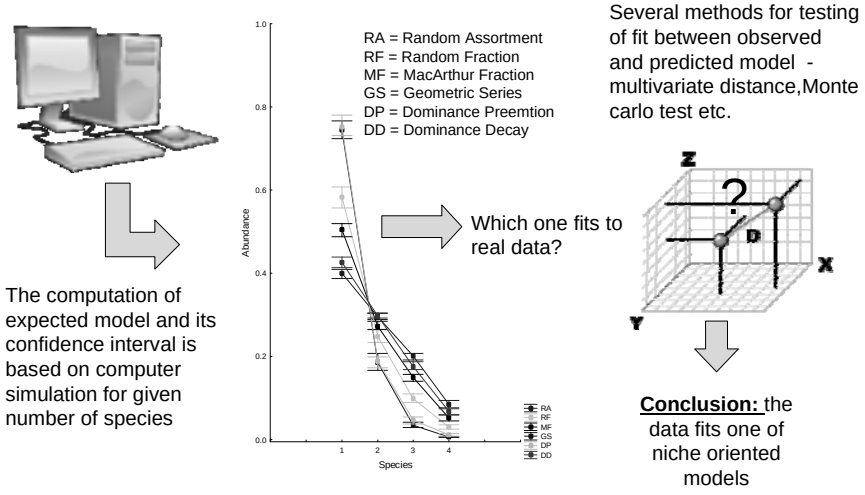


Fig. 6. Computer assisted estimation of niche oriented species-abundance models as proposed by Tokeshi (1990, 1993)

Fortunately, the recently developed niche-oriented species-abundance modeling (Tokeshi, 1990, 1993) allows an effective indication of species heterogeneity with a consideration of the relationship between resource and species-abundance pattern. A basic overview of these models is given in Table 3 and in Figure 5. Based on computer-assisted simulations (see scheme in Fig. 6), these models can be quite effectively fit to the species-abundance data of relatively small or even very small communities (less than 5 unique species). We can unambiguously mark these models and their mathematical background as revolutionary, because they released the modeling from the prison of sample size calculations. And additionally, the proposed models recognized the species-abundance patterns that imply an important relationship between the biology of given genera and their sensitivity to the environmental stress. The niche-oriented models are defined through a sequential breakage or filling of total niche (Fig. 7):

- **Geometric series:** the 1st species is supposed to preempt a fraction *k* of the total niche, the 2nd species *k* of the reminder, the 3rd one again *k* ….. .; then the relative abundances of species form a geometric series with a standard formula
- **Random fraction:** a sequential division of a niche in a random fashion (each fraction is

randomly selected for a random and uniform division)

- **McArtur fraction:** a simultaneous random breakage of the niche into several species (special type of Broken Stick model – see also Figure 5)
- **Random assortment:** suitable for highly dynamic communities under a varying environment; abundances of different species are not mutually related at all (apparently a consequence of non-correspondence between niche fragmentation and abundances of species)
- **Dominance preemption:** the model represents general and universal concept of dominance; 1st species exerts its dominance by occupying more than half of the total niche, and leaves the reminder to be exploited by the 2nd species in the same manner (the fraction *k* here represents the occupied proportion).

Tokeshi in his original papers (1990, 1993) gave an excellent biological reasoning for each individual model, including analyses of consequences for resource/niche fractioning. The models are open to a further scientific discussion, applications and possibly to the mathematical development. The strategic advantage of this approach, i.e. the ability to estimate the species-abundance profile in really small components of large communities, has still not been adequately addressed in literature.

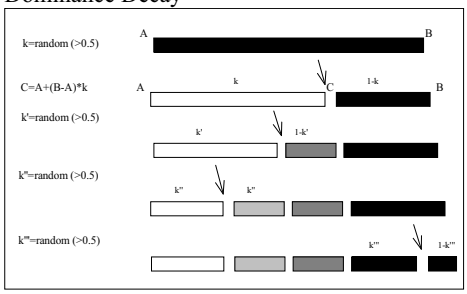
The experimental design of environmental and ecotoxicological monitoring is under a permanent pressure to reduce cost of routinely performed campaigns. Although it might seem improbable, a taxonomic survey is economically demanding, time-consuming and laborious. Moreover, a survey of large biological communities depends on experts in various taxonomic units. Therefore, we must search for an indicator species or at least a small assemblage of such species. The niche-oriented approach opened a very interesting challenge in this field. Using this approach, we would be able to relate diversity patterns to changeable environmental conditions.

3 Biodiversity of large communities

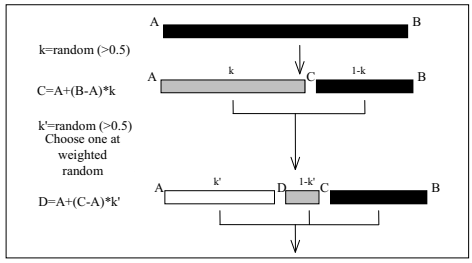
3.1 Biodiversity of large communities: sample size vs. information value

Let us focus our methodical comments on large biological communities, typically sampled in large-scale ecosystem studies. Samples with many species strengthen the estimation of common biodiversity indices and even the simplest counting of species can serve as a basis for the comparison of different systems (species richness as a marker of ecosystem status). However, the situation is not so ideal when we try to study the structure of the community and species-abundance profiles. Large communities comprise numerous taxonomic units, with very different evolutionary or feeding strategy. It can be even impossible to interpret overall species-abundance profiles is such heterogeneous field. The fractionation of the whole assemblage and separation of more homogeneous components can be apparently recommended as a very effective solution. However, the optimization of stochastic techniques for analyses of biodiversity of communities, which have been simultaneously fractionated according to different criteria, is a rather neglected area.

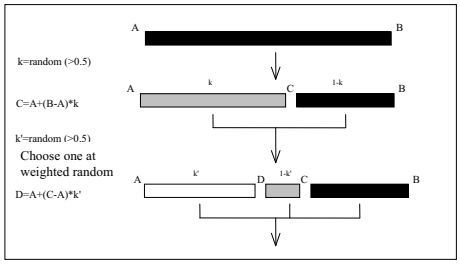
Dominance Decay



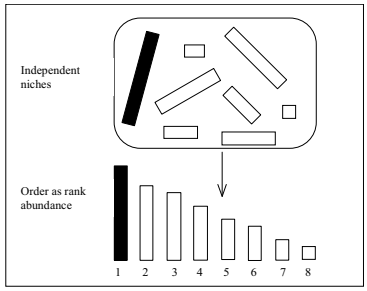
MacArthur Fraction



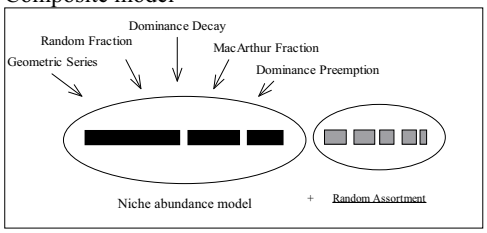
Random Fraction



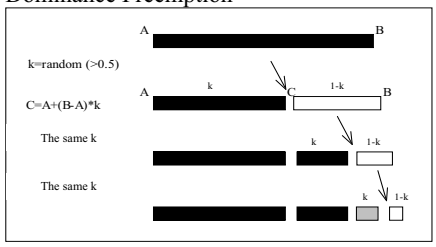
Random Assortment



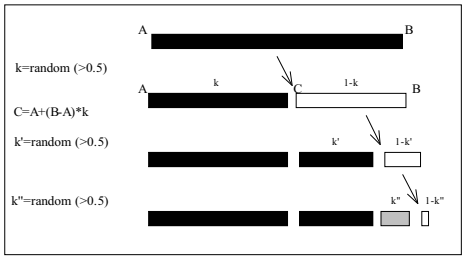
Composite model



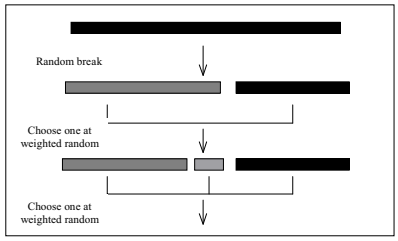
Dominance Preemption



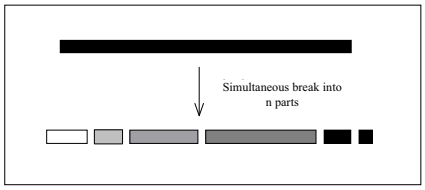
Geometric Series



Power Fraction.



Broken Stick



Overlapping niche

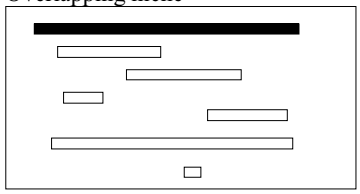


Fig. 7. Scheme presentation of some niche oriented species-abundance models

The fractionation can help us to exclude some evidently outlying units that should not be mixed with the others. Applying the fractionation can also detect components sensitive to environmental changes. It is surprising that such studies separating sub-components of biological systems with subsequent biodiversity measurement are still relatively rare. Any separation of detailed components of the whole community is accompanied with the following methodical problems or aspects:

- *necessity of biological, not only analytical expertise*: many communities are really complex and complicated, and the fractionation must respect already given biological (taxonomic) boundaries and limits
- *correction for multiple comparison*: the fractionation can be redundant, i.e. leading to several mutually overlapping sub-components; the comparison of different samples then faces the problem of multiple comparison and should be corrected for an error of I^{st} type
- *analyses of biodiversity based on small assemblage of species* or assemblages of species of different size; in that case, however, we must carefully select the proper methodology

Apparently, the fractionation of biological communities is required and valuable, although it generates methodical problems. Whenever we meet such complex and heterogeneous mixture of components, we should base the biodiversity analysis on robust techniques. Such approach is able to maintain the comparability of outcomes. For such purpose, we can recommend:

- *common species richness measures*: although they are rather of a poor information value, the interpretation is robust and straightforward
- *cumulative species-abundance curves* and associated Q statistics: a very robust estimate with Shannon-like interpretation; the estimate is functional for even very small communities
- *niche-oriented species-abundance models* as a methodology usable for very small assemblage of species; outcomes lead to a biologically relevant interpretation

3.2 Biodiversity of large communities: metazoan parasites of fish as example

This shortened case study represents an attempt to define the environmental indicative potential of biodiversity of monogenean parasites on the basis of hierarchically-structured species-abundance data. For this purpose, *Monogenea* were selected as one of the most numerous and diverse group of ectoparasites infecting fish (here Chubb). At present, there are estimated 3,000 monogenean species which have been described in about 1,500 species of fishes. The study is basically comparative and work with parasite communities from two river sites, one reference and one polluted. Primary data were taken and simplified from the paper Dusek et al. (1998).

It is apparent from a graphical display in Figure 8 that *Monogenea* formed the most abundant group of ectoparasites in both compared sites. Apart from a similar proportion of ectoparasites or *Monogenea* in both sites, there was a significant difference in the distribution of non-monogenean ectoparasites, namely the group *Bivalvia*. While *Bivalvia* formed a relatively abundant part of ectoparasites in the

unpolluted site, these species represented only a minor part of parasitofauna in the anthropogenic site (Fig. 8). This difference seemed to be easily explained on the basis of indirect life cycles of bivalvian molluscs: a substantially decreased number of their free living stages was stated in the anthropogenic site, apparently due to organic pollution. Therefore, the accurate assessment of any environmental change at the community level was very complicated due to the dominant position of *Bivalvia* in the control site, and the applied separate evaluation of *Monogenea* appeared to be both taxonomically and ecologically meaningful.

The applied fractionation scheme was further justified by the direct development of monogeneans without the participation of intermediate hosts. The need to extract taxonomically comparable groups led to the separation of ectoparasites and particularly *Monogenea* in the first two steps and to a subsequent subdivision of *Monogenea* according to different taxonomic or ecological criteria, as is shown in Figure 8. The study solves the two principal complications associated with analyses of biodiversity in such a hierarchical system. The first one is a profound taxonomic heterogeneity at higher levels of organization which could mask environmentally induced changes and mechanisms. And secondly, the statistical evaluation became increasingly more difficult as the fractionation led to small assemblages of closely related species.

Considering community level and separated ectoparasites, the increased values of Shannon diversity index H´ clearly proved a more equal distribution of species in the anthropogenic site. Values of dominance D, Margalef´s index M and a portion of rare species were numerically higher in the control than in the anthropogenic community, but there were nearly no indices of statistical significance of these trends (data not shown). The seemingly illogically lowered dominance and increased homogeneity of species distribution in the stressed site reflected the top position of parasites in the aquatic ecosystem. The parasites integrate adverse effects of complex and naturally varied stresses, influencing also the other components of aquatic ecosystems. The observed changes could be a result of two opposite way of stress influence, i.e. an effect on the entire parasite community and a parallel reduction of fish defense mechanisms against pathogenic agents (Thomas, 1990; Moriarty, 1993; Gelnar and Špakulová, 1997).

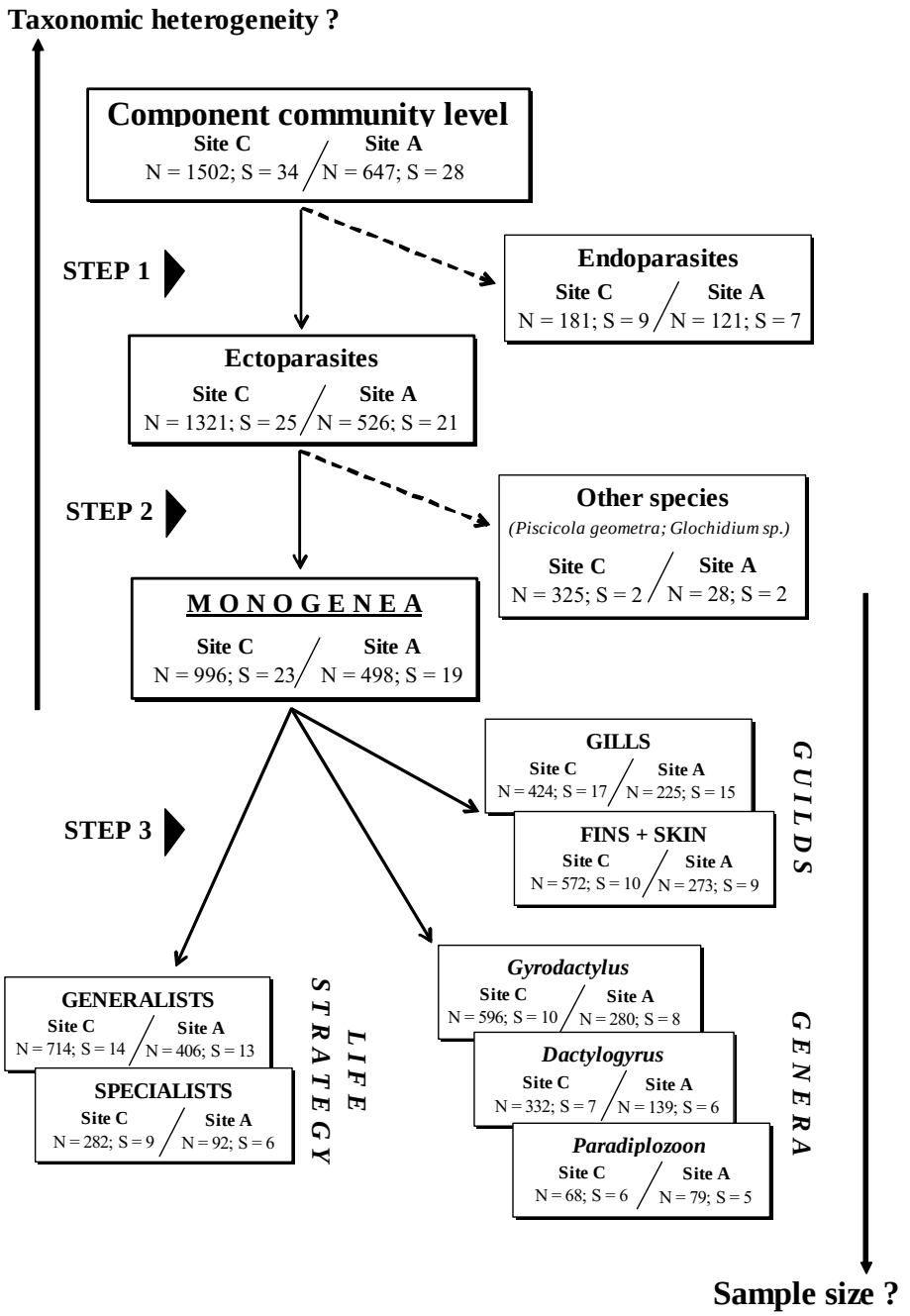


Fig. 8. Case study on fish parasites: scheme of relevant fractionation of the whole community (C: Control site; A: Site under anthropogenic stress)

Table 4. Species richness and heterogeneity measures evaluated for community and for subgroups of metazoan parasites [1]

Assemblage (category) of species	Shannon´s index (H´)			Dominance (in %)		Margalef´s (M)	
	C	A	t test [2]	C	A	C	A
Component community	2.75	2.89	3.279 *	20.9	14.8	4.5	4.1
Endoparasites	1.13	1.43	2.404 *	70.1	52.0	1.5	1.3
Ectoparasites	2.53	2.68	1.993 *	23.7	18.2	3.3	3.1
Monogenea	2.59	2.55	0.951	14.8	19.2	3.1	2.8
Life strategy							
Generalists	2.12	2.25	2.794 *	20.7	53.2	1.9	1.9
Specialists	2.25	1.31	3.757 *	40.0	23.6	1.4	1.2
Genera							
Dactylogyrus	1.47	1.03	4.416 *	40.1	69.0	1.0	1.0
Gyrodactylus	1.92	1.87	0.628	24.8	26.	1.4	1.2
Paradiplozoon	1.66	1.75	0.713	36.5	32.9	1.1	0.9
Guilds							
Gills	2.09	1.99	1.060	31.3	42.6	2.6	2.5
Fins + skin	1.93	1.90	0.634	23.9	27.1	1.4	1.4

[1] Two compared sites (**C** - control site; **A** - anthropogenic site) with hierarchically separated subgroups of metazoan parasites.

[2] t statistic as a result of testing differences between the sites in Shannon´s heterogeneity measure (approximate t-test, Zar, 1984; ***** marks the category with significantly different values H´ in the sites: $p < 0.01$).

Surprisingly, the statistical significance of these patterns was not detected when biodiversity analyses were carried out for the separated *Monogenea* (Table 4). Therefore, the limited development of bivalvian molluscs due to affected free living stadia appeared to influence the already described difference in biodiversity between the sites at the component community level. Based upon the set of questions induced by this outcome, a further inspection of monogenean group was necessary. Monogenean parasites grouped according to their specificity revealed the most significant differences between the sites (Table 4). In comparison with the control site, a significantly less homogeneous distribution of monogenean specialists (lower H´ values) was associated with sharp changes in the dominance, and a portion of rare species was found in the polluted site. The biodiversity of generalists revealed clearly the opposite pattern.

To conclude, the previously described results document the value of the fractionation of biological community and associated methodical problems as well. The separation of monogenean parasites and their stratification according to life strategy (generalists vs. specialists) led to the definition of an environmentally sensitive component of the whole community. Both simple species richness estimates and species heterogeneity measures agreed in this conclusion.

The consequent analyses of the study were focused on species-abundance models, both stochastic and niche-oriented. The stochastic models (log series and log-normal) confirmed already described features, i.e. a more homogeneous distribution of species with middle-ranged abundance, decreased species richness in the anthropogenic site

as compared to the control site, the opposite behavior of specialists and generalists and no significant environmentally induced shifts in diversity within the inhabited guilds. Unfortunately, the estimates of the parameters suffered from small samples, except for the whole community level and separated ectoparasites. However, niche-oriented approach found distinct models in different Monogenean genera and related the observed changes of control and polluted sites to the biology of these species. For more information see the study of Dusek et al. (1998).

4 Ecological risk assessment and biodiversity measures

4.1 Definition of risk assessment and the role of biodiversity

Ecological and human risk assessment can be characterized from the viewpoint of informatics as a complicated processing of heterogeneous data (mostly retrospectively collected from various sources) leading to the probabilistic estimation of some uncertain (prospective approach) or, on the other hand, a relatively certain (retrospective approach) risk event. Key methodical steps of the whole process are summarized in Figure 9 and can be simply defined as follows:

1. Problem formulation and hazard identification. Introduction to any reasonably designed study. It includes the recognition of the area of interest, the collection and aggregation of required information and a preliminary focus on the identified principal pollutants (stressors), the source of contamination and the most vulnerable environmental components and biological receptors.
2. Multi-component exposure assessment. Exploration, identification and quantification of important exposure pathways. This includes modeling and summaries of accessible data, as well as empirical estimates of environmental concentrations of proposed key pollutants.
3. Biological effect evaluation. The empirical stage focused on concentration-related or dose-related reactions of biological systems. The principal aim is to get parametric measures that identify biologically dangerous concentration levels. The process should not be limited only to laboratory testing, it works with ecosystem monitoring as well. Whenever we have access to regional or national bio-monitoring network, we should use this data as very powerful information background.
4. Risk characterization is a completely computational process that leads to the probabilistic estimate of the risk. In fact, it consists of a stochastic aggregation of data from all the preceding methodical blocks.

In other words, there are many inputs required and only limited number of outputs provided, however always with a serious impact.

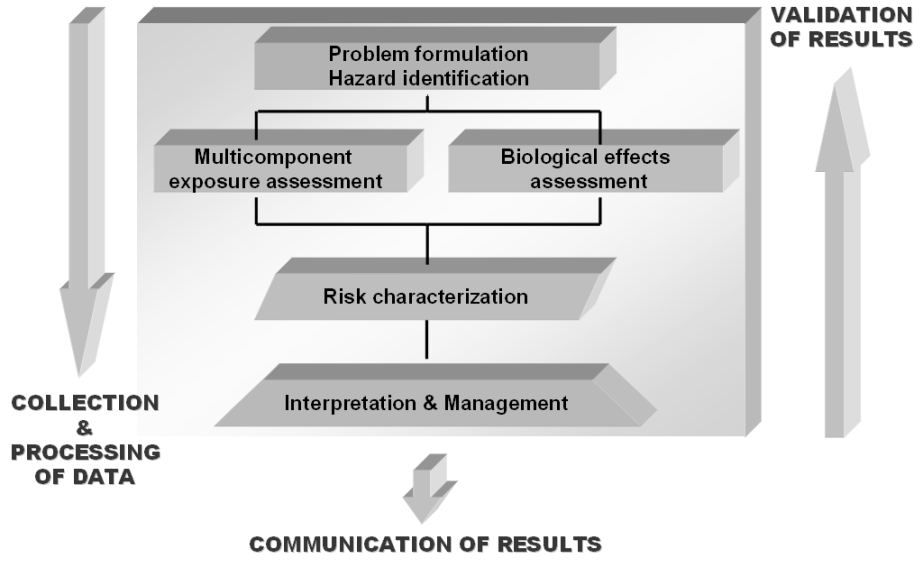


Fig. 9. Environmental risk assessment (EcoRA) and key methodical steps

The biodiversity as an intrinsic attribute of biological communities and ecosystems plays a fundamental role in the whole process of environmental risk assessment. It can be employed in several steps:

- The problem formulation often requires a definition of the status ("health") of ecosystems in the area of interest. Here, the biodiversity measures contribute substantially because they have a long-term "memory", i.e. the structure and diversity of biological communities can reflect the stressed impact that had been performed a long time ago. The biodiversity can also help to describe potential "hot spots", i.e. sites in the assessed area with a highly probable stress influence.
- Simple mapping of biodiversity in the area of interest can help to localize places with improbable findings – these sites can be subsequently subjected to chemical monitoring to prove exposure to toxic compounds.
- The biodiversity has also its place in the evaluation of biological effects of potential stressors.
- The biodiversity might be one of the most important end-points in the subsequent biological monitoring of the area of interest.

