Exploring the value of museum data for use in species distribution modeling:

Data limitations and how to tackle them

BENTE STØA

Dissertation presented for the degree of Philosophiae Doctor

Natural History Museum

Faculty of Mathematics and Natural Sciences

University of Oslo

2014

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 1508

ISSN 1501-7710

Cover: Inger Sandved Anfinsen.
Printed in Norway: AIT Oslo AS.

Produced in co-operation with Akademika Publishing.
The thesis is produced by Akademika Publishing merely in connection with the thesis defence. Kindly direct all inquiries regarding the thesis to the copyright holder or the unit which grants the doctorate.

Acknowledgements

This study was supported by the Norwegian Research Council MILJØ2015 grant 183318 to Vladimir Gusarov. A big thank to NFR for financing this PhD.

As I am writing this I am very close to obtaining my PhD and I am forever grateful to all of you who in one way or the other have helped me during the process. Thanks to my supervisors, Vladimir Gusarov, Jogeir N. Stokland and Rune Halvorsen: Vladimir for giving me the opportunity to do a PhD, for good discussions and for always keeping your office door open for me, Jogeir for giving me a kickstart by involving me in your pseudo-absence paper, for nice talks and good laughs, and Rune for being a rock during the whole period and for always keeping an eye on the goal, namely me obtaining my PhD in time.

I would like to give a warm thank to the insect group, Lars Ove, Leif, Karsten, Geir, Hallvard and Eirik (and all the other who have temporarily been working in the insect department) who instantly made me feel at home when I first arrived at the museum. I have had countless cups of coffee, countless discussions and countless laughs with you and I have enjoyed your company a lot.

To all past and present members of the GEco group thanks for inspiration, motivation, scientific input, knowledge-sharing, good seminars, lunches at Hovedgården and for simply being such a great group to be part of. You all mean a lot to me!

A special thanks to Sabrina. You have helped me on innumerous occations with GiS, R, spelling checks or just dragging me out of the office to see something else than my computer screen during lunch time. I know you have been just as busy as I have, but you have still always taken the time to help. That means a lot! More than one time I would have been totally lost without your help. Also thanks for all the good conversations.

Thanks to my former office neighbors at the zoological museum, Judith, Raul and Louis for being such nice roomies.

To Silje and Lovisa who started their PhD at about the same time as me, thanks for nice talks and encouragement along the way!

Thanks to my best friends Vivian, Ingrid, Dina and Hanne for believing in me, for giving me social breaks, laughter and lots of fun!

To my parents; thanks for teaching me the values of seeking knowledge, working hard and appreciating nature, all of which are values that got me where I am today. I would also like to thank you for the invaluable help and support during the years as a single mum PhD student. I would never have made it without you!

Thanks to my parents in law for always being warm, welcoming and unselfish and for “babysitting” Eirik and the girls from time to time so that I have been able to work on my PhD.

To Eirik; colleague, friend, boyfriend and husband (chronologically listed); thanks for being exactly who you are! The first thing you said to me when we were introduced at the museum the 18th of September 2008 was; “If there is anything you need, just come to me!” That is what I have done ever since, also when it comes to helping me with my PhD, and you never let me down. Thank you!

Last but not least, a big thanks to my two girls, Elisa and Sigrid. You are the two most important persons in my life. Thanks for forcing me to think about something other than work, for making me stress down and enjoy life. Thanks to you it has been easy to always stay positive during these five years since I started my PhD.

Contents

1. List of papers

Paper I. Støa, B., Halvorsen, R., Gusarov, V. I. A framework for assessment of sampling bias in presence-only data and its effects on species distribution models. Manuscript.

Paper II. Støa, B., Mazzoni, S., Halvorsen, R. Assessment of sampling bias in presence-only data and its effects on species distribution models: practical examples. Manuscript.

Paper III. Stokland, J.N., Halvorsen, R., Støa, B., 2011. Species distribution modeling - Effect of design and sample size of pseudo-absence observations. Ecological Modeling 222, 1800–1809.

Paper IV. Støa, B., Halvorsen, R., Stokland, J. N., Gusarov, V. I. How much is enough? Amount of presence data and performance of species distribution models. Manuscript.

2. Summary

Museum data is a great resource for species distribution modeling (SDM), and may consequently be of potential value for nature management and conservation purposes. Museum data are, however, not designed for SDM and may be suboptimal for several reasons. This thesis addresses challenges associated with the use of museum data, such as sampling bias (paper I and II), the sample size and the design of pseudo-absence/ background data (paper II and III) and the sample size of presence data (paper IV). The understanding of how these factors affect species distribution models needs to be based on ecological theory. In paper I sampling bias is defined and explained with reference to gradient analytic reasoning and macro-ecological biogeographic theory. A set of presence observations is defined as sampling biased if the frequency distribution of *observed* presence along major environmental complex-gradients deviates from the frequency distribution of the *true* presence of the species. Based on this definition, we provide a framework for assessment of sampling bias in presence-only datasets, typically used for species distribution modeling, by use of frequency-of-presence curves. With access to presence/absence data, sampling bias can be identified by statistical methods. In cases without availability of such data, indications of sampling bias can be obtained by visual inspection of frequency-of-observed-presence (FOP) curves, using theoretical frequency-of-presence (TFP) curves as a reference (see table 1 in paper I for definitions of different frequency-of-presence curves). These reference curves are typically smooth and unimodal (deduced from generalization of empirical frequency-of-presence curves and by taking expert knowledge about the species into account). Shape differences between the two curves may indicate sampling bias in the response data. Four types of sampling bias can be expected: (1) under-sampling, i.e., lower-than-true FOP in intervals along a gradient; (2) over-sampling, i.e., higher-than-true FOP in intervals along a gradient; (3) peripheral sampling gap, i.e., lack of presence observations at one or both of the species' tolerance limits; and (4) stochasticity, i.e., unsystematic variation in sampling effort. It is important to be aware that local minima and maxima on the FOP curve may represent real properties of the species in the study area. The extent to which strong cases for or against sampling bias can be made will depend on the quality of the information that forms the basis for the TFP curve.

We also provide a framework for assessment of effects of sampling bias on species distribution models by comparison between FOP- and predicted relative-frequency-of-presence (PRFP) curves generated from model output. The PRFP curve may deviate from the TFP curve for two reasons; the training dataset may be biased, or the modeling method may not be able to parameterize the FOP curve. With the exception of stochastic variation

in training data, SDM methods cannot be expected to be able to distinguish sampling bias from real properties of the data. The PRFP curves should therefore reproduce the FOP curve with a degree of detail that matches the level of generalization required by the purpose of the study (Halvorsen 2012), regardless of the FOP curve being realistic or not. Datasets that give rise to unrealistic FOP curves because of probable sampling bias should not be used for SDM.

In paper II FOP- and PRFP curves were plotted to explore the extent to which museum data used for SDM contain sampling bias, to address the assumption that background target group (BTG) data with similar bias as presence data may be obtained from museum databases and to explore the effects of sampling bias on Maximum Entropy (MaxEnt) models differing with respect to model complexity. Almost all the FOP curves deviated from the expected smooth, unimodal curve shapes. Curves calculated with BTG data, which is applied in SDM as a way of correcting for sampling bias, showed the strongest deviations from the TFP curves. The BTG approach was found to give rise to complex, often ecological meaningless FOP curves, essentially modeling the relationship between ecological conditions in sites where the focal species had been sampled and the ecological conditions in sites where taxonomically related species had been sampled. PRFP curves were in general similar to the corresponding FOP curves, most strongly so for curves for the most complex models. This demonstrates good ability of the MaxEnt method to fit response models to the data, while at the same time points out that the appropriateness of these models depends on the quality of the response data. However, results from papers I and II also show that, in some cases, complex models overfit to irregularities in the presence data, and that simpler models fit the more general responses to the environmental variables, without significantly loosing predictive power in terms of AUC. Such complex models are difficult to interpret ecologically and may not be transferable to other areas or future climates.

In paper III the effect of varying pseudo-absence data in SDM was explored by using empirical data for four real species and simulated data for two imaginary species. Pseudo-absence data generated by different sampling designs and in different numbers were added to assess their relative importance for the SDM output. The number of pseudo-absences had minimal effect on the predictive performance (measured by AUC), while sampling design strongly influenced the AUC values. This was attributed to the relationship between the environmental range of the pseudo-absences (i.e. the extent of the environmental space being considered) and the environmental range of the presence observations (i.e.

under which environmental conditions the species occurs). Results from paper II and IV show that uncritical use of BTG data in SDM should be avoided and that plotting species FOP curves, calculated by different kinds of background datasets, can be a valuable aid in the selection of an ecologically informative background.

Scarcity of presence data can also be a major obstacle for the modeling of species distributions. Accordingly, knowing the minimum number of presences required to obtain reliable distribution models is of fundamental importance for applied use of SDM. In paper IV the critical sample size (CSS) sufficient for non-random predictions of species distributions was assessed. Large presence datasets for thirty insect species were used to produce reference distribution models. Models based on replicated subsamples of different size drawn randomly from the full dataset were compared to the reference model using the index of vector similarity (*IVS*). Clearly non-random models were obtained with as few as 10 presences for 90% of the species and 15 presences for 97% of the species. We recommend using a minimum sample size of 10–15 presences for Maxent modeling to obtain additional information about a species distribution.

Small presence datasets (paper IV), the lack of absence data (paper III) and the choice of pseudo-absence/background data (paper II) can all be related to the issue of sampling bias and the distribution of frequency of presence along environmental gradients (paper I). The dataset is biased if the pseudo-absence/background dataset and the presences together fail to reflect the frequency of true presence of the model target. Moreover, with small presence (or pseudo-absence/background) datasets a high occurrence of stochasticity (the fourth type of sampling bias) is expected.

The development of SDM methods has opened up for numerous possibilities within applied ecology and conservation biology. These methods have been seen as “shortcuts’ for inferring past, present and future species distributions, as well as exploring species-environment relationships and dealing with phylogenetic questions, with the click of a button. This shortcut may, however, turn out to be deceptive in cases where the data quality is poor or data are sparse. A firm foothold in ecological theory, good understanding of how the modeling methods work, how the data are handled, and how models are to be interpreted are all essential for predicting when and why limitations in museum data become obstacles for making reliable distribution models.

3. Introduction

Today species are disappearing at a faster rate than has been seen earlier in the Earth's history (Barnosky et al. 2011) and there is scientific consensus that most species extinctions are now caused by human actions such as land use changes, pollution, human caused climate change and introduction of alien species to new areas. One of today's greatest challenges is to slow down the rate at which species are disappearing. Two obstacles are making this task difficult; the Linnean and the Wallacean shortfalls (Whittaker et al. 2005), i.e. the lack of knowledge about which and how many species populate the earth and the lack of knowledge about how they are distributed. Species distribution modeling (SDM) has been developed as a promising tool to address these challenges in diverse ways: to find new populations (Bourg et al. 2005, Guisan et al. 2006, Pearson et al. 2007) and new species (Raxworthy et al. 2003), to increase the knowledge of how species distributions will change under an altered climate regime (Thomas et al. 2004, Elith and Leathwick 2009, Austin and Van Niel 2011), or if introduced to new areas (Peterson 2003, Bean et al. 2012), to identify suitable habitats for reintroduction of species (Osborne and Seddon 2012), to explore species-environment relationships (Osborne and Seddon 2012) etc. The term SDM refers to correlative methods that link georeferenced species presence (and sometimes also absence) data with relevant environmental variables in order to predict the species distribution (Guisan and Zimmermann 2000, Elith et al. 2006).

In the last couple of decades we have seen a strong proliferation of scientific papers focusing on SDM and, as for most new tools, problems and challenges concerning the methodology arise along the way. Questions such as which algorithms to use (Elith et al. 2006), how grain and extent of the study area influence the models (Boulangeat et al. 2012), which environmental variables to apply, how biotic factors influence the models (Boulangeat et al. 2012, Wisz et al. 2013), how to deal with small datasets (Pearson et al. 2007, Bean et al. 2012, Hanberry et al. 2012) or the lack of information about species absence data (Stokland et al. 2011, Barbet-Massin et al. 2012, Golicher et al. 2012) and how to detect and mitigate sampling bias (Kadmon et al. 2003, Kadmon et al. 2004, Loiselle et al. 2008, Phillips et al. 2009, Bean et al. 2012) have all been discussed in numerous scientific papers. All of these questions need to be properly answered for SDM tools to be safely applied for practical, e.g., management purposes.

This thesis focuses on some of the greatest challenges concerning the application of museum data for distribution modeling purposes; *the quality and quantity of the training data*, particularly; the challenges of biased data (paper I and II), the sample size and the design of pseudo-absence/background data (paper II and III) and the sample size of

presence data (paper IV). In all the papers, and especially in the two first, the focus is also on what Austin (2007) calls the ecological model, i.e. the ecological understanding of the modeling process, a topic including the choices of model settings. These are all issues that should be addressed and understood before applying SDM for practical purposes.

3.1. Museum data: Properties and limitations in the context of species distribution modeling

Museum data consist of records associated with objects held in museum collections and are the largest source of biodiversity data in the world. There may be as many as 3 billion objects (http://www.sciencedaily.com/releases/2014/02/140226132750.htm) in the world's natural history museums, but only a small fraction of these are digitized and georeferenced. Museum data span not only space, but time as well. Collections made in habitats that no longer exist or of species now extinct are among the qualities that make museum data a virtual goldmine for those interested in the distribution of species.

Most objects are accompanied by information about the locality where the object was found, the taxonomy of the object and sometimes the technique used for sampling. Some of the objects are sampled by museum staff, but most objects are collected in by amateurs. This results in a great variation in the quality of the information related to the objects. In an SDM context, one of the greatest problems is poor (inaccurate) information about the sampling locality. This is especially challenging when looking at old data. Earlier the focus was rarely on exactly where an organism was found, but rather on variations within species, which species it was, etc. As a result, museum staff has had to georeference these objects retrospectively by interpreting the often rather insufficient locality information supplied by the collectors. One example is the use of the locality name "Dovre' for all mountain areas in Southern Norway, although Dovre *sensu stricto* is a specific mountain area in Central South Norway. Such specimens will inevitably have a very low georeferencing precision. Because the traditional use of museum data has been for taxonomic studies, the focus when managing the collections has mostly been on knowing what is in the collections and approximately where the material was collected, determined to country or region. Nowadays, however, most collectors own a global positioning system (GPS) unit and locality information tends to include geographic coordinates.

In addition to varying precision of the locality information, a problem with applying museum data for SDM is that they are normally not sampled for this purpose. Ideally, sampling should be random or designed to sample the whole environmental gradient, which is to be studied. Museum data is, however, likely to be biased towards areas that are easily accessible for biologists, that are known for their interesting fauna and flora, or

otherwise of special interest. As a consequence, distribution models trained with such data will be inaccurate. One example of biased sampling is the sampling of insects, with sweep nets, from cars, thus giving a heavy roadside bias in the resulting data. It is well known that the insect collections at the Natural History Museum of the University of Oslo are geographically biased to the south, were most people live. Although many authors have addressed sampling bias-related question in an SDM context (Loiselle et al. 2008, Veloz 2009, Boakes et al. 2010, Costa et al. 2010, Robertson et al. 2010, Wolmarans et al. 2010, Anderson and Gonzalez 2011, Feeley and Silman 2011a, McCarthy et al. 2011, Merckx et al. 2011, Yackulic et al. 2013) there is still a strong need for tools that can aid detection of bias and effects of such bias on distribution models, as well as guidelines for use of biased data (or when to leave data unused) in SDM (Boakes et al. 2010). Development of such tools and guidelines requires theoretical in-depth understanding of what sampling bias is and why and how it affects species distribution models, including an unambiguous definition of sampling bias in an SDM context.

Another type of bias associated with museum data is the temporal bias. Because at any time a limited number of collectors typically provide a large fraction of the objects (of a specific broad taxonomic group) delivered to the museum, objects will to a large extent represent the taxonomic groups of interest to the most productive collectors. Similarly, charismatic species or species of economic interest will be over-represented in the collections. Figure 1 shows that a disproportionately large fraction of the museum specimens from the four large Natural History Museums in Norway is made up by charismatic species groups such as vertebrates and vascular plants. If we restrict ourselves to specimens for which information is available in digital form, the bias is even stringer. The flipside of this is that few data are available, not only for rare species, which are hard to detect, but also for common species and species that, for some reason, are of little interest to the majority of collectors. Small datasets have been found to be a major obstacle for modeling species distributions (Lim et al. 2002, Papeş and Gaubert 2007, Feeley and Silman 2011b, Feeley and Silman 2011a, Kamino et al. 2012). The effect of sample size on species distribution models and the existence of a minimum presence sample size needed to generate reliable models have been widely debated. There exists a general agreement that the models' accuracy increases when sample size increases (Cumming 2000, Pearce and Ferrier 2000, Stockwell and Peterson 2002, Reese et al. 2005, Hernandez et al. 2006, Wisz et al. 2008), but no general consensus regarding the amount of presence data needed to

reliably predict a species' distribution has yet been reached.

