

# MODELLING PRESENCE OF SWAMP FOREST AND FOREST DWELLING BIRDS IN A BOREAL FOREST RESERVE USING AIRBORNE LASER SCANNING

MODELLERING AV UTBREDELSEN TIL SUMPSKOG OG FOREKOMSTEN AV FUGL I ET SKOGRESERVAT VED HJELP AV FLYBÅREN LASERSKANNING

**JOHN WIRKOLA DIRKSEN**



# Modelling presence of swamp forest and forest dwelling birds in a boreal forest reserve using airborne laser scanning

Modellering av utbredelsen til sumpskog og forekomsten av fugl i et skogreservat ved hjelp av flybåren laserskanning

Philosophiae Doctor (PhD) Thesis

John Wirkola Dirksen

Department of Ecology and Natural Resource Management  
Norwegian University of Life Sciences

Ås 2013



Thesis number 2013:52  
ISSN 1503-1667  
ISBN 978-82-575-1152-4

## **PhD supervisors**

Professor Mikael Ohlson

Norwegian University of Life Sciences, Department of Ecology and Natural Resource Management, post-box 5003, 1432 Ås, Norway

Professor Rune Halvorsen

University of Oslo, Natural History Museum, Department of Botany, post-box 1172 Blindern,  
0318 Oslo, Norway

**Dr. Katrine Eldegard**

Norwegian University of Life Sciences, Department of Ecology and Natural Resource Management, post-box 5003, 1432 Ås, Norway

Professor Erik Næsset

Norwegian University of Life Sciences, Department of Ecology and Natural Resource Management, post-box 5003, 1432 Ås, Norway

Professor Terje Gobakken

Norwegian University of Life Sciences, Department of Ecology and Natural Resource Management, post-box 5003, 1432 Ås, Norway

## **PhD Evaluation committee**

Professor Lena Gustafsson

Department of Ecology, Swedish University of Agricultural Sciences, post-box 7044, 750 07 UPPSALA, Sweden

Professor Ross Hill

Bournemouth University, School of Applied Sciences, Poole BH12 5BB, Dorset, England

## Acknowledgements

This thesis has been submitted in completion of my PhD studies at the department of Ecology and Natural Resource Management (INA) at the Norwegian University of Life Sciences (UMB). UMB funded my PHD.

I would like to thank my supervisors: Mikael Ohlson, Rune Halvorsen, Katrine Eldegard, Erik Næsset and Terje Gobakken: Mikael, my main supervisor, for giving me the opportunity to start as a PhD student in the first place, for always having the large overview of my study and for always redirecting my thoughts back into forest ecology. Rune, for introducing me to the ecology of swamp forests and MaxEnt, for your patience and for always being available. Katrine, for showing great enthusiasm for my PhD project even long before you were officially included into the supervisor group, for discussions and small talks. Erik and Terje for introducing me to and answering my questions about ALS, for commenting on the manuscripts, and for “giving” me Hans Ole as a sparring partner.

In addition I will highlight two persons with whom I have worked really close the last years: Sabrina Mazzoni and Hans Ole Ørka. Sabrina, thank you for all the hours we spend working with MaxEnt, for your good mood and for a friendship which I hope will continue to grow. Hans Ole you are a true master of R, ALS and ArcGIS, without you help I would not have come this far. In addition to all of your technical skills you have the mindset of an ecologist, and I hope you will continue to include aspects of ecology into your future studies, thank you for all your help.

I would also like to thank Trygve Danbolt, Christian Bianchi Strømme and Ole Martin Bollandsås for their effort during parts of the field work. Vegard Lien, Vegard Bakkestuen and Lars Erikstad for helping me out with ArcGIS. Lars Østbye Hemsing and Anders Bryn for discussions about MaxEnt. Rune Østergaard Pedersen, Endre Hofstad Hansen, Livue Ene, Marius Hauglin and Nadja Stumberg for discussions about ALS and MaxEnt.

All the other colleagues with whom I have had contact should be thanked as well, for easing up tense shoulders with a good sense of humor during lunch breaks etc. Special thanks go to Jens Thaulow, Markus Sydenham, Cathrine Glosli, Even Bergseng, Richard Bischof, Marcus Elfström, Massimo Bidussi, Rafael Leandro De Assis, Rebekka Lundgren, Vidar Selås, Espen Arestøl, Ørjan Totland, Yngvar Gauslaa, Sigmund Hågvar and finally Christian Bianchi Strømme for hours of training pursuing our common goal: to become a barstarzz.

To Anita my dear wife goes the biggest thanks, you have stood by my side through nine years of education at UMB, you have tolerated my absent-mindedness, endless hours of loneliness and frustrations over my dedication to my study. I promise that from now on I will be more present for you. Hege, Rikard and Mads you were just kids when I started this journey but are now all independent individuals with great personalities, thank you for your happiness and the tolerance you have showed for my work. Lukas, all tiden vi har tilbrakt sammen på Aschjemgård har vært ubeskrivelig verdifull for meg, du fikk meg til å glemme alt om jobben når intet annet kunne. I love you all and look forward to spend more time with you. I will also thank my twin brother Gert and my parents for supporting me throughout these years.

For me this PhD is yet another lap of a journey within forestry which started back in 1995 and which is far from finished.

John Wirkola Dirksen, Ljønerhagan, 2013



# Table of Contents

Acknowledgements.....	iii
Abstract.....	vii
Sammendrag.....	ix
List of Papers.....	xi
Synopsis.....	2
1. Introduction.....	4
1.1 Distribution modelling.....	4
1.2 MaxEnt.....	5
1.3 Remote sensing: airborne laser scanning.....	7
1.4 Airborne laser scanning in ecological studies.....	8
1.5 Research objectives.....	8
2. Study area and data collection.....	10
2.1 Study area.....	10
2.2 Study objects in focus.....	12
2.2.1 Boreal swamp forests.....	12
2.2.2 Forest dwelling birds.....	14
2.3 Collection of field data.....	15
2.3.1 Swamp forest registrations.....	15
2.3.2 Bird census.....	15
2.4 Acquisition of airborne laser scanning data.....	16
2.4.1 ALS derived metrics for paper 1.....	17
2.4.2 ALS derived metrics for paper 2.....	17
2.4.3 ALS derived metrics for paper 3.....	18
2.5 Acquisition of spectral data.....	18
3. Methods.....	19
3.1 Modelling procedures Paper 1.....	19
3.2 Modelling procedures Paper 2.....	20
3.3 Modelling procedures Paper 3.....	20
3.4 Evaluation data.....	21
3.4.1 Independent evaluation data set for paper 1.....	21
3.4.2 Independent evaluation data set for paper 2.....	22
3.4.3 Evaluation data for paper 3.....	22
4. Results and discussion.....	23
4.1 MaxEnt modelling of swamp forest distributions overfitting and importance of spatial autocorrelation in response data.....	23
4.2 Structural differences between boreal swamp forests and non-paludified forest in a forest reserve measured by airborne laser scanning.....	25
4.3 Modelling richness and presence of bird species in a boreal forest reserve using airborne laser scanning and aerial images.....	26
5. Conclusions.....	29
6. References.....	32
Paper 1	
Paper 2	
Paper 3	



## Abstract

Since the millennium there has been a huge increase in ecological studies which deals with distribution modelling, and it is a fact that distribution modelling has branched off from mainstream ecology and is now considered an independent branch within ecological science, as a key subfield of Conservation Biology under the name: Conservation Biogeography. When developing distribution models regardless of whether it is for the distribution of species or nature types it is essential to have environmental data of high quality covering large geographical areas. Such data is available through remote sensing. Airborne laser scanning (ALS) is a relatively new remote sensing technique providing modellers with 3D data which can be used as proxies for environmental variables e.g. terrain slope, vegetation height and density of forest canopies. Within the last decade ALS has been implemented in many studies which deals with distribution modelling.

In this thesis the main objective was to implement ALS data in ecological studies to gain more knowledge about the object in focus and to create guidelines for future use of ALS in similar studies. This thesis consists of three studies which all were conducted within an old-growth forests reserve in southeast Norway. Specific objects were addressed in the individual papers: For Paper 1 and 2 the main objective was to develop distribution models for boreal swamp forests. The two studies differed in the sense that Paper 1 used variables which described the topography of the terrain as environmental variables. Models were developed using the distribution modelling software Maxent. Two distinct MaxEnt modelling methods were compared, the standard MaxEnt procedure (SMP) where models were created using default settings in the software and the alternative MaxEnt procedure (AMP) where default settings were deactivated and manual selection of variables were conducted when models were built. In Paper 2 proxies for forest structures were used as environmental variables when GLM models were built. In Paper 3 the object in focus was forest dwelling birds. Models were developed for both total species richness and eight single bird species. Distribution models were developed using both GLM and Random Forest. We used ALS variables to describe forest structure, and variables derived from spectral images to describe tree composition in each of 157 counting stations.

Highly informative distribution models were developed in Paper 1. SMP models tend to be overfitted compared to AMP models. Slope, topographic position, ruggedness and curvature play an important role for the location of swamp forests. In Paper 2 we found that

forest structures differ between the two nature types. The mean vegetation height is approximately 3 m lower within swamp forests compared to non-paludified forests. The best model classified 62% of the cells correctly. For Paper 3 variables which described the density of the canopy within certain one meter sections turned out to be the most important sort of variables for six out of the eight most abundant bird species and species richness. The best distribution models were created for Goldcrest (*Regulus regulus*) and Wren (*Troglodytes troglodytes*), for the other species and species richness the models predicted poorer but still in accordance with previous field measures. We conclude that ALS data has an important role to play in the future development of distribution models and thereby Conservation Biogeography.

We recommend future modellers to always evaluate distribution models on an independent evaluation data set before implementing them into management plans.

## Sammendrag

Siden årtusenskiftet har det vært en økende interesse for bruken av utbredelsesmodellering i økologiske studier. Dette har resultert i at utbredelsesmodellering har blitt etablert som en egen disiplin innen økologiskforskning under navnet: Bevaringsbiogeografi. Uansett om man bygger utbredelses-modeller for ulike arter eller naturtyper, er det essensielt at man bruker data av høyeste kvalitet som dekker store geografiske områder. Denne type data er tilgjengelig fra flybåren laserskanning (ALS).

Hovedformålet med denne avhandlingen var å implementere ALS data i ulike typer av økologiske studier for å øke kunnskapen om bruken av ALS i denne typen studier, samt utarbeide mulige fremtidige retningslinjer for bruk av ALS data i liknende studier. Denne avhandlingen består av tre uavhengige manuskripter som alle bygger på ALS data og felldata registrert i et gammelskogreservat i Sørøst-Norge. I første manuskript var hovedformålet og predikere utbredelsen av boreal sumpskog ved hjelp av variabler som beskriver topografien i studieområdet. To forskjellige utbredelsesmodeller ble bygget ved hjelp av utbredelsesmodellingsverktøyet Maxent. En type modeller bygger på forhåndsinnstillinger i programmet en annen type modeller bygger på manuell seleksjon av variabler. I det andre manuskriptet ble ALS-data brukt til å lage variabler som beskriver skogstrukturen. Det ble undersøkt om skogstrukturen i sumpskog skiller seg fra skogstrukturen i ikke-sumpskog. Skogstrukturvariablene ble brukt til å lage generelle lineære modeller (GLM) for å finne en mulig sammenheng mellom sumpskog og skogstruktur. Disse modellene ble evaluert mot uavhengig evalueringsdata. I det tredje manuskriptet ble forekomsten av fugl modellert. Informasjon om skogstruktur og treartssammensetning fra 157 fugletakseringspunkter ble brukt til å bygge GLM-modeller og Random Forest-modeller for både total artsrikdom og åtte enkeltarter.

Utbredelsesmodellene som ble bygget ved hjelp av forhåndsinnstillingene i manus nr. 1 inneholdt flere variabler enn de alternative modellene. De variabler som best beskriver forekomsten av sumpskog er: helling, topografisk posisjon, mikrotopografi, samt informasjon om terrengets krumning. Resultatene fra manus nr. 2 viser at skogstrukturen er forskjellig mellom sumpskog og ikke-sumpskog, i gjennomsnitt er vegetasjonen 3 meter lavere i sumpskog. Den beste modellen fra dette studiet klassifiserte 62% av cellene fra et uavhengig evalueringsdatasett korrekt. Den type variabel som best beskriver tilstedeværelsen av fugler er en tetthetsvariabel som beskriver tettheten av kronen i bestemte én metersseksjoner. De

beste utbredelsesmodellene ble bygget for fuglekonge (*Regulus regulus*) og gjerdesmett (*Troglodytes troglodytes*). For de andre artene og artsrikdom ble det bygget mindre gode modeller, men disse modellene inneholdt variabler som i tidligere studier har vist seg å være viktige for de respektive artene. Konklusjonen på denne avhandlingen er at vi anbefaler videre bruk av data fra flybåren laserskanning i arbeidet med utbredelsesmodellering av både enkeltarter og naturtyper. Vi anbefaler fremtidige modellbrukere å alltid evaluere sine distribusjonsmodeller mot et uavhengig (evaluerings)datasett før de implementeres i forvaltningsplaner.

## List of Papers

### Paper 1

**Dirksen J. W.**, Halvorsen R., Mazzoni R., Næsset E., Gobakken T. and Ohlson, M. MaxEnt modelling of swamp forest distributions: overfitting and importance of spatial autocorrelation in response data. *Manuscript*.

### Paper 2

**Dirksen J. W.**, Halvorsen R., Ørka H. O., Eldegard K., Næsset E., Gobakken T. and Ohlson, M. Structural differences between boreal swamp forests and non-paludified forest in a forest reserve measured by airborne laser scanning. *Submitted manuscript*.

### Paper 3

**Dirksen J. W.**, Eldegard K., Ørka H. O., Halvorsen R., Næsset E., Gobakken T. and Ohlson, M. Modelling richness and presence of bird species in a boreal forest reserve using airborne laser scanning and aerial images. *Submitted manuscript*.



# Synopsis



# 1. Introduction

Understanding how and why species and nature types distribute as they do has always been of great interest for ecologists (MacArthur and MacArthur 1961, MacArthur and Wilson 1967). We know that species distribute along environmental gradients which affect their ecological niche in different ways, both on a local and global scale (Økland et al. 2001, Elith et al. 2006, Wollan et al. 2008). One way to gain more knowledge about the relationship between species or nature types and the environmental gradients affecting these is to build species distribution models (niche-based modelling, ecological niche modelling and habitat suitability models is just other names for species distribution models). In this thesis I will refer to the species distribution models as simply distribution models (or prediction models) since I use distribution modelling not only for species but also for nature types.

Since the millennium there has been an almost exponential growth in published studies which deals with distribution modelling (Lobo et al. 2010) and distribution modelling is now considered an independent branch within ecological science, as a key subfield of conservation biology under the name: Conservation Biogeography (Whittaker et al. 2005). This growth in publications has been boosted by technological innovations (Miller et al. 2004), access to digitalized environmental data (Telenius 2011) and new remote sensing techniques (Lefsky et al. 2002, Turner et al. 2003, Aplin 2005).

## 1.1 Distribution modelling

Before building a distribution model there are two important issues that needs to be taken into consideration: 1) Modelling purpose. The purpose with a distribution model can be three fold: i) to build spatial prediction models (SPM), these models are evaluated on their abilities to create reliable spatial predictions for the object in focus; ii) to build ecological response models (ERM), these models are evaluated on their ability to find relationships between response variables and a specific set of environmental variables, resulting in increasing ecological knowledge for the object in focus; iii) to build projective distribution models, these models are evaluated on their ability to make correct predictions outside of the study area and/or into the future with e.g. different climate scenarios (Halvorsen 2012). Within this thesis I will focus on prediction and explanatory models only. 2) Type of data. It is common to differentiate between two different types of data: presence only and presence-absence data. Presence only data brings information about the presence locations i.e. places where the

object in focus e.g. a species has actually been recorded no information about where the species was not recorded i.e. was absent is available, whereas presence-absence data bring information about both presences and absences. For both presence only data and presence-absence data it is preferable that the presence observations are gathered randomly from a large geographical area, unfortunately this is rarely the case. Presence observations are often prone to bias since samples are gathered from a small geographical area or easily accessible areas e.g. near roads, cities, rivers and nature reserves, so-called preferential sampling (Raes and ter Steege 2007, Merckx et al. 2011). Preferential sampling opens up for the possibility that parts of a species niche is not represented within the presence data i.e. the whole gradient for a certain environmental variable has not been accounted for when search for the species was conducted. When this is the case the distribution model which have been built, using the biased samples, will lead to a systematic under-prediction for the species i.e. the species is likely to be found in areas which is not predicted by the model (Raes and ter Steege 2007).

The purpose with the model and the data type available can work as a guideline when searching for an appropriate modelling method among the innumerable modelling method available (e.g. Elith et al. 2006). Recently several comparative studies have tried to find the best overall method for distribution modelling (Elith et al. 2006, Hernandez et al. 2006, Pearson et al. 2007, Peterson et al. 2007, Williams et al. 2009, Mateo et al. 2010, Merckx et al. 2011). In many of these studies a relatively new method MaxEnt (Maximum entropy, Phillips et al. 2004) stands out as the best or among the best (Elith et al. 2006, Phillips et al. 2006, Pearson et al. 2007).

## **1.2 MaxEnt**

After its release the Maxent software (Phillips et al. 2004) was quickly accepted as a reliable distribution modelling method and therefore became popular among ecologists (e.g. Wollan et al. 2008, Edvardsen et al. 2011, Erikstad et al. 2013). There are several reasons for MaxEnt's popularity: 1) its top ranking in several comparative studies; 2) it creates reliable models with presence only data (Phillips et al. 2004, Phillips and Dudík 2008); 3) it requires only a few presence observations to build reliable models (Hernandez et al. 2006, Pearson et al. 2007) and finally it is free of cost and user friendly (Phillips 2011). The fact that MaxEnt requires only a few presence observations makes is an advantage when creating distribution models for rare species e.g. species which are found in natural museum collections around the world (Graham et al. 2004, Mateo et al. 2010). MaxEnt has been used to predict the

distribution of different species: e.g. red listed plants (Edvardsen et al. 2011), birds (Smart et al. 2012, Paper 3), insects (Tognelli et al. 2009) and fungi (Wollan et al. 2008) but also to predict areas with certain nature types (Riordan and Rundel 2009, Bässler et al. 2011, Paper 1 and 2), areas likely to be invaded by alien species (Giovanelli et al. 2008) and areas prone to wildfires (Parisien and Moritz 2009). Even the distribution of the North American Sasquatch (i.e. Bigfoot) has been predicted by MaxEnt (Lozier et al. 2009).

Even though MaxEnt has been a success and has been implemented in many studies worldwide, some modellers have started to question MaxEnt's reliability when it comes to creating unbiased and non overfitted models (Raes and ter Steege 2007, Merckx et al. 2011, Halvorsen 2013). The term overfitting can be explained in different ways: i) "Overfitting means that a model adjusts to specific random features or noise of the training data but, works poorly on other datasets" Packalén et al. (2012); ii) "Overfitting generally occurs when a model is excessively complex, such as having too many degrees of freedom in relation to the amount of data available" (Merckx et al. 2011); iii) "when number of predictors  $\geq$  number of observations" (Guisan and Thuiller 2005). Halvorsen (2012) differentiate between three types of overfitting: Type I, that a more complex model has lower predictive performance on independent data than a simpler model; Type II, that a more complex model is similar (in the meaning 'not significantly better') in predictive performance on independent data than a simpler model; and Type III, that a more complex model with higher predictive performance on independent data than a simpler model fails to fit realistic overall ecological response curves. As a tool to prevent overfitting in MaxEnt models the  $\ell_1$  regularization ( $\lambda = 1$ ) has been implemented as default (Phillips et al. 2006). This regularization should help create simpler models, according to the principle of parsimony: that the simplest model is the best. It is believed that the regularization multiplier is one of the main reasons for MaxEnt's good performance (Elith et al. 2006, Phillips and Dudík 2008, Elith et al. 2011). However, recently the default settings for the regularization multiplier has been questioned by several authors which claims that the default settings results in overfitted models (Raes and ter Steege 2007, Anderson and Gonzalez 2011, Merckx et al. 2011, Halvorsen 2013). Regardless of modelling purpose, data type or  $\ell_1$  regularization settings it is essential to have environmental data of high quality covering large geographical areas when developing distribution models. Such data is available from remote sensing (Lefsky et al. 2002, Turner et al. 2003).

### **1.3 Remote sensing: airborne laser scanning**

Remote sensing defined as: “Any method of observing the earth’s surface without being directly in contact with it” (Strand et al. 2007). According to Wang et al. (2010) remote sensing will play a large role in the future not just in distribution modelling but also in the everyday life as well: “It is believed that remote sensing will develop in a path similar to that of computer science, which has penetrated all aspects of human life”.

Aerial photographs were the first remote sensing technique, and it has since 1964 been used to conduct forest inventories worldwide (Loetsch and Haller 1964). From the mid – 1970s and until today remote sensing satellites has made it possible to achieve images covering larger spatial areas than what is possible with aerial photographs (Strand et al. 2007). Both aerial photographs and satellite images are still used in forest inventories today, though being challenged by the new remote sensing technique: airborne laser scanning (ALS).

ALS is an active remote sensing technique, which can provide information about the three dimensional structures of a forest canopy by measuring the time span from a laser pulse is emitted from an aircraft mounted laser scanner and until its echo is received by the laser scanner after the pulse has been reflected from an object near the ground or the ground surface itself. The distance between the sensor and the point of reflection is estimated from the recorded time span between emission and reflection of the laser pulse. The spatial location of the individual echoes can be accurately determined since the exact position and orientation of the sensor is provided by use of global satellite navigation systems and internal navigation systems. The result from an ALS acquisition is a three dimensional point cloud, each point being an echo with its own unique position in an x, y, and z coordinate system. Ground echoes were classified using the progressive Triangular Irregular Network (TIN) densification algorithm (Axelsson 2000) of the TerraScan software (Terrasolid LTD., 2004). A TIN was created from the x and y coordinates and corresponding heights of the laser echoes classified as ground echoes. The relative height of each echo was computed as the difference between the height of the individual return and the TIN.

ALS have been used in operational forest inventories in Norway since 2002 (Næsset et al. 2004), and in many other countries on a routine basis since around 2005 (McRoberts et al 2010). In such inventories important biophysical stand variables such as mean tree height, timber volume, and diameter distribution are easily estimated from ALS data with high precision and high spatial resolution and yet with a large geographical coverage, often 500-

1000 km<sup>2</sup> or more (Næsset 2007; Gobakken and Næsset 2004). In addition, a numerous number of research articles dealing with e.g. biomass change (Næsset et al. 2013, Bollandsås et al. 2013) estimation of above- and below-ground biomass (Næsset and Gobakken 2008), estimations of forest maturity (Weber and Boss 2009), characterization of successional stage (Falkowski et al. 2009), detection of small single trees (Thieme et al. 2011) and estimation of single-tree branch biomass (Hauglin et al. 2013) have used ALS data.

#### **1.4 Airborne laser scanning in ecological studies**

During the last decade ecologists have recognized ALS as a tool to derive metrics which can be used as proxies for the three dimensional structures of forests (Lefsky et al. 2002, Turner et al. 2003, Aplin 2005, McDermid et al. 2005, Wang et al. 2010). Traditionally measures for these structures have been both time consuming and difficult to obtain especially for large geographical areas, but now proxies for e.g. successional stages (Falkowski et al. 2009), understory density (Hill and Broughton 2009, Martinuzzi et al. 2009, Wing et al. 2012), canopy gaps (Koukoulas and Blackburn 2004), standing dead trees (Bater et al. 2009, Kim et al. 2009), tree species (Holmgren et al. 2008, Ørka et al. 2009) and the vertical distribution of canopy density (Turner et al. 2003, Lefsky et al. 2002) can easily be obtained from ALS data. Even though ALS has been used in several studies focusing on mammals, e.g. studies on bats (Jung et al. 2012), mule deer (Coops et al. 2010), and Pacific fisher (*Martes pennant* Erxleben), it is within bird ecology that ALS has been the most implemented (e.g. Clawges et al. 2008, Graf et al. 2009). This is not surprisingly since bird species are known to respond to forests structures (Cramp 1988, 1992, 1993, 1994, Hill et al. 2004, Broughton et al. 2006, Bergen et al. 2007, Bellamy et al. 2009, Graf et al. 2009, Hinsley et al. 2009, Müller et al. 2010). Data derived from ALS has also been used to detect nature types of high biological importance (Mücher et al. 2009, Vehmas et al. 2009, Bässler et al. 2011)

#### **1.5 Research objectives**

The main objective of this thesis was to implement ALS data in three different ecological studies, all carried out in an old-growth forests reserve in southeastern Norway. In paper 1 the distribution of boreal swamp forests was modelled using terrain variables derived from a digital elevation model. In paper 2 differences in forest structure between swamp forests and non-paludified forest were modelled. In paper 3 a bird survey was conducted and explanatory

and prediction models for species richness and eight single-species created. The individual goals for each paper are described below.

Paper 1 – “*MaxEnt modelling of swamp forest distributions overfitting and importance of spatial autocorrelation in response data*” – This study has three main aims: Firstly: MaxEnt models built with default settings was compared against models build through a manual stepwise forward selection approach to test whether or not MaxEnt models build with default settings was overfitted. Models were built using ALS derived metrics describing the terrain in the study area. Secondly: we tested whether or not spatial autocorrelation in the response variable (presence of swamp forest) influenced the predictive performance of MaxEnt models and thirdly: we identified key factors determining the occurrence of swamp forests.

Paper 2 – “*Structural differences between boreal swamp forests and non-paludified forest in a forest reserve measured by airborne laser scanning*” – This study has three main aims: firstly: to test the hypothesis that swamp forest differ from the surrounding non-paludified forest in structural properties that can be recorded on at least one of the scales: 25 m<sup>2</sup>, 225 m<sup>2</sup> or 625 m<sup>2</sup>. Secondly: to model structural differences between swamp forests and surrounding non-paludified forests and thirdly: we evaluate the developed models on two independent evaluation data sets to test how well they performed.

Paper 3 – “*Modelling richness and presence of bird species in a boreal forest reserve using airborne laser scanning and aerial images* “ – The main aims of this paper were firstly: to assess the utility of ALS metrics for explaining bird species richness and single-species occurrence. Secondly: to assess the utility of ALS metrics for predicting bird species richness and single-species occurrence and thirdly: to compare the ability of ALS-derived predictors against aerial imagery for predicting richness and single-species occurrence. A secondary aim was to assess the relative utility of different statistical techniques for prediction modeling (generalized linear models and Random Forest).

## 2. Study area and data collection

### 2.1 Study area

The study area, which was the same for all three papers, was situated in Østmarka Nature Reserve (59°50' N, 11° 02' E, 190–368 m.a.s.l.), which is located in the southern boreal zone of Southeast Norway (Moen 1999) (Figure 1).

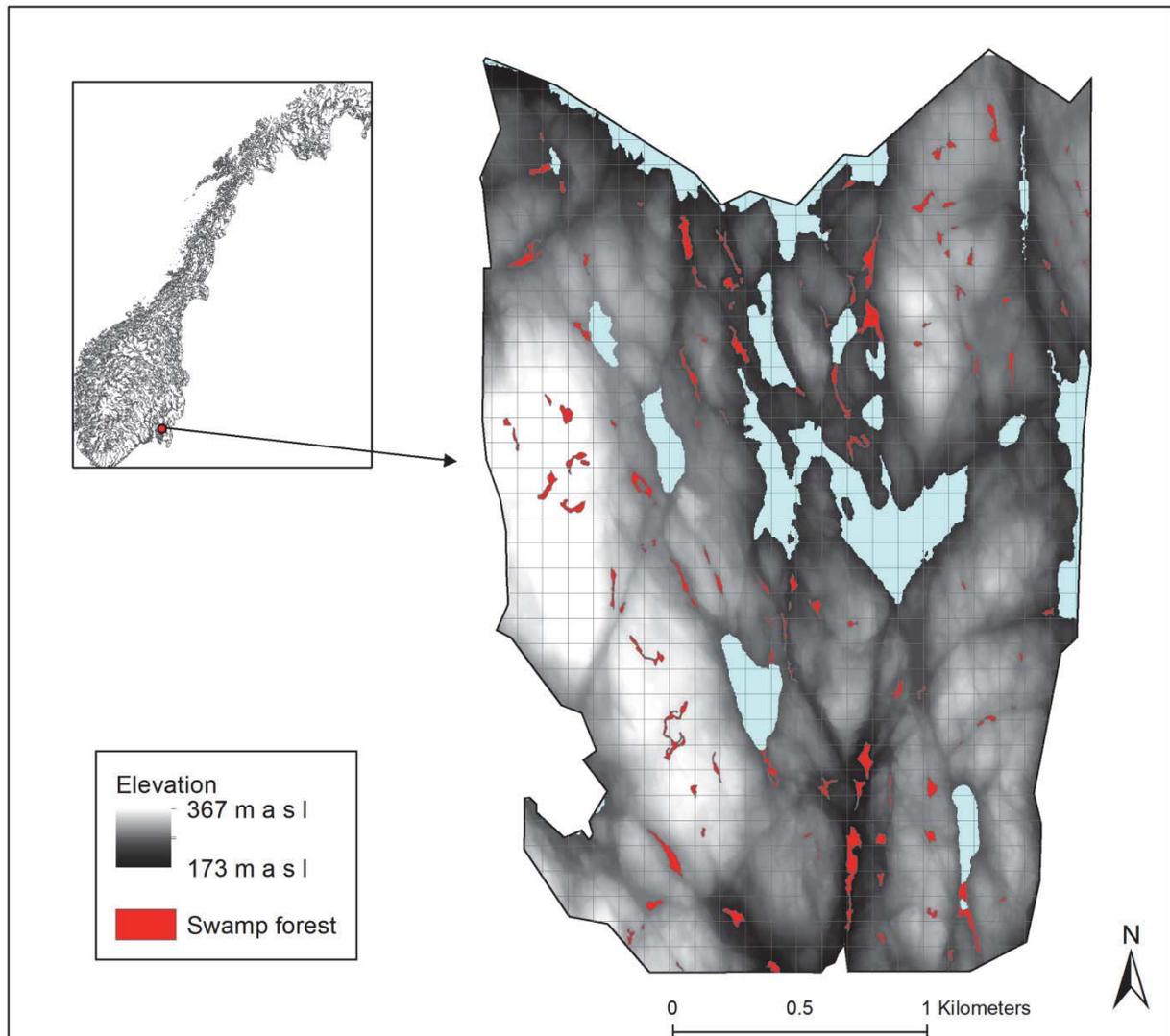


Figure 1. The location of the study area in the Østmarka Nature Reserve in Southeast Norway. The red polygons show the 121 swamp forest polygons. Lakes appear in blue. Altitude is indicated by shading; low-lying areas are dark and higher elevations have light grey shading. A raster (grey lines) with cells of size 100 m × 100 m is superimposed on the map to help illustrate the size of the polygons.

The study area comprised approximately 6 km<sup>2</sup>. Annual mean precipitation and temperature (1961–90 normals) at the nearest meteorological stations (Enebakk, ca. 10 km SE of the area, and Hakadal–Blikrudhagan, ca. 30 km N of the area) were 816 mm and 3.9 °C, respectively (Aune 1993, Førland 1993). The bedrock is of proterozoic age and quartz diorite, tonalite and gneisses of diverse origins dominate (Sigmond et al. 1984). The topography of the study area is dominated by north–south facing ridges and valleys (Figure 1). On finer scales, the landscape is dissected by ravines, resulting in a very broken topography with structures on several scales and a mixture of ridges, valleys, stream beds, large boulders, steep cliffs and more gentle slopes (Figure 2).



Figure 2. Fine scale terrain variation within Østmarka nature reserve.

Parts of the reserve have been protected since 1990 and additional areas were protected in 2002 so that the whole reserve now covers approximately 12.5 km<sup>2</sup>. The reserve is dominated

by old-growth spruce (*Picea abies* L. Karst) forests (Fig). Pine (*Pinus sylvestris* L.) dominates on nutrient poor and shallow soils on ridges, and deciduous trees, mainly birches (*Betula pubescens* Ehrh. and *B. pendula* Roth.), are scattered throughout the study area.

Even though selective logging has been conducted within the study area until around 1940 the forest has retained many of the characteristics found in primeval forests i.e. multi layered canopy, mixture of tree species, good spacing between the trees, open areas and large amounts of dead wood (Figure 3). The study area, which covers the center of the reserve, was conceptualized as a rasterized geographical space with resolution (grid-cell size) of 5 m, resulting in a total of 233774 grid cells this raster was used in paper 1 and 2.

## **2.2 Study objects in focus**

Paper 1 and 2 focused on boreal swamp forests whereas paper 3 focused up on the bird community within the study area.

### **2.2.1 Boreal swamp forests**

Boreal swamp forests (Figure 4, hereafter named swamp forests) are in this thesis defined as forest which grows on peat deposits, with a high water table which varies with precipitation. Swamp forests are mainly located within valleys, depressions and along stream beds, but are also occasionally located on hills and ridges large enough to contain minor depressions. In Fennoscandian swamp forests spruce is the dominating tree species. Several studies conducted on Fennoscandinavian swamp forests back in the 1990's emphasized the biological importance of swamp forests in the boreal forests landscape (Seegerström et al. 1994, Kuusinen 1996, Ohlson et al. 1997, Hörnberg et al. 1998). Sjöberg and Ericson (1992) argued that: "Forested wetlands generally should be preserved, since otherwise late successional forest stages will disappear from the managed forest landscape outside reserves. This is the best possibility of ensuring the survival of species dependent on forest stand continuity and characterized by low dispersal abilities". Hörnberg et al. (1998) draws the conclusion that swamp forests are to be considered: "biodiversity hotspots in the boreal forest landscape". Timonen et al. (2010) classifies spruce wetland as a woodland key habitat in northern Europe.



Figure 3. Large variation in forest structure is found within the study area.

That swamp forests are rich in biodiversity is nicely exemplified with the results from a study conducted by Ohlson et al. in 1997. In this study as many as 517 species were registered within 10 small remnants of swamp forests in Sweden (148 vascular plants, 131 mosses, 64 hepatics, 142 lichens and 32 wood-inhabiting fungi). This high number of species becomes even more remarkable when the average size of these 10 swamp forests was only 2 hectares. The bryophyte species alone resembles 33% of all bryophyte species registered in Sweden. Another example which documents the importance of swamp forests for forest dwelling birds is a study from 1993 where Sundberg compared bird species richness and number of

individuals between swamp forests and managed forest. He found that swamp forests had 42% more species and 55% more individuals than managed forests.

There are several explanations for why swamp forests are so rich in biodiversity: 1) long forest continuity. Even though swamp forests are no longer considered to be truly fire-free refugia the fire frequencies within swamp forests are considerably lower within swamp forests than in ordinary boreal forests (Hörnberg et al. 1995, Ohlson et al. 2006) resulting in a more stable environment where species can accumulate and survive over time; 2) a hydrotopographical gradient. Within a relatively small area it is possible to have the whole gradient from 100 percent water saturation to dry areas e.g. on hummocks and tree bases. This gradient in water saturation results in many different ecological niches where many different species will be present (Hörnberg et al. 1998, Økland et al. 2008); 3) a nutrition gradient. As species distribute along the water saturation gradient they will distribute along a nutrition gradient as well (Økland et al. 2008). In a single swamp forest it is possible to have nutrition poor area as well as nutrition rich areas e.g. rich areas could be located close to an underground spring or areas where drainage from higher elevated areas are rich in nutrition; 4) dead wood. Large quantities of dead wood also play an important role for the high amount of species found within swamp forests (Ohlson et al. 1997).