References

1. Anscombe, F. J. Sampling theory of the negative binomial and logarithmic series distributions. Biometrika 37: 358-382 (1950)

2. Barnhart, Benjamin J. (1989). "DOE Human Genome Program". Human Genome Quarterly **1**: 1.
3. Berger, W. H., and Parker, F. L. Diversity of planctonic foraminifera in deep sea sediments. Science 168: 1345-1347 (1970).
4. Bliss, C. L., and Fisher, R. A. Fitting the binomial distribution to biological data and a note on the efficient fitting of the negative binomial. Biometrics 9: 176-206 (1953).
5. Clifford, H. T., and Stephenson, W. An Introduction to Numerical Classification. Academic Press, London. 229 pp. (1975)
6. DeLisi, Charles (2001). "Genomes: 15 Years Later A Perspective by Charles DeLisi, HGP Pioneer". *Human Genome News* **11**: 3–4.
7. DeLong, D.C.: Defining biodiversity. Wildl. Soc. Bull. 24 (1996) 738-749
8. Dušek, L., Gelnar, M., Šebelová, Š. Biodiversity of parasites in a freshwater enviroment with respect to pollution: metazoan parasites of chub (Leuciscus cephalus L.) as a model for statistical evaluation. International Journal for Parasitology 28: 1555-1571 (1998)
9. Fausch K. D., Lyons J., Karr J. & Angermeier P. L. 1990. Fish community as indicators of environmental degradation. In: *Biological Indicators of Stress in Fish* (Edited by Adams S. M.). pp. 123 - 144. American Fisheries Symposium 8, Bethesda, Maryland.
10. Fisher, R. A., Corbet, A. S., and Williams, C. B. The relation between the number of species and the number of individuals in a random sample from an animal population. J. Anirn. Ecol. 12: 42-58. (1943)
11. Gelnar. M. & Špakulová M. 1997. A checklist of the monogenean parasites reported from fishes in the Czech and Slovak Republics. *Helminthologia* **34:** 189.
12. Gray J. S. & Pearson T. H. 1982. Objective selection of sensitive species indicative of pollution-induced change in benthic communities. 1. Comparative methodology. *Mar. Ecol. Prog. Ser.* **9:** 111-119.
13. Kempton R. A. & Taylor L. R. 1978. The Q-statistic and the diversity of floras. *Nature* 262: 818 - 820.
14. Krebs, C.J. Ecological Methodology. Harper & Row, New York. (1989)
15. Legendre, P., and Legendre, L. Numerical ecology. Elseviere Science BV, Amsterodam. (1998)
16. Magurran, A. Ecological diversity and its measurement. Croom Helm, London. pp. 177 (1983)
17. MacArthur, R. H. On the relative abundance of bird species. Proc. Natl. Acad. Sci. USA 43:293-295 (1957)
18. Mandelbrot, B. B. The Fractal Geometry of Nature. Freeman, San Francisco. (1982)
19. May, R. M. Patterns of species abundance and diversity. Ecology and Evolution of communities, pp. 81-120. Belknap/ Harvard University Press, Cambridge. (1975)
20. Motomura, I. A statistical treatment of associations (in Japanese and cited in May, 1975). Jpn. J.Zool., 44: 379-83 (1932)
21. Moriarty F. 1993. *Ecotoxicology. The study of Pollutants in Ecosystem*. Academic Press. London.
22. Pielou, E. C. An Introduction to Mathematical Ecology. Wiley, New York (1969)
23. Pielou, E. C. Ecological diversity. John Wiley & Sons, New York. 165 pp (1975)
24. Preston, F.W. The commonness, and rarity, of species. Ecology, 29: 254-83 (1948)
25. Routledge R. D. 1980. The form of species abundance distributions. J. Theor. Biol. 82: 503-515.
26. Simpson, E. H. Measurement of diversity. Nature Lond. J 163: 688. (1949)
27. Sugihara, G. Minimal community structure: an extrapolation of species abundance patterns.Am. Nat. 116: 770-787 (1980)

28. Thomas P. 1990. Molecular and biochemical responses of fish to stressors and their potential use in environmental monitoring. In: *Biological Indicators of Stress in Fish.* (Edited by Adams S. M.) pp. 9 - 28. American Fisheries Symposium 8, Bethesda, Maryland.
29. Tokeshi M. 1990. Niche apportionment or random assortment: species abundance patterns revisited. *J. Anim. Ecol*. **59:** 1129-1146.
30. Tokeshi, M., Species abundance Patterns and Community structure. Advances in ecological research vol. 24: 111-186 (1993)
31. Washington H. G. 1984. Diversity, biotic and similarity indices: a review with special relevance to aquatic ecosystems. *Water Res.* **18:** 653-694
32. Whittaker, R.H. Evolution and measurement of species diversity. Taxon, 21, 213-51. (1972)
33. Zahrádková S., Soldán T., Helešic J., Dušek L., Landa V., 1998, Distributional and quantitative patterns of Ephemeroptera and Plecoptera in the Czech Republic: A possibility of detection of long-term environmental changes of aquatic biotopes. 305 p., Folia, Masaryk University, Czech Republic.
34. Zipf, G. K. Human Behaviour and the Principle of Least Effort. Hafner, New York (1965)

Population dynamics – a source of diversity

Zdeněk Pospíšil

Masaryk University, Faculty of Science, Department of Mathematics and Statistics,
Kotlářská 2, Brno, Czech Republic,
pospisil@math.muni.cz,
WWW home page: http://www.math.muni.cz/~pospisil/

Abstract. The contribution presents several simple models of population dynamics, both with the discrete time (difference equations) and with the continuous one (ordinary differential equations). The models are realized using free software – spreadsheet Open Office Calc and R-language.

1 Introduction

Biological communities are formed by populations. Such a trivial observation can serve as a starting point to consideration on diversity of natural biocenoses that can be understand as *a number* of populations constituting community, *abundances* of these populations, *amount of relations* among them or all of these characteristics together. One can put questions such that: Can a single population in a community survive? On which conditions? What type of intra- and inter-population interactions promote or inhibit abundance or even presence of a population? Why do organisms become extremely abundant one year and then seem to disappear a few years later? Why do population outbreaks in particular species happen more or less regularly in certain locations, but only irregularly (or never at all) in other locations?

One of possible tools for dealing such questions is mathematical modeling of population growth. The mathematical population dynamics constitutes a classical part of the mathematical biology. It goes back to the Leonhard Euler's *Introductio in analysin infinitorum* (1748) and continues by seminal works *Recherches mathématiques sur la loie d'accroissement de la population* by Pierre-François Verhulst (1845), *Elements of physical biology* by Alfred J. Lotka (1925), *Leçons sur la théorie mathématique de la lutte pur la vie* by Vito Volterra (1931), *Sula teoria di Volterra della lota per l'esistenza* by Andrej N. Kolmogorov (1936) or *On the use of matrices in certain population mathematics* by Patrick H. Leslie (1945), cf. [1]. In the present time, it pose a well established theory with a huge amount of literature. E.g., the books [9, 6, 10, 12, 11] can serve as good and accessible introductory texts to the area of mathematical population models. The issues of population dynamics in a larger frame of mathematical biology are dealt in the books [8, 3, 7, 2].

The aim of this contribution is not to survey the mentioned literature and to present comprehensive mathematical theory in the background of models in

population dynamics but to show how one can "play" with some simple models and this way she or he can understand some population principles and to gain certain insight into processes occurring in ecology of populations. And, secondary, to show that such a "play" requires no expansive "toys": one can have a lot of fun with a free software.

For the purpose, the subsequent section summarize the very first mathematical models of the growth of one isolated population and of interacting populations. The third section shows realization of population models with the non-overlapping generations (dynamical models with discrete time, difference equations) in a spreadsheet. It is based on ideas from the book [5] and the examples are performed by the Open Office Calc spreadsheets, [13]. The last section deals with models with overlapping generations (continuous time models, ordinary differential equations) and their realization by scripts of R-language, [14]. It utilizes some material presented in the book [4].

2 Deterministic models of population dynamics

We start with a "first principle of population dynamics" [11, p. 100]: *A size of population increase or decrease exponentially provided an environment where all of individuals live is constant.* This "law" is similar to the Newton's law of inertia. Both the laws are abstract and their manifestation in the real world is not observable; there is no mechanical motion without friction and there is no constant environment for population growth – at least since each population transforms its environment. But the physical law is very useful and the law of population growth should be so. Therefore, we are going to elaborate it in a more tractable way.

First, let us suppose that the time goes by distinct step, that is a time instant t is an element of the set $\{0, 1, 2, \dots\}$. In another words, there is a "natural" time unit expressing a duration of life – newborns appear at the beginning of the period and die at its end. Such an assumption is tenable e.g. for annual plants or for insects, models established this way are called *models with non-overlapping generations.* The basic principle can be expressed in the form

$$x(t) = x_0 q^t,$$

where x_0 denotes initial size of population, $x_0 = x(0)$, and q is a positive coefficient, rate of change of the population size during the unit time interval. Consequently, $x(t+1) = x_0 q^{t+1} = x_0 q^t q = qx(t)$. This way, we obtain basic discrete equation of population growth in the form

$$x(t+1) = qx(t), \quad x(0) = x_0. \tag{1}$$

An alternative assumption consists in the idea that newborns appear and individuals die at any time, i.e. the time passes in a continuous way from a beginning, a time instant t is an element of the interval $[0, \infty)$ of reals. A typical example of such population is the human one. Population models based on

continuous time are called *models with overlapping generations.* Now, the basic principle reads

$$x(t) = x_0 \mathrm{e}^{pt},$$

where x_0 denotes initial size of population anew, p is a real parameter. The derivative of the function x is $x'(t) = x_0 p \mathrm{e}^{pt} = px(t)$, hence the basic continuous equation of population growth takes a form of

$$x' = \frac{\mathrm{d}x}{\mathrm{d}t} = px, \quad x(0) = x_0. \tag{2}$$

The solutions of the initial value problems (1) and (2) coincide for $t \in \{0, 1, 2, \dots\}$ if $p = \ln q$, or, equivalently, $q = \mathrm{e}^p$. Let us note, that the population size increases for $p > 0$ ($q > 1$) and that it decreases for $p < 0$ ($0 < q < 1$).

In the rest of this section, we deal with the differential equation models. However, the same considerations might be provided also for models with discrete time.

In real situation, no population growths according to the idealized equation (1) or (2), that is, no population is non-influenced. At least, it may be influenced by itself. More precisely, the coefficient p (or the rate q) depends on population size, $p = p(x)$ (or $q = q(x)$). Such dependence may be diverse.

First, suppose that a large population consumes almost all resources of its environment and, subsequently, it starves and become extinct; the population exhibits intraspecific competition. That is, $p(x)$ is a decreasing function. Moreover, we can assume that $\lim\limits_{x \to 0+} p(x) > 0$ (small population does not exploit resources and its size increases) and there is a value K such that $p(x) < 0$ for $x > K$. The positive parameter K denotes *carrying capacity* of the environment, i.e. the maximal population size such that he population survives. The simplest function possessing these properties is the linear one,

$$p(x) = r\left(1 - \frac{x}{K}\right).$$

Here, parameter r denotes the maximal possible growth rate of population, the so called *intrinsic growth rate.* This way, we obtain the Verhuls logistic differential equation

$$x' = rx\left(1 - \frac{x}{K}\right). \tag{3}$$

It can be rewritten to the form

$$x' = x(r - ax),$$

where $a = r/K$.

Alternatively, we can suppose that a small population is not able to survive, only population large enough prospers, e.g. females find males for mating, adults individuals are able to protect offsprings against predators etc. The population exhibit intraspecific cooperation, the so called Allee effect. Now $p(x)$ is an increasing function such that $\lim\limits_{x \to 0+} p(x) < 0$ and there exists a positive *survival*

threshold ϑ such that $p(x) > 0$ for $x > \vartheta$. Once more, the simplest function possessing such properties is the linear one,

$$p(x) = d\left(\frac{x}{\vartheta} - 1\right).$$

The positive parameter d is called *intrinsic death rate.* The equation modeling a growth of population exhibiting intraspecific cooperation is as follows

$$x' = dx\left(\frac{x}{\vartheta} - 1\right). \tag{4}$$

A natural population can exhibit both intraspecific cooperation and intraspecific competition; the Allee effects prevails over the intraspecific competition in a small population whereas the situation is reversed in a large one. Hence, $p(x) < 0$ for $0 < x < \vartheta$ and $p(x) > 0$ for $x > K$; now $K > \vartheta$. This consideration may lead to the equation

$$x' = dx\left(\frac{x}{\vartheta} - 1\right)\left(1 - \frac{x}{K}\right). \tag{5}$$

One can object that the population models (3), (4), (5) are excessively simply. There is no evidence that the dependencies of the rate p on the population size x are linear or quadratic. Hence, we can modify the Verhulst logistic equation (3) by an additional parameter b to the form

$$x' = rx\left(1 - \left(\frac{x}{K}\right)^b\right). \tag{6}$$

The parameter b expresses a "strength" of dependence of p on x: if $b > 1$ then an impact of population size to its growth, i.e. the intraspecific competition, is mild when the population is small and it is stronger when the population is greater; if $0 < b < 1$ than the intraspecific competition in a small population prevails the one in a greater population; if $b < 0$ then $p(x)$ is an increasing function with $p(x) < 0$ for $0 < x < 1$ and $p(x) > 0$ for $x > K$, hence the equation (6) models an evolution of population exhibiting the Allee effect and the parameter K represents the survival threshold. Consequently, the equation (6) generalizes both the model (3) and the model (4).

A single population does not constitute any community and so, models of single population growth seems to be irrelevant for study of diversity. But the consideration provided may inspire a line of inquiry of interacting population. Hence, let us consider a community formed by n populations and denote by the symbol $x_i = x_i(t)$ a size of the i-th population in a time instant t. Now, we write down the following system of ordinary differential equation as an analogy to the basic equation (2)

$$x_i' = x_i p_i(x_1, x_2, \ldots, x_n), \quad i = 1, 2, \ldots, n; \tag{7}$$

the growth rates p_i depend on sizes of all of the populations forming community. The system (7) is called the Kolmogorov one. The type of dependence of the rate

p_i on the population size x_j determines a kind of impact of the j-th population to the i-th one, or, conversely, the type of the interaction of the i-th population with the j-th one determines a sort of dependence of p_i on x_j. In a more concrete way:

$\frac{\partial p_i}{\partial x_j} > 0$**:** the j-th population promotes a growth of the i-th one, j-th population is a commensal of the i-th one;

$\frac{\partial p_i}{\partial x_j} < 0$**:** the j-th population inhibits a growth of the i-th one, j-th population is an amensal of the i-th one;

$\frac{\partial p_i}{\partial x_j} > 0$ **and** $\frac{\partial p_j}{\partial x_i} > 0$**:** the i-th and the j-th populations exhibit interspecific cooperation, they are mutualistic, symbiotic;

$\frac{\partial p_i}{\partial x_j} < 0$ **and** $\frac{\partial p_j}{\partial x_i} < 0$**:** the i-th and the j-th populations exhibit interspecific competition;

$\frac{\partial p_i}{\partial x_j} < 0$ **and** $\frac{\partial p_j}{\partial x_i} > 0$**:** the j-th population is predator (consumer, parasite) feeding on the i-th population, the i-th population represents prey (resource, host) for the j-th population.

$\frac{\partial p_i}{\partial x_i} < 0$**:** the i-th population exhibit intraspecific competition;

$\frac{\partial p_i}{\partial x_i} > 0$**:** the i-th population exhibit intraspecific cooperation.

Moreover, we can classify populations by the value $p_i(0, 0, \ldots, 0)$ which express the growth rate of the i-th population provided on condition that the all of the impact to population growth are excluded, that is, the intrinsic growth rate of the i-th population. If $p_i(0, 0, \ldots, 0) > 0$ then the i-th population is self-supporting (autotroph, producer), otherwise it depends on other populations, the population is consumer (predator, parasite).

The simplest special case of the general system (2) is the one with coefficient p_i that depends on population sizes $x_1, x_2, \ldots, x_n$ linearly. Namely, the system of the form

$$x_i' = x_i(r_1 - a_{i1}x_1 - a_{i2}x_2 - \cdots - a_{in}x_n), \quad i = 1, 2, \ldots, n, \tag{8}$$

which is called the Lotka-Volterra one. Obviously,

$$\frac{\partial p_i}{\partial x_j} = -a_{ij},$$

hence, we can classify the interactions appearing in the modeled community by signs of the coefficient a_{ij}.

The Lotka-Volterra system (8) can be rearranged to the form

$$x_i' = x_i\,(r_i - a_{ii}x_i) - x_i(a_{i1}x_1 + \cdots + a_{ii-1}x_{i-1} + a_{ii+1}x_{i+1} + \cdots + a_{in}x_n), \quad i = 1, 2, \ldots, n.$$

The first term on the right hand side describes the evolution of the isolated i-th population and the second one expresses the impact of other population in the community to the i-th population growth. If $r_i > 0$ and $a_{ii} > 0$, we can denote $K_i = r_i/a_{ii}$ and $\psi_i(x_1, x_2, \ldots, x_n)$ to obtain the system of ordinary differential equations

$$x_i' = r_i x_i \left(1 - \frac{x_i}{K_i}\right) - \psi_i(x_1, x_2, \ldots, x_n), \quad i = 1, 2, \ldots, n. \tag{9}$$

This system can be interpreted so that the isolated i-th population evolves according to the Verhulst logistic equation (3) and this evolution is modified by the impact of other populations.

Predation represents a fundamental relation among population. Therefore, let us mention it in more details. Let $x = x(t)$ denote a size of a prey (resource, plant) population and $y = y(t)$ denote a size of predator (consumer, herbivore) population. Let us assume that the prey population exhibit a intraspecific competition. Hence, the evolution of the prey population size can be modeled by an equation of the type (9), namely

$$x' = rx \left(1 - \frac{x}{K}\right) - \psi(x, y).$$

The term $\psi(x, y)$ expresses an amount of prey destroyed by predator population of size y in a unit time provided that the prey population is of the size x. The function ψ should possess the following properties:

- $\psi(0, y) = 0 = \psi(x, 0)$ for all x, y: if no prey is available the predator destroys none of them, if no predators are present no prey is destroyed;
- $\dfrac{\partial \psi(x, y)}{\partial x} \geq 0$, $\dfrac{\partial \psi(x, y)}{\partial y} \geq 0$ for all x, y: if more prey is available predators do not destroy less of them, if more predators are present they do not destroy less of prey;
- there exists a positive constant S such that $\psi(x, 1) \leq S$ for all x: one predator is able to destroy a limited amount of prey, it hunts a prey until it is satisfied and the constant S represents a level of its satiety.

One can easily verify that all of these properties are satisfied by the function

$$\psi(x, y) = Sy\varphi(x),$$

where φ is a non-decreasing function possessing properties

$$\varphi(0) = 0, \quad \lim_{x \to \infty} \varphi(x) = 1;$$

The function φ is called *trophic function* or *functional response of predator to a prey density.* The function

$$\varphi(x) = \frac{x^k}{x^k + a^k},$$

where a and k are positive parameters, satisfies the conditions and it is widely used for the purpose. The parameter a expresses a size of prey population that can "half-satisfy" one predator.

Let us assume further that if no prey is available then the predator population starves and, consequently, it dies out; we denote the death rate by d. If predators destroy prey they transform it to their population growth with an efficiency κ. Hence, the evolution of the predator population size can be modeled by the equation

$$y' = -dy + \kappa\psi(x, y).$$

This way, we obtain the system of ordinary differential equations for the predator-prey interaction in the form

$$x' = rx\left(1 - \frac{x}{K}\right) - Sy\varphi(x), \tag{10}$$

$$y' = -dy + \kappa Sy\varphi(x), \tag{11}$$

which is called the Gause-type predator-prey model.

3 Discrete time models

The basic equation (1) can be easily solved by a recursive procedure: starting with $x(0) = x_0$ we compute $x(1) = qx(0)$, from $x(1)$ we compute $x(2) = qx(1)$ and so on. This simple observation suggests that a convenient tool for solving a discrete equation is a spreadsheet, in particular, the Open Office Calc. Supposing that the cells `B2` and `D1` contains the values $x(0)$ and q, respectively, we can put the formula `=$D$1*B2` to the cell `B3` to obtain the value $x(1)$ here. Then, we can copy the cell `B3` to the cells `B4`, `B5`, `B6`, and so on. Then, the solution can be visualized by inserting a graph, see Fig. 1. The analogous computations can be provided with various population models with non-overlapping generations.

Now, let us think about models of populations exhibiting an intraspecific competition. The consideration leading to such models are similar to that provided during derivation of the Verhulst equation (3). We suppose that the growth rate q in the equation (1) is a decreasing function of the population size x, i.e. $q = q(x)$, such that $\lim\limits_{x\to 0+} q(x) > 1$ and there exist a population size K such that $q(x) > 1$ for $x < K$ and $q(x) < 1$ for $x > K$. The simplest choice is the linear dependence, that is

$$q(x) = \varrho - \frac{\varrho - 1}{K}x,$$

where $\varrho = \lim\limits_{x\to 0+} q(x)$. This option leads to the equation

$$x(t+1) = x(t)\left(\varrho - \frac{\varrho - 1}{K}x(t)\right). \tag{12}$$

Putting $q = \mathrm{e}^p$ where p denotes the growth rate appearing in the Verhulst model, we obtain the equation

$$x(t+1) = x(t)\exp\left[r\left(1 - \frac{x(t)}{K}\right)\right]. \tag{13}$$

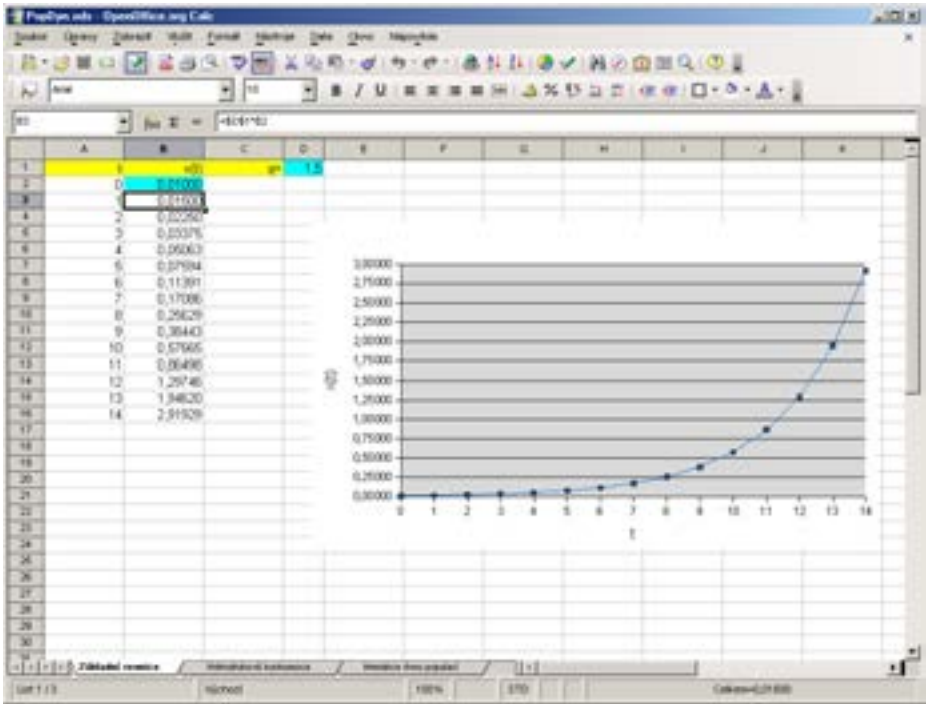


Fig. 1. Solution of the basic equation (1) in the spreadsheet Open Office Calc.