Last, but not least, because museum data differ from observational data by being verifiable i.e., that the organism is sampled and stored as a physical object, these data do not include absence data, i.e. reliable data on where the species is *not* present. In the absence of absence data, pseudo-absence or 'background' data are typically created to make possible the use of modeling methods that require data to which presence data are contrasted (group discriminating methods). Several principles for generating pseudo-absence data have been proposed during the last two decades (Stockwell and Peters 1999, Hirzel et al. 2001, Zaniewski et al. 2002, Elith and Leathwick 2007), one of which being the use of presence observations for other species in the group of species to which the target species belongs (background target group; BTG) (Phillips and Dudík 2008, Phillips et al. 2009). BTG has been launched as a means of mitigating the effects of bias in presence datasets on SDM models. The idea is that, by using background data with similar bias as the presence-only data, the estimated frequencies of observed presence will be closer to the true frequencies of presence. The degree to which use of BTG will improve distribution models will depend on the extent to which the assumption of similar biases in training and BTG data holds true (Phillips et al. 2009).

Natural history museums and databases all over the world contain large amounts of information of where different species of animals and plants are located. Despite the deficiencies of museum data with respect to bias, data-set size and the lack of absence data, these data can be of great value for SDM as long as we are able to assess the limitations of the data and interpret our models accordingly.

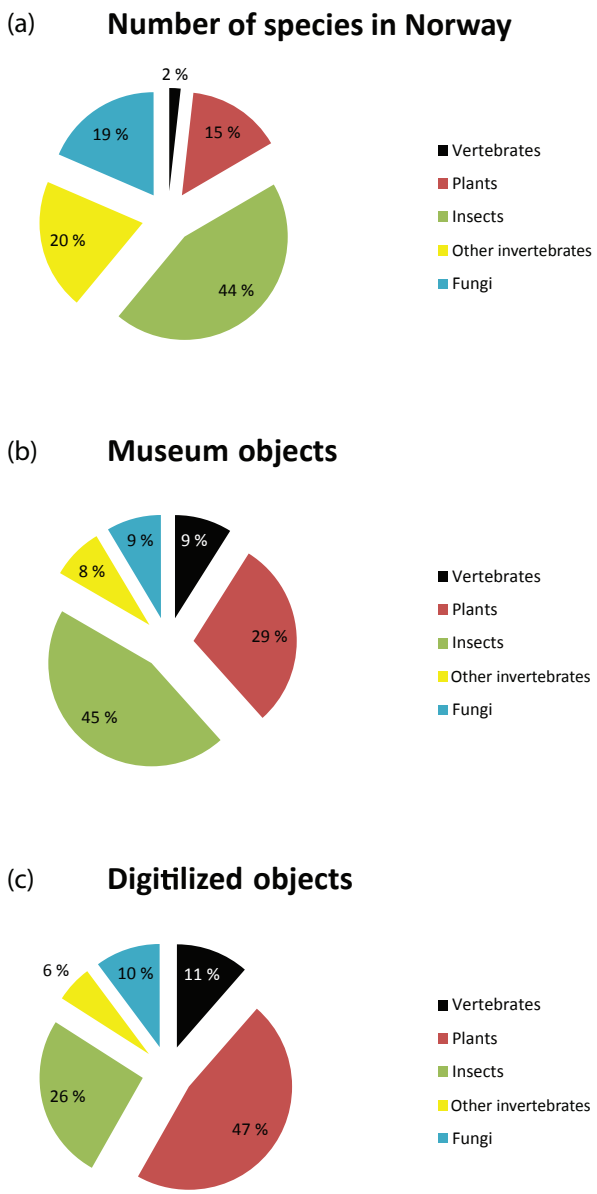

Figure 1. : Chart showing (a) the percent of all Norwegian species belonging to each organism group, (b) the percent of all museum objects in the four largest Natural History Museums in Norway belonging to each organism group, (c) the percent of all digitalized objects from the four largest Natural History Museums in Norway belonging to each organism group.

3.2. Theory

A good grasp of ecological theory, including knowledge about how species distributions are limited, is essential for appropriate use and skillful interpretation of species distribution models (Austin 2002, Guisan and Thuiller 2005, Austin 2007, Halvorsen 2012). Here follows a brief introduction to some topics that are of particular relevance to understanding how properties of the data influence distribution modeling results: gradient analysis, niche theory, source-sink dynamics and biotic interactions. Albeit these topics are all interrelated, I will treat them separately in the next sub-chapters.

3.2.1. Gradient analysis

The species distribution model is a generalization of the ecological response of the species to selected environmental variables, in environmental space. Knowledge of how species are distributed in conceptual environmental space is therefore essential for evaluating the quality of the response data used in distribution modeling and for interpreting model output. Species-environment relationships can be explained by the so-called gradient analytic perspective (Ter Braak and Prentice 1988, Halvorsen 2012), which predicts that species respond to environmental complex-gradients, i.e., sets of environmental gradients that act on species in concert, rather than to each single environmental variable individually (Whittaker 1956). This is because environmental variables of importance for the survival of the organisms tend to be correlated with other environmental variables. Furthermore, a few major complex-gradients do normally account for most of the variation in species composition in a given ecosystem.

Species' responses to complex-gradients are typically unimodal, because species have developed preferences for specific parts of the environmental space. With increasing 'environmental distance' from the optimum of a species along a major complex-gradient, habitat suitability and species performance decrease gradually (Whittaker 1956, Austin and Smith 1989). The tolerance of the species to conditions that vary along important environmental gradients determines the width of the species' response curve (e.g. Dahl and Birks, 1998), which, in turn, determines the range of the species' distribution along the gradient. The commonness of the species within its distributional range is given by the height of its response curve. Many ecological textbooks have described species response curves as unimodal and symmetrically bell-shaped (Giller 1984, Begon et al. 1990, Krebs 1994). The relationship between a species' tolerance and its commonness is the topic of abundance-occupancy theory, which predicts the two to be positively correlated (Gaston and Lawton 1988, Brown et al. 1996, Gaston et al. 2000). Hanski (1982) proposes that most species belong to one of two categories; core species, characterized by wide tolerance

and high abundance, and satellite species, characterized by narrow tolerance and low abundance. The CURS model of Collins et al.(1993) opens for the existence also of urban species with restricted distributions but high local abundance, and widespread but rare, rural, species, although fewer species belong to the urban and rural than to the core and satellite species categories.

The symmetrical, unimodal pattern can be displaced by other response curve shapes for several reasons, such as the influence of other species (Austin and Smith 1989, Austin and Gaywood 1994, Austin 2002), dispersal constraints (Primack and Miao 1992, Spens et al. 2007, Hatteland et al. 2013), the influence of other important complex-gradients (Halvorsen 2012) or by local adaptations (Westley et al. 2013). The degree to which a response curve is symmetric thus depends on the nature and the scaling of the underlying gradient and the species' response to it (Økland 1986, Økland 1992). The response-curve shape is also strongly affected by the grain and the extent of the study area (Loehle 2012); the tendency for a curve to become flat-topped (platykurtic) will increase with increasing commonness and increasing grid-cell size, while a truncated response curve will result if the full range of the species along the gradient in question is not contained within the study area (Halvorsen 2012).

3.2.2. Niche theory

What is really being estimated in SDM? This question has been widely debated in SDM literature over the last decades and the 'niche' is a key concept in many papers (Guisan and Zimmermann 2000, Pulliam 2000, Guisan and Thuiller 2005, Soberon and Peterson 2005, Araújo and Guisan 2006, Soberón 2007, McInerny and Etienne 2012a, McInerny and Etienne 2012b, McInerny and Etienne 2012c). The understanding of this concept seems fundamental for the interpretation of species distribution models, and a short summary of the debate over the validity of the niche concept in SDM is given here.

The ecological niche of a species was originally defined in two conceptionally different ways, which are now referred to as the Grinellian niche (Grinnell 1917), which encompasses the environmental conditions that determine a species' distribution, and the Eltonian niche (Elton 1927), which emphasizes the functional attributes of species and their corresponding trophic position. Hutchinson (1957) developed the Grinnellian niche concept further by describing the niche as an n-dimensional hypervolume in which every point corresponds to an environmental state (combination of important environmental conditions), which permit the species to exist indefinitely. He further distinguished between the fundamental niche, defined without including the effects of competition, and the realized niche, described as the fundamental niche reduced by the effects of competition. Competition comprises

interactions with a negative outcome for both organisms. Depending on the relative magnitude of the outcomes for each of the interacting organisms, competitive interactions can be ordered along a gradient from symmetric (equal outcomes for both species) to asymmetric. It remains unclear whether or not other biotic interactions than competition are considered part of Hutchinsons fundamental niche, or if they are to be interpreted as factors shaping the realized niche (Halvorsen 2012).

Most authors have argued that the realized niche is estimated in SDM studies. The rationale behind this is that georeferenced species presence and/or absence data used in the modeling are already constrained by biotic interactions, so that they cannot be used to model the fundamental niche of the species (Guisan and Zimmermann 2000, Pearson and Dawson 2003, Kearney 2006). Soberon and Peterson (2005) do, however, claim that the resulting model may approach the fundamental niche of the species in certain cases. To be specific these cases occur when biotic interactions are not assumed to influence and when absences are chosen outside of the fundamental niche.

There are several issues complicating the use of the niche terms in SDM. Pulliam (2000) points out that the realized niche may be larger than the fundamental niche, when taking source-sink theory into account (see chapter on metapopulation dynamics below). Kearney (2006) argues that correlative methods applied in SDM are not suitable to quantify the niche, because the niche concept implies understanding how biotic and abiotic variables affect the fitness of an organism. Jiménes-Valverde et al. (2008) suggest that, because of the above mentioned uncertainties regarding how to understand the fundamental and realized niches, it would be preferable simply to speak of potential and realized distributions in the context of SDM. The potential distribution of a species refers to places where a species could occur (because the values of the relevant environmental variables are appropriate for the survival of the species), while the realized distribution of a species refers to places where the species actually occurs. Araújo and Guisan (2006) propose to discard the fundamental and realized niche concepts altogether, accepting that any characterization of the niche is an incomplete description of the abiotic and biotic factors determining a species distribution. McInerny and Etienne (2012b) formulate the ultimate goal of the niche concept as such: '*Niche* is a term that should support communication and understanding in ecology by helping us to re-characterize what living things do in models.' As of now it seems that the term 'niche' complicates the communication and understanding in ecology, rather than being a support.

As long as the niche term remains complicated and ambiguous, I agree with the above mentioned authors in their conclusion that it cannot be applied alone to describe

the output of SDM studies. I have therefor chosen not to use it in my scientific works. The debate of what is being estimated in SDM studies has, however, been valuable to follow, in order to conceptionally better understand how species distributions are limited.

3.2.3. Metapopulation theory and source-sink dynamics

In the preceding chapter about niche theory I mention that the understanding of the term niche is complicated by metapopulation dynamics. Metapopulation theory shows that the distribution of populations is dynamic and shifting through time due to local extinctions and dispersal (Hanski and Simberloff 1997, Hanski 1998, Hanski and Ovaskainen 2000). This explains how a species may regularly be found in unsuitable habitats or be absent from suitable habitats. A metapopulation is a 'population of populations' made up by a shifting mosaic of populations, linked by dispersal (the extent to which dispersal occurs and influences the populations will vary).

Populations found in habitats where reproduction is insufficient to balance local mortality may continue to exist because of immigration from higher-productive areas nearby. Pulliam (1988) terms such habitats 'sink' and 'source' habitats, respectively. At the same time species are sometimes absent from suitable habitats due to limited dispersal capacity, dispersal barriers or insufficient time to disperse (Primack and Miao 1992, Spens et al. 2007, Hatteland et al. 2013), or due to local extinctions (Harrison 1991, Matthies et al. 2004, Ree and Smith 2008).

A metapopulation can be made up of several short-lived populations, where the distribution of the species changes substantially from generation to generation, it can be made up by a few source populations and several sink populations, fluctuating with the arrivals of immigrants, or the distribution of populations can be relatively stable (Primack 2010).

Understanding these dynamics of populations, on different spatial and temporal scales, is important for conservation biologists and nature managers, because the destruction of a source population may result in the extinction of numerous sink populations and because habitat fragmentation limits the dispersal necessary to recolonize a habitat after local extinction (Hanski et al. 1996). In the same way this insight is important for distribution modelers by explaining that species' distributions are dynamic, by providing a conceptual model for this dynamics, and by explaining that all presences do not indicate suitable habitats, and that all absences do not indicate unsuitable habitats.

3.2.4. Biotic interactions

In some of the sub-chapters above I have mentioned that species respond, not only to abiotic (non-living, 'environmental') factors, such as temperature and precipitation, but also to biotic factors (the influence of other living organisms).

An organism's effect on another organism can be intraspecific (a species' effect on itself) or it can be interspecific (a species' effect on other species) (Elton 1927, McInerny and Etienne 2012c). My focus here will be on interspecific biotic interactions, as these are the most relevant for SDM.

Biotic interactions can essentially be of five different types; negative effects on both species (competition), positive effects on both species (mutualism), negative effect on one species and positive effects on the other (predation, parasitism and contramensalism), positive effects on one of the species and no effect on the other (facilitation and commensalism) and negative effects on one species and no effect on the other (amensalism) (Halvorsen 2012, McInerny and Etienne 2012c).

In the context of SDM it is important to keep in mind that interspecific interactions are neighbor phenomena, i.e., interactions that take place between individuals. In order to affect the distribution of a species, biotic interactions with similar outcomes have to take place between many individuals and over a large area. The larger the study area, and the coarser a grid that is used to rasterize this area, the more individuals have to be involved in interactions to affect the distribution, and hence, the distribution model (Halvorsen 2012). The SDM literature contains several examples of effects of biotic interactions on different spatial scales (by spatial scale I here specifically mean the linear grain of the study area (grain being the size, in geographical space, of one observation unit). I here use the terminology of spatial scales of Halvorsen (2012), defining the micro scale as 0.1–1 m, the local scale as 1–1,000 m, the regional scale as 1,000–1,000,000 m and the global scale as more than 1,000,000 m. Effects of biotic interactions are assumed mostly to influence species distributions on micro to local scales (Pearson and Dawson 2003, Hortal et al. 2010). Effects on distribution models obtained for local-scale data are shown for competition (Meier et al. 2010, Boulangeat et al. 2012), facilitation (Boulangeat et al. 2012), mutualism (Gutiérrez et al. 2005), and parasitic/amensalistic relationships such as effects of the availability of host plants for butterflies (Pellissier et al. 2012). Effects of biotic interactions on regional-scale distribution models have, however, also been shown; for competition (Leathwick and Austin 2001, Anderson et al. 2002), facilitation (Heikkinen et al. 2007), effect of predation (Hebblewhite et al. 2005) and the availability of host plants (Araújo and Luoto 2007, Schweiger et al. 2012) and prey (Redfern et al. 2006).

There are some difficulties associated with introducing biotic effects as variables in SDM. A biotic variable such as the presence of a potentially interacting species may represent a (real) biotic interaction, or it may act as a proxy for an unidentified environmental variable (Austin 2002). Separating these two cases is difficult because interspecific biotic interactions can only be detected at a spatial scale where organisms meet and interact (Huston 2002), and this scale will differ between organism groups, being broader for larger, more mobile species, such as mammals, and finer (and typically much finer than the scale at which SDMs are performed) for small, sessile species, such as plants.

The literature gives no unambiguous answer as to whether all biotic interactions should be considered as shaping the realized niche, or some of them should be considered part of the fundamental niche. I propose to treat biotic interactions, which are absolutely essential for a species (obligate biotic interactions), as part of the fundamental niche, while other biotic interactions (non-obligate) should be considered as part of the realized niche. Examples of the first can be the availability of an obligate prey species for specialist parasitic or predator species (or host plants for a specialist herbivore species), while examples of the latter are the distribution of predators or competitors.

3.2.5 Purposes of distribution modeling

According to the gradient analytic perspective, a species' response to important environmental gradients is generally unimodal, while niche theory defines the fundamental niche of a species as an n-dimensional hypervolume (with each dimension being an environmental variable), in which every point corresponds to a combination of environmental conditions that permits the species to exist indefinitely. The realized niche is the fundamental niche constrained by biotic interactions. Metapopulation dynamics complicate the relationship between realized niche and realized distribution even further, by predicting that species may regularly be absent from suitable habitats while present in unsuitable habitats. In the same way biotic interactions and metapopulation dynamics can be incorporated into a gradient analytic perspective. Biotic interactions may influence the shape of the response curve of the species along a gradient, by reducing (or increasing in cases of positive interactions) the tolerance of a species, or by reducing (or increasing) the abundance of the species within the tolerance limits. These effects are referred to as an amplitude response and a magnitude response, respectively, by Halvorsen (2012) and will increase towards finer spatial scales. It is, however, important to stress that the perspectives of gradient analysis and niche theory are fully compatible, merely being two different ways of presenting the same issue.

Whether or not effects of factors related to metapopulation dynamics and biotic interactions should be accounted for in SDM, depends on the modeling purpose. Spatial response modeling (SPM) and ecological response modeling (ERM) can be seen as end-points of a continuous gradient from less to more general modeling purposes, requiring models that are closely fit to the data and models that express general species-environment relationships, respectively (Halvorsen 2012). The main purpose of ERM is to model the relationship between the distribution of a target species and a set of environmental variables. The focus is to find and understand *general patterns* in the overall ecological response of the modeled target to explanatory variables in environmental space. SPM, on the other hand, aims at modeling the distribution of the target species in a specific study area in a specific time interval, modeled by use of a set of explanatory variables. The main purpose is to optimize the fit between model predictions and the true distribution of the species in this area at this point in time. SPM thus addresses relationships in geographical space.

