

# Vegetation type distribution modeling with independent data evaluation - a local-scale study from the boreal-alpine transition zone

Heidrun Asgeirsdatter Ullerud

NORWEGIAN UNIVERSITY OF LIFE SCIENCES  
Department of Ecology and Natural Resource Management  
Master Thesis 30 credits 2013





## **Preface**

This thesis is the final work of my master degree in Management of Natural Resources at the Norwegian University of Life Sciences. It has been supported financially by the Cultour-project (NFR 189977/I10) at The Norwegian Forest and Landscape Institute.

I would like to thank my supervisors; Anders Bryn, Kari Klanderud and Lars Østbye Hemsing for guidance and support throughout the process of planning and writing this thesis. I would especially like to thank Anders for suggesting the topic of this thesis and for being so enthusiastic about mapping and modeling that it has spread to me.

Many other people also deserve thanks; Michael Angeloff for teaching me FYSAK, Odd Braaten and Hanne Gro Wallin for technical support, Frauke Hofmeister for helping with GIS, Rune Halvorsen for showing me how to evaluate the Maxent predictions with independent data, Johnny Hofsten for lending me his cabin in the study area, as well as Morten Ekeberg and Nils Karbø (CTO) at Blom Geomatics AS for preparing the CIR-aerial photos free of charge. The staff at the Department of Ecology and Natural Resource Management has been helpful whenever I have been stuck in ArcMap or R. I am grateful for all your help!

Writing this thesis has been a process of learning. For every bit I understood, I realized there was more to know. It has been interesting to follow the research published on distribution modeling and Maxent, since it is a topic that is continuously in change. I wanted to write a thesis that would be interesting for, or useful to, someone. I think I succeeded, and hope others will agree!

Norwegian University of Life Sciences

Ås 10<sup>th</sup> May 2013

---

Heidrun Asgeirsdatter Ullerud



## Summary

This study investigated the performance of distribution modeling (DM) for vegetation types. Two frame-areas were mapped. One area was used to train models in Maxent, a recommended method for DM. The other area was used for model projection and evaluation by independent data. Models were created for six vegetation types, two from each of the ecosystems present in the area; forest, wetland and mountain. For each ecosystem one locally common and one locally rare vegetation type was modeled. AUC was used as the model selection criteria. Environmental variables were selected through a backwards selection scheme, where variables contributing by less than 0.005 to the AUC-value were excluded. Model complexity in Maxent was limited by allowing only three transformations; linear, quadratic and threshold, and setting the regularization multiplier to eight.

The results showed that modeling of vegetation types and projecting the models locally to a neighboring area was possible. However, the resulting models varied greatly in predictive performance between the vegetation types, as well as in number of environmental variables included and the number of parameters in the final models. With the AUC-values from training, models for rare types were found to have better predictive performance than models for common types, and a significant negative relationship was found between the number of points used to train the model and the AUC-value.

The models' predictions were evaluated with independent data. The resulting AUC-values were found to be a better representation of the predictive performance than the training AUC-values, since the training AUC-values seemed to be affected by the characteristics of the training data. After evaluation the AUC-values portrayed less variation in model predictive performance. All six vegetation types had models that were characterized as good or excellent regardless of differences in occurrence, ecosystem, variation in environmental variables, number of points used to train the model and the complexity of the models.



## Content

Preface.....	I
Summary .....	III
Content.....	V
1 Introduction.....	1
1.1 Vegetation mapping .....	1
1.2 Distribution modeling.....	3
2 Materials.....	5
2.1 Study Area .....	5
2.2 Environmental variables.....	7
3 Methods .....	8
3.1 Vegetation mapping .....	8
3.2 Distribution modeling.....	9
4 Results .....	12
5 Discussion .....	16
5.1 High performance in distribution modeling of vegetation types.....	16
5.2 Variation within the vegetation types.....	17
5.3 The effect of ecosystem .....	19
5.4 Model selection criteria.....	19
5.5 Evaluation with independent data .....	20
5.6 Spatial challenges .....	21
5.7 The effect of model complexity in Maxent .....	22
5.8 Limitations posed by proxies and non-existing map layers .....	24
6 Conclusion .....	27
7 References.....	28
Appendices .....	31





# 1 Introduction

## 1.1 Vegetation mapping

There is an increasing need for reliable land cover information in nature management. Land cover and land-use maps that capture important ecological aspects of the nature, such as vegetation maps, are also important for the documentation and monitoring of changes in nature. Vegetation mapping has a long tradition in the European countries (Biondi 2011), and is based on recognition of predefined types more or less related to plant communities (Bryn 2006). Vegetation types for detailed mapping are often defined using combinations of common plant species and indicator species (Fremstad 1997), whereas survey mapping systems lean more on the vegetation physiognomy and structure that can be detected using aerial photos (Gudjonsson 2010; Ihse 2007). The ultimate goal of vegetation mapping is to capture ecological variation at a given scale, in order to provide information for different purposes such as nature management.

Vegetation maps are recognized as one of the best existing maps for portraying location of natural resources and conditions (Rekdal & Bryn 2010). Many sectors depend on information present in, or derived from, maps of vegetation types. Farmers need information on land resources in order to find potential and carrying capacity for agriculture and foraging (Gudjonsson 2010). The nature management needs information based on the vegetation in order to find important areas for species and selection of management schemes. Also, the municipalities and private land owners need land cover information for planning of infrastructure, areas for settlement and outdoor areas for recreation. Vegetation maps are valuable for assessing land-use changes and analyzing impacts (Gudjonsson 2010). Vegetation maps can also be used to extract proxies for many environmental variables, such as water availability, species richness, snow cover and soil nutrients.

The vegetation types and the present distribution of them within a landscape, is a result of the realized ecological response. All historical and recent biotic-, abiotic- and human-factors and interactions have influenced the given distribution (Halvorsen 2012a). Such a summary of the generally valid ecological relationships in geographical space has been used as a variable in many modeling studies (Franklin 2009), and is a good base for a projection of a model in time or space (Halvorsen 2012a). However, the low extent of coverage of vegetation maps is a major obstacle for this approach, especially in Norway, where only approximately ten percent of the land area has been mapped (Rekdal & Bryn 2010). To have a higher coverage of maps from all parts of Norway would be very useful both for nature management and modeling.

The first map from Norway that included all the present vegetation types within the study area, was published in 1937 (Mork & Heiberg 1937). At that time the mapping was based mainly on phytosociological systems (Biondi 2011). After some decades with more focus on ecology and landscape (Biondi 2011), vegetation mapping was again in focus in the seventies. Phytosociology was still the basis for the mapping systems that were developed. The Norwegian Forest and Landscape Institute (TNFLI) developed their own mapping system and became one of the main actors in the field of vegetation mapping. They were given the national responsibility for vegetation mapping in Norway in the eighties (NOU 1983).