The equation (13) appears to be more adequate as a model of population growth that the equation (12) since it admits positive solutions only provided $x(0) > 0$; to the contrary, the equation (12) yields $x(1) < 0$ for $x(0) > K\varrho/(\varrho - 1) > 0$.

A "spreadsheet solution" of the equation (13) is displayed on Fig. 2. It shows that a single equation can produce solutions with various properties depending on values of parameters. Great intrinsic growth rate r produces oscillations in abundance; such a behavior is typical, e.g., for small rodents and it is called r-strategy. Small rate r yields a monotone growth of population size to the value of carrying capacity K; such behavior is called K-strategy[1]. The equation (13) admits also regular oscillations; e.g., for $K = 10$ and $r = 2.5$, the population size repeats itself every four time units after some time from beginning $t = 0$.

According to the considerations provided in the previous section, the communities consisting of two species may be modeled by the system of difference

[1] K-strategy is typical for great mammals. But for such animals, the overlapping of generations appears always. This observation demonstrates that the ecological strategies are not simple manifestations of population dynamic equations

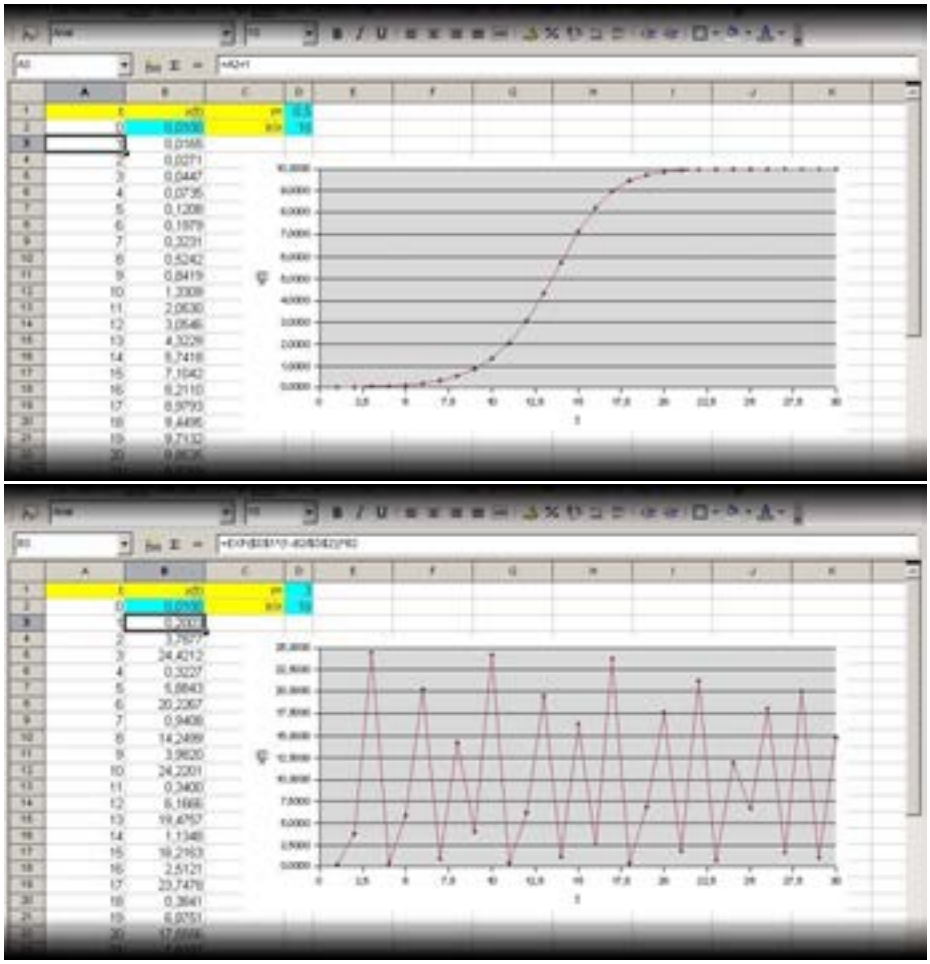


Fig. 2. Solution of the equation (13) – model of population with non-overlapping generation that exhibit an intraspecific competition. "Small" value of the intrinsic growth rate r yields monotone solution tending to the value K (above), "large" r produces irregular oscillations of population size (down).

equations (recurrence formulae)

$$x(t+1) = x(t)\exp\left(r_1 - a_{11}x(t) - a_{12}y(t)\right), \tag{14}$$
$$y(t+1) = y(t)\exp\left(r_2 - a_{21}x(t) - a_{22}y(t)\right). \tag{15}$$

The signs of the parameters a_{ij}, $i, j = 1, 2$ determines a type of interaction between populations. If all of them are positive, the equations describe interspecific competition of two populations exhibiting intraspecific competition. If $a_{ii} > 0$, $i = 1, 2$ and $a_{ij} < 0$ for $i \neq j$, the equations are about symbiosis of two self-limiting populations. If $a_{ij}a_{ji} < 0$ for $i \neq j$, the equations express predator-prey interaction. Using the spreadsheet, one can experiment with the system.

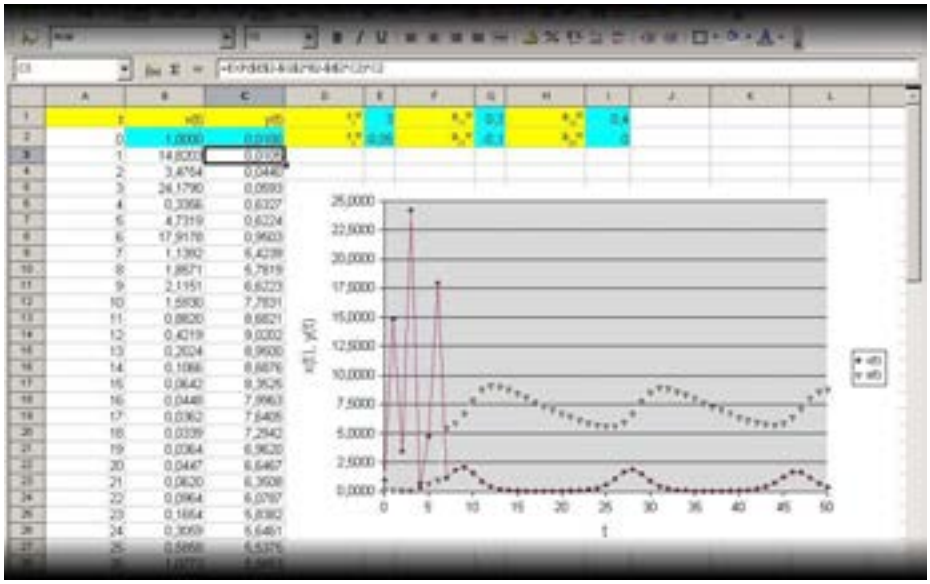


Fig. 3. Solution of the system (14), (15) with the parameters $r_1 = 3$, $r_2 = -0.5$, $a_{11} = 0.3$, $a_{12} = 0.4$, $a_{21} = -0.1$, $a_{22} = 0$. The system models an invasion of predators to a territory occupied by prey population asserting an r-strategy.

Fig. 3 presents the following predator-prey model:

$$x(t+1) = x(t)\exp\left(3 - 0.3x(t) - 0.4y(t)\right),$$
$$y(t+1) = y(t)\exp((-0.05 + 0.1x(t))\,;$$

here $x(t)$ and $y(t)$ denotes a size of the prey and of the predator population, respectively. The population of prey without a presence of predators, i.e. with $y(t) = 0$, evolves according to the equation (13) with the parameters $r = r_1 = 3$, $K = r_1/a_{11} = 3/0.3 = 10$. That is, $x(t)$ is the size of population modeled on Fig. 2 down. The predators are specialized to the prey species involved to

the model since they are not able to survive without it, $r_2 = -0.05 < 0$. But presence of prey population decreases the growth rate of the predator population, $-a_{21} = 0.1 > 0$. The initial value of predator population is small, this option models an invasion of predators to a territory occupied by the prey population. We can see that the predator population eliminates the irregular oscillations of the prey abundance. The both population coexists and their sizes vary in certain limits. These variations are such that the maxima of prey abundance are followed by the maxima of predator abundance after some time (approximately three time units).

4 Continuous time models

The software R can be supplemented by the package `deSolve` for numerical solution of differential equations. We illustrate the use of it by several examples.

4.1 Basic equation (2)

First of all, we load the library by the command

```
library(deSolve)
```

The model (2) possesses one parameter p, the growth rate. We set it to the value 1.5. Next, we need to specify the initial value, i.e. the initial size of population x_0. Let us assume that the initial population is small, $x_0 = 0.01$:

```
parameters <- c(p=1.5)
state <- c(x=0.1)
```

The subsequent modeling step consist in setting the initial and terminal time for solution and in specifying the time step for numerical integration of differential equation:

```
time <- seq(0,5,by=0.01)
```

The right hand side of the equation (2) is defined by the function

```
ODE <- function(t,state,parameters){
  with(as.list(c(state,parameters)),{dx <- p*x
                                     list(c(dx))})}
```

and the numerical solution of the initial value problem is computed and stored to the variable `out` by the command

```
out <- ode(state,time,func=ODE,parms=parameters)
```

Finally, we can plot the obtained solution:

```
plot(out[,"time"],out[,"x"],type="l",lwd=2,bty="l",
     xlab="time",ylab="x")
```

The result is displayed on Fig. 4 left. We can see that the solution is exponential curve, indeed, that is the population follows the "fundamental law". We can also investigate the impact of parameter p value to the solution. We vary the parameter

```
parameters1 <- c(p=1.1)
parameters2 <- c(p=0.5)
parameters3 <- c(p=-0.1)
parameters4 <- c(p=-1)
```

set the time scale and the initial value

```
time <- seq(0,2,by=0.01)
state <- c(x=1)
```

solve the equation with the specified parameters

```
out1 <- ode(state,time,func=ODE,parms=parameters1)
out2 <- ode(state,time,func=ODE,parms=parameters2)
out3 <- ode(state,time,func=ODE,parms=parameters3)
out4 <- ode(state,time,func=ODE,parms=parameters4)
```

and plot the result

```
plot(out1,out2,out3,out4,lty=c(1,2,3,4),lwd=2,
     col="black",bty="l")
legend(0,max(out1[,"x"]),lty=c(1,2,3,4),lwd=2,
       legend=c("p=1.1","p=0.5","p=-0.1","p=-1"))
```

The result is displayed on Fig. 4 right. No surprisingly, the solution of the equation with the parameter $p = 1.1$ grows unboundely, the solution with $p = 0.5$ grows as well but more slowly, the solutions with negative values of parameter p tend to zero, the one with $p = -1$ deceases as well but more quickly.

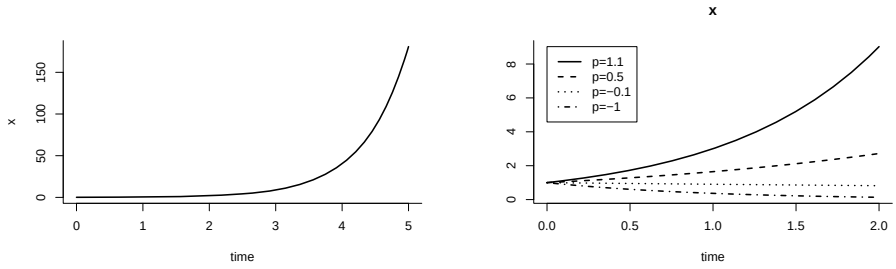


Fig. 4. Solution of the equation (2) with $x_0 = 0.1$ and $p = 1.5$ (left) and with initial value $x_0 = 1$ and various values of p (right).

4.2 Verhulst logistic equation

The equation (3) depends on two parameters r and K

```
parameters <- c(r=1,K=1)
```

the right hand side is defined by the function

```
ODE <- function(t,state,parameters){
  with(as.list(c(state,parameters)),{dx <- r*x*(1-x/K)
                                     list(c(dx))})}
```

We set time scale and the initial value $x_0 = 0.05$ to simulate growth of a small population:

```
time <- seq(0,10,by=0.1)
state1 <- c(x=0.05)
```

and solve the equation

```
out1 <- ode(state1,time,func=ODE,parms=parameters)
```

We also change the initial value to "a large size" of population and solve the equation one more time

```
state2 <- c(x=2)
out2 <- ode(state2,time,func=ODE,parms=parameters)
```

The plot of solutions

```
plot(out1[,"time"],out1[,"x"],type="l",lwd=2,bty="l",
     xlab="time",ylab="x",ylim=c(0,2))
points(out2[,"time"],out2[,"x"],type="l",lwd=2,lty=2)
```

we supplement with a line parallel to the time axis in the height $K = 1$

```
abline(h=1,lty=2)
```

From the result displayed on Fig. 5 left we can see that the small population growth along to a S-shaped curve to the carrying capacity $K = 1$, the great population decreases and it tends to the capacity K as well.

Now, we will check a impact of parameters r, K to the shape of the growth curve. For the purpose, we set four choices of parameters and set the initial value

```
parameters1 <- c(r=1,K=1)
parameters2 <- c(r=1,K=2)
parameters3 <- c(r=0.5,K=1)
parameters4 <- c(r=2,K=1)
state <- c(x=0.05)
```

and solve the four initial value problems for the equation (3)

```
out1 <- ode(state,time,func=ODE,parms=parameters1)
out2 <- ode(state,time,func=ODE,parms=parameters2)
out3 <- ode(state,time,func=ODE,parms=parameters3)
out4 <- ode(state,time,func=ODE,parms=parameters4)
```

We plot the solutions

```
plot(out1,out2,out3,out4,lwd=2,lty=c(1,2,3,4),col="black",bty="l")
legend(0,2,lty=c(1,2,3,4),lwd=2,
       legend=c("r=1, K=1","r=1, K=2","r=0.5, K=1","r=2, K=1"))
```

and display them on Fig. 5 right. We can see that the parameter K determines the limit of growth of modeled population and the parameter r control a speed of reaching the limit K. Moreover, the convergence of the population size to the limit K is monotone. This means that populations with overlapping generations assert a K-strategy (see comment to the equation (13) in the previous section).

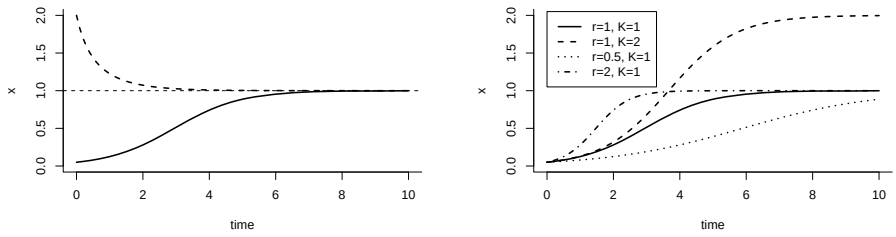


Fig. 5. Solution of the equation (3) with two different initial values x_0 (left) and with various values of parameters r, K (right).

4.3 Lotka-Volterra two-species systems

A Lotka-Volterra model (8) of two interacting species is of the form

$$x_1' = x_1\left(r_1 - a_{11}x_1 - a_{12}x_2\right),$$
$$x_2' = x_2\left(r_2 - a_{21}x_1 - a_{22}x_2\right).$$

The right hand side of the two dimensional system in the R-language environment is given by the function

```
LotkaVolterra <- function(t,state,parameters){
  with(as.list(c(state,parameters)),{
    dx1 <- x1*(r1-a11*x1-a12*x2)
    dx2 <- x2*(r2-a21*x1-a22*x2)
    list(c(dx1,dx2))})}
```

Let us consider the interspecific competition of two population exhibiting the intraspecific competition. That is, all of parameters a_{ij}, $i, j = 1, 2$ are positive. Let both the population have the same growth rate and carrying capacity provided they are isolated, $r_1 = r_2$, $a_{11} = a_{22}$, and let they differ in "strength of competition pressure", a_{12}, a_{21}. Hence, we set four different sets of the system parameters and initial values

```
parameters1 <- c(r1=1,r2=1,a11=1,a12=0.8,a22=1,a21=0.5)
state1 <- c(x1=0.01,x2=0.01)
parameters2 <- c(r1=1,r2=1,a11=1,a12=1.25,a22=1,a21=0.5)
state2 <- c(x1=1,x2=0.01)
parameters3 <- c(r1=1,r2=1,a11=1,a12=1.25,a22=1,a21=1.2)
state3 <- c(x1=1,x2=1)
parameters4 <- c(r1=1,r2=1,a11=1,a12=1.25,a22=1,a21=1.2)
state4 <- c(x1=0.02,x2=0.01)
```

The first choice represents mild interspecific competition and a situation when two small populations colonize a territory. The second one models an invasion of a population to a territory occupied by a resident population and the invading population x_2 exhibit strong competitive pressure a_{12} to the resident population x_1 whilst it exhibit only mild competitive pressure a_{21}. The third and the fourth choices model competition of the populations with strong interspecific competition. One choice describes disappearance of a barrier between two niches occupied by the two population, the other invasion of the two population to a vacant territory.

The time scale for the four simulations can be common:

```
time <- seq(0,30,by=0.1)
```

The solutions of the initial value problems are obtained by the commands

```
out1 <- ode(state1,time,func=LotkaVolterra,parms=parameters1)
out2 <- ode(state2,time,func=LotkaVolterra,parms=parameters2)
out3 <- ode(state3,time,func=LotkaVolterra,parms=parameters3)
out4 <- ode(state4,time,func=LotkaVolterra,parms=parameters4)
```

and they are plotted to one figure:

```
split.screen(c(2,2))
screen(1)
plot(out1[,"time"],out1[,"x1"],type="l",lwd=2,bty="l",
     ylim=c(0,max(out1[,"x1"],out1[,"x2"])),
     xlab="time",ylab=expression(list(x[1],x[2])),
     main="a_12 = 0.8,  a_21 = 0.5")
points(out1[,"time"],out1[,"x2"],type="l",lwd=2,lty=2)
screen(2)
plot(out2[,"time"],out2[,"x1"],type="l",lwd=2,bty="l",
     ylim=c(0,max(out2[,"x1"],out2[,"x2"])),
```

```
     xlab="time",ylab=expression(list(x[1],x[2])),
     main="a_12 = 1.25,  a_21 = 0.5")
points(out2[,"time"],out2[,"x2"],type="l",lwd=2,lty=2)
screen(3)
plot(out3[,"time"],out3[,"x1"],type="l",lwd=2,bty="l",
     ylim=c(0,max(out3[,"x1"],out3[,"x2"])),
     xlab="time",ylab=expression(list(x[1],x[2])),
     main="a_12 = 0.8,  a_21 = 1.2")
points(out3[,"time"],out3[,"x2"],type="l",lwd=2,lty=2)
screen(4)
plot(out4[,"time"],out4[,"x1"],type="l",lwd=2,bty="l",
     ylim=c(0,max(out4[,"x1"],out4[,"x2"])),
     xlab="time",ylab=expression(list(x[1],x[2])),
     main="a_12 = 1.25,  a_21 = 1.2")
points(out4[,"time"],out4[,"x2"],type="l",lwd=2,lty=2)
close.screen(all=TRUE)
```

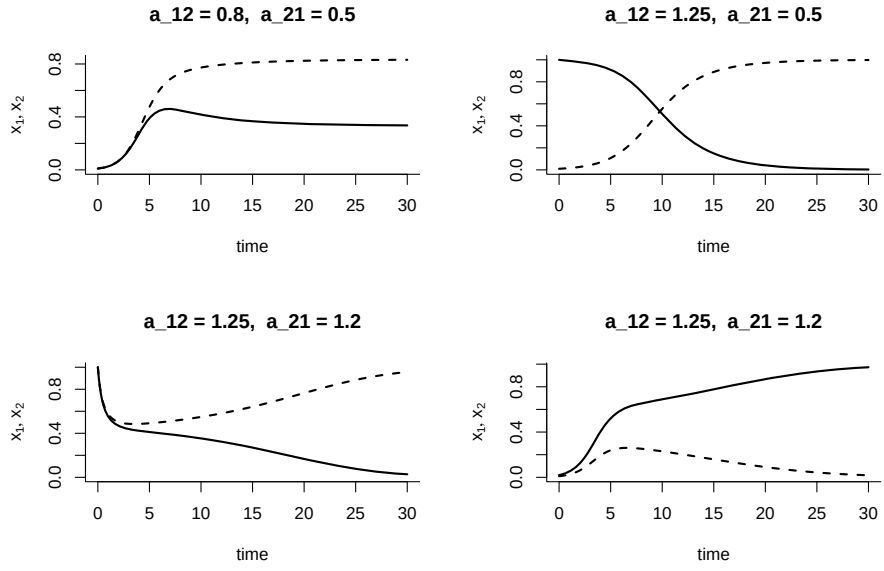


Fig. 6. Solution of the system (8) with $n = 2$, $r_1 = r_2 = K_1 = K_2 = 1$ and various values of interaction parameters a_{12}, a_{21} and initial values: $a_{12} = 0.8$, $a_{21} = 0.5$, $x_1(0) = 0.01$, $x_2(0) = 0.01$ (upper left); $a_{12} = 1.25$, $a_{21} = 0.5$, $x_1(0) = 1$, $x_2(0) = 0.01$ (upper right); $a_{12} = 0.8$, $a_{21} = 1.2$, $x_1(0) = 1$, $x_2(0) = 1$ (down left); $a_{12} = 1.25$, $a_{21} = 1.2$, $x_1(0) = 0.02$, $x_2(0) = 0.01$ (down right). Solid line: $x_1(t)$, dotted line: $x_2(t)$.

The result is displayed on Fig. 6. We can see that the populations exhibiting mild interspecific competition can coexist with sizes less than state of equilibrium for each of them being isolated. A population putting strong competitive pressure excludes a population exerting a mild one. If the both populations exhibits strong competitive pressure then one of them excludes the other, but it depends on the initial sizes of the populations which one of them goes to extinction, the competitive pressures does not determine it.

Now, we will simulate the predator-prey interaction. Let us suppose that the prey population exhibits an intraspecific competition and that the predator is specialized on it, that is the predator population goes to extinction if no prey is available. The predator population exhibits neither intraspecific competition nor cooperation. Hence, the model (8) takes the form

$$x' = x(r - a_{11}x) - a_{12}xy, \tag{16}$$

$$y' = -dy + a_{21}xy, \tag{17}$$

where x and y denote a size of the prey and of the predator populations, respectively, and all of the parameters are positive. We solve the system simply by two commands

```
time <- seq(0,40,by=0.1)
out1 <- ode(c(x1=1,x2=0.01),time,
            func=LotkaVolterra,
            parms=c(r1=1,a11=1,a12=1,r2=-0.5,a21=-0.8,a22=0))
```

i.e. we model an invasion of predators to a territory occupied by the prey population. The result is plotted on Fig. 7 left. The predator population survives in the territory and it diminish the size of prey population.

Now, we can put a question whether predators are able to control a size of pray population without an intraspecific competition, i.e. without a self-limiting mechanism. In other words, we check the system

$$x' = rx - a_{12}xy, \tag{18}$$

$$y' = -dy + a_{21}xy. \tag{19}$$

We obtain its solution by the command

```
out2 <- ode(c(x1=0.5,x2=0.5),time,
            func=LotkaVolterra,
            parms=c(r1=1,a11=0,a12=1,r2=-0.5,a21=-0.8,a22=0))
```

The result displayed on Fig. 7 right shows that sizes of the both populations oscillate between certain extreme values; maxima of prey abundances are followed by maxima of that of predators.