The ERM – SPM duality can be discussed in light of the currently ongoing debate over the validity of the ecological niche term for SDM (Guisan and Zimmermann 2000, Pulliam 2000, Guisan and Thuiller 2005, Soberon and Peterson 2005, Araújo and Guisan 2006, Soberón 2007, McInerny and Etienne 2012a, McInerny and Etienne 2012b, McInerny and Etienne 2012c). Because SDMs with the ERM purpose aim at modeling general species-environment relationships, it can be argued that ERM models are estimating the fundamental niche, or the potential distribution, of the species. This is, however, a problematic argument because the presence/absence or presence-only data used to generate the model are drawn from the realized distribution of the species. According to the conceptual model of Soberon and Peterson (2005) [modified in Soberón (2007)], a species distribution is limited by three classes of factors (Fig. 2); abiotic (region A), biotic (region B) and factors related to accessibility (region M). Region A corresponds to the fundamental niche (FN) of the species, while the intersection of region A and B ($A \cap B$) corresponds to the realized niche (RN) of the species. The realized distribution (Jiménes-Valverde et al. 2008) of the species is, however, also limited by the accessibility of the area (M). This is determined by the dispersal ability of the species, dispersal barriers and (anthropogenic) means of introduction. In practice the species can be found throughout the entire region M, due to the presence of sink populations (Pulliam 2000), although populations with positive fitness (source populations) are only encountered in the area where all three regions intersect.

While metapopulation theory assumes that the environment consists of discrete patches of suitable habitat surrounded by uniformly unsuitable habitats, modern SDM methods assign a continuous variable of probability of occurrence to each pixel into which the study area is gridded. This improves the realism of model predictions (compared to approaches which provide binary predictions), by attributing low probability of occurrence to sink habitat areas. The degree to which SDMs are influenced by presence observations in sink habitats will depend on how large a fraction of the dataset such observations contribute.

In addition to the three above-mentioned classes of factors (abiotic, biotic and factors related to accessibility), deficient sampling may also contribute to imperfect representation of the fundamental niche by an ERM-purpose SDM.

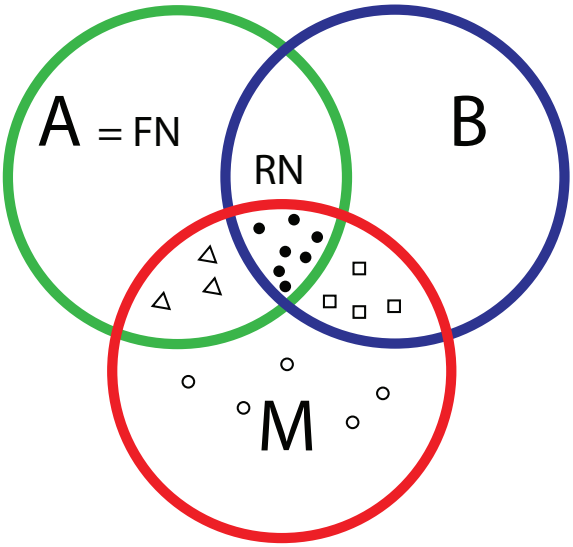

Figure 2. : A representation of factors affecting the distribution of a species, redrawn from Soberón (2007). Region A represents the geographical area where the abiotic environment is suitable for the species. Region B represents the geographical area where the biotic factors that affect resource use and biotic interactions would allow the species to persist. M represents the total area that has been or is accessible to the species within a time period of interest. Solid circles represent source populations. Open triangles are sink populations with negative growth rates due to competitive exclusion. Open squares represent sink populations due to negative intrinsic growth rates. Open circles are combinations of the above. FN = fundamental niche and RN = realized niche.

The degree to which each of these factors will influence a distribution model is also affected by the scale of the study. Dispersal barriers are expected to be rare when the extent of the study area is small (Soberon and Peterson 2005). Limitations and potentially imperfect sampling are, hence, the main reasons why ERM models cannot be interpreted as representations of the fundamental niche at broader scales. At finer scales, biotic interactions and metapopulation dynamics may also contribute to mismatch between predictions from an ERM model and the fundamental niche. In cases where biotic interactions occur among many individuals over a large area, biotic interactions may also influence the distribution of a species on a broad scale. This is especially evident when the biotic interaction is obligate [such as the availability of host plants for host-specific insects (Araújo and Luoto 2007, Schweiger et al. 2012)].

It is argued that ERM models should be simple, e.g., in terms of number of model parameters, in order to express fundamental species-environment relationships (Halvorsen 2012), so that effects of biotic interactions, limited accessibility, source-sink dynamics and insufficient sampling are ideally smoothed out.

SPM models are, by definition, less general than ERM models (Halvorsen 2012). Complex and flexible modeling tools may improve such models, and it can easily be argued that SDM models are representations of the realized niche of the species, with effects of biotic interactions taken into account (Guisan and Zimmermann 2000, Pearson and Dawson 2003, Kearney 2006). Nevertheless, temporal and geographical accessibility and metapopulation dynamics may also affect SPM models. The modeled entity is consequently not the realized niche, but rather the realized distribution of the species (Jiménes-Valverde et al. 2008). The only obstacle left for modeling the realized distribution, then, is imperfect sampling.

ERM and SPM models are suitable for different applied purposes. SPM models do, by definition, specifically describe a species' distribution in a specific study area at a given time-point (at which the data used to obtain the model are sampled). Models based on species-environment relationships characterized by idiosyncrasies of the study area, such as dispersal-barriers, anthropogenic influence or biotic interactions, tend not to be valid when projected onto another study area or to another point in time (Barbosa 2009, Wolmarans et al. 2010, Schweiger et al. 2012). In cases where the study area is large, these relationships may not even be transferable between different parts within the study area. SPM models are the obvious choice for practical cases in which the purpose is to model suitable habitat in the area from which data for training the model is obtained, given that the species' relationship to the environment can be expected to be reasonably homogeneous

throughout the study area. Applications of SPM models can include assistance in finding new populations of a species (Bourg et al. 2005, Guisan et al. 2006, Pearson et al. 2007), finding suitable habitat for species reintroductions (Osborne and Seddon 2012) and nature reserves (Loiselle et al. 2003), and finding areas for cultivation of species for food, biofuels, etc. (Barney and DiTomaso 2010, Evans et al. 2010). ERM models, on the other hand, express general species-environment relationships, more stable in space and time, e.g., with the purpose of projecting to other study areas and points in time (Halvorsen 2012). ERM models are thus preferable for modeling invasive species in new areas, species distributions in an altered climate regime, for species niche conservancy studies and for exploring species-environment relationships.

3.3. What determines the distribution of insect species?

The distribution of insects have been found to be limited by different factors at different spatial scales (Hortal et al. 2010). As for most organisms abiotic factors and biogeographic processes (speciation, extinction, isolation and long distance dispersal) are the most important factors on a global to regional scale, while biotic factors and metapopulation dynamics are more important on a local to micro scale (Cabeza et al. 2010, Hortal et al. 2010). A clear limit between scales at which abiotic factors matters more than biotic factors and vice versa can, however, hardly be drawn. This will depend, inter alia, on the dispersal capabilities of the species and the heterogeneity of the physical environment (e.g. Holway et al. 2002) .

One abiotic factor of particular importance for the distribution of insects is temperature. Because insects, with few exceptions, are ectothermic, their body temperature is ultimately determined by the ambient temperature. This in turn influences the speed and efficiency of their vital biological processes (such as development, metabolism, ecdysis and reproduction). It has been shown that temperature extremes are more important than temperature averages in defining species distributions, both in temperate regions and in the tropics (Overgaard et al. 2014). Several other temperature-related descriptors have also been found to correlate with insect distributional limits, e.g., degree days (Hawkins and Porter 2003, Luoto et al. 2006, Pellissier et al. 2012), mean temperatures (Luoto et al. 2006) and the amounts of solar radiation (Pellissier et al. 2012).

Another important abiotic factor limiting insect distributions is humidity (Fink and Völkl 1995, Hill et al. 1999). Factors related to humidity and precipitation can influence insect species directly by altering their water balance, and different species have different behavioral and physiological strategies for coping with unfavorable humidity regimes (Chown et al. 2011). Larger, more heavily sclerotized insects are less susceptible

to desiccation or waterlogging than smaller, more delicate species (Schowalter 2012). Similarly, immature stages may be much more vulnerable to dry conditions than adults (Chown et al. 2011). An insect's water balance and exposure to desiccation is also linked to temperature. In general, temperature is a more important distribution-limiting factor in cool climates, whereas the availability of water is more critical in warmer climates (Hawkins et al. 2003).

Temperature and precipitation may also have indirect effects on the species through habitat alterations, which again affects resource quality and availability, and exposure to predation and parasitism. Some herbivorous insects may for example be drawn to plants stressed by drought (Schowalter 2012). Other abiotic factors which may influence species distributions are wind (Fink and Völkl 1995) and soil chemistry (Schowalter 2012).

Land cover / habitat availability has also been found to influence insect distributions (Hill et al. 1999, Ulrichs and Hopper 2008). Land cover is likely to be an indirect predictor of insect distributions, reflecting the effects of temperature, humidity and soil structure on the organisms' physiology. Land cover may also reflect the availability of resources, such as host-plants for parasitic or plant-eating insects. The fact that land-use changes has been listed as one of the five biggest threats to the worlds biodiversity, insects included (Kålås et al. 2010), indicates that habitat availability is crucial. Habitat destruction is linked to metapopulation dynamics because habitat fragmentation limits the dispersal between different habitat patches (Hanski et al. 1996).

3.4. Aims of the thesis

The overall aim of the thesis is to explore how limitations in the response data, i.e. the quality and quantity of the data used to train or parameterize SDMs, affect the outcome of the modeling.

Specific goals are:

- to provide a theoretically founded understanding of what sampling bias is and to explore its effects on species distribution models (paper I and II)

- to explore how the choice of background data affects species distribution models (paper II and III)

- to examine how the number of presence observations affects species distribution models, and to assess if a general minimum sample size required to obtain useful SDMs, can be found (paper IV)

4. General material and methods

4.1. Study area

The study area for the analyses of all four papers was the mainland of Norway, comprising 323,782 km^2, from 58 to 71° N and from 5 to 31 ° S, from sea level to 2469 m a.s.l. This area is especially suited for exploring properties of distribution models because of the high topographical and geological diversity and the variation in intensity of human land-use, e.g. for agricultural purposes, which bring about variation in environmental conditions and species composition over a large range of spatial scales (Halvorsen 2012). The degree of variation over short distances is rare, not only within the Nordic countries, but also in a global context (Moen 1999).

Norway contains strong regional bioclimatic gradients in temperature and oceanicity-continentality (Moen 1999, Bakkestuen et al. 2008). The temperature gradient follows the range of vegetation zones represented in Norway (from boreonemoral to high alpine) (Fig. 3a), while the oceanicity-continentality gradient follows the range of vegetation sections (from the strongly oceanic to the slightly continental) (Fig. 3b).

The main land-cover types in Norway are non-forested land (46%; mainly situated above and north of the tree line), forest (38%) and mires and lakes (6 % each).

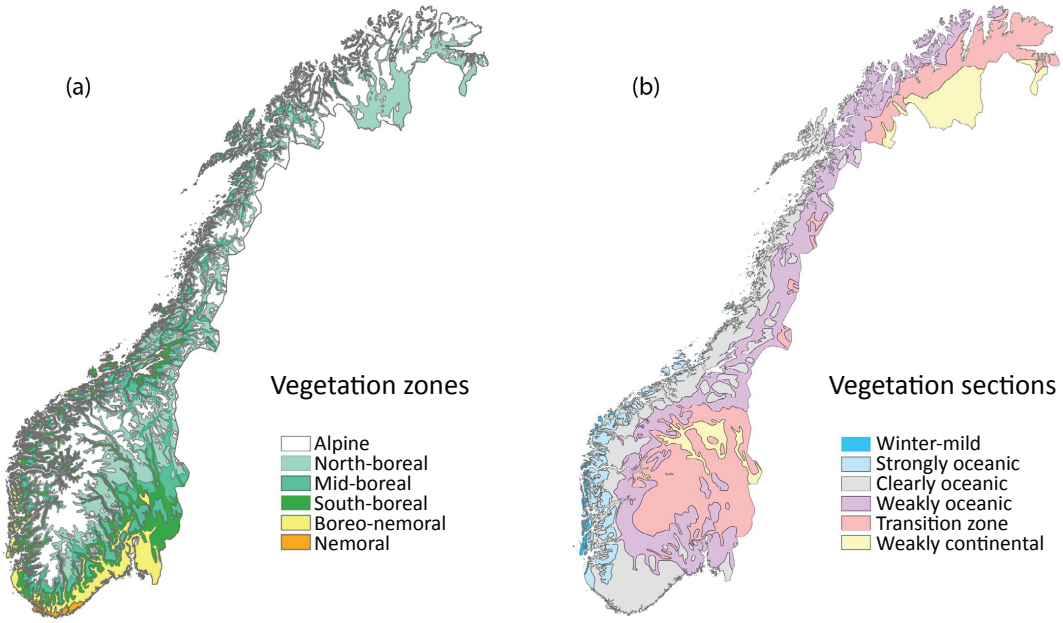

Figure 3. Map of (a) the vegetation zones and (b) the vegetation sections of Norway (Moen 1999).

4.2. Study organisms; presence and background target group datasets

Insects were chosen as the only study organisms in two of the four papers of this thesis. One paper is a theoretical paper, without any analyses, while one paper includes two insect species as well as two fungi species (Table 1).

Insects are particularly interesting because they comprise more than half of the worlds described species (Chapman 2009) and because many of them are represented by few presence records (New 2009). Furthermore insects represent a wide variety of life histories, dispersal abilities, size and shapes, which can potentially influence their responses to environmental variables and, hence, their distributions. For the analyses in paper IV ten species from each of three orders (Coleoptera, Lepidoptera and Diptera) were chosen (Fig. 4). This was done to include species with different dispersal abilities, feeding habits, sizes and shapes.

Coleoptera: Based on numbers of described species, beetles are the most diverse group of organisms on Earth. In most species the elytra cover the membranous flight wings and the abdomen. In this way the beetles are protected against predation and environmental stress (Foottit and Adler 2009). Of the families represented in paper 4, two of them, Curculionidae and Cerambycidae are associated with wood and are known to sometimes cause damage to trees. The species belonging to Nitidulidae and Elateridae are herbivorous, while the Cantharidae species is a predator.

Diptera: Diptera are commonly called true flies or two-winged flies and are among the most diverse insect orders, not only in species richness, but also in structure, habitat selection and life habits. The larvae of most diptera species can be considered aquatic in the broadest sense, because they require a moist to wet environment within the tissues of living plants, in decaying organic material, as parasites or parasitoids of other animals or in association with bodies of water (Foottit and Adler 2009). The diptera species in paper IV include species from the families Conopidae (parasitic), Asilidae (predatory) and Syrphidae (adults feed on pollen and nectar, while larvae are either saphrotrophs or predatory).

Lepidoptera: Lepidoptera is a large insect order, including moths and butterflies. They inhabit all kinds of terrestrial habitats, but almost all species are associated with higher plants, as they feed on nectar. Lepidopterans are soft bodied and fragile and, hence, exposed to predation. Almost all species have some form of membranous wings and are good dispersers. Some species are even migratory. However, they have to be warm in order to fly and this is further dependent on their environment (Foottit and Adler 2009). Lepidopterans are easy targets for SDM studies because they are charismatic and relatively easily identifiable. Therefore Lepidoptera species are well represented in museums

and databases. The lepidopteran species used in paper IV include five species from the superfamiliy papilionoidea (the true butterflies); from the families Papilionidae, Pieridae and Lycaenidae, in addition to three species of the family Geometridae, one species from the family Lasiocampidae and one from Zygaenidae.

The choice of species was also based on the desire to include species with different distribution patterns. I tried to include wide-spread species as well as species with more limited distributions. This was, however, not easy, as an additional requirement was a minimum sample size of 40 presences, for subsampling to be possible. Therefore, few red listed species were included, with the exception of three species within the order of Lepidoptera; *Thecla betulae* (Fig. 5 b) (category NT), Glaucopsyche alexis (Fig. 5a) (category NT) and *Aporia crataegi* (category VU) .

Paper II was written after paper IV, and all datasets from paper IV were already available. In order to avoid having to present too many response curves in paper II, we only wanted to analyze a subset of the species used in paper IV. These were chosen to fit in a 3x3 design, with three species from each of three orders. Within each order we selected one species with a restricted distribution, one with an intermediate distribution and one with a broad distribution. If presented with more than one alternative from each distribution class, we selected the one with the largest presence dataset.

Most of the insect presence data used in this thesis (3267 georeferenced presences altogether) were extracted from the database of the insect collection at the Natural History Museum, University of Oslo, Norway (89 %). Additional data were obtained from the insect collections at Bergen Museum (University of Bergen, Norway) (6 %) and the Museum of Natural History and Archaeology (Norwegian University of Science and Technology, Trondheim) (5 %).