### **2.2.2 Forest dwelling birds**

As mentioned in the introduction, ALS data has often been used to gain more knowledge about how bird distributions are related to different forest structures, but that has never been done for species located in Fennoscandia. We therefore decided to use the same ALS data as was used to predict the location of swamp forests in paper 1 and 2 to predict the habitat preferences for the hazel grouse (*Tetrastes bonasia* L.), a forest dwelling grouse species which has a preference for swamp forests, in Paper 3. Unfortunately a pilot study revealed that hazel grouse was difficult to detect within the study area, the explanation for this could be twofold: (1) either the hazel grouse population within the study area consisted of few individuals or (2) the broken topography within the study area resulted in large territories for the individual birds which again made it difficult to detect the birds. Based on these facts we decided that it would be too risky to rely only on hazel grouse in our study. As a consequence of this we changed the objectives of Paper 3 from just focusing on hazel grouse to perform a survey of the whole bird community in the study area. The bird community

within the study area is relatively species-poor, a total of 28 species were recorded during the bird censuses (see section 2.3.2)

## **2.3 Collection of field data**

### **2.3.1 Swamp forest registrations**

The location of swamp forests were registered during summer and fall 2010. I systematically searched the entire study area for swamp forests larger than 100 m<sup>2</sup>. When swamp forests were encountered the edges of these were drawn manually onto a topographic map (1:1000) in the field. The locations of the edges were found by use of a handheld Garmin eTrex H Global Positioning System (GPS) receiver. The recorded swamp forest edges were digitized as polygons in a geographical information system project, using ArcMap 9.3 (Anonymous 2008). A total of 121 swamp forest polygons were recorded (Figure 1). The centroid cells for each of these 121 polygons constituted one presence-only data set, while another presence-only data set included all 7175 cells included in the 121 swamp forest polygons both data sets were used as training data in Paper 1. For Paper 2 we used a presence-absence data set, where the 121 centroid cells from the 121 swamp forests polygons constituted the presence observations and 121 randomly assigned non-paludified forest cells constituted the absence cells.

### **2.3.2 Bird census**

To collect the bird data we followed the random stratified point transect technique (Bibby 2000). A total of 157 counting stations, with a radius of 50 m, were visited twice between the 8<sup>th</sup> of May and the 10<sup>th</sup> of June 2012, by the same experienced observer, between sunrise and 10:00 AM. At each visit the observer recorded birds for five minutes. Only birds recorded by visual and/or audio observations within the 50 m radius were registered. To avoid counting the same male twice the centers of the counting stations were positioned at least 150 meters apart. The 50 m radius was determined by visual estimation. To ensure effective sampling of the study area i.e. to capture as much of the variation in vegetation structure (height) and composition (tree species) as possible the counting stations were stratified, into different strata: spruce, pine and deciduous trees and low, middle and tall trees (see Paper 3 for further details). These strata should resemble habitat characteristics that are known to be important

for birds in general (e.g. Sutherland et al. 2004). Bird counting was terminated if it was raining and/or the wind-speed was above 8 meters/second.



Figure 4. Illustrations of boreal swamp forests included in the study.

#### **2.4 Acquisition of airborne laser scanning data**

The same ALS data set was used for all three papers. The acquisition of the ALS data was performed 2<sup>th</sup> of July 2010 by the contractor Blom Geomatics AS, Norway (see Table 1 for technical information). In Paper 2 and 3 echoes below 0.5 m were considered as ground hits and echoes above 0.5 m were considered vegetation hits. We used a threshold of 0.5 m since

the ellipsoidal height accuracy of the TIN model was expected to be around 20–30 cm (Kraus and Pfeifer 1998, Reutebuch 2003) in addition a small measurement error (the z-error) will always be present in the individual echoes ( $z < 10$  cm; Baltsavias 1999) large stones and logs that occurred scattered throughout the study area could also be confused with shrubs.

Table 1. Technical details for the ALS acquisition.

	Technical specifications
Aircraft	PA31 Piper Navajo
Laser scanner instrument	Optech ALTM-Gemini
Average speed ( $\text{m s}^{-1}$ )	80
Flying altitude (m)	900
Scan side overlap (%)	55
Scan frequency (Hz)	55
Swath width (m)	440
Max. number of echoes per pulse	4
Mean pulse density ( $\text{m}^{-2}$ )	5
Half-scan angle (degrees)	13.8
Footprint diameter (cm)	23
Pulse repetition frequency (kHz)	100

#### 2.4.1 ALS derived metrics for paper 1

From a digital elevation model (DEM) we derived metrics which described the topography of the terrain. A total of 11 variables describing the terrain were used as environmental variables (EVs) in the MaxEnt modelling of swamp forest distribution in the study area.

#### 2.4.2 ALS derived metrics for paper 2

To describe as much of the variation in the three dimensional forest structure as possible we derived 24 variables: mean height, kurtosis of heights, coefficient of variation of heights and standard deviation of height, nine percentiles, nine variables describing the canopy density and one variable describing shrub density between 0.5 and 2 m. Canopy density was computed by dividing the 95<sup>th</sup> percentile minus 0.5 into 10 vertical bins of equal height and calculating the proportion of echoes within a certain bin to the total number of vegetation echoes. The shrub density was computed as the proportion of echoes between 0.5 and 2 m to total number of vegetation echoes. We used two echo categories: ‘first’ echoes and ‘last’ echoes (‘first’ echoes were constituted by single echoes and first of many echoes; ‘last’ echoes were constituted by single and last of many echoes) resulting in a total of 48 variables. Each of the variables was analyzed on three spatial scales: 25  $\text{m}^2$ , 225  $\text{m}^2$  (3x3 cells) and 625

m<sup>2</sup> (5x5 cells). Variables describing the two broadest scales were derived using a mowing window approach.

### **2.4.3 ALS derived metrics for paper 3**

For each of the 157 counting stations a total of 66 variables describing the forest structure were computed from the ALS data. Variables which described the height of the vegetation was: maximum height, mean height, standard deviation, coefficient of variation, kurtosis of height, skewness of height in addition nine percentiles were computed for all vegetation echoes (i.e. echoes above 0.5 m). Two different measures for canopy density were computed: canopy density and density layers. Canopy density was computed in the same manner as described for Paper 2 in section 2.4.3. Density layers relates to the proportion of echoes registered within a certain 1 m layer (e.g. between 1-2 m, 2-3 m or 10-11 m) to total amount of vegetation echoes for each individual counting station, calculations were done for a total of 40 layers. Finally, two variables described the density of the shrub layer: proportion of pulses between 0.5 and 2 m to total amount of vegetation echoes, and proportion of pulses between 0.5 and 3 m to total amount of vegetation echoes.

### **2.5 Acquisition of spectral data**

The Vexcel sensor acquired both panchromatic and four multispectral bands (red, green, blue, and infrared). To co-register the ALS and image data, each laser echo coordinate were mapped to the image planes of the Vexcel UltraCam digital images using a rotation matrix (Mikhail et al. 2001). The digital numbers from all bands of the pixel position of the laser echo were assigned to the respective laser echo in the original scale. If the laser echo pixel position occurred on more than one image, the mean band value of the pixels from all images was assigned to the respective laser echo. Furthermore, since only 'first' echoes will be reflected close to the canopy surface and all other subsequent echoes will penetrate into the canopy image pixel values were only assigned to the first returns. From each of the four bands of the UltraCam imagery descriptive features (mean, standard deviation, coefficient of variation, skewness, kurtosis and percentiles (10th, 30th, . . . , 90th) were used to compute the relative band values (the individual band divided by the sum of all bands), and the band ratios (one band divided by another band).

## 3. Methods

### 3.1 Modelling procedures Paper 1

The MaxEnt software (Phillips et al. 2006, Phillips and Dudík 2008, Hijmans and Elith 2011), version 3.3.3k (Phillips 2011), was used for modelling the distribution of swamp forests in the study area as a response to the 11 EVs. MaxEnt was run in batch mode via a set of customised R scripts (Mazzoni & Halvorsen unpublished material).

The main steps of MaxEnt modelling, after presence only data for the modelled target and EVs have been prepared, are (Halvorsen 2013): (1) to transform EVs to derived explanatory variables (DVs; for a full explanation of how the DVs are derived refer to appendix B in Paper 1); (2) to choose a model selection strategy, or regularisation method; and (3) to choose a method for internal model performance assessment to be used in step (2); and (4) to parameterize the model. The transformation step (1) is required to open for modelling nonlinear relationships between the modelled target and the EVs. The choice of model selection strategy (2) implies choosing one specific among many alternative models; model selection is typically carried out by an iterative process by which the selected model performance statistic (3) is used to assess the contribution of EVs, directly or via the DVs that represent the EVs, and, thus, to rank alternative models. A specific threshold value for the performance statistic, the model improvement criterion, may be used to test the hypothesis that more complex models do not improve a simpler reference model significantly. In this study we wanted to test two different types of MaxEnt models: standard maxent procedure (SMP) and alternative maxent procedure (AMP) models.

To create SMP models default settings for the MaxEnt software was retained i.e. transformation of EVs into derived variables (DVs) was performed as an integrated part of the model parameterisation procedure. Model selection was carried out by  $\ell_1$ -regularisation and SMP models were obtained for six values of the regularisation multiplier  $\lambda = 1, \lambda = 2, \lambda = 4, \lambda = 8, \lambda = 16, \lambda = 32$ .

AMP models were parameterised using the aforementioned R-scripts for stepwise forward variable selection. This method for model selection includes three phases (see appendix A Paper 1): (1) derivation of a parsimonious set of DVs from each EV; this set of DVs was used to represent the EV in the next model-building phases; (2) building a MaxEnt

model without interactions by sequentially adding EVs until no more EVs can be found that increases model performance; and (3) building a MaxEnt model with interactions by sequentially adding product variables [products between DVs representing the EVs included in the final model from phase (2)] to the final model from phase (2) until no more EVs can be found that increases model performance. The  $F$ -ratio test developed for MaxEnt by Halvorsen (2013) was used for comparison of nested MaxEnt models, with significance level  $\alpha$  as model improvement criterion (hence the regularisation multiplier was deactivated i.e.  $\lambda = 0$ ). Seven different  $\alpha$  were used:  $\alpha_1 = 0.25$ ,  $\alpha_2 = 0.1$ ,  $\alpha_3 = 0.05$ ,  $\alpha_4 = 0.01$ ,  $\alpha_5 = 1.0e-03$ ,  $\alpha_6 = 1.0e-04$ ,  $\alpha_7 = 1.0e-05$ ,  $\alpha_8 = 1.0e-07$  and  $\alpha_9 = 1.0e-11$ .

All SMP and AMP models were created using two different numbers of presence observations i.e. response variables: 121 presence observations which resemble the centroid of each of the 121 swamp forest polygons, and 7175 presence observations which resemble all the grid cells within the 121 swamp forest polygons.

### **3.2 Modelling procedures Paper 2**

To detect structural differences between swamp forests and non-paludified forests we performed Kruskal-Wallis tests for all variables from both first and last echoes on each of the three spatial scales (25m<sup>2</sup>, 225m<sup>2</sup> and 625m<sup>2</sup>). We used  $\alpha = 0.05$  as the significance level.

One explanatory logistic regression model [generalised linear models (GLM) with binary response variable and binomial distribution of errors; (Venables and Ripley 2002)] was obtained for each of the spatial scales by best-subset selection, as implemented in the R-package *glmulti*. The criteria for selection of the best model were an internal ranking of the models BIC values (the Bayesian Information Criterion, or Schwarz' criterion; Schwartz 1978). When models were developed first and last echoes were pooled together. Since the main purpose with these models was to gain more knowledge about how forest structure differs between swamp forest and non-paludified forest, we decided to include a maximum of four variables into these models using the R-package *glmulti* (best subset selection).

### **3.3 Modelling procedures Paper 3**

Explanatory models and prediction models were created for both species richness and each of the eight most abundant bird species. To create the explanatory models ALS variables were used as explanatory variables whereas for the prediction models three different groups of explanatory variables were used: 1) ALS variables, 2) spectral image variables and 3) ALS

and spectral image variables pooled (ALS+spectral). Response variables used in the modeling were species richness ( $y$  = number of species) and binary presence of each of the eight most abundant bird species ( $y = 1$  means that the species was observed while  $y = 0$  means not observed i.e. absent). For each response variable, we used GLM to create both an explanatory model and a prediction model. In addition we used Random Forest to create additional prediction models, for each of the response variables, in order to compare the predictive abilities and performance of two different prediction modeling techniques.

To build the explanatory models for each of the eight species we fitted a logistic regression model, for species richness we fitted a Poisson regression model. For all models we used the R-package *glmulti* to ensure that the models included a maximum of four variables. We created prediction models for species richness and single-species in the same way as when we created explanatory models the only difference was that up to 10 variables were allowed to be included in the prediction models (prediction models build using GLM are hereafter referred to as GLM10 models). Random Forest models (hereafter RF models) were created using the random forest algorithm (Breiman 2001). The *randomForest* package in R (Liaw and Wiener 2002) was used to build the models for regression (species richness) and classification (presence-absence of single-species).

### **3.4 Evaluation data**

The best way to evaluate models, and thereby get reliable estimates of the models predictive performance, is by use of an independent presence/absence test data set i.e. presence and absence locations are sampled independently of the training data (Elith et al. 2006, Austin 2007, Veloz 2009, Edvardsen et al. 2011, Halvorsen 2012).

#### **3.4.1 Independent evaluation data set for paper 1**

Grid cells, which were included in the evaluation data set, were obtained by creating a pilot distribution model using logistic regression with manual forward stepwise selection among individually significant EVs. All EVs that were found to be significant in the pilot model were used to obtain a pilot MaxEnt model, for swamp forests in the study area. As a result of this MaxEnt model each grid cell was assigned a value between zero and one (cells with a value close to one had the highest probability of being a swamp forest). We divided the cells into nine classes. Grid cells which potentially could be included in the evaluation data set

were selected as follows: Starting with class one (the class which contained the cells with highest probability of containing swamp forest) and proceeding to class seven, 10 cells were selected at random from each class. Additionally, 15 cells were selected from each of classes eight and nine, a higher number of cells from these two classes were included in the evaluation data set because they represented a higher proportion of the total number of grid cells, resulting in a total of 100 cells. This procedure was repeated four times resulting in a total of 400 potential evaluation cells, out of these 203 cells were included in the evaluation data set. These cells were visited in the fall 2011 to determine if they contained swamp forest or non-paludified forest. We used (AUC) [the area under the receiver operating characteristic (ROC) curve (Pearce and Ferrier 2000)] as the evaluation criteria when models were evaluated on the evaluation data set.

### **3.4.2 Independent evaluation data set for paper 2**

For evaluation of the explanatory models we created two different independent evaluation data sets. The presence cells i.e. cells containing swamp forests were the same for both of the evaluation data sets, these cells were encountered in the following way: for each of the 121 swamp forest polygons cells which was surrounded by other swamp forest cells in each of the four cardinal directions were located. One-hundred-and-three swamp forests had several cells which fulfilled this criterion. From each of these swamp forests we randomly picked a single cell resulting in 103 presence cells. Two sets of absence cells were derived for evaluation of the models. The first set included 103 absence cells, which were positioned no closer than 5 m and no longer away than 10 m from the edge of the swamp forests which contained the presence cells. The second set enclosed 103 cells which were randomly picked from all the non-paludified forests cells within the study area. We wanted to use these two different sets of absence cells to see if the model predictions differed between non-paludified forests located close to swamp forests versus non-paludified forests situated longer away from the swamp forests. The evaluation data sets were named: ‘close evaluation data set’ and ‘random evaluation data set’, respectively. As a measure for the proportion of variation explained by each model we used the modified version of  $R^2$  proposed by Nagelkerke (1991).

### **3.4.3 Evaluation data for paper 3**

Unfortunately we did not have an independent evaluation data set which could be used to evaluate the models developed in paper 3 and was therefore compelled to perform a 10 fold

cross-validation. For evaluation of the prediction models we used Cohen's kappa ( $\kappa$ ) and AUC in parallel. It was not possible to compute AUC and kappa values for GLM10 and RF models describing species richness since species richness is a count response variable. For evaluation of the explanatory models we used the modified  $R^2$  proposed by Nagelkerke (1991).

## 4. Results and discussion

### 4.1 MaxEnt modelling of swamp forest distributions overfitting and importance of spatial autocorrelation in response data

Our results clearly demonstrate that MaxEnt models built with default setting for the MaxEnt software (SMP models) are overfitted regardless of how overfitting is defined (Halvorsen 2013). SMP models consistently included more EVs and DVs than did the AMP models, even though both types of models were initialized with the same 11 EVs. Number of DVs included in the SMP models differs widely when the regularisation multiplier ( $\lambda$ ) was changed. The SMP models which used  $\lambda = 1$  (default setting in MaxEnt, and the least strict criteria for including additional DVs into the model) included as many as 101 and 240 DVs when developed for 121 and 7175 presence observations, whereas SMP models created with  $\lambda = 32$  included 9 and 28 DVs, respectively. The AMP models which used  $\alpha = 0.25$  (the least strict model selection criteria for AMP models) included only 14 and 12 DVs when developed for 121 and 7175 presence observations, AMP models created with  $\alpha = 1E-11$  included 9 and 12 DVs, respectively. SMP models built with  $\lambda=1$  always had the highest training AUC values among the SMP models, whereas when evaluated on the independent evaluation data these models had the second lowest test AUC. The fact that the SMP models which uses the default setting for the regularisation multiplier ( $\lambda = 1$ ) are overfitted, and predicts the location of swamp forests poorer than simpler SMP models is important findings since this default value originally was implemented to avoid overfitting in MaxEnt models and to make MaxEnt more users friendly (Phillips and Dudik 2008), several studies even claim that the good performance of MaxEnt models in comparative studies is due to the default settings for the regularization multiplier (Elith et al. 2006, Phillips and Dudík 2008, Elith et al. 2011). No doubt the default settings in MaxEnt have made the software more user friendly: a mini review of 87 papers, performed by Halvorsen (2013) showed that the majority of studies which used MaxEnt for distribution modelling used the default settings

without questioning its right, unfortunately these studies might have emphasized models which are in fact overfitted and do not predict distribution of the object in focus properly (Halvorsen 2013 and references herein).

Our results are supported by other studies which also question the default settings in MaxEnt (Raes and ter Steege 2007, Anderson and Gonzalez 2011, Merckx et al. 2011, Warren and Seifert 2011). Merckx et al. (2011) build models with default settings for the regularisation multiplier and concluded that these models were overfitted. Anderson and Gonzalez (2011) conclude that the optimal regularization multiplier for their MaxEnt models often was higher than the default, similar to our results. Several aspects of the training data can influence the outcome of the modelling exercise and both Merckx et al. (2011) and Anderson and Gonzalez (2011) emphasize that data sets which includes data from preferential sampling or data which contains only a few presence observations are more prone to overfitting since the full niche of a species is not represented in the data. Our data set is influenced by neither preferential sampling, nor has it too few observations. Our training data includes a minimum of 121 presence observations which is a lot more than what is needed to create reliable MaxEnt models (e.g. Hernandez et al. 2006, Anderson and Gonzales 2011). In addition, we search for swamp forests with equally intensity throughout the whole study area to avoid preferential sampling. Even though it is not well understood how spatial autocorrelation among presence observations i.e. response variables, affect distribution models (Dormann et al. 2007, Santika and Hutchinson 2009). Merckx et al. (2011) mention spatial autocorrelation among the presence observations as a potential reason for getting overfitted models. Spatial autocorrelation will affect distribution models in the same manner as preferential sampling. In our study we looked at the affects of spatial autocorrelation among presence observations but were not able to show any significant differences in the predictive performance for models which were build using 121 presence observations which were not spatially autocorrelated as response variables and models which included 7175 highly autocorrelated presence observation as response variables.

To get a better understanding of how the distribution of swamp forests is affected by the terrain formations, complies with the ERM purpose of distribution modelling. For a model to be appropriate for ERM it should include only a few EVs but at the same time predict the object in focus with a high accuracy (Halvorsen 2013). The best ERM model in our study included four EVs: slope, topographic potion index, plan curvature and vector ruggedness measure. The single most important EV for describing the distribution of swamp

forests is slope, nearly all swamp forests are situated on slopes between 0 and 3°. The topographic position index variable indicates that swamp forests are located in valleys or depressions and not on ridges. Plan curvature distinguishes between convex and concave slopes and its presence in the model indicate that swamp forests are mainly situated on concave slopes. The vector ruggedness measure variable indicates that the surface of swamp forests is smoother than the surrounding landscape a result of the in-filling of depressions in the landscape by peat accumulation during swamp forest development (Ohlson et al. 2006). These EVs accord with those pointed out as important for swamp forest development in previous studies (see Ohlson et al. 2006, Hörnberg et al. 2012)

#### **4.2 Structural differences between boreal swamp forests and non-paludified forest in a forest reserve measured by airborne laser scanning**

In this study we show that forest structure do differ between swamp forests and non-paludified forest on all three spatial scales: 25m<sup>2</sup>, 225m<sup>2</sup> and 625m<sup>2</sup>, mean vegetation height, standard deviation of height and the percentiles was significantly different for all the models. The mean vegetation height for swamp forests was 3.1 m lower than in non-paludified forests when first echoes covering 25m<sup>2</sup> were analyzed. The fact that we were able to show that swamp forest vegetation is significantly lower within swamp forests supports the hypothesis we set out to test: ‘that vegetation in swamp forests is lower than vegetation in non-paludified forest’. Even though the variables which were included in the models were highly significant, the predictive performance of these models varied in an inconsistent manner among spatial scales and evaluation data sets e.g. the highest and lowest proportion of correctly predicted cells was found for the model developed for the largest spatial scale, when evaluated on the random evaluation data set this model predicted 62% of the presence/absence cells correctly, whereas it only 54% of the cells correctly when evaluated on the close evaluation data set.

We believe that there are at least two issues which are responsible for the difficulties we experienced when trying to distinguish between forest structures in the two nature types: 1) size of the swamp forests: most of the swamp forests within the study area are relatively small (less than 1000 m<sup>2</sup>). Swamp forests which cover only a small geographical area are more prone to erroneous classification of the ALS echoes than larger swamp forests (Næsset et al. 2013). Imagine a scenario where a large spruce tree is growing just outside of a swamp forest, most of its crown is situated outside the swamp forest but a proportion of the branches are stretching into the swamp forest. When echoes are reflected from the branches

which are stretching into the swamp forests these echoes are erroneously recorded as echoes reflected from vegetation growing in the swamp forest, resulting in an unnaturally increase in the mean vegetation height for the swamp forest; 2) heterogeneous forest structure: there were distinct differences in forest structure both between and within individual swamp forests in the study area. Two swamp forests could have very different forest structures: one swamp forest could be dominated by small spruce and birch trees, whereas another could be entirely dominated by large spruce trees. Differences in structure could also be pronounced within individual swamp forests, e.g. with large trees in one end and an almost treeless area in the other end (Figure 4). These differences in structure between different swamp forests and within the same swamp forest corroborate the results in Hörnberg et al. (1995) and Økland et al. (2001) and complicate the distinction between the two nature types.

We are not aware of any other studies which have used ALS data to compare forest structure between boreal swamp forests and non-paludified forest in Fennoscandia or other places in the world. A few other studies have analyzed ALS data from related ecosystem: mires or forested wetlands (i.e. forested areas where repeated flooding is common but continuously standing water is lacking) but these studies have mainly looked upon geomorphic characters of these nature types (Hopkinson et al. 2005, Richardson et al. 2010, Korpela et al. 2009). For example Korpela et al. (2009) among other things used ALS data to model the surface patterns of mires to detect hummocks and hollows in Finland. Other studies have used ALS data to distinguish between different nature types: Mücher et al. (2009) and Bässler et al. (2011) mapped the distribution of nature 2000 nature types. Vehmas et al. (2009) predicted the location of herb-rich forest stands. Vehmas et al. (2009) reports that herb-rich stands were classified correctly in 55% of the cases which is approximately the same chances as we had for classifying swamp forests correctly in our study. In their study the size of the stands had no effect on the classification, most likely because the variation in forest structures found in rich versus poor forest stands are more pronounced than between swamp forests and non-paludified forests in our study area.

#### **4.3 Modelling richness and presence of bird species in a boreal forest reserve using airborne laser scanning and aerial images**

That ALS variables were significantly related to both species richness and occurrence of six of the eight single bird species is in accordance with several other studies which have used

ALS data as explanatory variables in models describing the distribution of birds (e.g. Goetz et al. 2007, Graf et al. 2009, Lesak et al. 2011, Tattoni et al. 2012, Smart et al. 2012).

One or more density layer variables was included in each of the seven developed explanatory models indicating that certain canopy heights and the density of the canopy around this height is important for the individual species and species richness. We believe the advantage of density layer is that they express the density within a certain layer independently of the canopy height in the counting stations whereas canopy density variables are dependent on the maximum height of the canopy. It has been shown that different species relate differently to certain sections of the canopy and therefore will be affected by the density of branches in the particular layer(s) (Alatalo et al. 1987). We have not seen any other studies which have divided the canopy into vertical 1 m bins as we did for the density layers, but our results show that this could be an important variable to include in future bird studies. Vegetation height has turned out to be the most important variable, regardless of forest type, in other studies which used ALS data to predict overall species richness of forest-dwelling birds (Clawges et al. 2007, Goetz et al. 2007, Lesak et al. 2011).

The best explanatory model was created for Goldcrest and explained 27% of the variation in occurrence. In field-based studies Goldcrest have shown to have a preference for relatively dense well grown coniferous stands (Cramp 1992) this corresponds well with the ALS variables included in the model: the 70<sup>th</sup> and 90<sup>th</sup> height percentiles and the density layer between 28 and 29 m. In contrast to our results, Müller et al. (2009) had low predictive power ( $R^2$  close to zero) when modelling occurrence for Goldcrest in spruce forest in Germany, the best variable being standard deviation of mean vegetation height. That differences between important habitat characteristics are found for the same species is not unusual when two study sites are situated several hundred kilometers from each other (e.g. Cramp 1992, 1993). As another example we predicted the occurrence of Willow warbler (*Phylloscopus trochilus*) in a forest landscape dominated by coniferous forest whereas Bellamy et al. (2009) predicted the occurrence of Willow warbler in woodland areas in England dominated by mixed deciduous forest. In both their and our study reliable prediction models were created but a comparison between these models makes no sense. The same is the case for Great tit (*Parus major*) for which the habitat preferences have been investigated in several studies in mixed deciduous forests in England (e.g. Hinsley et al. 2002, Hill et al. 2004). The 2009 study by Müller et al. also looked upon Great tit and since this survey was conducted in an area dominated by spruce, it is easier to compare their results with ours. They registered Great tits in many

different habitats and no specific ALS variables could be related to the presence of Great tit, results which are in accordance with ours. Great tit are considered a generalist species which is present in almost any habitat which have at least some trees or shrubs and it is therefore reasonable that no ALS variables stands out as particularly important.

The best prediction model were developed for the Wren (AUC = 0.8), useful models were constructed for Chaffinch, Robin, Goldcrest, Willow warbler, Tree pipit and species richness. It was not possible to create prediction models for Great tit and Siskin. Only minor differences in AUC or kappa values were found between corresponding models built with ALS, spectral or ALS+spectral variables. For three of the species in our study the best models contained ALS+spectral variables for two species the best model contained ALS variables alone and for one species the best model contained spectral variables alone. It is therefore difficult to state which group of variables is best at predicting the occurrence of birds from our study alone. Goetz et al. (2007) also built models for bird species richness with the same three variable groups and concluded that models built with ALS variables performed better than model built purely on spectral variables whereas no major differences were found between models built only with ALS variables and model built with both ALS and spectral variables. A study by Müller et al. (2009) which compared ALS variables, spectral variables and traditional field measurements, concluded that ALS variables were more useful for modelling the potential abundance of single bird species than were field measurements and spectral variables. If an overall conclusion of which variables groups to use should be based on our study in addition to the results from the study by Goetz et al. and Müller et al. ALS should in my opinion be preferred in future studies.

Irrespective of GLM10 and RF models are evaluated on AUC or kappa, RF models tend to have slightly better predictive ability than the GLM10 models. This result is in accordance with the result by Cutler et al. (2007) which compared GLM and RF models for nest sites for cavity nesting birds. An important distinction between GLM and RF models is RF method applies to classification and prediction and not to creation of explanatory models (Cutler et al. 2007).

For most prediction models obtained for single bird species in this study the predictive ability was relatively low, with AUC values < 0.7 and kappa values below 0.3 (considered as poor to slight fit according to Araújo et al. (2005) and Landis and Koch (1977), respectively). One possible cause for these results is 'Detectability noise' in the response variables (bird occurrence) data. In many cases it is not possible to detect all species

of interest within a certain area of interest e.g. a counting station like in our study. If a species is actually present in a counting station, but not registered, it results in a false absence in the dataset (MacKenzie 2006, Lobo 2010). Lobo (2010) distinguishes between three different kind of absences: (1) contingent absences which is defined as areas which might actually be suitable for the species of interest but which is not occupied due to dispersal limitations, historical factors, local extinctions, biotic interaction or the fact that the habitat is too small, (2) environmental absences is a fact when the environmentally and climatically conditions are not suitable for the species and finally (3) methodological absences caused by bias and scarceness gathering of the training data e.g. preferential sampling. In our case the environmental and methodological absence will play an important role. In addition it is likely that birds will pass undetected if they hide and stay quiet when a counting station is visited (Bibby 2000). Finally, the degree of elusiveness differs between species, time of the year, the time during the day that recordings are made and weather conditions. The experience of the observer can also influence the probability that a certain species is recorded. Although we tried to counteract or minimize these factors in our study design, detectability is likely to be less than 100 % and to vary among species since variation in detectability is unavoidable. Another cause for our results might be the size of the counting stations (7850 m<sup>2</sup>). In large counting stations the explanatory power will be reduced since not all parts of a large site will be equally suitable for a species. In our study area, high within counting station variation in important habitat characteristics is expected to be an important reason for suboptimal prediction models.

When building models which relates birds to their habitat it is important to remember that birds are not solidly focusing on habitat heterogeneity when selecting habitat, they also depend on e.g. inter and intra specific competition (Alatalo et al. 1987), access to food (Hinsley et al. 2008), risk of predation (Møller 1988, Martin 1993) and changes in weather conditions (Hinsley et al. 2006).

## **5. Conclusions**

The main conclusions from this thesis highlight the advantages of implementing ALS data in ecological studies.

In Paper 1 good distribution models were created for boreal swamp forests, when terrain variables computed from a digital elevation model, derived from ALS, were used as

explanatory variables in the novel distribution modelling software: MaxEnt. Not only did we create good distribution models, we also showed how important it is to ask questions about novel methods reliability. We examined the reliability of the default settings in MaxEnt and showed that the use of default settings results in overfitted models. We therefore recommend future studies to use the alternative approach for developing MaxEnt models as outlined in Paper 1. At least future users of MaxEnt, which continues to develop models using the default settings, should be aware of the high risk of creating overfitted models and as a minimum use different settings for the regularisation multiplier (easy to change in the software), and evaluate the models on independent evaluation data to find the least overfitted model, before implementing them into any kind of management plans. To complement the results from Paper 1, we in Paper 2 developed models to distinguish between swamp forest and non-paludified forest solidly on differences in the forest structure. Result from Paper 2 was not as convincing as those from Paper 1, but we show that e.g. mean vegetation height is lower in swamp forests compared to the surrounding forest. The developed models correctly classified between 54 and 62% of the cells when evaluated on independent evaluation data. An obvious future study would be to combine the variables from the best model in Paper 1 with the variables from the best model in Paper 2 and use these terrain and forest structure variables as environmental variables in a new MaxEnt model developed with the alternative MaxEnt procedure. This will most likely improve the best model from Paper 1 since one of the draw backs with models from Paper 1 is that they cannot distinguish between swamp forests and open mire, a task which is easily accomplished if forest structure variables are included in the model. Future studies working with the distribution of nature types in general and boreal swamp forests in particular can benefit from our results.

Paper 3 is the first study which combines ALS data and spectral images with the distribution of birds in Fennoscandia. Explanatory models were successfully created for six single bird species and species richness, no models were built for Great tit and Siskin (*Carduelis spinus*). Within all the explanatory models one or more density layer variables were included indicating that one or more 1 m sections of the canopy were of particular interest for the individual species and overall species richness. We recommend future studies to implement density layer variables as a variable which combines important aspects of both vegetation height and density in one. Good prediction models were created for Goldcrest and Wren, for the other species and species richness only moderate models were built. We argue that even moderate models can add new knowledge to how birds distribute within a certain

area, but it is important to remember that birds are not only focusing on habitat heterogeneity when selecting habitat, they also depend on e.g. inter and intra specific competition, access to food, risk of predation, and changes in weather conditions. If only data from one remote sensing technique, either ALS or spectral images, should be used in future studies of birds we recommend the use of ALS. Future studies which use ALS in bird studies might consider using single-tree variables as an alternative to the area-based approach for acquiring ALS derived variables used in the present study it might though be challenging to merge single-tree variables with standard sampling techniques for birds. It is however, possible that single-species bird studies based on data collected by use of other sampling techniques (e.g. territory mapping), may benefit from using the single-tree approach.

