Species Distribution Models

Ecological Applications for Management of Biodiversity

Simon Hallstan
Faculty of Natural Resources and Agricultural Sciences
Department of Aquatic Sciences and Assessment
Uppsala

Licentiate Thesis
Swedish University of Agricultural Sciences
Uppsala 2011

Cover: Boreal landscape (photo: S. Hallstan).

ISBN 978-91-576-9037-1

Print: SLU Service/Repro, Uppsala 2011

Species Distribution Models. Ecological Applications for Management of Biodiversity

Abstract

Species distribution models are a group of methods often used to estimate consequences of global change, to assess ecological status and for other ecological applications. The main idea behind species distribution models is that the geographical distributions of species can, to a large part, be explained by environmental factors and that species distributions therefore can be predicted in time or space. For robust and reliable applications, models need to be based on sound ecological principles, predictions need to be as accurate as possible, and model uncertainties need to be understood.

Two approaches are available for modelling entire species communities: (1) each species can be modelled individually and independently of other species or (2) community information can be incorporated into the models. The first study in this thesis compares these two modelling approaches for predicting phytoplankton assemblages in lakes. The results showed that predictive accuracy was higher when species were modelled individually. The results also showed that phytoplankton can be used for model-based assessment of ecological status. This finding is important because phytoplankton is required for assessing the ecological status of European water bodies according to the European Water Framework Directive.

Dispersal barriers in the landscape or limited dispersal ability of species might be a reason for species being absent from suitable habitats, and these factors might therefore affect model accuracy. The second study in this thesis examines the influence of dispersal and the spatial configuration of ecosystems on prediction accuracy of benthic invertebrate and phytoplankton distribution and assemblage composition. The results showed only a minor influence of spatial configuration and no effect of flight ability of invertebrates on model accuracy. However, the models used may partly account for dispersal constraints, since dispersal-related factors, such as lake surface area, are included as predictor variables. The result also showed that composition of littoral invertebrate assemblages was easier to predict at sites located in well-connected lake systems, possibly because the relatively unstable littoral zone necessitates a need for species to re-colonize disturbed habitats from source populations.

Keywords: biodiversity, environmental assessment, species distribution models

Author's address: Simon Hallstan, SLU, Department of Aquatic Sciences and Assessment, P.O. Box 7050, 750 07 Uppsala, Sweden
E-mail: simon.hallstan@slu.se

Contents

List of Publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I Hallstan, S., Johnson, R.K., Willén, E. and Grandin, U. (2011). Comparison of classification-then-modelling and species-by-species modelling for predicting lake phytoplankton assemblages (*manuscript*).

II Hallstan, S., Johnson, R.K. and Sandin, L. (2011). Small effects of dispersal-related factors on species distribution model accuracy (*manuscript*).

The contribution of Simon Hallstan (SH) to the papers included in this thesis was as follows:

I SH was involved in the design of the study, and solely responsible for model development and statistical analyses. Writing and interpretation were mainly done by SH, but in conjunction with his co-authors.

II SH was responsible for the design of the study, and solely responsible for model development and statistical analyses. Writing and interpretation were mainly done by SH, but in conjunction with his co-authors.

Abbreviations

AUC	Area under curve
E	Expected taxa richness
O	Observed taxa richness
RIVPACS	River InVertebrate Prediction and Classification System
SD	Standard deviation
SDM	Species distribution model
TPP	True positive proportion

1 Introduction

The spatial distribution of plants, animals and other organisms has always been a central interest for ecologists. In fact, in one of the most well-known textbooks on ecology, Krebs (1972) states that:

> ...ecology is the scientific study of the interactions that determine the distribution and abundance of organisms. We are interested in where organisms are found, how many occur there, and why...

In the latter parts of the 19th and early parts of the 20th century, the distribution of organisms were studied within zoology, limnology, palaeontology and other fields of ecology, but interactions between scientists from the different scientific disciplines were at best infrequent. It was not until the end of the 20th century that biogeography was recognised as a separate discipline (Brown, 2004). Key works such as MacArthur and Wilson's island biogeography (MacArthur & Wilson, 1967) created interest in several of the traditional fields and contributed to the emergence of biogeography as a discipline of its own.

Biogeographical theories, in other words theories of how species are distributed and what controls these distributions, were early on suggested to be useful for ecological applications, such as design of nature reserves (e.g. Diamond, 1975). A number of applications have been and are currently being developed with the aim of predicting the distribution of species in time or space, and are accordingly referred to as species distribution models (SDMs; also called niche models, suitable habitat models and bioclimatic envelope models). The main idea behind SDMs is that it is possible to relate environmental factors to species distributions using statistical functions or algorithms. The SDMs can then be applied to new environmental data to interpolate or extrapolate the species distribution in space or time, for

applications such as risk assessments of biological invasions or estimations of the effect of global change on biodiversity. Already in the 1920s, Johnstone (1924; cited in Mack, 1996) used correlations between the distribution of an invasive cactus species and environmental variables to predict the potential spread of the cactus species in Australia. Another early example, also from Australia, is the prediction of the distribution of crop species by Nix and co-workers (Nix *et al.* 1977; cited in Guisan & Thuiller, 2005). However, it was not until the end of the 20th century and the beginning of the 21st century that SDMs generated a wider interest.

During the last decades, developments in computer technology making it easier to handle large amounts of data, and infrastructure making the data easily accessible through the Internet, have facilitated the increasing interest in SDMs. Species records from, for example, museum collections, surveys and monitoring programs have been digitalized and made freely available via the Internet by networks such as the *Global biodiversity information facility* (GBIF; www.gbif.org), which currently contains 267 374 767[1] records of species occurrences, and the Swedish *Species Gateway* (Swedish *Artportalen*; www.artportalen.se), which currently contains 27 514 942[1] recorded species observations. Environmental data such as climate (e.g. Worldclim, www.worldclim.org) and land use (e.g. Corine, Commission of the European Communities, 1995) have also become easily available in digital form.