The systems for mapping were gradually developed alongside the focus of ecological research and most Nordic systems are now mainly based on ecological gradient perspectives (Bryn 2006). The scale of the mapping will affect which gradients that influence the vegetation in an area. On a regional scale the vegetation sector and zones are important (Bakkestuen et al. 2008), whilst at the local scale factors such as slope, exposure, soil moisture, snow cover and nutrient availability in the substrate will influence the vegetation. Often, also human influence in an area will be of importance for understanding the variation in nature (Bryn & Hemsing 2012).

Today there are two main scales for mapping of vegetation in use in Norway, Sweden, Iceland and Finland. One scale is used for a number of survey mapping systems (e.g. Andersson 2010; Gudjonsson 2010), in Norway mainly handled by TNFLI and used for mapping at scales 1:25 000 - 50 000 (Rekdal & Bryn 2010). Standard procedure for mapping with this system is use of aerial photos in field. The other scale of mapping is used for a number of detailed mapping systems (Fremstad 1997; Pålsson 1998). Fremstad (1997) has developed a more detailed system for Norway that is used for mapping at scales 1:5000 - 25 000. The systems are linked hierarchically so that all units from the detailed level can be located within units at a survey level. Mapping of nature types rather than mapping based mainly on vegetation, has gradually received more attention (Halvorsen et al. 2009). Nature types are useful as they can describe nature also where vegetation is absent, such as in cold water coral reefs. The EU Habitat Directive (1992) demands mapping of nature types, and although Norway has not signed the directive, a project for mapping of nature types has been started (Halvorsen et al. 2009). However, this system for mapping of nature types is not yet commonly used by mappers.

The areas in Norway where vegetation has been mapped are mainly located in the mountainous regions in south-central Norway, Nordland and in inland parts of Finnmark (*Vegetasjonskart - dekning*). Vegetation mapping can be done at a pace of 3 – 5 km<sup>2</sup> per day using the most general

system, the survey system of TNFLI. Mapping the entire land area of Norway is approximated to cost 1 billion NOK, even if this survey system is used (Strand & Rekdal 2010).

The Norwegian area frame survey of land cover and outfield land resources (AR18x18) is an attempt to provide more information about land resources (Strand 2013). In the AR18x18 system a grid with a spacing of 18 km has been laid over Norway, and in every grid corner vegetation has been mapped in a plot of 0.9 km<sup>2</sup> (Strand & Rekdal 2010; Strand 2013). The AR18x18 system will provide reporting of land cover statistics for Norway (Strand & Rekdal 2010), but does not increase the cover of vegetation maps very much. However, the AR18x18 system could be used to speed up the process of mapping vegetation, for example by using the mapped plots to train models for vegetation types and use these to generate vegetation maps for the surrounding areas (extrapolation). These maps will still need to be validated with field-data, but if such modeling is found to be possible, the speed of generating vegetation maps might be greatly increased.

## **1.2 Distribution modeling**

Distribution modeling (DM) could be a possible method for increasing the progress of vegetation mapping. In DM, observations of a defined target (e.g. species, vegetation type) are combined with digital maps of relevant ecological variables in order to create a model that predicts probability of presence for the target in a defined area. The output of DM is a map representation of model predictions as well as measurements of performance for the model.

The pioneers in DM performed the studies manually using graphs and calculating gradients (Whittaker 1960). The introduction of Geographical Information Systems (GIS) in the eighties and the following rapid improvements and innovations in GIS has increased the access to digital environmental variables, as well as the amount of information that can be processed in DM (Franklin 2009). This has given DM a great boost and quickly turned it into a separate field of ecology (Halvorsen 2012a).

There is a great variety of approaches and targets used in DM, and the terminology and classification of the modeling varies as much as the range of applications (Ferrier et al. 2002; Franklin 2009).

Modeling can aim at explaining ecological relationships in nature, or in different ways try to locate species or habitats (Halvorsen 2012a). There are many different types of DM-methods, including regression methods such as generalized additive models and generalized linear models, envelope-style and distance-based methods, as well as newer, less established methods for DM such as machine learning methods (Elith et al. 2006). Some methods use presence-only data, others include absences. The methods for evaluation of the models also differ greatly. Several of the methods have

proved to be quite successful in predicting the occurrence of the modeled target (Elith et al. 2006). Indifferent to how the modeling is classified, the purpose of the study and the characteristics of the modeled target are important issues when deciding methods and settings (Halvorsen 2012a).

Species distribution has been the most common modeling target for DM (Bekkby et al. 2002; Edvardsen et al. 2011; Marino et al. 2011; Parolo et al. 2008; Wollan et al. 2008). However, it has also been used for totally different topics, for example modeling of the potential for expansion of forest following land-use change in Norway (Bryn et al. 2013). Distribution modeling has also been applied to model land-cover types (Dobrowski et al. 2008) and different species assemblages, such as vegetation types (Cawsey et al. 2002; Ferrier et al. 2002; Hemsing & Bryn 2012; Weber 2011).

Vegetation types have different characteristics and the ease of modeling will vary with how well the distribution of the vegetation types is explained by the included environmental variables, but also on how strictly the vegetation types have been defined. Common types might cover a larger range of environmental variables, making specific criteria for the distribution hard to find. Rare vegetation types are often strongly correlated with a more narrow range on several ecoclines (Halvorsen 2012a), and this might make them easier to model.

The purpose of this study is to explore the ability of predicting the vegetation type in an area through modeling. Models will be trained with vegetation samples from one area before they are used to project results into a comparable neighboring geographical area. The projections will be evaluated with independent data to find the predictive performance of the models outside the area of training.

*This study will examine the certainty by which survey vegetation types can be projected locally by use of presence-only distribution modeling. Secondly it will investigate whether locally rare vegetation types are easier to model than common, and thirdly find the effect of variance in environmental variables on modeling ability. The fourth objective of the study is to investigate whether vegetation types from some ecosystems are easier to model than others. Additional problems to be addressed are to identify the most important environmental variables when modeling vegetation types and to identify environmental variables that should be provided as digital wall-to-wall maps in order to improve results and progress of DM.*