## 6. References

- Alatalo, R. V. et al. 1987. Exploitation competition influences the use of foraging sites by tits - experimental-evidence. — *Ecology* 68: 284-290.
- Anderson, R. P. and Gonzalez, I. J. 2011. Species-specific tuning increases robustness to sampling bias in models of species distributions: an implementation with Maxent. — *Ecological Modelling* 222: 2796-2811.
- Anonymous 2008. ArcGIS, ver. 9.3. — Inc. Redlands, California
- Aplin, P. 2005. Remote sensing: ecology. — *Progress in Physical Geography* 29: 104-113.
- Araújo, M. B. et al. 2005. Validation of species-climate impact models under climate change. — *Global Change Biology* 11: 1504-1513.
- Austin, M. 2007. Species distribution models and ecological theory: A critical assessment and some possible new approaches. — *Ecological Modelling* 200: 1-19.
- Axelsson, P. 2000. DEM generation from laser scanner data using adaptive TIN-models. — *International Archives of Photogrammetry and Remote Sensing* 33: 110-117.
- Bater, C. W. et al. 2009. Estimation of standing dead tree class distributions in northwest coastal forests using lidar remote sensing. — *Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere* 39: 1080-1091.
- Baltsavias, E. P. 1999. Airborne laser scanning: basic relations and formulas. — *Isprs Journal of Photogrammetry and Remote Sensing* 54: 199-214.
- Bellamy, P. E. et al. 2009. Willow Warbler *Phylloscopus trochilus* habitat in woods with different structure and management in southern England. — *Bird Study* 56: 338-348.
- Bergen, K. M. et al. 2007. Multi-dimensional vegetation structure in modeling avian habitat. — *Ecological Informatics* 2: 9-22.
- Bibby, C. J. 2000. Bird census techniques. — Academic Press.
- Breiman, L. 2001. Random forests. — *Machine Learning* 45: 5-32.
- Broughton, R. K. et al. 2006. Marsh Tit *Poecile palustris* territories in a British broad-leaved wood. — *Ibis* 148: 744-752.
- Bässler, C. et al. 2011. LiDAR as a rapid tool to predict forest habitat types in Natura 2000 networks. — *Biodiversity and Conservation* 20: 465-481.
- Clawges, R. et al. 2007. Use of a ground-based scanning lidar for estimation of biophysical properties of western larch (*Larix occidentalis*). — *International Journal of Remote Sensing* 28: 4331-4344.
- Clawges, R. et al. 2008. The use of airborne lidar to assess avian species diversity, density, and occurrence in a pine/aspen forest. — *Remote Sensing of Environment* 112: 2064-2073.
- Coops, N. C. et al. 2010. Assessing the utility of lidar remote sensing technology to identify mule deer winter habitat. — *Canadian Journal of Remote Sensing* 36: 81-88.
- Cramp, S. 1992. Handbook of the birds of Europe, the Middle East and North Africa : the birds of the Western Palearctic. — Oxford University Press.
- Cramp, S. 1993. Handbook of the birds of Europe, the Middle East and North Africa : the birds of the Western Palearctic. — Oxford University Press.
- Cutler, D. R. et al. 2007. Random forests for classification in ecology. — *Ecology* 88: 2783-2792.
- Dormann, C. F. et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. — *Ecography* 30: 609-628.
- Edvardsen, A. et al. 2011. A fine-grained spatial prediction model for the red-listed vascular plant *Scorzonera humilis*. — *Nordic Journal of Botany* 29: 495-504.
- Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. — *Ecography* 29: 129-151.
- Elith, J. et al. 2011. A statistical explanation of MaxEnt for ecologists. — *Diversity and Distribution*. 17: 43-57.

- Erikstad, L. et al. 2013. Impact of Scale and Quality of Digital Terrain Models on Predictability of Seabed Terrain Types. — *Marine Geodesy* 36: 2-21.
- Falkowski, M. J. et al. 2009. Characterizing forest succession with lidar data: An evaluation for the Inland Northwest, USA. — *Remote Sensing of Environment* 113: 946-956.
- Giovanelli, J. G. R. et al. 2008. Predicting the potential distribution of the alien invasive American bullfrog (*Lithobates catesbeianus*) in Brazil. — *Biological Invasions* 10: 585-590.
- Goetz, S. et al. 2007. Laser remote sensing of canopy habitat heterogeneity as a predictor of bird species richness in an eastern temperate forest, USA. — *Remote Sensing of Environment* 108: 254-263.
- Graf, R. F. et al. 2009. Habitat assessment for forest dwelling species using LiDAR remote sensing: Capercaillie in the Alps. — *Forest Ecology and Management* 257: 160-167.
- Graham, C. H. et al. 2004. New developments in museum-based informatics and applications in biodiversity analysis. — *Trends in Ecology & Evolution* 19: 497-503.
- Gobakken, T. and Næsset, E. 2004. Estimation of diameter and basal area distributions in coniferous forest by means of airborne laser scanner data. — *Scandinavian Journal of Forest Research* 19: 529-542.
- Guisan, A. and Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat models. — *Ecology letters* 8: 993-1009.
- Halvorsen, R. 2012. A gradient analytic perspective on distribution modelling. — *Sommerfeltia* 35: 1-165.
- Halvorsen, R. 2013. A strict maximum likelihood explanation of MaxEnt, and some implications for distribution modelling. — *Sommerfeltia* 36: 1-132.
- Hauglin, M. et al. 2013. Estimating single-tree branch biomass of Norway spruce with terrestrial laser scanning using voxel-based and crown dimension features. — *Scandinavian Journal of Forest Research* 28: 456-469.
- Hernandez, P. A. et al. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. — *Ecography* 29: 773-785.
- Hijmans, R. J. and Elith, J. 2011. Species distribution modelling with R. — *The R foundation for statistical computing*.
- Hill, R. A. and Broughton, R. K. 2009. Mapping the understorey of deciduous woodland from leaf-on and leaf-off airborne LiDAR data: A case study in lowland Britain. — *Isprs Journal of Photogrammetry and Remote Sensing* 64: 223-233.
- Hill, R. A. et al. 2004. Predicting habitat quality for Great Tits (*Parus major*) with airborne laser scanning data. — *International Journal of Remote Sensing* 25: 4851-4855.
- Hinsley, S. A. et al. 2002. Quantifying woodland structure and habitat quality for birds using airborne laser scanning. — *Functional Ecology* 16: 851-857.
- Hinsley, S. A. et al. 2006. The application of lidar in woodland bird ecology: Climate, canopy structure, and habitat quality. pp. 1399-1406.
- Hinsley, S. A. et al. 2008. Effects of structural and functional habitat gaps on breeding woodland birds: working harder for less. — *Landscape Ecology* 23: 615-626.
- Hinsley, S. A. et al. 2009. Bird species distributions across woodland canopy structure gradients. — *Community Ecology* 10: 99-110.
- Holmgren, J. et al. 2008. Species identification of individual trees by combining high resolution LIDAR data with multi-spectral images. Taylor & Francis Ltd, pp. 1537-1552.
- Hopkinson, C. et al. 2005. Vegetation class dependent errors in lidar ground elevation and canopy height estimates in a boreal wetland environment. — *Canadian Journal of Remote Sensing* 31: 191-206.
- Hörnberg, G. et al. 1995. Stand dynamics, regeneration patterns and long-term continuity in Boreal old-growth *Picea abies* swamp-forests. — *Journal of Vegetation Science* 6: 291-298.

- Hörnberg, G. et al. 2012. Fire as an important factor for the genesis of boreal *Picea abies* swamp forests in Fennoscandia. — *Holocene* 22: 203-214.
- Hörnberg, G. et al. 1998. Boreal swamp forests. — *Bioscience* 48: 795-802.
- Jung, K. et al. 2012. Moving in three dimensions: effects of structural complexity on occurrence and activity of insectivorous bats in managed forest stands. — *Journal of Applied Ecology* 49: 523-531.
- Kim, Y. et al. 2009. Distinguishing between live and dead standing tree biomass on the North Rim of Grand Canyon National Park, USA using small-footprint lidar data. — *Remote Sensing of Environment* 113: 2499-2510.
- Korpela, I. et al. 2012. Understory trees in airborne LiDAR data — Selective mapping due to transmission losses and echo-triggering mechanisms. — *Remote Sensing of Environment* 119: 92-104.
- Koukoulas, S. and Blackburn, G. A. 2004. Quantifying the spatial properties of forest canopy gaps using LiDAR imagery and GIS. — *International Journal of Remote Sensing* 25: 3049-3071.
- Kraus, K. and Pfeifer, N. 1998. Determination of terrain models in wooded areas with airborne laser scanner data. — *Isprs Journal of Photogrammetry and Remote Sensing* 53: 193-203.
- Kuusinen, M. 1996. Importance of spruce swamp-forests for epiphyte diversity and flora on *Picea abies* in southern and middle boreal Finland. — *Ecography* 19: 41-51.
- Landis, J., R. and Koch, G., C. 1977. The measurement of observer agreement for categorical data. — *Biometrics* 33: 159-174.
- Lefsky, M. A. et al. 2002. Lidar remote sensing for ecosystem studies. — *Bioscience* 52: 19-30.
- Lesak, A. A. et al. 2011. Modeling forest songbird species richness using LiDAR-derived measures of forest structure. — *Remote Sensing of Environment* 115: 2823-2835.
- Liaw, A. and Wiener, M. 2002. Classification and Regression by randomForest. *R News* 2(3), 18-22.
- Lobo, J. M. et al. 2010. The uncertain nature of absences and their importance in species distribution modelling. — *Ecography* 33: 103-114.
- Loetsch, F. and Haller, K. E. 1964. Statistics of forest inventory and information from aerial photographs. — *BLV*.
- Lozier, J. D. et al. 2009. Predicting the distribution of Sasquatch in western North America: anything goes with ecological niche modelling. — *Journal of Biogeography* 36: 1623-1627.
- MacArthur, R. and MacArthur, J. W. 1961. On bird species-diversity. — *Ecology* 42: 594-598.
- MacArthur, R. H. and Wilson, E. O. 1967. *The theory of island biogeography*. — Princeton University Press.
- MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L. and Hines, J. E. 2006. *Occupancy estimation and modeling. Inferring patterns and dynamics of species occurrence*. — Academic Press.
- Martin, T. E. 1993. Nest predation among vegetation layers and habitat types - revising the dogmas. — *American Naturalist* 141: 897-913.
- Martinuzzi, S. et al. 2009. Mapping snags and understory shrubs for a LiDAR-based assessment of wildlife habitat suitability. — *Remote Sensing of Environment* 113: 2533-2546.
- Mateo, R. G. et al. 2010. Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absences and target-group absences from natural history collections. — *Diversity and Distribution* 16: 84-94.
- McDermid, G. J. et al. 2005. Remote sensing for large-area habitat mapping. — *Progress in Physical Geography*. 29: 449-474.
- McRoberts, R. E. et al. 2010. Advances and emerging issues in national forest inventories. — *Scandinavian Journal of Forest Research* 25: 368-381.
- Merckx, B. et al. 2011. Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. — *Ecological Modelling*. 222: 588-597.
- Mikhail, E. M. et al. 2001. *Introduction to modern photogrammetry*. — Wiley.

- Miller, J. R. et al. 2004. Spatial extrapolation: The science of predicting ecological patterns and processes. — *Bioscience* 54: 310-320.
- Moen, A. 1999. National Atlas of Norway: Vegetation. — Norwegian Mapping Authority.
- Møller, A. P. 1988. Nest predation and nest choice in passerine birds in habitat patches of different size - a study of magpies and blackbirds. — *Oikos* 53: 215-221.
- Mücher, C. A. et al. 2009. Modelling the spatial distribution of Natura 2000 habitats across Europe. — *Landscape and Urban Planning* 92: 148-159.
- Müller, J. et al. 2009. Using airborne laser scanning to model potential abundance and assemblages of forest passerines. — *Basic and Applied Ecology* 10: 671-681.
- Müller, J. et al. 2010. Composition versus physiognomy of vegetation as predictors of bird assemblages: The role of lidar. — *Remote Sensing of Environment* 114: 490-495.
- Nagelkerke
- Hinsley, S. A. et al. 2006. The application of lidar in woodland bird ecology: Climate, canopy structure, and habitat quality. pp. 1399-1406.
- Nagelkerke, N. J. D. 1991. A note on a general definition of the coefficient of determination. — *Biometrika* 78: 691-692.
- Næsset, E. et al. 2004. Laser scanning of forest resources: The Nordic experience. — *Scandinavian Journal of Forest Research* 19: 482-499.
- Næsset, E. 2007. Airborne laser scanning as a method in operational forest inventory: Status of accuracy assessments accomplished in Scandinavia. — *Scandinavian Journal of Forest Research* 22: 433-442.
- Næsset, E. and Gobakken, T. 2008. Estimation of above- and below-ground biomass across regions of the boreal forest zone using airborne laser. — *Remote Sensing of Environment* 112: 3079-3090.
- Næsset, E. et al. 2013. Model-assisted estimation of change in forest biomass over an 11 year period in a sample survey supported by airborne LiDAR: A case study with post-stratification to provide "activity data". — *Remote Sensing of Environment* 128: 299-314.
- Ohlson, M. et al. 1997. Habitat qualities versus long-term continuity as determinants of biodiversity in boreal old-growth swamp forests. — *Biological Conservation* 81: 221-231.
- Ohlson, M. et al. 2006. The macroscopic charcoal record in forested boreal peatlands in southeast Norway. — *Holocene* 16: 731-741.
- Økland, R. H. et al. 2001. Vegetation-environment relationships of boreal spruce swamp forests in Østmarka Nature Reserve, SE Norway. *Sommerfeltia*, 29. p 190.
- Økland, R. H. et al. 2008. Species richness in boreal swamp forests of SE Norway: The role of surface microtopography. — *Journal of Vegetation Science* 19: 67-74.
- Ørka, H. O. et al. 2009. Classifying species of individual trees by intensity and structure features derived from airborne laser scanner data. — *Remote Sensing of Environment* 113: 1163-1174.
- Packalén, P. et al. 2012. Variable selection strategies for nearest neighbor imputation methods used in remote sensing based forest inventory. — *Canadian Journal of Remote Sensing* 38: 1-13.
- Parisien, M. A. and Moritz, M. A. 2009. Environmental controls on the distribution of wildfire at multiple spatial scales. — *Ecol. Monogr.* 79: 127-154.
- Pearce, J. and Ferrier, S. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. — *Ecological Modelling* 133: 225-245.
- Pearson, R. G. et al. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. — *Journal of Biogeography* 34: 102-117.
- Peterson, A. T. et al. 2007. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. — *Ecography* 30: 550-560.

- Phillips, S. J. and Dudík, M. 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. — *Ecography* 31: 161-175.
- Phillips, S. J. et al. 2004. A maximum entropy approach to species distribution modeling. — In: Anonymous (ed), *Proceedings of the 21st international conference on machine learning*. ACM Press, pp. 655-662.
- Phillips, S. J. et al. 2006. Maximum entropy modeling of species geographic distributions. — *Ecological Modelling* 190: 231-259.
- Phillips, S. J. 2011. A brief tutorial on Maxent. — AT&T Research, Princeton, NJ
- Raes, N. and ter Steege, H. 2007. A null-model for significance testing of presence-only species distribution models. — *Ecography* 30: 727-736.
- Reutebuch, S. E. et al. 2003. Accuracy of a high-resolution lidar terrain model under a conifer forest canopy. — *Canadian Journal of Remote Sensing* 29: 527-535.
- Richardson, M. C. et al. Analysis of airborne LiDAR surveys to quantify the characteristic morphologies of northern forested wetlands. — *Journal of Geophysical Research-Biogeosciences* 115: 16.
- Riordan, E. C. and Rundel, P. W. 2009. Modelling the distribution of a threatened habitat: the California sage scrub. — *Journal of Biogeography* 36: 2176-2188.
- Santika, T. and Hutchinson, M. F. 2009. The effect of species response form on species distribution model prediction and inference. — *Ecological Modelling* 220: 2365-2379.
- Schwarz, G. 1978. Estimating the dimension of a model. — *Annals of statistics* 6: 461-464.
- Smart, L. S. et al. 2012. Three-dimensional characterization of pine forest type and red-cockaded woodpecker habitat by small-footprint, discrete-return lidar. — *Forest Ecology and Management* 281: 100-110.
- Segerström, U. et al. 1994. Disturbance history of a swamp forest refuge in northern Sweden. — *Biological Conservation* 68: 189-196.
- Sigmond, E. M. O. et al. 1984. Berggrunnskart over Norge 1: 1 000 000. — *Norg. geol. Unders.*
- Sjöberg, K. and Ericson, L. 1992. Forested and open wetland complexes. — In: Hansson, L. (ed), *Ecological principles of nature conservation*. Elsevier, pp. 326-351.
- Strand, H. et al. (eds.) 2007. *Sourcebook on Remote Sensing and Biodiversity Indicators* — Secretariat of the Convention on Biological Diversity, Montreal, p 203.
- Sundberg, S. 1993. wet forests - the Scandinavian "rain forest". — *Fåglar i Uppland* 20: 65-80.
- Sutherland, W. J. et al. 2004. *Bird ecology and conservation. A handbook of techniques*. — Oxford University Press.
- Tattoni, C. et al. 2012. Can LiDAR data improve bird habitat suitability models? — *Ecological Modelling*. 245: 103-110.
- Telenius, A. 2011. Biodiversity information goes public: GBIF at your service. — *Nordic Journal of Botany* 29: 378-381.
- Thieme, N. et al. 2011. Detection of small single trees in the forest-tundra ecotone using height values from airborne laser scanning. — *Canadian Journal of Remote Sensing* 37: 264-274.
- Timonen, J. et al. 2010. Woodland key habitats in northern Europe: concepts, inventory and protection. — *Scandinavian Journal of Forest Research* 25: 309-324.
- Tognelli, M. F. et al. 2009. An evaluation of methods for modelling distribution of Patagonian insects. — *Revista Chilena De Historia Natural* 82: 347-360.
- Turner, W. et al. 2003. Remote sensing for biodiversity science and conservation. — *Trends in Ecology & Evolution* 18: 306-314.
- Vehmas, M. et al. 2009. Identification of boreal forest stands with high herbaceous plant diversity using airborne laser scanning. — *Forest Ecology and Management* 257: 46-53.
- Veloz, S. D. 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. — *Journal of Biogeography* 36: 2290-2299.
- Venables, W. N. and Ripley, B. D. 2002. *Modern applied statistics with S*. — Springer.

- Wang, K. et al. 2010. Remote Sensing of Ecology, Biodiversity and Conservation: A Review from the Perspective of Remote Sensing Specialists. — *Sensors* 10: 9647-9667.
- Warren, D. L. and Seifert, S. N. 2011. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. — *Ecological Applications* 21: 335-342.
- Weber, T. C. and Boss, D. E. 2009. Use of LiDAR and supplemental data to estimate forest maturity in Charles County, MD, USA. — *Forest Ecology and Management* 258: 2068-2075.
- Whittaker, R. J. et al. 2005. Conservation Biogeography: assessment and prospect. — *Diversity and Distribution* 11: 3-23.
- Williams, J. N. et al. 2009. Using species distribution models to predict new occurrences for rare plants. — *Diversity and Distribution* 15: 565-576.
- Wing, B. M. et al. 2012. Prediction of understory vegetation cover with airborne lidar in an interior ponderosa pine forest. — *Remote Sensing of Environment* 124: 730-741.
- Wollan, A. K. et al. 2008. Modelling and predicting fungal distribution patterns using herbarium data. — *Journal of Biogeography* 35: 2298-2310.

# Paper 1



# **MaxEnt modelling of swamp forest distributions: overfitting and importance of spatial autocorrelation in response data**

## **Authors**

John Wirkola Dirksen<sup>1\*</sup>, Rune Halvorsen<sup>2</sup>, Sabrina Mazzoni<sup>2</sup>, Erik Næsset<sup>1</sup>, Terje Gobakken<sup>1</sup> and Mikael Ohlson<sup>1</sup>.

## **Addresses**

<sup>1</sup>) Norwegian University of Life Sciences, Department of Ecology and Natural Resource Management, PO Box 5003, NO-1432 Ås, Norway

<sup>2</sup>) University of Oslo, Natural History Museum, Department of Botany, PO Box 1172 Blindern,  
NO-0318 Oslo, Norway

## **\*) Corresponding author**

John Wirkola Dirksen

E.mail [john.dirksen@umb.no](mailto:john.dirksen@umb.no)

Phone: +47 64 96 57 31

## Abstract

In this study, we investigate the relationship between model selection criteria and potential overfitting in distribution modelling MaxEnt methods. We do this by applying two different approaches when modelling the occurrence of swamp forests, using terrain variables derived from a LiDAR based digital elevation model. Furthermore, effects of spatial autocorrelation in the response and explanatory variables are studied. Parallel models were built by manual forward stepwise selection (AMP models) of variables, using an  $F$ -ratio test as selection criteria, and by the default option for model selection and transformation of variables in the Maxent software (SMP models) by  $\ell_1$ -regularization. A wide range of significance levels  $\alpha$  in  $F$ -ratio tests, and regularization multipliers  $\lambda$  were used. All models were evaluated by use of the AUC criterion, as applied to an independent evaluation data set as well as to the training data. Even though the training data did not contain systematic sampling bias, training AUC was considerably higher than test AUC for all models. Accordingly, the importance of using independent data for evaluation is emphasized. SMP models turned out to be very complex, with up to 240 variables derived from the original 11 variables. AMP models were much simpler, with up to 14 derived variables. No significant differences in performance between SMP and AMP models were found, indicating that the regularisation procedure may not be the main or only reason for MaxEnt's good performance in comparative tests. The concept of overfitting is discussed with reference to these results. We argue that the modelling purpose has to be taken into account when decisions about overfitting of distribution models are made. Additionally, predictive performance did not differ between models made with different numbers of presence observations, indicating that spatial autocorrelation in the response variable did not influence the models' predictive performance. Finally, the most important variables for explaining swamp-forest presence were slope and terrain position. This is consistent with current explanations of swamp forest formation. Thus we conclude that distribution modelling can be a valuable tool in conservation and management of this and other nature-types.

## Introduction

Proper management of biodiversity requires precise information about the distribution of rare or threatened nature types and species. The need for efficiency in locating rare species and nature types (e.g., Edwards et al. 2005, Guisan et al. 2006, Williams et al. 2009, Gogol-Prokurat 2011) is one among several reasons for the rapid growth of distribution modelling as a discipline within ecology and conservation biology over the last decade (Guisan and Zimmermann 2000, Austin 2007, Elith and Graham 2009, Franklin 2010). A wide spectre of statistical methods and software implementations have been developed for distribution modelling (Guisan and Zimmermann 2000, Elith et al. 2006), amongst which the maximum entropy (MaxEnt) modelling method has increased strongly in popularity since it was made available in the free, user-friendly MaxEnt software (Phillips et al. 2004, 2006). The popularity of MaxEnt for distribution modelling has been boosted by the method's consistent ranking among the best in comparative studies (Elith et al. 2006, Hernandez et al. 2006, Phillips et al. 2006, Wisz et al. 2008, Mateo et al. 2010). Another reason for MaxEnt's popularity is the functionality to apply presence-only data (Phillips and Dudík 2008), which means that enormous amounts of digital data now available from natural history collections worldwide, e.g., through the Global Biodiversity Information Facility (GBIF; Telenius 2011), can be utilized (Graham et al. 2004). Furthermore, MaxEnt is reported to produce good prediction models with small sample sizes, i.e., when relatively few presence locations are available (Hernandez et al. 2006, Wisz et al. 2008, Mateo et al. 2010)

Despite documented good performance in practical distribution modelling, MaxEnt is reported to be susceptible to spatial autocorrelation in the response variable and its ability to avoid overfitting of the data has been questioned (Raes and ter Steege 2007, Merckx et al. 2011). A model is commonly regarded as overfitted to the data when its predictive performance is poorer than that of simpler models (Guisan and Thuiller 2005, Merckx et al. 2011, Halvorsen 2013). However, overfitting can be understood in additional ways according to modelling purpose. Halvorsen (2013) distinguishes between two different modelling purposes; spatial prediction modelling (SPM) and ecological response modelling (ERM). The purpose of a distribution modelling study is SPM if models are to be evaluated exclusively by their ability to produce good spatial predictions. Whereas, the purpose is ERM if, the focus of the study is to understanding the relationship between the response variable and environmental variables of importance for the modelled target's distribution. With this in

mind, three types of overfitting can be defined: Type I, that a more complex model has lower predictive performance on independent data than a simpler model; Type II, that a more complex model is similar (in the meaning ‘not significantly better’) in predictive performance on independent data than a simpler model; and Type III, that a more complex model with higher predictive performance on independent data than a simpler model fails to fit realistic overall ecological response curves. Thus, type III is only relevant in terms of ERM purposes (Halvorsen 2012).

Several model selection strategies, or regularization methods (Reineking and Schröder 2006), can be used to fine-tune model complexity with modelling purpose and avoid overfitting to the training data. The standard selection strategy implemented in MaxEnt software is a shrinkage method of the lasso penalty type, often referred to as  $\ell_1$ -regularization (Tibshirani 1996). This procedure is often regarded as one of the main reasons for MaxEnt’s good performance for SPM (e.g., Elith et al. 2006, Phillips and Dudík 2008, Elith et al. 2011). There are, however, indications that MaxEnt models created by default settings tend to be strongly overfit to the data (Raes and ter Steege 2007, Anderson and Gonzalez 2011, Merckx et al. 2011), regardless of modelling purpose (Halvorsen 2013). Comprehensive comparisons with alternative model selection methods, such as the manual forward stepwise selection procedure often used in maximum likelihood estimation approaches (e.g., Hastie et al. 2009), still remain to be carried out.

The main aim of this study was to test the following hypothesis: "The standard MaxEnt procedure (SMP), i.e. models created by use of default settings including  $\ell_1$ -regularization, and with all variables of interest included, are not overfitted, neither from an SPM nor from an ERM point of view, when compared to models created by the alternative MaxEnt procedure (AMP) using stepwise forward selection of variables". A proper test of this hypothesis requires an evaluation test dataset, which is collected independently of the data used to train the model and that consists of presence or absence observations of the modelled target (Austin 2007, Edvardsen et al. 2011, Peterson et al. 2011, Halvorsen 2013). Evaluation data that meet these requirements are, however, rarely collected because of resource constraints (but see, e.g., Newbold et al. 2010, Rebelo and Jones 2010, Edvardsen et al. 2011, Gogol-Prokurat 2011). The main reason why independent data are mandatory for proper evaluation of distribution models is that preferentially sampled presence-only data, which is the most common type of data used to train distribution models (Franklin 2009, Peterson et al. 2011), tend to be strongly biased e.g., towards easily accessible sites (Pearson

et al. 2007, Wollan et al. 2008) which again results in sampling/collection bias (Raes and ter Steege 2007, Merckx et al. 2011, Halvorsen 2012). One type of bias in such data is patchy distribution of presence observations, typically resulting in pairs or groups of presence observations located so close to each other that they are no longer spatially independent (Veloz 2009, Anderson and Gonzalez 2011, Merckx et al. 2011). Independently collected presence/absence data efficiently circumvents all kinds of potentially confounding effects of sampling bias on model evaluation results, including eventual effects of spatial autocorrelation in the training data, how spatial autocorrelation affect the performance of distribution models is still not well known or understood (Dormann et al. 2007, Santika and Hutchinson 2009). We used a sample of presence/absence data collected according to a design in which a priori model predictions were used as auxiliary data in the design phase for testing our second hypothesis: "Spatial autocorrelation in the response variable (presence-only observations for the modelled phenomenon) does not influence the predictive performance of MaxEnt models".

We accomplished these aims by using swamp forests in a study area in SE Norway as our study object. Swamp forests were chosen as study objects because of the considerable conservation interest associated with them (Ohlson et al. 1997, Hörnberg et al. 1998). Motivated by the need for more knowledge about how topography of the terrain and hydrological conditions determines swamp forest occurrence (Sjörs 1948b, Økland 1989, Gorham 1983), for evaluation of the models' appropriateness for the ERM purpose, the following secondary aims were also addressed: to assess the extent to which swamp forest sites can be identified by distribution modelling methods (i.e., an SPM purpose); and to identify key factors determining the occurrence of swamp forests.

## **Material and methods**

### **Study area**

The study was carried out in the Østmarka Nature Reserve (Fig. 1) (59°50' N, 11° 02' E, 190–368 m.a.s.l.), which is located in the southern boreal zone of Southeast Norway (Moen 1999). Annual mean precipitation and temperature (1961–90 normals) at the nearest meteorological stations (Enebakk, ca. 10 km SE of the area, and Hakadal–Bliksrudhagan, ca. 30 km N of the area, respectively) were 816 mm and 3.9 °C, respectively (Aune 1993, Førland 1993).

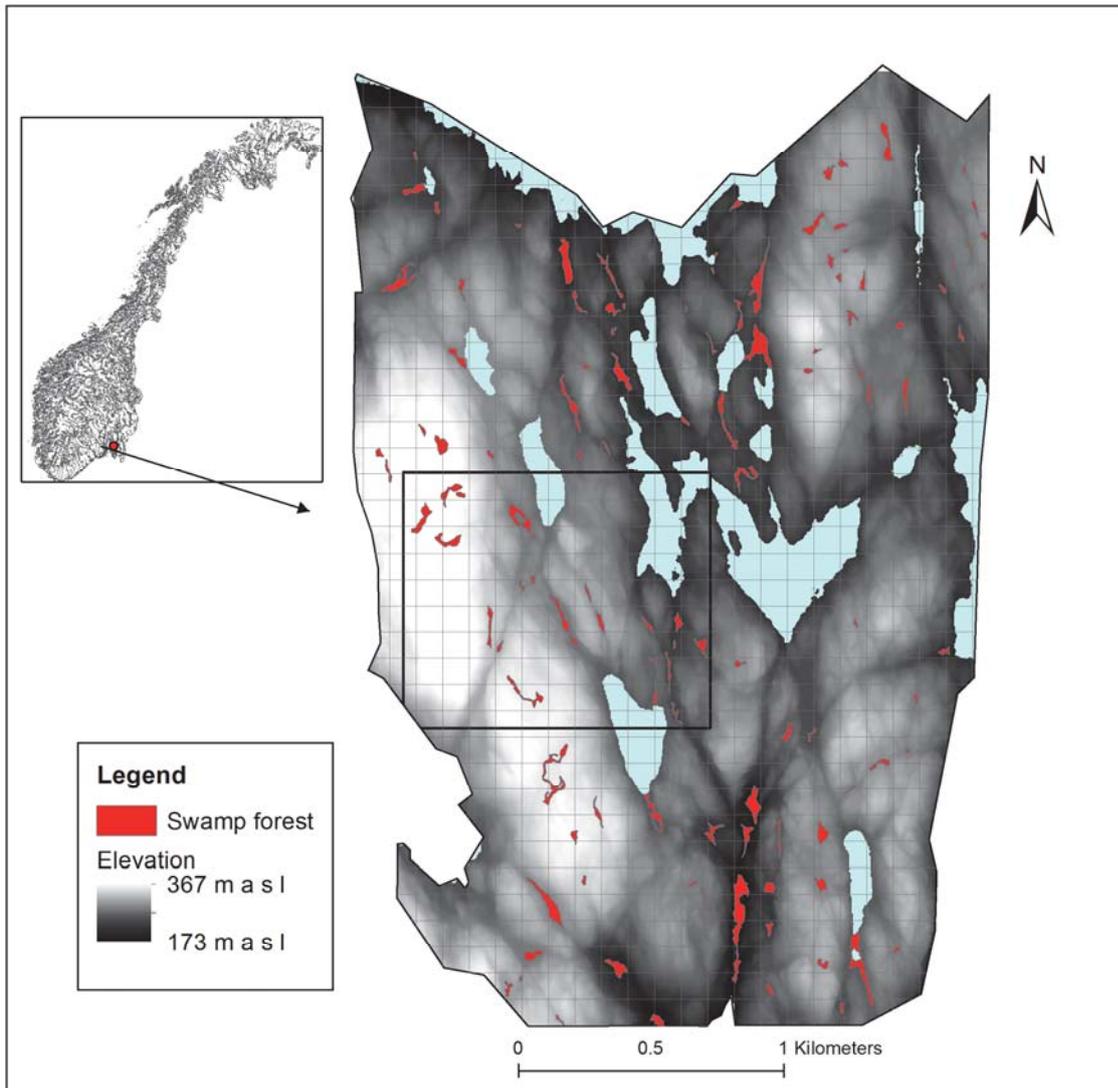


Figure 1. The Østmarka study area and its location in Southeast Norway. The red polygons show the 121 swamp forests recorded by systematic search; these polygons comprise the presence-only training data. A raster (grey lines) with cells of size  $100\text{ m} \times 100\text{ m}$  is superimposed on the map to help illustrate the size of the polygons. Lakes appear in blue. Altitude is indicated by shading; low-lying areas are dark and higher elevations have light grey shading. The black square delimits the area from which a preliminary distribution model was created to extract the cells which should be included in the independent evaluation data set.

The bedrock is of proterozoic age and quartz diorite, tonalite and gneisses of diverse origins dominate (Sigmond et al. 1984). The topography of the study area is dominated by ridges and valleys in the N–S direction. On finer scales, the landscape is dissected by minor valleys, resulting in a broken topography with structures on several scales and a mixture of ridges, valleys, stream beds, large boulders, steep cliffs and more gentle slopes. The study area comprised approximately  $8\text{ km}^2$ .

The study area was conceptualized as a rasterized geographical space with resolution (grid-cell size) of  $5 \text{ m} \times 5 \text{ m}$  which, given the fine-scaled topography of the study area, was expected to capture variation in the topographically conditioned hydrological factors responsible for swamp forest occurrence (cf. Sjörs 1948a, Økland 1989, Økland et al. 2001). The total number of grid cells in the area, cells located in lakes excluded, was 233 774.

### **Presence-only training data**

Swamp forests larger than  $100 \text{ m}^2$  were systematically searched for in the field in the summer and fall of 2010 by a manual protocol. Edges of all swamp forests, that were found, were drawn manually onto a topographic map (1:1000) in the field, the location of the edges were found by use of a handheld Garmin eTrex H Global Positioning System (GPS) receiver. The recorded swamp forest edges were digitized as polygons in a geographical information system project, using ArcMap 9.3 (Anonymous 2008). A total of 121 swamp forest polygons were recorded (Fig. 1).

In order to evaluate the second hypothesis, that spatial autocorrelation in the response variable does not influence the performance of MaxEnt models, we derived two sets of response variables from the digitized polygons: (i) response variable 1, consisting of the 121 grid cells containing the centres of gravity (centroids) of the 121 swamp forest polygons; and (ii) response variable 2, consisting of all 7175  $5 \text{ m} \times 5 \text{ m}$  grid cells in which swamp forest was recorded as present. The first set resembles a typical response variable in species distribution modelling, whereby each locality is represented by one presence observation (often referred to as samples in the literature e.g. Merckx et al. 2011). Whereas the second set, with 7175 presence observations, is meant to represent a very strongly spatially autocorrelated response variable.

### **Environmental explanatory variables**

Airborne laser scanning (ALS) data were acquired for the study area 2 July 2010. Technical details about the scanner and the scanning attributes are shown in Table 1. The initial processing of the data was accomplished by the contractor (Blom Geomatics, Norway). Planimetric coordinates (x and y) and ellipsoidal height values were computed for all echoes. Ground echoes were found and classified using the progressive Triangular Irregular Network (TIN) densification algorithm (Axelsson 2000) of the TerraScan software (Anon. 2005).

Table 1. Technical details for the ALS acquisition.

	Technical specifications
Platform	PA31 Piper Navajo
Sensor	Optech ALTM-Gemini
Mean flying speed (m s <sup>-1</sup> )	80
Flying altitude above ground (m)	900
Side overlap (%)	55
Scan frequency (Hz)	55
Swath width (m)	440
Max. number of echoes per pulse	4
Pulse density (m <sup>-2</sup> )	5
Half-scan angle (degrees)	13.8
Footprint diameter (cm) <sup>a</sup>	19
Pulse repetition frequency (kHz)	100
Date of acquisition	2 July 2010
Contractor	Blom Geomatics AS, Norway

<sup>a</sup> Computed after (Baltsavias) 1999 based on mean acquisition settings.

A TIN was created from the planimetric coordinates and corresponding heights of the laser echoes classified as ground points. The ellipsoidal height accuracy of the TIN model was expected to be around 20-30 cm (Kraus and Pfeifer 1998, Reutebuch et al. 2003). The TIN was converted to a 5 m x 5 m raster digital elevation model (DEM). Descriptors of terrain shape, obtained for each cell in the 5 m x 5 m raster, were therefore used as environmental explanatory variables (EVs) in the modelling of swamp forest distribution in the study area. A total of 11 EVs were used in this study (Table 2).