4.4 Gause-type predator-prey model

The system (10), (11) contains five numerical parameters r, K, S, d, κ and one functional parameter – trophic function φ. Hence, we first need to specify the function φ:

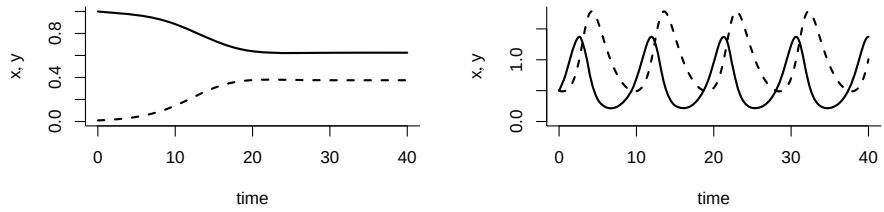


Fig. 7. Solution of the Lotka-Volterra predator-prey model; solid line represents prey abundance, dotted line represents predator abundance. Left: the system (16), (17) with $r = a_{11} = a_{12} = 1$, $d = 0.5$, $a_{21} = 0.8$, $x(0) = 1$, $y(0) = 0.01$. Right: the system (18), (19) with $r = a_{12} = 1$, $d = 0.5$, $a_{21} = 0.8$, $x(0) = 1$, $y(0) = 0.01$.

```
phi <- function(x){x/(x+0.1)}
```

The right hand sides of the equations (10), (11) can be evaluated by the function:

```
Gause <- function(t,state,parameters){
  with(as.list(c(state,parameters)),{pom <- S*y*phi(x)
                                     dx <- r*x-b*x^2-pom
                                     dy <- -d*y+kappa*pom
                                     list(c(dx,dy))})}
```

First, we set the parameter values to $r = K = S = 1$, $d = 0.5$, $\kappa = 0.6$ and the initial values to $x(0) = 1$, $y(0) = 0.01$. Such option models invasion of a predator population to an environment occupied by a prey population. We solve the initial value problem

```
time <- seq(0,200,by=0.1)
out1 <- ode(c(x=1,y=0.01),time,func=Gause,
            parms=c(r=1,b=1,S=1,d=0.5,kappa=0.6))
```

and plot the obtained solution to Fig. 8 upper left. We see that the sizes of the both populations tends to certain values with dumped oscillations.

Now, we slightly change the value of prey carrying capacity to the value $K = 0.8$ (hence $b = r/K = 1.25$):

```
out2 <- ode(c(x=0.8,y=0.01),time,func=Gause,
            parms=c(r=1,b=1.25,S=1,d=0.5,kappa=0.6))
```

and plot the solution to Fig. 8 upper right. The solution achieves the stabilized values monotone, without any oscillations. The third choice consists in expansion of the prey carrying capacity to the value $K = 1.25$:

```
out3 <- ode(c(x=1.25,y=0.01),time,func=Gause,
            parms=c(r=1,b=0.8,S=1,d=0.5,kappa=0.6))
```

The solution is plotted on Fig. 8 down left. The sizes of the both populations oscillates about certain values, the maxima of prey population precedes the maxima of the predator one. The simulations presented one example of the so called "paradox of enrichment": an increase of carrying capacity (enrichment of resources for the producer population) may destabilize a community.

Finally, we let the values of parameters and set the initial values near the average abundances of prey and predator populations obtained from the former solution:

```
out4 <- ode(c(x=mean(out3[,"x"]),y=mean(out3[,"y"])),time,
            func=Gause,parms=c(r=1,b=0.8,S=1,d=0.5,kappa=0.6))
```

The result plotted on Fig. 8 down right is very similar to the previous one. The sizes of populations tends to the same oscillation as in the previous case.

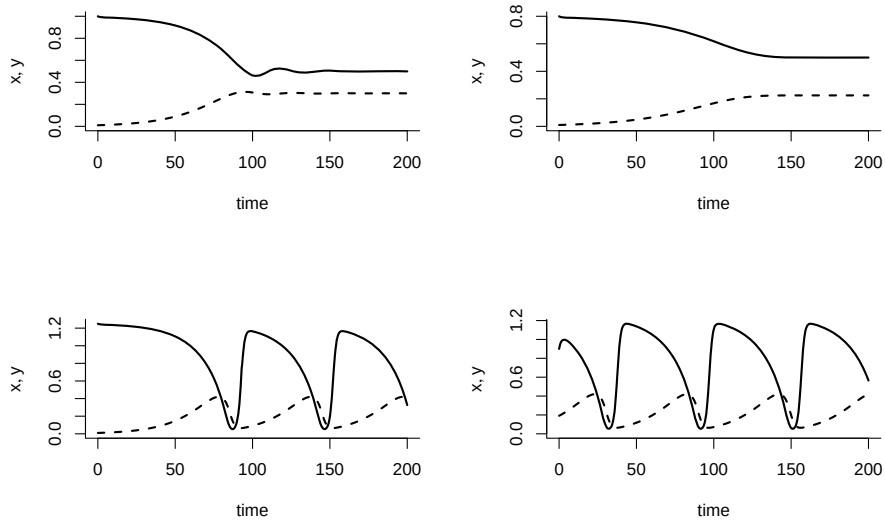


Fig. 8. Solution of the Gause-type predator-prey model (10), (11) with the parameters $r = S = 1$, $d = 0.5$, $\kappa = 0.6$ and the trophic function $\varphi(x) = x/(x + 0.1)$. Prey abundance is plotted by the solid line, predator abundance by the dotted line. Carrying capacity for the prey population varies: $K = 1$ (above left), $K = 0.8$ (above right), $K = 1.25$ (down). The solutions plotted down differs in initial values only.

4.5 Coexistence of three populations

Let us consider a community consisting of three self-supporting species that exhibit mutual interspecific competition and intraspecific competition as well.

Such a community can be modeled by the three dimensional Lotka-Volterra system

$$\begin{aligned} x_1' &= x_1\left(r_1 - a_{11}x_1 - a_{12}x_2 - a_{13}x_3\right), \\ x_2' &= x_2\left(r_2 - a_{21}x_1 - a_{22}x_2 - a_{23}x_3\right), \\ x_3' &= x_3\left(r_3 - a_{31}x_1 - a_{32}x_2 - a_{23}x_3\right), \end{aligned}$$

where all of the parameters are positive. We choice the parameters such that all of the growth rates r_i and all of the intraspecific competition coefficients a_{ii} equal unity and the parameters of interspecific competition satisfy the conditions

$$a_{12} > 1 > a_{21}, \quad a_{13} < 1 < a_{31}, \quad a_{23} > 1 > a_{32}.$$

These inequalities imply that none of the two-species sub-community cannot persist; the first population would exclude the second one, the second population would exclude the third one and the third population would exclude the first one (see simulation provided in the subsection 4.3). We will simulate an invasion of three such species to a vacant territory. The system is solved and the solution is plotted by the following R-language script:

```
LotkaVolterra <- function(t,state,parameters){
  with(as.list(c(state,parameters)),{
         dx1 <- x1*(r1-a11*x1-a12*x2-a13*x3)
         dx2 <- x2*(r2-a21*x1-a22*x2-a23*x3)
         dx3 <- x3*(r3-a31*x1-a32*x2-a33*x3)
         list(c(dx1,dx2,dx3))})}
parameters <- c(r1=1,a11=1,a12=1.25,a13=0.8,r2=1,a22=1,
                a21=0.5,a23=1.2,r3=1,a33=1,a31=1.2,a32=0.9)
state <- c(x1=0.05,x2=0.05,x3=0.05)
time=seq(0,150,by=0.1)
out <- ode(state,time,func=LotkaVolterra,parms=parameters)

yl <- 1.15*max(out[,"x1"],out[,"x2"],out[,"x3"])
plot(time,out[,"x1"],type="l",lwd=2,bty="l",ylim=c(0,yl),
     xlab="time",ylab=expression(list(x[1],x[2],x[3])))
points(time,out[,"x2"],type="l",lwd=2,lty=2)
points(time,out[,"x3"],type="l",lwd=2,lty=3)
legend(0,yl,lwd=c(2,2,2),lty=c(1,2,3),yjust=0.5,horiz=TRUE,
       legend=c(expression(x[1]),expression(x[2]),
                expression(x[3])))
```

The solution is displayed on Fig. 9 left. All of the three populations persist and their abundances oscillate about certain values. Hence, this community can serve as an example of the phenomenon that an increasing complexity leads to larger stability.

We conclude considerations on population models by simulation of a community consisting of a predator feeding on two prey populations linked by competition. Such a community can be modeled by the following system of ordinary

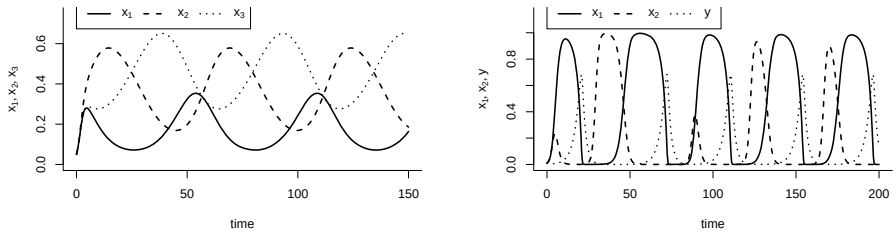


Fig. 9. Left: solution of the Lotka-Volterra model of three competing species; an example of stability arising from complexity. Right: solution of model of predator feeding on two competing species; an example of predator mediated coexistence.

differential equations:

$$\begin{aligned} x_1' &= x_1\,(r1 - a_{11}x_1 - a_{12}x_2) - S_1 y\varphi_1(x_1),\\ x_2' &= x_2\,(r2 - a_{21}x_1 - a_{22}x_2) - S_2 y\varphi_2(x_2),\\ y' &= -dy + \kappa_1 S_1 y\varphi_1(x_1) + \kappa_2 S_2 y\varphi_2(x_2). \end{aligned}$$

Here, x_1 and x_2 denote sizes of the prey populations, y a size of predator population. The sub-community formed by the two prey populations evolves according to the two-dimensional competitive Lotka-Volterra system. Each of the two predator-prey sub-communities evolves like Gause-type predator-prey system. Let us assume that the competitive subsystem is such that the first population excludes the second one, that is $a_{12} < 1 < a_{21}$ provided that $r1 = r2 = a_{11} = a_{22} = 1$. Let the functional response of predator to the first and to the second prey population be of the form

$$\varphi_1(x_1) = \frac{x_1}{x_1 + a_1}, \quad \text{and} \quad \varphi_2(x_2) = \frac{x_2^2}{x_2^2 + a_2},$$

respectively. We simulate a situation where all of the three populations invade to a vacant territory by the following script:

```
phi1 <- function(x){x/(x+0.1)}
phi2 <- function(x){x^2/(x^2+0.1)}
P2P <- function(t,state,parameters){
  with(as.list(c(state,parameters)),{
    F1x <- phi1(x1)
    F2x <- phi2(x2)
    dx1 <- x1*(r1-a11*x1-a12*x2)-S1*y*F1x
    dx2 <- x2*(r2-a21*x1-a22*x2)-S2*y*F2x
    dy <- y*(-d+kappa1*S1*F1x+kappa2*S2*F2x)
    list(c(dx1,dx2,dy))})}
time=seq(0,200,by=0.1)
```

```
out <- ode(c(x1=0.01,x2=0.01,y=0.01),
           time,func=P2P,
           parms=c(r1=1,a11=1,a12=0.5,r2=1,a22=1,a21=1.8,
             d=0.5,S1=1,S2=0.1,kappa1=0.9,kappa2=0.5))

yl <- 1.15*max(out[,"x1"],out[,"x2"],out[,"y"],na.rm=TRUE)
plot(time,out[,"x1"],type="l",lwd=2,bty="l",ylim=c(0,yl),
     xlab="time",ylab=expression(list(x[1],x[2],y)))
points(time,out[,"x2"],type="l",lwd=2,lty=2)
points(time,out[,"y"],type="l",lwd=2,lty=3)
legend(0,yl,lwd=c(2,2,2),lty=c(1,2,3),yjust=0.5,horiz=TRUE,
       legend=c(expression(x[1]),expression(x[2]),
         expression(y)))
```

The result is plotted on Fig. 9 right. We see the coexistence of the three populations, their abundances oscillate. This is an example of predator mediated coexistence of competing species. The second prey population would be excluded by the first one but with presence of predator the both competing populations survives. This means, that the predator is in a sense obligatory mutualist of its prey.

References

1. Bacaër N.: A Short history of Mathematical Population Dynamics. Springer-Verlag: London, 2011
2. Britton N. F.: Essential Mathematical Biology. Springer-Verlag: London, 2003
3. Edelstein-Keshet L.: Mathematical Models in Biology. Random House: New York, NY, 1988
4. Henry M., Stevens H.: A Primer of Ecology with R. Springer: Dordrecht, Heidelberg, London, New York, 2009
5. Kreith K., Chakerian, D.: Iterative Algebra and Dynamic Modeling: a curriculum for the third millennium. Springer-Verlag: New York, 1999
6. Kot M.: Elements of Mathematical Ecology. Cambridge University Press: Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, 2001
7. Murray J. D.: Mathematical Biology I: An Introduction. Springer-Verlag: Berlin, Heidelberg, 2001
8. Pielou E. C.: An Introduction to Mathematical Ecology. J.Willey&Sons, New York, NY, 1969
9. Svirezhev Iu. M., Logofet D.O.: Stability of Biological Communities. Mir: Moscow, 1983
10. Thieme H. R.: Mathematics in Population Biology. Princeton University Press: Princeton and Oxford, 2003
11. Tkadlec E.: Populační ekologie. Struktura, růst a dynamika populací. Univerzita Palackého v Olomouci, Olomouc 2008
12. Turchin P.: Complex Population Dynamics: a theoretical/empirical synthesis. Princeton University Press: Princeton, NJ, 2003
13. `http://www.openoffice.cz/`
14. `http://www.R-project.org/`

Traditional Measures of Diversity, Their Estimates and Sensitivity to Changes

Martin Horáček and Jana Zvárová *

Center of Biomedical Informatics, Department of Medical Informatics,
Institute of Computer Science AS CR,
Pod Vodárenskou věží 2, 182 07 Prague, Czech Republic
horacek@euromise.cz

Abstract. The following article deals with a family of diversity measure functions known as traditional measures of diversity. We deal with sample estimates of traditional measures of diversity, we develop a new estimator and compare its behavior to two established estimators in a simulation study. We also introduce a function that can be used to evaluate the sensitivity of a given diversity measure to changes in a population.

1 Introduction

This paper is devoted to a family of diversity measures called traditional measures of diversity. This family consists of the diversity measures that are functions of the probabilities $(p_1, \ldots, p_r) = \boldsymbol{p}$, where p_i denotes the probability that the investigated feature in an individual randomly chosen from a given population belongs to the class i. Let the number of classes $r \in \mathbb{N}$ be known. We denote the domain of $\boldsymbol{p}$ as Δ^r, i.e.

$$\Delta^r = \left\{ (p_1, \ldots, p_r) : \sum_{i=1}^{r} p_i = 1,\ p_i \geq 0\ \forall\, i \right\}.$$

Denote $\boldsymbol{d}_j^r$ the vector $(\underbrace{0, \ldots, 0}_{j-1}, 1, 0, \ldots, 0) \in \Delta_r$. A traditional measure of diversity H should also satisfy that $H(\boldsymbol{d}_1^r) \geq 0$ and H is a Schur-concave function. As was shown in Horáček (2009), these two properties ensure that $H(\boldsymbol{p})$ is symmetric, nonnegative, $H(\boldsymbol{p})$ is minimal when $\boldsymbol{p} = \boldsymbol{d}_1^r$ and maximal when p_j are identical, equal to $1/r$ for all j.

Most of the traditional measures of diversity are related to the f-entropies proposed by Zvárová (1974) and furter studied in Zvárová, Vajda (2006).

* This work was supported by the project 1M06014 of the Ministry of Education, Youth and Sports of the Czech Republic.

2 Traditional Measures of Diversity

In this section we introduce some of the most common traditional diversity measures like Simpson's index, Shannon's entropy, Renyi's entropy of order α, Hill's index and others. Further we study sample estimates of selected traditional measures of diversity.

2.1 The Most Common Traditional Measures of Diversity.

The most often mentioned and used diversity measures include the number of alleles (or species)

$$H_0(\boldsymbol{p}) = \sum_{i=1}^{r} I_{(0,1]}(p_i) - 1,$$

the Simpson's index

$$H_2(\boldsymbol{p}) = 1 - \sum_{i=1}^{r} p_i^2$$

and the Shannon's entropy

$$H_1(\boldsymbol{p}) = -\sum_{i=1:p_i>0}^{r} p_i \ln p_i.$$

These three indices are generalized by the family of power entropies

$$H_\alpha(\boldsymbol{p}) = (\alpha - 1)^{-1} \left(1 - \sum_{i=1}^{r} p_i^\alpha\right), \quad \text{when } \alpha > 0, \alpha \neq 1,$$

defined as limits when $\alpha = 0$ (number of alleles) and $\alpha = 1$ (Shannon's entropy). When $\alpha = 2$, we get the Simpson's index.

Another frequently mentioned and used indices include the $\boldsymbol{\gamma}$-entropic function

$$H_{\mathrm{A}(\gamma)}(\boldsymbol{p}) = (1 - \gamma)^{-1} \left[1 - \left(\sum_{i=1}^{r} p_i^{1/\gamma}\right)^{\gamma}\right], \quad \text{when } \gamma > 0, \gamma \neq 1,$$

Hill's index

$$H_{\mathrm{H}(\alpha)}(\boldsymbol{p}) = \left(\sum_{i=1}^{r} p_i^\alpha\right)^{\frac{1}{1-\alpha}}, \quad \text{when } \alpha > 0,\ \alpha \neq 1$$

and Rényi's entropy of order $\boldsymbol{\alpha}$

$$H_{\mathrm{R}(\alpha)}(\boldsymbol{p}) = (1 - \alpha)^{-1} \ln \left(\sum_{i=1}^{r} p_i^\alpha\right), \quad \text{when } \alpha > 0, \alpha \neq 1.$$

2.2 Sample Estimates of Traditional Measures of Diversity

Let $\boldsymbol{p} = \{p_i, \ldots, p_r\} \in \Delta^r$ be a vector of unknown probabilities p_i that a randomly chosen individual has allele of type A_i on a given locus. For the sake of simplicity, assume that there is only one allele on every locus. We deal with estimates of a diversity function in the form

$$H(\boldsymbol{p}) = F\left(\sum_{i=1}^{r} h(p_i)\right),$$

where F and h are an arbitrary continuous functions. The estimate is done on the basis of relative frequencies $\hat{\boldsymbol{p}}_n = (X_1/n, \ldots, X_r/n) = (\hat{p}_i, \ldots, \hat{p}_r)$ of alleles observed in a sample of n individuals selected from the population randomly with replacement. In that case, the cumulative distribution of the vector $\boldsymbol{X} = (X_1, \ldots, X_r)$ is multinomial $\mathrm{M}(n, \boldsymbol{p})$.

The most commonly used estimator, often called "plug-in" estimator, consists in simply replacing the unknown probabilities p_i with the observed relative frequencies $\hat{p}_i$. However, despite $\hat{p}_i$ is an unbiased estimate of p_i, the plug-in estimator is generally not unbiased.

Sometimes, the bias could be easily corrected. For example, the mean value of the plug-in estimate of Simpson's index is

$$\begin{aligned}
\mathsf{E}H_2(\hat{\boldsymbol{p}}_n) &= 1 - n^{-2}\sum_{i=1}^{r}\mathsf{E}X_i^2 = 1 - n^{-2}\sum_{i=1}^{r}\left[\mathsf{var}X_i + (\mathsf{E}X_i)^2\right] \\
&= 1 - n^{-2}\sum_{i=1}^{r}\left[np_i(1-p_i) + n^2p_i^2\right] \\
&= \left(1 - n^{-1}\right)H_2(\boldsymbol{p}).
\end{aligned}$$

Thus, the unbiased estimate of Simpson's index is given by

$$\hat{H}_2(\hat{\boldsymbol{p}}_n) = n(n-1)^{-1}H_2(\hat{\boldsymbol{p}}_n).$$

However, it often difficult to find an unbiased estimate of other diversity measure functions. For example, it can be shown that Shannon's index doesn't have an unbiased estimate (Blyth 1959). Hence, several authors dealt with this problem and suggested more sophisticated estimators. We start from the estimator proposed by Bonachela et al. (2008) that is called the balanced estimator. We suggest a modification of this estimator that takes into account the likely distribution of values of p_i in the interval $[0, 1]$.

Bonachela et al. proposed their estimator in the form

$$\hat{H}(\boldsymbol{X}) = F\left(\sum_{i=1}^{r}\zeta(X_i)\right),$$

where the function ζ is chosen to minimize

$$\Phi_\zeta^2(p_i) = [\mathsf{E}(\zeta(X_i) - h(p_i))]^2 + \mathsf{var}(\zeta(X_i))$$

possibly weighed by a function $w(p_i)$ when we have some prior knowledge about the distribution of values $p_i \in [0,1]$. This way, if we ignore the possible influence of the function F and the correlations, we can simultaneously reduce the variance and the square of bias of the estimate. The weighted average error is then given by

$$\bar{\Phi}^2_\zeta(p_i) = \int_0^1 \Phi^2_\zeta(p_i) w(p_i) \mathrm{d}p_i. \tag{1}$$

The necessary condition for minimality is a zero value of the derivations

$$\frac{\delta}{\delta\zeta(k)} \bar{\Phi}^2_\zeta(p_i) = 0, \quad k \in \{0, \ldots, n\}.$$

Therefore, we choose such ζ that

$$\frac{\delta}{\delta\zeta(k)} \int_0^1 \Big[h^2(p_i) - 2h(p_i) \sum_{j=0}^{n} P(X_i = j)\zeta(j) + \sum_{j=0}^{n} P(X_i = j)\zeta^2(j) \Big] w(p_i) \mathrm{d}p_i = 0$$

which can be simplified to

$$\int_0^1 \Big[\zeta(k) P(X_i = k) - h(p_i) P(X_i = k) \Big] w(p_i) \mathrm{d}p_i = 0.$$

If we use the relation

$$P(X_i = k) = \binom{n}{k} p_i^k (1 - p_i)^{n-k},$$

we get

$$\zeta(k) = \frac{\int_0^1 h(p_i) w(p_i) \binom{n}{k} p_i^k (1 - p_i)^{n-k} \mathrm{d}p_i}{\int_0^1 w(p_i) \binom{n}{k} p_i^k (1 - p_i)^{n-k} \mathrm{d}p_i}. \tag{2}$$

Bonachela et al. (2008) derived the form of balanced estimator for Shannon's and power entropies with the weight function equal to 1 on the whole interval $[0,1]$ of possible values of p_i. However, this choice doesn't reflect the distribution of values p_i very well for $r > 2$. When $r >> 2$, most p_i: $i \in \{1, \ldots, r\}$ are very small and only one p_i can be possibly greater than 0.5.

Based on this reasoning, we suggest to choose the weight function $w(p_i)$ proportionally to the marginal density of random variable Y_1 when the corresponding random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_r)$ is uniformly distributed on Δ^r.

This marginal density is proportional to

$$\begin{aligned} f(y_1) &\propto \int_0^{1-y_1} \ldots \int_0^{1-y_1-\ldots-y_{r-2}} \mathrm{d}y_{r-1} \ldots \mathrm{d}y_2 \\ &= \frac{(1-y_1)^{r-2}}{(r-2)!}, \end{aligned}$$

which is (outside a multiplicative constant) a density of Beta distribution $\mathsf{B}(1, r-1)$.

We found the ζ function that minimizes (1) with $w(p_i)$ chosen as $(1-p_i)^{r-2}$ and we derived the corresponding estimator. We call this new estimator a β-estimator. We describe the derivation of the β-estimator for Shannon's entropy, i.e. when $h(p_i) = -p_i \ln p_i$ and $F(x) = x$. The symbols Γ, B and Ψ denote the Gamma, Beta and Digamma functions, respectively.