## 2 Materials

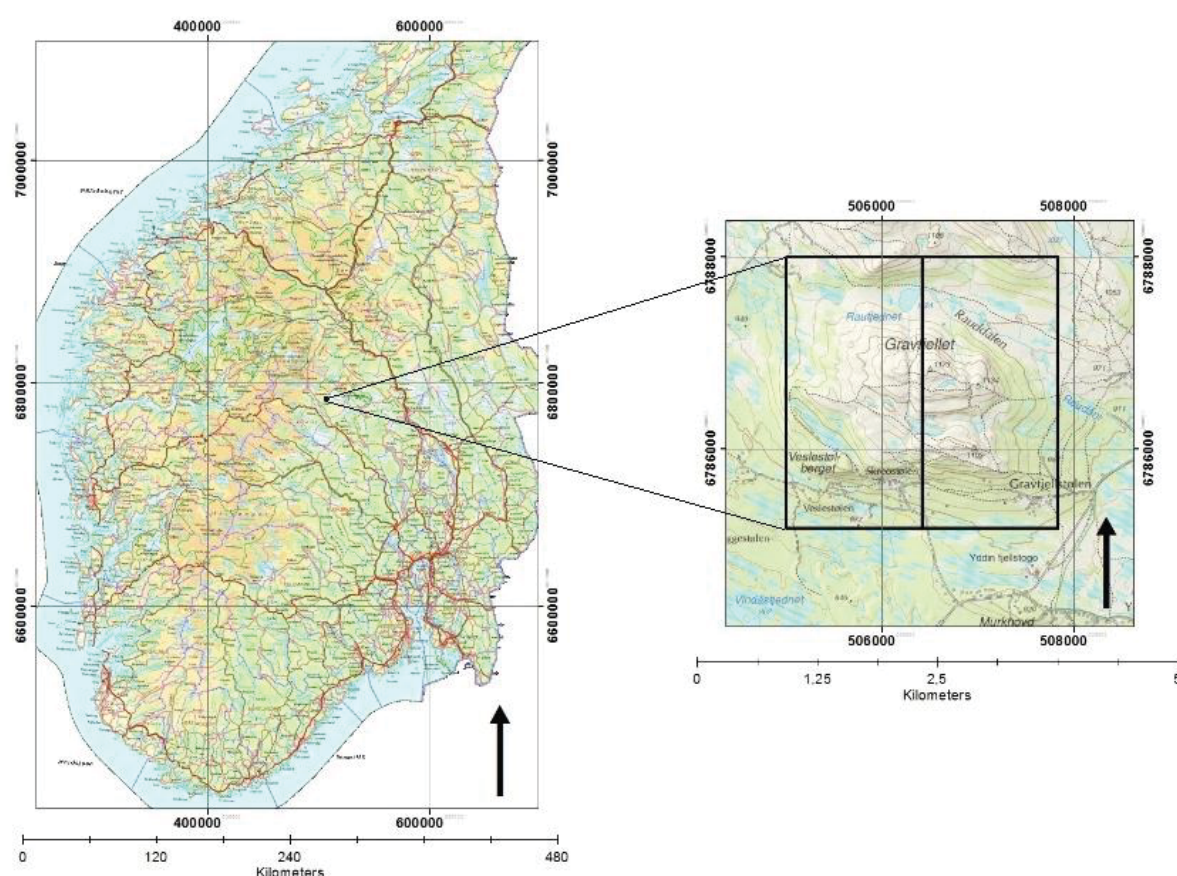
### 2.1 Study Area

#### 2.1.1 Physical location of the area

The study area is located at Gravfjellet, south-east of the village Beitostølen, in the middle of Øystre Slidre municipality, north-east in the district of Oppland, in south-central Norway (Table 1 and Figure 1). The study area consists of two rectangular frame-areas. One frame-area was used for training the model, while the other was used for projection and evaluation by independent data.

**Table 1** Area, position and altitude of the frame-areas that make up the study area.

Frame-area	Area (km <sup>2</sup> )	Centre coordinates (WGS84/UTM32N)	Altitudinal range (m a.s.l.)
Training	4.0	6786585N/505708E	849-1169
Projection	4.0	6786585N/507123E	881-1173



**Figure 1** Maps showing the position of the study area in south-central Norway, as well as the detailed location of the study area with its two neighboring frame-areas. The left frame-area was used for training, the right for evaluation by independent data. Map projection WGS 84/UTM 32N. Maps from geoNorge (*Topografisk rasterkart 2WMS*).

### 2.1.2 Nature and climate

The vegetation within the study area spans from the northern boreal to the low alpine zone (Moen 1999). The area is in the transition vegetation section (OC – Oceanic Continental), where eastern plants are most common, but weakly western plants can be found (Moen 1999). Zone and section are influenced locally by topographic variation, and the scale makes topographic variation a more important factor for determining the vegetation in this study (Bakkestuen et al. 2008; Moen 1999). The mean annual temperature at the closest weather station, Løken in Volbu, is -0.8 °C, and there is an average of 590 mm precipitation per year (Table 1).

North boreal spruce forests, rich in nutrients, are found mainly in the south-facing, low-lying areas to the south of Gravfjellet. The valley north of Gravfjellet is dominated by different constellations of wetlands. Mountain birch forests dominate east and west of Gravfjellet, with elements of wetlands in moist areas. Mountain vegetation types dominate in the higher elevated areas, with little soil or vegetation on the highest mountain tops. In poorly drained areas, wetlands are found almost all the way to the top of the mountain, since biomass production is still high enough for the creation of peat at this elevation.

**Table 2 Monthly mean temperatures (°C) and precipitation (mm) for the meteorological station that is closest to the study area, Løken in Volbu (521 m a.s.l.). Temperature is the monthly normal from 1961 to 1990, while precipitation is the monthly normal from 2003 to 2012. Distance to study area is approximately 10 km. Meteorological data from eKlima (eKlima).**

	Jan	Feb	March	April	May	June	July	Aug	Sept	Oct	Nov	Des
<b>Temperature</b>	-9.9	-8.4	-4.1	0.8	6.8	11.7	13.1	11.8	7.1	2.7	-4.1	-8.4
<b>Precipitation</b>	43	27	32	24	44	64	74	70	59	64	52	37

### 2.1.3 Cultural influence

The study area is located in a region that for many centuries has been extensively utilized for traditional summer dairy farming (Bryn & Daugstad 2001). Animals have grazed within the area, fodder-plants have been utilized for haymaking and firewood has been collected. This has had large impact on the vegetation in the region (Hemsing & Bryn 2012). Regrowth of trees has effectively been hindered by domestic grazing, giving more open areas and a lower forest line than what is given by nature (Axelsen 1975). This tradition of using all available resources in the outlands has been in retreat the last 50 years (Almås et al. 2004; Ihse 2007; MacDonald et al. 2000), causing semi-natural landscapes to be overgrown by young, reappearing forests (Bryn et al. 2013; Ihse 2007).

There used to be three summer dairy farms within the study areas, all located in southern parts of the frame-areas (Axelsen 1975). Two of these were abandoned before 1975 (Axelsen 1975), and currently there is no summer dairy farm that receives economic support in the area (*Beitelag - seter*).



However, some in-fields in the south of the study area are still in use for production of grass, six fields are located in the frame-area for training and three in the frame-area for projection. Cattle are grazing in some of these fields. Sheep also graze freely in the area, but there are not enough animals to avoid regrowth. A number of new holiday cottages have been built within the study area, mostly located in the southern parts. However, ignoring the area immediately surrounding the cottages, these do not hinder regrowth. The main changes in cultural influence have not happened recently, and effects of regrowth have been seen for more than 40 years (Axelsen 1975). The vegetation in the study area has not yet reached its climax; regrowth is slowly changing the vegetation and landscape in the region (Hemsing & Bryn 2012).

## 2.2 Environmental variables

Nine environmental variables were used for training the models and for projection of the models to the neighboring frame-area for independent evaluation (Table 3). The modeling resolution was 5 x 5 m for all variables. No relevant digital map layers of soil nutrients or soil moisture existed, so the closest proxies available were used (Table 3).