### **Maximum entropy modelling (MaxEnt)**

The MaxEnt software (Phillips et al. 2006, Phillips and Dudík 2008, Hijmans and Elith 2011), version 3.3.3k (Phillips 2011), was used for MaxEnt modelling. MaxEnt was run in batch mode via a set of customised R scripts (Mazzoni et al., in prep) to model the distribution of swamp forests in the study area as a response to the 11 EVs. The MaxEnt method is based on the maximum entropy principle (Jaynes 1957b, Jaynes 1957a, Jaynes 1982) by which a target probability distribution is estimated by finding the probability distribution of maximum entropy. Essentially, MaxEnt finds the combination of variables which best describes the distribution of the modelled target. This method can be explained from a machine-learning perspective (Phillips et al. 2004, 2006, Phillips & Dudík 2008 and Elith et al. 2011) or from a maximum likelihood estimation perspective (Halvorsen 2013).

Table 2. Description of the environmental explanatory variables (EVs) used in this study.

Variable Name	Description
Slope	Calculated from the difference in height between a single cell and its eight neighbor cells.
Topographic Position Index (TPI) for three different neighborhoods: TPI3, TPI5, and TPI10	Position of cell relative to surrounding cells; positive values indicate convex terrain. The TPI index (Jenness 2006) is naturally scale dependent; i.e., if a cell is considered to be convex or concave depends on the scale used to analyse the landscape. The scale addressed by the index is determined by the neighborhood used in the analysis. The TPI values reflect the difference between the elevation in a particular cell and the average elevation of the cells around that cell. Neighborhoods of 3, 5, and 10 cells, denoted TPI3, TPI5, and TPI10, were used.
Vector Ruggedness Measure (VRM) for three different neighborhoods: VRM3, VRM5, and VRM10	VRM combines slope and aspect into a single measure that describes the ruggedness of the terrain around a cell (Sappington et al. 2007). VRM is naturally scale dependent in the same way as TPI. Neighborhoods of 3, 5, and 10 cells denoted VRM3, VRM5, and VRM10, respectively, were used.
Plan curvature	The curvature at a cell perpendicular to the slope (Kimerling et al. 2009).
Profile curvature	The curvature at a certain cell in the direction of the slope (Kimerling et al. 2009).
Curvature	The curvature of a raster surface, i.e., a combination of plan and profile curvatures (Kimerling et al. 2009).
Flow accumulation	The accumulated weight of all cells flowing into each down slope cell in the output raster.

### Modelling procedures

The main steps of MaxEnt modelling, after presence only data for the modelled target and EVs have been prepared, are (Halvorsen 2013): (1) to transform EVs to derived explanatory variables (DVs; for a full explanation of how the DVs are derived refer to appendix B); (2) to choose a model selection strategy, or regularisation method; and (3) to choose a method for internal model performance assessment to be used in step (2); and (4) to parameterize the model. The transformation step (1) is required to open for modelling nonlinear relationships between the modelled target and the EVs. The choice of model selection strategy (2) implies choosing one specific among many alternative models; model selection is typically carried out by an iterative process by which the selected model performance statistic (3) is used to assess the contribution of EVs, directly or via the DVs that represent the EVs, and, thus, to rank alternative models. A specific threshold value for the performance statistic, the model improvement criterion, may be used to test the hypothesis that more complex models do not improve a simpler reference model significantly.

The main aim of this study is to compare MaxEnt models obtained by two different model selection procedures, each with different settings (tuning parameters) for several options (variants of the method). These two procedures are referred to as the alternative and the standard MaxEnt procedures, AMP and SMP, respectively.

AMP models were parameterised using the aforementioned R-scripts for stepwise forward variable selection. This model selection method includes three major phases (see appendix A): (1) derivation of a parsimonious set of DVs from each EV; this set was used to represent the EV in the next model-building phases; (2) building a MaxEnt model without interactions by sequentially adding EVs until no more EVs can be found that increases model performance; and (3) building a MaxEnt model with interactions by sequentially adding product variables [products between DVs representing the EVs included in the final model from phase (2)] to the final model from phase (2) until no more EVs can be found that increases model performance. The *F*-ratio test developed for MaxEnt by Halvorsen (2013) was used for comparison of nested MaxEnt models, with significance level  $\alpha$  as model improvement criterion. The AMP models were obtained by use of Maxent software by setting the regularisation multiplier to zero ( $\lambda = 0$ ), thus detaching the  $\ell_1$ -regularisation procedure. AMP models were obtained for all combinations of the two response variables (121 samples and 7175 samples) and seven model improvement criteria ( $\alpha_1 = 0.25$ ,  $\alpha_2 = 0.1$ ,  $\alpha_3 = 0.05$ ,  $\alpha_4 = 0.01$ ,  $\alpha_5 = 1.0e-03$ ,  $\alpha_6 = 1.0e-04$ ,  $\alpha_7 = 1.0e-05$ ,  $\alpha_8 = 1.0e-07$ ,  $\alpha_9 = 1.0e-11$ ).

For the AMP model with the highest test AUC (i.e., upon evaluation on independent evaluation data), the sequence of models obtained at the end of each set expansion step (step 3c in appendix A) was subjected to evaluation by the independent evaluation data set in order to investigate if a more complex AMP model performed better than simpler models (models with fewer DVs).

SMP models were obtained by use of default settings in the Maxent software, by which transformation of EVs into DVs of six feature types is performed as an integrated part of the model parameterisation procedure and model selection is carried out by  $\ell_1$ -regularisation or the lasso penalty (Tibshirani 1996), a shrinkage method which evaluates DVs as such without taking into account from which EVs they are derived (see Phillips & Dudík 2008, Elith et al. 2011 and Halvorsen 2013 for a more detailed description). SMP models were obtained for all combinations of the two response variables (121 cells or 7175 cells) and six values of the regularisation multiplier ( $\lambda_1 = 1$ ,  $\lambda_2 = 2$ ,  $\lambda_3 = 4$ ,  $\lambda_4 = 8$ ,  $\lambda_5 = 16$ ,  $\lambda_6 = 32$ ).

For each model (both AMP and SMP), the number  $m$  of DVs with nonzero parameters, the number of EVs from which these DVs were derived, and the number of interaction (product) variables in the final model, were retrieved from the NN.lambdas output file produced by the Maxent software (Warren & Seifert 2011). The 'raw' output format was used for both AMP and SMP models. Both AMP and SMP models were evaluated on an independent presence-absence evaluation data set.

### **Evaluation of MaxEnt models by use of independent presence-absence data**

An evaluation dataset that consisted of true presence-absence information was obtained by visiting 203 grid cells in the field in the fall of 2011. Positions for cells to be included in the evaluation data set were obtained in the following way. (1) A pilot distribution model I was obtained by use of logistic regression [GLM with binomial errors (Venables and Ripley 2002) with manual forward stepwise selection among individually significant variables, as implemented in R version 2.14.1 (R Core Development Team 2009)], following recommendations of Wollan et al. (2008). All 11 EVs were used untransformed. (2) EVs that were found to be significant in pilot model I were used to obtain a pilot MaxEnt model II for swamp forests in the study area by the SMP procedure ( $\lambda = 1$  and default settings for all other options). The pilot models were obtained by use of presence observations for swamp forests in 524 5 m  $\times$  5 m grid cells in the 25 252 grid cells located on land in a 1-km square centrally placed in the study area (Fig. 1). This area was chosen because it contained the full range of variation in elevation and terrain shapes encountered in the study area (results not shown). Model predictions (RPPP values in raw output format) from pilot model II were obtained for the entire study area. (3) We divided the set of RPPP values for all grid cells into nine classes, each class spanning a fixed RPPP range. (4) Potential positions for grid cells to be included in the evaluation data set were selected as follows: (i) Starting with class one (the highest RPPP values) and proceeding till class seven, 10 cells were selected at random from each class. To avoid spatial autocorrelation in the evaluation data set, cells were not allowed to be located within a buffer zone with radius = 40 m around an already chosen cell. (ii) Additional, 15 cells were selected from each of classes eight and nine. A higher number of cells from these two classes were included in the evaluation data set because they represented a higher proportion of the total number of cells. Procedures (i) and (ii) were repeated four times to select a total of 400 cells, all of which were visited in the field. Recordings of presence or

absence of swamp forests in the 203 grid cells that could be located with a precision of  $\pm 2$  cm by use of a Real Time Kinematic (RTK) GPS made up the evaluation data set.

We used the AUC value (the area under the receiver operating characteristic (ROC) curve, derived from the test data) to evaluate the performance (Pearce and Ferrier 2000) of all AMP and SMP models. For the AMP model with the highest test AUC, the entire sequence of models obtained at the end of each set expansion step (see appendix A Step 3c) was subjected to evaluation, in order to compare the performance of AMP models along a model complexity gradient. AUC values were obtained for the evaluation as well as the training data sets, which are referred to as test AUC and training AUC, respectively. The expected (test or training) AUC value for a model that assigns RPPP values randomly to grid cells, regardless of presence or absence of the modelled target, is 0.5. AUC values are often translated into qualitative characterizations of model goodness-of-fit.

Following Lobo et al. (2008), AUC is often stated to be a suboptimal model performance criterion for distribution models. We used AUC because it evaluates model performance over all possible thresholds for conversion of RPPP values to binary predictions, with equal weight to predicted presences and predicted absences. These properties of the AUC are in line with the general purpose of comparing models in our study, for which none of the five reasons of Lobo et al. (2008) for not using AUC apply: (1) that AUC ignores if the model is well calibrated [the term model calibration is used for the degree to which the numerical accuracy of model predictions corresponds to real probabilities of presence; cf. Pearce & Ferrier (2000)]; model calibration is irrelevant to our aim which is to evaluate the ability of different models to separate presence from absence grid cells, i.e. to provide RPPP values that *rank* grid cells so that real presence grid cells have higher RPPP values than real absence cells, because the ranking of RPPP is independent of how well the model is calibrated; (2) that AUC summarises model performance over regions of the ROC space in which one rarely operates, and (3) that AUC weighs omission and commission errors equally is irrelevant because when the prevalence of the modelled target is unknown and the models' general performance is addressed, no reasons then exist for unequal weighting of the two types of errors, or, which is equivalent, to emphasise some regions in the ROC space above others; (4) that AUC does not give information about the spatial distribution of model errors, which is irrelevant because we want to evaluate the overall predictive performance of the models; and (5) that the AUC is influenced by the total extent of the study area is irrelevant for our study because the extent is kept constant throughout the study.

## Response functions

In order to judge the ERM performance of the different models and, notably, to detect Type III overfitting, response functions for swamp forest to selected EVs were calculated from model output in 'probability-ratio' output format. Generalized additive models (GAMs; e.g., Hastie et al. 2009) with up to 10 degrees of freedom were used to model  $\hat{q}$  as a function of each EV.

## Results

AUC values and additional information for all 30 MaxEnt models are summarized in Table 3. The labeling of models in Table 3 is used throughout the result chapter when specific models are referred to.

### Relationship between model selection procedure and test AUC

The SMP model with highest test AUC (model No. 3, AUC = 0.8684) was built for 121 samples and regularisation multiplier  $\lambda = 4$ . For SMP models built for 7175 samples, the highest test AUC, AUC = 0.8633, was obtained for model No. 21, with  $\lambda = 32$ . The AMP model with highest test AUC, AUC = 0.8627, was model No. 13, built for 121 samples by internal performance criterion  $\alpha = 1 \cdot 10^{-5}$  in the  $F$ -ratio test, while for AMP models built for 7175 samples the highest test AUC = 0.8618 was obtained for model 30, with  $\alpha = 1 \cdot 10^{-11}$ . A majority of SMP models created with 121 samples and AMP models created with 7175 samples had test AUC > 0.86.

Among models built for 121 samples, four out of six SMP models (models 2, 3, 4 and 5) had higher test AUC than the best out of nine AMP models (model No. 13). The difference in test AUC between SMP and AMP models created with 121 samples was barely significant (Two-sample unpaired Wilcoxon-Mann-Whitney test against the two-tailed alternative hypothesis:  $p = 0.045$ ,  $n = 30$ ). Among models built for 7175 samples, the highest test AUC was found for SMP model No. 21 while, ranked by test AUC, the next five models (models 25, 26, 27, 28 and 30) were AMP models. The difference in test AUC between model selection procedures for models built for 7175 samples was not significant (Wilcoxon-Mann-Whitney test:  $p = 0.155$ ,  $n = 30$ ).

## Relationship between sample size and test AUC

Five out of the six SMP models built for 121 samples had higher test AUC than the corresponding SMP model built for 7175 samples, only for  $\lambda = 32$  did a model built with 7175 samples perform better than the corresponding model built for 121 samples (Fig. 2, Table 3; Two-sample paired Wilcoxon-Mann-Whitney test against the two-tailed alternative hypothesis:  $p = 0.063$ ,  $n = 12$ ).

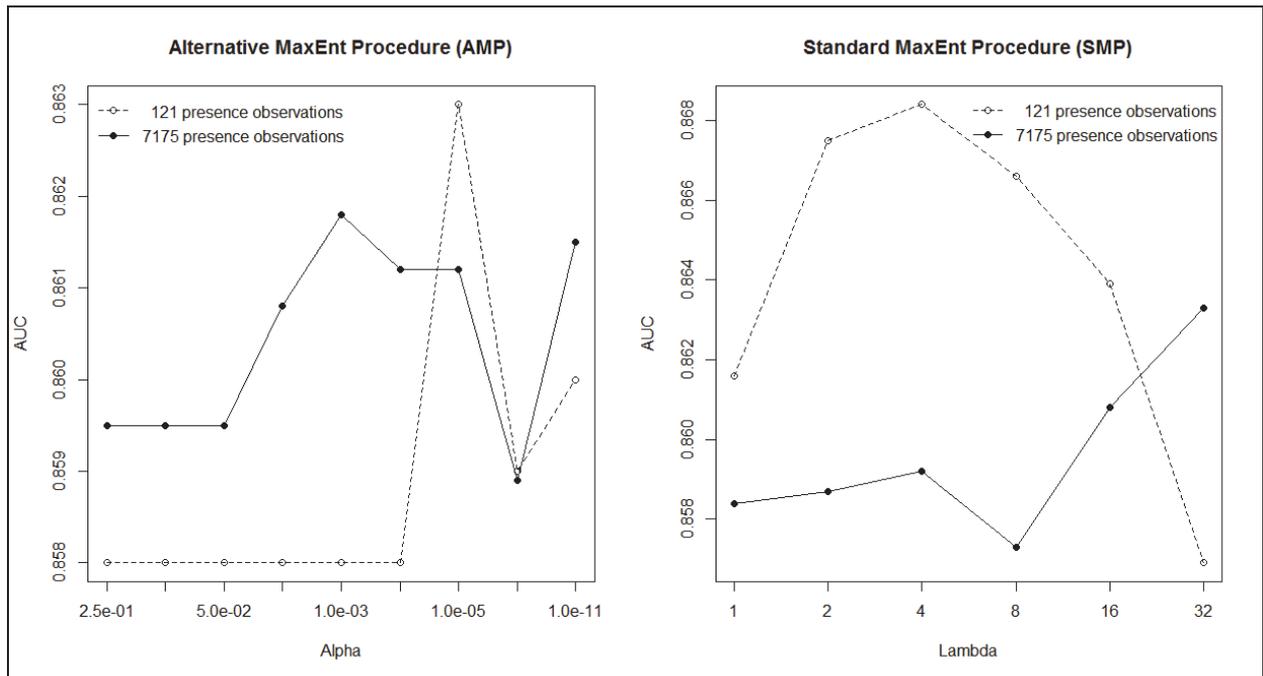


Figure 2. The effect of model selection strategy on model performance, measured as test AUC. The two model selection strategies are the standard MaxEnt procedure (SMP), with different regularisation parameter values, lambda  $\lambda$ , and the alternative MaxEnt procedure (AMP), by which explanatory variables are selected by a forward selection procedure, using the  $F$ -ratio test with different significance levels, alpha  $\alpha$ , as model comparison criterion. For both model selection procedures, results are shown for models built by use of 121 and 7175 presence observations.

The opposite pattern was found for AMP models (Fig. 2, Table 3): eight out of nine AMP models built for 7175 samples had higher test AUC than the corresponding AMP model built for 121 samples (Two-sample paired Wilcoxon-Mann-Whitney test against the two-tailed alternative hypothesis:  $p = 0.017$ ,  $n = 18$ ). Only one 121-sample model (model No. 13) had a clearly higher test AUC than the corresponding 7175-sample model, but the test AUC of this model was the highest test AUC encountered among all AMP models (Table 3). No consistent overall difference in test AUC was found between models built for 121 and 7175 samples, when the 15 pairs of models built for different modelling procedures were compared

(Two-sample paired Wilcoxon-Mann-Whitney test against the two-tailed alternative hypothesis:  $p = 0.712$ ,  $n = 30$ ).

Table 3. MaxEnt models for swamp forests in Østmarka, based upon 121 and 7175 presence observations and built using two contrasting model selection procedures: The standard MaxEnt procedure (SMP) refers to models obtained by use of  $\ell_1$ -regularisation with different values for the regularization multiplier  $\lambda$  (lambda). The alternative MaxEnt procedure (AMP) refers to models obtained by forward selection of explanatory variables and derived variables, with different values for the significance level  $\alpha$  (alpha) in an  $F$ -ratio test, used as internal performance criterion. Training and test AUC values refer to AUC values obtained for the data set used to build the model, and for the independent evaluation data set, respectively. Presence observations = the number of presence observations used to build the model. EV = number of explanatory variables represented by DVs in the final model; DV = number of variables derived from EVs; P = product (interaction) variables. R = model rank according to test AUC.

No	Model procedure	Presence observations	Lambda ( $\lambda$ ) Alpha ( $\alpha$ )	Training AUC	Test AUC	AUC difference	EV	DV	P	R
1	SMP	121	$\lambda = 1$	0.9666	0.8616	0.1050	11	101	1	8
2	SMP	121	$\lambda = 2$	0.9266	0.8675	0.0591	10	51	3	2
3	SMP	121	$\lambda = 4$	0.9570	0.8684	0.0886	9	25	4	1
4	SMP	121	$\lambda = 8$	0.9250	0.8666	0.0584	8	13	4	3
5	SMP	121	$\lambda = 16$	0.9512	0.8639	0.0873	9	13	5	4
6	SMP	121	$\lambda = 32$	0.9231	0.8569	0.0662	7	9	3	30
7	AMP	121	$\alpha = 0.25$	0.9456	0.8584	0.0872	7	14	6	22
8	AMP	121	$\alpha = 0.1$	0.9456	0.8584	0.0872	7	14	6	23
9	AMP	121	$\alpha = 0.05$	0.9456	0.8584	0.0872	7	14	6	24
10	AMP	121	$\alpha = 0.01$	0.9456	0.8584	0.0872	7	14	6	25
11	AMP	121	$\alpha = 0.001$	0.9456	0.8584	0.0872	7	14	6	26
12	AMP	121	$\alpha = 0.0001$	0.9456	0.8584	0.0872	7	14	6	27
13	AMP	121	$\alpha = 0.00001$	0.9429	0.8627	0.0802	6	9	5	6
14	AMP	121	$\alpha = 0.0000001$	0.9429	0.8589	0.0840	6	9	4	19
15	AMP	121	$\alpha = 1E-11$	0.9427	0.8596	0.0831	6	9	3	14
16	SMP	7175	$\lambda = 1$	0.9465	0.8584	0.0881	11	240	2	28
17	SMP	7175	$\lambda = 2$	0.9216	0.8587	0.0629	11	211	2	21
18	SMP	7175	$\lambda = 4$	0.9388	0.8592	0.0796	11	122	3	18
19	SMP	7175	$\lambda = 8$	0.9199	0.8573	0.0626	8	100	3	29
20	SMP	7175	$\lambda = 16$	0.9338	0.8608	0.0730	7	46	4	12
21	SMP	7175	$\lambda = 32$	0.9186	0.8633	0.0553	9	28	6	5
22	AMP	7175	$\alpha = 0.25$	0.9167	0.8595	0.0572	7	12	5	15
23	AMP	7175	$\alpha = 0.1$	0.9167	0.8595	0.0572	7	12	5	16
24	AMP	7175	$\alpha = 0.05$	0.9167	0.8595	0.0572	7	12	5	17
25	AMP	7175	$\alpha = 0.01$	0.9165	0.8608	0.0557	8	11	6	13
26	AMP	7175	$\alpha = 0.001$	0.9166	0.8618	0.0548	7	10	6	7
27	AMP	7175	$\alpha = 0.0001$	0.9155	0.8612	0.0543	8	11	5	10
28	AMP	7175	$\alpha = 0.00001$	0.9155	0.8612	0.0543	8	10	5	11
29	AMP	7175	$\alpha = 0.0000001$	0.9149	0.8589	0.0560	8	11	4	20
30	AMP	7175	$\alpha = 1E-11$	0.9147	0.8615	0.0532	7	12	4	9

### **Relationship between number of EVs, DVs and test AUC**

AMP models included significantly fewer DVs than did SMP models (Two-sample unpaired Wilcoxon-Mann-Whitney test against the two-tailed alternative hypothesis:  $p = 0.0008$ ,  $n = 30$ ), while the number of EVs varied little (from 6 to 11) among models (Two-sample paired Wilcoxon-Mann-Whitney test against the two-tailed alternative hypothesis:  $p = 0.784$ ,  $n = 12$ ). In SMP models, the number of EVs and DVs included in models decreased consistently with increasing regularisation parameter  $\lambda$ , from 11 and 11 EVs, and 101 and 240 DVs for  $\lambda = 1$  with 121 and 7175 samples, respectively, to 7 and 9 EVs, and 9 and 28 DVs, respectively, with  $\lambda = 32$ . SMP models built for 7175 samples consistently included more DVs than the corresponding models built for 121 samples (Two-sample paired Wilcoxon-Mann-Whitney test against the two-tailed alternative hypothesis:  $p = 0.031$ ,  $n = 12$ ). In contrast, AMP models included 6–8 EVs and 9–14 DVs, regardless of significance level  $\alpha$  used as internal performance criterion or the number of samples upon which the model was built (Table 3).

The large number of DVs in SMP models with low  $\lambda$  was due to inclusion of many hinge- and threshold-type DVs, while the number of interaction (product) variables never exceeded 6 (Table 3). The number of interaction variables was significantly higher in AMP than in SMP models (3–6 and 1–6, respectively; Two-sample unpaired Wilcoxon-Mann-Whitney test against the two-tailed alternative hypothesis:  $p = 0.0008$ ,  $n = 30$ ). No relationship between the number of DVs in models and test AUC was found (Kendall's rank correlation coefficient:  $\tau = -0.0986$ ,  $p = 0.468$ ,  $n = 30$ ).

### **Relationship between strictness of the model selection criterion and test AUC**

Test AUC varied very little as a function of strictness of the model selection criterion (the regularisation parameter  $\lambda$  in SMP models and the significance level  $\alpha$  of the  $F$ -test in AMP models, respectively) within the series of SMP and AMP models (Fig. 2). No general relationship between test AUC and each of  $\lambda$  and  $\alpha$  was found (Fig. 2). For AMP models, test AUC peaked at intermediate  $\alpha$  values ( $\alpha = 0.001$  and  $\alpha = 1 \cdot 10^{-5}$  for models built for 7175 and 121 samples, respectively; Fig. 2), while for SMP models, test AUC was unimodally related to  $\lambda$  and peaked for  $\lambda = 4$  for the model built for 121 samples, while for the model built for 7175 samples the highest test AUC was obtained for the strictest regularization,  $\lambda = 32$ .

### **Relationship between training AUC and test AUC**

All models had values for training AUC  $> 0.91$ . Training AUC values were consistently more than 0.05 units higher than the corresponding test AUC values (Table 3). The highest training AUC values were found for the most complex SMP models, i.e. models with regularisation parameter  $\lambda = 1$ , regardless of the number of samples used to build the model (Table 3). For corresponding SMP models built on 121 and 7175 samples, no significant difference between training and test AUC was found (Two-sample paired Wilcoxon-Mann-Whitney test against the two-tailed alternative hypothesis:  $p = 0.156$ ,  $n = 12$ ). AMP models built for 121 samples had consistently higher difference between training and test AUC than models built for 7175 samples (Two-sample unpaired Wilcoxon-Mann-Whitney test against the two-tailed alternative hypothesis:  $p = 0.009$ ,  $n = 18$ ). No general relationship between training and test AUC was found (Kendall's rank correlation coefficient:  $\tau = 0.1069$ ,  $p = 0.630$ ,  $n = 30$ ).

### **Comparison among AMP models that make up a sequence of increasing model complexity**

The best performing AMP model, as evaluated by test AUC was the 6-EV model obtained with interaction (0.8627; Fig. 3). The number of DVs with nonzero coefficients was the same, 9, in both models because five of the nine DVs in the 6-EV model without interactions obtained zero-valued coefficients in the model with interaction DVs. Interestingly, the 4-EV model which included Slope, TPI05, Curv\_plan and VRM05, had a higher test AUC than did the 6-EV model without interactions; test AUC values were 0.8527 and 0.8519, respectively (Fig. 3). Test AUC was reduced to 0.8402 for the 3-EV model (not including VRM05), while remaining almost unchanged for the 2-EV model (not including VRM05 and Curv\_plan, AUC = 0.8385, Fig. 3). Even the model with Slope as the only EV had a fairly high test AUC value (AUC = 0.7901; Fig. 3). Training AUC followed a slightly different pattern, with very gentle reduction from the 6-EV model without interactions (training AUC = 0.9361) to the 2-EV model (training AUC = 0.9280). The model with Slope as the only EV had training AUC = 0.8888.

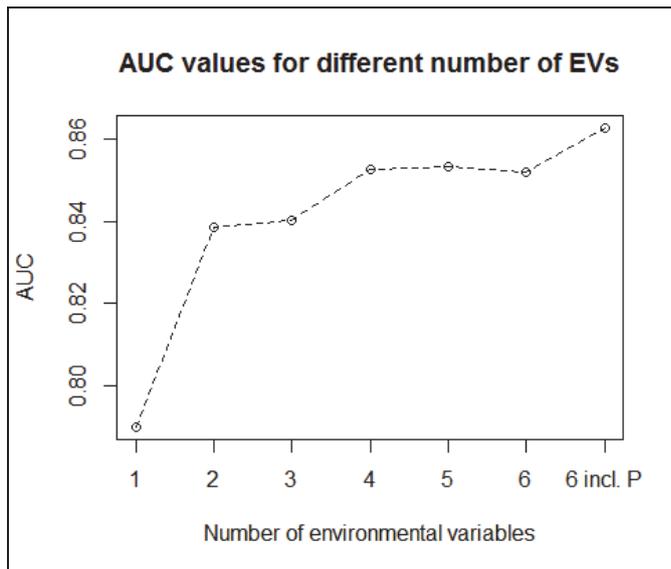


Figure 3. The relationship between number of environmental variables (1–6) included in AMP models created with 121 samples, using  $\alpha = 10^{-7}$  in the  $F$ -ratio test as internal performance criterion, measured as test AUC. Interaction (product variables) is allowed in the '6 incl. P' model only.

### Response functions

The modelled response functions for swamp forests with respect to the four main variables (Slope, TPI05, Curv\_plan and VRM05) for the 2-EV, 4-EV, best AMP and the best SMP models are shown in Fig. 4. The range of output values  $\hat{q}$  and the complexity of response-curve shapes increase with increasing model complexity (towards the right in Fig. 4). Suitability increased slightly towards steeper slopes (upper row in Fig. 4 for the 2-EV model), while a monotonous response to slope is shown for all other models, most strongly for the 4-EV model. Response curves to TPI and Curv\_plan changed from near indeterminate (weak trends) for the 2-EV model to distinctly hinge-shaped and distinctly unimodal, respectively, when Curv\_plan and VRM5 were included (i.e., from 2-EV to 4-EV models). Except for small increases in the complexity of response-curve shapes, response curves differed very little between the 4-EV, the best AMP and the best SMP models.

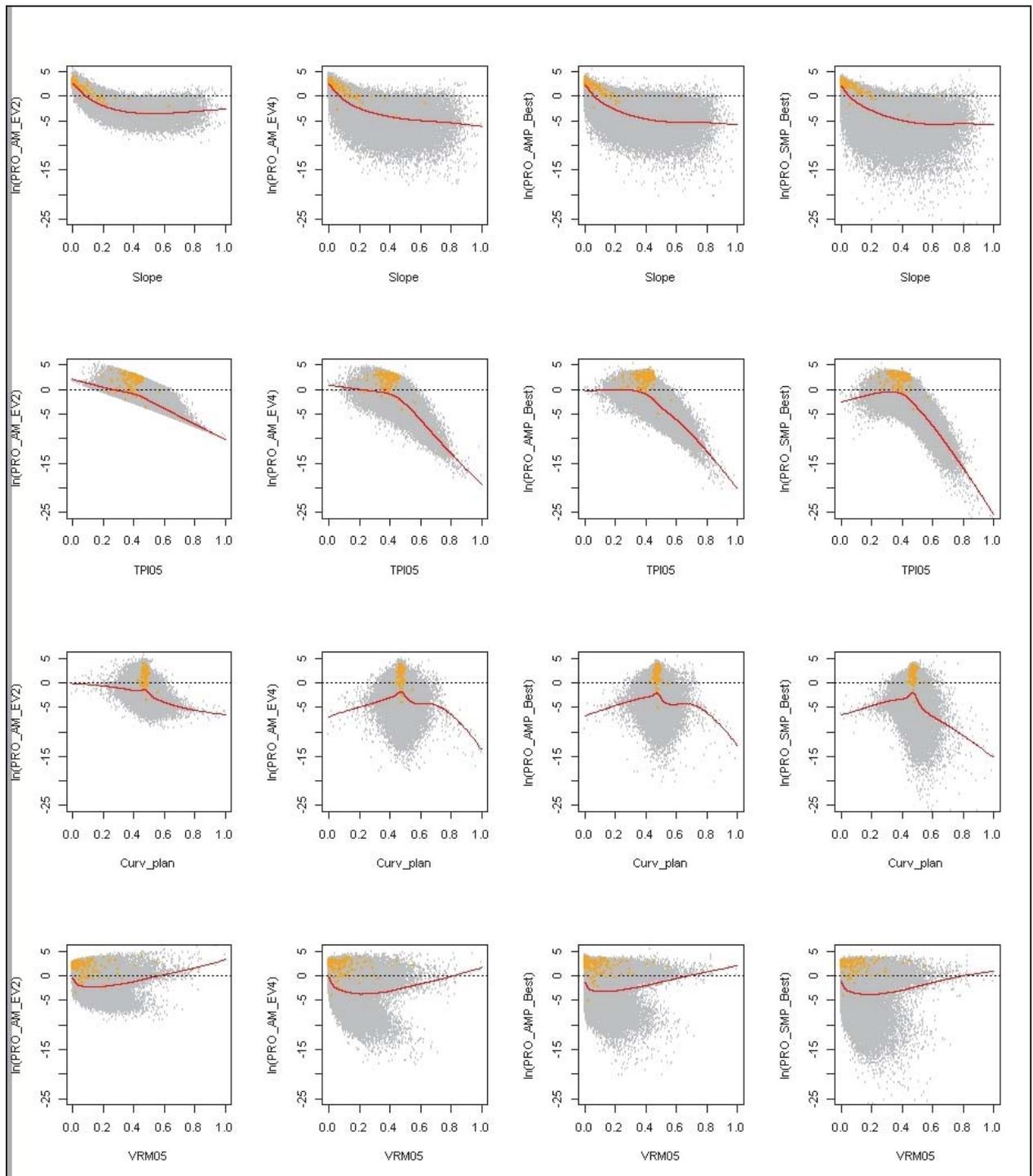


Figure 4. Response functions for swamp forests with respect to four environmental variables (EVs), Slope, TPI05, Curv\_plan and VRM05, based upon probability-ratio output  $\hat{q}$  for 2-EV, 4-EV, best AMP and best SMP models based upon 121 presence observations. Orange dots = presence observations; Grey dots = uninformed background observations; red line = trendline obtained by GAM with 10 degrees of freedom allowed.

## Discussion

### Overfitting of MaxEnt models

Consistent differences in the number of environmental variables (EVs) and derived variables (DVs), and in training and test AUC values, were found among alternative MaxEnt procedure (AMP) models and standard MaxEnt procedure (SMP) models built by use of the same 11 EVs and the same sets of presence observations, but that differed with respect to regularization criteria [alpha ( $\alpha$ ) and lambda ( $\lambda$ ) values, respectively, for the model improvement criteria used in the model selection processes]. Most notably, SMP models obtained by weak regularization (low  $\lambda$  value) included the largest number of DVs and at the same time had lowest test AUC values. This indicates that models built with the Maxent software's default settings, include considerably more EVs and DVs than needed to produce a model with good predictive capability, and hence are overfitted by any definition. Widely used definitions of overfitting are: 1) when models have lower predictive ability on independent data than a simpler model (e.g., Hastie et al. 2009) and 2) that they are overfit if predicting poorer than a model of comparable complexity (Merckx et al. 2011, Halvorsen 2012).

Judgement of overfitting of distribution models should, however, take the purpose of modelling into account (Jiménez-Valverde et al. 2008, Stokland et al. 2011, Halvorsen 2012). Most notably, the distinction between SPM and ERM made by Halvorsen (2012) is important. Models which suffer from Type I overfitting are considered poor models both from the SPM and ERM perspectives. From the gradient analytic perspective on distributions of species and nature types (Halvorsen 2012), it follows that good models for the ERM purpose should have few EVs and relatively few DVs. Preference for models with few EVs follows from the fact that few, often only two or three, major complex-gradients, i.e. sets of more or less strongly correlated environmental variables (Whittaker et al. 2005), are usually sufficient to describe and understand variation in species' responses to the environment. Preference for models with few DVs follows from response curves with respect to these major complex-gradients typically being relatively simple, unimodal or truncated unimodal (Oksanen and Minchin 2002, Austin 2005, Halvorsen 2012). Accordingly, Halvorsen (2012) suggested in the context of distribution modelling with the ERM purpose that the concept of overfitting should be broadened also to include cases in which a model includes many variables which do not improve the model's predictive performance substantially (Type II

overfitting), as exemplified by more complex models with similar or almost similar AUC values than simpler models. Furthermore, Halvorsen (2012) also argued from a theoretical point of view that also models that do not suffer from Type II overfitting may be too complex to comply with standards for good ERM models (Type III overfitting; Halvorsen 2012).

Our results show that modelling by the SMP procedure, which implies construction of several derived variables (DVs) of different types from each EV and model selection by  $\ell_1$ -regularisation, tends to result in complex models with many DVs unless much stricter regularisation than the default regularisation parameter  $\lambda = 1$  is applied (Table 3). The most complex SMP models, with suboptimal test AUC values, clearly satisfy the criteria given above for Type I overfitting and are suboptimal for SPM as well as for ERM purposes. We do, however, argue that even the best SPM models suffer from overfitting type II because their test AUC values are almost similar to that of simpler models, and that they also seem to suffer from Type III overfitting because the response function is unrealistically complex and because the complexity of the model prevents interpretation of ecological relationships of the target (see below). From an SPM perspective, as many as 13 out of the 30 models had test AUC  $> 0.86$ , and many of these models are simple in terms of DV numbers and therefore accord with type I overfitting avoidance.