1.1 Global change and environmental management

Another reason for the increased interest in SDMs is the ongoing impairment of global biodiversity and consequent threats to ecosystem services (Harrison *et al.*, 2010) and human livelihood and wellbeing, which have prompted the need for new tools for management of biodiversity (Olden *et al.*, 2006). Despite ambitious initiatives such as the *2010 Biodiversity Target* (Secretariat of the Convention on Biological Diversity, 2010), which stated that the loss of biodiversity should have come to an end by 2010, the situation is still severe. It is estimated that the current extinction rate is 100 to 1 000 times the historical average (Pimm *et al.*, 1995). Freshwater habitats seem to be more sensitive than terrestrial (Dudgeon *et al.*, 2006). Ricciardi and Rasmussen (1999) estimated that the extinction rate of North American freshwater fauna is five times higher than the extinction rate of terrestrial fauna. One reason for the high extinction rates could be the problems

[1] 2011-02-02

associated with management: rivers and lakes are not only affected by direct pressures, but also by anthropogenic disturbances in the entire catchment (Lake *et al.*, 2007).

Freshwater ecosystems are vital to human livelihood and wellbeing – water is used for everything from drinking and irrigation to power production and recreational activities. Although freshwater ecosystems cover only 0.01% of the Earth's surface (Balian *et al.*, 2008), the value of ecosystem services provided by freshwater ecosystems has been estimated to make up 20% of the total value provided by all ecosystem services on Earth (Costanza *et al.*, 1997). Furthermore, freshwater ecosystems contain a disproportionate amount of the Earth's biodiversity: 9.5% of the planet's described animal species and 35% of the described vertebrate species are found in freshwater habitats (Balian *et al.*, 2008).

For sound management of biological resources, it is essential to monitor and assess the status of ecosystems throughout the world. Most assessment methods are based on the concept of comparing the current state with a reference state, which could range from *pristine* to *best attainable* (Stoddard *et al.*, 2006). Biological records of pre-industrial or pre-agricultural conditions are often not available and therefore the biological reference state must be derived from other sources, for example palaeoecological studies or unimpacted ecosystems believed to be similar to the studied ecosystems. In aquatic environmental assessment, modelling approaches are commonly used to assess pre-anthropogenic states. The general idea of models for reference communities is that it is possible to predict the species composition that a site would have if it was unaffected by human-induced pressures. This is done by creating a model for species composition using sites defined as references and environmental variables unaffected by human activities. The obtained "natural" or "reference" species composition can then be compared with the actual (observed) species composition. One of the most used approaches to model reference state species composition is the *River InVertebrate Prediction and Classification System* (RIVPACS), which was first developed in the United Kingdom during the late 1970s and 1980s (Moss *et al.*, 1987). Similar approaches have also been developed for other countries, taxa and habitats, for example invertebrates in lakes in Sweden (Johnson, 2003) and invertebrates, fish (Carlisle *et al.*, 2008) and benthic diatoms (Cao *et al.*, 2007) in streams in the United States.

For species distribution models to be successful tools in applied ecology it is desirable to further improve models and analyze sources of model error. Much research has been focused on development and evaluation of new statistical methods (see e.g. Elith *et al.*, 2006). However, several studies have

shown that the differences in model performance often are greater between species than between modelling methods (e.g. Guisan *et al.*, 2007; Thuiller, 2003). It is therefore important to know if there are systematic errors in model performance caused by differences in ecological properties of species and ecosystems. There are indications that species characteristics such as range size (Newbold *et al.*, 2009), niche width (Kadmon *et al.*, 2003) and prevalence (Franklin *et al.*, 2009) affect model accuracy, although the effect varies between studies. For example, Marmion *et al.* (2009a) found that models in their study were less accurate for more common species and argued that species with low prevalence often have narrow biological niches and therefore are easier to model. By contrast, Seoane *et al.* (2005) argued that model accuracy, found to be low for relatively rare species, was explained as an artefact of the model used; the regression trees in their study were better calibrated with more observations ("presences"). Another species characteristic that could potentially affect model accuracy is dispersal ability. For example, even though the environment at a site is suitable for a species, it is not certain that the species has been able to colonize the habitat. The relative importance of the environment and dispersal for species distributions and assemblage composition has been debated in ecology (e.g. Mazaris *et al.*, 2010).

Another issue relevant to SDMs, which have been extensively debated within ecology, is whether species respond to changes in environment as individuals or if discrete species assemblages (communities) exists (McIntosh, 1995). These two different views become important when models are developed for predictions of entire community compositions. Either all species in the communities can be modelled individually (species-by-species modelling) or community information could be incorporated into the models (e.g. by the so called classification-then-modelling approach). Advocates of species-by-species modelling argue that species respond individually to environmental gradients and that discrete communities do not exist in nature, whereas advocates of classification-then-modelling argue that species interactions can be included and improve model performance (Olden *et al.*, 2006; Ferrier *et al.*, 2002).

2 Objectives

The main objective of this thesis was to assess possible improvements of species distribution models and methods for ecological assessment by answering the following questions:

- Can phytoplankton reference-state composition be modelled? (paper I)
- Does the species-by-species modelling approach improve model accuracy compared to the community-based RIVPACS approach? (paper I)
- Is the distribution of species with poor dispersal ability less well predicted than the distribution of species with relatively high dispersal ability? (paper II)
- Is the species composition of ecosystems that are relatively difficult for organisms to colonize, such as small ecosystems and isolated ecosystems, more difficult to model than larger ecosystems and ecosystems with high connectivity? (paper II)

3 Methods

3.1 Data

The data used in these two studies were collected within the Swedish environmental monitoring programs.

3.1.1 Water chemistry

Surface water (approximately 0.5 m depth) samples were collected at a mid-lake station in each lake. Water samples were collected with a Plexiglas sampler and kept cool during transport to the laboratory. Samples were analysed for variables indicative of acidity (e.g. pH and SO_4^{2-} concentration), nutrients (e.g. total nitrogen, total phosphorus), water colour (absorbance of filtered water measured at 420 nm), as well as total organic carbon and concentrations of major base cations. All water analyses were done at the Department of Aquatic Sciences and Assessment according to international (ISO) or European (EN) standards when available (Wilander *et al.*, 2003).