To generate background target group (BTG) datasets for each species in paper II and the beetle species in paper III, all geographically referenced presences of their respective taxonomic families in Norway, extracted from the database of the insect collection at the Natural History Museum, University of Oslo, Norway were used. The BTG datasets for the fungi in paper III were generated from all geographically referenced presences of their taxonomic family in Norway, extracted from the Norwegian GBIF database. This was done because it can be assumed that specimens from the same taxonomic family at the same museum (or from the same database) are sampled by the same group of collectors, with approximately the same sampling technique. For each taxonomic family there are often a few collectors contributing data to the museum, and the sampling effort has often been concentrated around the ‘favorite sampling areas’ of these collectors. Data for species from the same family may therefore share the same sampling bias.

Although the datasets for all four papers cover the entire country, sampling intensity has generally been higher in the south east (Fig. 6). This is related to the fact that the majority of collectors lives in this part of Norway.

Table 1. List of modeled species included in the different papers, with their systematic position, ecological range of the collected presences, number of presences, and number of presences for the family of which they belong (BTG).

Species	Family	Order	Ecological range	No. presences	No. BTG	Paper II	Paper III	Paper IV
Ips acuminatus	Curculionidae	Coleoptera	Broad	65				X
Leptura maculata	Cerambycidae	Coleoptera	Restricted	58	3292		X	X
Meligethes aeneus	Nitidulidae	Coleoptera	Restricted	60	443	X		X
Otiorhynchus nodosus	Curculionidae	Coleoptera	Broad	82				X
Pogonocherus hispidus	Cerambycidae	Coleoptera	Restricted	44				X
Rhagium mordax	Cerambycidae	Coleoptera	Intermediate	111	1228	X		X
Rhagonycha limbata	Cantharidae	Coleoptera	Broad	113	312	X		X
Selatosomus aeneus	Elateridae	Coleoptera	Broad	84				X
Strophosoma capitatum	Curculionidae	Coleoptera	Intermediate	46				X
Tetrops praeusta	Cerambycidae	Coleoptera	Intermediate	41				X
Anoplodera sexguttata	Cerambycidae	Coleoptera	-	31	3292		X	
Conops quadrifasciatus	Conopidae	Diptera	Restricted	133				X
Dioctria hyalipennis	Asilidae	Diptera	Restricted	95	375	X		X
Eristalis arbustorum	Syrphidae	Diptera	Broad	137				X
Eristalis interrupta	Syrphidae	Diptera	Broad	92				X
Eristalis intricaria	Syrphidae	Diptera	Broad	194	652	X		X
Eristalis pertinax	Syrphidae	Diptera	Intermediate	141				X
Laphria flava	Asilidae	Diptera	Intermediate	73				X
Neoitamus socius	Asilidae	Diptera	Restricted	93				X
Sicus ferrugineus	Conopidae	Diptera	Intermediate	210				X
Volucella bombylans	Syrphidae	Diptera	Intermediate	103	598	X		X
Aporia crataegi	Pieridae	Lepidoptera	Intermediate	57				X
Glaucopsyche alexis	Lycaenidae	Lepidoptera	Restricted	84	1541	X		X

Heterothera serraria	Geometridae	Lepidoptera	Intermediate	49				X
Lasiocampa trifolii	Lasiocampidae	Lepidoptera	Restricted	45				X
Parnassius apollo	Papilionidae	Lepidoptera	Intermediate	88	211	X		X
Pieris napi	Pieridae	Lepidoptera	Broad	346	926	X		X
Thecla betulae	Lycaenidae	Lepidoptera	Restricted	43				X
Xanthoroe annotinata	Geometridae	Lepidoptera	Restricted	112				X
Xanthoroe decoloraria	Geometridae	Lepidoptera	Restricted	238				X
Zygaena exulans	Zygaenidae	Lepidoptera	Restricted	169				X
Fomitopsis rosea	Fomitopsidaceae	Polyporales	-	676	35.892		X	
Xylobolus frustulatus	Stereaceae	Russulales	-	310	35.892		X	

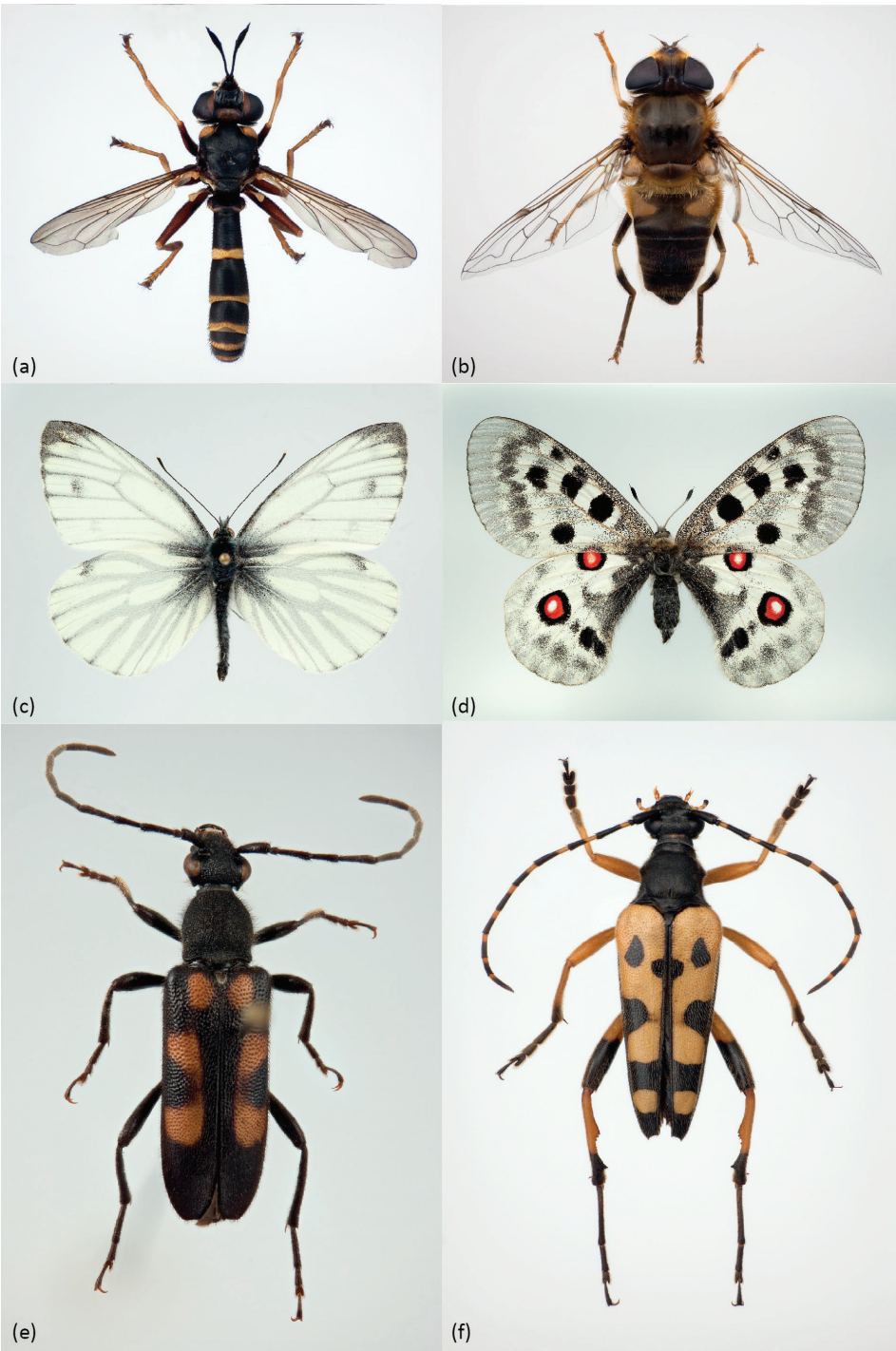

Figure 4. (a) *Conops quadrifasciatus* (paper IV), (b) *Eristalis pertinax* (paper IV), (c) *Pieris napi* (paper II and IV), (d) *Parnassius apollo* (paper II and IV), (e) *Anoplodera sexguttata* (paper III) and (f) *Leptura maculata* (paper III and IV) (Photo: Karsten Sund).

Figure 5. The red listed species; (a) *Glaucopsyche alexis* (paper II and IV) and (b) *Thecla betulae* (paper IV) (Photo: Karsten Sund).

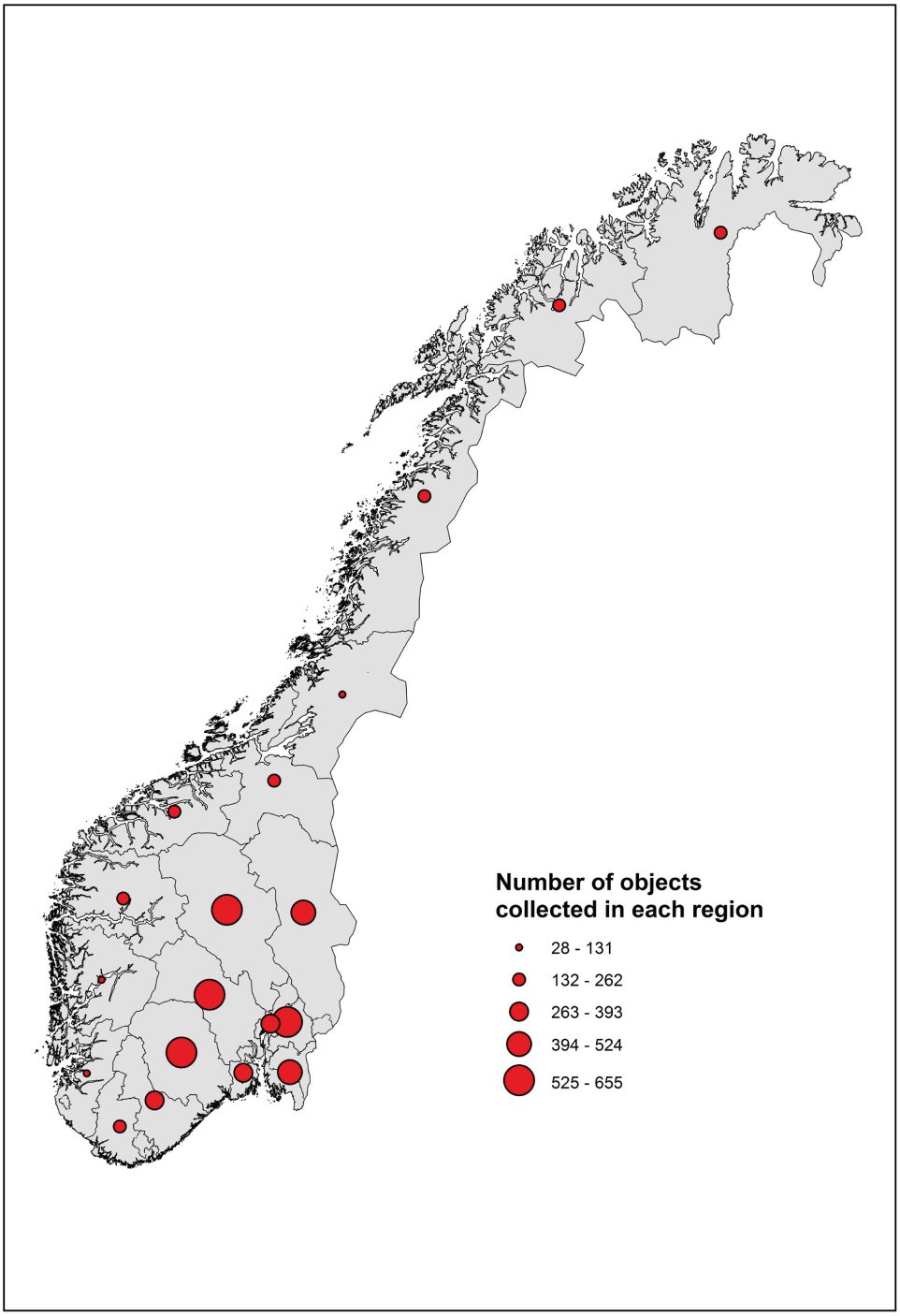

Figure 6. Number of objects from the five background target group families used in paper II, collected in each region of Norway.

4.3. Environmental variables

In paper II and IV we used variables representing the two most important regional bioclimatic gradients in Norway (Moen 1999, Bakkestuen et al. 2008), the oceanicity-continentality gradient (bioclimatic sections) and the summer temperature gradient (bioclimatic zones), as explanatory variables. The bioclimatic gradients were expressed as step-less variables for which values were available for all 1×1 km grid cells in the rasterized study area with center-point not falling on sea. The two variables were obtained by Bakkestuen et al. (2008) as the directions of best fit of vectors for ordered classes for vegetation sections and vegetation zones according to Moen (1999), in the space spanned by the first two axes of the principal component analysis (PCA) ordination of 54 climatic, topographical, hydrological and geological variables. The bulk of these 54 variables were derived from 1-km resolution raster climatic data based on the 1960-90 normals (Aune 1993, Førland 1993), compiled by the Meteorological Institute (Tveito et al. 1997) while terrain data (based upon a 100-m resolution digital elevation model) were obtained from the Norwegian Mapping Authorities, hydrological data from the Norwegian Water Resources and Energy Directorate (Beldring et al. 2002) and geological data were obtained from the Norwegian Geological Survey (based on vector data scale 1:1,000,000). The two resulting PCA axes accounted for 63% of the variation in the dataset (Bakkestuen et al. 2008). PCA efficiently reduces the dimensionality of matrices of correlated, multicollinearly related variables (Bakkestuen et al. 2008). PCA ordination is therefore commonly used prior to species distribution modeling to extract a few, step-less, axes of variation from large matrices of environmental variables. These variables are then used as environmental predictor variables in the subsequent species distribution model (Barve et al. 2011).

For the analyses in paper III the two above-mentioned variables were selected in addition to terrain ruggedness (the mean elevation difference between adjacent 100 m x 100 m grid cells within the 5 km x 5 km grid cells of the rasterized study area, calculated by a standard procedure (TRI in ArcView GIS 9.1; (Riley et al. 1999), forest cover (fraction of grid cell covered by forest according to the digital map series N50 from the National Mapping Authorities of Norway), solar radiation in April (maps of estimated potential solar irradiance in April, rasterized from vector format maps scale 1:7,000,000; (Aune 1993) and July precipitation (July mean values for precipitation, based on the 1961–90 normal, where the original estimates obtained for a 1 km x 1 km grid were averaged; (Tveito et al. 1997). These variables were selected from a candidate set of 12 environmental predictor variables available for all of Norway, rasterized to 5 x 5 km resolution, of which no predictor was allowed to have a Pearson`s product-moment correlation coefficient higher than 0.7 with any other selected predictor.

4.4. Modeling methods

4.4.1. MaxEnt and the Maxent software

Most of the data available for species DM are of the presence-only type, i.e., datasets in which observations of species absences are lacking (Franklin 2009, Stokland et al. 2011). Accordingly, many of the commonly used DM methods are adapted to use presence-only data (Franklin 2009). One such method is Maximum Entropy (MaxEnt) (Phillips et al. 2006, Phillips and Dudík 2008, Elith et al. 2011, Halvorsen 2013) which, based upon the maximum entropy principle (Jaynes 1957a, b), estimates a probability distribution for the modeled target over the set of all *n* grid cells of a rasterized study area (Phillips et al. 2006, Phillips and Dudík 2008, Elith et al. 2011). MaxEnt has proved to give models with acceptable predictive ability even when few presence records are available, i.e., when the sample size is small (Elith et al. 2006, Hernandez et al. 2006, Marini et al. 2010, Rebelo and Jones 2010, Peterson et al. 2011). The most popular software for MaxEnt as a method for distribution modeling is Maxent (Phillips et al. 2004, Phillips et al. 2006, Phillips and Dudík 2008, Phillips et al. 2012) (note the distinction between the software Maxent and the modeling method MaxEnt). Maxent version 3.3.1 is used in paper IV, while the more recent Maxent version 3.3.3k (http://www.cs.princeton.edu/~schapire/maxent) is used in paper II. Default Maxent settings implies automatic generation of derived variables of up to five types [‘auto features’ in the terminology of Phillips et al. (2006)]. A derived variable is a variable derived from the raw explanatory variable by transformation, i.e. by a mathematical operation. Which types of derived variables that are generated in each case depends on the number of presence observations: linear variables are used for all sets; quadratic variables are used for sets with ≥ 10 presence observations; hinge variables for sets with ≥ 15 presence observations; and product and threshold variables are used for sets with ≥ 80 presence observations. The most parsimonious model is then selected by the ℓ_1-regularisation (Hastie et al. 2009) or lasso penalty (Tibshirani 1996) method which operates by parameter shrinkage (Reineking and Schröder 2006, Halvorsen 2013). By this method, a regularization multiplier, set by the user, regulates the stringency of the variable selection procedure. The default regularization multiplier in Maxent is 1. When a smaller regularization multiplier is selected, the model will be closer fit to the data. The use of default parameterization in Maxent has been questioned by several authors (Raes and ter Steege 2007, Anderson and Gonzalez 2011, Merckx et al. 2011, Warren and Seifert 2011), and Halvorsen (2013) suggests that the complex response curves produced with default Maxent settings and the large number of derived variables with nonzero parameters listed in the *NN.lambdas* output file from Maxent software indicates that Maxent models tend to be overfit.

In paper IV the raw output format was used. This consists of a set of values that sum to unity for the total training dataset of presence and uninformed background observations. In paper II model output was expressed in probability-ratio output (PRO) format (Halvorsen, 2013), obtained by multiplication of each raw output value with the total number of background observations. PRO output values differ from raw output by being independent of N, the number of background observations and by attributing a specific interpretation to the output value of 1 (which is the average output value for all background observations).