First, replace $h(p_i)$ and $w(p_i)$ with the appropriate forms and calculate the integral in the denominator of equation (2)

$$\zeta(X_i) = \frac{\int_0^1 h(p_i) p_i^{X_i} (1-p_i)^{n-X_i+r-2} \mathrm{d}p_i}{B(X_i+1, n-X_i+r-1)}.$$

The partial derivation of Beta function satisfies

$$\frac{\delta}{\delta x} B(x,y) = B(x,y)[\Psi(x) - \Psi(x+y)],$$

and the numerator can be expressed as

$$\begin{aligned}
&\int_0^1 h(p_i) p_i^{X_i} (1-p_i)^{n-X_i+r-2} \mathrm{d}p_i \\
&= -\int_0^1 p_i \ln(p_i) p_i^{X_i} (1-p_i)^{n-X_i+r-2} \mathrm{d}p_i \\
&= -\lim_{\alpha \to 0} \int_0^1 \frac{p_i^\alpha - 1}{\alpha} p_i^{X_i+1} (1-p_i)^{n-X_i+r-2} \mathrm{d}p_i \\
&= \lim_{\alpha \to 0} \frac{1}{\alpha} [B(X_i+2, n-X_i+r-1) - B(X_i+\alpha+2, n-X_i+r-1)] \\
&= B(X_i+2, n-X_i+r-1)[\Psi(n+r+1) - \Psi(X_i+2)].
\end{aligned}$$

Therefore, the ζ function follows

$$\begin{aligned}
\zeta(X_i) &= \frac{X_i+1}{n+r}[\Psi(n+r+1) - \Psi(X_i+2)] \\
&= \frac{X_i+1}{n+r} \sum_{k=X_i+2}^{n+r} \frac{1}{k}
\end{aligned}$$

and the β-estimator of Shannon's entropy is

$$\hat{H}(\boldsymbol{X}) = \sum_{i=1}^{r} \frac{X_i+1}{n+r} \sum_{k=X_i+2}^{n+r} \frac{1}{k}.$$

The β-estimate for power entropies, whose satisfy $F(x) = x$ and $h(p_i) = (\alpha-1)^{-1}(p_i - p_i^\alpha)$, could be calculated in a similar manner. With the weight function chosen as $w(p_i) = (1-p_i)^{r-2}$, the β-estimator of power entropies satisfies

$$\hat{H}_\alpha(\boldsymbol{X}) = (\alpha-1)^{-1} \left[1 - \sum_{i=1}^{r} \frac{B(n+r, \alpha)}{B(X_i+1, \alpha)} \right].$$

On Fig. 1 to Fig. 2, we can see a comparison of the β-estimator, Bonachela's original balanced estimator and the plug-in estimator in a population with 6 possible different alleles distributed as $\boldsymbol{p} = (24/50, 11/50, 9/50, 3/50, 2/50, 1/50)$. The figures show the average sample mean and sample variance computed out of 300 trials. The figures were done in R (2011). Generally, when $r > 2$ and the probabilities p_i are not all equal or nearly equal, the β-estimator has a lesser bias and a lower variance than the other two estimators.

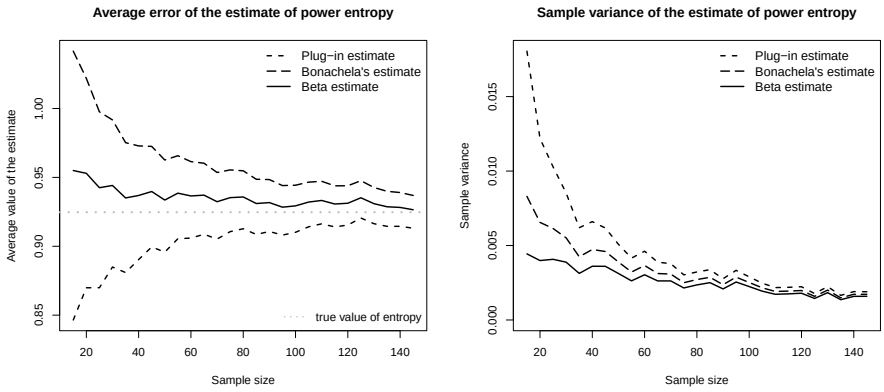


Fig. 1. The average sample mean and sample variance of estimate of power entropy $H_{3/2}$ given a sample size.

3 Sensitivity to Changes

The indices used to measure genetic diversity differ more or less in their qualities. Their characteristic that is frequently of interest is the rate of the change in value of diversity measure connected to changes in frequencies of alleles of a chosen gene. Several authors dealt with this problem, namely Boyle et al. (1990), who were interested mostly in the empirical results, and Izsak (1996), who tried to construct a sensitivity measure on a theoretical background. On the suggestion of I. Vajda, we propose a sensitivity measure that is similar as the Izsak sensitivity, but is easier to compute and has a clearer interpretation.

Define the *sensitivity of diversity measure H to changes in the j-th group* as

$$S_H(\boldsymbol{p}|\boldsymbol{d}_j^r) = \lim_{\epsilon \to 0_+} rp_j \frac{H((1-\epsilon)\boldsymbol{p} + \epsilon \boldsymbol{d}_j^r) - H(\boldsymbol{p})}{\epsilon}.$$

Calculated this way, if the sensitivity to changes of a diversity measure H in a j-th allele A $S_H(\boldsymbol{p}|\boldsymbol{d}_j^r)$ is 3 times greater than the sensitivity to changes $S_H(\boldsymbol{p}|\boldsymbol{d}_k^r)$ in a k-th allele B, it means that a small, say a 10 % increase in the frequency

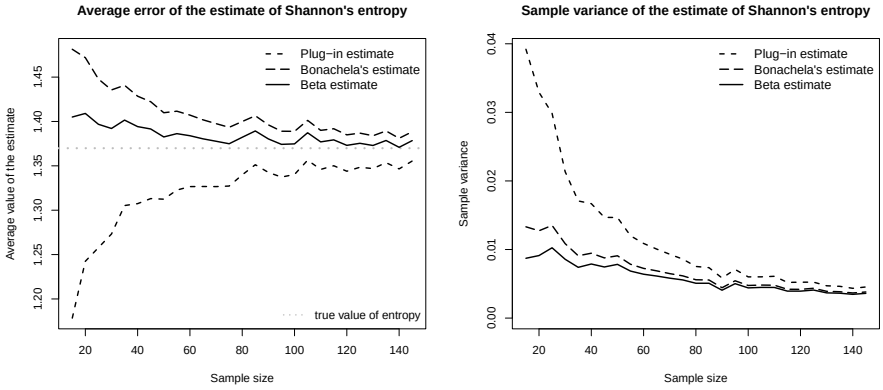


Fig. 2. The average sample mean and sample variance of estimate of Shannon's entropy H_1 given a sample size.

of allele A results in about 3 times greater change of value of H than a 10 % increase in the frequency of allele B.

3.1 Sensitivity of Power and Renyi's Entropies

We have derived the form of sensitivity function for several traditional diversity measures in Horáček (2009). Here we present the sensitivity of power entropies and of the Renyi's entropy of order α. If all $p_i > 0$, the sensitivity of power entropies satisfies

$$S_{H_\alpha}(\boldsymbol{p}|\boldsymbol{d}_j^r) = \frac{\alpha}{\alpha - 1} \sum_{i=1}^{r} p_i^{\alpha-1}(p_i - \delta_{ij})$$

when $\alpha \neq 1$, and

$$S_{H_1}(\boldsymbol{p}|\boldsymbol{d}_j^r) = \frac{\alpha}{\alpha - 1}$$

where $\delta_{ij} = 1$ iff $i = j$, otherwise $\delta_{ij} = 0$. The sensitivity of Renyi's entropy equals to

$$S_{H_{R(\alpha)}}(\boldsymbol{p}|\boldsymbol{d}_j^r) = \frac{\alpha}{1 - \alpha} \frac{\sum_{i=1}^{r} p_i^{\alpha-1}(\delta_{ij} - p_i)}{\sum_{i=1}^{r} p_i^{\alpha}}.$$

If we choose to compare sensitivities of different diversity indices, it is suitable to normalize them to a common scale. The most natural way to do this is to divide them by $[\max(H) - \min(H)]$, where $\max(H)$ (resp. $\min(H)$) is the maximal (minimal) value of corresponding diversity index H on Δ^r. We call the quantity

$$\frac{S_H(\boldsymbol{p}|\boldsymbol{d}_j^r)}{\max(H) - \min(H)}$$

the relative sensitivity (of diversity measure H to changes in the j-th group).

A comparison of the relative sensitivity of various power entropies and of Renyi's entropies in a population with $\boldsymbol{p} = (24/50, 11/50, 9/50, 3/50, 2/50, 1/50)$ is shown in Fig. 3. We can see for example that with increasing α, Renyi's entropy seems to be more responsive to relative changes in the more frequent alleles, but less sensitive to relative changes in the less frequent alleles. This behavior could influence the choice of diversity index for a given problem.

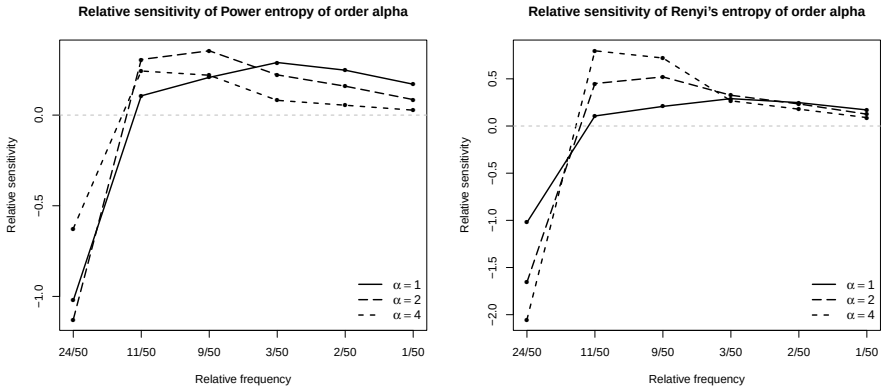


Fig. 3. Comparison of the sensitivity to changes of power entropies

4 Conclusions

We dealt with the traditional measures of diversity and investigated some of their properties. We introduced a new estimator of traditional diversity measures and we showed that it compares favorably to two established estimators, namely to the plug-in estimator and to the balanced estimator proposed Bonachela et al. (2008). We also presented a new way to measure sensitivity of measures of diversity to changes thet could be helpful when we want to select an appropriate measure of diversity for a given study.

References

Blyth, C. R.: Note on estimating information. Annals of Math. Stat. **30** (1959) 71–79

Bonachela, J. A., Hinrichsen, H., Muñoz, M. A.: Entropy estimates of small data sets. J. of Phys. A: Math. and Theor. **41** (2008) 1–9

Boyle, T. P., Smillie, G. M., Anderson, J. C., and Beeson, D. R.: A sensitivity analysis of nine diversity and seven similarity indices. Research Journal Water Pollution Control Federation **62** (1990) 749–762

Izsak, J.: Sensitivity Profiles of Diversity Indices. Biom. J. **38** (1996) 921–930

Horáček, M.: Measures of biodiversity and their applications. Master thesis, Charles university, Prague, supervisor J. Zvárová (2009)

R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. *http://www.R-project.org* (2011)

Zvárová J.: On Measures of Statistical Dependence. Časopis pro pěstování matematiky **99** (1974) 15–29

Zvárová J., Vajda I.: On Genetic Information, Diversity and Distance. Methods of Inform. in Medicine, **2** (2006) 173–179

Biological Diversity of Benthic Macroinvertebrates as a Tool for Water Management

Světlana Zahrádková[1] and Jiří Jarkovský[2]

[1] Department of Botany and Zoology, Masaryk University, Faculty of Science, Kotlářská 2, 611 37 Brno, Czech Republic
[2] Institute of Biostatistics and Analyses, Masaryk University, Kamenice 3, 625 00 Brno, Czech Republic

Abstract. The ecological status assessment of surface waters, required by the Directive 2000/60/EC, is based on the analyses of various biological quality elements (fishes, benthic macroinvertebrates etc.). The biological diversity analyses play an important role in the assessment system supporting the construction of surface water types which serves as a frame for the type specific assessment and primarily, the classification of the status is based on metrics which reflects taxonomical or functional diversity of biological quality elements. The background of the assessment system and examples of suitable metrics are presented.

Keywords: taxonomic diversity, functional diversity, metrics, diversity indices

1 Introduction

Water is an essential component of all ecosystems and there is no doubt about its importance for humankind. The quality of water is of the same importance as its quantity and accessibility and the attention is paid to these aspects globally. Within the European Union the principles of water policy incl. the protection of aquatic ecosystems are defined in the Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy (Water Framework Directive - WFD) [1]. The directive deals with all types of continental waters and emphasizes biological and ecological aspects in the status assessment and in management generally. Among other things, the directive requires each Member State to assess the ecological status of surface waters within units called water bodies. The status should be classified using five classes (high, good, moderate, poor, and bad). The classification is based on the comparison with a reference status (status with the minimal anthropogenic impact).

The ecological status assessment is based on the analysis of various biological quality elements (e.g. phytobenthos, macrophytes, benthic

macroinvertebrates, and fishes). High ecological status is defined separately for each of these elements. The following definition is used for benthic macroinvertebrates:

- The taxonomic composition and abundance correspond totally or nearly totally to undisturbed conditions.
- The ratio of disturbance sensitive taxa to insensitive taxa shows no signs of alteration from undisturbed levels.
- The level of diversity of invertebrate taxa shows no sign of alteration from undisturbed levels.

Metrics are used to assess the status of biological quality elements. Metric is a measurable part or process of a biological system empirically shown to change in value along a gradient of human influence. It reflects specific and predictable responses of the community to human activities, either to a single factor or to the cumulative effects of all events and activities within a watershed [2].

Only metrics capable of discriminating between "stressed" and "unstressed" conditions are used. Metrics that clearly respond to specific pollutants or stressors are most useful as a diagnostic tool. Furthermore, the metrics used should cover diverse aspects of structure, composition, health and function of the aquatic biota.

Metrics are combined in multimetric indices (MMI). Before being used for the Multimetric Index, each metric result must be transferred into a value between 0 and 1. For a detailed description of the multimetric approach see [3] and [2].The value of the MMI is used for the classification of the water body – i.e. for its assignment to the ecological status class.

The WFD requires so called type-specific approach. Types of surface waters are defined by abiotic environmental variables (e.g. altitude, size, geology etc.). These abiotic water types should be relevant to characteristic species assemblages.

Biological diversity is therefore reflected both in setting a framework for the evaluation (types) and the actual assessment of water status.

2 Taxonomic and functional diversity

Biological diversity can by analysed at different levels. Species (or higher taxa) level in terms of species richness and species diversity is used for the assessment of ecological status of water bodies. Beside this evaluation oriented on the taxonomic units, the functional composition is analysed: each species can be assigned to the functional group according to its specific characteristics or preferences (traits), e.g. to the functional feeding groups

such as predators or grazers. Trait is a well-defined, measurable property of organisms, usually measured at the individual level and used comparatively across species. A functional trait is one that strongly influences organismal performance [4]. Relative abundances of functional trait attributes within a sample of a biological quality element are usually used for the assessment of ecological status (e.g. the share of predators within the sample of benthic macroinvertebrates). Traits are linked to various environmental gradients, e.g. temperature gradient (maximal body size, fecundity, time of emergence of insects etc.); hydrologic gradient (rheophily, body shape, oviposition etc.); chemical gradient (e.g. respiration type). Traits can be common within the higher taxa, usually at the level of genera, which could bring a big advantage of the trait approach, the reduced necessity of precise species identification. Unfortunately, it is not true in general and, therefore, the combination of metrics based on both taxonomic and functional structure and identification at the species level (if possible) is the optimal solution for the assessment of ecological status.

3 Typology, type-specific assessment and reference conditions

The type specific approach required by the WFD means that the reference status should be defined for individual water body types (a small stream at higher altitude differs fundamentally from a large lowland river). Waters should be split into a reasonable number of types. The types are delimited geographically (e.g. by ecoregions or sea drainage basins) and defined by the combination of several environmental variables. Each variable is divided into several categories (e.g. three or four categories are used for altitude). Metric values related to the species assemblages in reference status should have minimal variability within a type and should be different from those in other types.

Information on biological diversity (taxonomic and functional) is used for the development of the typology, both for the selection of suitable environmental variables and their categorisation and for the verification of intra-type and inter-type variability. The selection of appropriate variables is done using different methods such as Principal Component Analysis or SEM analysis. The categorisation of environment variables represents an effort to identify those parts of the gradient in which important changes in biodiversity and/or metrics usually occur. Proposed abiotic typologies should be compared with the classifications of communities according to their taxonomic and/or functional structure. The similarity of both approaches is desirable.

The defined types form a basis for the definition of the reference conditions that are optimally derived from real data from pristine sites. If such data sets do not exist, modelling or expert judgement can be used.

The classification of the ecological status is set within the types, too. There are two important boundaries: the high/good status boundary which is essential for the definition of reference conditions, and the good/moderate status boundary which is of high practical importance – the objective of water management is to achieve at least good status of all waters and, therefore, the results of the classification of water status determine management actions.

4 Metrics used for the ecological status assessment

The metrics based on benthic macroinvertebrates as a frequently used biological element are used here. The following metric types can be distinguished (for formulas see [5]):

(i) Composition/abundance metrics: all metrics giving the share of a taxon or taxonomic group in relation to the total number of individuals counted, all metrics giving the abundance of a taxon or taxonomic group, metrics comparing reference and observed taxa (e.g. similarity indices).
The percentage of abundance of three families - Ephemeroptera, Plecoptera, Trichoptera and percentage of Plecoptera are the composition/abundance metrics usually best correlating to hydromorphological parameters [6]. The decrease indicates mainly various habitat losses, absence of stabile substrata etc.

(ii) Richness/diversity metrics: all metrics giving the number of taxa within a certain taxon (including the total number of taxa), all diversity indices (e.g. Shannon-Wiener index or Margalef index).
Among the richness/diversity metrics the number of Plecoptera taxa and number EPT taxa generally correlates to hydromorphological parameters and also the Shannon-Wiener index is well correlated to them [6] as well as to general degradation.

(iii) Sensitivity/tolerance metrics: all metrics giving the ratio of taxa sensitive and insensitive to stress in general or to a certain stress-type, either using presence/absence or abundance information, e.g. saprobic index which reflects an intensity of organic pollution.

(iv) Functional metrics: all metrics addressing the characteristics of taxa other than their taxonomic definition (biological or ecological traits, ecological guilds): feeding types, habitat preferences, ecosystem type preferences, current preferences, life-history parameters, body-size parameters; they can be based on taxa abundance or richness; e.g. RETI (Rhithron Feeding Type Index - based on proportion of trophic guilds in a sample). The longitudinal zonation measures or RETI are well correlated with hydrological impact like influence of large dams and current preference measures correlates usually to hydromorphological parameters too.

For examples of metrics which are actually used in the Czech Republic see [7].

5 Significance of biodiversity for the surface water management

Biological diversity has an important role for the surface water management in intentions of Water Framework Directive. The analyses of biodiversity support construction of surface water types (typology) which serves as a frame for type specific assessment of status. The classification of the ecological status is based on metrics which predominantly reflects taxonomical or functional diversity of biological quality elements. The results of the classification of water status determine management actions which should lead through six-year river basin management plans to gradual improvement of our environment.

References

1. Council of the European Communities: Directive 2000/60/EC, Establishing a framework for community action in the field of water policy. European Commission PE-CONS 3639/1/100 Rev 1, Luxembourg (2010)
2. CEN/TR 16151: Water quality - Guidance on the design of Multimetric Indices. Technical Report (2011)
3. Karr, J.R., Chu, E.W: Restoring Life in Running Waters: Better Biological Monitoring. Island Press, Washington D.C. (1999)
4. McGill, B.J., Enquist, B.J. Weiher, E., Westoby M.: Rebuilding community ecology from functional traits. Trends in Ecology and Evolution, 21, 178 --185 (2006)
5. AQEM consortium: Manual for the application of the AQEM method. A comprehensive method to assess European streams using benthic macroinvertebrates, developed for the purpose of the Water Framework Directive. Version 1.0 (2002)
6. Hering, D., Meier, C., Rawer-Jost, C., Feld Ch.K., Biss, R., Zenker, A., Sundermann, A., Lohse, S., Boehmer, J.: Assessing streams in Germany with benthic invertebrates: selection of candidate metrics. Limnologica 34, 398--415 (2004)
7. Opatřilová, L., Kokeš, J., Syrovátka, V., Němejcová, D., Zahrádková, S.: Hodnocení tekoucích vod podle makrozoobentosu: popis a vývoj metodiky. VTEI 53, 7--9 (2011)

Species Structure Analysis of the Database of the Czech Forest Site Classification System

Václav Zouhar[1,2], Klára Komprdová[3], Jiří Komprda[3], Milan Sáňka[3], Ondřej Hájek[4], Jiří Jarkovský[3,5], Tereza Kalábová[3,5]

[1] Department of Forest Protection and Wildlife Management, Mendel University in Brno, Zemědělská 3, CZ-613 00, Brno, Czech Republic
[2] Forest Management Institute Brandýs nad Labem, branch Brno, Vrázova 1, CZ-616 00, Brno, Czech Republic
[3] RECETOX (Research Centre for Toxic Compounds in the Environment), Kamenice 126/3, CZ–625 00 Brno, Czech Republic
[4] Department of Botany and Zoology, Masaryk University, Kotlářská 2, CZ-611 37 Brno, Czech Republic
[5] Institute of Biostatistics and Analyses, Masaryk University, Kamenice 126/3, CZ–625 00 Brno, Czech Republic
komprdova@recetox.muni.cz

Abstract. The objective of this study was to assess the importance of the vegetation structure for the individual levels of the Czech forest site classification system (CFSCS) and classify samples from the database of these units. By using species, a classification analysis was carried out by applying the Random Forests method on all CFSCS levels. From the analysis results it can be determined which typological units are well defined in terms of vegetation and which overlap.

Keywords: Czech forest site classification system, Random Forests

1 Introduction

Forest site classification was created and serves as the basis for determining forest management measures in forests as well as operating and manufacturing targets through Forest Management Plans and Guidelines. Its significance grew even further in the new political and environmental relations (after 1989) when it also became the basis for forest ecosystem assessments, forest valuation and for framing caretaking plans for areas subject to special protection [1].

Forest Typology, being an important component of forest management throughout the development period, was introduced in a compact manner in the publication "Typologický systém ÚHÚL" (Czech forest site classification system) [2], where the fundamental properties of the system and the units thereof were described on a more or less general level. The system units were defined based on the author's empirical experience with the application of the then-known scientific knowledge. Subsequently, the CFSCS underwent minor adjustment in 1973. Since then, no major

changes have been made to this system. Even though the Czech Forest Site Classification System was published several times as part of expert papers [3][4][5], it has not been assessed, modified or changed on a comprehensive basis since its establishment.

Insufficient or general definitions in the publication of K. Plíva lead to an inconsistent understanding of the system by its users. The result is inaccurate or varying application of these units in practice, mostly in the preparation of the Forest Site Type Map. This gives rise to the issue of inconsistent contents and formal aspects in the characteristics of CFSCS units. If we realize that CFSCS is the main instrument in the differentiation of our forests for the needs of forest management, funding policies, forest valuation, decision-making of environmental protection authorities and state forest management bodies, then this condition of the typological system is alarming. The extensive Database of Czech Forest Site Classification System of the Forest Management Institute (FMI) is hence an instrument to improve this condition, as the data it contains (almost 50,000 samples from terrain research plots) can be subjected to a comprehensive assessment of the CFSCS and precise definitions of the units can be made. Also, the methods of a practical Forest Site Typology application can be developed for the user.