3.1.2 Phytoplankton

Phytoplankton was sampled in August of each year by taking a water sample from the epilimnion (0–4 m) using a Plexiglas tube sampler (3 cm diameter). August phytoplankton samples are often used for environmental assessment in Sweden (Anonymous 2000), as the August assemblages are regarded as relatively stable (Reynolds 1988) and are comparable between the south and the north of the country. In lakes with a surface area >1 km^2, a single mid-lake site was used for sampling. In lakes with a surface area <1 km^2, five random epilimnetic water samples were taken and mixed to form a composite sample from which a subsample was taken. The samples were preserved with acid Lugol's iodine solution (Throndsen 1978). Phytoplankton counts were made using an inverted light microscope and

the modified Utermöhl technique commonly used in the Nordic countries (Olrik 1989). Taxa were identified to the lowest taxonomic unit possible (usually species).

3.1.3 Invertebrates

Benthic invertebrates were collected from two habitats in late autumn (October–November) each year. Littoral samples were collected using standardized kick sampling (European Committee for Standardization 1994) with a hand net (0.5 mm mesh size). A composite sample consisting of five standardized kick samples (20 s duration, 0.25 m × 1 m long at about 0.5 m depth, total area 1.25 m^2) was taken from hard-bottom, vegetation-free sites in each lake. Profundal samples consisted of five replicate Ekman samples (~247 cm^2) taken in the deepest area of the lake. Invertebrate samples were preserved with 95% ethanol in the field (final concentration approximately 70%). The samples were processed by sorting with 10 times magnification in the laboratory. Invertebrates were identified and counted using dissecting and light microscopes. Organisms were identified to the lowest taxonomic unit possible, generally to species level. Sorting and taxonomical identification were done according to quality control and assurance protocols.

3.1.4 Geography and climate

Besides water chemistry, environmental data were acquired from digital maps (altitude, lake surface area, catchment area) and from the *Swedish Institute for Climate and Hydrology* (temperature, precipitation and runoff). Land use data were acquired from *Corine* (Commission of the European Communities, 1995).

3.2 Methods used for paper I

In the first study, the performance of two modelling approaches for predicting reference-state phytoplankton community composition was compared. The two modelling approaches differ in that one utilizes community information (a RIVPACS-type model), whereas the other fits a model for each species independently of other species (species-by-species models).

For both modelling approaches the *random forest* modelling method was used (Prasad *et al.*, 2006; Liaw & Wiener, 2002; Breiman, 2001). Random forest is an improved version of a method called *classification trees* (Breiman *et*

al., 1984). A classification tree model is constructed by identifying and splitting the single best predictive variable into two subsets of observations. Subsets are split again and splitting is repeated until each "node" contains only one category (i.e. presences or absences) or a pre-selected minimum number of sites. The *random forest* method combines several trees using a technique called bootstrap aggregating (bagging). Each tree is constructed independently of the previous tree, using a random subset of predictor variables for each split and a bootstrap sample of sites for each tree. The prediction is decided by majority vote. The random forest modelling technique has been successfully used in several ecological studies (e.g. Cutler *et al.*, 2007).

For the community model the *River InVertebrate Predictive and Classification System* (RIVPACS) method was used (Moss *et al.* 1987). The RIVPACS approach is based on classification of sites using species composition. Traditionally *multiple discriminant analysis* has been used to derive equations that predict the site-group membership, but other statistical methods can be used. In this study, random forest was used. The modelling functions predict the probability for each site to belong to one of the pre-defined biological groups. The probability for individual species is then calculated as the prevalence of the species in each group, weighted by the probability of the site belonging to the corresponding group.

For the species-by-species modelling approach random forest was used, with the same setting and the same environmental variables as in the RIVPACS model. One random forest model was fitted for each species.

A null model, which assumes that species distribution and assemblages composition is independent of environmental gradients, was also used. According to the null model, the occurrence probability for each species equals the prevalence of the focal species in the calibration data. In other words, if a species is found in half of the calibration lakes there is a 50% probability that the species exists in any given lake, regardless of the lake's nutrient levels, pH, climate and other environmental factors.

When the probability of species occurrences have been estimated, different indices can be calculated. The most common index derived from the RIVPACS model is taxonomic completeness, also called the observed to expected (O/E) ratio. Simply put, this is an index of how many of the reference-state species that are recorded as being present at a site. The expected richness (E) is the sum of all occurrence probabilities above a selected threshold and the observed richness (O) is the number of species that both have an occurrence probability above the threshold and are found at the site. The ratio of O to E then indicates how much the species

composition of a site deviates from the reference-state species composition. A value near 1.0 indicates that the studied site is close to its "reference state" and values considerably lower than 1.0 indicate that the site is impaired. Taxonomic completeness (O/E) was calculated for the references lakes using all possible thresholds from 0 to 0.50 in intervals of 0.01 (51 different thresholds).

A dissimilarity measure similar to the Bray-Curtis index, called BC (Van Sickle, 2008), was calculated as an alternative to O/E. BC compares the predicted probabilities for occurrence and the observed assemblage taxon-specifically, and includes all taxa regardless of occurrence probability. The BC index ranges from 0 to 1, and high values indicates a large difference between observed species assemblage and predicted reference-state species assemblage.

The models were also compared on a species basis using AUC, the Area Under the Receiver-Operator Curve (ROC), a metric commonly used in SDM studies because it is insensitive to species prevalence and because it does not require a threshold value to convert probabilities to presence-absence (Manel *et al.*, 2001). The value of AUC ranges from 0 to 1, with high values indicating more accurate models.