4.4.2. An alternative MaxEnt process: A manuel forward selection method

In paper II additional MaxEnt models (referred to as ecological response models; ERM) were obtained for each species by a two-phase stepwise forward selection procedure as outlined by Halvorsen (2013). The idea behind this approach was to generate simpler models with fewer parameters, by which ecologically more realistic response curves could be obtained, a to make the steps in the variable selection process more traceable.

By this alternative procedure, default ℓ_1-regularisation is replaced by a model optimization criterion based upon comparison of variations accounted for by nested MaxEnt models (Halvorsen 2013). Forward selection of variables can be described as a process with three phases: (1) testing each single derived variable for individually significant contribution to explaining variation in the response; (2) selection of a parsimonious set of the derived variables passing step (1) to represent each explanatory variable; and (3) selection of a parsimonious set of explanatory variables, each represented by the sets of derived variables obtained in step (2). For comparison of two nested MaxEnt models, we used an *F*-ratio test with the significance level $\alpha = 1 \cdot 10^{-6}$. All types of derived variables ('features' in the terminology of Phillips et al. 2006), except product (interaction) variables, were generated. In addition, a sixth variable type, the deviation variable, was obtained for all combinations of species and explanatory variables for which the FOP curve (after smoothing) had a distinct peak. This peak was used as an estimate for the species' optimum along the gradient. The deviation variable was obtained as the absolute value difference between a grid cell's explanatory variable value and the estimated optimum. After generating derived variables, steps (1)–(2) were performed, as described above. Model output was expressed in PRO format (Halvorsen 2013). A set of customized R scripts (S. Mazzoni & R. Halvorsen, unpubl. material) was used together with the Maxent software (Phillips et al. 2006, Phillips and Dudík 2008).

4.4.3. Boosted regression trees (BRT)

In paper III boosted regression trees (BRT) (Friedman et al. 2000, Friedman 2002, Hastie et al. 2005, Elith et al. 2008) is applied as the modeling method. BRT is a machine learning method that in comparative studies has been shown, in general, to perform among the best (Elith et al. 2006). It combines the strengths of regression trees (models that relate a response to predictors by recursive binary splits) (Hastie et al. 2005) with those of boosting (an adaptive method for combining many simple models into a combined model with improved predictive performance) (Friedman 2002).

Tree-based models partition the environmental space into rectangles using a series of rules to identify regions having the most homogenous responses to the predictors. Then the mean response for observations in each region is fit. A tree is built by recursive binary splits, i.e. the two subsets resulting from each split are again each split into two subsets. These subsets are described in terms of their homogeneity in the response variable (reduction in deviance is one measure of homogeneity) (Elith et al. 2008, Franklin 2009). It is recommended to grow a large tree and then 'prune' it , i.e. remove the splits that add the least to overall subgroup homogeneity (Hastie et al. 2009).

Boosting is a forward stagewise procedure for improving model accuracy and works by repeatedly sampling the data with replacement and developing trees for each dataset. Each observation sampled is weighted to have a higher probability of selection if it is modeled poorly by the existing collection of trees. The final BRT model thus consists of numerous simple trees that can be understood as an additive regression model (Elith et al. 2008).

The model-building process performs best if it gradually improves the predictive performance of the model, and the contribution of each tree is hence shrunk by a learning rate that is less than one. The tree complexity controls whether interactions are fitted. The learning rate and the tree complexity determine together the number of trees required for optimal prediction.

In BRT, the relative contribution of the environmental variables is based on the number of times the variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees (Friedman and Meulman 2003). The fitted functions are visualized in partial dependence plots showing the effect of a variable on the response.

5. Main findings

5.1. The frameworks

In this thesis a set of presence observations is defined as sampling biased if the frequency distribution of presence observations along major environmental complex-gradients deviates from the frequency distribution of the modeled target's true presence in the environmental space. Based on this definition, we provide two frameworks; one for assessment of sampling bias in presence-only datasets and one for assessment of effects of sampling bias on species distribution models.

According to the definition of sampling bias above, a set of presence observations for a species in a study area contains sampling bias if the species frequency-of-observed-presence (FOP) curve deviates from its frequency-of-true-presence (FTP) curve. By definition, the FTP curve requires presence/absence information for all observation units in the study area. Because this information is rarely available, FTP curves can, in almost all cases, not be generated. The second-best option is to estimate empirical frequency-of-presence (EFP) curves, using an independently collected sample of presence/absence data (Edwards et al. 2005, Edwards et al. 2006, Veloz 2009, Edvardsen et al. 2011, Erikstad et al. 2013). In cases where presence/absence data are lacking altogether we propose to use theoretical-frequency-of-presence (TFP) curves, based on ecological theory, as a reference with which FOP curves are compared (cf. Vaughan and Ormerod 2003). Deviations between FOP and TFP curves may indicate sampling bias. Local minima and maxima on the FOP curve may, however, also represent real properties of the species in the study area and the extent to which strong cases for or against sampling bias can be made will depend on the quality of the information that forms the basis for the TFP curve.

In the bias *effects* framework effects of bias can be visualized by plotting the predicted-relative-frequency-of-presence (PRFP) curve. PRFP curves should reproduce the FOP curve with a degree of detail that matches the level of generalization required by the purpose of the study (Halvorsen 2012), regardless of the FOP curve being realistic or not.

In paper II the two frameworks are tested on nine different species differing with regards to taxonomic affiliation and distribution patterns. We find that comparisons between FOP- and TFP curves indeed show indications of sampling bias in presence-only data. The largest deviations between the two curves are seen in species with wide distributions. For the species with narrower distributions FOP curves are unimodal, yet not entirely smooth.

The PRFP curves for the nine study species did, in most cases, resemble the FOP curves closely. This was true for both ERM- and SPM models, also in cases where FOP curves were complex and ecologically unrealistic. This shows that MaxEnt models in general are efficient at reproducing the FOP curve patterns, and that data quality is more limiting than modeling method for the species distribution modeling output.

5.2. Choice of background/pseudo-absence

The topic of background / pseudo-absence was treated both in paper II and paper III. As paper III was the first to be written, the frameworks had yet to be developed, so the main means of comparison between different models was AUC. We found that sampling design influenced strongly the predictive performance of the models. The models based on randomly selected and fixed-grid pseudo-absence observations were very similar with respect to AUC values, while models based on the BTG design obtained significantly lower AUC values than the models based on random and fixed pseudo-absence observations. Varying the pseudo-absence design also influenced the relative importance of the environmental variables, which was similar in the fixed grid models and the random design models, but different in the BTG design models. In the latter the oceanicity-continentality variable generally scored higher than the temperature variable, while the reverse was true for the random and fixed designs.

In paper II we applied the framework for assessment of sampling bias in presence-only datasets to evaluate which of the two background designs, uninformed background (UB) and background target group, led to the most realistic models. We found that the FOP curves calculated by use of BTG background differed more or less strongly, and in a seemingly unpredictable manner, from FOP curves calculated by use of UB background.

Results from paper II and III show that the use of BTG background, which has been launched as a means to mitigate effects of sampling bias (Phillips and Dudík 2008, Phillips et al. 2009), may, in some cases, actually introduce bias or increase already existing bias in the data.

5.3. Sample size of presences and background points

In the last paper of this thesis the critical sample size sufficient for generating a nonrandom prediction of species distribution was assessed by generating MaxEnt models for datasets of different sizes for 30 insect species. Models based on replicated random subsamples were compared to reference models, based on the full datasets, using the index of vector similarity (*IVS*). Nonrandom models were in most cases obtained from

datasets even with very few presences; for 97% of species clearly nonrandom models were made with less than 15 presences. Critical sample sizes were not found to be significantly correlated with the size of the full dataset, distributional range or taxonomic affiliation.

In paper III we found that the number of pseudo-absences had minimal effect on the predictive performance of the models. The number of generated pseudo-absences did, however, have a direct effect on the predicted probability. A low number of pseudo-absence observations led to a modeled distribution with a relatively large area of high relative predicted probability of presence, whereas the opposite was true when the number of pseudo-absence observation was large. However, when the prevalence (i.e. the number of presence observations as a fraction of all presence plus all pseudo-absence observations, see Jiménez-Valverde et al. 2009) was used as a threshold for converting the predicted degree of presence into a black and white presence–absence map, the area with predicted presence was very similar for a wide range of prevalence values.

From the results from paper III and IV I draw the conclusion that sample size of presences or background/pseudo-absences are not as important as the quality of the data.

5.4. Model complexity

The effect of model complexity was studied in paper II. The PRFP curves produced in paper II did, in most cases, resemble the shapes of the FOP curves, both in the case of simple ERM models and more complex SPM models. The number of derived predictors in SPM models was significantly higher than in ERM models. This did, however, not result in significantly better models, as AUC values for SPM models were not significantly higher than AUC values for the ERM models.

6. Discussion

6.1. Data quality

6.1.1. Assessment of sampling bias in presence data (affecting the data model)

Data quality may be suboptimal for several reasons, including geo-referencing uncertainties (Graham et al. 2008), misidentification of species (Lozier et al. 2009) and sampling bias (Kadmon et al. 2003, Loiselle et al. 2008, Phillips et al. 2009, Yackulic et al. 2013). The focus in paper I and II is on the latter.

Methods and approaches proposed for detecting and quantifying sampling bias fall into two groups: those that address geographical sampling bias and those that address environmental sampling bias. Many studies addressing geographical bias do not investigate whether or not the observed geographical bias results in environmental bias (Hijmans et al. 2000, Veloz 2009, Costa et al. 2010). Furthermore, among studies in which the relationship between the geographical and environmental bias is addressed, a general consensus about how these two, conceptually different biases, relate is lacking (Kadmon et al. 2004, Loiselle et al. 2008, Wolmarans et al. 2010, McCarthy et al. 2011, Merckx et al. 2011). Geographical sampling bias is addressed, e.g., by comparing the distribution of distances of presence observations to specific geographic features with the distances of a set of random points to the same features (Hijmans et al. 2000, Reddy and Dávalos 2003). Features tested for being sources of geographical bias include roads, rivers, cities and 'biodiversity hotspots', the latter approached by comparing the number of observed species presences inside proposed conservation priority areas with the number of observed presences in adjacent non-priority areas (Reddy and Dávalos 2003). Such comparisons rest on the assumption that a species true prevalence does not differ between the compared areas. Furthermore, for geographical sampling bias to be of relevance for assessment of sampling bias (response bias) as defined above, the assumption that geographical bias translates into environmental bias has to be justified. Several direct approaches to assessment of bias in environmental space are proposed. Kadmon et al. (2004) and Loiselle et al. (2008) apply the Kolmogorov-Smirnov test (Massey 1951) to evaluate the difference between the frequency distributions for binned climatic variables in a set of localities where observations (collections) have been made and a random set of localities, taking differences as indications of sampling bias. None of these approaches can, however, be used to assess sampling bias as defined above because they are based upon assumptions about the distribution of the frequency-of-true-presence of the modeled species in environmental space. These assumptions cannot be substantiated by use of presence-only data when the distribution of sampling effort is unknown. Because SDM takes place in environmental space we define a presence

dataset to contain sampling bias if the frequency distribution of observed presence along major environmental complex-gradients deviates from the frequency distribution of its true presence in the environmental space (paper I). Our definition of sampling bias establishes one and only one standard for an unbiased sample: *the true distribution of frequency of presence of the modeled species in environmental space.* From this follows that the key to understanding sampling bias and to finding indications of sampling bias in presence data is to explore the FOP curve of the species.

By examining the FOP curves of nine insect species with respect to the oceanicity-continentality gradient (paper II) we find examples of under- and oversampling well within the tolerance limits of a species (bias type 1 and 2 from paper I), as well as unsystematic variation or stochasticity (bias type 4 from paper I). Stochasticity is likely due to the amount of presences not being large enough to adequately recapture the true response of the species to the gradient. From the definition of sampling bias from paper I we conclude that such irregularities, even if they are minor, must be interpreted as sampling bias, as they give rise to erroneous estimates of frequency of presence along the gradient. With respect to the temperature gradient, all FOP curves are right-truncated. This is likely to result from the species having their optimum outside the sampled portion of the gradient. The truncated response to the temperature gradient may also be strengthened as a result of over-sampling in the high temperature extreme of the gradient. The warmest (south-eastern) parts of Norway contain the most densely populated areas and sampling intensity is particularly high there.

It is important to keep in mind that it is impossible to set up a concrete parameterization of theoretical frequency-of-presence (TFP) curves, with which to compare the FOP curves. Consequently it is impossible to make a formal test of the extent to which a FOP curve deviates from the corresponding TFP curve. The TFP curves are deduced from gradient analytic and macro-ecological biogeographic theory, which again is build up from numerous empirical studies. Sampling bias in data used for SDM can only be *verified* if an independent presence-absence test dataset is applied to generate an EFP curve, with which the FOP curve can be compared. Nevertheless, results obtained in paper I and II show that the proposed framework for assessment of sampling bias in presence-only data by visual comparison of frequency-of-observed-presence (FOP) curves with theoretical frequency-of-presence (TFP) curves can provide useful *indications* of systematic sampling bias.

6.1.2. Choice of background dataset (affecting the data model)

The choice of background data is heavily debated in SDM literature and several background designs have been proposed. Examples are random selection (Stockwell and

Peters 1999), random selection with geographic-weighted exclusion (Hirzel et al. 2001, Dudık et al. 2005), random selection with environmental-weighted exclusion (Zaniewski et al. 2002), use of locations that have been visited where the target species was not observed (Elith and Leathwick 2007), and use of observations for the group of species to which the target species belongs (Ponder et al. 2001, Phillips and Dudík 2008). It is well-known that the choice of background data influences the outcome of the SDM exercise (Thuiller et al. 2004, Chefaoui and Lobo 2008, VanDerWal et al. 2009, Lobo et al. 2010). A clear guide for the selection of background data has, however, not been made available. This thesis offers a tool for the selection of appropriate background data, using species frequency-of-observed presence curves.

In paper II we interpret the observed discrepancies between FOP-UB curves and TFP curves for all species with respect to the main complex-gradients to be caused mainly by sampling bias. We therefor examine whether the use of BTG reduces or removes the discrepancy between FOP and TFP curves. If the BTG data are biased to the same degree as the presence data (Phillips et al. 2009), the FOP-BTG curves should approximate FTP better than FOP-UB curves. However, we find that FOP-BTG curves deviate even more strongly from the expected TFP curves than FOP-UB curves and in all cases, in which a unimodal FOP curve is found for the UB approach, the unimodal pattern weakens, or disappears entirely, when applying BTG data. One clear example is *R. mordax*, for which FOP-curve shapes change from unimodal (truncated unimodal along the temperature gradient) to a monotonous shape without any trace of unimodality left, for both gradients. This is not surprising, considering that the FOP-BTG shows the relationship between the ratios of the frequency by which the focal species is recorded and the frequency by which species of the BTG is recorded, as a function of position along the underlying complex-gradient. Taxonomically related species often have similar ecological requirements and this will result in a monotonous shaped FOP curve, as seen for *R. mordax*.

Our results show that FOP-BTG curves are not more similar to TFP curves than FOP-UB curves, and hence apparently do not reduce sampling bias in Norwegian insect data. This contrasts the common view that applying BTG data is an effective way of mitigating sampling bias in presence-only data used for DM by group discriminative methods (Phillips and Dudík 2008, Phillips et al. 2009, Mateo et al. 2010a, Bystriakova et al. 2012, Yackulic et al. 2013). The idea of using BTG in distribution modeling, as a means of correcting for imperfect sampling was introduced by Ponder et al. (2001). It has since then been used by some authors (Lütolf et al. 2006, Elith and Leathwick 2007), but it was not until Phillips and Dudík (2008) introduced the use of BTG as a new extension in the software Maxent, that BTG really gained a reputation of being an effective way of mitigating effects of sampling

bias in SDM. In many studies BTG is, however, applied, without exploring the data prior to modeling and without testing other background designs for comparison (Cord and Rödder 2011, Graham et al. 2011, Crall et al. 2013, Lohmann et al. 2013). Furthermore, in several studies where different background designs are compared, models generated with BTG data are found to be poorer than those generated by other background designs (Heibl and Renner 2012, Millar and Blouin-Demers 2012). Bystriakova et al. (2012), on the other hand, find BTG models to perform better than UB models when trained with species data from western and central Europe and tested on species data from Ukraine. To our knowledge, no studies have yet addressed the BTG approach from the perspective offered by comparison between response curves generated with UB or BTG data. Our results suggest that this issue should be carefully addressed in future studies and that use of BTG for DM should be based upon careful examination of the two types of FOP curves.

The effects of using different background designs in SDM will also be apparent on the modeling results. In paper II values of AUC drop significantly from the UB approach to the BTG approach. Differences in AUC are also observed in paper III, where the random- and the fixed designs produce almost identical results, while use of the BTG design results in smaller AUCs. However, AUC may be misleading for measuring the accuracy of distribution models for several reasons, one of which being that the extent of the study area and the environmental distance between the presences and absences, to a large degree, determines the AUC value (Lobo et al. 2008). AUC is measuring how efficient the model is at discriminating presences from (pseudo-)absences. With the BTG design the environmental range of the pseudo-absences is typically much smaller and less segregated from the environmental range of the presences than in the fixed and random design. This causes AUC to be lower, when the BTG design is used, for all species in papers II and III. In paper II the poor discriminatory power of the BTG models and hence the drop in AUC values are reflected also in the FOP curves; in many cases there are no apparent relationship between the species' presences and the environmental variables and in some cases either one of the variables do not contribute at all to explaining the distribution. Thus, I conclude that the drop in AUC is not merely an artefact due to the study area becoming smaller, but reflects an actual reduced ability of the BTG models to reproduce the FTP of the species. Another factor influencing AUC is whether AUC is calculated using the training dataset or an independent test dataset. In almost all cases AUC will be lower when calculated with an independent test dataset. To be able to properly compare AUC values obtained with UB data with AUC values obtained with BTG data an independent test dataset should be applied.