**Table 3 Environmental variables used in the study, where they originated from, how they were created and the environmental processes they were assumed to be proxies for. All the environmental variables were on a continuous scale.**

<b>Environmental variable</b>	<b>Generated from</b>	<b>Original resolution</b>	<b>Transformation to 5 x 5 m resolution</b>	<b>Proxy for</b>
<b>DEM (m)</b>	LiDAR*	1 x 1 m	Aggregate	Temperature, topographic variation
<b>Vegetation height (m)</b>	LiDAR*	1 x 1 m	Aggregate	Vegetation height
<b>Slope (°)</b>	DEM, Slope-function in ArcMap 10.1	1 x 1 m	Aggregate	Soil moisture, soil characteristics
<b>Curvature (0.01 m)</b>	DEM, Curvature-function in ArcMap 10.1	50 x 50 m**	Aggregate and resample	Exposure to wind, water and soil nutrient runoff direction
<b>IR-photos in 3 bands (RGB-color values)</b>	Color infrared (CIR) aerial photos***	0.3 x 0.3 m	Resample	Land cover variation, vegetation productivity
<b>Intensity (watts)</b>	LiDAR*	1 x 1 m	Aggregate	Ground/vegetation cover
<b>Wetness index</b>	Single flow algorithm on a digital terrain model****	25 x 25 m	Resample	Groundwater

\*The LiDAR scanning was done in 2007/08 and had an average of 0.7 points/m<sup>2</sup>. Frauke Hofmeister at TNFLI processed the LiDAR data.

\*\*When creating the curvature layer, the DEM was first aggregated to 50 x 50 m resolution, and then resampled back to 5x5m after creating the layer. The curvature function in ArcMap only considers 9 raster squares at a time. The aggregation was done in order to portray the curvature in all areas, also those that are weakly curving, in a way that optimizes modeling results.

\*\*\* CIR-pictures provided by Blom Geomatics AS, the details for the photos are the same as for the aerial photos used for mapping (Appendix 4).

\*\*\*\* The wetness index was created by Eva Solbjørg Flo Heggem at TNFLI.

The CIR aerial photos were split into three separate bands, each band representing the color value of red, green or blue for each raster cell. If joined as one layer, all the variation present in the photo would not be communicated.

## **3 Methods**

### **3.1 Vegetation mapping**

Vegetation in both frame-areas, making up the study area, was mapped by Heidrun A. Ullerud in the beginning of August 2012. Both frame-areas were mapped using the standard of TNFLI (Rekdal & Larson 2005). This standard has 54 categories of vegetation, of which 44 are vegetation types and 10 are other land cover categories. Additional signs were used to identify other types of variation in nature. Mosaics were used when two vegetation types occurred interchangeably together in an area and they both covered more than 25 percent of the area. According to the guidelines, vegetation types covering areas smaller than 0.01 km<sup>2</sup> should not be included in the mapping unless they were part of a mosaic. In order to portray more of the variation, smaller polygons were allowed in the mapping, the smallest with an area of 0.002 km<sup>2</sup>. Of the mapped polygons, 15 percent cover less than 0.01 km<sup>2</sup>. The largest polygon was 0.4 km<sup>2</sup>.

The vegetation polygons were drawn in field, using a portable lens stereoscope and dual color aerial photos from September 2010 (Appendix 4), printed in a scale of 1:25 000. Aerial photos were used to help in the delineation of polygons and to aid in the interpretation of the distribution of the vegetation types. However, the registration of the vegetation type was done by direct field-inspection. In this project the vegetation was mapped by a person with little experience in mapping vegetation. Supervision during the field-work by two independent, experienced mappers was provided to ensure quality of mapping. The resulting map was also compared with a map created by an experienced mapper to eliminate mistakes.

The aerial photos with manual registering of vegetation types were scanned, orthorectified and digitalized. The digitalization of the maps was done in FYSAK (Version E20). After digitalization the polygon borders were corrected by comparing the digitized map with high resolution orthophotos.



Six vegetation types were chosen for detailed studies through a predefined study design. There were three ecosystems present within the study area; wetland, forest and mountain. Two vegetation types were selected from each ecosystem; one widespread and one rare within the study area (Table 4). All the six vegetation types for detailed studies were present in both frame-areas; the frame-area for training and frame-area for projection.

**Table 4 Name and code of the six modeled vegetation types, the ecosystem they belong to, as well as the occurrence of each within the study area. The proportion of the study area covered by each vegetation type is also included.**

Vegetation type	Ecosystem	Occurrence	Proportion of study area (%)
Lichen heath	Mountain	Rare	4
Dwarf shrub heath	Mountain	Common	27
Blueberry birch forest	Forest	Common	24
Meadow spruce forest	Forest	Rare	4
Fen	Wetland	Common	9
Mud-bottom fens and bogs	Wetland	Rare	2

### 3.2 Distribution modeling

Several methods are available for DM. Maximum entropy modeling, a method for presence-only DM, was used in this study. It is often described as a machine learning method (Phillips et al. 2006), but can also be explained as a maximum likelihood method (Halvorsen 2012b). The software Maxent (version 3.3.3k, October 2011) was used for creating the maximum entropy models. Given presence-only records of a specific target and environmental variables for the study area, Maxent creates a model for the distribution of the target with parameters based on the value of the environmental variables in the presence-cells (Elith et al. 2011; Phillips & Dudik 2008). This model is used to predict areas where the target might be present.

Using ArcMap 10.1, sets of presence-only records were generated from the vegetation map of the frame-area for training, one set for each vegetation type (Table 5). Polygons that fulfilled the criteria for each specific vegetation type were the base for the generation of points, and in these polygons presence-points were laid in a grid with a square size of 10 m. Points were also generated in mosaic polygons where the vegetation type to be modeled was the main vegetation type in the mosaic, as well as in some of the polygons with additional signs (column 3 in Table 5). Types with mosaics or additional signs originating from a different ecosystem than the vegetation type targeted for modeling, for example trees in wetlands, were not used to create presence-points (Table 5). Polygons with bare ground were not used to create presence-points for Lichen heath, while polygons with high

lichen-cover were used. Polygons with high lichen-cover were also used to create presences in Dwarf shrub heath, but polygons with trees, *Salix* spp, or bare ground were not. In Blueberry birch forest, polygons with Norway spruce (*Picea abies*) were used to create presences, while polygons with sparse forest were not, as they would give a wrong representation of the vegetation height for forests in general. In Meadow spruce forest, polygons with deciduous trees were used to create presences as the types are not fundamentally different and this was also a necessity in order to give large enough areas for the modeling. For Fen and Mud-bottom fens and bogs only mosaics with other wetland types were used to create presences in addition to the specific vegetation type itself.

**Table 5** The number of training points generated in each vegetation type, as well as which mosaics and additional signs that were, and were not, used when creating the set of presence-points for each type. The same rules for use of mosaics and additional signs were applied when generating evaluation points in the frame-area for projection. Codes are explained in Appendix 1 and 2.