SDMs can produce two kinds of erroneous predictions: (1) a species can be predicted as absent when it is actually present and (2) a species can be predicted as present when it is actually absent. Because the O/E index could be misleading due to omissions of the former, the proportion of species correctly predicted as present was calculated (true positive proposition; TPP)

The models were also applied to lakes affected by either eutrophication or acidification and the BC index was calculated to determine the effect of these stressors on phytoplankton assemblages.

The methods are described in detail in the original paper (paper I).

3.3 Methods used for paper II

In the second study, the difference in model accuracy between species with different dispersal ability and between ecosystems with different spatial configuration was examined. The occurrence probabilities for 164 littoral and 44 profundal invertebrate taxa and 129 phytoplankton taxa in 105 lakes were predicted. The modelling was performed using the modelling platform BIOMOD (Thuiller & Lafourcade, 2010) in the R-software (R Development Core Team, 2010). Six different modelling methods were used, all of which have been shown to perform well for ecological predictions (e.g. Elith *et al.*, 2006). The methods represent three classes of modelling methods, namely

classification (*classification tree analysis* and *mixture discriminant analysis*), regression (*generalized linear models* and *multivariate adaptive regression splines*) and machine learning (*generalized boosting models* and *random forests*). For each species, the occurrence probabilities produced by the six models were averaged; *mean occurrence probability* is a consensus method that has been shown to be robust in other ecological studies (Marmion *et al.*, 2009b). Cross-validation was used to obtain AUC values independent of calibration for all species distribution predictions (taxa-AUC) and for all assemblage predictions (lake-AUC).

In order to analyze differences in model performance between species with different dispersal abilities, the invertebrate taxa were divided into three classes according to their flight ability (no, low, or high). Odonata were the largest (flying) organisms in the dataset and were classified as having high flight ability; other insects were classified as having low flight ability, and wingless invertebrates were classified as having no flight ability. *T*-tests were used to test differences in taxa-AUC between the three groups. For taxa recorded from both littoral and profundal habitats the highest AUC was used.

Geographical descriptors of a lake's spatial configuration were used to test if accuracy in assemblage prediction differed between ecosystems. Altitude, distance to neighbour lakes, total surface area of other lakes and streams in the catchment and number of lakes within buffer zones (100 m, 200 m, 500 m, 1 000 m, 5 000 m and 10 000 m) were used as indicators for connectivity, and lake surface area and catchment size were used as indicators of ecosystem size. Spearman correlations between lake-AUC and connectivity and ecosystem size were calculated to test the importance of spatial configuration on model performance.

The methods are described in detail in the original paper (paper II).

4 Results & Discussion

4.1 Comparison of RIVPACS modelling and species-by-species modelling

Predicted taxa assemblage compositions were similar between the RIVPACS model and the random forest models. The species-by-species models were more accurate according to both the dissimilarity measure BC, the proportion of true positives and AUC; the latter metric is the most frequently used measure for determining model accuracy of species distribution models (Figure 1). Contrastingly, according to taxonomic completeness (O/E), which is often used for RIVPACS modelling, the RIVPACS model was the most accurate approach for most probability thresholds tested, although no significant difference in mean O/E_{00} was found.

The community-types (cluster groups) used in RIVPACS could be seen as unrealistic, since it is unclear whether such groups exist in nature (see review by McIntosh, 1995). However, the use of community-types has also been argued as one of the strengths of RIVPACS models, since information on species interactions (i.e. co-occurrences and exclusion) is indirectly included in the groups (Johnson, 2000). Furthermore, the groups used by RIVPACS are not used as discrete entities, but instead the probabilities of membership to all groups are used to calculate occurrence probabilities. Whether or not community information improves models might depend on the species being modelled. For example, Leathwick *et al.* (2006) found small differences in mean accuracy between models (*multivariate adaptive regression splines*) used with and without a community addition, but showed that for rare species the community model performed better. However, the results from this study showed that species-by-species models were more accurate for predictions of rare species, compared to the RIVPACS model.

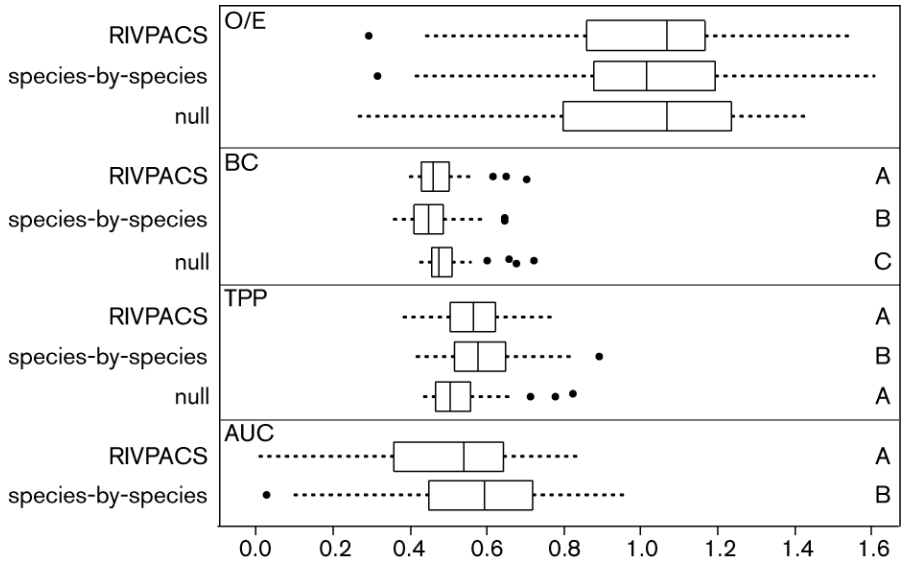

Figure 1. Boxplot of model accuracy for assemblage composition in the reference lakes, measured as O/E_{00}, BC and TPP, and AUC for species distributions in the reference lakes. O/E_{00} is a measure of taxonomic completeness (accurate models have O/E close to 1.0), BC is a dissimilarity measure between modelled and observed assemblage (accurate models have low BC values), true positive proportion (TPP) is the proportion of species found in a lake predicted to be present (accurate models have high TPP), and AUC, area under receiver operator curve, is a measure of the models' ability to separate presences from absences (accurate models have high AUC). Significantly different means between models according to ANOVA with post-hoc Tukey-Kramer test are indicated by different letters.