These results provide another case of failure of the  $\ell_1$ -regularization procedure, which is implemented as standard in MaxEnt, to prevent overfitting. Our results thus accord with results of some other studies (Raes and ter Steege 2007, Merckx et al. 2011, Halvorsen 2013), adding to a growing pile of evidence which indicates that the frequent claims in the literature that the regularization procedure is a major reason for good performance of MaxEnt in practical distribution modelling (e.g. Elith et al. 2006, 2011, Phillips et al. 2006, Parolo et al. 2008, Phillips & Dudík 2008) does not hold true.

As an alternative to the potentially overfit SMP models, we created AMP models by forward stepwise selection of EVs and DVs to be included in the models. The AMP models obtained by use of the  $F$ -ratio test as model comparison method, with different significance levels  $\alpha$  as performance assessment criterion, included fewer EVs and DVs than did the SMP models. The best AMP model, obtained for 121 presence observations with  $\alpha = 1 \cdot 10^{-5}$ , included as few as six EVs and nine DVs and had a test AUC value of 0.8627, which is comparable with the test AUC value of 0.8684 for the best SMP model obtained for 121 presence observations with  $\lambda = 4$ . The latter model included nine EVs and 25 DVs. The two models are more or less indistinguishable from the point of view of SPM modelling: while

the best SMP model had slightly higher test AUC, it is also clearly more complex than the best AMP model and the difference in test AUC was so small that it may be a result of random variation in the evaluation data set (cf. Fig. 2). Applying Ockham's razor, the best AMP model should be preferred also from the perspective of SPM purposes. From a strict ERM perspective, the best AMP model in terms of test AUC is also the simplest and certainly the best among the 30 models obtained as end-products by the AMP and SMP procedures. However, judgment of the response curves in Fig. 4, together with consideration of the number of EVs and DVs in the models, indicates that even this model suffered from Type III overfitting. This shows that even the strictest model comparison criterion applied in this study,  $\alpha = 1 \cdot 10^{-11}$ , did not result in models as simple as desired from the ERM point of view.

The forward selection procedure leaves behind a sequence of models of increasing complexity, created in the process that leads to a final model. Examination of this sequence for the best AMP model shows that the steady increase in explained variation in the training data (reflected in the increase of training AUC) is not necessarily followed by increase of test AUC on independent evaluation data (Fig. 3). While the overall best AMP model (model No. 13; samples = 121,  $\alpha = 1 \cdot 10^{-5}$ , test AUC = 0.8627, EVs = 6, DVs = 9) had the highest test AUC value, test AUC of the equally complex 6-EV model without interactions was lower than that of the simpler 4-EV model (0.8519 vs. 0.8527). Large difference in test AUC was found also between the models which contained 1 and 2 EVs (i.e., when TPI05 is added to Slope), and between models which contained 3 and 4 EVs (i.e., when Plan curvature and VRM05 are added to Slope and TPI05). From the AUC alone it is difficult to tell which of the 2-EV model with test AUC = 0.8385 or the 4-EV model with test AUC = 0.8527 that is the best model from the ERM perspective. However, comparison of response curves (Fig. 4) suggests that the 4-EV model should be preferred because the slight bimodality of the response to slope in the 2-EV model is corrected in the 4-EV model and a more realistic distinctly unimodal response to plan curvature is obtained.

### **Model evaluation**

We found a large, consistent difference between training and test AUC for all the 30 MaxEnt models, regardless of model selection procedure and strength of regularisation. While test AUC can be interpreted as the probability that a randomly chosen presence observation has higher RPPP value than a randomly chosen absence point in the evaluation set (Hanley and McNeil 1982), training AUC measures the probability that a randomly chosen presence

observation in the training data set has higher RPPP value than a randomly chosen uninformed background observation (Phillips et al. 2006). While the theoretical maximum value for test AUC is 1, training AUC has a maximum of  $1 - C/2$  where  $C$  is the prevalence of the modelled target in the study area (Wiley 2003, Phillips et al. 2006). The estimated prevalence of swamp forests in Østmarka Nature Reserve is about 3 % (7175 presence grid-cells out of a total of 233 774 grid cells), thus the theoretical maximum value for training AUC is about 0.985. Accordingly, the probability of randomly drawing a non-swamp forest grid cell with higher RPPP value than a swamp forest grid cell ( $1 - AUC$ ) is about 0.14 for the independent test data set (test AUC  $\approx$  0.86) and only 0.02–0.07 for the training data (training AUC = 0.915–0.965). In general, differences between training and test AUC values may reflect strong sampling bias in the training data, the most commonly cited reason for low performance of distribution models, a tendency of models to be inappropriately closely fit to the training data and/or a tendency for performance measures based on training data to be misleading (e.g., Austin 2007, Veloz 2009, Halvorsen 2012).

The training data used for our study differs from most training data sets used for distribution modelling by being more or less free of sampling bias: the presence-only data set was collected by a careful and systematic search for swamp forests throughout the entire study area, resulting in the registration of 121 swamp forests. The only sources of error expected to be present in the training data are the minor uncertainties regarding the precision of the hand held GPS used to draw borders for the swamp forests and, possibly, digitizing errors. Thus, the main reason for the observed difference between training and test AUC is likely to be a general tendency for models to be too closely fit to the training data (i.e. models predict the location of the swamp forests already included in the training data very well, but is not as good in predicting the locations of swamp forests in general, resulting in lower AUC when evaluated on independent data) and/or a tendency for performance measures based upon training data to be overly optimistic. We do not consider close fit to training data to be a result of a low number of presence observations in our study. One hundred and twenty one presence observations is a relatively high number compared to other distribution modelling studies, and higher than the recommended minimum number in several studies (Hernandez et al. 2006, Wisz et al. 2008, Mateo et al. 2010, Hanberry et al. 2012). Thus, the realised distribution of the modelled target in environmental variable space should be adequately represented by presence observations in the training data (Anderson and Gonzalez 2011). However, the occasional occurrence of presence observations with values for Slope and other

EVs that deviated strongly from the bulk of observations (Fig. 4) may have influenced the models strongly. An example of strong influence of presence observations with unusual values for environmental variables is shown by the 2-EV model, where Slope has an inappropriately shaped, bimodal response curve.

The strong increase in training AUC values with decreasing  $\lambda$  for SMP models, and the relationship between training AUC and model complexity, demonstrate that overfitting is likely to be an important reason for discrepancy between training and test AUC values. This accords with the conclusion of the previous section, that all the 30 SMP and AMP models are too complex to be appropriate for the ERM purpose, and indicates a general tendency for distribution models to be too closely fit to training data to reflect general relationships between modelled target and the environment. This is particularly problematic when modelling results are to be used for projection in space or time (PPM; projective distribution modelling; Halvorsen 2012).

Our result that inappropriately complex models, revealed by large differences between training and test AUC (high values for training AUC and low values for test AUC), are obtained by SMP with default settings for the regularisation multiplier ( $\lambda = 1$ ) in the MaxEnt software, give empirical support to the conclusions of reviews of current MaxEnt practice by Yackulic (in press) and Halvorsen (2013), that using default regularisation settings is likely to result in models with suboptimal predictive performance. Furthermore, our results indicate that the tendency of MaxEnt models to be overfit to the training data will fail to be detected in most cases, because independently collected data are rarely used for evaluation. Our study thus provides strong empirical support for the frequently expressed view, but rare practice (e.g., Newbold et al. 2010, Edvardsen 2011) that distribution models should be evaluated by use of independent evaluation data (Elith et al. 2006, Austin 2007, Veloz 2009, Newbold et al. 2010, Edvardsen et al. 2011, Halvorsen 2013).

### **Effects of number of presence observations and spatial structure in training data on model performance and model complexity**

MaxEnt models built by use of 7175 presence observations with the default SMP procedure were consistently more complex than models built by use of 121 presence observations. This tendency for models to be increasingly overfit with increasing number of presence observations is a logical outcome of the increasingly fine detail in which the relationship between observed presence and background can be described when the number of presence

observations increase. Models obtained by the AMP procedure do not show this tendency to be increasingly overfit when the number of presence observations increases, most likely because the AMP procedure implies a strong control over model complexity at each step in the forward selection procedure.

We did not find any consistent difference in performance on independent evaluation data (test AUC) between models trained on 7175 presence observations (all grid cells in which swamp forest was recorded as present) and models trained on centroids of the 121 swamp forests, all other settings and options equal and the same environmental variables used. However, the overall best SMP and AMP models were both obtained by use of 121 presence observations. The 121- and 7175-presence observations response data sets differ by the latter having strong fine-scaled spatial structure. Spatial structure in ecological data is common (Legendre 1993), and data used for distribution modelling are, in general, no exception to this (Chapman 2010, Dormann 2011); both the response and environmental variables are likely to be spatially structured (Segurado et al. 2006). This is also the case for the environmental explanatory variables used in our study, which mostly reflect terrain form in a landscape with main topographic features much larger than the grain size of 5 m used in our study (Økland et al. 2001). At a first glance, our results may be interpreted in support of the predictive performance of distribution models being unaffected by spatial autocorrelation in the response variable. This is, however, not necessarily the case. In the context of statistical testing (e.g., by the  $F$ -ratio test applied in the AMP procedure), it is not spatial structure in the raw response variable that matters, but eventual spatial autocorrelation in the residuals (Kühn and Dormann 2012). The distinction between spatial dependence and spatial autocorrelation as two components of spatial structure (Borrough 1987) is important; the former referring to spatial structure resulting from dependence of the response variable on spatially structured variables (e.g., dependence on swamp forest development on spatially structured terrain forms), the latter referring to spatially structure in the processes inherent in shaping the distribution of response variable, i.e. swamp forest encroachment (Segerström et al. 1994, Ohlson et al. 2006, Weckström et al. 2010). The spatial structure in swamp forest presence patterns in the 7175-observations data set is likely to be due to spatial dependence on topographic factors, which is fully accounted for by all AMP and SMP models. This indication that none of our models are likely to be burdened with spatial autocorrelation of residuals thus supports other findings (Kühn and Dormann 2012). This also explains why model performance may be unaffected by spatial structure in the raw response data as such,

and also explains why spatial structure (often referred to as 'spatial autocorrelation') in the response data have been found to lead either to inflation of AUC (Segurado et al. 2006, Veloz 2009) or to decrease in AUC (Merckx et al. 2011). Our results support the view of (Dormann 2011) that no specific measures need to be taken if presence of spatial autocorrelation does not affect modelling results, but also highlight the importance of ensuring that no such effect is present. Furthermore, our results show that increasing the number of presence observations does not necessarily improve the predictive performance of distribution models.

### **Understanding swamp forest distributions: contributions from the distribution models**

To understand how the distribution of the modelled target is shaped by important environmental factors conforms with the ERM purpose of distribution modelling. Ecological response modelling calls for the simplest model with reasonably high predictive performance, in our case found to be the 4-EV AMP model which includes Slope, TPI05, Plan curvature and VRM05. Among these variables, Slope and TPI05 were selected first by the AMP procedure. Slope is clearly the best single predictor of swamp forest distribution, as demonstrated by the test AUC of 0.79 for the model with slope as the only EV. Nevertheless, the models' predictive performance increased considerably, to test AUC > 0.85, when more variables were included. The obvious reasons for this is that inclusion of the other variables improved the modelled response to slope, partly by improving the separation of swamp forest from non-swamp forest grid cells in nearly flat terrain (shown by the wider vertical amplitude of the grey cloud of background points for low slopes for the 4-EV than for the 2-EV model; almost all swamp forests are situated on slopes between 0 and 3°), and partly by improving predictions for steeper slopes for which presence observations are few (note the change from a slightly bimodal to a monotonously decreasing response curve in Fig. 4; which better reflects the fact that steep slopes almost never support swamp forests; Økland et al. 2001). The contribution of the topographic position index (TPI) to model performance reflects the fact that swamp forests are located in valleys and not on ridges. The contribution of the vector ruggedness measure (VRM) reflects the more smooth surface of swamp forests than of the surrounding mineral soil, a result of the in-filling of landscape depressions during peat accumulation and swamp forest development (Ohlson et al. 2006). The bimodal response curve to VRM shown by all models (Fig. 4) is therefore likely to result from a few presence observations with high ruggedness: a monotonously decreasing curve, resembling the curve for slope, should be expected. Plan curvature distinguishes between convex and concave

slopes. The slight skew to the left in the response to this variable may reflect occasional occurrences of swamp forests on long slopes in shallow but distinctly concave-shaped depressions.

The EVs identified by our AMP models to be good predictor variables of local swamp forest distribution accord with those pointed out as important for swamp forest development in previous studies (see Ohlson et al. 2006, Hörnberg et al. 2012, and references therein). This adds to previous studies showing that distribution modelling can be a valuable tool in conservation and management of swamp forests and other nature-types that are biodiversity hotspots (Riordan and Rundel 2009, Weber 2011). However, occurrence of other nature-types under conditions that cannot be distinguished from those identified by the best MaxEnt ERM model as suitable for swamp forests by the environmental variables included in the model, will reduce the applicability of the model. For swamp forests in SE Norway this is the case for open mires, which are found adjacent to several lakes in the study area as well as in the central parts of large peatland complexes with swamp forests along their margins, hence sharing the 'topographic signal' of swamp forests. Presence of open mires in the study area therefore certainly adds to the observed commission error, i.e., that swamp forest presence is predicted for cells that support open mires. The possibility to separate swamp forests from open mires by use of vegetation structural variables, e.g., as obtained from ALS measurements, should be explored further (Lefsky et al. 2002, Lim et al. 2003, Korpela et al. 2009, Vehmas et al. 2009, Bässler et al. 2011).

### **Conclusion: recommendations for distribution modelling by MaxEnt and future studies**

Our results demonstrate a strong tendency for the default standard MaxEnt procedure (use of  $\ell_1$ -regularisation with a value of 1 for the regularisation multiplier) to result in overly complex models with lower than optimal predictive performance. We therefore strongly encourage comparisons of the stepwise forward selection procedure used in this study with other model selection procedures (e.g., see Reineking and Schröder 2006, Halvorsen 2013), using a large specter of data sets. Meanwhile, and regardless of modelling purpose and model selection procedure, we recommend using a strict model performance assessment criterion and that user's take control over the important process by which explanatory variables are transformed to derived variables. Finally, our results clearly show that truly independent presence-absence data are mandatory for reliable evaluation of distribution models.

## **Acknowledgements**

We thank Vegar Bakkestuen\*, Lars Erikstad<sup>^</sup>, Vegard Lien<sup>#</sup> and Hans Ole Ørka<sup>#</sup> for assisting us in ArcGIS, R and ALS. Anders Bryn\*<sup>~</sup>, Lars Østbye Hemsing<sup>~</sup> with MaxEnt-related questions. Ole Martin Bollandsås<sup>#</sup> and Christian Bianchi Strømme<sup>#</sup> for help with the last part of the field work.

\* The Norwegian Natural History Museum, <sup>^</sup> Norwegian institute for water science,

<sup>#</sup> Norwegian University of Life Sciences, <sup>~</sup> Norwegian institute for forests and landscape.

## References

- Anderson, R. P. and Gonzalez, I. J. 2011. Species-specific tuning increases robustness to sampling bias in models of species distributions: an implementation with Maxent. — *Ecological Modelling* 222: 2796-2811.
- Anon. 2005. TerraScan user's guide. Terrasolid Ltd.
- Anonymous 2008. ArcGIS, ver. 9.3. — Inc. Redlands, California
- Aune, B. 1993. Temperaturnormaler : normalperiode 1961-1990. — DNMI.
- Austin, M. 2007. Species distribution models and ecological theory: A critical assessment and some possible new approaches. — *Ecological Modelling* 200: 1-19.
- Austin, M. P. 2005. Vegetation and environment: discontinuities and continuities. — In: Maarel, v. d. (ed), *vegetation ecology*. Blackwell pp. 52-84.
- Axelsson, P. 2000. DEM generation from laser scanner data using adaptive TIN models. — *Int. Arch. Photogramm. Remote Sens.* 33(B4): 111-118.
- Baltsavias, E. P. 1999. Airborne laser scanning: basic relations and formulas. — *Isprs Journal of Photogrammetry and Remote Sensing* 54: 199-214.
- Borrough, P. 1987. Spatial aspects of ecological data. — In: Jongman, R. H. G. et al. (eds), *Data Analysis in Community and Landscape Ecology*. Pudoc, Wageningen, The Netherlands, pp. 213-251.
- Bässler, C. et al. 2011. LiDAR as a rapid tool to predict forest habitat types in Natura 2000 networks. — *Biodiversity and Conservation* 20: 465-481.
- Chapman, D. S. 2010. Weak climatic associations among British plant distributions. — *Global Ecology and Biogeography*. 19: 831-841.
- Dormann, C. 2011. Modelling species' distributions. — In: Jopp, F. et al. (eds), *Modelling complex ecological dynamics: an introduction into ecological modelling*. Springer, pp. 179-196.
- Dormann, C. F. et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. — *Ecography* 30: 609-628.
- Edwardsen, A. et al. 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. et al. 2005. Model-based stratifications for enhancing the detection of rare ecological events. — *Ecology* 86: 1081-1090.
- Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. — *Ecography* 29: 129-151.
- Elith, J. and Leathwick, J. R. 2009. *Species Distribution Models: Ecological Explanation and Prediction Across Space and Time*. *Annual Review of Ecology Evolution and Systematics*. Annual Reviews, pp. 677-697.
- Elith, J. and Graham, C. H. 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. — *Ecography* 32: 66-77.
- Elith, J. et al. 2011. A statistical explanation of MaxEnt for ecologists. — *Diversity and Distribution*. 17: 43-57.
- Franklin, J. 2009. *Mapping species distributions: spatial inference and prediction*. — Cambridge University Press.
- Franklin, J. 2010. Moving beyond static species distribution models in support of conservation biogeography. — *Diversity and Distribution*. 16: 321-330.
- Førland, E. J. 1993. Nedbørnormaler: normalperiode 1961-1990. — Instituttet.
- Gogol-Prokurat, M. 2011. Predicting habitat suitability for rare plants at local spatial scales using a species distribution model. — *Ecological Applications* 21: 33-47.
- Gorham, A. J. P. 1983. *Ecosystems of the World 4A. Mires: swamp, bog, fen and moor*. — Elsevier.
- Graham, C. H. et al. 2004. New developments in museum-based informatics and applications in biodiversity analysis. — *Trends in Ecology & Evolution* 19: 497-503.

- Guisan, A. and Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. — *Ecological Modelling* 135: 147-186.
- Guisan, A. and Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat models. — *Ecological Letters*. 8: 993-1009.
- Guisan, A. et al. 2006. Using niche-based models to improve the sampling of rare species. — *Conservation Biology* 20: 501-511.
- Halvorsen, R. 2012. A gradient analytic perspective on distribution modelling. — *Sommerfeltia* 35: 1-165.
- Halvorsen, R. 2013. A strict maximum likelihood explanation of MaxEnt, and some implications for distribution modelling. — *Sommerfeltia* 36: 1-132.
- Hanberry, B. B. et al. 2012. Sample sizes and model comparison metrics for species distribution models. — *Ecological Modelling* 227: 29-33.
- Hanley, J. A. and McNeil, B. J. 1982. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. — *Radiology* 143: 29-36.
- Hastie, T. J. et al. 2009. *The elements of statistical learning: data mining, inference, and prediction.* — Springer Science + Business Media.
- Hernandez, P. A. et al. 2006. The effect of sample size and species characteristics on performance of different species distribution modelling methods. — *Ecography* 29: 773-785.
- Hijmans, R. J. and Elith, J. 2011. *Species distribution modelling with R.* — The R foundation for statistical computing.
- Hörnberg, G. et al. 1998. Boreal swamp forests. — *Bioscience* 48: 795-802.
- Hörnberg, G. et al. 2012. Fire as an important factor for the genesis of boreal *Picea abies* swamp forests in Fennoscandia. — *Holocene* 22: 203-214.
- Jaynes, E. T. 1957a. Information theory and statistical mechanics. — *Physical Review* 106: 620-630.
- Jaynes, E. T. 1957b. Information theory and statistical mechanics II. — *Physical Review* 108: 171-190.
- Jaynes, E. T. 1982. On the rationales of maximum-entropy methods. — *Proceedings of the Institute of Electrical and Electronics Engineers* 70: 939-952.
- Jiménez-Valverde, A. et al. 2008. Not as good as they seem: the importance of concepts in species distribution modelling. — *Diversity and Distribution*. 14: 885-890.
- Korpela, I. et al. 2009. Airborne small-footprint discrete-return LiDAR data in the assessment of boreal mire surface patterns, vegetation, and habitats. — *Forest Ecology and Management* 258: 1549-1566.
- Kraus, K. and Pfeifer, N. 1998. Determination of terrain models in wooded areas with airborne laser scanner data. — *ISPRS Journal of Photogrammetry and Remote Sensing* 53: 193-203.
- Kühn, I. and Dormann, C. F. 2012. Less than eight (and a half) misconceptions of spatial analysis. — *Journal of Biogeography*. 39: 995-998.
- Lefsky, M. A. et al. 2002. Lidar remote sensing for ecosystem studies. — *Bioscience* 52: 19-30.
- Legendre, P. 1993. Spatial autocorrelation - trouble or new paradigm. — *Ecology* 74: 1659-1673.
- Lim, K. et al. 2003. LiDAR remote sensing of forest structure. — *Progress in Physical Geography* 27: 88-106.
- Lobo, J. M. et al. 2008. AUC: a misleading measure of the performance of predictive distribution models. — *Global Ecology and Biogeography* 17: 145-151.
- Mateo, R. G. et al. 2010. Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absences and target-group absences from natural history collections. — *Diversity and Distribution*. 16: 84-94.
- Merckx, B. et al. 2011. Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. — *Ecological Modelling* 222: 588-597.
- Moen, A. 1999. *National Atlas of Norway: Vegetation.* — Norwegian Mapping Authority.
- Newbold, T. et al. 2010. Testing the accuracy of species distribution models using species records from a new field survey. — *Oikos* 119: 1326-1334.

- Ohlson, M. et al. 1997. Habitat qualities versus long-term continuity as determinants of biodiversity in boreal old-growth swamp forests. — *Biological Conservation* 81: 221-231.
- Ohlson, M. et al. 2006. The macroscopic charcoal record in forested boreal peatlands in southeast Norway. — *Holocene* 16: 731-741.
- Oksanen, J. and Minchin, P. R. 2002. Continuum theory revisited: what shape are species responses along ecological gradients? — *Ecological Modelling* 157: 119-129.
- Parolo, G. et al. 2008. Toward improved species niche modelling: *Arnica montana* in the Alps as a case study. — *Journal of Applied Ecology* 45: 1410-1418.
- Pearce, J. and Ferrier, S. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. — *Ecological Modelling* 133: 225-245.
- Pearson, R. G. et al. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. — *Journal of Biogeography* 34: 102-117.
- Peterson, A. T. et al. 2011. *Ecological niches and geographic distributions*. — Princeton University Press.
- Phillips, S. 2012. Inferring prevalence from presence-only data: a response to 'Can we model the probability of presence of species without absence data?'. — *Ecography* 35: 385-387.
- Phillips, S. J. et al. 2004. A maximum entropy approach to species distribution modelling. — In: Anonymous (ed), *Proceedings of the 21st international conference on machine learning*. ACM Press, pp. 655-662.
- Phillips, S. J. et al. 2006. Maximum entropy modelling of species geographic distributions. — *Ecological Modelling* 190: 231-259.
- Phillips, S. J. and Dudík, M. 2008. Modelling of species distributions with Maxent: new extensions and a comprehensive evaluation. — *Ecography* 31: 161-175.
- Phillips, S. J. 2011. A brief tutorial on Maxent. — AT&T Research, Princeton, NJ
- Raes, N. and ter Steege, H. 2007. A null-model for significance testing of presence-only species distribution models. — *Ecography* 30: 727-736.
- Rebelo, H. and Jones, G. 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.
- Reineking, B. and Schröder, B. 2006. Constrain to perform: regularization of habitat models. — *Ecological Modelling* 193: 675-690.
- Reutebuch, S. E. et al. 2003. Accuracy of a high-resolution lidar terrain model under a conifer forest canopy. — *Canadian Journal of Remote Sensing* 29: 527-535.
- Riordan, E. C. and Rundel, P. W. 2009. Modelling the distribution of a threatened habitat: the California sage scrub. — *Journal of Biogeography* 36: 2176-2188.
- Santika, T. and Hutchinson, M. F. 2009. The effect of species response form on species distribution model prediction and inference. — *Ecological Modelling* 220: 2365-2379.
- Seegerström, U. et al. 1994. Disturbance history of a swamp forest refuge in northern Sweden. — *Biological Conservation* 68: 189-196.
- Segurado, P. et al. 2006. Consequences of spatial autocorrelation for niche-based models. — *Journal of Applied Ecology* 43: 433-444.
- Sigmond, E. M. O. et al. 1984. *Berggrunnskart over Norge 1: 1 000 000*. — Norg. geol. Unders.
- Sjörs, H. 1948a. *Myrvegetation i Bergslagen*. — Almqvist & Wiksell International.
- Sjörs, H. 1948b. *Myrvegetation i Bergslagen*. — *Acta Phytogeographica Suecica* 21: 1-299.
- Stokland, J. N. et al. 2011. Species distribution modelling-Effect of design and sample size of pseudo-absence observations. — *Ecological Modelling* 222: 1800-1809.
- Telenius, A. 2011. Biodiversity information goes public: GBIF at your service. — *Nordic Journal of Botany* 29: 378-381.

- Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. — *Journal of the Royal Statistical Society* 58: 267-288.
- Vehmas, M. et al. 2009. Identification of boreal forest stands with high herbaceous plant diversity using airborne laser scanning. — *Forest Ecology and Management* 257: 46-53.
- Veloz, S. D. 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. — *Journal of Biogeography* 36: 2290-2299.
- Venables, W. N. and Ripley, B. D. 2002. *Modern applied statistics with S*. — Springer.
- Ward, G. et al. 2009. Presence-only data and the EM algorithm. — *Biometrics* 65: 554-563.
- Warren, D. L. and Seifert, S. N. 2011. Ecological niche modelling in Maxent: the importance of model complexity and the performance of model selection criteria. — *Ecological Applications* 21: 335-342.
- Weber, T. C. 2011. Maximum entropy modelling of mature hardwood forest distribution in four U.S. states. — *Forest Ecology and Management* 261: 779-788.
- Weckström, J. et al. 2010. Climatic influence on peatland formation and lateral expansion in sub-arctic Fennoscandia. — *Boreas* 39: 761-769.
- Whittaker, R. J. et al. 2005. Conservation Biogeography: assessment and prospect. — *Diversity and Distribution*. 11: 3-23.
- Wiley, E. O., McNyset, K.M., Peterson, A.T., Robins, C.R. & Stewart, A.M. 2003. Niche modelling and geographic range predictions in the marine environment using a machine-learning algorithm. — *Oceanography* 16: 120-127.
- Williams, J. N. et al. 2009. Using species distribution models to predict new occurrences for rare plants. — *Diversity and Distribution*. 15: 565-576.
- Wisz, M. S. et al. 2008. Effects of sample size on the performance of species distribution models. — *Diversity and Distribution*. 14: 763-773.
- Wollan, A. K. et al. 2008. Modelling and predicting fungal distribution patterns using herbarium data. — *Journal of Biogeography* 35: 2298-2310.
- Yackulic, C. B., Chandler, R., Zipkin, E.F., Royle, J.A., Nichols, J.D., Grant, E.H.C., Veran, S. in press. Presence-only modelling using MAXENT: when can we trust the inferences. — *Methods in Ecology and Evolution*. 4: 236-243.
- Økland, R. H. 1989. Hydromorphology and phytogeography of mires in inner Østfold and adjacent part of Akershus, SE Norway, in relation to regional variation in SE Fennoscandian mires. — *Statens naturvetenskapliga forskningsråd*.
- Økland, R. H. et al. 2001. Vegetation-environment relationships of boreal spruce swamp forests in Østmarka Nature Reserve, SE Norway. *Sommerfeltia*, 29. p 190.

Appendix A. Outline of a manual procedure for forward stepwise selection of explanatory variables (EVs) and derived variables (DVs) in MaxEnt. ISDV = individually significant DV (Modified from Halvorsen 2013: Table 4)

Step	Term	Description
<b>1</b>	<b>Initial steps: preparatory phase</b>	
1a	Specification of method for internal model performance assessment	Select a method for comparing two nested MaxEnt models (the <i>F</i> -ratio test was used for this study), including a threshold ( <i>model improvement criterion</i> ) to be used to decide if a more complex Maxent model is better than a simpler model (nine different significance levels $\alpha$ were used as thresholds in this study).
1b	Construction of continuous DVs	Transform each continuous EV into continuous DVs, i.e., DVs of the L, M and D types (appendix 2).
1c	Construction of spline DVs	Transform each EV into variables of the spline type (HF, HR, and T). Because (almost) infinitely many DVs of each type can be constructed by shifting the position of the knot, a method for limiting the number of DVs is required (operationalisation of this step is explained in the text).
<b>2</b>	<b>Analysis of single EVs and DVs phase</b>	
2a	Single-variable modelling step	For each EV and DV resulting after Step 1c, make a 'plain' single-variable MaxEnt model (i.e., a model obtained without $\ell_1$ -regularisation, using the specifications in Step 1a).
2b	Single-variable test step	Compare each single-variable MaxEnt model with the MaxEnt null model by use of the method and the criterion specified in Step 1a.
2c	Single-variable selection step	Select all ISDVs, i.e., EVs and DVs that satisfy the criterion specified in Step 1a, for use in Step 3, and leave out all other EVs and DVs for which no ISDVs could be obtained.
<b>3</b>	<b>Selection of parsimonious sets of DVs for each EV</b>	
3a	Finding the best DV in each set	For each set of ISDVs derived from the same EV, select the ISDVs that performs best in the single-variable test of Step 2b. For sets with only one ISDV, this ISDV makes up the parsimonious set of DVs for this EV. For all other EVs, proceed to step 3b.
3b	Model improvement test step	For each additional ISDV in each set, make a two-variable MaxEnt model by adding this ISDV to the best ISDV in the set. Use the criterion specified in Step 1a to compare each two-variable model with the one-variable model for the best ISDV in the set.
3c	Set expansion step	Consider the following three cases: (i) If no additional ISDV is found in Step 3b that satisfies the criterion specified in Step 1a, the parsimonious set of DVs consists of the best ISDV only. (ii) If one ISDV satisfies the criterion, the parsimonious set consists of this ISDV and the best ISDV. (iii) If more than one DV satisfies the criterion, select the one which performs best in the two-variable test of Step 3b and repeat Step 3b by comparing three-variable models with the best two-variable model. Repeat the process until no more ISDVs can be added to the set.
<b>4</b>	<b>Finding the best MaxEnt model without interactions</b>	
4a	EV test step	Compare MaxEnt models for each EV represented by a parsimonious set of ISDVs with the MaxEnt null model. Select the best EV according to the criterion specified in Step 1a.
4b	Model improvement test step	Similar to Step 3b, but applied to EVs represented by parsimonious sets of DVs instead of single DVs
4c	Model expansion step	Similar to Step 3c, but applied to EVs represented by parsimonious sets of DVs instead of single DVs.
4d	Termination step	The best MaxEnt model without interactions is found when neither more EVs represented by parsimonious sets of DVs can be found that improve the model, as judged by the criterion specified in Step 1a.
<b>5</b>	<b>Finding the best MaxEnt model with Interactions</b>	
5a	Construction of	Specify the number <i>s</i> of continuous EVs for which interaction DVs are to be

	interaction DVs	created. This will be the first $s$ EVs selected in step 4a. Interaction DVs are obtained as the products of all combinations of DVs in all $s(s-1)/2$ pairs of EVs.
5b	Interaction DV test step	For each interaction DV obtained in Step 5a, make a MaxEnt model by adding this DV to the DVs in the best MaxEnt model without interactions, obtained in Step 4d.
5c	Model improvement test step	Similar to Step 3b, but applied to interaction DVs
5d	Model expansion step	Similar to Step 3c, but applied to interaction DVs
5e	Termination step	The best MaxEnt model with interactions is found when neither more interaction DVs can be found that improve the model, as judged by the criterion specified in Step 1a.

Appendix B. Transformation of continuous, explanatory variables (EVs;  $Z_j$ , which are ranged onto a [0,1] scale) into derived variables (DV) of different types, grouped into three main types: C = continuous; S = spline and I = interaction). The transformation procedure occurs in two steps, of which formulae for the first step, transformation into 'raw' derived variables (rDVs)  $X_k$ , is shown in the rightmost table column. In a second step, the DVs used as input to MaxEnt modelling,  $X_k$ , are obtained by linear ranging of rDVs onto a [0,1] scale. AMP = alternative MaxEnt model; SMP = standard MaxEnt model.  $n$  = number of observed presences.  $s$  = number of EVs. The maximum number of DVs that might be derived from each EV was 16.

DV types	DVT Code	Description and specification Term	Number of DVs			Transformation function for DVs
			AMP	SMP	DVs	
C	L	linear	1	1	1	$x_{ki}' = z_{ji}$
C	M	monotonous	1	1	1	AMP: $x_{ki}' = \begin{cases} e^{z_{ji}} & \text{if skew}(Z_j) < 0 \\ \ln(c + z_{ji}) & \text{if skew}(Z_j) > 0 \end{cases}$
C	D	deviation	2	0	0	SMP: $x_{ki}' = (z_{ji})^2$ $x_{ki}' = (z_{ji} - \bar{z}_j)^a$
S	HF	forward hinge	0-4	un-limited	un-limited	$x_{ki}' = \begin{cases} 0 & \text{if } z_{ji} < z_{j0} \\ \frac{z_{ji} - z_{j0}}{\max(z_{j1}) - z_{j0}} & \text{if } z_{ji} \geq z_{j0} \end{cases}$
S	HR	reverse hinge	0-4	un-limited	un-limited	$x_{ki}' = \begin{cases} \frac{z_{j0} - z_{ji}}{z_{j0} - \min(z_{j1})} & \text{if } z_{ji} \leq z_{j0} \\ 0 & \text{if } z_{ji} > z_{j0} \end{cases}$
S	T	threshold	0-4	$n-1$	$n-1$	$x_{ki}' = \begin{cases} 1 & \text{if } z_{ji} \geq z_{j0} \\ 0 & \text{if } z_{ji} < z_{j0} \end{cases}$
I	P	product	0	$s(s-1)$	$s(s-1)$	$x_{ki}' = z_{ji} \cdot z_{vi}$



# Paper 2



**Structural differences between boreal swamp forests and non-paludified forest in a forest reserve measured by airborne laser scanning**

**Authors**

John Wirkola Dirksen<sup>1\*</sup>, Rune Halvorsen<sup>2</sup>, Hans Ole Ørka<sup>1</sup>, Katrine Eldegard<sup>1</sup>, Erik Næsset<sup>1</sup>, Terje Gobakken<sup>1</sup> and Mikael Ohlson<sup>1</sup>.