Vegetation type	No. of points	Mosaics and signs used when creating presences	Excluded mosaics and signs
Lichen heath	1286	2c, 2cv, 2cx	2cv<, 2cv>, 2cx>
Dwarf shrub heath	11614	2e, 2ev, 2ex	2es, 2e&, 2e}, 2e*
Blueberry birch forest	9728	4b, 4b*, 4b/c	4b], 4b]>
Meadow spruce forest	2122	7c, 7c&	None
Fen	2708	9c, 9c/a, 9c/b, 9c/d	9c&, 9c/2e&, 9c/3bs, 9c/4c, 9c/8d&
Mud-bottom fens and bogs	1170	9d, 9d/a	None

Maxent models, one for each vegetation type, were run with the created set of presences and the environmental variables for the frame-area for training. Model selection was done by integrating internal model assessment methods in a backwards stepwise selection process (Halvorsen 2012a). In this study the main purpose was to find a model that best predicts the spatial distribution of different vegetation types when projected to a neighboring geographical area. The area under the curve (AUC) of the receiver operating characteristic (ROC) (Phillips et al. 2006), given by Maxent, was used for distinguishing between the models relative predictive ability (Franklin 2009; Halvorsen 2012a; Pearce & Ferrier 2000). The ROC curve is obtained by plotting the species' true positive rate on the y-axis, and the false positive rate on the x-axis for all possible thresholds (Phillips et al. 2006). Maxent compensates for lack of absences by characterizing a random sample of background cells as (pseudo-)absences (Elith et al. 2011). The training AUC can be described as a value of the model's ability to differentiate between presences and (pseudo-)absences, and give presences a higher relative

predicted probability of presence (Halvorsen 2012a; Stokland et al. 2011). Data resubstitution (Jackknifing) was used to find variables that did not contribute to the model (Halvorsen 2012a). Explanatory variables that contributed less than 0.005 to the AUC-value were excluded, following Stokland et al. (2011).

The same settings in Maxent were used for all six vegetation types. To avoid model overfitting, the regularization multiplier was set to eight. This value was chosen based on experiences made when modeling with this dataset. The raw output format was used, as it represents the Maxent exponential model without any transformation (Phillips et al. 2005). Maxent creates model parameters based on the environmental variables and transformations of these. Three types of transformations were permitted; linear, threshold and quadratic. Linear features are the continuous environmental variables without any transformations, while threshold allow fitting of more arbitrary functions based on the environmental variables (Phillips & Dudik 2008). Threshold features gives the value one if the variable is above, and zero if the variable is below a given threshold (Phillips et al. 2006). The quadratic transformation is the square of the linear environmental variable (Phillips & Dudik 2008). These transformations were selected in order to reduce the number of parameters created for each model, while still allowing enough complexity to model the actual responses represented in nature. Product- and hinge transformations, other functions that are available in Maxent, were not activated.

Maxent uses the models created in the frame-area for training and the environmental variables for the frame-area for projection to generate map representations of model predictions for the frame-area for projection (extrapolation). The Maxent prediction results based on presence-only data are given as relative predicted probability of presence (RPPP), and in order to translate these to predicted probability of presence (PPP), the predictions were evaluated with independent data. The independent data were generated from the vegetation map by spreading random points in the frame-area for projection, and attaching information on presence and absence of vegetation types to each point. The same rules for additional signs and mosaics used when generating presence-points for training were also applied when assigning presence or absence to the random points in the evaluation (Table 5). The evaluation was done twice, first with 2000, then with 4000 random points, in order to increase the robustness in the results.

The PPP AUC-value is a measure for the predicted probability of presence. Models with AUC-values below 0.5 are no better than a random model, and models with a value of 1.0 predicts presences only where actual presences are found (Fielding & Bell 1997). Following Araújo et al (2005), a five-grade scale was used to classify model results by categorizing the AUC-values (Table 6).

**Table 6 Five-grade scale used for classifying model results based on the AUC-value (Araujo et al. 2005)**

Category/model value	Fail	Poor	Fair	Good	Excellent
AUC-values	< 0.6	0.6 - 0.7	0.7 - 0.8	0.8 - 0.9	> 0.9

The values of the environmental variables of 1170 training-points from each vegetation type were used to analyze the internal variation in the vegetation types. ANOVA, with the environmental variable as the dependent variable and vegetation type as the independent variable, was used for all nine environmental variables to test if the variance was similar across the six vegetation types. A Tukey Post Hoc test was used to show which specific vegetation types that did not differ significantly in variation of the environmental variables. A numerical measure of the variation in the environmental variables (MVEV) for each vegetation type was created by ranking the standard deviations (SDs) within each environmental variable, and calculating the average ranking value of the environmental variables that were included in the final model. The smallest possible value of MVEV was one, and the largest six.

The AUC-values were tested to search for significant differences in the predictive value of the models based on MVEV, ecosystem or occurrence. The Shapiro-Wilk normality test indicated normality for the different types of AUC-values and the MVEV, but not for the number of presence-points for training, nor the number of parameters in the models. Ecosystem and occurrence were tested using ANOVA. Linear regressions were used to investigate relationships between the AUC-values of the models and MVEV, as well as relationships between the AUC-values and the number of points used to train the models. Linear regressions were also used to examine if there were relationships between the AUC-values and the number of parameters in the final models, and between the number of parameters in the final models and the number of points used for training the models. A significance level of 0.05 was applied when interpreting the results. All statistics was done using RStudio Version 0.97.318.

## 4 Results

AUC-values of the modeled vegetation types varied from 0.905 to 0.634 after training (Table 7). With one exception, AUC-values for all vegetation types increased after evaluation with independent data. The PPP AUC-values varied from 0.973 to 0.819 (Table 7). The maximum increase was 0.232 for Dwarf shrub heath, changing the model result from poor to good. Also the model for Blueberry birch forest changed from a poor to a good model after evaluation with independent data. The AUC-value for the model for Meadow spruce forest increased from good to excellent, while the AUC-value for

the model for Mud-bottom fens and bogs reduced from excellent to good after evaluation with independent data.

The final models differed in which and how many environmental variables were used, with a maximum of six, and a minimum of two (Table 7). Also, the contributions of the environmental variables to the models varied for each vegetation type. DEM was the only variable included in all models, and it was the most important variable in five of the models (Table 7). Vegetation height was included in three models. Curvature was included in two models, and was the variable that contributed the most to the model for Fen. The number of parameters in the final models also varied greatly (Table 7).