When the models were applied to lakes judged to be impaired according to water chemistry and land use variables, all models, including the null model, resulted in significantly lower BC values for both acidified and eutrophied lakes, compared to reference lakes. Hence, both modelling approaches could be used in lake management to gauge the effects of human-induced stress on boreal lake ecosystems.

4.2 Effect of dispersal-related factors on species distribution model accuracy

According to the postulated hypothesis, model accuracy would be lower for species with low or no flying ability compared to organisms with high

dispersal ability. However, no effect of invertebrate flight ability on model accuracy was found (Figure 2). Furthermore, no effect of ecosystem size on model accuracy was detected (Table 1), although it is well known within ecology that larger ecosystems are easier to colonize (MacArthur & Wilson, 1967). Probably, most taxa are not dispersal-limited and have had time to colonize all suitable habitats within the studied area. Another reason for the small effect of dispersal, ecosystem size and connectivity is that the models developed here partly account for dispersal and colonization, because factors indicating colonization potential, such as lake surface area, were included as predictor variables. In contrast to flight ability and ecosystem size, some effects of connectivity on the accuracy of predictions were found for phytoplankton assemblages and littoral invertebrate assemblages, which were less well predicted for isolated lakes (Table 1). Possibly, the difference observed in model performance between the two invertebrate habitats could be because the environment of littoral habitats are relatively unstable (e.g. due to abrasion of ice and wave action) compared to profundal habitats. Littoral invertebrate composition might therefore be more dependent on re-colonization and connectivity. The effect of connectivity on phytoplankton predictions are more difficult to explain; microorganisms are often believed to have "ubiquitous dispersal" (Finlay, 2002), although this is debated (see e.g. Foissner *et al.*, 2003), and phytoplankton dispersal mechanisms are poorly studied (Kristiansen, 1996).

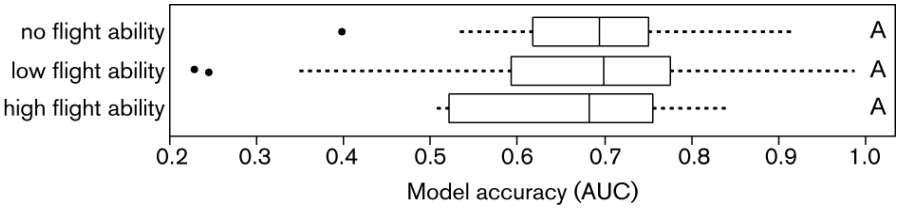

Figure 2. Boxplot of model accuracy per taxa grouped according to their flight ability. Lettering indicates significantly different means (*t*-test, $p \leq 0.05$).

Table 1. *Correlations (Spearman ρ) between per lake prediction accuracy (Area under receiver operator curve, AUC) and indicators of ecosystem size and spatial configuration. Distance neighbour is the distance to the closest and 2nd, 5th and 10th closest neighbour lake and N lakes is the number of lakes with 100, 200, 500, 1 000 and 5 000 meter buffer zones round the lake shoreline. Variables that were significant are shown in bold: * p≤0.05, ** p≤0.01, ***p≤0.001*

	Littoral	Profundal	Phytoplankton
lake surface area	-0.19	-0.05	0.15
catchment area	-0.13	0.05	0.14
water (excl. study lake)	-0.07	0.08	0.13
altitude	-0.05	-0.14	**0.53*** **
distance neighbour 1	**-0.27****	-0.01	-0.04
distance neighbour 2	**-0.23***	0.04	-0.02
distance neighbour 5	**-0.21***	0	-0.08
distance neighbour 10	**-0.21***	-0.01	-0.16
N lakes 100 m	0.01	-0.03	-0.09
N lakes 200 m	-0.01	-0.02	-0.02
N lakes 500 m	0.11	-0.02	0.08
N lakes 1000 m	0.08	-0.11	0.1
N lakes 5000 m	0.18	0.03	**0.26****
N lakes 10000 m	0.14	0.05	**0.32*** **

5 Conclusions

Biodiversity throughout the world is under immense threat (e.g. Sala *et al.*, 2000; Pimm *et al.*, 1995). Models for predicting species distributions and assemblage composition, based on a long history of ecological research, could be valuable tools for management and policy development. Already today, species distribution models are used to predict consequences of global change on biodiversity (e.g. Thuiller *et al.*, 2005), to assess the ecological status of ecosystems (e.g. Hargett *et al.*, 2007) and to assess the risk of biological invasions (e.g. Peterson & Vieglais, 2001). However, improvements in modelling techniques and further insight into the ecological theory behind the distribution of species could facilitate additional applications and improve the accuracy of model predictions. In this thesis, I conclude that:

1. phytoplankton can be used in modelled-based assessment of ecological status;
2. the species-by-species model approach can produce more accurate models than the classification-then-modelling (community) approach;
3. AUC and BC are better metrics for model evaluation than O/E;
4. neither dispersal ability nor ecosystem size affects model accuracy;
5. connectivity affects model accuracy of littoral invertebrates and phytoplankton assemblages.

These findings add to current knowledge of species distribution models and uncertainties associated with model predictions, and can hopefully aid the further development of valuable tools for biodiversity management. For example, models for environmental assessment could be improved if species are modelled independently, although community-models could possibly be improved if multiple organism groups (e.g. phytoplankton and zooplankton)

are included in the model development. Furthermore, to account for dispersal-related errors, model developers should, if possible, include predictor variables such as ecosystem size and altitude in their models. However, the questions studied in this thesis should be further examined with other taxa and ecosystems, and complementary methods, such as genetics and experiments, should be used to gain a better understanding of the ecological processes that influence the distribution of organisms.