The objective of this study was to assess the importance of the vegetation structure for the individual levels of the Czech forest site classification system: vegetation tiers (=altitudinal vegetation zones) (VT), edaphic series (ES), edaphic categories (EC) and forest site type complexes (FSTC)) and classify samples from the database of these units. This is therefore the first independent assessment of an expert system based on terrain data. By using taxons, a classification analysis was carried out by applying the Random Forests method on all CFSCS levels. From the analysis results, it can be determined which CFSCS units are well defined in terms of vegetation and which overlap.

1.1 Data Sources

The data set for CFSCS level classification includes 48,439 typological samples in 39,157 Forest site research plot. The Czech forest site classification system consists of hierarchical levels of forest site typological units. Forest site type complexes (178) are given by the combination of nine vegetation tires (1-9) and natural pine forest habitats and 25 edaphic categories. Edaphic categories are defined by soil properties important for management. Vegetation tiers characteristic with their woody structure are the foundation units for indirect representation of altitudinal climate (Fig.1).

SERIES	category		Forest vegetation (altitudinal) belt									categ.
			0 Pine	1 Oak	2 Beech-Oak	3 Oak-Beech	4 Beech	5 Fir-Beech	6 Spruce-Beech	7 Beech-Spruce	8 Spruce	9 Dwarf Pine	
EXTREME	*xerothermal*	X	**0X** - Dealpine Pine	**1X** - Cornelian Cherry-Oak	**2X** - Cornelian Cherry-Beech-Oak	**3X** - Cornelian Cherry-Oak-Beech	**4X** - Dealpine Beech						X
	scrub	Z	**0Z** - Relickt Pine	**1Z** - Scrub Oak	**2Z** - Scrub Beech-Oak	**3Z** - Scrub Oak-Beech	**4Z** - Scrub Beech	**5Z** - Scrub Fir-Beech	**6Z** - Scrub Spruce-Beech	**7Z** - Scrub Beech-Spruce	**8Z** - Rowan-Spruce	**9Z** - Dwarf Pine	Z
	skeletal	Y	**0Y** - Ravine Pine		**2Y** - Skeletal Beech-Oak	**3Y** - Skeletal Oak-Beech	**4Y** - Skeletal Beech	**5Y** - Skeletal Fir-Beech	**6Y** - Skeletal Spruce-Beech	**7Y** - Skeletal Beech-Spruce	**8Y** - Skeletal Spruce	**9Y** - Skeletal alpine tundra	Y
ACIDIC	*nutrient-very poor*	M	**0M** - Nutriet-very poor (Oak)-Pine	**1M** - Nutriet-very poor Pine-Oak	**2M** - Nutriet-very poor Beech-Oak	**3M** - Nutriet-very poor Oak-Beech	**4M** - Nutriet-very poor Beech	**5M** - Nutriet-very poor Fir-Beech	**6M** - Nutriet-very poor Spruce-Beech	**7M** - Nutriet-very poor Beech-Spruce	**8M** - Nutriet-very poor Spruce		M
	acidic	K	**0K** - Acidic (Oak-Beech)-Pine	**1K** - Acidic Oak	**2K** - Acidic Beech-Oak	**3K** - Acidik Oak-Beech	**4K** - Acidic Beech	**5K** - Acidic Fir-Beech	**6K** - Acidic Spruce-Beech	**7K** - Acidic Beech-Spruce	**8K** - Acidic Spruce	**9K** - Acidic Dwarf Pine	K
	compacted-acid	I		**1I** - Compacted-acid (Hornbeam-) Oak	**2I** - Compacted-acid Beech-Oak	**3I** - Compacted-acid Oak-Beech	**4I** - Compacted-acid Beech	**5I** - Compacted-acid Fir-Beech	**6I** - Compacted-acid Spruce-Beech				I
	stony-acidic	N	**0N** - Spruce-Pine and/or Pine-Spruce	**1N** - Stony-acid (Hornbeam-) Oak	**2N** - Stony-acid Beech Oak	**3N** - Stony-acid Oak-Beech	**4N** - Stony-acid Beech	**5N** - Stony-acid Fir-Beech	**6N** - Stony-acid Spruce-Beech	**7N** - Stony-acid Beech Spruce	**8N** - Stony-acid Spruce		N
NUTRIENT-RICH	*nutrient-medium*	S		**1S** - Sandy (Hornbeam-) Oak	**2S** - Nutrient-medium Beech-Oak	**3S** - Nutrient-medium Oak-Beech	**4S** - Nutrient-medium Beech	**5S** - Nutrient-medium Fir-Beech	**6S** - Nutrient-medium Spruce-Beech	**7S** - Nutrient-medium Beech-Spruce	**8S** - Nutrient-medium Beech-Spruce		S
	water-deficient	C	**0C** - Serpentine Pine	**1C** - Water deficient Hornbeam-Oak	**2C** - Water deficient Beech-Oak	**3C** - Water deficient Oak-Beech	**4C** - Water deficient Beech	**5C** - Water deficient Fir-Beech					C
	slope-stony nutrient-medium	F				**3F** - Slope-stony nutrient-medium	**4F** - Slope-stony nutrient-medium Oak-Beech	**5F** - Slope-stony nutrient-medium Beech	**6F** - Slope-stony nutrient-medium	**7F** - Slope-stony nutrient-medium Beech	**8F** - Slope-stony nutrient-medium		F
	loamy	H		**1H** - Loess Hornbeam-Oak	**2H** - Loamy Beech-Oak	**3H** - Loamy Oak-Beech	**4H** - Loamy Beech	**5H** - Loamy Fir-Beech	**6H** - Loamy Spruce-Beech				H
	nutrient-rich	B		**1B** - Nutrient-rich Hornbeam-Oak	**2B** - Nutrient-rich Beech-Oak	**3B** - Nutrient-rich Oak-Beech	**4B** - Nutrient-rich Beech	**5B** - Nutrient-rich Fir-Beech	**6B** - Nutrient-rich Spruce-Beech	**7B** - Nutrient-rich Beech-Spruce			B
	limestone	W		**1W** - Limestone Hornbeam-(Beech)-Oak	**2W** - Limestone Beech-Oak	**3W** - Limestone Oak-Beech	**4W** - Limestone Beech	**5W** - Limestone Fir-Beech					W
MAPLE	*enriched-colluvial*	D		**1D** - Enriched-colluvial Hornbeam-Oak	**2D** - Enriched-colluvial Beech-Oak	**3D** - Enriched-colluvial Oak-Beech	**4D** - Enriched-colluvial Beech	**5D** - Enriched-colluvial Fir-Beech	**6D** - Enriched-colluvial Spruce-Beech				D
	stony-colluvial	A		**1A** - Stony-colluvial Maple-Hornbeam-Oak	**2A** - Stony-colluvial Maple-Beech-Oak	**3A** - Stony-colluvial Lime-Oak-Beech	**4A** - Stony-colluvial Lime-Beech	**5A** - Stony-colluvial Maple-Beech	**6A** - Stony-colluvial Maple-Spruce-Beech	**7A** - Stony-colluvial Maple-Beech-Spruce	**8A** - Stony-colluvial Maple-Spruce		A
	talus	J		**1J** - Hornbeam-Maple		**3J** - Lime-Maple		**5J** - Talus (Elm-Ash)-Maple	**6J** - Talus (Elm-Spruce)-Maple				J
ASH	*floodplain*	L		**1L** - Elm floodplain	**2L** - Stream floodplain	**3L** - Ash-Alder		**5L** - Montane Ash-Alder	**6L** - Speckled Alder floodplain				L
	ravine	U		**1U** - Poplar floodplain		**3U** - Maple-Ash		**5U** - Moist Ash-Maple					U
	moist to wet	V		**1V** - Moist to wet Hornbeam-Oak	**2V** - Moist to wet Beech-Oak	**3V** - Moist to wet Oak-Beech	**4V** - Moist to wet Beech	**5V** - Moist to wet Fir-Beech	**6V** - Moist to wet Spruce-Beech	**7V** - Moist to wet Beech-Spruce	**8V** - Moist to wet Maple-Spruce		V
GLEYED	*nutrient-medium*	O	**0O** - Nutrient-medium Fir-Oak-Pine	**1O** - Nutrient-medium Lime-Oak	**2O** - Nutrient-medium Fir-(Beech)-Oak	**3O** - Nutrient-medium Fir-Oak-Beech	**4O** - Nutrient-medium Oak-Fir	**5O** - Nutrient-medium (Beech-) Fir	**6O** - Nutrient-medium Spruce-Fir	**7O** - Nutrient-medium Fir-Spruce	**8O** - Nutrient-medium Spruce		O
	acidic	P	**0P** - acidic Fir-Oak-Pine	**1P** - Acidic Birch-Oak	**2P** - Acidic Oak	**3P** - Acidic Fir-Oak	**4P** - Acidic Oak-Fir	**5P** - Acidic Fir	**6P** - Acidic Spruce-Fir	**7P** - Acidic Fir-Spruce	**8P** - Acidic Spruce		P
	nutrient-poor	Q	**0Q** - Nutrient-poor Fir-Oak-Pine	**1Q** - Nutrient-poor Birch-Oak	**2Q** - Nutrient-poor Oak	**3Q** - Nutrient-poor Fir-Oak	**4Q** - Nutrient-poor Oak-Fir	**5Q** - Nutrient-poor Fir	**6Q** - Nutrient-poor Spruce-Fir	**7Q** - Nutrient-poor Fir-Spruce	**8Q** - Nutrient-poor Spruce		Q
WET	*nutrient-poor wet*	T	**0T** - Nutrient-poor wet Birch-Pine	**1T** - Nutrient-poor wet Birch-Alder	**2T** - Nutrient-poor wet Fir-Oak	**3T** - Nutrient-poor wet Fir-Spruce-Oak	**4T** - Nutrient-poor wet Oak-Fir	**5T** - Nutrient-poor wet Oak-Spruce-Fir	**6T** - Nutrient-poor wet Spruce/Fir	**7T** - Nutrient-poor wet Fir/Spruce	**8T** - Nutrient-poor wet Scrub Spruce		T
	nutrient-medium wet	G	**0G** - Nutrient-medium wet Spruce-Pine	**1G** - Willow-Alder	**2G** - Nutrient-medium wet Fir-Oak	**3G** - Nutrient-medium wet Fir (-Spruce)-Oak	**4G** - Nutrient-medium wet Oak-Fir	**5G** - Nutrient-medium wet Fir	**6G** - Nutrient-medium wet Spruce-Fir	**7G** - Nutrient-medium wet Fir-Spruce	**8G** - Nutrient-medium wet Spruce		G
	paet	R	**0R** - Peat Pine		**3R** - Acidic relict Spruce		**4R** - Nutrient-medium Relict Spruce	**5R** - Peat Pine-Spruce	**6R** - Nutrient-medium Peat Spruce	**7R** - Acidic Peat Spruce	**8R** - Raised Bog Spruce	**9R** - Raised Bog Dwarf Pine	R

Fig. 1. Czech forest site classification system (CFSCS).

Most forest site typological samples come from 1950 to 1980. The sampling covers forest areas of the Czech Republic in a rather representative way, with the exception of certain areas (e.g. Ore Mountains) (Fig.2). The average distance of the closest plots was 350m. Almost 90% of samples have the size of 400-500sqm and 20% of plots were sampled more than once.

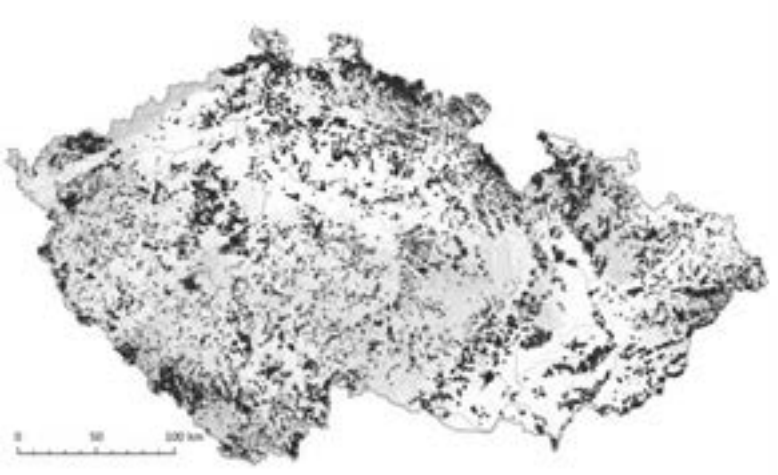


Fig. 2. Distribution of forest site research plot used for the analyses based on species structure.

2 Methodology

2.1 Random Forests Method for Classification

Random Forests are the extension to decision trees. They may be applied for classification and regression, and remove certain difficulties stemming from the use of trees, mostly their instability [6]. Forests are mostly used for classification and the determination of predictors' significance. For classification forests, each observation is classified into one of the dependent variable categories based on predictors. The classification result is achieved by a majority vote. The most well-known forests include Bagging, Boosting and Random Forests.

Random Forests [7] are one of the most recent techniques applied for classification issues. This technique was developed for sets containing a large amount of variables and samples. A big advantage is that the variables may be continuous or categorical and may correlate. Since this is a non-parametric technique, no specific distribution of variables is required. The only disadvantage is the testing of the settings of the individual parameters of this method.

A Random Forest consists of a varying number of trees. For the Random Forests method, binary trees of the CART [8] type are used. Training sets for the individual trees T are bootstrapped selections from the data set L. Bootstrapped selections are created by a random selection with replacement of size n and each of them is used for building one tree (in our case the "classification" tree). Observations not contained in the i-th bootstrapped selection L_i for growing a T_i tree, are used for estimating the generalization (overall) error of this tree. These estimations are called OOB (out-of-bag) estimations. In Random Forests the objective is not to grow an optimum tree, but, on the contrary, large trees are grown which are no longer pruned. When selecting the branching for the respective node, then certain m of input variables (predictors) $X_1,..,X_M$ are selected from M and the best branching of the respective tree is searched only among the branching based on selected m variables. See more in [7]. Random Forests therefore uses both bagging and a random variable selection [9] for building a tree.

The model created this way allocates importance to each predictor (species) in a 0-100 range which determines how suitable this variable is for the classification of a dependent variable (e.g. forest-type group).

Random Forests produces several importance measurements of variables. One of the most precise determinations of this quantity is a measurement of misclassification rate, where the variable values in the tree node are randomly permutated.

For the forest algorithm, the right number of variables (m) and the number of trees (*ntree*) in the forest must be selected. The determination of these parameters is experimental to a certain degree and requires experience. The conventional method is to carry out experiments with different settings of these parameters, so that a forest can be obtained which shows the lowest possible error rate. Since testing is very time-consuming (especially in the case of sets containing thousands of samples), it makes sense to select as many trees as is sufficient for optimum classification. Therefore, a larger number of trees are tested first. After a certain period of time, the trees start to converge into a correct value for the OOB estimate.

With the help of forests, we are able to find out not only the classification percentage for the respective category and the values of variable importance, but we can also determine the suitable species combination for the prediction into categories of typological levels and prototype categories.

We know the percentage of correct classification for each sample. The prototype is the "core" in the category, containing samples which are classified by trees with an accuracy of over 50%. By using these samples, we can calculate e.g. the median and quartile (for species coverage) and/or percentage representation (for presence-absence taxons) for each dependent variable category. The prototype hence provides an overall idea of the relationship between variables and classification.

For each typological level, the setting combinations *ntree* = {100, 300, 500,1000,2000} a m = {1,2,3,4,5,6,7,8,9,10} were tested. The overall forest error rate for biotic classification was stable for 600 trees and did not change any further. The minimum setting of the *ntree* value is therefore 600 trees in our case. The m parameter was different for different typological levels.

Since trees are left to grow big (the usual number of samples in the terminal node is 5) and are not pruned retrospectively, it is obvious that samples within the same terminal node will be very similar. For each sample pair it can be calculated how

many times the samples occurred within the same terminal node. This rate is called "proximity" or "closeness" and ranges from 0 to 1 where 1 indicates maximum proximity (e.g. when the samples occurred together in the terminal node of 800 out of 1000 trees, the proximity equals 0.8). If proximity is counted mutually from all samples, we obtain the association matrix. The similarity association matrix ("proximity") may be used further for multi-dimensional analyses, because it meets all the conditions. Also outlier samples may be defined by means of proximity. As outlier samples are considered those that show the smallest proximity to samples within its classification category. Besides proximity measurement, also a probability measurement may be applied.

2.2 Analysis of Forest Site Type Complexes Overlapping

This classification was calculated for edaphic categories and forest tiers, the combination of which forms Forest site type complexes (FSTC), the most important unit in the classification system.

After making all the taxonomic adjustments a total of 791 species were available. Different taxonomic levels were used for the classification. By doing so, it can be determined what taxonomic level better defines the CFSCS units.

Typological records, samples in the same site (n = 3410) or within a distance of less than 500m (n = 2317), were used as testing records in the classification, so that autocorrelation influence and duplication of the information in classification analysis could be avoided.

Data on species in two variants were used for the classification analysis. In the first instance, only the presence/absence within a site is recorded for plant species. In the second instance, their actual coverage is identified. With standard settings of the analyses where the same weight is allocated to all classification units, groups with a smaller amount of records were underestimated. This was caused mostly by the distribution of these categories in the environment (e.g. edaphic category "R") and partly also by data collection. Therefore, further classification variants were calculated, with weighting of the categories. When creating the model, records for the testing and training set were selected more times for less represented categories than better represented ones. This may however also result in a certain distortion of results if the irregularity of categories is caused by data collection and not the representation in an environment. Also, there is bigger loss of variability in these groups and a slight overestimation of classification results. Still, despite these uncertainties, the results obtained by category weighting are more precise and reflect more suitable environment conditions than the results obtained by non-weighted classification.

For all CFSCS levels, the percentage of correct classification was determined for all of the above mentioned settings (a total of 8 variants of classification analyses for each level).

3 Results

3.1 Classification of Edaphic Categories

Edaphic categories contain a total of 25 categories. The data set for edaphic category analysis contained 17,979 plots and 791 species. The number of species required for the best possible classification into edaphic categories by the Random Forest method was 86. Figure 3 shows the 50 most significant species.

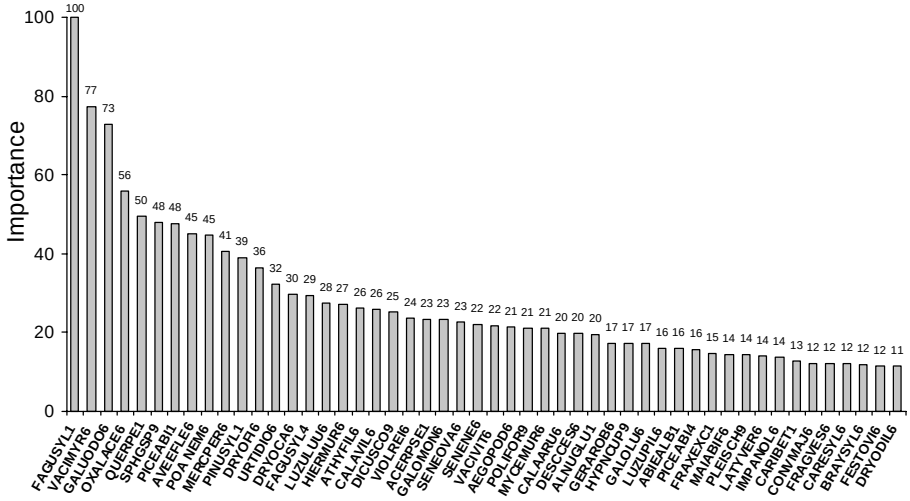


Fig. 3. Significance of species for classification into edaphic categories (based on coverage).

The results of classification into edaphic categories for coverage and presence-absence are very similar. For edaphic categories R, Z, S, Q, C and Y the percentage of classification by using coverage was approximately about 10% higher than with presence-absence (Fig. 4).

The best classified Edaphic categories by the Random Forests method were X, W, L, U and J (for abundance also F and I) where the correct classification percentage (hereafter only CP) was over 70%.

Other, relatively well-classified categories were Y, R, T, C (CP > 60%), but only for the classification with species coverage – for the presence-absence Edaphic categories with percentage classification between 60% and 70% were not represented at all.

The least well-defined edaphic category was K. Categories with low classification were D, B, N, S and A.

In the classification there was an overlap in EC from the G series - wet , specifically between T, G and R edaphic category. Category K overlaps in its species structure mostly with EC within the K series - acidic and with Y. In another, less

well-defined S category, there is an overlap with F, Y and N. Category A is most often classified as J, F and W, and category Q as M and T. Even though the M and W categories showed a high percentage of classified samples, we cannot say that these are well defined categories, because also other EC are classified into them. A relatively well-defined EC based on species structure appears to be X into which however also C is classified with 13%, L with a slight overlap with U, and I (as for coverage it overlaps only with K and as for presence-absence also with C).

The results of the edaphic categories show clearly that the overlap between Forest site type complexes will exist not only within a single edaphic category, but also among FSTC from other edaphic categories.

presence/absence

Edaphic Series			EXTREME			ACIDIC				NUTRIENT-RICH						MAPLE			ASH			GLEYED			WET		
Number of plots	Percentage of classification (%)	Edaphic Categories	X	Z	Y	M	K	N	I	S	F	C	B	W	H	D	A	J	L	U	V	O	P	Q	T	G	R
76	92	X	92									5		3													
625	29	Z	4	29	9	22	2	6	4		1	7						1					1	2	2	7	3
156	57	Y		3	57	8	1	11	1	1	3	1			1			3					4		1	2	3
541	71	M		3	5	71	2	3	2			1											1	7	2	1	2
2422	10	K		5	13	20	10	14	10	2	1	3											6	5	1	7	1
514	32	N		3	24	4	5	32	6	3	4	3						1					5	1	1	8	1
383	69	I		2	5	2	1	2	69	1		5		1	1							2	6	3			
1790	19	S	3	3	12	2	2	9	7	19	12	6	3	2	2	1	1	2	1	1	3	4	2			2	1
336	69	F			6			3		3	69		1	2		1	2	3		3	4		1			3	
675	46	C	13	4	2	2		1	9	1	1	46	1	5	2		1	1				1	2	4			
2246	33	B	2		2			1	1	2	12	4	35	11	6	4	5	8	1	4	3	1					
114	82	W	5							1	1	1		82	4	1	2	1		4							
1148	52	H	6		1			1	8	1	2	5	2	5	52	1		1	1	2	1	9	1				
1078	36	D	2								6	1	2	12	3	36	2	7	7	15	3	3					
1209	24	A	2	1	4			1		1	12	4	2	13	1	1	24	27		4	2					1	
509	70	J	3		1						6	1	1	6		1	3	70		6							
836	77	L														1			77	11	1				3	3	1
201	71	U									2			4		1		2	15	71	2						
524	39	V			4			1			8		1	1		1	1	2	9	14	39	6	2		1	7	2
698	52	O	1		3	1		1	8		1			1	2	2			1	2	4	52	8	2	1	7	1
301	37	P		2	4	5	1	4	7										1		1	4	37	10	8	12	3
181	43	Q		1		26		1	7														4	43	14	2	3
217	55	T		1	1	4		1											4				3	11	55	7	12
617	47	G		1	2	3		2											8	1	3	2	5	4	13	47	10
525	47	R		2	2	3		1											1	1			1	3	21	18	47

Fig. 4. Classification of edaphic categories and their overlap due to vegetation. Numbers are percentage of classification, empty cells are zeros.