**Table 7 AUC-values from model training and after projection with 2000 and 4000 random points as well as the number of parameters in the final models (obtained from the lambda-file created by Maxent). The models are categorized based on the PPP AUC-values. The table also shows how many and which environmental variables were used in the final models, and the percent contribution of each.**

	<b>Lichen heath</b>	<b>Dwarf shrub heath</b>	<b>Blueberry birch forest</b>	<b>Meadow spruce forest</b>	<b>Fen</b>	<b>Mud- bottom fens and bogs</b>
<b>AUC – training</b>	0.905	0.634	0.667	0.857	0.788	0.903
<b>AUC proj. 2000</b>	0.970	0.857	0.819	0.957	0.874	0.860
<b>AUC proj. 4000</b>	0.973	0.866	0.833	0.961	0.885	0.839
<b>Model category</b>	Excellent	Good	Good	Excellent	Good	Good
<b># parameters</b>	6	13	39	5	15	7
<b>Percent contribution of environmental variables</b>						
DEM	90	55.1	50.3	98	8.7	48.9
Vegheight		44.9	49.7		21.8	
Slope					22.7	41.9
Curvature					31.5	9.2
IR – Blue	4.4				6	
IR – Green					9.4	
IR – Red	5.6					
Intensity				2		
Wetness						
<b>Total</b>	100	100	100	100	100	100

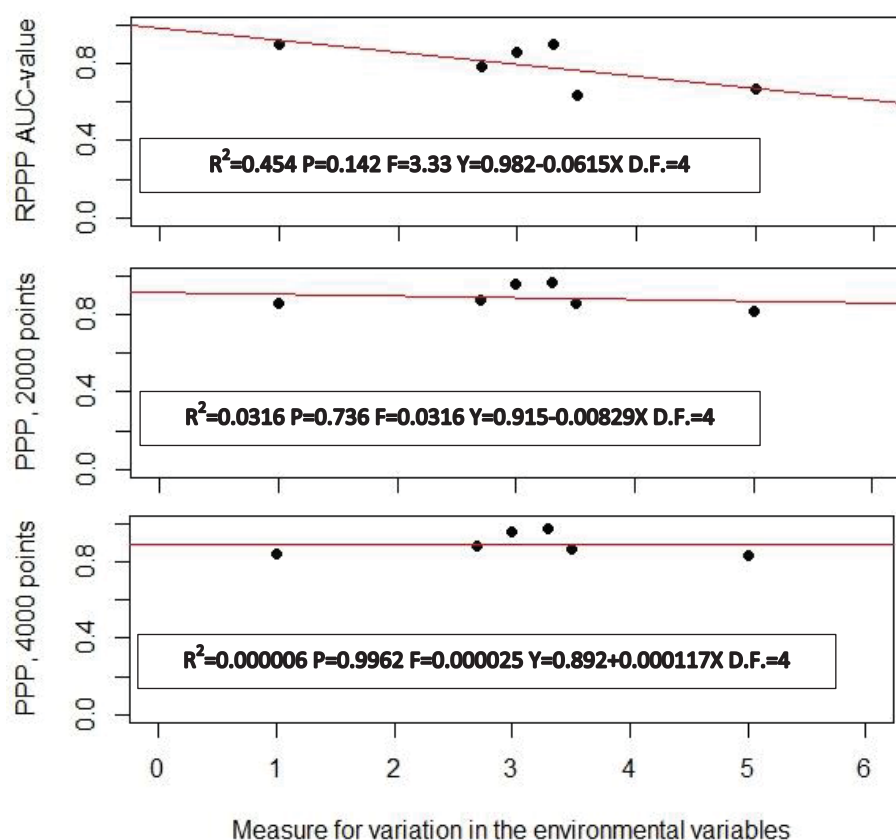
The ANOVAs gave p-values smaller than 2e-16 for all vegetation types and showed that the environmental variation differed across vegetation types. The Tukey Post Hoc test showed that only some variables for a few vegetation types did not differ significantly (Table 8). The MVEV was smallest for the wetland types, while the largest MVEV was found for Blueberry birch forest (Table 8).

**Table 8** Mean and standard deviation (SD) of each environmental variable for all vegetation types. R is the ranking of the SD for each environmental variable, where 1 is the smallest and 6 is the largest. Vegetation types where the range of the environmental variables did not differ significantly (Tukey Post Hoc test) are marked with bold italics. Environmental variables with grey shading are the ones that were used in the final model. The measure of variation in environmental variables (MVEV) is the average ranking value of the variables included in the final Maxent model. A small value shows that the vegetation type had small SD's compares with the other models SD's.

	Lichen heath			Dwarf shrub heath			Blueberry birch forest			Meadow spruce forest			Fen			Mud-bottom fens and bogs		
	Mean	SD	R	Mean	SD	R	Mean	SD	R	Mean	SD	R	Mean	SD	R	Mean	SD	R
DEM (m)	1121	24	3	1078	36	4	1011	38	5	889	20	2	996	70	6	1034	17	1
Vegetation height (m)	<b>0.0</b>	<b>0.0</b>	<b>1</b>	<b>0.0</b>	<b>0.1</b>	<b>3</b>	<b>0.9</b>	<b>0.7</b>	<b>5</b>	1.1	0.8	6	0.1	0.3	4	<b>0.0</b>	<b>0.1</b>	<b>2</b>
Slope (°)	<b>12</b>	5	3	<b>12</b>	<b>6</b>	<b>4</b>	14	8	5	17	9	6	7	5	2	3	3	1
Curvature (0.01m)	0.3	0.4	6	0.1	0.3	4	0.1	0.4	5	-0.1	0.3	3	-0.2	0.2	2	-0.2	0.2	1
Blue (RGB-colorvalue)	144	34	5	138	35	6	134	34	4	112	33	3	137	20	1	129	20	2
Green (RGB-colorvalue)	<b>140</b>	<b>32</b>	<b>4</b>	134	32	5	130	30	3	105	38	6	147	23	1	<b>140</b>	<b>23</b>	<b>2</b>
Red (RGB-colorvalue)	<b>185</b>	<b>24</b>	<b>2</b>	192	25	4	173	27	5	169	33	6	<b>193</b>	<b>21</b>	<b>1</b>	178	25	3
Intensity (watts)	<b>109</b>	<b>26</b>	<b>2</b>	111	27	3	<b>71</b>	<b>26</b>	<b>1</b>	<b>68</b>	<b>29</b>	<b>4</b>	<b>106</b>	<b>33</b>	<b>6</b>	81	32	5
Wetness	-212	115	1	-146	158	2	-130	165	3	-52	189	4	0	227	5	29	254	6
Sum of range ranking for included param.	10			7			10			6			16			3		
MVEV	3.3			3.5			5.0			3.0			2.7			1.0		



None of the AUC-values were significantly related to the MVEV (Figure 2). Although not significant, the RPPP AUC-values seemed to decrease with increasing MVEV (Figure 2).