**Addresses**

<sup>1)</sup> Norwegian University of Life Sciences, Department of Ecology and Natural Resource Management, PO Box 5003, NO-1432 Ås, Norway

<sup>2)</sup> University of Oslo, Natural History Museum, Department of Botany, PO Box 1172 Blindern,  
NO-0318 Oslo, Norway

**\*<sup>)</sup> Corresponding author**

John Wirkola Dirksen

E.mail [john.dirksen@umb.no](mailto:john.dirksen@umb.no)

Phone: +47 64 96 57 31

## **Abstract**

Swamp forests are considered to be biodiversity hotspots in the Scandinavian boreal forests, yet little is known about structural differences between swamp forests and non-paludified forest. We used airborne laser scanning (ALS) to measure if forest structure differed between swamp forests and non-paludified forest in an old-growth forest reserve in south eastern Norway. Twenty-four structure variables were analyzed on three spatial scales: 25m<sup>2</sup>, 225m<sup>2</sup> and 625m<sup>2</sup>. Mean height, standard deviation of height and all percentiles differed significantly between swamp forests and non-paludified forest on all three spatial scales. On average the vegetation in swamp forests was 3.1m lower than in non-paludified forest. A logistic regression model was developed for each of the three spatial scales. These models were used to predict the location of swamp forests on two independent evaluation data sets. The proportion of correctly predicted presence and absence cells for the two evaluation data sets was 54–62 %. The individual variables which were included in the models differed between the three different scales. Although the predictive power of the prediction models was only moderate, they may still found a basis for development of prediction maps, which may be improved through field validation and subsequent improvement of models.

## Introduction

Swamp forests cover only a small fraction of the boreal forest landscape in Scandinavia and are known to be biodiversity hotspots (Ohlson et al. 1997, Hörnberg et al. 1998). At least three important environmental factors are known to contribute to the high species richness found in swamp forests: 1) a gradient in plant nutrient availability (Økland et al. 2001), 2) a gradient in water saturation related to surface microtopography (Økland et al. 2008) and 3) the amount of dead wood (Ohlson et al. 1997). Furthermore, Økland et al. (2001) found that the species composition of swamp forests shows patterns related to tree spacing, i.e., that the vegetation under trees differs from the vegetation between trees. This accords with the general view that forest structure is of vital importance for forest-dwelling species (Brokaw and Lent 1994, Spies 1998) and opens for the possibility that the tree-layer structure in swamp forests may act as a fourth factor that may explain the high number of species typically found in swamp forests. Even though the species composition and biological diversity of Scandinavian swamp-forests have been addressed in several studies over the last decades (Hornberg et al. 1997, Ohlson et al. 1997, Hörnberg et al. 1998, Økland et al. 2001, 2008), the extent to which the forest structure of swamp forests actually differs from that of non-paludified boreal forests is poorly known. However, Hörnberg et al. (1995) found large variation among ten boreal spruce-dominated swamp forests with respect to tree densities and tree-size distributions, indicating that swamp forests may lack specific, distinguishing structural properties. High potential importance and lack of knowledge motivates strong needs for an in-depth comparative study of forest structure between swamp forests and non-paludified forests.

One reason why so little is known about structure of swamp forests is that, until recently, measuring forest structure in a proper way has been both challenging and time consuming. Structural characteristics like tree size and age, mean volume of living trees per hectare, number of standing dead trees per hectare, mean volume of dead wood on the ground, number of trees above a certain diameter, and number of stumps, are commonly used structural attributes of boreal forests (Hofgaard 1993, Kuuluvainen et al. 1998, Lie 2009). These measures do, however, reflect age and size distribution rather than the actual three-dimensional structure of the forest. Expansion of our knowledge about forest structural properties, including how forest types differ in these properties, requires methods for

assessing three-dimensional forest structural properties, including distribution of biomass in three dimensions. This information can now be obtained from airborne laser scanning (ALS) data.

Airborne laser scanning is a remote sensing technique, which provides measurements in three dimensions by measuring the time span from a laser pulse is emitted from a laser scanner mounted on an aircraft until its echo is received by the instrument after having been reflected from an object near the ground or the ground surface itself. The distance between the sensor and the object on which the laser pulse was reflected, is estimated from the time span recorded between emission and reflection of the laser pulse. Because the exact position and orientation of the sensor is provided by use of global satellite navigation systems and internal navigation systems, the spatial location of the echoes can be accurately determined. The result from an ALS acquisition is a three dimensional point cloud, each point being an echo with its own unique position in an x, y, and z coordinate system. Echoes returned from the ground are used to derive a terrain model, and a measure for vegetation height is derived by subtracting the terrain model from non-ground echoes. ALS has been used as a standard method in operational forest inventories in Norway since 2002 (Næsset et al. 2004). In such inventories important biophysical stand variables such as mean tree height, timber volume, and diameter distribution can easily be estimated from ALS data with high precision and high spatial resolution and yet with a large geographical coverage, often 500-1000 km<sup>2</sup> or more (Næsset 2002, Gobakken and Næsset 2004). Over the last decade, ALS has also increasingly been used for ecological applications because of the possibilities ALS offers for describing three-dimensional forest structures (Lim et al. 2003, Turner et al. 2003, Aplin 2005, Vierling et al. 2008) over large areas (Nelson et al. 2005). Structures of ecological importance estimated by ALS are e.g. successional stages (Falkowski et al. 2009), understory (Hill and Broughton 2009, Martinuzzi et al. 2009, Wing et al. 2012), canopy gaps (Koukoulas and Blackburn 2004), standing dead trees (Bater et al. 2009, Kim et al. 2009), tree species and canopy density (Turner et al. 2003). Although most ecological studies so far that have made use of ALS data have focused on habitat requirements of individual species (Hill et al. 2004, Müller et al. 2009, Coops et al. 2010, Smart et al. 2012) some recent studies have focused on identification of structures that can assist localization of specific nature types (Mücher et al. 2009, Vehmas et al. 2009, Bässler et al. 2011).

Low-growing and less dense vegetation than in the surrounding forest may, potentially, enable identification of swamp forests by use of ALS data. The reason for this is twofold. Firstly, the anoxic soil conditions in waterlogged peaty soil limits tree growth in swamp forests and results in small, stunted trees with fewer branches and less foliage than forests on non-paludified ground. Secondly, swamp forests tend to be characterized by microtopographic (hummock-hollow) variation with trees of most species, including the characteristic species of swamp forests, Norway spruce (*Picea abies* L. Karst), growing on hummocks. Since hummocks typically cover only a small proportion of the area of most swamp forests, the canopy of swamp forests is likely to be more open than that of surrounding forests. Furthermore, observations in the field (J.W. Dirksen, pers. obs.) suggest that deciduous shrubs are more common and cover larger areas in swamp forests than in surrounding non-paludified forests. If these general assumptions about forest structural differences between swamp forests and the surrounding non-paludified forests hold true, swamp forests may potentially be separated from non-paludified forests by ALS-derived metrics describing vegetation height, canopy density and shrub density. An important advantage of ALS data compared to other remote sensing data in this respect is that ALS data can be aggregated on different scales, thus the same ALS data can be used to derive structural patterns at different spatial scales.

This study has three main objectives: 1) to test the hypothesis that swamp forest differ from the surrounding non-paludified forest in structural properties that can be recorded in observation units of at least one of the scales: single grid-cells of 5×5 m (25 m<sup>2</sup>), 3×3 grid cells (15×15 m, 225 m<sup>2</sup>) and 5×5 grid cells (25×25 m, 625 m<sup>2</sup>); 2) to model structural differences between swamp forests and surrounding non-paludified forests; and 3) to evaluate the developed models on two independent evaluation data sets that contrast swamp forests, the first evaluation data set with absence cells located close to the swamp forests and the second evaluation data set with absence cells randomly selected within the study area.

## **Materials and methods**

### **Study area**

The study was carried out in the Østmarka Nature Reserve (59°50′ N, 11° 02′ E, 190–368 m.a.s.l.), which is located in the southern boreal zone of Southeast Norway (Moen 1999)

(Fig. 1). The study area comprised approximately 6 km<sup>2</sup>. Annual mean precipitation and temperature (1961–90 normals) at the nearest meteorological stations (Enebakk, ca. 10 km SE of the area, and Hakadal–Bliksrudhagan, ca. 30 km N of the area) were 816 mm and 3.9 °C, respectively (Aune 1993, Førland 1993). The bedrock is of proterozoic age and quartz diorite, tonalite and gneisses of diverse origins dominate (Sigmond et al. 1984). The topography of the study area is dominated by ridges and valleys in the north–south direction. On finer scales, the landscape is dissected by minor valleys, resulting in a broken topography with structures on several scales and a mixture of ridges, valleys, stream beds, large boulders, steep cliffs and more gentle slopes. Parts of the reserve have been protected since 1990 and additional areas were protected in 2002 so that the reserve now covers approximately 12.5 km<sup>2</sup>. Until around 1940 selective logging was performed in areas now located within the reserve. The reserve is dominated by old-growth Norway spruce forests. Scots pine (*Pinus sylvestris* L.) dominates on nutrient poor and shallow soils on ridges. Deciduous trees, mainly birch (*Betula pubescens* Ehrh. and *B. pendula* Roth.), occur scattered throughout the study area. The study area, which covers the center of the reserve, was conceptualized as a rasterized geographical space with resolution (grid-cell size) of 5 m, with a total of 233774 grid cells.

### **Training data**

During the summer of 2010 field work was conducted to locate swamp forests inside the study area. Swamp forests were defined as water-saturated areas dominated by trees, mostly spruce but also scattered birch, with *Sphagnum* mosses covering the ground. A total of 121 swamp forests with a minimum size of 100 m<sup>2</sup> were recorded (Fig. 1). The perimeter of each swamp forest was mapped to scale 1:1 000 in the field by walking the edges of each swamp forest and recording coordinates with a handheld Garmin eTrex H Global Positioning System (GPS, precision ±5 m) receiver. The swamp forest polygons delimited in the field were digitized using ArcMap 9.3 (ESRI 2011). A set of 2×121 observations was used as training data in the analyses: the centroid cells for each of the 121 swamp forest polygons constituted the presence observations (i.e. grid cells containing swamp forest), whereas 121 cells were randomly selected among the grid cells not registered as swamp forests, and these cells constituted absence observations (i.e. absence of swamp forest).

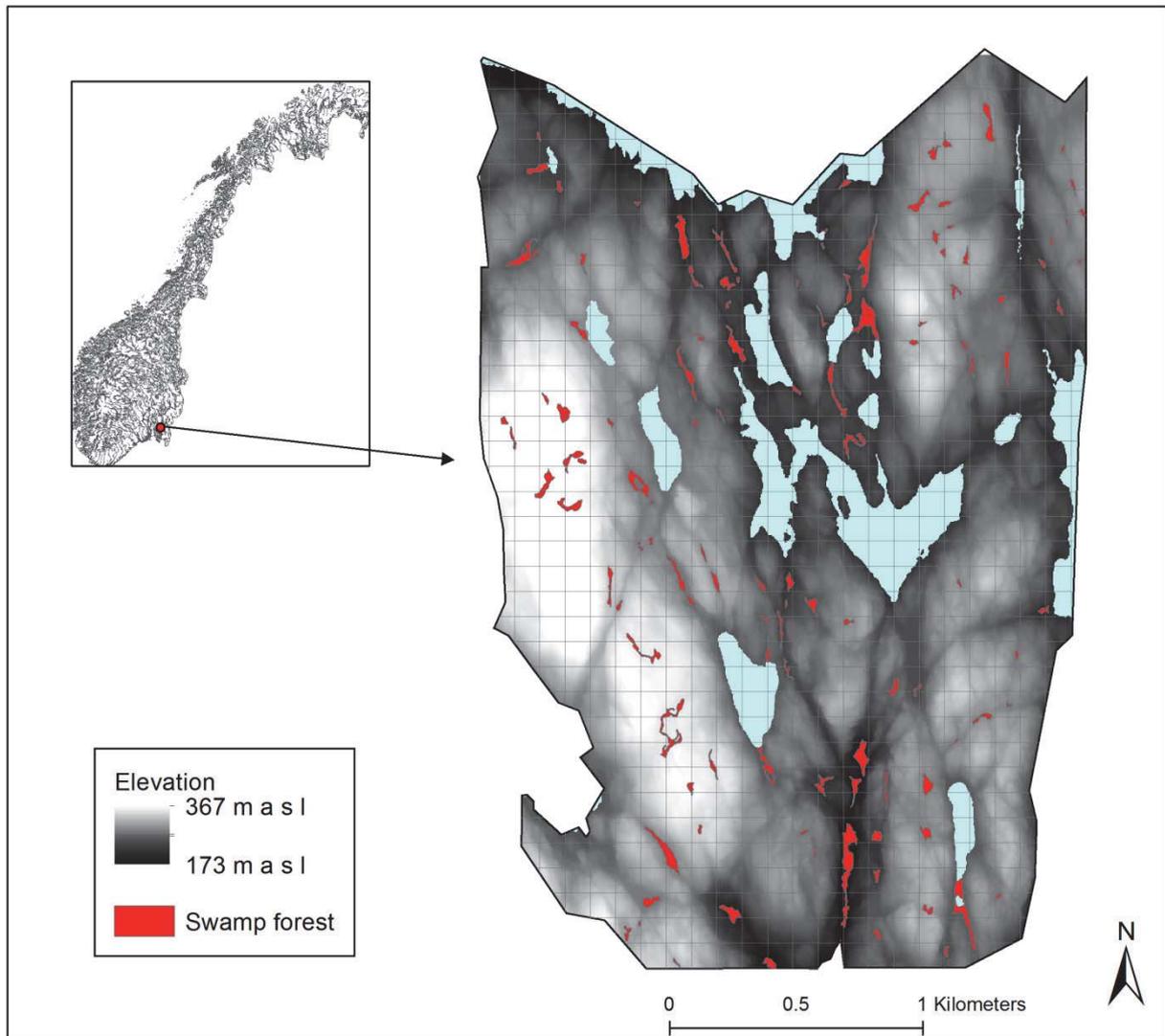


Figure 1. The location of the study area in the Østmarka Nature Reserve in Southeast Norway. The 121 swamp-forest polygons are shown in red. Lakes appear in blue. Altitude is indicated by shading; from dark-shaded low-lying areas to white areas at high altitudes. A raster (grey lines) with cells of size  $100 \times 100$  m is superimposed on the map to illustrate the size of the polygons.

### ALS derived metrics

The acquisition of ALS data used to derive variables for each grid cell in the entire study area was performed 2 July 2010 by the contractor Blom Geomatics AS, Norway (see Table 1 for technical information). For all echoes we computed the planimetric coordinates (x and y) and ellipsoidal height. Ground echoes were found and classified using the progressive triangular irregular network (TIN) densification algorithm (Axelsson 2000) of the TerraScan software. A TIN was created from the planimetric coordinates and corresponding heights of the laser

echoes classified as ground points. An uncertainty of 20-30 cm was expected for the ellipsoidal height in the TIN model (Kraus and Pfeifer 1998, Reutebuch et al. 2003).

Table 1. Technical details for the airborne laser scanning (ALS) acquisition.

	Technical specifications
Aircraft	PA31 Piper Navajo
Laser scanner instrument	Optech ALTM-Gemini
Average speed (m s <sup>-1</sup> )	80
Flying altitude (m)	900
Scan side overlap (%)	55
Scan frequency (Hz)	55
Swath width (m)	440
Max. number of echoes per pulse	4
Mean pulse density (m <sup>-2</sup> )	5
Half-scan angle (degrees)	13.8
Footprint diameter (cm)	23
Pulse repetition frequency (kHz)	100

The TIN was converted to a digital elevation model (DEM) grid cells being 5 x 5 m). This resolution was chosen because a coarser resolution, e.g. of 10 m, was considered to be too large compared to the minimum size of a registered swamp forests of 100 m<sup>2</sup> whereas a finer resolution, e.g. of 1 m, was considered too small to reflect important structures in the vegetation.

Twenty-four ALS derived metrics, hereafter called variables, describing different aspects of the forest structure, were derived separately for first and last echoes (in the case where only a single echo was detected this was counted both as a first and as a last echo). All hits above 0.5 m were considered as vegetation hits in the calculations below. We decided to use a threshold of 0.5 m for several reasons: 1) this threshold matches the ellipsoidal height accuracy of the TIN model of 20–30 cm (Kraus and Pfeifer 1998, Reutebuch et al. 2003), 2) a small uncertainty will always be present in the individual echoes, z-error < 10 cm (Anon 2013), 3) echoes which are classified as ground returns may actually be located a few decimeter above the actual ground since these echoes are often reflected from the ground vegetation and not from the actual ground (Næsset et al. 2004), and 4) to avoid stones and logs, which occur scattered throughout the study area, to be confused with small shrubs.

The 24 variables were: mean height (Hmean), skewness of heights (Hskew), kurtosis of heights (Hkurt), the coefficient of variation of heights (Hcv), the standard deviation of heights (Hsd), and nine percentiles (notation: 10<sup>th</sup> percentile, 20<sup>th</sup> percentile, ..., 90<sup>th</sup>

percentile). Furthermore, nine density variables  $D_n$  ( $n = 1, \dots, 9$ ) were derived for each grid cell by splitting the range between 0.5 m and the highest-situated echo into ten equally large vertical bins and subsequently calculating  $D_n$  as the proportion of all returns coming from bins above bin  $n$ . The proportion of echoes between 0.5 and 2 m was used as a proxy for shrub density.

The ability to predict the occurrence of swamp forests from ALS-derived variables may depend on which spatial scale the ALS variables are extracted. Therefore, we analyzed the variables described above on three scales:  $25\text{m}^2$  (single grid cell),  $225\text{m}^2$ , and  $625\text{m}^2$ . While variables calculated for  $25\text{m}^2$  give information mainly about the vertical distribution of echoes, variables calculated for the two larger sampling scales also take into account the horizontal distribution of echoes. Variables at the two broadest scales were obtained by a moving window filter approach. The value for the center cell in each of these filter windows was calculated using two different operators: mean or standard deviation (sd). If e.g. the mean operator were used to derive a value for the median vegetation height in the center cell in a  $225\text{m}^2$  window the value was calculated as the mean for all nine measures of median vegetation height within the filter.

### **Statistical analysis, modeling and model evaluation**

The Kruskal-Wallis test was applied to all 24 variables to identify variables that differed between swamp forests and non-paludified forests.

One explanatory logistic regression model [generalised linear models (GLM) with binary response variable and binomial distribution of errors; (Venables and Ripley 2002)] was obtained for each of the spatial scales by best-subset selection, as implemented in the R-package *glmulti* (R version 2.14.2, R Core Development Team 2009), using BIC (the Bayesian Information Criterion, or Schwarz' criterion; Schwartz 1978) as selection criterion. Since the main purpose of this modelling was to identify differences between swamp-forest and non-paludified forests *in general*, we wanted to avoid overly complex models (Halvorsen 2012). This was accomplished by limiting the number of variables in these models to a maximum of four. The modified version of  $R^2$  proposed by Nagelkerke (1991), which measures the proportion of variation explained by each model, was used for internal model performance assessment.

Scientific studies have frequently emphasized the importance of using an independent data set when evaluating models (Elith et al. 2006, Austin 2007, Edvardsen et al. 2011, Halvorsen 2013). Two independent evaluation data sets were created for the study area in the following way: for each of the 121 swamp forests we tried to locate a single grid cell to be used as presence cell. The only criterion for these cells was that they should be surrounded by other swamp forest cells in each of the four cardinal directions. This criterion was used to ensure that presence cells resemble the interior of a swamp forest. A total of 103 swamp forests were large enough to fulfill this criterion, resulting in 103 presence cells to be included into the evaluation data. Two different sets of absence cells were derived for evaluation of the models. The first set of absence cells was positioned close to the 103 swamp forests which contain presence cells (hereafter called: close cell). These absence cells were located no closer than 5 m and no further away than 10 m from the swamp forest border. The second set of absence cells (also 103 cells) were picked at random from all the non-paludified swamp forests cells in the study area (hereafter called: random cell).

## **Results**

### **Structural differences**

Variables that describe vegetation height (mean height, standard deviation of height and the percentiles) differed significantly between swamp forests and non-paludified forests, regardless of spatial scale (Table 2). The mean vegetation height, derived from first echoes, was 3.1 m lower in swamp forests than in non-paludified forests. The same trend was found for all percentiles. Only six density variables showed significant difference between swamp forests and non-paludified forests, of which four were means from the 225m<sup>2</sup> window (Table 2).

### **Logistic regression models**

Except for mean height, which was included in two of the models, the variables included in the three different models (different spatial scales) were not the same (Table 3). The individual variables included in the models were all highly significant (Table 3). Modified R<sup>2</sup> values for each of the models ordered from the finest to the broadest scale, were 0.22, 0.20, 0.16, respectively (Table 3).

## Model evaluation

The boxplots and confusion matrixes show that none of the models predicted presence and absence of swamp-forests well (Fig. 2 and Fig. 3). The proportion of correctly predicted presences and absences in the training data was 62–66 %, while corresponding values for the two evaluation data sets were 54–62 %, without clear systematic variation among the two evaluation data sets (Fig 3). No systematic difference in predictive ability were found for the models when evaluated on the two evaluation data sets, except for the coarse (625 m<sup>2</sup>) scale, at which the performance in the random-cell set was considerably higher (62 % correctly predicted) than in the close-cell set (54 %).

Table 2. P-values as a result of the Kruskal-Wallis test run for each of the variables, on three different spatial scales (25m<sup>2</sup>, 225m<sup>2</sup> and 625m<sup>2</sup>).

Metric	First echoes	Last echoes	First echoes		Last echoes		First echoes		Last echoes	
	25m <sup>2</sup>	25m <sup>2</sup>	225m <sup>2</sup>	225m <sup>2</sup>	225m <sup>2</sup>	225m <sup>2</sup>	625m <sup>2</sup>	625m <sup>2</sup>	625m <sup>2</sup>	625m <sup>2</sup>
			mean	sd	mean	sd	mean	sd	mean	sd
Hskew	0.069	0.018	0.0809	0.0153	0.0667	0.6180	0.0357	0.1581	0.0648	0.5317
Hkurt	0.122	0.614	0.9583	0.3391	0.1166	0.3542	0.9334	0.4609	0.0544	0.3200
Hcv	0.453	0.442	0.8249	0.5079	0.5937	0.3373	0.9101	0.2446	0.3648	0.3147
Hmean	0.000	0.000	0.0001	0.0131	0.0005	0.0043	0.0006	0.0379	0.0018	0.0401
Hsd	0.000	0.000	0.0001	0.6896	0.0000	0.8796	0.0009	0.9319	0.0008	0.3254
P90	0.000	0.000	0.0001	0.5648	0.0002	0.1888	0.0004	0.7154	0.0007	0.3155
P80	0.000	0.000	0.0001	0.4609	0.0002	0.1901	0.0005	0.3855	0.0011	0.1945
P70	0.000	0.000	0.0001	0.3354	0.0003	0.1095	0.0005	0.2141	0.0014	0.1035
P60	0.000	0.000	0.0001	0.2021	0.0003	0.0384	0.0005	0.0959	0.0017	0.0645
P50	0.000	0.000	0.0001	0.0871	0.0004	0.0093	0.0006	0.0316	0.0016	0.0195
P40	0.000	0.000	0.0001	0.0118	0.0006	0.0012	0.0005	0.0090	0.0023	0.0056
P30	0.000	0.000	0.0001	0.0015	0.0016	0.0008	0.0006	0.0025	0.0030	0.0073
P20	0.000	0.004	0.0003	0.0001	0.0026	0.0001	0.0009	0.0005	0.0027	0.0027
P10	0.001	0.042	0.0018	0.0000	0.0070	0.0005	0.0035	0.0004	0.0044	0.0028
Dn1	0.252	0.039	0.3628	0.2067	0.6376	0.1533	0.4070	0.4587	0.5773	0.0923
Dn2	0.914	0.992	0.9436	0.7556	0.4631	0.7347	0.7893	0.6468	0.4643	0.3726
Dn3	0.471	0.420	0.5245	0.8092	0.1957	0.9261	0.6039	0.8191	0.1738	0.6936
Dn4	0.083	0.128	0.1381	0.7528	0.0875	0.7822	0.2733	0.7361	0.1192	0.5150
Dn5	0.084	0.065	0.0214	0.3085	0.0495	0.9043	0.0515	0.7992	0.1166	0.3245
Dn6	0.233	0.169	0.0236	0.1791	0.0288	0.7361	0.0139	0.3085	0.0637	0.8263
Dn7	0.264	0.263	0.1888	0.0337	0.2669	0.2484	0.0604	0.3755	0.1864	0.8536
Dn8	0.978	0.799	0.7417	0.2134	0.6521	0.2484	0.3987	0.8134	0.4498	0.9232
Dn9	0.937	0.742	0.2839	0.1476	0.4443	0.3542	0.6284	0.8092	0.4498	0.7766
Shrub	0.470	0.843	0.8651	0.6855	0.4133	0.9509	0.6337	0.8941	0.1876	0.7389

Table 3. The three logistic regression models for swamp-forest presence, obtained by best-subset selection of variables on each of the three spatial scales (25 m<sup>2</sup>, 225 m<sup>2</sup> and 625 m<sup>2</sup>). Explanation for all coefficients are found in the “ALS derived metrics section”.

Scale	Coefficient	Echo category	Operator	Estimate	<i>p</i> -value	Nagelkerke R <sup>2</sup>
25m <sup>2</sup>	Intercept			1.39	0.0021	0.22
	Hmean	First		-1.21	0.0001	
	Hmean	Last		1.09	0.0005	
	Dn6	First		11.38	0.0005	
	Dn6	Last		-11.26	0.0012	
225m <sup>2</sup>	Intercept			2.10	0.0001	0.20
	Hmean	First	sd	-2.19	0.0001	
	90 <sup>th</sup> percentile	First	mean	-0.70	0.0006	
	60 <sup>th</sup> percentile	First	sd	1.97	0.0001	
	70 <sup>th</sup> percentile	Last	mean	0.66	0.0029	
625m <sup>2</sup>	Intercept			11.13	0.0000	0.16
	Hkurt	First	mean	1.01	0.0030	
	Hsd	Last	mean	-0.58	0.0001	
	Dn5	First	mean	-9.36	0.0010	
	Dn5	Last	sd	-12.24	0.0009	

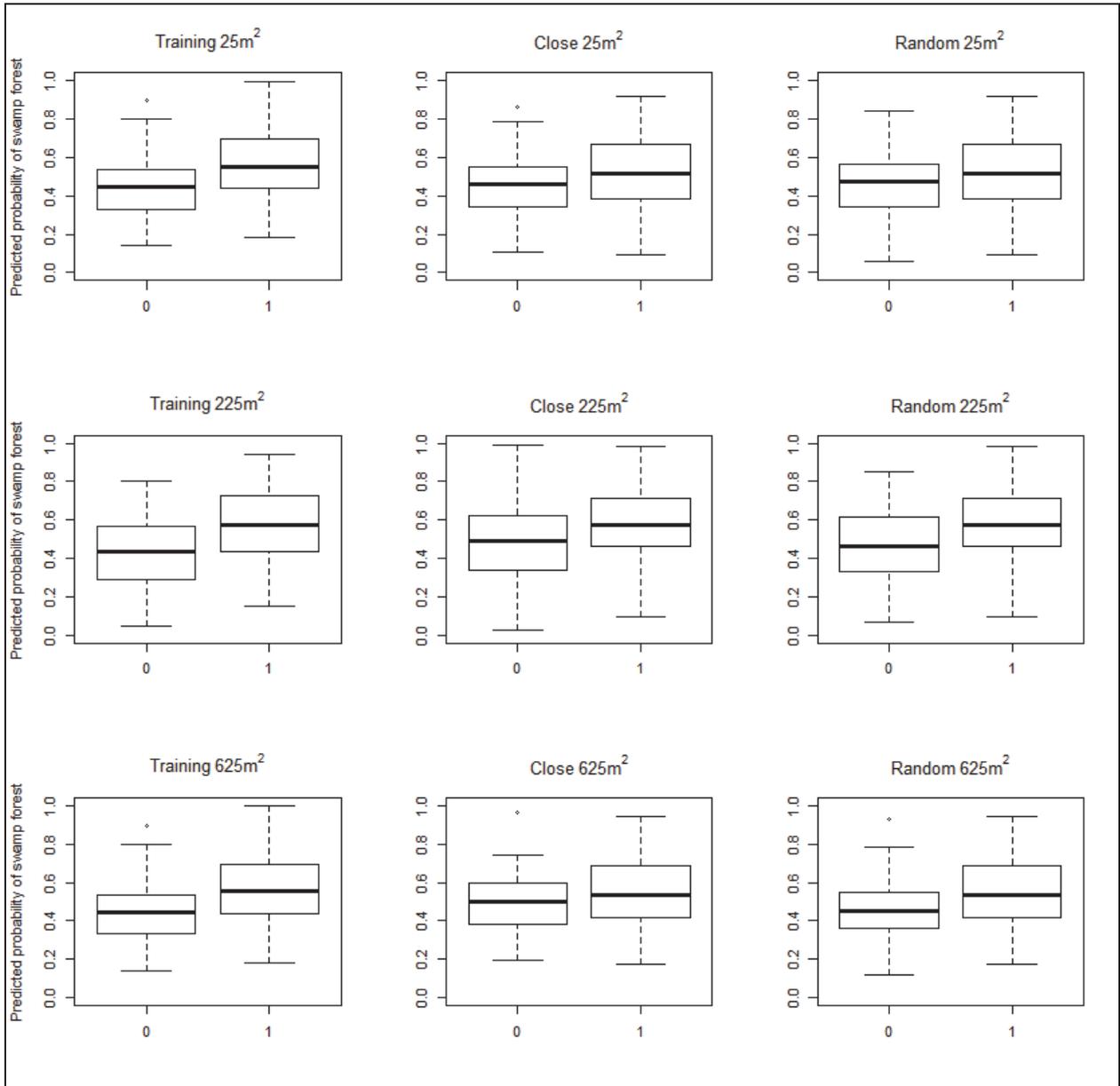


Figure 2. Proportion of grid cells for which swamp-forest presence is predicted by the three logistic regression models (covering three spatial scales: 25m<sup>2</sup>, 225m<sup>2</sup> and 625m<sup>2</sup>), in each of three data sets: the training data set (242 observations), and two evaluation data sets with 206 observations each (the Close and Random data sets are explained in the text).

Training 25m <sup>2</sup>				Close 25m <sup>2</sup>				Random 25m <sup>2</sup>			
66 %	0	1		58 %	0	1		55 %	0	1	
0	82	39	121	0	65	38	103	0	58	45	103
1	44	77	121	1	48	55	103	1	48	55	103
	126	117	159		113	94	120		106	101	113
Training 225m <sup>2</sup>				Close 225m <sup>2</sup>				Random 225m <sup>2</sup>			
62 %	0	1		59 %	0	1		60 %	0	1	
0	74	47	121	0	54	49	103	0	57	46	103
1	46	75	121	1	36	67	103	1	36	67	103
	120	123	149		90	117	121		93	114	124
Training 625m <sup>2</sup>				Close 625m <sup>2</sup>				Random 625m <sup>2</sup>			
63 %	0	1		54 %	0	1		62 %	0	1	
0	81	40	121	0	51	52	103	0	66	37	103
1	50	71	121	1	42	61	103	1	42	61	103
	131	112	152		93	114	112		108	99	127

Figure 3. Confusion matrixes for the three regression models (covering three spatial scales: 25m<sup>2</sup>, 225m<sup>2</sup> and 625m<sup>2</sup>) evaluated on the training data (242 observations) and two independent evaluation data sets with 206 observations each. A confusion matrix contains information about actual and predicted classifications (Kohavi and Provst 1998). Description for the Close and Random evaluation data sets can be found in the “Statistical analysis, modeling and model evaluation” section.

## Discussion

Results from the 240 Kruskal-Wallis single-variable tests reveal structural differences between swamp forests and non-paludified forests in the study area. All percentiles, mean height and standard deviation of height variables showed highly significant differences between swamp forests and non-paludified forest on each of the three spatial scales covering 25m<sup>2</sup>, 225m<sup>2</sup> and 625m<sup>2</sup> (Table 2). The mean difference between these significant variables when measured within the swamp forests and non-paludified forest indicates that the vegetation is lower in swamp forests than in non-paludified forests. These results support the hypothesis that the vegetation of swamp forests is lower than that of non-paludified forests. Only six out of 100 tests of density variables were significant at the  $\alpha = 0.05$  level. The hypothesis that the vegetation of swamp forests is less dense than the vegetation of non-paludified forest was therefore not supported. Even though it was hypothesized that the

density of the shrub layer (echoes from structures between 0.5 and 2 m above ground) would differ between swamp forests and non-paludified forest, our results did not support this hypothesis. Failure of ALS-derived variables to distinguish between a well-developed understory, and lack of an understory, also contrasts results of recent studies (Bollandsås et al. 2008, Martinuzzi et al. 2009, Korpela et al. 2012, Wing et al. 2012). Since deciduous shrub is found within most of the swamp forests and not in the non-paludified forest, we argue that some structures within the non-paludified forest, which resembles the echoes reflected from the deciduous shrub, probably obscure our results. Issues which are important to take into consideration when trying to detect understory, is openness of the old-growth forest we have studied, which mostly contains well-spaced trees, and the fact that ALS data are just sets of points in three dimensions. An open forest will on one hand enhance the probability of detecting understory shrubs in open swamp forests using ALS-derived variables, whereas openness in a non-paludified forest results in trees which tend to have branches situated close to the ground. Since ALS data is just points situated in a three dimensional coordinate system, it is difficult to distinguish between echoes reflected from actual shrub or from low hanging branches (mainly from Norway spruce). This confusion is thought to be the most likely reason for our inability to detect differences in understory development between swamp forests and non-paludified forests.

Even though the results from the Kruskal-Wallis test showed promising results in regard of structural differences between swamp forests and non-paludified forests, the multi-variable prediction models were unable to assign grid cells to swamp forest or non-paludified forest with higher accuracy than around 60 % (Fig 3). Furthermore, the sets of variables included in these models varied among spatial scales and model performance varied in an inconsistent manner among spatial scales and evaluation data sets so that no overall best model could be identified. A possible reason for the much poorer predictions by the broad-scale (625 m<sup>2</sup>) model in the close-cell evaluation data set than in the random-cell set is that the moving window in the former case extends over the swamp forest border, thereby also including signals from within swamp forests when variables were calculated

We think that there are two main reasons for why it is difficult to separate between swamp forest sites and non-paludified forests by forest structural characteristics measured by ALS-derived metrics: Firstly, the size of the swamp forests matters. Most of the swamp forests found in the study area are relatively small, with considerable variation in shape.

Several swamp forests contain parts that are less than five meters wide, causing strong edge effects. It is likely that trees which are actually growing outside the swamp forests have a proportion of their crown projected into the swamp forest and therefore echoes from these trees will erroneously be recorded as belonging to the swamp forest and vice versa. The edge effect could perhaps have been reduced somewhat by use of a larger basic grid-cell size than  $5 \times 5$  m grid cells, even though the grid cell size accorded to the minimum legible delineation

area (MLD) 
$$Pixel\ size = \sqrt{\frac{MLD}{4}} = \sqrt{\frac{100m}{4}} = 5m$$
 (formula presented by Hengl 2006).