Based on samples (with the probability of classification into the edaphic category over 0.5 – (see fig 15 and 16) the percentage representation (in the case of presence-absence) and median (in the case of coverage) in edaphic categories were calculated for all species used for the classification. This is called edaphic category prototype. It can also be determined this way for which EC the respective taxon was important.

3.2 Classification of Vegetation Tiers

For vegetation tier (VT) and natural pine forest habitats classification, it was sufficient to reach a combination of 27 species to achieve the best classification. Very similar results can be achieved by combining different species with similar environmental conditions. Figure 5 shows the 30 most important species for classification.

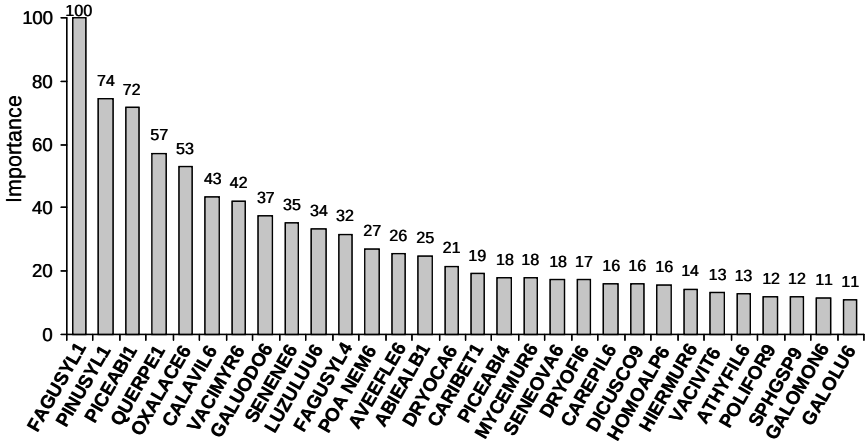


Fig. 5. Importance of species for vegetation tiers (coverage).

The results for presence-absence and coverage were again very similar. The difference in classification results for these variants is < 10% for all vegetation tiers. Vegetation tiers and natural pine forests habitats were very well classified thanks to their species structure. Most often there is an overlap with the neighboring vegetation tier. The best classified edaphic categories are natural pine forests habitats (91%) and vegetation tiers 1 and 9. Vegetation tier 9, however, contains very few samples (n = 53). Pine forests are also the best defined edaphic categories, since no other vegetation tiers overlap with them to a major extent. The least well-classified stage is vegetation tier 3, followed by (in classification through species presence) VT 7 and 4. The biggest overlap persists in VT 7 → 8 (31% for presence, 26% for coverage), as shown in Fig. 6.

presence-absence			Predicted values									
Number of plots	Correct classification (%)	VT	0	1	2	3	4	5	6	7	8	9
1507	91	0	91	1	1	1			1			5
1759	81	1	2	81	14	1				1		
2342	70	2	4	18	70	7	1					
3577	55	3	3	7	14	55	14	5	2			
2151	58	4	3	1	2	11	58	18	5	1		
2852	61	5	3	1		4	10	61	19	1		
1538	66	6	3				1	13	66	13	3	
1599	56	7	2						9	56	31	2
603	74	8							1	14	74	11
53	98	9									2	98

Fig. 6. Classification of vegetation tiers and their overlap due to vegetation. Numbers are percentage of classification, empty cells are zeros.

3.3 Forest site type complexes Classification

The set for the analysis of Forest site type complexes (FSTC) contained 29,854 typological plots, divided into 178 FSTC. The most significant variables for the determination of FSTC were the combinations of important variables from edaphic categories and vegetation tiers (+ natural pine forests), through which Forest site type complexes are clearly defined (Fig. 7). These were the following climate variables: altitude, annual precipitation, average annual temperature and vegetation period duration (amount of days with temperature over 8^{O}C) as well as these soil-related variables: soil type (80), geomorphological unit, humus form of soil and soil texture, the most significant variable of which is the soil type.

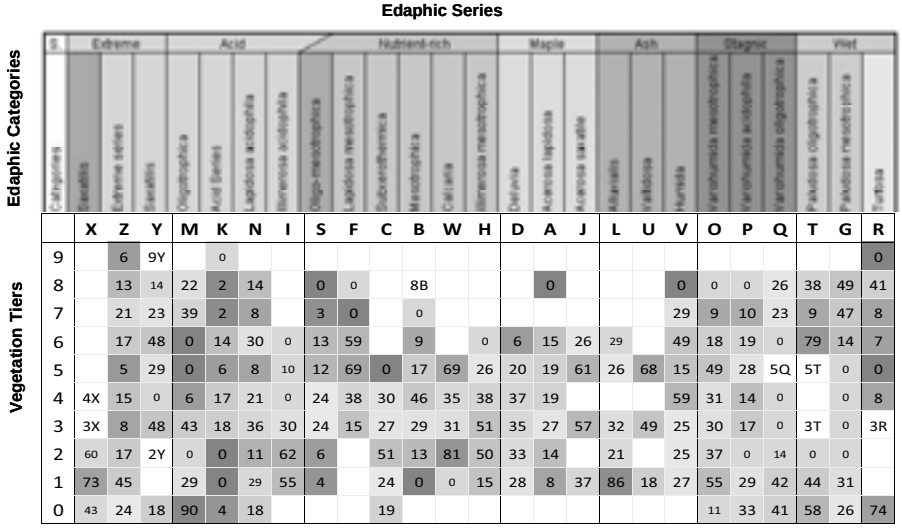


Fig. 7. Classification of Forest site type complexes based on vegetation (presence-absence). Numbers are percentage of correct classification.

From the total of 178 FSTC, the biotic classification discovered 20 FSTC containing ≤ 10 samples for analysis. In another 12 FSTC, the number of samples for analysis was ≤ 20. The classification percentage for FSTC with this low number of samples cannot be considered significant; this data is informative only. Similarly, for an FSTC between 10 and 20 records, any interpretation must be approached carefully. The comparison of FSTC classification is shown only for an FSTC where the number of records is > 20. FSTC with low classification percentages overlap with very similar FSTC (in terms of vegetation or soil characteristics).

4 Conclusions

The preparation and assessment of the Database of Czech Forest Site Classification System demonstrated the strengths and weaknesses of the characteristics of the classification system of forest site typology. Following data evaluation, system modifications can be designed and implemented, and thus improved. An important aspect of possible adjustments is the preservation of basic characteristics of the CFSCS, which is a practical utility for forest management needs. The monitoring of forest site research plot is an important and integral part of forest site typology. Without an extensive database featuring the state of forest ecosystems, the Czech forest site classification system cannot be administered in an objective manner and no models of implications of the climate change on forests in the Czech Republic can be prepared. The analyses imply that there are insufficient samples for certain Forest site type complexes, which reduces the utility of the database for analyses. In order to

improve the utility of the database and make future analyses more accurate, data must be collected on monitoring areas, so that all units (Forest site type complexes) are described with min. 20 records. FSTC with a low percentage of correctly classified samples (this applies to almost one half of all assessable FSTC) are mostly similar FSTC (e.g. FSTC with neighboring forest vegetation zone and/or in same EC). The results can be therefore used as a proposal to merge certain FSTC. Another important result was the identification of FSTC that were defined only with difficulties. Also, the question is whether criteria other than those in the database shall be involved for decision-making for FSTC classification in the field. The application of the Random Forests method therefore proved to be very efficient when used for large data sets. This was not only because of the accuracy of classification into units, but also thanks to additional informations for the categories, such as outlier samples, determination of a set of parameters suitable for sampling in the field and also rules for clear sample categorization.

Acknowledgments. The research was supported by CETOCOEN (CZ.1.05/2.1.00/ 01.0001) project, granted by the European Union and administered by the Ministry of Education, Youth and Sports of the Czech Republic (MEYS), by MEYS (MSMT0021622412).

References

1. Forest Management Institute Brandýs nad Labem, http:// www.uhul.cz
2. Plíva, K.: Typological system of Forest Management Institute. Forest Management Institute Brandýs nad Labem, Brandýs nad Labem (1971)
3. Randuška, D., Vorel, J., Plíva, K.: Fytocenológia a lesnícka typológia. (Phytocenology and Forest Typology) Príroda, Bratislava (1986)
4. Viewegh, J.: Classification of forest plant communities, Czech University of Life Sciences Prague, Prague (2003)
5. Průša, E.:, Growing Forests on Typological Foundations, Kostelec nad Černými lesy (2001)
6. Hastie, T., Tibshirani, R., Friedman, J. H.: The elements of statistical learning : data mining, inference, and prediction. 2nd ed. Springer, New York (2009)
7. Breiman, L.: Random forests. Machine Learning 45, 5-32 (2001)
8. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Chapman and Hall, New York (1984)
9. Klaschka, J., Kotrč, E.: Klasifikační a regresní lesy. Classification and Regression Forests, Proceedings from conference ROBUST (2004)

Biodiversity: from Genetics to Geography, from Mathematics to Management

Computational Biology Students' Abstracts

Metaheuristic Optimization Methods for Magnetic Resonance Image Registration

Petr Dluhoš

Faculty of Science, Masaryk University, Brno

Abstract. The aim of this paper is to present the common global optimization methods used in medical image registration and to propose an implementation of one of them – the genetic algorithm. The genetic algorithm for 3-D affine registration is implemented in MATLAB in package Statistical Parametric Mapping version 8 (SPM8). The efficiency of this algorithm is tested on magnetic resonance images of human brain and compared with the conventional algorithm for affine registration used in SPM8. Experimental results show that the genetic algorithm gets better performance for highly misregistred images where the initial solution is far from the optimal one and gets comparably good results as the conventional algorithm in the case of slightly misregistred images.

Keywords: MRI, linear registration, affine registration, optimization, metaheuristic methods, genetic algorithm

1 Introduction

Registration is one of the basic methods for integration of image data. It is used as a crucial step in many tasks resulting in image comparison, finding specific structures or gathering new information about displayed object. The quality of results of tasks using medical image analysis, such as image segmentation, construction of anatomical or functional atlases, localization of pathologic tissues or designing of surgeries and evaluating their impact highly depend on quality of prior image registration. This is the reason for a big effort to enhance the quality of automatic image registration methods.

There exist reliable methods for finding parameters of linear image registration when a sufficiently good solution is known which can serve as a starting point for the registration. This coarse solution can be provided by an expert in relevant area. However, with increasing tendency of automatization of image processing methods arises also the demand for automatic run of registration algorithm without the need for human intervention. In some cases, such as real-time registration during an operation, it is the only possible solution. These are the situations suitable for using global registration methods which can avoid suboptimal solutions and find the desirable solution without any initial knowledge.

1.1 Metaheuristic optimization methods

Metaheuristic optimization methods are global non-gradient methods. Optimal solution is searched by an iterative process and it is usually worked with whole population of solutions instead of only one. Many of these methods use random variables in the process which leads to a non-deterministic behavior. However, practical results indicate that metaheuristic methods can be very efficient for solving some classes of problems.

There are four main metaheuristic optimization methods which are used for magnetic resonance (MR) image registration: Tabu search which uses memory for storing recently visited states to avoid local optima (this states are called tabu) [1][2], simulated annealing which was inspired by annealing in metallurgy, a technique for creating crystals. It is based on gradual cooling of the material to give it an opportunity to stabilize in the state with the lowest energy. This corresponds to a decreasing chance for the algorithm to accept a state with better functional values [3][4]. Third method used for optimization in medical image registration is particle swarm method [5]. In this method, a population of particles moves in the search space of the optimization problem, each particle representing one solution. Particles change their velocities according to the quality of the represented solution and the solutions of the neighbors. Last group of metaheuristic algorithms are the generic or evolutionary algorithms [6]. These methods are inspired by the evolutionary processes in nature where the pressure of the environment and competition leads to adaptation of the population and general improvement of the average fitness of its individuals. In the algorithm solutions represent the individuals whose fitness is derived from the functional values of the optimized function. New solutions are made from the old ones by using three basic operations – mutation, crossover and selection.

2 Methods

It was chosen an algorithm for 12-parametic affine registration of MR brain images from the MATLAB package SPM8 and the optimization procedure of this algorithm was replaced by an implementation of a genetic algorithm. This new algorithm works with a population of 500 individuals, each of them representing one solution (coded as twelve real numbers – triplets of parameters for translation, rotation, shear and scale). Selection was done by roulette wheel selection algorithm – probability of being selected to the next generation is proportional to the order of the individual among the others according to their fitness. Mutation was implemented as a small chance that each parameter would be randomly changed with respect to specific boundaries. Crossover was realized by exchanging some subintervals of two 12-tuples representing two individuals.

2.1 Experiments

Performance of the newly implemented genetic algorithm was tested on a set of manually misregistred MR images (T1-weighted, 512x512x160 voxels) and the

results of the rigid registration were compared to the original algorithm from SPM8. Less exhaustive experiments were done for more general affine registration and intersubject registration. Detailed results of the experiments can be found in the section Results.

3 Results

The testing images were divided into several groups – three groups for the intrasubject rigid registration according to the number of nonzero parameters in searched transformation (out of six parameters of rigid transformation): one parameter, two parameters, complex rigid registration (six parameters), and two groups for nonrigid registration: one group for general intrasubject affine registration and one for intersubject affine registration. So for the first three groups, only translation and rotation parameters were used, more parameters were included for the last two groups. Results of the first three groups (with known optimal solution) were classified as success or failure according to the absolute difference between found and best values of affine parameters (success means less then 1.5 mm or less then 0.05 rad) and can be seen in Table 1. Results of the nonrigid registration and intersubject registration were evaluated by mean squared error of differences of intensities of corresponding voxels in both images (Table 2).

Table 1. Results of the rigid registration performed by genetic algorithm (GA) and original algorithm. Testing group = number of nonzero parameters in searched transformation, Number of cases = number of performed experiments, Percent of cases classified as success = percent of successful registrations.

Testing group	Number of cases	Percent of cases classified as success for GA / original algorithm
one parameter	18	100% / 67%
two parameters	9	89% / 67%
six parameters	3	100% / 33%

Table 2. Results of the intrasubject and intersubject affine registration performed by genetic algorithm. Testing group = dividing of experiments to inter- and intra-subject registration, Average MSE = average achieved mean squared error of differences of intensities of corresponding voxels in both images.

Testing group	Average MSE
intrasubject affine registration	0.235
intersubject affine registration	0.586

4 Conclusion

The presented genetic algorithm performed significantly better than the original algorithm in the task of rigid registration while the computation time was slightly longer (minutes). It was able to find the optimal transformation even for highly misregistred images which the original algorithm cannot handle. With the increasing number of parameters of the affine transformation, the time demands increased and the precision got worse. So the ideal usage of the algorithm would be to find a coarse transformation in the case we do not have any good initial solution. This coarse solution could then serve as an initial transformation for some good local algorithm.

The implemented genetic algorithm appeared to be not as good for the general affine transformation probably due to the high dimension of search space. Possible solution and a way for further improvement of the algorithm could be finding a better settings of parameters (population size, mutation and crossover rates, selection rule) or implementing registration in more steps (multiresolution approach).

References

1. Glover F.: Tabu search - Part I. ORSA Journal on Computing. 1, 190--206 (1989)
2. Chelouah R., Siarry P.: Tabu search applied to global optimization. Eur. J. Oper. Res. 123, 256--270 (2000)
3. Kirkpatrick S., Gelatt C.D., Vecchi M.P.: Optimization by simulated annealing. Science. 220, 671--680 (1983)
4. Vanderbilt D., Louie S.G.: A Monte carlo simulated annealing approach to optimization over continuous variables. J. Comput. Phys. 56, 259--271 (1984)
5. Dorigo M. et al., http://www.scholarpedia.org/article/Particle_swarm_optimization
6. Coley S.A.: An introduction to genetic algorithms for scientists and engineers. Singapure: World Scientific (1999)

Stochastic Modelling of Mortality of Patients with Acute Heart Failure

Eva Jakubcová

Faculty of Science, Masaryk University, Brno

Abstract. Acute heart failure (AHF) is a disease with complicated etiology, difficult diagnosis and high mortality. The treatment is demanding and economically very costly. AHF together with chronic heart failure is disease that we consider to be the epidemic of 21st century. Due to these facts case studies of acute heart failure are often being gathered in registries throughout the world. The aim of this study was to analyze hospital mortality and long-term survival on real data from the Czech registry AHEAD (Acute HEArt failure Database) and to create a review of AHF registries in the literature and compare their results with the AHEAD.

Keywords: Acute heart failure, logistic regression, mortality, odds ratio, AHEAD, survival analysis.

1 Introduction

Acute heart failure is a sudden incurred disorder of cardiac function (or its sudden setback) when the heart is unable to pump enough blood. The consequence of this disorder is congestion of blood in the lungs and others organs and a lack of oxygenated blood supply to organs.

According to the severity of symptoms, AHF is divided into 6 types: right HF, heart failure with a high output, mild HF, pulmonary edema, hypertensive HF, and cardiogenic shock.

AHF has a high mortality and affects increased amount of people in consequence of aging population and successful treatment of others diseases. It is the worldwide problem and in Czech Republic it is the registry AHEAD, which collects data about patients with AHF and also served as a data source for this study. Data of primohospitalizations from the 4153 patients from eight cardiocenters with Cath Lab facilities and information about long-term survival was used.

2 Methods

Basic patient characteristics, risk factors of in-hospital mortality and long-term survival were evaluated. Review of AHF registries was created for the purpose of comparison with registry AHEAD.

2.1 In-hospital mortality

In-hospital mortality was assessed using logistic regression whose coefficients are estimated by method of maximum likelihood and odds ratio was used to interpretation. At first individual variables were evaluated by univariate analysis, missing values were eliminated from the analysis. Statistically significant variables were used as a basis for multivariate analysis. Prior to multivariate analysis variables with many missing values and redundant variables were excluded. The remaining variables were evaluated by backward stepwise elimination method of logistic regression - separately for patients with cardiogenic shock and for patients without this syndrome.

2.2 Survival analysis

Survival analysis was computed by method of Kaplan-Meier and was stratified according to etiologies and syndromes. For comparison among individual subgroups log rank test was used. Landmark survival analysis from admission, and 30 day after admission was computed.

3 Results

3.1 Risk factors of in-hospital mortality

The greatest risk for patients with cardiogenic shock are aortic stenosis with odds ratios 3.8 (1.466, 9.692) and acute renal failure where the chances of death is nearly three times higher than in patients without this comorbidity.

In patients without cardiogenic shock acute renal failure belongs to the most risk factor again with odds ratios 7.5 (4.057, 13.715), increased levels of C-reactive protein in the blood above 10 mg/L - these patients have up to 5 times higher chance of death in comparison with patients with lower levels of this protein in the blood, low systolic blood pressure is also risk factor.

3.2 Long-term survival

From the difference of the survival curves stratified according to basic syndromes and evaluated from the first admission and after 30 days of admission we could conclude that the most patients dying during hospitalization is due to cardiogenic shock. If they survive the first 30 days, significant relationship to the syndromes cannot be seen anymore.

3.3 Review of registries

Registries collect data on the demographics of patients, the diagnosis and treatment and follow-up, then the data are evaluated statistically and can provide valuable information.
Following studies were compared with the AHEAD registry:
American Studies: ADHERE [1], OPTIMIZE-HF [2]
European studies: EHFS II [3], EFICA [4], FINN-AKVA [5], studies in Zurich and Helsinki [6]
Asian Studies: ATTEND [7], Thai ADHERE [8]

3.4 Comparison of registries with database AHEAD

AHEAD registry is not significantly different from others in terms of the basic description of the patients. But of all the registries that show the value of in-hospital mortality the AHEAD has the highest value (12.7%) while others reached a maximum of 8%. Probably this value is very influenced by varying structure of syndromes in patients of individual registers, especially the ratio of patients with cardiogenic shock which is the main cause of in-hospital mortality.

High age, low systolic blood pressure and increased levels of creatinine in blood were the most often mentioned in the registries from the risk factors of mortality. All these factors also have been identified as the risk factors in the registry AHEAD.

Some results may be influenced by the ethnicity of the population. In the OPTIMIZE-HF study was found that black patients have on average higher blood pressure and the occurrence of AHF predominate in them in women.

4 Conclusions

Logistic regression and survival analysis methods are important procedure in the statistical evaluation of mortality and survival not only in patients with the acute heart failure. During the study register AHEAD was analyzed (descriptive analysis and analysis of mortality factors) and the review of registers AHF was created.

In-hospital mortality (12.7%) is higher than in others registries. This value is probably affected by the proportion of patients with cardiogenic shock.

Among the most risk factors of patients with cardiogenic shock belong aortic stenosis and acute renal failure, in patients without this syndrome it is especially acute renal failure, increased level of C-reactive protein in blood and low systolic blood pressure. These factors, except for aortic stenosis occur in others registries.

The results of short-term and long-term survival were different, especially in AHF syndromes (cardiogenic shock) which have an impact mainly on in-hospital mortality.

Register AHEAD is an important AHF registry, in terms of number of patients it is the 3rd largest registry within others, but it is necessary to be careful during interpretation by reason that it may not be fully representative for Czech Republic because the analysis is limited to data from large centers.

References

1. Adams, K. F., Fonarow, G. C., Emerman, C. L., et al.: Characteristics and outcomes of patients hospitalized for heart failure in the United States: Rationale, design, and preliminary observations from the first 100,000 cases in the Acute Decompensated Heart Failure National Registry (ADHERE). American Heart Journal, vol. 149, pp. 209-216 (2005)
2. Gheorghiage, M., Abraham, W. T., Albert, N. M., et al.: Systolic Blood Pressure at Admission, Clinical Characteristics, and Outcomes in Patients Hospitalized With Acute Heart Failure. JAMA, vol. 296, pp. 2217-2226 (2006)
3. Neieminem, M. S., Brutsaert, D., Dickstein, K., et al.: EuroHeart Failure Survey II (EHFS II): a survey on hospitalized acute heart failure patients: description of population. European Heart Journal, vol. 27, pp. 2725-2736 (2006)
4. Zannad, F., Mebazza, A., Juilliére, Y., et al.: Clinical profile, contemporary management and one-year mortality in patients with severe acute heart failure syndromes: The EFICA study. European Journal of Heart Failure, vol. 8, pp. 697-705 (2005)
5. Siirila–Waris, K., Lassus, J., Melin, J., et al.: Characteristics, outcomes, and predictors of 1-year mortality in patients hospitalized for acute heart failure. European Heart Journal, vol. 27, pp. 3011-3017 (2006)
6. Rudigera, A., Harjolab, V. P., Müllera, A., et al.: Acute heart failure: Clinical presentation, one-year mortality and prognostic factors. European Journal of Heart Failure, vol. 7, pp. 662-670 (2005)
7. Sato, N., Kajimoto, K., Asai, K.,et al.: Acute decompensated heart failure syndromes (ATTEND) registry. A prospective observational multicenter cohort study: Rationale, design, and preliminary data. American Heart Journal, vol. 159, pp. 949-955.e1 (2010)
8. Laothavorn, P., Hengrussamee, K., Kanjanavanit, R., et al.: Thai Acute Decompensated Heart Failure Registry (Thai ADHERE). CVD Prevention and Control, vol. 5, pp. 89-95 (2010)

QT Interval Detection in Electrocardiogram Signal

Jitka Jirčíková[1], supervisor: Ing. Pavel Jurák, CSc.[2]

[1] Faculty of Science, Masaryk University, Brno
[2] Institute of Scientific Instruments, Academy of Sciences of the Czech Republic, Brno

Abstract: This study describes methods of T-wave end detection in electrocardiogram (ECG) signals. The most widely used method of T-wave end detection is compared with three other methods of T-wave end detection and one method of T-wave maximum detection. Differences between these four methods of T-wave end detection were statistical tested. The method of T-wave maximum detection was tested and verified on real data sets.

Keywords: T-wave, electrical activity of the heart, electrocardiography, electrocardiogram (ECG), RR interval, QT interval, T-wave detection, ECG filtration.