6 Sammanfattning på svenska

Människans nyttjande av jordens resurser innebär stora påfrestningar för den biologiska mångfalden. Många arter har försvunnit, och ekosystemtjänster som är nödvändiga för människors försörjning och välbefinnande är hotade. För förvaltning av den biologiska mångfalden krävs att ekologiska teorier kan användas för att skapa praktiska verktyg. Vilka faktorer som styr djurs, växters och andra organismers utbredning har länge studerats inom ekologin. Dessa kunskaper ligger till grund för en samling ekologiska verktyg som kallas för artutbredningsmodeller eller habitatmodeller. Grundidén med artutbredningsmodeller är att organismers utbredning till stor del styrs av miljöfaktorer, och att det därför är möjligt att med hjälp av matematiska och statistiska modeller prediktera organismers utbredning inom områden där man känner till miljöförhållandena. Artutbredningsmodeller används bland annat för att förutsäga möjliga konsekvenser av klimatförändringar, för att bedöma risken för spridning av främmande arter och för att bedöma tillståndet i miljön (ekologisk status).

Målet med detta avhandlingsarbete var att förbättra kunskaperna om artutbredningsmodeller genom att jämföra olika metoder för att modellera artsammansättningar (studie 1) och undersöka hur arters och ekosystems egenskaper påverkar modellers noggrannhet (studie 2).

Vid bedömning av ekologisk status med hjälp av modeller brukar vanligtvis hela organismgruppers artsammansättning modelleras. Det finns då två alternativ: antingen kan varje arts utbredning modelleras enskilt, eller så kan information om artsamhället inkluderas i modellerna. I den första studien i avhandlingen undersöktes modeller för växtplankton i sjöar, och resultaten visade att artspecifika modeller var mer noggranna än modeller som byggde på artsamhället sammansättning. Resultaten visade också att växtplankton kan användas för modellbaserade bedömningar av tillståndet i sjöar, vilket är viktigt eftersom EU:s vattendirektiv föreskriver att

växtplankton ska användas för klassning av ekologisk status av unionens sjöar.

En mängd olika tekniker (statistiska metoder, algoritmer) har utvecklats för att prediktera artutbredningar, men flera studier har visat att skillnaden i noggrannhet är större mellan olika arter än mellan olika modeller. För att modellprediktioner ska vara tillförlitliga är det viktigt att förstå vad skillnaderna mellan arter beror på. En av anledningarna till denna variation kan vara att en del arter inte lyckas sprida sig till alla lämpliga habitat på grund av begränsad spridningsförmåga.

Resultaten från den andra studien i avhandlingen, i vilken modeller för bottenfauna (makroevertebrater) och växtplankton i sjöar undersöktes, visade att utbredningen för evertebrater utan vingar eller med sämre flygförmåga inte var svårare att prediktera än utbredningen för evertebrater med god flygförmåga. Det var inte heller svårare att prediktera artsammansättningen i sjöar som antogs vara svårare att kolonisera, det vill säga mindre sjöar och sjöar med mindre avrinningsområden. Resultaten visade dock att artsammansättningen av evertebrater i litoralzonen var svårare att prediktera i sjöar som har relativt få andra sjöar i närheten. Det kan bero på att miljön i litoralzonen är instabil och att populationerna där lättare slås ut, och därför är beroende av återkoloniseringar från närliggande populationer.

Resultaten från den här avhandlingen bidrar med kunskaper om artutbredningsmodeller och modelprediktioners osäkerhet, och kan förhoppningsvis användas för att förbättra artutbredningsmodeller som verktyg för förvaltning av den biologiska mångfalden.

Acknowledgement

Att slutföra den här avhandlingen hade inte varit möjligt utan ett stort antal personer – tack till alla som gett hjälp och stöd under de senaste åren.

Tack speciellt till min handledare, Richard Johnson, som har läst och kommenterat manus, diskuterat, och gett stöd för mitt deltagande i kurser och konferenser. Tack också till mina biträdande handledare, Ulf Grandin och Leonard Sandin som båda har varit till stor hjälp med att få klart avhandlingen. Tack också till Ulf för att jag har fått jobba med många spännande projekt vid sidan av avhandlingen – vandrarmusslor, sjöhjortron, fladdermöss och MVA-kursen.

Jag vill också passa på att tacka Eva Willén och Daniel Larson för att de introducerade mig till institutionen som handledare för mitt examensarbete. Eva har också bidragit med värdefulla kunskaper om växtplankton.

Att sitta vid datorn och analysera data och skriva manus hade varit mindre roligt utan mina trevliga rumskompisar Ina Bloch, Maria Khalili, Atlasi Daneshvar och Ana Villa. Tack för sällskapet de senaste åren. Detsamma gäller såklart alla doktorander på institutionen.

Jag vill också tacka alla som gett mig chansen att se naturen på riktigt genom att be mig om hjälp med fältarbete, i snö och solsken, i jordbruksdiken och på Mälaren. Tack till Jenny Rydh-Stenström, Karin Johansson, Peter Carlson, Emma Göthe, Micke Östlund, Jocke Ahlgren och Anders Gustavsson för trevliga dagar i fält.

Några av de definitiva höjdpunkterna under forskarutbildningen har varit alla resor. Tack till alla som gjort mig sällskap under resorna, ni är för många för att nämna.

Tack också till Britta Lidström, Annika Lundberg, Hasse Eurell och Herman Paz för hjälp med reseräkningar och andra praktiska saker på institutionen.

Bert Karlsson, Anders Danielsson-Stenström och Barbro Sandin har varit hjälpsamma med att ta fram data från databaserna och Jakob Nisell med att ta fram GIS-data.

Arbetet med avhandlingen hade definitivt varit tråkigare utan Simon K, Martin, Jakob L, Jocke, Hampus, Peter, Salar, Stefan, Stephan, Ronald och alla andra innebandyspelare.

Slutligen, tack till familj och vänner utanför den akademiska världen för er hjälp och ert stöd.

References

Anonymous (2000). *Environmental quality criteria. Lakes and watercourses*. Stockholm: Swedish Environmental Protection Agency (Rapport / Naturvårdsverket; 5050). ISBN 91-620-4913-5.