A change in the relative contribution of the environmental variables is also observed in paper III when changing the design of the pseudo-absence observations and, notably, the range of environmental variation spanned by presence and pseudo-absence observations. These findings correspond to the results of VanDerWal et al. (2009) who find that fine-scaled environmental predictors are suppressed and coarsely varying climate factors become dominant when the environmental space for sampling pseudo-absences is enlarged. Although the contribution of each environmental variable is not quantified in paper II, the effect of varying the background design is visualized by observing the changes in FOP and PRFP curves in relation to the temperature gradient. With the BTG approach large parts of the gradients left side are excluded. The BTG-FOP- and BTG-PRFP curves with respect to the temperature gradient are hence closer to uniform than the UB-FOP- and UB-PRFP curves. Accordingly temperature contributes less to explaining the distribution of the species in the BTG approach.

Figure 7. (a) *Sicus ferrugineus* (paper IV) and (b) *Rhagium mordax* (paper II and IV) (Photo: Karsten Sund).

6.1.3. The assessment of effects of sampling bias on species distribution models (the statistical model)

Paper II and III show that the choice of background data (paper II) and pseudo-absence data (paper III) strongly influence the outcome of the modeling exercise, in terms of AUC (paper II and III), the shape of the FOP curve, and hence the PRFP curve (paper II), the relative importance of environmental variables (paper III) and the size of the area with high predicted degree of presence. In paper III, the random and fixed grid sampling designs always give similar results, while the results obtained by BTG sampling design are deviating.

It is important that effects of sampling bias and effects of shortcomings in the modeling method are kept separate (Papers I and II). A group-discriminative distribution modeling method should be considered as good if it reproduces a FOP curve as detailed as required by the purpose of the study (Halvorsen 2012), regardless of the FOP curve being realistic or not. The fact that most of the differences between model results in paper II are observed between models trained with different background (or pseudo-absence) data and not between models parameterized in different ways, shows that the data quality, rather than the modeling method, is the limiting factor for SDM performance.

One means for mitigating sampling bias is to avoid overfitting the model to the data. The extent to which a model is overfit is, however, related to the modeling purpose. An SPM model is overfit when it is closely fit to irregularities in the training data (resulting from sampling bias), but not when fit to 'irregularities' (compared to theoretical response curves) caused by factors such as biotic interactions, metapopulation dynamics or the influence by complex-gradients other than those used for modelling. In practice, it is more or less impossible to separate these causes of irregularities by their appearance on FOP curves. An ERM model, on the other hand, is overfit when it reflects patterns other than the response of the modelled target to anything else than the environmental variables (complex-gradients) applied in the modeling. Halvorsen (2012) recognizes three types of overfitting: (1) that a more complex model has lower predictive performance on independent data than a simpler model; (2) that a more complex model is similar in predictive performance on independent data than a simpler model; and (3) that a more complex model with higher predictive performance on independent data than a simpler model fails to fit realistic overall ecological response curves. ERM models are considered to be overfit by all three types of overfitting, while SPM models are, however, only considered to be overfit in the first case. This leads to the conclusion that overfitting is much easier to mitigate in an ERM model. From inspection of FOP patterns it can, however, not be decided *with absolute certainty* whether sampling bias is present or whether deviant curves represent real properties of the modeled species, even when the modeling purpose is ERM.

The PRFP curves produced in paper II do, in most cases, resemble the shapes of the FOP curves, both from the ERM and SPM perspectives. For the species with the most distinctly unimodal FOP curves, the MaxEnt models predict unimodal responses to the bioclimatic gradients, in which unsystematic variation is effectively smoothed out. For the species with FOP curves showing clearer examples of under- or over-sampling, these are in most cases reflected also in the MaxEnt PRFP curves. SPM models are more closely fit to the data than ERM models only for a few species (judged by visual inspection of the FOP and PRFP curves).

The largest difference between ERM- and SPM models is the number of derived predictor variables used to parameterize the models. For SPM models, this number is significantly higher than for ERM models, while AUC values are not significantly higher for SPM models than for ERM models. Complex SPM models often result from inclusion of many threshold- and hinge-type variables (Phillips and Dudík 2008) in order to make PRFP curves fit FOP curves more closely. Nevertheless, the PRFP curves for the two models are often very similar. This shows that simple models with very few parameters (variables) can be as good as more complex models in representing (relatively) general features of FOP curves. We hence support previous advises against building very complicated SPM models, which seem neither to assist understanding of the focal species' relationship to the environment nor to enhance predictive power (Anderson and Gonzalez 2011, Halvorsen 2013, Syfert et al. 2013). It is also important to keep in mind that not all modeling methods are appropriate choices both for ERM and SPM. MaxEnt and BRT are listed as two of the best in several papers comparing the predicitive ability of SDM methods (Elith et al. 2006, Hirzel et al. 2006, Franklin 2009). These comparisons do, however, refer to models fit with the SPM purpose of modeling, as only SPM models can be evaluated by predictive performance in geographical space. An ERM model on the other hand must be judged by its ability to express the overall ecological response of the modeled target to the selected environmental variables in ecological space (Halvorsen 2012). This is usually done by fitting simple and smooth functions and can be accomplished using methods like generalized linear models (GLM), generalized additive models (GAM) or other functions by maximum likelihood methods (Austin et al. 1994, Oksanen 1997, Jansen & Oksanen 2013). ERM models have to be explicitly parameterized, i.e., the relationship between predictor and response variables has to be given by a parameterized mathematical function (Austin et al. 1994, Oksanen 1997, Jansen and Oksanen 2013). Machine-learning methods like boosted regression trees (BRT), and other ensemble forecasting methods (e.g., BIOMOD), which are not explicitly parameterized, are therefore not appropriate for ERM. MaxEnt, which has been described alternatively as a machine-learning method (Phillips 2008,

Phillips and Dudík 2008) and as a statistical learning method (Elith et al. 2011), but which can alternatively be explained by strict application of the maximum likelihood principle (Halvorsen 2013), is therefore appropriate both for ERM and SPM (Halvorsen 2012).

6.2. How does the data quantity affect the species distribution models?

6.2.1. Presence datasets

Scarcity of presence data is a major obstacle for modeling species distributions (Lim et al. 2002, Papeş and Gaubert 2007, Feeley and Silman 2011b, Feeley and Silman 2011a, Kamino et al. 2012). The effect of sample size on species distribution models and the existence of a minimum presence sample size needed to generate reliable models are widely debated. There exists a general agreement that the models' accuracy increases when sample size increases (Cumming 2000, Pearce and Ferrier 2000, Stockwell and Peterson 2002, Reese et al. 2005, Hernandez et al. 2006, Wisz et al. 2008), but a general consensus regarding the amount of presence data needed to reliably predict a species' distribution is lacking. Recommendations span from 5–10 observations (Hernandez et al. 2006, Pearson et al. 2007) via 10–30 (Stockwell and Peterson 2002, Pearson et al. 2007, Wisz et al. 2008, Mateo et al. 2010b) to more than 200 (Hanberry et al. 2012) presence observations being needed. The contrasting conclusions from these studies may result from differences with respect to characteristics of the study areas, resolution and extent of the study area (Loe et al. 2012), modeling method (Dupin et al. 2011) and environmental variables used in the modeling (Syphard and Franklin 2009), as well as from different criteria for determining what constitutes an acceptable distribution model (Hanberry et al. 2012).

Differences in species characteristics, such as distribution patterns, are reported to be of importance for the predictability of species' distributions by DM methods (Hernandez et al. 2006, Guisan et al. 2007, Mateo et al. 2010b, Stokland et al. 2011). In these studies specialist species, i.e., species with a narrow distribution in environmental variables space and restricted distributions in geographical space, are found to be easier targets for SDM than generalist species, i.e., species with a broader distribution in the two conceptual spaces (geographical and environmental). Accordingly, fewer presences are needed to obtain acceptable distribution models for the former. The argument underpinning this view is that generalists have broader ecological requirements and that more presences are needed to represent the entire range of suitable environmental conditions for such species.

In paper IV we assess the minimum number of presences required to obtain non-random distribution models. In order to make non-arbitrary and verifiable conclusions about an eventual critical sample size (CSS), we determine thresholds based on

comparisons with randomly generated models. The idea behind this choice is that a model performing better than random (i.e., the model is more similar to the reference model than a randomly generated model) provides useful information about the modeled species.

We find that models based on subsamples become more and more similar to the corresponding reference models as presence sample size increases; however, nonrandom models are in most cases obtained from datasets with very few presences. For the 30 species there are no significant correlations between CSS and the species' relative ecological range, the number of presences in the full dataset or taxonomic order. This suggests that our generally low CSS estimate is a robust result, of general validity.

The results from paper IV should be understood in the light of the knowledge gained from the two papers discussing sampling bias (papers I and II). Removing presences from a dataset resembles the situation of sampling bias in the sense that stochasticity (sampling bias type 4 from paper I) may be introduced. This is because every FOP value is calculated from a smaller number of presences and is hence infested with an increased degree of uncertainty. The low CSS obtained in paper IV confirm that MaxEnt models, even if they are complex, are able to effectively smooth out stochasticity in the data.

6.2.2. Pseudo-absence/background datasets

In paper III we show that the number of pseudo-absence observations has a large effect on the relative predicted probability of presence, which in turn has a strong effect on the resulting binary distribution map. When following the advice of Cramer (1999) to set the threshold equal to the prevalence, however, the distribution maps generated by different amounts of pseudo-absences, in most cases, become very similar.

Although the main aims of paper II concern sampling bias, the results also opens up for a discussion on background dataset sample size. Larger datasets lead to more robust estimates of FOP and reduce stochasticity bias. For some of the species the BTG datasets are quite small. To our knowledge, the number of observations needed to form a robust BTG dataset has so far not been addressed. Stokland et al. (2011) do, however, find that the effect of varying the number of pseudo-absence observations in BRT (boosted regression trees) models from 64 to over 4000 records is small compared to that of sampling design and properties of the focused species. Moreover, Mateo et al. (2010a) claim that SDMs may be improved by use of BTG datasets consisting of as few as 15 observations. Our results, as well as theoretical reasoning, give reasons to seriously question this claim. However, the problem attached to BTG is not primarily the size of the dataset, but rather the distribution of the BTG observations along the gradient. With smaller datasets, more stochasticity bias will be introduced.

6.3. Potential limitations

6.3.1. The lack of an independent test dataset

In paper I we emphasize that the eventual existence of sampling bias cannot normally be conclusively confirmed without access to a reliable test dataset of independent presence/absence data for the species in the study area. A limitation of all the three subsequent papers of this thesis (papers II, III and IV) is the lack of an independent test dataset. Irregularities in the FOP curves may not always indicate sampling bias (as discussed in paper I) and the final answer to the question if such irregularities result from sampling bias or other sources, such as biotic interactions, the absence of species in habitable sites or the presence of sink populations in unhabitable sites, are likely to be found by evaluating the model on an independent dataset. Such data makes possible generating an empirical frequency-of-presence (EFP) curve, as outlined in paper I. I therefore, in all possible cases, strongly recommend evaluating the model using an independent test dataset. Nevertheless, even though irregularities in the FOP curve do not provide indisputable evidence for bias in presence datasets, a relatively smooth and unimodal FOP curve is a useful *indication* that a sample of presences for the modelled target is *un*biased.

In paper III, the effects of combinations of pseudo-absence designs and sample sizes on SDM results are compared. Many of the conclusions in paper III are based on AUC values. AUC is here used to measure the ability of a model to correctly predict presences and to predict absence for pseudo-absences used to train the model. If an independent dataset had been applied for model evaluation, AUC would most likely have been lower. Given that the BTG approach generated realistic models, it is also possible, in theory, that the difference between AUC values for the BTG design and the random and fixed designs would have been levelled out, due to more false absences, and hence a decrease in AUC, in the random and fixed designs.

In paper IV the models based on the full datasets for the species are used as reference models, with which models based on subsamples are compared. As no independent datasets are available for evaluating these reference models, we cannot claim that they represent the true distributions of the species. This study therefore shows how the predicted distribution changes with decreasing presence sample size, not in relation to the true distribution, but in relation to a reference distribution, which may or may not be imperfectly sampled.

6.3.2. The inclusion of only two environmental variables

It can be argued that a limitation of the thesis is that only two environmental variables, oceanicity-continentality and temperature, are applied in the modeling exercises of paper II and IV. However, gradient analytic theory predicts that species do not respond to single environmental gradients, but to environmental complex-gradients, i.e., sets of correlated environmental variables (Whittaker 1967). Furthermore, only a few major complex-gradients normally account for most of the variation in species' composition that can be explained environmentally. The step-less oceanicity-continentality gradient and the step-less temperature gradient summarize the co-variation of several topographical, hydrological, geological and climatic variables (Bakkestuen et al. 2008) and correspond to the two main bioclimatic gradients used in expert classifications of Norway into biogeographical regions: vegetation sections and vegetation zones (Moen 1999).

In paper III a candidate set of 12 environmental predictor variables is used as the starting point for variable selection, of which six of these are selected (by the criterion that no predictor should have a Pearson's product-moment correlation coefficient higher than 0.7 with any other selected predictor). Of these six, only two or three variables are found to contribute strongly (about 80% of the variation explained by all variables) to the resulting models. The oceanicity-continentality and the temperature gradients themselves, or variables strongly related to these, were always among the strongly contributing variables. This indicates that the oceanicity-continentality and temperature gradients are the most appropriate choices for a reduced set of environmental predictor variables for distribution modeling at a fine regional scale in Norway.

In paper II we use data for variation at the fine regional scale (1 km x 1 km) to judge the performance of the bias assessment and bias effects frameworks proposed in paper I. Performance of the frameworks should also be judged on broader (coarse regional–global) and finer (local–micro) scales. On a broader scales, distributions are expected to be limited by geophysical processes such as (historical) continental plate movements, sea-level changes, mountain-chain upfoldings, and glacial cycles, over very wide time spans (Willis and Whittaker 2002). On finer scales, as outlined in the theory chapter, the influence of biotic interactions and metapopulation dynamics on observed distributions are expected to increase. A shift from unimodal to polymodal or irregular curve shapes *may* occur when these factors are important, but a decrease or increase in the tolerance of a species, or a decrease or increase in the abundance of the species within the species' tolerance limits, is expected to be observed more commonly, even in these cases (Halvorsen 2012).

7. Perspectives and conclusive remarks

In paper I we define a set of presence observations for a modeled target to contain sampling bias if the frequency distribution of presence observations along major environmental complex-gradients deviates from the frequency distribution of the modeled target's true presence in the environmental space. From this I conclude that small presence datasets (paper IV), the lack of absence data (paper III) and the choice of pseudo-absence/ background data (paper II and III) all relate to the issue of sampling bias and the distribution of frequency-of-presence along environmental gradients (paper I). One or both of the presence- and background/pseudo-absence datasets will be biased if these two, together, fail to reflect the true frequency ratio of the model target as a function of position along the gradient. I therefore recommend beginning the modeling exercise by plotting the FOP along major complex gradients of importance to the species. If the presence datasets are too small and/or biased, or if the selected background is not suitable for revealing the species-environment relationship, this will be evident by the FOP curve deviating from the expected smooth and unimodal shape.

The comparison between theoretical response curves and species' FOP curves is found to be a promising procedure for encountering indications of sampling bias in presence data in paper II. Interestingly the BTG approach, which is applied in SDM as a way of correcting for sampling bias, is found to give rise to complex, often ecological meaningless FOP curves, essentially modeling the relationship between ecological conditions found in sites where the focal species has been sampled and the ecological conditions in sites where taxonomically related species have been sampled (paper II). The BTG approach is also addressed in paper III, where it was contrasted against a fixed background design and a random background design. While the random and fixed designs produce almost identical model results, the BTG models deviate from the other two. In paper III we conclude that when the main purpose of a study is to produce broad-scale distribution maps, pseudo-absences should include environmental conditions from areas where a species does not occur to obtain complete map coverage. When the purpose is to investigate effects of environmental conditions within the distribution area, however, one should not include pseudo-absences from environmental conditions far outside the ecological tolerance of a species; as environmental factors working on a broader scale will tend to mask the effect of other, more local ecological factors. A prerequisite in the last scenario is that the species tolerance limits along the gradient are contained in the area covered by pseudo-absences (or background data). If not the FOP curve may become monotonous and non-informative.

In addition to addressing sampling bias and choice of background for SDM, this thesis addresses the presence sample size question. Paper IV shows that nonrandom models in most cases are obtained with very few presences available. Only 3 out of the 30 modeled species require more than 10 presences to obtain a nonrandom model, and more than 15 presences are required for only 1 out of 30 species. Nonrandom models can be interpreted as 'useful' in the sense that they add to the information available prior to modeling and, accordingly, may serve as valuable starting points for further studies of poorly known species with few known presence records. Generally, with small presence (or pseudo-absence/background) datasets, a high occurrence of the fourth type of sampling bias; unsystematic variation or stochasticity (paper I) is expected.