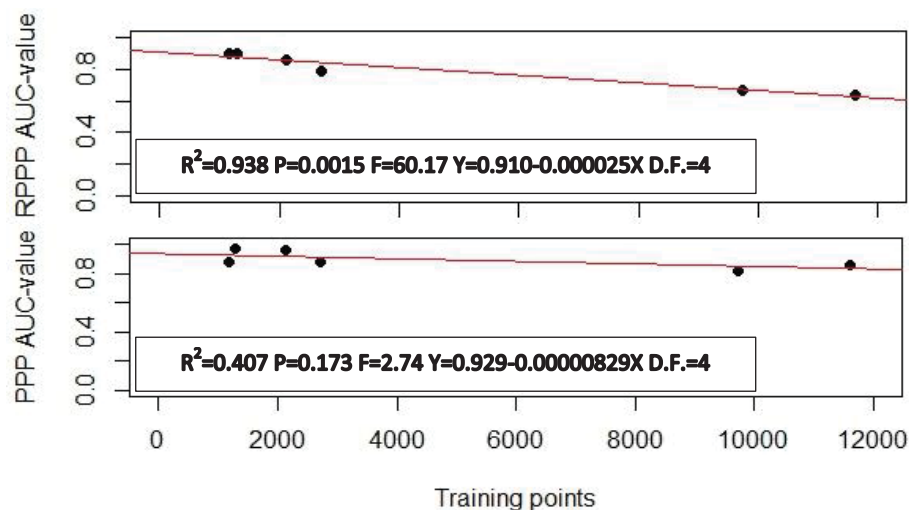
**Figure 2** Linear regression results for the relationship between AUC-values and the measure of variation in environmental variables (MVEV). The measure of variation in environmental variables (MVEV) is the average ranking value of the variables included in the final Maxent model (Table 8). A small value shows that the vegetation type had small SD's compares with the other models SD's.

The relationship between AUC and occurrence was significant for the RPPP-values, but not significant for the PPP-values (Table 9). There was no significant relationship between the AUC-values and ecosystem (Table 9).

**Table 9** ANOVA-results for relations between ecosystem (mountain, wetland and forest), occurrence (common and rare) and AUC-values for the vegetation types.

Independent variable	Training		2000pts		4000pts	
	F-value	P-value	F-value	P-value	F-value	P-value
<b>Occurrence</b>	15.12	0.018	4.249	0.108	1.924	0.238
<b>Ecosystem</b>	0.209	0.823	0.203	0.826	0.336	0.738

There was a significant relationship between the RPPP AUC-values and the number of points used to train the model (Figure 3). This relationship was no longer significant after converting AUC-values from RPPP to PPP (Figure 3).



**Figure 3** Linear regression results for the relationship between RPPP AUC-values (upper graph) as well as PPP AUC-values (lower graph) and number of training points. Only the PPP-values generated with 2000 points is shown, but the regression for the 4000 points PPP AUC-values showed similar results.

The AUC-values had no significant relation to the number of parameters in the final models, neither the RPPP-, nor the PPP-values. There was also no significant relationship between the number of points used to train the models and the number of parameters (Table 10).

**Table 10** Linear regressions results between the AUC-values from training (RPPP), both projected values (PPP) and the number of parameters in final models (Table 7), as well as result of regression between the number of parameters in the final models and the number of points used to train the models (Table 5).

Dependent variable	Independent variable	Intercept	$\beta_1$	$R^2$	F	DF	P-value
RPPP AUC	Parameters	0.88	-0.006	0.45	3.24	4	0.146
PPP AUC – 2000	Parameters	0.94	-0.003	0.51	4.13	4	0.112
PPP AUC - 4000	Parameters	0.93	-0.003	0.36	2.24	4	0.209
Parameters	Points	1626	225	0.39	2.51	4	0.188

## 5 Discussion

### 5.1 High performance in distribution modeling of vegetation types

The results of this study showed that vegetation types can be modeled and projected into an ecologically comparable neighboring area by use of presence-only DM. The evaluation of the projection with independent data showed that the projection was successful; all models for the six different vegetation types were described as good or excellent following the guidelines of Araujo et al. (2005). The results indicated that there is a potential for training models and subsequently use them as a basis for vegetation maps in new areas close to the training area. This is especially interesting as mapped plots of the AR18x18 system exist throughout Norway. The result of this study suggests a potential for using these plots to train models that can indicate the vegetation types in an



area surrounding each plot. However, the mapped plots in the AR18x18 system are smaller than the frame-area used for training in this study, and the area between the plots is larger than the frame-area for projection in this study (Strand 2013). This gives less data for training using AR18x18 data (compared with the projection area), which also subsequently leads to larger degree of extrapolation. Whether the entire area between the plots will be possible to model in a good way has so far not been tested, but is a possible task for further studies.

The ANOVA between RPPP AUC-values and occurrence showed that models for locally rare vegetation types have better predictive capacity than models for common types. This result is supported by many DM studies for different targets (Elith et al. 2011; Lobo et al. 2008; Merckx et al. 2011; Stokland et al. 2011). However, the ANOVA results were not significant for any of the PPP AUC-values. The relationship found between RPPP AUC-values and occurrence was probably an effect of the negative relationship between RPPP AUC-values and the number of points used to train the model. Evaluation with independent data makes the AUC-values more independent of the characteristics of the training-points, showing that evaluation is necessary in DMs using AUC as the main model selection criteria (Edvardsen et al. 2011).

The DEM was clearly the most resourceful environmental variable when modeling vegetation types in this area. The vegetation type to be modeled and its characteristics decided which other variables were used. Four of the six vegetation types in this study were modeled using only variables derived from LiDAR data. If DM is to be used in a large scale for Norway, LiDAR data with a resolution of at least 5 x 5 m for the entire Norway would represent a great advantage. Presently however, LiDAR is only available for a few regions in Norway, and this could hinder the progress of DM.

This study adds to the knowledge of DM of targets originally located in polygons. This is useful also outside the field of nature/vegetation types mapping, for example for DM of animals with a large home range (Franklin 2009). The study also improves the knowledge of evaluation with independent data, and supports the findings by Lobo et al. (2008) that vegetation types with a low RPPP AUC-value, can give good projection results.

## **5.2 Variation within the vegetation types**

The survey vegetation types were not very homogenous, compared with the types in more detailed mapping systems, as they were mapped at a fairly coarse scale with few types. Some areas had vegetation types with elements from several other types, and no rules could be set that covered them all. Exclusion of some mosaics and non-typical types was in this study seen as a necessity in order to train the model to recognize typical types, and this probably increased the resulting models' predictive performances.

There is no common practice for calculating the internal variation of the modeled target, so there are no values available for comparison with the values for variation found in this study. The MVEV quantified only the variation in the variables included in the final models, and was probably also affected by which mosaics and non-typical vegetation types were included in the model training data. The sequence obtained by the MVEV coincides with experience from field, and the variation was also in accordance with the description of the vegetation types provided by the field-guide (Rekdal & Larson 2005). However, the magnitude of variation was not as expected. The MVEV for Dwarf shrub heath was expected to be as high as for Blueberry birch forest. Vegetation height was one of two variables used to model these types. The variation in Dwarf shrub heath could be artificially low since all types with *Salix* spp and trees were excluded from Dwarf shrub heath. For Lichen heath the variation was larger than expected. This could be because two out of three explanatory variables were IR-bands and although areas with bare rock were excluded, the vegetation type still includes areas with and without high lichen cover, and the color values will differ.