1 Introduction

This study considers methods of T-wave end detection in electrocardiogram (ECG) signals. There is a detailed description of a procedure of QT interval measurement. The most widely used method of T-wave end detection uses the interleaving of a straight line along the descending part of the T-wave and its intersection with the isoelectric line (standard method). The QT interval is the phase of depolarization and repolarization of ventricles. Automatic measurement of QT intervals can uncover serious life-threatening genetic diseases or poor physiological conditions resulting from certain drugs.

The differences between the four methods of T-wave end detection were tested and discussed. The final part of this study concerns a method of T-wave maximum detection. This was tested and compared with the standard method.

2 Methods

Data for testing the four methods was taken from 12 healthy volunteers. This data was measured at St. Anne's University Hospital in Brno. Lead II of a 12-lead of stress measurement of ECG was used. Stress is specific cycling in this situation. The four methods compared were the method of the first minimum, the local minimum method, the method of derivation and the standard method mentioned previously. The special program ScopeWin QT from the Institute of Scientific Instruments of the Academy of Sciences of the Czech Republic was used for the purposes of detection. The MATLAB system was used for statistical testing. The ECG signal

was pre-processed by a low-pass filter of 0.8 Hz, a high-pass filter of 48 Hz and a floating window (20 points width) and was subsequently segmented.

The First Minimum Method. This method identifies the T-wave end as the point of the first minimum from the left limit in the detection area.

The Local Minimum Method. This method works by analogy. It identifies the T-wave end as the point of the global minimum inside the detection area.

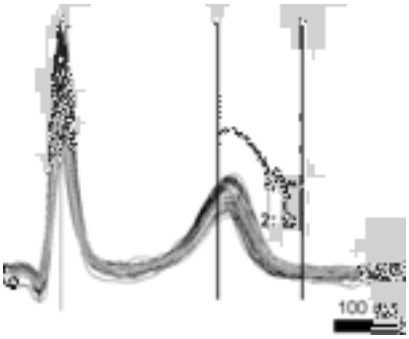


Fig. 1. Methods of T-wave end detection – the First Minimum (2:) and the Local Minimum (1:)

The Standard Method. This uses the interleaving of a straight line along the descending part of the T-wave and its intersection with the isoelectric line for marking the T-wave end.

The Method of Derivation. This method tells us that the T-wave end is the point of, for example, a 20 % fall of the descending part of the T-wave. The derivative signal is used for detection.

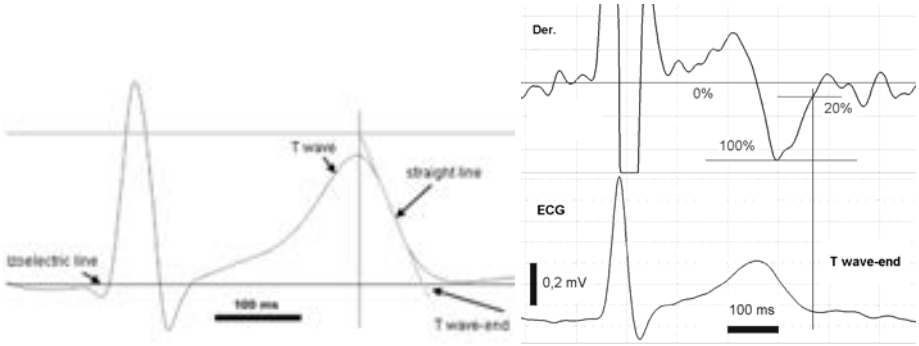


Fig. 2. An algorithm of the Standard Method (*left*) and the Method of Derivation (*right*)

All these methods require the presence of a detectable minimum in the T-wave end area for their algorithm. These methods provide incorrect results if this condition is not met. This is the reason for developing a new method that detects the T-wave maximum.

T-wave maximum detection was performed using the MATLAB system. Data from ECG measurements of mental stress (counting) in lead V3 and data from measurement of physical stress (cycling) in lead V4 was used for testing this method. Data was taken from one chosen volunteer. In the first case there is a T-wave end without a good minimum, while in the second case there is a good profile but also a lot of noise and heart-rate variability. Good results using the standard method were expected in the second case.

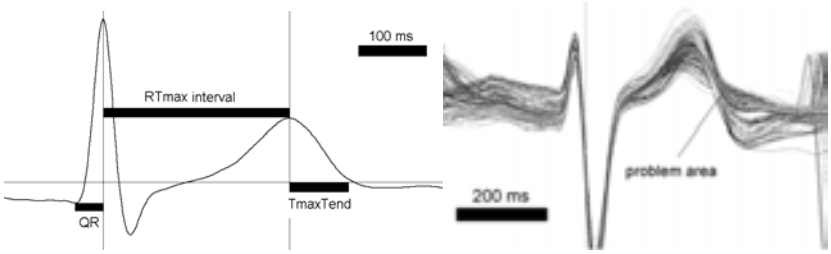


Fig. 3. Construction of QT interval for T-wave maximum detection (*left*); T-wave end without minimum in mental stress measurement (*right*)

3 Results

3.1 A comparison of the methods

The graph below shows the average length of QT intervals. This was detected by four different methods from data from 12 healthy volunteers. We can see that the standard method measures shorter QT intervals than the other methods.

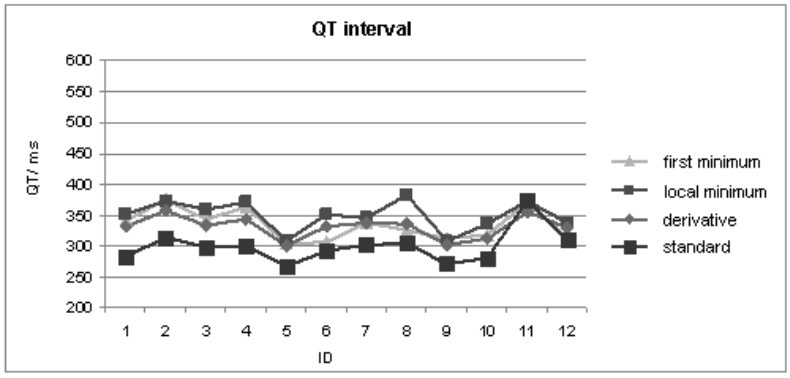


Fig. 4. The average values of QT intervals from 12 people using 4 methods

The statistical significance of the differences in QT intervals was tested by MATLAB. A non-parametric sign test with Bonferroni's correction was used because of violation of expectations (Gaussian distribution and data independence). We proved a significant difference between the derivative method and the method of local minimum and a difference between the standard method and all other methods at a significance level of 0.05/ 6 = 0.0083.

3.2 T-wave maximum detection

T-wave maximum detection was performed on data from mental stress and physical stress, both from one chosen subject. In the first case there is a T-wave end without a good minimum, and in the second case there is a good profile but also a lot of noise and heart-rate variability. Good results were expected of the standard method in the second case.

If heart rate increases (RR interval is shorter), the QT interval must also be shorter. Prolongation of the QT interval is not expected at this time. This is evidence of the failure of the standard method. We can see this in Figure 5, left panel. The right panel shows the second situation – physical stress. There is correct behaviour of both curves.

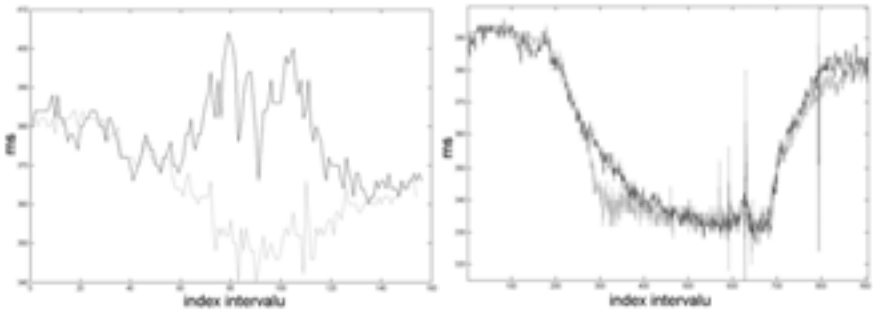


Fig. 5. Length of QT intervals measured by the standard method in ScopeWin QT (*black*) and by the method of T-wave maximum detection in MATLAB (grey). The left panel is taken from mental stress measurement and the right panel from physical stress measurement.

The different detection at the time of the rapid onset of stress in the second measurement was surprising to us. This area was studied, and we find a changing length of the T_{max}-T_{end} interval during the signal. This finding is good material for further study.

4 Conclusion

We do not know which method detects the correct length of the QT interval. The practical significance of this must be consulted with doctors. If we compare methods of QT interval detection (method of the first minimum, method of the local

minimum, method of derivation, and method using interleaving of a straight line along the descending part of the T-wave and its intersection with the isoelectric line – standard method), we verify the expectation that the standard method detects a shorter QT interval. At a significance level of 0.05 we also proved a significant difference between the derivative method and the method of local minimum.

The right trend in QT interval detection was confirmed in the last part of this study for a new method of T-wave maximum detection. This method was very good for detection in a signal without a minimum, and was also comparable in quality with the standard method (in ScopeWin QT) in an ordinary signal.

Bayesian Coalescence Analysis of Rabies Virus in China, USA and Europe

Jiří Moravec

Faculty of Science, Masaryk University, Brno

Abstract. Rabies virus causes approximately 55 000 deaths, mainly in Africa and Asia. Reconstruction of population history would expose spreading of Rabies and its response on vaccination programs and other means of animal control. We have reconstructed population history of 453 RABV sequences coding N-protein from China, USA and Europe using BEAST program for Bayesian coalescence analysis. Furthermore, bayesian skyline plot was constructed using Tracer. Bayesian coalescence analysis suggest introduction of RABV into USA at the very beginning of European colonization. Results have also shown increase of European RABV population in 1920s with increased trend in 1960s, probably due to adaptation and introduction of new hosts. Increase of dynamics in China and USA is probably caused by intensive sequencing, rather than real changes in population structure, although vaccination-caused peaks are presented.

Keywords: rabies, RABV, Bayesian coalescence analysis, population history, bayesian skyline plot

1 Introduction

Rabies virus causes approximately 55 000 deaths, mainly in Africa and Asia. Although some developed countries have been proclaimed rabies-free due to intensive vaccination and animal control programs, risk of rabies has not been completely eliminated and due to animal trade and natural human and animal migration, rabies has managed to return and spread.

Coalescence method is another way, beside incidence, to gain information of population dynamic of rabies. Bayesian coalescence method [1] was successfully used in this area, from analyzing of local populations of dog RABV in Middle and West Africa [2] or population of RABV in China [3].

Aim of this work is to reconstruct population history of three distinctive populations with different historic development, populations of China, Europe and USA, visualize them by Bayesian skyline plot [4] and evaluate their development with connection to rabies-related events (such as vaccination, animal control program or introduction of new hosts).

2 Materials and Methods

Complete sequences coding N-protein were acquired from GenBank database. Only non-vaccine strains with acquirable date and place of isolation were used, resulting in total 453 sequences, 17 from Europe, 178 from China and 258 from USA. Sequences were aligned using ClustalW [5] version 1.82 through ClustalW-XXL[1] web service.

Bayesian coalescence analysis was handled by BEAST[2] [6] version 1.6.1. Three demographic models (constant population size, exponential grow and Bayesian skyline) and three molecular clock models (strict molecular clock, relaxed log-normal and relaxed exponential molecular clock) were used along with GTR+Γ substitution model. Number of generations in the MCMC run was set to 10^8, recording every 10^4 sample. One run for combination of each dataset, demographic model and molecular clock model, 27 runs in total were executed on MetaCentrum3 computational network.

Corresponding bayes factors of each model were then compared using Tracer4 to find best model for every dataset and Bayesian skyline plot was constructed for each dataset and model of molecular clock to reveal demographic changes.

3 Results and Discussion

Bayes factors have shown that no demographic model was better than others. Constant molecular clock for Chinese and American dataset were slightly better than either relaxed molecular clocks. For European dataset, strict molecular clock model for constructing Bayesian skyline plot was selected as the most simple hypothesis.

Bayesian skyline plot has revealed possible introduction of European RABV into the USA in the very beginning of its colonization. Additionally to that, rapid decrease of population size in the 1940s correlate with start of national vaccination programs [7], accompanied by a number of similar rapid decreases and increases. This could be artifact caused by intensive sequencing and obtaining of new strains.

1 http://www.ch.embnet.org/software/ClustalW-XXL.html
2 http://beast.bio.ed.ac.uk/Main_Page
3 http://metavo.metacentrum.cz/
4 http://beast.bio.ed.ac.uk/Tracer

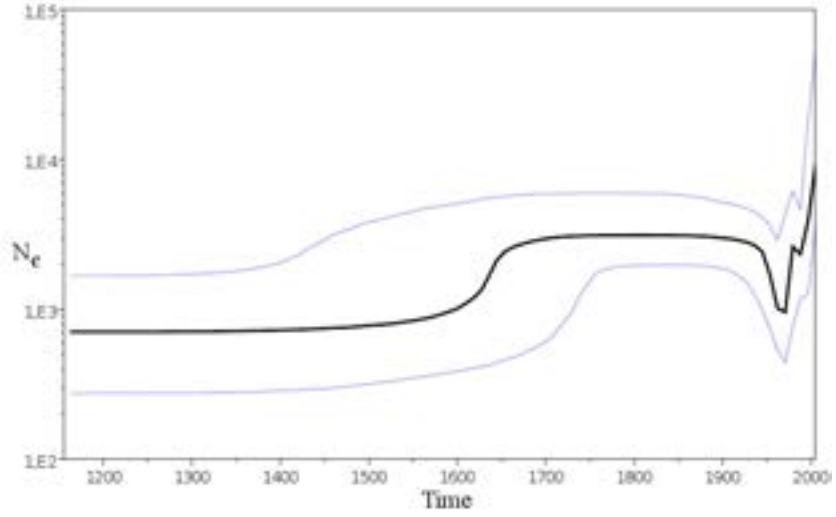


Fig. 1. Bayesian skyline plot for dataset from USA. The x axis is in calendar years and the y axis show effective population. The thick line show median of highest posterior density while the gray lines its 95% interval.

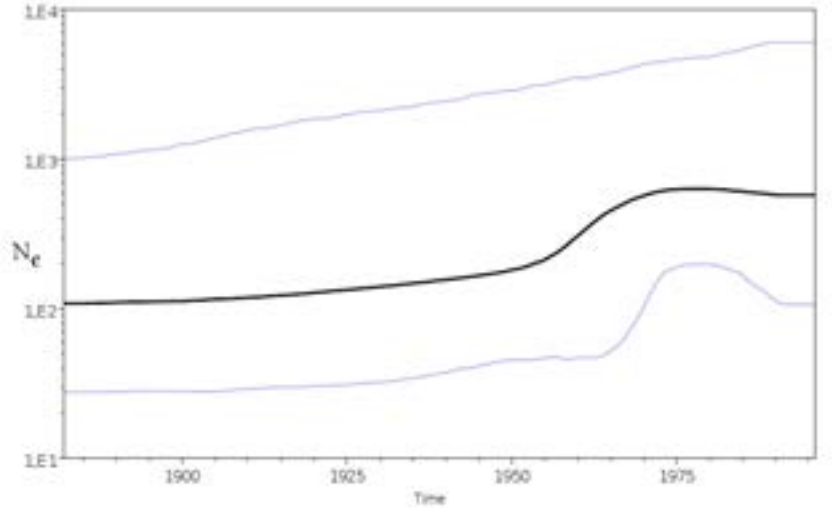


Fig. 2. Bayesian skyline plot for dataset from Europe. The x axis is in calendar years and the y axis show effective population. The thick line show median of highest posterior density while the gray lines its 95% interval.

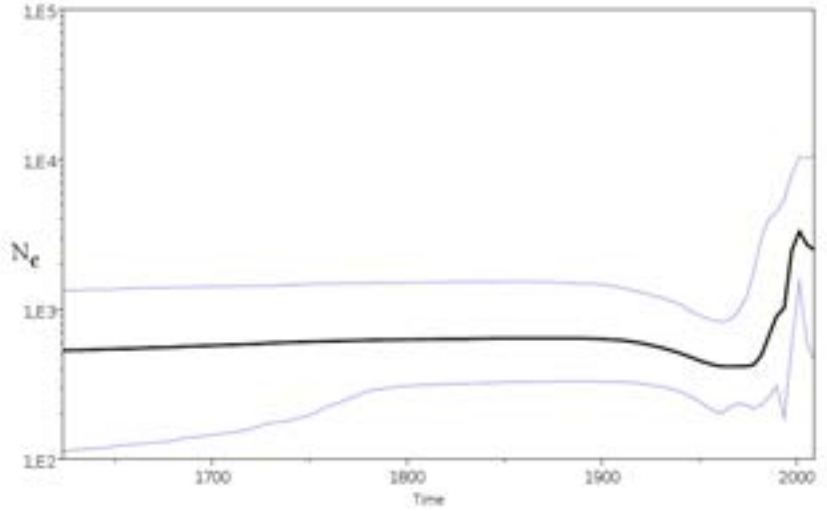


Fig. 3. Bayesian skyline plot for dataset from China. The x axis is in calendar years and the y axis show effective population. The thick line show median of highest posterior density while the gray lines its 95% interval.

For European dataset, Bayesian skyline plot showed increase of RABV population from the 1920s with increasing trend in the 1960s with its peak around the 1970s. There is well-documented adaptation of RABV to foxes on Russian-Polish borders in 1940s, that has spread through Europe, reaching France in 1968 and Northern Italy in early 1980s [7]. Furthermore, in the 1920s Raccoon dog (*Nyctereutes procyonoides*) was introducted in Russia. It has then spread into most of Northern Europe [7]. No sudden change due to vaccination is probably caused by high level of environment cultivation in Europe and low level of rabies there.

Bayesian skyline plot for the Chinese dataset shows decrease of RABV population from the 1920s to the 1960s. We didn't find any rabies-related event, that would explain this decline. Furthermore, there is same situation as in USA, rapid increase and decrease of Chinese RABV population from 1980s. Although there is evidence of increased level of RABV in China, we don't find doubling of its population realistic and ascribe it to similar phenomenon we have encountered with USA dataset, intensive sequencing. In closer examination, we can identify particular waves of vaccination, notably in 2001, when new type of vaccine was introduced [8], although in that year, incidence of rabies have increased due to higher cost of new vaccine and its lower availability. Another way to explain the increase of RABV in China is its relation to turbulent economic changes, intensive animal market and town immigration.

4 Conclusion

We have reconstructed population history of RABV in Europe, China and USA and connected most of changes in estimated population with historic rabies-related events. However, the problem of distortion caused by intensive sequencing persists. There is also area for future depth research of Chinese population, given the turbulent changes in recent past and large amount of data.

Acknowledgments

The access to the MetaCentrum computing facilities provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" LM2010005 funded by the Ministry of Education, Youth, and Sports of the Czech Republic is highly appreciated.

References

1. Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., Solomon, W. Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data. Genetics., 161, 3, 1307-1320 (2002)
2. Talbi, C., Holmese, E. C., Benedictis, P., Faye, O., Nakoune, E., Gamatie, D., Diarra, A., Elmamy, B. O., Sow, A., Adjogoua E. V., Sangare, O., Dundon, W. G., Caupa, I., Sall, A. A., Bourhy, H. Evolutionary history and dynamics of dog rabies virus in western and central Africa. J Gen Virol. 90, 4, 783-791 (2009)
3. Ming, P., Yan, J., Rayner, S., Meng, S., Xu, G., Tang, Q., Wu, J., Luo, J., Yang, X. A history estimate and evolutionary analysis of rabies virus variants in China. J Gen Virol. 91, 3, 759-764 (2010)
4. Drummond, A. J., Rambaut, A., Shapiro, B., Pybus, O. G. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. Molecular Biology and Evolution. 22, 5, 1185-1192 (2005)
5. Thompson, J., Higgins, D., Gibson, T. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 22, 4673-4680 (1994)
6. Drummond, A., Raumbaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology. 7, 1, 214 (2007)
7. Finnegan, C. J., Brookes, S. M., Johnson, N., Smith, J., Mansfield, K. L., Keene, V. L., McElhinney, L. M., Fooks, A. R. Rabies in North America and Europe. J R Soc Med. 95, 1, 9-13 (2002)
8. Song, M., Tang, Q., Wang, D. M., Mo. Z. J., Guo, S. H., Li, H., Tao, X. Y., Rupprecht, C., Feng, Z. J., Liang, G. D. Epidemiological investigations of human rabies in China. BMC Infectious Diseases. 9, 1, 210 (2009)

Parametric Survival Models

Michal Uher

Faculty of Science, Masaryk University, Brno

Abstract. In survival analysis, non-parametric methods are widely used and they have become very popular because of plainness of their application. However there may be settings in which non-parametric methods are not available and that is why a parametric approach is also very important. Parametric methods have some advantages over non-parametric ones but an important assumption has to be verified – the probability distribution of survival time has known parametric form. Aim of this contribution is to summarize issues of basic parametric regression models used in survival analysis (e.g. the most important distributions of survival time, AFT and proportional hazard form of model) and discuss an adequacy of used models within individual cancer diagnosis.

Keywords: survival analysis, censoring, parametric regression model, accelerated failure model, proportional hazard model, Akaike information criterion

1 Introduction

Survival analysis includes important methods used in modern medicine and clinical research especially in oncology. These methods provide information about survival time of patients which is essential when assessing a quality of health care for cancer patients. This assessment is necessary because it is often a disease with high mortality and an expensive treatment.

A characteristic of survival data is censoring [2] which occurs when the value of survival time is only partially known.

2 Methods and materials

In my work, methods of survival analysis are presented as non-parametric or fully parametric. Then the best known non-parametric estimators of survival function are mentioned: the Kaplan-Meier estimator [3] and life-tables based estimator [4]. The main part of my thesis deals with parametric regression models in survival analysis [2]. These models allow to quantify an effect of explanatory variables on survival of patients and can also be used to predict future values of survival. Regression models based on exponential, Weibull, log-normal and generalized gamma distribution are presented closely. Then two forms of models are mentioned: accelerated failure time

form (AFT) which assumes that the effect of covariate is to multiply survival time by constant; and proportional hazard form (PH) which assumes that the effect of covariate is to multiply hazard by constant. Comparison of Akaike information criterions [1] is presented as a way to judge which of used models fits the data best.

3 Results

Methods presented in previous section were applied to data from the Czech National Cancer Registry (CNCR) including cancer patients with diagnosis C18-C21, C25, C34, C50 and C61 diagnosed in 1989-2008. Individual parametric estimates of survival function were compared with Kaplan-Meier estimator. These comparisons show that data of patients with prostate cancer (C61) can be modeled with Weibull distribution. In other cases, occurrence of proportion of cured patients causes complication with fitting parametric model. Thus only data of patients with cancer diagnosed in fourth stage (where proportion of cured patients is assumed to be zero) can also be modeled with parametric models. All these applications were done by using software R and STATA.

4 Conclusion

In this work, I presented basic non-parametric and parametric methods used in survival analysis and applied them to the real data from CNCR. Besides C61 and cancer diagnosed in fourth stage, some problems with fitting parametric models have occurred. I assume this might be solved by using another advanced models.

References

1. Anderson, D.R.: The Model Based Inference in the Life Sciences, Springer (2008)
2. Hosmer, D.W., Lemeshow, S., May, S.: Applied survival analysis: regression modeling of time-to-event data. Wiley-Interscience, 2nd ed. Hoboken, N.J. (2008)
3. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. J. Amer. Statist. Assn. 53, 457-481 (1958)
4. Marubini, E., Valsecchi, M.G.: Analysing survival data from clinical trials and observational studies. Institute of medical statistics and biometry, University of Milan, Italy (2004)