Secondly, there were major structural differences among and within individual swamp forests in the study area. Typically, two swamp forest patches could have very different forest structures: one swamp forest could be dominated by small spruce and birch trees, whereas another could be entirely dominated by large spruce trees. This difference in structure could also be pronounced within individual swamp forests, e.g. with large trees in one end and an almost treeless area in the other end of the swamp forest patch, only a few meters apart. These differences in structure between different swamp forests and within the same swamp forest corroborate the results in Hörnberg et al. (1995) and Økland et al. (2001). If we had performed a stratification of the non-paludified forest and split the forest into spruce and pine forest it might have improved the predictions, since pine forest are thought to have a lower average height than spruce forests.

The results of this study complements the results of a parallel study conducted in the same study area (Dirksen et al., submitted ms), showing that ALS-derived terrain variables can be used to develop good models for the distribution of swamp forests. Nevertheless a challenge when predicting the location of swamp forests using only terrain variables is to distinguish between swamp forests and areas with open mires. Although the models from this study were not very good at predicting the location of swamp forests, we believe that including ALS derived variables which describe forests structure may improve the performance of the models developed with terrain variables only, since structure variables will help distinguish between swamp forests and tree-less open mires.

Although the predictive power of the prediction models was only moderate, it is important to remember that this kind of models can also play an important role in prediction modelling through field validation of prediction maps and successive improvement of the models (Guisan et al. 2006). Even from average prediction maps it will be possible to locate

areas that have a higher or lower probability of containing the object in focus e.g. swamp forests. These maps are of value for field workers since they can use them as guidelines for where to search for swamp forest and thereby save time. If new areas with swamp forests are encountered through the field registrations it will be possible to improve the initial prediction model by including these new observations into the group of predictor variables and create a new and improved prediction model.

## References

- Anon 2013. Gemini Summary Specification Sheet. Optech Inc.
- Aplin, P. 2005. Remote sensing: ecology. — *Prog. Phys. Geogr.* 29: 104-113.
- Austin, M. 2007. Species distribution models and ecological theory: A critical assessment and some possible new approaches. — *Ecol. Model.* 200: 1-19.
- Axelsson, P. 2000. DEM generation from laser scanner data using adaptive TIN-models. — *Int Arch Photogramm Remote Sens* 33: 110-117.
- Bater, C. W. et al. 2009. Estimation of standing dead tree class distributions in northwest coastal forests using lidar remote sensing. — *Can J For Res* 39: 1080-1091.
- Bollandsås, O. M. et al. 2008. Measures of spatial forest structure derived from airborne laser data are associated with natural regeneration patterns in an uneven-aged spruce forest. — *Forest Ecol Manag.* 255: 953-961.
- Brokaw, N. and Lent, R. (eds.) 1994. Vertical structure.
- Bässler, C. et al. 2011. LiDAR as a rapid tool to predict forest habitat types in Natura 2000 networks. — *Biodivers and Conserv* 20: 465-481.
- Coops, N. C. et al. 2010. Assessing the utility of lidar remote sensing technology to identify mule deer winter habitat. — *Can J Remote Sens.* 36: 81-88.
- Edvardsen, A. et al. 2011. A fine-grained spatial prediction model for the red-listed vascular plant *Scorzonera humilis*. — *Nord J Bot.* 29: 495-504.
- Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. — *Ecography* 29: 129-151.
- ESRI 2011. ArcGIS Desktop: Release 10.1. Redlands, CA: Environmental Systems Research Institute.
- Falkowski, M. J. et al. 2009. Characterizing forest succession with lidar data: An evaluation for the Inland Northwest, USA. — *Remote Sens Environ.* 113: 946-956.
- Gobakken, T. and Næsset, E. 2004. Estimation of diameter and basal area distributions in coniferous forest by means of airborne laser scanner data. — *Scand J For Res.* 19: 529-542.
- Guisan, A. et al. 2006. Using niche-based models to improve the sampling of rare species. — *Conserv Biol* 20: 501-511.
- Halvorsen, R. 2012. A gradient analytic perspective on distribution modelling. — *Sommerfeltia* 35: 1-165.
- Halvorsen, R. 2013. A strict maximum likelihood explanation of MaxEnt, and some implications for distribution modelling. — *Sommerfeltia* 36: 1-132.
- Hengl, T. 2006. Finding the right pixel size. — *Comput Geosci* 32: 1283-1298.
- Hill, R. A. et al. 2004. Predicting habitat quality for Great Tits (*Parus major*) with airborne laser scanning data. — *Int J Remote Sens* 25: 4851-4855.
- Hill, R. A. and Broughton, R. K. 2009. Mapping the understorey of deciduous woodland from leaf-on and leaf-off airborne LiDAR data: A case study in lowland Britain. — *ISPRS J Photogramm.* 64: 223-233.
- Hofgaard, A. 1993. Structure and regeneration patterns in a virgin *Picea-abies* forest in northern Sweden. — *J Veg Sci.* 4: 601-608.
- Hörnberg, G. et al. 1995. Stand dynamics, regeneration patterns and long-term continuity in Boreal old-growth *Picea abies* swamp-forests. — *J Veg Sci.* 6: 291-298.
- Hörnberg, G. et al. 1997. Influence of bryophytes and microrelief conditions on *Picea abies* seed regeneration patterns in boreal old-growth swamp forests. — *Can J For Res.* 27: 1015-1023.
- Hörnberg, G. et al. 1998. Boreal swamp forests. — *Bioscience* 48: 795-802.
- Kim, Y. et al. 2009. Distinguishing between live and dead standing tree biomass on the North Rim of Grand Canyon National Park, USA using small-footprint lidar data. — *Remote Sens Environ.* 113: 2499-2510.
- Kohavi, R. and Provost, F. Glossary of terms. — *Mach Learn.* 30: 271-274.

- Korpela, I. et al. 2012. Understory trees in airborne LiDAR data — Selective mapping due to transmission losses and echo-triggering mechanisms. — *Remote Sens Environ.* 119: 92-104.
- Koukoulas, S. and Blackburn, G. A. 2004. Quantifying the spatial properties of forest canopy gaps using LiDAR imagery and GIS. — *Int J Remote Sens.* 25: 3049-3071.
- Kraus, K. and Pfeifer, N. 1998. Determination of terrain models in wooded areas with airborne laser scanner data. — *ISPRS J Photogramm.* 53: 193-203.
- Kuuluvainen, T. et al. 1998. Structure of a pristine *Picea abies* forest in northeastern Europe. — *J Veg Sci.* 9: 563-574.
- Lie, M. H. 2009. Cutting history, forest structure and biodiversity in boreal forests. — Norwegian University of Life Sciences.
- Lim, K. et al. 2003. LiDAR remote sensing of forest structure. — *Prog. Phys. Geogr.* 27: 88-106.
- Martinuzzi, S. et al. 2009. Mapping snags and understory shrubs for a LiDAR-based assessment of wildlife habitat suitability. — *Remote Sens Environ* 113: 2533-2546.
- Moen, A. 1999. National Atlas of Norway: Vegetation. — Norwegian Mapping Authority.
- Mücher, C. A. et al. 2009. Modelling the spatial distribution of Natura 2000 habitats across Europe. — *Landscape Urban Plan* 92: 148-159.
- Müller, J. et al. 2009. Using airborne laser scanning to model potential abundance and assemblages of forest passerines. — *Basic Appl. Ecol.* 10: 671-681.
- Nagelkerke, N. J. D. 1991. A note on a general definition of the coefficient of determination. — *Biometrika* 78: 691-692.
- Næsset, E. 2002. Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. — *Remote Sens Environ.* 80: 88-99.
- Næsset, E. et al. 2004. Laser scanning of forest resources: The Nordic experience. — *Scand J For Res.* 19: 482-499.
- Nelson, R. et al. 2005. Locating and estimating the extent of Delmarva fox squirrel habitat using an airborne LiDAR profiler. — *Remote Sens Environ.* 96: 292-301.
- Ohlson, M. et al. 1997. Habitat qualities versus long-term continuity as determinants of biodiversity in boreal old-growth swamp forests. — *Biol Conserv.* 81: 221-231.
- Økland, R. H. et al. 2001. Vegetation-environment relationships of boreal spruce swamp forests in Østmarka Nature Reserve, SE Norway. *Sommerfeltia*, 29. p 190.
- Økland, R. H. et al. 2008. Species richness in boreal swamp forests of SE Norway: The role of surface microtopography. — *J Veg Sci* 19: 67-74.
- Reutebuch, S. E. et al. 2003. Accuracy of a high-resolution lidar terrain model under a conifer forest canopy. — *Can J Remote Sens.* 29: 527-535.
- Schwarz, G. 1978. Estimating the dimension of a model. — *Ann Stat.* 6: 461-464.
- Sigmond, E. M. O. et al. 1984. Berggrunnskart over Norge 1: 1 000 000. — *Norg. geol. Unders.*
- Smart, L. S. et al. 2012. Three-dimensional characterization of pine forest type and red-cockaded woodpecker habitat by small-footprint, discrete-return lidar. — *Forest Ecol Manag.* 281: 100-110.
- Spies, T. A. 1998. Forest structure: A key to the ecosystem. — *Northwest Sci.* 72: 34-39.
- Turner, W. et al. 2003. Remote sensing for biodiversity science and conservation. — *Trends Ecol Evol.* 18: 306-314.
- Vehmas, M. et al. 2009. Identification of boreal forest stands with high herbaceous plant diversity using airborne laser scanning. — *Forest Ecol Manag.* 257: 46-53.
- Venables, W. N. and Ripley, B. D. 2002. Modern applied statistics with S. — Springer.
- Vierling, K. T. et al. 2008. Lidar: shedding new light on habitat characterization and modeling. — *Front Ecol Environ.* 6: 90-98.
- Wing, B. M. et al. 2012. Prediction of understory vegetation cover with airborne lidar in an interior ponderosa pine forest. — *Remote Sens Environ* 124: 730-741.



# Paper 3



# **Modelling richness and presence of bird species in a boreal forest reserve using airborne laser scanning and aerial images**

## **Authors**

John Wirkola Dirksen<sup>1\*</sup>, Rune Halvorsen<sup>2</sup>, Hans Ole Ørka<sup>1</sup>, Katrine Eldegard<sup>1</sup>, Erik Næsset<sup>1</sup>,  
Terje Gobakken<sup>1</sup> and Mikael Ohlson<sup>1</sup>.

## **Addresses**

<sup>1</sup>) Norwegian University of Life Sciences, Department of Ecology and Natural Resource  
Management, PO Box 5003, NO-1432 Ås, Norway

<sup>2</sup>) University of Oslo, Natural History Museum, Department of Botany, PO Box 1172  
Blindern,  
NO-0318 Oslo, Norway

## **\*<sup>1</sup>) Corresponding author**

John Wirkola Dirksen

E.mail [john.dirksen@umb.no](mailto:john.dirksen@umb.no)

Phone: +47 64 96 57 31

## **Abstract**

During the last decade airborne laser scanning (ALS) has showed its potential for modelling environmental variables that determine forest structure. Thus, distribution models for forest dwelling birds may potentially be improved by access to three-dimensional structural information obtainable from ALS data. This is the first study in Scandinavia that addresses bird habitats in an old-growth boreal forest using ALS data. We build explanatory GLM models and prediction models by GLM and Random Forest (RF) for species richness and the eight most abundant species within the study area. Vertical distribution of the canopy was the most important variable for both species richness and single-species. The best explanatory model explained 27% of the variation in occurrence. Prediction models were evaluated by cross validation, and both the area under the curve (AUC) and Cohen's kappa were computed. AUC values for the different species ranged between 0.48-0.80 for GLM and 0.57-0.77 for RF. Kappa values ranged between 0-0.22 for GLM and 0-0.36 for RF. Prediction models developed with both ALS and spectral images made better predictions than models based on only one of these data sources, indicating that access to ALS data adds significantly to predictability of bird distributions in boreal forests.

## **Introduction**

Informed management actions aimed at preventing biodiversity loss depend on knowledge about how species are related to their habitats, both in human-modified habitats, and in habitats modulated by natural disturbance dynamics (Whittaker 2005, Elith et al. 2006). During the last decade an increasing number of scientists have developed species distribution models that aim at explaining and predicting species occurrence patterns, by combining observations of species presence and abundance with estimates of habitat characteristics (Guisan and Zimmermann 2000, Whittaker et al. 2005, Pearson et al. 2007, Elith and Leathwick 2009; Gogol-Prokurat 2011). Field-based studies of birds on fine spatial scales have revealed that the three-dimensional arrangement of the habitat strongly influences bird habitat use (e.g. MacArthur and MacArthur 1961, Brokaw and Lent 1999). For forest-dwelling birds, the three-dimensional structural complexity of the forest, including canopy height and tree species composition, influences the presence of single bird species as well as bird richness (MacArthur and MacArthur 1961, Willson 1974, Sutherland et al. 2004). However, there is still no general agreement as to how results obtained in fine-scale studies scale up to the population and landscape levels (Vierling et al. 2008). This problem of predicting broad-scale responses by up-scaling from fine-scale data has been a recurring theme in ecology (Miller et al. 2004) because traditional field work techniques are often labour-intensive and detailed information cannot be sampled over population-wide areas (Johnson and Gillingham 2008).

Remote sensing increasingly provides an effective tool for mapping and quantifying selected habitat characteristics over large areas (Horning et al. 2010). In order to expand the scale of their studies, ecologists have commonly used spectral data that provide two-dimensional representations of the forest. Such spectral images makes it possible to distinguish between several landscape structures, but vegetation types and horizontal information about the forest structure can only be vaguely inferred from spectral images (Turner et al. 2003, Bergen et al. 2009, Horning et al. 2010). In contrast, airborne laser scanning (ALS) technology provides an accurate and objective three-dimensional representation of the Earth's surface and notably, the vegetation on and above it (Lefsky et al. 2002, Næsset et al. 2004). During the last decade the number of ecologists that have used ALS-derived variables to describe the three dimensional structure of forests has increased

considerably (Turner et al. 2003, Aplin 2005, McDermid et al. 2005, Bässler et al. 2011). ALS is an active remote sensing technique, which provides measurements of forest structure or terrain forms in three dimensions. The operating principle of ALS is to measure the time span from a laser pulse is emitted from an aircraft mounted laser until its echo is received by the instrument after having been reflected from an object near the ground or by the ground surface itself. The time span recorded between emission and reflection of the laser pulse is used to estimate the distance between the sensor and the object on which the laser pulse was reflected. Because the exact position and orientation of the sensor is provided by use of global satellite navigation systems and internal navigation systems, the spatial location of the echoes can be accurately determined. The result from an ALS acquisition is a three dimensional point cloud, each point being an echo with its own unique position in an x, y, and z coordinate system. Echoes returned from the ground are used to derive a terrain model, and a measure for vegetation height is derived by subtracting the terrain model from non-ground echoes.

Since 2004 ALS has been fully operational for forest inventories in Scandinavia (Næsset 2004) and important forest characteristics like mean tree height (Næsset 1997a), timber volume (Næsset 1997b) and canopy height (Næsset and Bjercknes 2001) are now easily derived from ALS data. ALS has also been utilized in a range of ecological applications; e.g. for classification of forest successional stages (Falkowski et al. 2009), detection of understory scrubs (Hill and Broughton 2009, Martinuzzi et al. 2009, Wing et al. 2012), quantifying properties of canopy gaps (Koukoulas and Blackburn 2004) and estimation of standing dead trees (Bater et al. 2009, Kim et al. 2009). As birds in particular are exploiting the three dimensional 'landscape' of the forest canopy (e.g. Karr and Roth 1971, Holmes et al. 1979), it is not surprising that bird ecology is the sub-discipline within ecology where ALS has been used the most. For example, ALS data have been used to detect forest structures important for single bird species (Hill et al. 2004, Broughton et al. 2006, Hinsley et al. 2006, Graf et al. 2009, Smart et al. 2012), whole bird communities (Goetz et al. 2007, Clawges et al. 2008, Müller et al. 2009, Seavy et al. 2009) or different types of bird guilds (Goetz et al. 2007, Lesak et al. 2011).

Previous studies that have investigated the relationship between forest-dwelling birds and ALS-data have been conducted in many different types of forests e.g. broadleaf forests in England (Hinsley et al. 2002, Broughton et al. 2006, Bellamy et al. 2009). Spruce (*Picea*

*abies* L.), fir (*Abies alba* Mill) and beech (*Fagus sylvatica* L.) forests in Germany (Müller et al. 2009, Müller et al. 2010). Pine (*Pinus sylvestris* L.) and aspen (*Populus tremula* L.) forests in USA (Clawges et al. 2008). Spruce forests in Italy (Tattoni et al. 2012) and pine forests in the USA (Smart et al. 2012). Yet, despite the fact that ALS is increasingly used for scientific and commercial purposes within forest inventory in Fennoscandia (Næsset et al. 2004, Vehmas et al. 2009, Ørka et al. 2010), no studies have been performed in Fennoscandia in which relationships between bird data and habitat characteristics estimated by ALS have been modelled. Moreover, to our knowledge, no previous studies have explored potential relationships between ALS metrics and bird data in boreal forest reserves.

Regardless of forest type and bird species, previous studies which have used ALS data to build distribution models for forest dwelling birds have shown that measures of the vegetation height are important ALS variables that should be included in such models (Bellamy et al. 2009, Hinsley et al. 2009, Broughton et al. 2006). Our study was conducted in a boreal forest reserve, consisting largely of old-growth forest, and thus there was less variation in successional stages and vegetation height than usually found in managed forests. Therefore, we expected that canopy density variables would be relatively more important than pure height variables for predicting species richness and presence of single bird species.

In this paper, we explored the utility of ALS metrics to explain and predict bird species richness and presence of single bird species in a boreal forest reserve in southeast Norway. Our primary objectives were: (1) to assess the utility of ALS metrics for explaining bird species richness and single-species presence; (2) to assess the utility of ALS metrics for predicting bird species richness and single-species presence; (3) to compare the ability of ALS-derived metrics with metrics derived from more traditional remote sensing imagery for predicting bird species richness and single-species presence. A secondary objective was to assess the relative merits of two different statistical techniques for prediction modelling, i.e. Generalized Linear Models (GLM) and Random Forest (RF).

## **Material and methods**

### **Study area**

The study was carried out in the Østmarka Nature Reserve (59°50′ N, 11° 02′ E, 190–368 m above sea level), which is located in the southern boreal zone of Southeast Norway (Moen 1999).

The study area covers approximately 6 km<sup>2</sup> and is located within the core of the reserve. One part of the reserve was protected in 1990 while the reserve was expanded in 2002 to comprise approximately 12.5 km<sup>2</sup>. Until around 1940 selective logging was performed in some areas, but no clear-cutting has been performed in the reserve. The reserve is dominated by old-growth spruce forest, whereas pine forests occur on drought-exposed ridges and deciduous trees, mainly birch (*Betula sp.*) occur scattered throughout the study area. Mean annual precipitation and temperature (1961–1990 normals) for the nearest meteorological stations (Enebakk, ca. 10 km SE of the area, and Hakadal–Blikrudhagan, ca. 30 km N of the area) were 816 mm and 3.9 °C, respectively (Aune 1993, Førland 1993). The bedrock is of proterozoic age and quartz diorite, tonalite and gneisses of diverse origins dominate (Sigmond et al. 1984). The topography of the study area is dominated by north/south facing ridges and valleys. On a fine scale, the landscape is dissected by minor ridges and valleys, stream beds, large boulders, steep cliffs and more gentle slopes, resulting in a broken topography. The area is relatively species poor, a total of 28 bird species were recorded in this study, table 3 gives a summary of the most important habitat characteristics for the eight most common species.

### **Placement of bird counting stations**

To obtain measures of bird species richness and presence of single bird species in the study area, we used a stratified random point transect sampling technique (Bibby 2000). Stratification was used to ensure that a range of habitat conditions within the study area was represented and to capture the variation in vegetation structure (height) and composition (tree species), which are known to be important factors influencing bird-habitat associations (e.g. Sutherland et al. 2004). Five different strata were defined based on prior information about bird species presences in the study area, and assumptions about the species-habitat associations of the bird species within the area. The strata were: 1 – ‘tall’; 2 – ‘middle’; and 3

– ‘low’ spruce forest; 4 – ‘pine’ forest and 5 – ‘deciduous’ forest (Table 1). The three spruce forest canopy height strata: ‘tall’, ‘middle’ and ‘low’ were derived from a wall-to-wall acquisition of ALS data from the study area (see details below). We calculated the mean return height of all echoes above 0.5 m within  $5 \times 5$  m grid cells covering the whole study area and used this information to create polygons for areas dominated ( $> 50$ ) by cells with low (0–10 m), middle (10–20 m) and tall (20–40 m) vegetation. Thereafter, we identified areas dominated by Scots pine or deciduous trees (in this case mainly birch *Betula ssp*). Within the study area, pine forest is typically located on ridges. A digital elevation model was used to locate ridges that made up the ‘pine’ stratum. No continuous large areas dominated by deciduous trees were identified in the study area, but swamp forests typically hold a high proportion of deciduous trees compared to the surrounding forest and accordingly we used 121 swamp forests mapped for a separate research project in 2010 as a ‘deciduous’ stratum. Since the ‘low’, ‘middle’ and ‘tall’ height strata were dominated by spruce we did not define a separate spruce stratum. Areas with open water and open mires were excluded from this study.

A standardized procedure was used to allocate observation units (center points for counting stations) to the five different strata. For each of the five strata we randomly selected up to 50 counting stations. If a location was located so close to the border of a stratum that more than 50 % of the counting station was located outside of the strata or closer than 150 m to a counting station already picked, it was manually moved towards the interior of the polygon so approximately 70% of the counting station was within the strata if possible, if not, a replacement location was selected at random. This procedure was used to minimize edge effects and to ensure that counting stations were located at minimum 150 m apart. Since many of the deciduous and pine-dominated polygons were small, parts of counting stations (of 50 m radius) sometimes fell outside the stratum boundary. From the 50 random locations obtained for each stratum, a sample of counting stations was obtained by first selecting 10 counting stations at random from the ‘deciduous’ stratum, which covered the smallest area, followed by selecting 10 counting stations from the other strata in the order ‘pine’, ‘low’, ‘tall’ and ‘middle’, respectively. Then, 10 more counting stations from each stratum were added to the sample, in the same order: ‘deciduous’, ‘pine’, ‘low’, ‘tall’ and ‘middle’. When 20 counting stations had been located within each stratum it was not possible to locate ‘deciduous’ stratum polygons which were large enough to contain an additional counting

station. Additional counting stations were located in the ‘pine’, ‘low’, ‘middle’ and ‘tall’ stratum (Table 1) according to the defined criteria. The positions for center points of counting stations were located with a handheld GPS (Garmin 60CSx) in the field.

Table 1. Distribution of the 157 selected counting stations on five strata and characteristics of the strata according to vegetation height and tree species.

Stratum name	Tree species	Percent cover of total study area	Number of counting stations
tall	spruce	25	38
middle	spruce	58	43
low	spruce	11	30
pine	pine	4	26
deciduous	mixture of spruce and deciduous	2	20

### **Bird counting procedures**

The bird counting on the 157 counting stations was carried out between May 8<sup>th</sup> and June 10<sup>th</sup>, 2012. Each counting station was visited twice by the same, experienced, observer. Bird counting was carried out between sunrise and 10:00 AM (local summer time). At the center of each counting station, the observer recorded birds for five minutes at each visit. Each counting station had an ideal radius of 50 m, which was determined by visual assessment. Only birds recorded by visual and audio observations within the 50 m radius were registered. The observer practiced distance estimation before data collection started in order to increase the precision of the distance assessments. Moreover, the observer had experience from previous bird surveys using the same sampling and distance estimation technique. Singing males were registered as couples, and if one male and one female bird were observed, they were counted as one couple. Non-territorial birds (e.g. overflying birds) and flocks were not counted. Bird counting was only performed when it was not raining and/or the wind-speed was above 8 meters/second.

### **Airborne laser scanning**

Airborne laser scanning data covering the entire study area were acquired on July 2<sup>nd</sup>, 2010. Flight and acquisition parameters are shown in Table 2. For all echoes x and y coordinates and ellipsoidal height values were computed. Ground echoes were classified using the progressive Triangular Irregular Network (TIN) densification algorithm (Axelsson 2000) of the TerraScan software (Terrasolid LTD., 2004).

Table 2. Technical details for the ALS acquisition.

	Technical specifications
Platform	PA31 Piper Navajo fixed-wing aircraft
Sensor	Optech ALTM-Gemini
Mean flying speed (m s <sup>-1</sup> )	80
Flying altitude above ground (m)	900
Side overlap (%)	55
Scan frequency (Hz)	55
Swath width (m)	440
Max. number of echoes per pulse	4
Pulse density (m <sup>-2</sup> )	5
Half-scan angle (degrees)	13.8
Footprint diameter (cm) <sup>a</sup>	19
Pulse repetition frequency (kHz)	100
Contractor	Blom Geomatics AS, Norway

<sup>a</sup> Computed after (Baltsavias 1999) based on mean acquisition settings.

A TIN was created from the x and y coordinates and corresponding heights of the laser echoes classified as ground echoes points. The relative height of each echo was computed as the difference between the height of the individual return and the TIN.

Results of preliminary analyses revealed that splitting the data into first and last echoes which are treated separately in the analysis did not improve the predictions. We therefore analyzed all echoes together. Echoes below 0.5 m were considered as ground hits and echoes above 0.5 were considered as vegetation hits. We used a threshold of 0.5 m for the following reasons (1) the ellipsoidal height accuracy of the TIN model was expected to be around 20–30 cm (Kraus and Pfeifer 1998, Reutebuch 2003); (2) a small measurement error (the z-error) will always be present in the individual echoes ( $z < 10$  cm; Baltsavias 1999); and (3) large stones and logs that occurred scattered throughout the study area could be confused with shrubs.

A total of 66 variables were computed from the ALS echoes within the 50 m circular buffer around the 157 counting stations. Both traditional field-based studies, and studies which have used ALS derived variables, have shown that vegetation height is one of the most important variables influencing bird distribution (Seavy et al. 2009, Smart et al. 2012, Tattoni et al. 2012). From the height distribution we computed maximum height (Hmax), mean height (Hmean), standard deviation (Hsd), coefficient of variation (Hcv), kurtosis (Hkurt), and skewness (Hskew). The nine deciles (10<sup>th</sup> percentile, 20<sup>th</sup> percentile,...,90<sup>th</sup> percentile) were computed for all echoes above 0.5 m. Furthermore, two different sets of variables describing canopy density were derived: (1) canopy density; and (2) density layers. Canopy density variables were calculated as the proportion of echoes above a certain height bin to

total number of echoes. The height bins were defined by dividing the height for the 95<sup>th</sup> percentile minus 0.5 m into ten equally high vertical height bins (D1, D2,...,D9). Density layers measure the proportion of echoes registered within certain one-meter sections of the canopy to the total amount of echoes for each individual counting station. Since the highest echo was situated 40 m above ground, we decided to use 40 m as an overall maximum height measure, and divided the canopy into 40 equally vertical height bins. Thus, each bin (density layer) covers one meter of the canopy following the notation LD1, LD2, ..., LD40 of which the LD1 corresponds to the layer between 0 and 1 meter, LD2 the layer between 1 and 2 meter, and so on. The motivation for including both canopy density and density layers as variables in the analyses of bird distributions, was that vegetation architecture (height, structure and density of vegetation) often affects birds by providing perches, song posts and cover, and by limiting the bird's field of view and ability to fly and to find and capture prey (e.g., Sutherland et al. 2004). Moreover, different bird species relate differently to different parts of the tree, and therefore may be affected by the density of branches in the particular layers (e.g. Alatalo 1982). We hypothesized that density-layer variables describe the distribution of birds better than canopy density variables, since density-layer variables are independent of the height distribution of the echoes in each of the 157 counting stations, i.e., that one will get one value for a certain layer which is comparable with exactly the same layer at all other stations. On the other hand, canopy variables are dependent on the 95<sup>th</sup> percentile within each counting station and can therefore not be compared exactly among counting stations. Finally, two variables describing the density of the shrub layer (SD2 and SD3), important for certain bird species, were computed. SD2 was computed as the proportion of pulses between 0.5 and 2 m to the total amount of vegetation hits and SD3 was computed as the proportion of pulses between 0.5 and 3 m to total amount of vegetation hits. All variables were zero-skewness standardized (Økland et al. 2001) before analyses and ranged to a common 0–1 scale to enhance comparability of effect sizes for variables with different units of measurement.

### **Spectral images**

Fourteen spectral images were taken with a Vexcel UltraCam XP camera on August 8<sup>th</sup> 2010. The Vexcel sensor acquired both panchromatic and four multispectral bands (red, green, blue, and infrared). In the current study only multispectral imagery was used. Acquired from a

mean flight height 2850 meters above ground the Ground Sampling Distance (GSD) was 51 cm for the multispectral bands, and they were delivered according to the processing level 2 defined by Vexcel Imaging GmbH.

To co-register the ALS and image data, each laser echo coordinate were mapped to the image planes of the Vexcel UltraCam digital images using a rotation matrix (Mikhail et al. 2001). The digital numbers from all bands of the pixel position of the laser echo were assigned to the respective laser echo in the original scale. If the laser echo pixel position occurred on more than one image, the mean band value of the pixels from all images was assigned to the respective laser echo. Furthermore, since only first returns (first of many and single echoes) will be reflected close to the canopy surface, and all other subsequent echoes will penetrate into the canopy, image pixel values were only assigned to the first returns.

The rationale behind including image data in the analysis was to add information that better describes differences in species composition than ALS data. Therefore, variables which have been used in other studies to combine laser and imagery for species estimation and classification were used (Packalèn et al. 2009; Waser et al. 2011, Ørka et al. 2012). From each of the four bands of the UltraCam imagery descriptive features (mean, standard deviation, coefficient of variation, skewness, kurtosis and percentiles (10th, 30th, . . . , 90th) were used to compute the relative band values (the individual band divided by the sum of all bands), and the band ratios (one band divided by another band). Thus, a total of 240 image features were computed for each counting station.

## **Statistical analyses**

### **Modelling**

Response variables used in the modelling were species richness ( $y$  = number of species) and binary presence  $y$  of each of the eight most abundant bird species ( $y = 1$  means that the species was observed while  $y = 0$  means not observed). For each response variable, we used GLM (e.g., Venables & Ripley 2002) to create both an explanatory model and a prediction model. In addition we developed a Random Forest (RF; Breiman 2001) prediction model for each species, in order to compare the predictive abilities and performance of two different prediction modelling techniques.

## **Explanatory models**

For each of the eight species we fitted logistic regression models (GLM models with binary response and binomial distribution of errors; Venables and Ripley 2002). For species richness we fitted a GLM model with positive integer values for the response, assuming a Poisson distribution of errors. Model reduction was done by best-subset selection, as implemented in the R-package *glmulti*. As explanatory variables in these models we only included ALS-derived variables. Since these models were to be used as explanatory models, they should be simple in terms of number of parameters (e.g., Halvorsen 2012), and thus a maximum of four variables were included in each model. The modified version of  $R^2$  proposed by Nagelkerke (1991) was used to evaluate the explanatory models.

## **Prediction models**

Two kinds of prediction models were created in parallel: GLM10 and Random Forest (RF) models. GLM10 models were developed in the same manner as the explanatory models, but with the specification that each GLM10 model was allowed to include 10 variables.

Random Forest models were created using the random forest algorithm (Breiman 2001). To account for imbalanced data, i.e., due to rare species, a balanced Random Forest procedure was used (Chen et al. 2004). Random Forest uses randomly selected subsets of the observations and explanatory variables to create multiple regression trees, and an average of these trees are used for making predictions of, e.g., single bird species. The *RandomForest* package in R (Liaw and Wiener 2002) was used to build the models for regression (species richness) and classification (presence or absence of single species). For each of the eight most abundant species, the ten most important variables ranked from the out-of-bag accuracy using all variables, i.e., by mean decrease in classification accuracy from a classification for each species and mean increase in mean square error for the species richness.

We compared GLM10 and RF models built from three different sets of explanatory variables: (1) all ALS-derived variables; 2) all spectral image-derived variables; and (3) all ALS-derived variables and spectral image-derived variables pooled (denoted ALS+spectral). Prediction models were evaluated by use of a 10-fold cross validation procedure. As measures of model performance we used Cohen's kappa ( $\kappa$ ) and AUC [the area under the receiver operating characteristic (ROC) curve (Pearce and Ferrier 2000)] in parallel. No AUC

and kappa values could be computed for GLM and RF models describing species richness since species richness is a count response variable.

Table 3 is a summary of the most important habitat characteristics for eight of the most common birds found in the reserve (extracted from Cramp 1988, 1992, 1993 and 1994).

Species	Habitat
Chaffinch <i>Fringilla coelebs</i>	In Fennoscandia, with density highest in deciduous forests, also occurs in spruce and pine forests.
Robin <i>Erithacus rubecula</i>	Preferred habitat includes cool shade, moisture, cover of at least medium height and not more than medium density; prefers patches or fringes of open ground and song posts that give adequate view without undue exposure. In Fennoscandia occupies mixed and coniferous forests, especially with moist moss-rich forest floor and ample dead wood.
Goldcrest <i>Regulus regulus</i>	In the breeding season strictly arboreal, occurring in more or less dense stands of well-grown conifers. Prefers spruce forest; pine forests are less attractive. Character and height of undergrowth is irrelevant. Inhabits broad-leaved forests only when at least some spruce trees occur intermixed.
Willow warbler <i>Phylloscopus trochilus</i>	Less dependent of mature woodland than of shrub, typical of secondary-growth forests, and of transitions to more open habitats. Readily settles on shrubby, bushy, or cleared ground in early phases of regrowth, and thus typical of fringe areas such as forest clearings. Although resorting to woods of many different tree species, both deciduous and coniferous, prefers sites which provide the most accessible insect food, birch in particular. Also found in younger conifer stands and in moist places with alder ( <i>Alnus ssp</i> ) and willow ( <i>Salix ssp</i> ), as well as forest glades. Rare in pure spruce forest.
Great tit <i>Parus major</i>	Dislikes pure coniferous forest, prefers mixed forest types and open stands or even sites with fragmented and scattered trees to dense, pure deciduous forest. High tree cover may be less important than the structure and density of the undergrowth; opportunities for ground feeding are particularly important.
Tree pipit <i>Anthus trivialis</i>	Ground-nester and ground feeder, typical of trees and shrubs that serve as look-outs and song-posts. Avoids both open treeless and shrubless habitats and sites where the density of woody vegetation leaves insufficient access to open, low herbage
Wren <i>Troglodytes troglodytes</i>	Suitable habitat offered by a wide variety of forests with low cover and good foraging opportunities. Among the most favourable habitats are moist deciduous or mixed forests with much undergrowth and thick cover. In the northern parts of its range, breeds in taiga and mixed forests or open woodland provided these contain dense clusters of shrubs, nettles, or other undergrowth, and heaps of dead vegetation and dry branches.
Siskin <i>Carduelis spinus</i>	Mainly occupies spruce but also pine forest, especially in areas with tall trees that are not growing too close to each other, sometimes occurs in mixed coniferous and broad-leaved forests.

## Results

A total of 28 bird species were observed within the 50 m radii of the 157 counting stations. Species richness varied from 0 to 5 with a median of 3 within each counting station, which was too low to allow for analyses of changes in species composition (e.g., by ordination methods). The eight most abundant species were recorded in 23 to 99 of the counting stations while each of the remaining species was observed in  $\leq 9$  counting stations. The distribution of observations on the five strata differed widely among species (Table 4).