This thesis indicates that SDM can be useful for practical purposes. In paper IV we demonstrate that although predictions get poorer with smaller presence sample sizes, useful predictions can be obtained with very few presence observations. Moreover, paper I and II outline a procedure for evaluating whether or not presence and pseudo-absence/ background datasets contain sampling bias and further, if they are suitable for generating informative species distribution models. In a practical SDM perspective it is essential to ensure that the modeling exercise does not result in suboptimal models. I recommend plotting the FOP curve prior to modeling. Moreover, test datasets, when available, should be used to evaluate the models. In paper IV we briefly address whether SDM can be used for assessment of red-list status. I conclude that SDM should not be used alone as an assessment method. The distribution is often conditioned on other factors than availability of suitable environment. These factors are rarely known and can hence not be accounted for in the modeling process. Nevertheless, I do consider SDM to be a useful indirect aid in identifying sites suitable for the species in question and hence guide further fieldwork. It may also be useful for other management purposes, such as predicting the distribution of invasive species, predicting species distributions in a future climate regime or finding habitats suitable for the reintroduction of species. For these purposes it is important to keep in mind that a model transferable in space and time is needed. Such models should not be too closely fit to idiosyncrasies in the dataset caused by sampling bias, idiosyncrasies specific to the study area etc. Papers I and II demonstrate that, in some cases, complex models overfit to irregularities in the presence data, and that simpler models fit the more general responses to the environmental variables, without significantly loosing predictive power in terms of AUC. These are the models required for projecting into other areas or future climates.

The development of SDM methods has opened up for numerous possibilities within applied ecology and conservation biology. These methods have been seen as 'shortcuts'

for inferring past, present and future species distributions, as well as exploring species-environment relationships and dealing with phylogenetic questions with the click of a button. This thesis shows that the shortcut may not be as short as hoped to be. In cases where the data quality is poor or data are sparse the shortcut may turn out to be a detour. A firm foothold in ecological theory, good understanding of how the modeling methods work, how the data are handled, and how models are to be interpreted are all essential for obtaining a good understanding of when these shortcomings are obstacles for making reliable distribution models.

I thank Rune Halvorsen, Eirik Rindal, Trine Bekkby, Sabrina Mazzoni, Niklaus E. Zimmerman and Alberto Jimenéz-Valverde for helpful comments on the introduction of this thesis. I would also like to thank Eirik Rindal for valuable help formatting the final manuscript.

8. References

Anderson, R. P., and I. J. Gonzalez. 2011. Species-specific tuning increases robustness to sampling bias in models of species distributions: an implementation with Maxent. Ecological Modelling **222**:2796–2811.

Anderson, R. P., A. T. Peterson, and M. Gómez-Laverde. 2002. Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. Oikos **98**:3–16.

Araújo, M. B., and A. Guisan. 2006. Five (or so) challenges for species distribution modelling. Journal of Biogeography **33**:1677–1688.

Araújo, M. B., and M. Luoto. 2007. The importance of biotic interactions for modelling species distributions under climate change. Global Ecology and Biogeography **16**:743–753.

Aune, B. 1993. Nasjonalatlas for Norge: Klima., Statens Kartverk.

Austin, M. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling **157**:101–118.

Austin, M. 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecological Modelling **200**:1–19.

Austin, M., A. Nicholls, M. Doherty, and J. Meyers. 1994. Determining species response functions to an environmental gradient by means of a β-function. Journal of Vegetation Science **5**:215–228.

Austin, M., and T. Smith. 1989. A new model for the continuum concept. Vegetatio **83**:35–47.

Austin, M. P., and M. J. Gaywood. 1994. Current problems of environmental gradients and species response curves in relation to continuum theory. Journal of Vegetation Science **5**:473–482.

Austin, M. P., and K. P. Van Niel. 2011. Improving species distribution models for climate change studies: variable selection and scale. Journal of Biogeography **38**:1–8.

Bakkestuen, V., L. Erikstad, and R. Halvorsen. 2008. Step-less models for regional environmental variation in Norway. Journal of Biogeography **35**:1906–1922.

Barbet-Massin, M., F. Jiguet, C. H. Albert, and W. Thuiller. 2012. Selecting pseudo-absences for species distribution models: how, where and how many? Methods in Ecology and Evolution **3**:327–338.

Barbosa, A. M. 2009. Transferability of environmental favourability models in geographic space: the case of the Iberian desman (*Galemys pyrenaicus*) in Portugal and Spain. Ecological Modelling **220**:747–754.

Barney, J. N., and J. M. DiTomaso. 2010. Bioclimatic predictions of habitat suitability for the biofuel switchgrass in North America under current and future climate scenarios. Biomass and Bioenergy **34**:124–133.

Barnosky, A. D., N. Matzke, S. Tomiya, G. O. U. Wogan, B. Swartz, T. B. Quental, C. Marshall, J. L. McGuire, E. L. Lindsey, K. C. Maguire, B. Mersey, and E. A. Ferrer. 2011. Has the Earth's sixth mass extinction already arrived? Nature **471**:51–57.

Barve, N., V. Barve, A. Jiménez-Valverde, A. Lira-Noriega, S. P. Maher, A. T. Peterson, J. Soberón, and F. Villalobos. 2011. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. Ecological Modelling **222**:1810–1819.

Bean, W. T., R. Stafford, and J. S. Brashares. 2012. The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. Ecography **35**:250–258.

Begon, M., J. L. Harper, and C. R. Townsend. 1990. Ecology: individuals, populations and communities, 2nd edition. Page 945. Blackwell Scienctific Publications, Boston.

Beldring, S., L. A. Roald, and A. Voksø. 2002. Avrenningskart for Norge. Årsmiddelverdier for avrenning 1961–1990. NVE-Dokument **2**:1–49.

Boakes, E. H., P. J. McGowan, R. A. Fuller, D. Chang-qing, N. E. Clark, K. O'Connor, and G. M. Mace. 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. PLoS biology **8**:e1000385.

Boulangeat, I., D. Gravel, and W. Thuiller. 2012. Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. Ecology Letters **15**:584–593.

Bourg, N. A., W. J. McShea, and D. E. Gill. 2005. Putting a CART before the search: successful habitat prediction for a rare forest herb. Ecology **86**:2793–2804.

Brown, J. H., G. C. Stevens, and D. M. Kaufman. 1996. The geographic range: Size, shape, boundaries, and internal structure. Annual Review of Ecology and Systematics **27**:597–623.

Bystriakova, N., M. Peregrym, R. H. Erkens, O. Bezsmertna, and H. Schneider. 2012. Sampling bias in geographic and environmental space and its effect on the predictive power of species distribution models. Systematics and Biodiversity **10**:305–315.

Cabeza, M., A. Arponen, L. Jäättelä, H. Kujala, A. Van Teeffelen, and I. Hanski. 2010. Conservation planning with insects at three different spatial scales. Ecography **33**:54–63.

Chapman, A. D. 2009. Numbers of living species in Australia and the world. Department of the Environment, Water, Heritage and the Arts, Canberra, Australia.

Chefaoui, R. M., and J. M. Lobo. 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. Ecological Modelling **210**:478–486.

Chown, S. L., J. G. Sørensen, and J. S. Terblanche. 2011. Water loss in insects: An environmental change perspective. Journal of insect physiology **57**:1070–1084.

Collins, S. L., S. M. Glenn, and D. W. Roberts. 1993. The hierarchical continuum concept. Journal of Vegetation Science **4**:149–156.

Cord, A., and D. Rödder. 2011. Inclusion of habitat availability in species distribution models through multi-temporal remote-sensing data? Ecological Applications **21**:3285–3298.

Costa, G. C., C. Nogueira, R. B. Machado, and G. R. Colli. 2010. Sampling bias and the use of ecological niche modeling in conservation planning: a field evaluation in a biodiversity hotspot. Biodiversity Conservation **19**:883–899.

Crall, A. W., C. S. Jarnevich, B. Panke, N. Young, M. Renz, and J. Morisette. 2013. Using habitat suitability models to target invasive plant species surveys. Ecological Applications **23**:60–72.

Cramer, J. S. 1999. Predictive performance of the binary logit model in unbalanced samples. Journal of the Royal Statistical Society: Series D (The Statistician) **48**:85–94.

Cumming, G. S. 2000. Using between-model comparisons to fine-tune linear models of species ranges. Journal of Biogeography **27**:441–455.

Dahl, E., and J. Birks. 1998. The Phytogeography of Northern Europe. Cambridge University Press, Cambridge.

Dudık, M., R. E. Schapire, and S. J. Phillips. 2005. Correcting sample selection bias in maximum entropy density estimation. Advances in neural information processing systems **17**:323–330.

Dupin, M., P. Reynaud, V. Jarošík, R. Baker, S. Brunel, D. Eyre, J. Pergl, and D. Makowski. 2011. Effects of the Training Dataset Characteristics on the Performance of Nine Species Distribution Models: Application to *Diabrotica virgifera virgifera*. PLoS ONE **6**:e20957.

Edvardsen, A., V. Bakkestuen, and R. Halvorsen. 2011. A fine-grained spatial prediction model for the red-listed vascular plant *Scorzonera humilis*. Nordic Journal of Botany **29**:495–504.

Edwards, T. C. J., D. R. Cutler, N. E. Zimmermann, L. Geiser, and J. Alegria. 2005. Model-based stratifications for enhancing the detection of rare ecological events. Ecology **86**:1081–1090.

Edwards, T. C. J., D. R. Cutler, N. E. Zimmermann, L. Geiser, and G. G. Moisen. 2006. Effects of sample survey design on the accuracy of classification tree models in species distribution models. Ecological Modelling **199**:132–141.

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

Elith, J., and J. Leathwick. 2007. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. Diversity and Distributions **13**:265–275.

Elith, J., and J. R. Leathwick. 2009. Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics **40**:677–697.

Elith, J., J. R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. Journal of Animal Ecology **77**:802–813.

Elith, J., S. J. Phillips, T. Hastie, M. Dudík, Y. E. Chee, and C. J. Yates. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions **17**:43–57.

Elton, C. S. 1927. Animal ecology. University of Chicago Press.

Erikstad, L., V. Bakkestuen, T. Bekkby, and R. Halvorsen. 2013. Impact of scale and quality of digital terrain models on predictability of seabed terrain types. Marine Geodesy **36**:2–21.

Evans, J. M., R. J. Fletcher, and J. Alavalapati. 2010. Using species distribution models to identify suitable areas for biofuel feedstock production. GCB Bioenergy **2**:63–78.

Feeley, K. J., and M. R. Silman. 2011a. The data void in modeling current and future distributions of tropical species. Global Change Biology **17**:626–630.

Feeley, K. J., and M. R. Silman. 2011b. Keep collecting: accurate species distribution modelling requires more collections than previously thought. Diversity and Distributions **17**:1132–1140.

Fink, U., and W. Völkl. 1995. The effect of abiotic factors on foraging and oviposition success of the aphid parasitoid, Aphidius rosae. Oecologia **103**:371–378.

Foottit, R. G., and P. H. Adler. 2009. Insect biodiversity: science and society. John Wiley & Sons.

Franklin, J. 2009. Mapping species distributions: spatial inference and prediction. Cambridge University Press, Cambridge.

Friedman, J., T. Hastie, and R. Tibshirani. 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics **28**:337–407.

Friedman, J. H. 2002. Stochastic gradient boosting. Computational Statistics & Data Analysis **38**:367–378.

Friedman, J. H., and J. J. Meulman. 2003. Multiple additive regression trees with application in epidemiology. Statistics in medicine **22**:1365–1381.

Førland, E. J. 1993. Nedbørsnormaler normalperiode 1961–1990. DNMI-rapport Klima **39**:1–63.

Gaston, K. J., T. M. Blackburn, J. J. D. Greenwood, R. D. Gregory, R. M. Quinn, and J. H. Lawton. 2000. Abundance–occupancy relationships. Journal of Applied Ecology **37**:39–59.

Gaston, K. J., and J. H. Lawton. 1988. Patterns in the distribution and abundance of insect populations. Nature **331**:709–712.

Giller, P. S. 1984. Community structure and the niche. Chapman and Hall London.

Golicher, D., A. Ford, L. Cayuela, and A. Newton. 2012. Pseudo-absences, pseudo-models and pseudo-niches: pitfalls of model selection based on the area under the curve. International Journal of Geographical Information Science **26**:2049–2063.

Graham, C. H., J. Elith, R. J. Hijmans, A. Guisan, A. Townsend Peterson, B. A. Loiselle, and G. The NCEAS Predicting Species Distributions Working. 2008. The influence of spatial errors in species occurrence data used in distribution models. Journal of Applied Ecology **45**:239–247.

Graham, J., C. Jarnevich, N. Young, G. Newman, and T. Stohlgren. 2011. How will climate change affect the potential distribution of Eurasian tree sparrows *Passer montanus* in North America? Acta Zool Sin **57**:648–654.

Grinnell, J. 1917. The niche-relationships of the California Thrasher. Auk **34**:427–433.

Guisan, A., O. Broennimann, R. Engler, M. Vust, N. G. Yoccoz, A. Lehmann, and N. E. Zimmermann. 2006. Using niche-based models to improve the sampling of rare species. Conservation biology **20**:501–511.

Guisan, A., and W. Thuiller. 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters **8**:993–1009.

Guisan, A., and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecological Modelling **135**:147–186.

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

Gutiérrez, D., P. Fernández, A. S. Seymour, and D. Jordano. 2005. Habitat distribution models: Are mutualist distributions good predictors of their associates? Ecological Applications **15**:3–18.

Halvorsen, R. 2012. A gradient analytic perspective on distribution modelling. Sommerfeltia **35**.

Halvorsen, R. 2013. A maximum likelihood explanation of MaxEnt, and some implications for distribution modelling. Sommerfeltia **36**.

Hanberry, B. B., H. S. He, and D. C. Dey. 2012. Sample sizes and model comparison metrics for species distribution models. Ecological Modelling **227**:29–33.

Hanski, I. 1982. Dynamics of regional distribution: the core and satellite species hypothesis. Oikos:210–221.

Hanski, I. 1998. Metapopulation dynamics. Nature **396**:41–49.

Hanski, I., A. Moilanen, and M. Gyllenberg. 1996. Minimum viable metapopulation size. American Naturalist:527–541.

Hanski, I., and O. Ovaskainen. 2000. The metapopulation capacity of a fragmented landscape. Nature **404**:755–758.

Hanski, I., and D. Simberloff. 1997. The metapopulation approach, its history, conceptual domain, and application to conservation. Metapopulation biology: ecology, genetics, and evolution:5–26.

Harrison, S. 1991. Local extinction in a metapopulation context: an empirical evaluation. Biological journal of the Linnean Society **42**:73–88.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The elements of statistical learning. Data mining, inference, and prediction. Second Edition. Springer, New York.

Hastie, T., R. Tibshirani, J. Friedman, and J. Franklin. 2005. The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer **27**:83–85.

Hatteland, B. A., S. Roth, A. Andersen, K. Kaasa, B. Støa, and T. Solhøy. 2013. Distribution and spread of the invasive slug *Arion vulgaris* Moquin-Tandon in Norway. Fauna norvegica **32**:13–26.

Hawkins, B. A., R. Field, H. V. Cornell, D. J. Currie, J.-F. Guégan, D. M. Kaufman, J. T. Kerr, G. G. Mittelbach, T. Oberdorff, E. M. O'Brien, E. E. Porter, and J. R. G. Turner. 2003. Energy, water and broad-scale geographic patterns of species richness. Ecology **84**:3105–3117.

Hawkins, B. A., and E. E. Porter. 2003. Does herbivore diversity depend on plant diversity? The case of California butterflies. The American Naturalist **161**:40–49.

Hebblewhite, M., E. Merrill, and T. McDonald. 2005. Spatial decomposition of predation risk using resource selection functions: an example in a wolf–elk predator–prey system. Oikos **111**:101–111.

Heibl, C., and S. S. Renner. 2012. Distribution models and a dated phylogeny for Chilean Oxalis species reveal occupation of new habitats by different lineages, not rapid adaptive radiation. Systematic Biology **61**:823–834.

Heikkinen, R. K., M. Luoto, R. Virkkala, R. G. Pearson, and J.-H. Körber. 2007. Biotic interactions improve prediction of boreal bird distributions at macro-scales. Global Ecology and Biogeography **16**:754–763.

Hernandez, P. A., C. H. Graham, L. L. Master, and D. L. Albert. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography **29**:773–785.

Hijmans, R. J., K. A. Garrett, Z. Huamán, D. P. Zhang, M. Schreuder, and M. Bonierbale. 2000. Assessing the geographic representativeness of genebank collections: the case of Bolivian wild potatoes. Conservation Biology **14**:1755–1765.

Hill, J. K., C. D. Thomas, and B. Huntley. 1999. Climate and habitat availability determine 20th century changes in a butterfly's range margin. Proceedings of the Royal Society of London. Series B: Biological Sciences **266**:1197–1206.

Hirzel, A., V. Helfer, and F. Metral. 2001. Assessing habitat-suitability models with a virtual species. Ecological Modelling **145**:111–121.