Balian, E.V., Segers, H., Leveque, C. & Martens, K. (2008). The freshwater animal diversity assessment: An overview of the results. *Hydrobiologia* 595, 627-637.

Breiman, L. (2001). Random forests. *Machine Learning* 45(1), 5-32.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). *Classification and regression trees*. New York, NY: Chapman & Hall, Inc. (The Wadsworth statistics/probability series.

Brown, J.H. (2004). Introduction. In: Lomolino, M.V.*, et al.* (Eds.) *Foundations of Biogeography. Classic Papers with Commentaries*. pp. 1-3. Chicago, IL: The University of Chicago Press.

Cao, Y., Hawkins, C.P., Olson, J. & Kosterman, M.A. (2007). Modelling natural environmental gradients improves the accuracy and precision of diatom-based indicators. *Journal of the North American Benthological Society* 26(3), 566-585.

Carlisle, D.M., Hawkins, C.P., Meador, M.R., Potapova, M. & Falcone, J. (2008). Biological assessments of Appalachian streams based on predictive models for fish, macroinvertebrate, and diatom assemblages. *Journal of the North American Benthological Society* 27(1), 16-37.

Commission of the European Communities (1995). *CORINE Land Cover*: Directorate-General Environment, Nuclear Safety and Civil Protection.

Costanza, R., d'Arge, R., de Groot, R., Farber, S., Grasso, M., Hannon, B., Limburg, K., Naeem, S., O'Neill, R.V., Paruelo, J., Raskin, R.G., Sutton, P. & van den Belt, M. (1997). The value of the world's ecosystem services and natural capital. *Ecological Economics* 387, 253-260.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A. & Hess, K.T. (2007). Random forests for classification in ecology. *Ecology* 88(11), 2783-2792.

Diamond, J.M. (1975). The island dilemma lessons of modern bio geographic studies for the design of natural reserves. *Biological Conservation* 7(2), 129-146.

Dudgeon, D., Arthington, A.H., Gessner, M.O., Kawabata, Z.I., Knowler, D.J., Leveque, C., Naiman, R.J., Prieur-Richard, A.H., Soto, D., Stiassny, M.L.J. & Sullivan, C.A.

(2006). Freshwater biodiversity: importance, threats, status and conservation challenges. *Biological Reviews* 81(2), 163-182.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberon, J., Williams, S., Wisz, M.S. & Zimmermann, N.E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29(2), 129-151.

European Committee for Standardization (1994). *Water quality-methods for biological sampling-guidance on handnet sampling of aquatic benthic macro-invertebrates. SS-EN-27-828.* Brussels, Belgium.

Ferrier, S., Drielsma, M., Manion, G. & Watson, G. (2002). Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. Community-level modelling. *Biodiversity and Conservation* 11(12), 2309-2338.

Finlay, B.J. (2002). Global dispersal of free-living microbial eukaryote species. *Science* 296(5570), 1061-1063.

Foissner, W., Struder-Kypke, M., van der Staay, G.W.M., Moon-van der Staay, S.Y. & Hackstein, J.H.P. (2003). Endemic ciliates (Protozoa, Ciliophora) from tank bromeliads (Bromeliaceae): a combined morphological, molecular, and ecological study. *European Journal of Protistology* 39(4), 365-372.

Franklin, J., Wejnert, K.E., Hathaway, S.A., Rochester, C.J. & Fisher, R.N. (2009). Effect of species rarity on the accuracy of species distribution models for reptiles and amphibians in southern California. *Diversity and Distributions* 15(1), 167-177.

Guisan, A. & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters* 8(9), 993-1009.

Guisan, A., Zimmermann, N.E., Elith, J., Graham, C.H., Phillips, S. & Peterson, A.T. (2007). What matters for predicting the occurrences of trees: Techniques, data, or species' characteristics? *Ecological Monographs* 77(4), 615-630.

Hargett, E.G., ZumBerge, J.R., Hawkinsc, C.P. & Olsonc, J.R. (2007). Development of a RIVPACS-type predictive model for bioassessment of wadeable streams in Wyoming. *Ecological Indicators* 7(4), 807-826.

Harrison, P.A., Vandewalle, M., Sykes, M.T., Berry, P.M., Bugter, R., de Bello, F., Feld, C.K., Grandin, U., Harrington, R., Haslett, J.R., Jongman, R.H.G., Luck, G.W., da Silva, P.M., Moora, M., Settele, J., Sousa, J.P. & Zobel, M. (2010). Identifying and prioritising services in European terrestrial and freshwater ecosystems. *Biodiversity and Conservation* 19(10), 2791-2821.

Johnson, R.K. (2000). RIVPACS and alternative statistical modelling techniques - accuracy and soundness of principles. In: Wright, J.F*., et al.* (Eds.) *Assessing the biological quality of fresh waters: RIVPACS and other techniques*. pp. 323-332. Ableside, Cumbria, U.K.: Freshwater Biological Association and Environment Agency.

Johnson, R.K. (2003). Development of a prediction system for lake stony-bottom littoral macro invertebrate communities. *Archiv Fur Hydrobiologie* 158(4), 517-540.

Johnston, T.H. (1924). The relation of climate to the spread of prickly pear. *Trans. R. Soc. South Aust.* 48, 269-295.

Kadmon, R., Farber, O. & Danin, A. (2003). A systematic analysis of factors affecting the performance of climatic envelope models. *Ecological Applications* 13(3), 853-867.

Krebs, C.J. (1972). *Ecology: the experimental analysis of distribution and abundance*. New York: Harper and Row.

Kristiansen, J. (1996). Dispersal of freshwater algae - A review. *Hydrobiologia* 336(1-3), 151-157.

Lake, P.S., Bond, N. & Reich, P. (2007). Linking ecological theory with stream restoration. *Freshwater Biology* 52(4), 597-615.