### Bird-habitat associations - explanatory models

Of the explanatory GLM models for species richness and the eight most abundant bird species, two models contained no variables (Table 5). The  $R^2$  values of these models range between 0.06 and 0.27 (Table 5). LD variables were included in all the models, two percentile variables and D variables were included in one model each (models for Goldcrest and Wren, respectively) (Table 5).

Table 4. Number of times each of the eight most abundant bird species were recorded in the each of the five strata: pine = areas dominated by pine; tall = areas dominated by spruce trees with a height between 20 and 40 m; middle = areas dominated by spruce trees with a height between 10 and 20 m; and low = areas dominated by spruce trees with a height below 10 m. deciduous = swamp forests, typically with a tall proportion of deciduous trees. Note that the number of counting stations within each stratum ranged from 20-43.

Strata name	No. stations	Species richness	Chaf finch	Robin	Gold crest	Willow warbler	Great tit	Tree pipit	Wren	Siskin
pine	26	67	17	7	5	10	11	7	3	6
tall	38	108	30	19	21	4	14	3	13	4
middle	43	135	29	23	25	22	17	7	3	9
low	30	65	13	8	7	15	7	12	0	3
deciduous	20	46	10	12	6	5	6	2	4	1

### Bird-habitat associations – prediction models

For six of the species, GLM10 models were obtained that included ALS variables (Figure 1), whereas only four models were developed using spectral and ALS+spectral variables (Figure 2). No GLM10 models were developed for Great tit and Siskin (Figure 2), resulting in a total of 14 GLM10 models. When GLM10 models were developed for single-species using ALS, spectral and ALS+spectral variables the number of variables included in the models were 1–4, 1–7 and 2–7, respectively. For species richness no GLM models were obtained for the

spectral and the ALS+spectral variables. RF models were created for all single bird species using variables from all three groups, resulting in a total of 27 RF models (Figure 2).

### **Evaluation of prediction models by cross-validation**

Generally RF models performed better than GLM models (Figure 2). The RF models created with ALS+spectral variables had the highest AUC and kappa values for Chaffinch, Robin and Tree pipit (Figures 2a, 2b and 2f, respectively) whereas for Willow warbler and Goldcrest the best models, also RF models, only included spectral variables and ALS variables, respectively (Figure 2d). For Wren, the best model was created by GLM, using ALS variables only (Figures 2c and 2g, respectively). For Great tit no GLM models were created and the three different RF models performed equally well. The model which had the highest AUC value (0.80) was the GLM10 model for Wren, using ALS variables only. The highest kappa value (0.36) was obtained for the RF models developed for Wren and Goldcrest.

Table 5. Results of GLM explanatory models for bird-habitat associations (Poisson regression for species richness and logistic regression for binary presence/absence data for of each of the eight most abundant bird species). Data consists of recordings in 157 counting stations in a boreal forest reserve. LD variables convey information about the density within a certain 1-m vertical section of the canopy; D variables describe canopy density; and percentile variables describe the height distribution of the canopy. # significant variables = Number of explanatory variables, which were significantly associated with the response when tested individually..  $R^2$  = the modified version of  $R^2$  proposed by Nagelkerke (1991). Estimate = beta. SE = standard error. Z = students z-value. df= degrees of freedom. p = p-value.

Response	# significant variables	$R^2$	Explanatory variable	Estimate	SE	Z	df	p
Species richness	4	0.09	LD18	0.4	0.50	2.72	1,155	0.0065
Chaffinch	32	0.11	LD18	2.4	0.69	3.47	1,155	0.0005
Robin	4	0.09	LD16	2.7	0.84	3.18	1,155	0.0015
Goldcrest	54	0.27	70 <sup>th</sup> percentile	26.4	5.97	4.43	3,153	<0.0001
			90 <sup>th</sup> percentile	-31.4	7.33	-4.30	3,153	<0.0001
			LD28	5.9	1.77	3.36	3,153	0.0008
Willow warbler	32	0.06	LD29	3.0	2.20	1.38	2,154	0.1692
			LD30	-1.8	2.09	-0.87	2,154	0.3872
Great tit <sup>+</sup>	0	-	-	-	-	-	-	-
Tree pipit	49	0.13	LD6	2.6	4.90	0.53	4,152	0.5950
			LD7	-6.7	5.07	-1.31	4,152	0.1870
			LD11	2.5	1.87	1.35	4,152	0.1760
			LD15	0.2	1.34	0.19	4,152	0.8530
Wren	39	0.21	D5	16.6	6.11	2.71	3,153	0.0067
			D7	-15.6	5.63	-2.77	3,153	0.0057
			LD9	-4.5	1.82	-2.50	3,153	0.0125
Siskin*	1	-	-	-	-	-	-	-

+No model was created for Great tit; no one of the explanatory variables were significantly related to the response.

\* No model with explanatory variables was built in the best subset selection, model with intercept only.

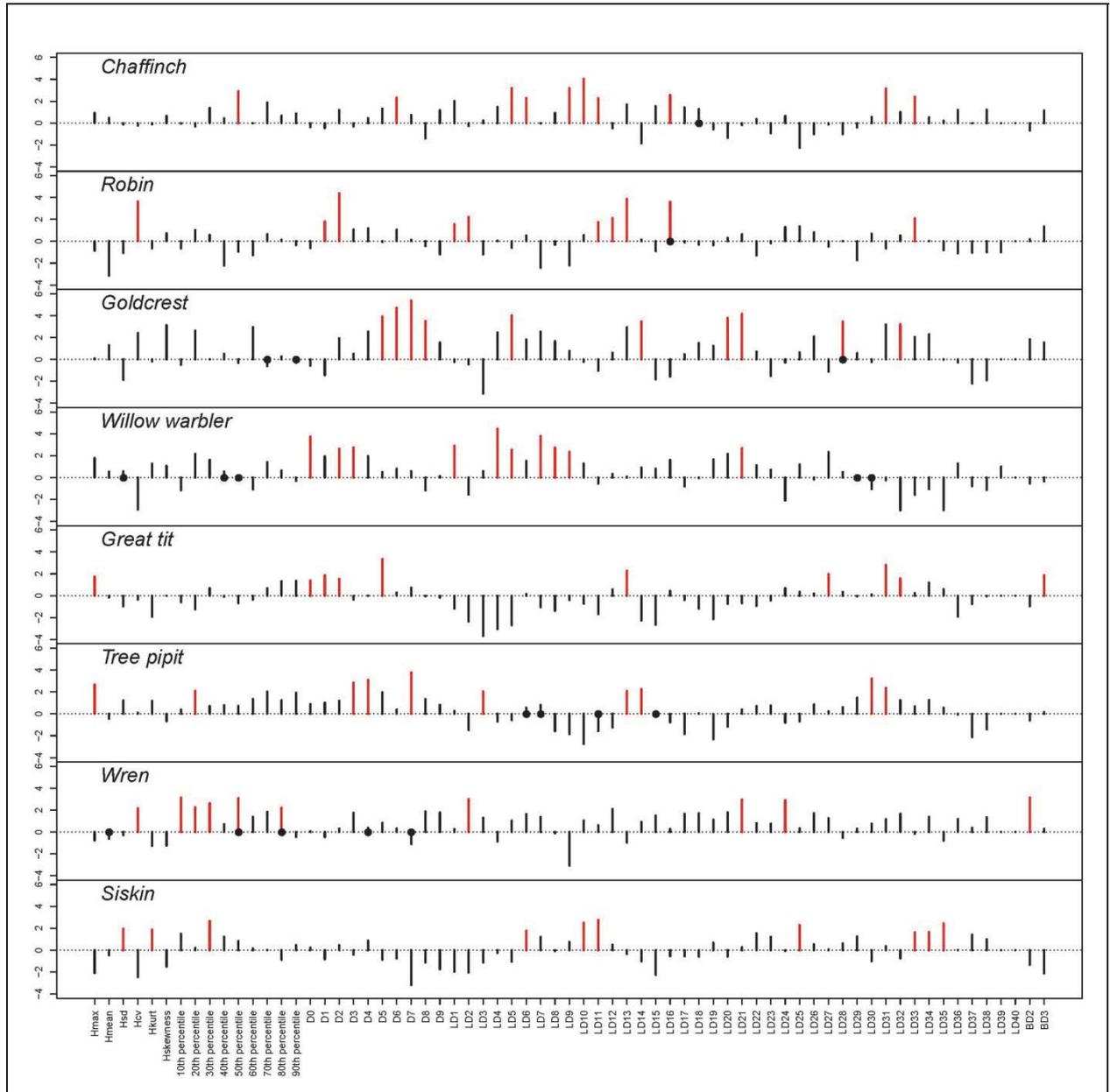


Figure 1. Importance of individual ALS variables in the Random Forest models developed for each of the eight most abundant species. Red bars indicate the 10 most important variables for each species. It has not been accounted for correlation or interactions among the most important variables for each species. The black dots represent variables which were included in the corresponding GLM10 model for each species (but note that GLM10 models were not developed for Great tit and Siskin). Variables in the H group (with names starting with a capital H) describe canopy height; variables in the percentile group describe the height distribution of the canopy; variables in the D group describe canopy density; variables in the LD group describe density layers; and variables in the SD group describe shrub density.

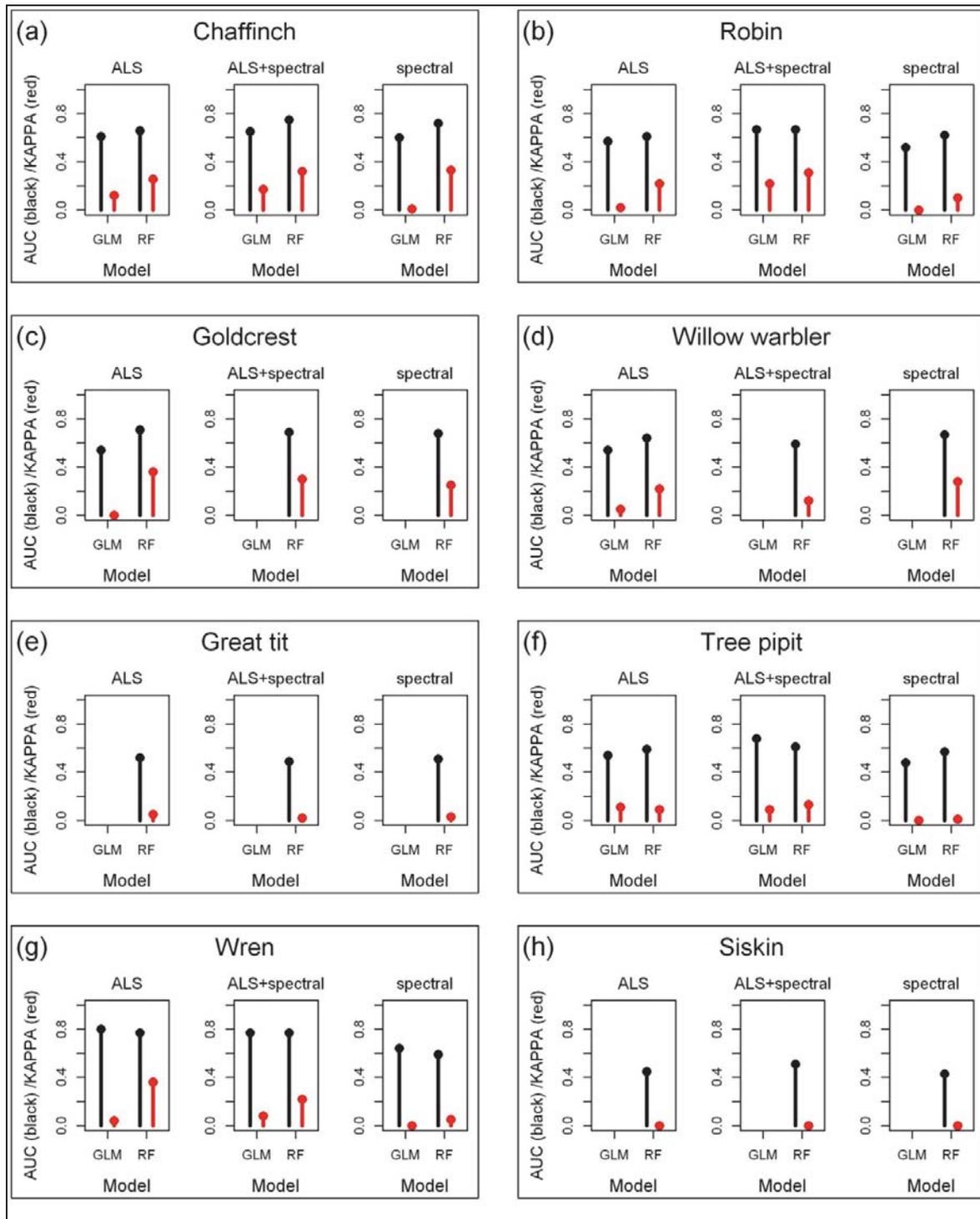


Figure 2. Evaluation of prediction models for the presence of bird species in boreal forest reserve. Models were constructed by use of three main groups of predictor variables derived from remote sensing data: (1) airborne laser scanning (ALS) variables; (2) variables derived from spectral images (spectral); and (3) ALS and spectral variables pooled (ALS+spectral). For each species and predictor-variable group, two types of prediction models were constructed: a generalized linear model (GLM, obtained by best subset selection of maximum 10 variables), and random forest (RF). Each model was evaluated by 10-fold cross-validation, with AUC and kappa as measures of predictive capability.

## Discussion

### Value of ALS-derived variables for explaining bird species richness and single-species presence

Species richness and presence of six of the eight single bird species were significantly related to ALS variables. The best explanatory model was developed for Goldcrest (27% of the variation in occurrence explained), the second-best model for Wren (21%), whereas 6–13% of the variation was explained for the other four species, and 9% for species richness. In our study, 72% of the registrations of Goldcrest were from the two strata ‘middle’ and ‘tall’ spruce forest, whereas only five registrations were from the ‘pine’ forest stratum (Table 4). The variables included in the model were the 70 and 90<sup>th</sup> height percentiles and LD28. The percentiles and LD28 indicate that presence of tall trees is important, while LD28 indicates that the crown density high above the ground is important as well. These findings fit well with previous knowledge from field-based studies, in which Goldcrest is mainly reported to prefer well-grown, more or less dense spruce forest stands (Table 3). In contrast to our results Müller et al. (2009) had low predictive power for the Goldcrest, in their study the most important ALS variable was standard deviation of mean tree height. In our study, 56% of the registrations of Wren were from ‘tall’ spruce forest. Previous field-based studies have found that Wren prefers mixed forests (Table 3), which are sparse in the study area. Our results indicate that old-growth spruce forest may also, at least under some circumstances, fulfill the requirements important for the Wren. Within the study area, areas with tall spruce trees are old-growth forests, which provide shade (a moist climate) and, typically, also dead wood. Moreover, these old trees often have living and dead branches relatively near the ground, creating cover near the ground, which is considered as another important habitat attribute for the Wren (Table 3). This is in accordance with the ALS variables included in the model: all of D5, D7 and LD9 describe the lower portion of the canopy. For the other single-species, and richness, only a small proportion of variation was explained (Table 5). Nevertheless, the ALS variables included in the models apparently accord fairly well with previous information about habitat requirements in Table 3. For Tree pipit, density of the lower canopy is important (Table 3), and LD6 and LD7 are included in the model. Robins prefer vegetation of at least medium height and not more than medium density (Table 3), and LD16 is included in the model. For the remaining species and species richness, we found no clear correspondence

between ALS variables included in the explanatory models (Table 5) and important habitat characteristics identified in field-based studies (Table 3). Although many studies have used ALS data to find important structural variables which describe the distribution of forest dwelling birds (e.g. Clawges et al. 2007, Lesak et al. 2011 Müller et al 2009, Broughton et al. 2007), only a few studies include the species analysed in our study. Among these, the study by Müller et al. (2009) included Chaffinch and found that tall and dense forests were important, as we did. Our result that neither height nor density variables were significant for Great tit contrasts results of other studies, which found that vegetation height was important for the distribution of this species (Hinsley et al. 2002, Müller et al. 2009).

Notably, all the explanatory models included one or more LD variables, i.e., variables that combine information about tree height and density. Our results thus show that the proportion of the echoes reflected from a certain vertical layer, and the density of the canopy in this particular layer, is important for the distribution of single-species.

### **Value of ALS-derived variables for predicting bird species richness and single-species presence**

The best prediction models were developed for the Wren, and these models were quite good (Figure 2). The second best prediction models were developed for Goldcrest (Figure 2). Moreover, although evaluation criteria indicated poor to fair fit according to criteria for goodness of fit given by Landis and Koch (1977) and Araújo et al (2006) [see, however, Halvorsen (2013) for a critical view of generalising goodness of fit based on these statistics], we find that useful prediction models could be constructed for Chaffinch, Robin, Willow warbler, Tree pipit and for species richness while attempts to construct prediction models for Great tit and Siskin failed. The most important variables for predicting species richness in our study were the LD variables which provide information about vegetation height and canopy density. Vegetation height was the most important variable, regardless of forest type, also in other studies in which ALS data are used to predict overall species richness of forest-dwelling birds (Clawges et al. 2007, Goetz et al. 2007, Lesak et al. 2011).

It is important to note that even relatively poor prediction models may, to some extent, assist detection of parts within a study area in which the occurrence probability of the modelled target is higher (or lower) than average. However, it is also important to note that bias in the data may make prediction models seem better than they really are. Evaluation of

prediction models by field data collected independently of the data used to build the models, is therefore important as an ultimate test of a model's predictive ability (Austin 2007, Halvorsen 2012).

### **Prediction models obtained by use of ALS-derived data compared to other models**

Only minor differences in AUC or kappa values were found between corresponding models built with ALS-derived variables, with variables derived from spectral data, and with ALS+spectral variables. However, for Chaffinch, Robin and Tree pipit models based upon a combination of ALS and spectral variables performed slightly better than pure ALS models and for Willow warbler the best model used spectral variables only. Even though Willow warbler is considered a generalist, it often prefers areas with deciduous trees, especially birches, which enhance insect availability (Table 3). The spectral image-derived variables were included in this study because they are known to be useful for determination of tree species (Hill & Thomson 2005). Our results may indicate that image-derived variables do in fact capture some of the variation in tree-species composition which is of importance for this species. The best model for Goldcrest and Wren were obtained using ALS variables only. Goetz et al. (2009) also built models for bird species richness with these three variable groups and concluded that models built with ALS were significantly better at predicting species richness than were models build from optical images alone whereas only small differences were detected between models build with ALS alone and models build with ALS and optical images pooled. A study by Müller et al. (2009), in which ALS, optical and traditional field data were compared, concluded that ALS data were more useful for modelling the potential abundance of single bird species than were field measurements and optical images. Although the differences among models obtained from the three groups of variables in our study (ALS, spectral, ALS+spectral) were not very large, prediction models developed with a combination of ALS and spectral images tended to make better predictions than models based on only one of these data sources.

### **Relative performance of different prediction modelling techniques**

Of the fourteen comparable GLM10 and RF models (for given combinations of response variable and predictor variable group), better performance based upon AUC was obtained for ten RF models and three GLM10 models, respectively. The corresponding values for

evaluation based upon kappa were 12 and one, respectively. Our results therefore indicate that RF models tend to have slightly better predictive ability than GLM models. This result accords with the result by Cutler et al. (2007) for comparison of GLM and RF models for nest sites for cavity nesting birds. It should, however, be noted that the RF method applies to classification and prediction and not to creation of explanatory models (Cutler et al. 2007). Moreover, although random forest models performed better than generalized linear models in our study, this was not invariably the case, and thus we recommend using more than one technique for prediction modelling.

### **Factors influencing the models' predictive ability**

The overall predictive ability of most prediction models obtained for bird species in this study was relatively low, as indicated by AUC values  $< 0.7$  and kappa values  $< 0.3$  (considered as poor to slight fit according to Araújo et al. (2005) and Landis and Koch (1977), respectively. Two factors which may contribute to lower the predictive ability of our models are: (1) 'Detectability noise' in response variables (bird presence) data. A typical feature of animal presence data is that single animals, populations or even entire species, fail to be observed and pass undetected in surveys (MacKenzie 2006). Lack of detection results in false absences in the dataset, i.e. that the modelled species fails to be recorded when actually present. The degree of elusiveness differs between species, and other factors can also influence the probability that a certain species is recorded, e.g., the time of the year, the time during the day that recordings are made, weather conditions, and the experience of the observer. Although we tried to counteract or minimise these factors in our study design, detectability is likely to be less than 100 % and to vary among species since variation in detectability is unavoidable. (2) The size of the counting stations was relatively large (7850 m<sup>2</sup>), at least compared to the plot size used in other ALS studies in Norway (typically ca. 200 m<sup>2</sup>; Næsset and Bjercknes 2001, Næsset 2002). Increased station size also reduces explanatory power because not all parts of a large site will be equally suitable for a species, e.g., because forest structural heterogeneity increase with increasing plot size. At least in our study area, high within counting-station variation in important habitat characteristics is expected to be an important reason for suboptimal prediction models.

Single-tree variables represent an alternative to the area-based approach for acquiring ALS-derived variables used in the present study. However, the random stratified point

transect sampling technique (Bibby 2000) used in this study, which is a widely used method for estimating bird presence, provides data which are more directly compatible with the area-based approach. It is however, possible that single-species bird studies based on data collected by use of other sampling techniques (e.g. territory mapping), may benefit from using the single-tree approach.

## **Conclusion**

Our results suggest that the utility of ALS variables may be better for predicting than for explaining bird occurrence in boreal forests dominated by natural disturbance dynamics. Prediction models developed with a combination of ALS and spectral images tended to make better predictions than models based on only one of these data sources, indicating that access to ALS data adds significantly to predictability of bird distributions in boreal forests. To our knowledge, this is the first study from unmanaged (i.e. not dominated by human disturbances) boreal forest, in an area of limited spatial extent, and more studies are needed before any general conclusions can be drawn. In particular, we recommend that future studies evaluate prediction models by use of bird data collected independently of the data used to construct the prediction models, in order to get an ultimate test of the models' predictive abilities.

## References

- Alatalo, R. V. 1982. Multidimensional foraging niche organization of foliage-gleaning birds in northern Finland. — *Ornis Scand* 13: 56-71.
- Aplin, P. 2005. Remote sensing: ecology. — *Prog Phys Geogr* 29: 104-113.
- Araújo, M. B. et al. 2005. Validation of species-climate impact models under climate change. — *Glob Change Biol* 11: 1504-1513.
- Aune, B. 1993. Temperaturnormaler : normalperiode 1961-1990. — DNMI.
- Austin, M. 2007. Species distribution models and ecological theory: A critical assessment and some possible new approaches. — *Ecol Model* 200: 1-19.
- Axelsson, P. 2000. DEM generation from laser scanner data using adaptive TIN-models. — *Int Arch Photogramm Remote Sens* 33: 110-117.
- Baltsavias, E. P. 1999. Airborne laser scanning: basic relations and formulas. — *Isprs Journal of Photogrammetry and Remote Sensing* 54: 199-214.
- Bater, C. W. et al. 2009. Estimation of standing dead tree class distributions in northwest coastal forests using lidar remote sensing. — *Can J For Res* 39: 1080-1091.
- Bellamy, P. E. et al. 2009. Willow Warbler *Phylloscopus trochilus* habitat in woods with different structure and management in southern England. — *Bird Stud* 56: 338-348.
- Bergen, K. M. et al. 2009. Remote sensing of vegetation 3-D structure for biodiversity and habitat: Review and implications for lidar and radar spaceborne missions. — *J Geophys Res Biogeosci* 114: 13.
- Bibby, C. J. 2000. Bird census techniques. — Academic Press.
- Breiman, L. 2001. Random forests. — *Machine Learning* 45: 5-32.
- Brokaw, N.V.L., Lent, R.A., 1999. Vertical structure. In: Hunter, M.L. (Ed.), *Maintaining Biodiversity in Forest Ecosystems*. Cambridge University Press, Cambridge, 373–395.
- Broughton, R. K. et al. 2006. Marsh Tit *Poecile palustris* territories in a British broad-leaved wood. — *Ibis* 148: 744-752.
- Bässler, C. et al. 2011. LiDAR as a rapid tool to predict forest habitat types in Natura 2000 networks. — *Biodivers and Conserv* 20: 465-481.
- Chen, C. et al. 2004. Using random forest to learn imbalanced data. Tech. Rep. Dept. Statistics, Univ. California Berkeley. pp. 1-12.
- Clawges, R. et al. 2008. The use of airborne lidar to assess avian species diversity, density, and occurrence in a pine/aspen forest. — *Remote Sens Environ* 112: 2064-2073.
- Cramp, S. 1988. *Handbook of the birds of Europe, the Middle East and North Africa : the birds of the Western Palearctic*. — Oxford University Press.
- Cramp, S. 1992. *Handbook of the birds of Europe, the Middle East and North Africa : the birds of the Western Palearctic*. — Oxford University Press.
- Cramp, S. 1993. *Handbook of the birds of Europe, the Middle East and North Africa : the birds of the Western Palearctic*. — Oxford University Press.
- Cramp, S. 1994. *Handbook of the birds of Europe, the Middle East and North Africa : the birds of the Western Palearctic*. — Oxford University Press.
- Cutler, D. R. et al. 2007. Random forests for classification in ecology — *Ecology* 88: 2783-2792.
- Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. — *Ecography* 29: 129-151.
- Elith, J. and Leathwick, J. R. 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annu Rev Ecol Syst Annual Reviews*, 677-697.
- Falkowski, M. J. et al. 2009. Characterizing forest succession with lidar data: An evaluation for the Inland Northwest, USA. — *Remote Sens Environ* 113: 946-956.
- Førland, E. J. 1993. Nedbørnormaler: normalperiode 1961-1990. — Instituttet.

- Goetz, S. et al. 2007. Laser remote sensing of canopy habitat heterogeneity as a predictor of bird species richness in an eastern temperate forest, USA. — *Remote Sens Environ* 108: 254-263.
- Gogol-Prokurat, M. 2011. Predicting habitat suitability for rare plants at local spatial scales using a species distribution model. — *Ecol Appl* 21: 33-47.
- Graf, R. F. et al. 2009. Habitat assessment for forest dwelling species using LiDAR remote sensing: Capercaillie in the Alps. — *Forest Ecol Manag* 257: 160-167.
- Guisan, A. and Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. — *Ecol. Model* 135: 147-186.
- Halvorsen, R. 2012. A gradient analytic perspective on distribution modelling. — *Sommerfeltia* 35: 1-165.
- Halvorsen, R. 2013. A strict maximum likelihood explanation of MaxEnt, and some implications for distribution modelling. — *Sommerfeltia* 36: 1-132.
- Hill, R. A. and Thomson, A. G. 2005. Mapping woodland species composition and structure using airborne spectral and LiDAR data. — *Int J Remote Sens* 26: 3763-3779.
- Hill, R. A. et al. 2004. Predicting habitat quality for Great Tits (*Parus major*) with airborne laser scanning data. — *Int J Remote Sens* 25: 4851-4855.
- Hill, R. A. and Broughton, R. K. 2009. Mapping the understorey of deciduous woodland from leaf-on and leaf-off airborne LiDAR data: A case study in lowland Britain. — *Isprs Journal of Photogrammetry and Remote Sensing* 64: 223-233.
- Hinsley, S. A. et al. 2002. Quantifying woodland structure and habitat quality for birds using airborne laser scanning. — *Funct Ecol* 16: 851-857.
- Hinsley, S. A. et al. 2006. The application of lidar in woodland bird ecology: Climate, canopy structure, and habitat quality. — *Photogramm Eng Rem S* 72: 1399-1406.
- Hinsley, S. A. et al. 2009. Bird species distributions across woodland canopy structure gradients. — *Community Ecol* 10: 99-110.
- Holmes, R. T. et al. 1979. Guild structure of the Hubbard Brook bird community - multivariate approach. — *Ecology* 60: 512-520.
- Horning, N. et al. 2010. Remote sensing for ecology and conservation. — Oxford University Press.
- Johnson, C. J. and Gillingham, M. P. 2008. Sensitivity of species-distribution models to error, bias, and model design: An application to resource selection functions for woodland caribou. — *Ecol Model* 213: 143-155.
- Karr, J. R. and Roth, R. R. 1971. Vegetation structure and avian diversity in several new world areas. — *Am Nat* 105: 423-435.
- Kim, Y. et al. 2009. Distinguishing between live and dead standing tree biomass on the North Rim of Grand Canyon National Park, USA using small-footprint lidar data. — *Remote Sens Environ* 113: 2499-2510.
- Koukoulas, S. and Blackburn, G. A. 2004. Quantifying the spatial properties of forest canopy gaps using LiDAR imagery and GIS. — *Int J Remote Sens* 25: 3049-3071.
- Kraus, K. and Pfeifer, N. 1998. Determination of terrain models in wooded areas with airborne laser scanner data. — *Isprs Journal of Photogrammetry and Remote Sensing* 53: 193-203.
- Landis, J., R. and Koch, G., C. 1977. The measurement of observer agreement for categorical data. — *Biometrics* 33: 159-174.
- Lefsky, M. A. et al. 2002. Lidar remote sensing for ecosystem studies. — *Bioscience* 52: 19-30.
- Lesak, A. A. et al. 2011. Modeling forest songbird species richness using LiDAR-derived measures of forest structure. — *Remote Sens Environ* 115: 2823-2835.
- Liaw, A. and Wiener, M. 2002. Classification and Regression by RandomForest. *R News* 2(3), 18-22.
- MacArthur, R. and MacArthur, J. W. 1961. On bird species-diversity. — *Ecology* 42: 594-&.

- MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L. and Hines, J. E. 2006. Occupancy estimation and modeling. Inferring patterns and dynamics of species occurrence. — Academic Press.
- Martinuzzi, S. et al. 2009. Mapping snags and understory shrubs for a LiDAR-based assessment of wildlife habitat suitability. — *Remote Sens Environ* 113: 2533-2546.
- McDermid, G. J. et al. 2005. Remote sensing for large-area habitat mapping. — *Prog Phys Geogr* 29: 449-474.
- Mikhail, E. M. et al. 2001. Introduction to modern photogrammetry. — Wiley.
- Miller, J. R. et al. 2004. Spatial extrapolation: The science of predicting ecological patterns and processes. — *Bioscience* 54: 310-320.
- Moen, A. 1999. National Atlas of Norway: Vegetation. — Norwegian Mapping Authority.
- Müller, J. et al. 2009. Using airborne laser scanning to model potential abundance and assemblages of forest passerines. — *Basic Appl Ecol* 10: 671-681.
- Müller, J. et al. 2010. Composition versus physiognomy of vegetation as predictors of bird assemblages: The role of lidar. — *Remote Sens Environ* 114: 490-495.
- Nagelkerke, N. J. D. 1991. A note on a general definition of the coefficient of determination. — *Biometrika* 78: 691-692.
- Packalèn, P. et al. 2009. A Two Stage Method to Estimate Species-specific Growing Stock. — *Photogramm Eng Rem S* 75: 1451-1460.
- Økland, R. H. et al. 2001. Vegetation-environment relationships of boreal spruce swamp forests in Østmarka Nature Reserve, SE Norway. *Sommerfeltia*, 29. p 190.
- Ørka, H. O. et al. 2010. Effects of different sensors and leaf-on and leaf-off canopy conditions on echo distributions and individual tree properties derived from airborne laser scanning. — *Remote Sens Environ* 114: 1445-1461.
- Ørka, H. O. et al. 2012. Simultaneously acquired airborne laser scanning and multispectral imagery for individual tree species identification. — *Can J For Res* 38: 125-138.
- Næsset, E. 1997a. Determination of mean tree height of forest stands using airborne laser scanner data. — *Isprs Journal of Photogrammetry and Remote Sensing* 52: 49-56.
- Næsset, E. 1997b. Estimating timber volume of forest stands using airborne laser scanner data. — *Remote Sens Environ* 61: 246-253.
- Næsset, E. and Bjercknes, K. O. 2001. Estimating tree heights and number of stems in young forest stands using airborne laser scanner data. — *Remote Sens Environ* 78: 328-340.
- Næsset, E. 2002. Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. — *Remote Sens Environ* 80: 88-99.
- Næsset, E. 2004. Accuracy of forest inventory using airborne laser scanning: Evaluating the first Nordic full-scale operational project. Taylor & Francis As, pp. 554-557.
- Næsset, E. et al. 2004. Laser scanning of forest resources: The Nordic experience. — *Scand J For Res* 19: 482-499.
- Pearce, J. and Ferrier, S. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. — *Ecol Model* 133: 225-245.
- Pearson, R. G. et al. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. — *J Biogeog* 34: 102-117.
- Reutebuch, S. E. et al. 2003. Accuracy of a high-resolution lidar terrain model under a conifer forest canopy. — *Can J For Res* 29: 527-535.
- Seavy, N. E. et al. 2009. Riparian bird response to vegetation structure: a multiscale analysis using LiDAR measurements of canopy height. — *Ecol Appl* 19: 1848-1857.
- Sigmond, E. M. O. et al. 1984. Berggrunnskart over Norge 1: 1 000 000. — *Norg geol Unders*

- Smart, L. S. et al. 2012. Three-dimensional characterization of pine forest type and red-cockaded woodpecker habitat by small-footprint, discrete-return lidar. — *Forest Ecol Manag.* 281: 100-110.
- Sutherland WJ, Newton I, Green RE (2004) *Bird ecology and conservation. A handbook of techniques.* Oxford University Press, Oxford.
- Tattoni, C. et al. 2012. Can LiDAR data improve bird habitat suitability models? — *Ecol. Model.* 245: 103-110.
- Turner, W. et al. 2003. Remote sensing for biodiversity science and conservation. — *Trends Ecol Evol* 18: 306-314.
- Vehmas, M. et al. 2009. Identification of boreal forest stands with high herbaceous plant diversity using airborne laser scanning. — *Forest Ecol Manag.* 257: 46-53.
- Venables, W. N. and Ripley, B. D. 2002. *Modern applied statistics with S.* — Springer.
- Vierling, K. T. et al. 2008. Lidar: shedding new light on habitat characterization and modeling. — *Front Ecol Environ* 6: 90-98.
- Waser, L. T. et al. 2011. Semi-automatic classification of tree species in different forest ecosystems by spectral and geometric variables derived from Airborne Digital Sensor (ADS40) and RC30 data. — *Remote Sens Environ.* 115: 76-85.
- Whittaker, R. J. et al. 2005. Conservation Biogeography: assessment and prospect. — *Divers. Distrib.* 11: 3-23.
- Willson, M. F. 1974. Avian community organization and habitat structure. — *Ecology* 55: 1017-1029.
- Wing, B. M. et al. 2012. Prediction of understory vegetation cover with airborne lidar in an interior ponderosa pine forest. — *Remote Sens Environ.* 124: 730-741.





ISBN 978-82-575-1152-4  
ISSN 1503-1667



NORWEGIAN UNIVERSITY OF LIFE SCIENCES  
NO-1432 Ås, NORWAY  
PHONE +47 64 96 50 00  
www.umb.no, e-mail: postmottak@umb.no