Hirzel, A. H., G. Le Lay, V. Helfer, C. Randin, and A. Guisan. 2006. Evaluating the ability of habitat suitability models to predict species presences. Ecological Modelling **199**:142–152.

Holway, D. A., A. V. Suarez, and T. J. Case. 2002. Role of abiotic factors in governing susceptibility to invasion: a test with Argentine ants. Ecology **83**:1610–1619.

Hortal, J., N. Roura-Pascual, N. Sanders, and C. Rahbek. 2010. Understanding (insect) species distributions across spatial scales. Ecography **33**:51.

Huston, M. A. 2002. Introductory essay: critical issues for improving predictions. Predicting species occurrences: issues of accuracy and scale:7–21.

Hutchinson, G. 1957. Concluding remarks.*in* Cold springs harbor symposium in quantitative biology.

Jansen, F., and J. Oksanen. 2013. How to model species responses along ecological gradients–Huisman–Olff–Fresco models revisited. Journal of Vegetation Science **24**:1108–1117.

Jaynes, E. T. 1957a. Information theory and statistical mechanics. Physical Review **106**:620–630.

Jaynes, E. T. 1957b. Information theory and statistical mechanics 2. Physical Review **108**:171–190.

Jiménes-Valverde, A., J. M. Lobo, and J. Hortal. 2008. Not as good as they seem: the importance of concepts in species distribution modelling. Diversity and Distributions **14**:885–890.

Jiménez-Valverde, A., J. Lobo, and J. Hortal. 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. Community Ecology **10**:196–205.

Kadmon, R., O. Farber, and A. Danin. 2003. A systematic analysis of factors affecting the performance of climatic envelope models. Ecological Applications **13**:853–867.

Kadmon, R., O. Farber, and A. Danin. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecological Applications **14**:401–413.

Kamino, L. H. Y., J. R. Stehmann, S. Amaral, P. De Marco, T. F. Rangel, M. F. de Siqueira, R. De Giovanni, and J. Hortal. 2012. Challenges and perspectives for species distribution modelling in the neotropics. Biology Letters **8**:324–326.

Kearney, M. 2006. Habitat, environment and niche: what are we modelling? Oikos **115**:186–191.

Krebs, C. 1994. Ecology: The experimental analysis of distribution and abundance. 4th Edition. Page 801. Harper & Row Publishers. New York.

Kålås, J. A., S. Henriksen, S. Skjelseth, Å. Viken, and (eds.). 2010. Environmental conditions and impacts for Red List species. Norwegian Biodiversity Information Centre, Norway.

Leathwick, J., and M. Austin. 2001. Competitive interactions between tree species in New Zealand's old-growth indigenous forests. Ecology **82**:2560–2573.

Lim, B. K., A. T. Peterson, and M. D. Engstrom. 2002. Robustness of ecological niche modeling algorithms for mammals in Guyana. Biodiversity and Conservation **11**:1237–1246.

Lobo, J. M., A. Jimenez-Valverde, and J. Hortal. 2010. The uncertain nature of absences and their importance in species distribution modelling. Ecography **33**:103–114.

Lobo, J. M., A. Jiménez-Valverde, and R. Real. 2008. AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography **17**:145–151.

Loe, L. E., C. Bonenfant, E. L. Meisingset, and A. Mysterud. 2012. Effects of spatial scale and sample size in GPS-based species distribution models: are the best models trivial for red deer management? European Journal of Wildlife Research **58**:195–203.

Loehle, C. 2012. Relative frequency function models for species distribution modeling. Ecography **35**:487–498.

Lohmann, G., K. Grosfeld, D. Wolf-Gladrow, A. Wegner, J. Notholt, and V. Unnithan. 2013. Ecosystems and climate change. Pages 113–118 Earth system science: Bridging the gaps between disciplines. Springer.

Loiselle, B. A., C. A. Howell, C. H. Graham, J. M. Goerck, T. Brooks, K. G. Smith, and P. H. Williams. 2003. Avoiding pitfalls of using species distribution models in conservation planning. Conservation Biology **17**:1591–1600.

Loiselle, B. A., P. M. Jørgensen, T. Consiglio, I. Jiménez, J. G. Blake, L. G. Lohmann, and O. M. Montiel. 2008. Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? Journal of Biogeography **35**:105–116.

Lozier, J., P. Aniello, and M. Hickerson. 2009. Predicting the distribution of Sasquatch in western North America: anything goes with ecological niche modelling. Journal of Biogeography **36**:1623–1627.

Luoto, M., R. K. Heikkinen, J. Pöyry, and K. Saarinen. 2006. Determinants of the biogeographical distribution of butterflies in boreal regions. Journal of Biogeography **33**:1764–1778.

Lütolf, M., F. Kienast, and A. Guisan. 2006. The ghost of past species occurrence: improving species distribution models for presence-only data. Journal of Applied Ecology **43**:802–815.

Marini, M., M. Barbet-Massin, L. Lopes, and F. Jiguet. 2010. Predicting the occurrence of rare Brazilian birds with species distribution models. Journal of Ornithology **151**:857–866.

Massey, F. J. 1951. The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association **46**:68–78.

Mateo, R. G., T. B. Croat, Á. M. Felicísimo, and J. Muñoz. 2010a. Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absences and target-group absences from natural history collections. Diversity and Distributions **16**:84–94.

Mateo, R. G., A. M. Felicísimo, and J. Muñoz. 2010b. Effects of the number of presences on reliability and stability of MARS species distribution models: the importance of regional niche variation and ecological heterogeneity. Journal of Vegetation Science **21**:908–922.

Matthies, D., I. Bräuer, W. Maibom, and T. Tscharntke. 2004. Population size and the risk of local extinction: empirical evidence from rare plants. Oikos **105**:481–488.

McCarthy, K. P., R. J. Fletcher Jr, C. T. Rota, and R. L. Hutto. 2011. Predicting species distributions from samples collected along roadsides. Conservation Biology **26**:68–77.

McInerny, G. J., and R. S. Etienne. 2012a. Ditch the niche–is the niche a useful concept in ecology or species distribution modelling? Journal of Biogeography **39**:2096–2102.

McInerny, G. J., and R. S. Etienne. 2012b. Pitch the niche – taking responsibility for the concepts we use in ecology and species distribution modelling. Journal of Biogeography **39**:2112–2118.

McInerny, G. J., and R. S. Etienne. 2012c. Stitch the niche–a practical philosophy and visual schematic for the niche concept. Journal of Biogeography **39**:2103–2111.

Meier, E. S., F. Kienast, P. B. Pearman, J.-C. Svenning, W. Thuiller, M. B. Araújo, A. Guisan, and N. E. Zimmermann. 2010. Biotic and abiotic variables show little redundancy in explaining tree species distributions. Ecography **33**:1038–1048.

Merckx, B., M. Steyaert, A. Vanreusel, M. Vincx, and J. Vanaverbeke. 2011. Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. Ecological Modelling **222**:588–597.

Millar, C. S., and G. Blouin-Demers. 2012. Habitat suitability modelling for species at risk is sensitive to algorithm and scale: A case study of Blanding's turtle, *Emydoidea blandingii*, in Ontario, Canada. Journal for Nature Conservation **20**:18–29.

Moen, A. 1999. National atlas of Norway: Vegetation. Norwegian Mapping Authority, Hønefoss.

New, T. R. 2009. Insect species conservation. Cambridge University Press, Cambridge, UK.

Økland, R. H. 1986. Rescaling of ecological gradients. II. The effect of scale on symmetry of species response curves. Nordic Journal of Botany **6**:661–670.

Økland, R. H. 1992. Studies in SE Fennoscandian mires: relevance to ecological theory. Journal of Vegetation Science **3**:279–284.

Oksanen, J. 1997. Why the beta-function cannot be used to estimate skewness of species responses. Journal of Vegetation Science **8**:147–152.

Osborne, P. E., and P. J. Seddon. 2012. Selecting suitable habitats for reintroductions: variation, change and the role of species distribution modelling. Reintroduction biology: integrating science and management:73–104.

Overgaard, J., M. R. Kearney, and A. A. Hoffmann. 2014. Sensitivity to thermal extremes in Australian Drosophila implies similar impacts of climate change on the distribution of widespread and tropical species. Global Change Biology.

Papeş, M., and P. Gaubert. 2007. Modelling ecological niches from low numbers of occurrences: assessment of the conservation status of poorly known viverrids (Mammalia, Carnivora) across two continents. Diversity and Distributions **13**:890–902.

Pearce, J., and S. Ferrier. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling **133**:225–245.

Pearson, R. G., and T. P. Dawson. 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Global Ecology and Biogeography **12**:361–371.

Pearson, R. G., C. J. Raxworthy, M. Nakamura, and A. Townsend Peterson. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogeography **34**:102–117.

Pellissier, L., J.-N. Pradervand, J. Pottier, A. Dubuis, L. Maiorano, and A. Guisan. 2012. Climate-based empirical models show biased predictions of butterfly communities along environmental gradients. Ecography **35**:684–692.

Peterson, A. T. 2003. Predicting the geography of species' invasions via ecological niche modeling. The quarterly review of biology **78**:419–433.

Peterson, A. T., J. Soberón, R. G. Pearson, R. P. Anderson, E. Martínez-Meyer, M. Nakamura, and M. B. Araújo. 2011. Ecological niches and geographic distributions. Princeton University Press, Princeton and Oxford.

Phillips, S., M. Dudik, and R. Schapire. 2012. A brief tutorial on Maxent. AT&T Labs-Research, Princeton University, and the Center for Biodiversity and Conservation, American Museum of Natural History.

Phillips, S. J. 2008. Transferability, sample selection bias and background data in presence-only modelling: a response to Peterson et al. (2007). Ecography **31**:272–278.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling **190**:231–259.

Phillips, S. J., and M. Dudík. 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography **31**:161–175.

Phillips, S. J., M. Dudík, J. Elith, C. H. Graham, A. Lehmann, J. R. Leathwich, and S. Ferrier. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications **19**:181–197.

Phillips, S. J., M. Dudík, and R. E. Schapire. 2004. A maximum entropy approach to species distribution modeling. Page 83 *in* Proceedings of the twenty-first international conference on Machine learning. ACM.

Ponder, W. F., G. A. Carter, P. Flemons, and R. R. Chapman. 2001. Evaluation of museum collection data for use in biodiversity assessment. Conservation Biology **15**:648–657.

Primack, R. B. 2010. Essentials of conservation biology. 5 edition. Sinaeur, Sunderland, MA.

Primack, R. B., and S. L. Miao. 1992. Dispersal can limit local plant distribution. Conservation Biology **6**:513–519.

Pulliam, H. R. 1988. Sources, sinks, and population regulation. The American Naturalist **132**:652–661.

Pulliam, H. R. 2000. On the relationship between niche and distribution. Ecology Letters **3**:349–361.

Raes, N., and H. ter Steege. 2007. A null-model for significance testing of presence-only species distribution models. Ecography **30**:727–736.

Raxworthy, C. J., E. Martinez-Meyer, N. Horning, R. A. Nussbaum, G. E. Schneider, M. A. Ortega-Huerta, and A. T. Peterson. 2003. Predicting distributions of known and unknown reptile species in Madagascar. Nature **426**:837–841.

Rebelo, H., and G. Jones. 2010. Ground validation of presence-only modelling with rare species: a case study on barbastelles *Barbastella barbastellus* (Chiroptera: Vespertilionidae). Journal of Applied Ecology **47**:410–420.

Reddy, S., and L. M. Dávalos. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. Journal of Biogeography **30**:1719–1727.

Redfern, J., M. Ferguson, E. Becker, K. Hyrenbach, C. P. Good, J. Barlow, K. Kaschner, M. F. Baumgartner, K. Forney, and L. Ballance. 2006. Techniques for cetacean–habitat modeling.

Ree, R. H., and S. A. Smith. 2008. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. Systematic Biology **57**:4–14.

Reese, G. C., K. R. Wilson, J. A. Hoeting, and C. H. Flather. 2005. Factors affecting species distribution predictions: A simulation modeling experiment. Ecological Applications **15**:554–564.

Reineking, B., and B. Schröder. 2006. Constrain to perform: regularization of habitat models. Ecological Modelling **193**:675–690.

Riley, S. J., S. D. DeGloria, and R. A. Elliot. 1999. A terrain ruggedness index that quantifies topographic heterogeneity. Intermountain journal of sciences **5**:23–27.

Robertson, M. P., G. S. Cumming, and B. F. N. Erasmus. 2010. Getting the most out of atlas data. Diversity and Distributions **16**:363–375.

Schowalter, T. 2012. Insect responses to major landscape-level disturbance. Annual review of entomology **57**:1–20.

Schweiger, O., R. K. Heikkinen, A. Harpke, T. Hickler, S. Klotz, O. Kudrna, I. Kühn, J. Pöyry, and J. Settele. 2012. Increasing range mismatching of interacting species under global change is related to their ecological characteristics. Global Ecology and Biogeography **21**:88–99.

Soberón, J. 2007. Grinnellian and Eltonian niches and geographic distributions of species. Ecology Letters **10**:1115–1123.

Soberon, J., and A. T. Peterson. 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. Biodiversity Informatics **2**.

Spens, J., G. Englund, and H. Lundqvist. 2007. Network connectivity and dispersal barriers: using geographical information system (GIS) tools to predict landscape scale distribution of a key predator (*Esox lucius*) among lakes. Journal of Applied Ecology **44**:1127–1137.

Stockwell, D., and D. Peters. 1999. The GARP modelling system: problems and solutions to automated spatial prediction. International journal of geographical information science **13**:143–158.

Stockwell, D. R. B., and A. T. Peterson. 2002. Effects of sample size on accuracy of species distribution models. Ecological Modelling **148**:1–13.

Stokland, J. N., R. Halvorsen, and B. Støa. 2011. Species distribution modelling - Effect of design and sample size of pseudo-absence observations. Ecological Modelling **222**:1800–1809.

Syfert, M. M., M. J. Smith, and D. A. Coomes. 2013. The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. PloS one **8**:e55158.

Syphard, A. D., and J. Franklin. 2009. Differences in spatial predictions among species distribution modeling methods vary with species traits and environmental predictors. Ecography **32**:907–918.

Ter Braak, C. J. F., and I. C. Prentice. 1988. A theory of gradient analysis. Advances in Ecological Research **18**:271–317.

Thomas, C. D., A. Cameron, R. E. Green, M. Bakkenes, L. J. Beaumont, Y. C. Collingham, B. F. Erasmus, M. F. De Siqueira, A. Grainger, and L. Hannah. 2004. Extinction risk from climate change. Nature **427**:145–148.

Thuiller, W., L. Brotons, M. B. Araújo, and S. Lavorel. 2004. Effects of restricting environmental range of data to project current and future species distributions. Ecography **27**:165–172.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological):267–288.

Tveito, O. E., E. J. Førland, B. Dahlström, E. Elomaa, P. Frich, I. Hanssen-Bauer, T. Jónsson, H. Madsen, J. Perälä, P. Rissanen, and H. Vedin. 1997. Nordic precipitation maps. DNMI-Reports **22**:1–22.

Ulrichs, C., and K. Hopper. 2008. Predicting insect distributions from climate and habitat data. BioControl **53**:881–894.

Van Der Wal, J., L. P. Shoo, C. Graham, and S. E. Williams. 2009. Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know? Ecological Modelling **220**:589–594.

Veloz, S. 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. Journal of Biogeography **36**:2290–2299.

Warren, D. L., and S. N. Seifert. 2011. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. Ecological Applications **21**:335–342.

Westley, P. A., E. J. Ward, and I. A. Fleming. 2013. Fine-scale local adaptation in an invasive freshwater fish has evolved in contemporary time. Proceedings of the Royal Society B: Biological Sciences **280**.

Whittaker, R. H. 1956. Vegetation of the Great Smoky Mountains. Ecological Monographs **26**:1–80.

Whittaker, R. H. 1967. Gradient analysis of vegetation. Biological Reviews **42**:207–264.

Whittaker, R. J., M. B. Araújo, J. Paul, R. J. Ladle, J. E. Watson, and K. J. Willis. 2005. Conservation biogeography: assessment and prospect. Diversity and Distributions **11**:3–23.

Willis, K. J., and R. J. Whittaker. 2002. Species diversity-scale matters. Science **295**:1245–1248.

Wisz, M. S., R. J. Hijmans, J. Li, A. T. Peterson, C. H. Graham, and A. Guisan. 2008. Effects of sample size on the performance of species distribution models. Diversity and Distributions **14**:763–773.

Wisz, M. S., J. Pottier, W. D. Kissling, L. Pellissier, J. Lenoir, C. F. Damgaard, C. F. Dormann, M. C. Forchhammer, J. A. Grytnes, and A. Guisan. 2013. The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. Biological Reviews **88**:15–30.

Wolmarans, R., M. P. Robertson, and B. J. van Rensburg. 2010. Predicting invasive alien plant distributions: how geographical bias in occurrence records influences model performance. Journal of Biogeography **37**:1797–1810.

Yackulic, C. B., R. Chandler, E. F. Zipkin, J. A. Royle, J. D. Nichols, E. H. Campbell Grant, and S. Veran. 2013. Presence-only modelling using MAXENT: when can we trust the inferences? Methods in Ecology and Evolution **4**:236–243.

Zaniewski, A. E., A. Lehmann, and J. M. Overton. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. Ecological Modelling **157**:261–280.

I

II

III

IV