Leathwick, J.R., Elith, J. & Hastie, T. (2006). Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological Modelling* 199(2), 188-196.

Liaw, A. & Wiener, M. (2002). Classification and Regression by randomForest. *R News* 2(3), 18-22.

MacArthur, R.H. & Wilson, E.O. (1967). *The Theory of Island Biogeography*. Princeton, N.J.: Princeton University Press.

Mack, R.N. (1996). Predicting the identity and fate of plant invaders: Emergent and emerging approaches. *Biological Conservation* 78(1-2), 107-121.

Manel, S., Williams, H.C. & Ormerod, S.J. (2001). Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology* 38(5), 921-931.

Marmion, M., Luoto, M., Heikkinen, R.K. & Thuiller, W. (2009a). The performance of state-of-the-art modelling techniques depends on geographical distribution of species. *Ecological Modelling* 220(24), 3512-3520.

Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R.K. & Thuiller, W. (2009b). Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions* 15(1), 59-69.

Mazaris, A.D., Moustaka-Gouni, M., Michaloudi, E. & Bobori, D.C. (2010). Biogeographical patterns of freshwater micro- and macroorganisms: a comparison between phytoplankton, zooplankton and fish in the eastern Mediterranean. *Journal of Biogeography* 37(7), 1341-1351.

McIntosh, R.P. (1995). Gleasons,H.a. Individualistic Concept and Theory of Animal Communities - a Continuing Controversy. *Biological Reviews of the Cambridge Philosophical Society* 70(2), 317-357.

Moss, D., Furse, M.T., Wright, J.F. & Armitage, P.D. (1987). The Prediction of the Macroinvertebrate Fauna of Unpolluted Running-Water Sites in Great-Britain Using Environmental Data. *Freshwater Biology* 17(1), 41-52.

Newbold, T., Reader, T., Zalat, S., El-Gabbas, A. & Gilbert, F. (2009). Effect of characteristics of butterfly species on the accuracy of distribution models in an arid environment. *Biodiversity and Conservation* 18(13), 3629-3641.

Nix, H., McMahon, J. & Mackenzie, D. (1977). Potential areas of production and the future of pigeon pea and other grain legumes in Australia. In: Wallis, E.S*., et al*. (Eds.) *The*

potential for pigeon pea in Australia. Proceedings of Pigeon Pea (Cajanus cajan (L.) Millsp.) Field Day. pp. 5/1–5/12. Queensland, Australia: University of Queensland.

Olden, J.D., Joy, M.K. & Death, R.G. (2006). Rediscovering the species in community-wide predictive modelling. *Ecological Applications* 16(4), 1449-1460.

Olrik, K.P., Blomqvist, P., Brettum, P., Cronberg, G., & Eloranta, P (1989). *Methods for quantitative assessment of phytoplankton in freshwaters, part I.* Stockholm: Swedish Environmental Protection Agency. ISSN Report 4860.

Peterson, A.T. & Vieglais, D.A. (2001). Predicting species invasions using ecological niche modelling: new approaches from bioinformatics attack a pressing problem. *Bioscience* 51(5), 363-371.

Pimm, S.L., Russell, G.J., Gittleman, J.L. & Brooks, T.M. (1995). The Future of Biodiversity. *Science* 269(5222), 347-350.

Prasad, A.M., Iverson, L.R. & Liaw, A. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems* 9(2), 181-199.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reynolds, C.S. (1988). The concept of ecological succession applied to seasonal periodicity of freshwater phytoplankton. *Verhandlungen der Internationalen Vereinigung für Theoretische und Angewandte Limnologie. Verhandlungen IVTLAP.* 23(2), 683-691.

Ricciardi, A. & Rasmussen, J.B. (1999). Extinction rates of North American freshwater fauna. *Conservation Biology* 13(5), 1220-1222.

Sala, O.E., Chapin, F.S., Armesto, J.J., Berlow, E., Bloomfield, J., Dirzo, R., Huber-Sanwald, E., Huenneke, L.F., Jackson, R.B., Kinzig, A., Leemans, R., Lodge, D.M., Mooney, H.A., Oesterheld, M., Poff, N.L., Sykes, M.T., Walker, B.H., Walker, M. & Wall, D.H. (2000). Biodiversity - Global biodiversity scenarios for the year 2100. *Science* 287(5459), 1770-1774.

Secretariat of the Convention on Biological Diversity (2010). *Global Biodiversity Outlook 3*. Montréal.

Seoane, J., Carrascal, L.M., Alonso, C.L. & Palomino, D. (2005). Species-specific traits associated to prediction errors in bird habitat suitability modelling. *Ecological Modelling* 185(2-4), 299-308.

Stoddard, J.L., Larsen, D.P., Hawkins, C.P., Johnson, R.K. & Norris, R.H. (2006). Setting expectations for the ecological condition of streams: The concept of reference condition. *Ecological Applications* 16(4), 1267-1276.

Thuiller, W. (2003). BIOMOD - optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology* 9(10), 1353-1362.

Thuiller, W. & Lafourcade, B. (2010). BIOMOD: species/climate modelling functions. R package version 1.1-3/r137. http://R-Forge.R-project.org/projects/biomod/.

Thuiller, W., Lavorel, S., Araujo, M.B., Sykes, M.T. & Prentice, I.C. (2005). Climate change threats to plant diversity in Europe. Proceedings of the National Academy of Sciences of the United States of America 102(23), 8245-8250.

Throndsen, J. (1978). Preservation and storage. In: Sournia, A. (Ed.) Phytoplankton manual. Monographs on oceanographic methodology. pp. 70-71 UNESCO.

Van Sickle, J. (2008). An index of compositional dissimilarity between observed and expected assemblages. Journal of the North American Benthological Society 27(2), 227-235.

Wilander, A., Johnson, R.K. & Goedkoop, W. (2003). Riksinventering 2000. En synoptisk studie av vattenkemi och bottenfauna i svenska sjöar och vattendrag. Uppsala (in Swedish): Department of Environmental Assessment; Rapport 2003:1.