The vegetation types that cover large areas and are common often have the same characteristics as generalist species; they can tolerate a larger range of environmental conditions. For generalist species Maxent does not necessarily manage to discriminate between the values of the environmental variable that the types tolerate and what they do not tolerate, making them harder to model (Lobo et al. 2008; Merckx et al. 2011; Stokland et al. 2011), and the same results were expected for vegetation types with a large MVEV. The results of this study did not show any significant relationship between AUC-values and MVEV. However, this result was opposed by a significant relationship between the RPPP AUC-values and number of points used to train the model, and that the mean RPPP AUC for rare vegetation types was significantly higher than the mean RPPP AUC for common types. The area covered and the occurrence can be seen as proxies for the internal environmental variation, since the internal variation is likely to increase with a larger distribution. These relationships were not significant for the PPP-values, thus confirming that environmental variation was not what determined the models prediction performance for vegetation types in this study.

Although the relation between AUC and MVEV was not significant, the two vegetation types with the highest MVEV obtained the lowest AUC-values, and their RPPP-value was categorized as poor. Lobo et al. (2008) found that high internal variation caused by tolerance for a large variety of environmental variables can give a low AUC-value, although the model predictions are valid. This was also the case in this study; the types with the largest MVEV were the models with the largest increase in AUC-values when evaluated with independent data.

### **5.3 The effect of ecosystem**

The results of this study show that type of ecosystem does not affect the modeling ability, neither in PPP or RPPP results. Thus, there were no particular properties of any of the ecosystems that made them harder to model than the others, and probably the distribution of the vegetation types in the three different ecosystems was determined largely by the same environmental variables. Since the vegetation types within the same ecosystem have some characteristics in common, vegetation types from the same ecosystem were expected to have similar predictive value. The PPP AUC-values for the vegetation types from forest ecosystems differed by at least 0.107, and the AUC-values for vegetation types from mountain ecosystems by 0.138. To my knowledge, no previous study has investigated the effect of differing ecosystems on predictive performance of vegetation type DM, but Roy et al. (2006) created a land cover map for India using a Holdridge life zone model. They found differing accuracies in the land cover map for different forest types. This supports my results, the differences in predictive value of vegetation types within an ecosystem can be as large as the differences between vegetation types belonging to different ecosystems.

The models for vegetation types from wetland ecosystems had similar predictive performance. This probably reflects that the two selected wetland types were ecologically similar, and that their distribution was regulated by the same environmental variables. However, the small difference between these types is not necessarily a sign that all wetland vegetation types can be modeled with the same predictive performance. Other wetland types have characteristics that differ from the two types modeled in this study, for example Bog can in some cases be more similar to Dwarf shrub heath than to Fen and Mud-bottom fens and bogs (Rekdal & Larson 2005). A model for Bog would be expected to have a lower predictive performance than the wetland types had in this study.

### **5.4 Model selection criteria**

Maximum entropy modeling was chosen as the modeling method in this study, since it has been found to be among the better methods for DM (Elith et al. 2006; Ortega-Huerta & Peterson 2008). The Maxent software was chosen since it performs well with presence-only samples (Elith et al. 2011), as well as being freeware and user-friendly. Maxent also provides several methods for internal model assessment such as the AUC-value (Halvorsen 2012b). AUC-value is one of the most common ways of reporting predictive performance (Elith et al. 2006). Researchers disagree on how high an AUC-value must be before the model is good, maybe due to a lack of theoretical foundation (Halvorsen 2012b). The AUC-scales are also dependent on the modeling purpose and data properties, and cannot necessarily be adopted by other studies (Swets 1988). However, most studies agree that models with AUC-values of above 0.8, as all PPP-values in this study, do have predictive value (Halvorsen 2012b; Merckx et al. 2011).

Using the AUC-value as the only model-selection criterion has been criticized (Lobo et al. 2008). The study in this thesis avoids being subject of this criticism in at least two ways; the extent of the study was limited (Lobo et al. 2008) and many presence-points were used to train the model. Though many points were used for all vegetation types, the number of training-points still varied, and this might affect the AUC-value. The default number of pseudo-absences was used for all models, and though the number varied between the models, it did not hinder a lower prevalence rate for the rare vegetation types. The prevalence is the proportion of presences in relation to the total number of points used to train the model (Franklin 2009). The number of absences was not reduced to balance the presences for rare vegetation types, as Stokland et al. (2011) found that keeping a low prevalence-rate for rare species gave more accurate predictions, with a more natural span in probability of presence.

With presence-only data, the maximum AUC-value that can be obtained is lower than one, and decreasing with the area covered by the species (Elith et al. 2011). A model with many presence-points needs to have better predictive value in order to obtain the same AUC-value as a model with fewer points (Stokland et al. 2011). Merckx et al. (2011) have found that the AUC-value decreases with the number of training points for a random model if no validation is applied. This study supports these findings, as there was a statistically significant relationship between the RPPP AUC-value and the number of points used to train the model. Vegetation types with many points, especially Blueberry birch forest and Dwarf shrub heath, cover large areas and the models representing them had the lowest AUC-values. Also, the vegetation types with the fewest points, Mud-bottom fens and bogs and Lichen heath had the highest AUC-values. This is a challenge when using the AUC-value as a model selection criterion.

## **5.5 Evaluation with independent data**

Both the frame-area for training and the frame-area for projection were mapped following the same method and with the same material, making the results comparable with respect to accuracy and resolution. The frame-areas were mapped by the same mapper in the same period of field-work, ensuring similar understanding of the vegetation, and the same decision criteria being used for non-typical types in both frame-areas. As the mapping was based on the vegetation only, the data were still independent. There might have been small elements of other vegetation types than the targeted vegetation type in the polygons used to create presence-points, and the converse in the polygons used to create absence-points. However, the effect of this was assumed to be small, so the quality of the collected data was considered to be presence-absence. Using half of the data as presence-only was done by ignoring the absence-observations. This allowed for investigation of the technical problems related to RPPP AUC-values as well as comparisons between RPPP and PPP AUC-values.

The PPP AUC-values showed that the predictions of the created models were valid also outside the frame-area where the models were trained. The AUC-values for the frame-area for projection did not show the same relationship with the number of training points as the RPPP AUC-values. This supports the theory that the PPP-value is a better measure for the quality of the model than the RPPP-values (Merckx et al. 2011). The criticisms of RPPP AUC-values as the only model selection criteria, are less likely to apply to PPP AUC-values. However, the number and distribution of the points might still affect the AUC-values for the projected area.

A very small number of points and skewed prevalence rate can affect the AUC-value directly if the prevalence rate is below 0.01 or above 0.99 (Jimenez-Valverde et al. 2009). This relationship was found for the training data, but could also be valid for the data-sets used in evaluation. In this study there were few presences relative to absences in the sets used for evaluation. For Mud-bottom fens and bogs the number of presences gives prevalence immediately above and below 0.01 for 2000 and 4000 random points respectively. An effect could be that this result is less robust than the others, and that the decrease observed for this vegetation type when changing from RPPP to PPP AUC-values should not be emphasized.

Jimenez-Valverde et al. (2009) also found that modeling results for models trained with less than 70 points of either presences or absences have lower predictive value. For the evaluation based on 2000 random points, three of the six vegetation types had less than 70 presence-points, while this was the case for only one type when evaluated with 4000 random points. The difference in projected AUC-values created with 2000 and 4000 points was small, yet the trend was clear. For all vegetation types except Mud-bottom fens and bogs, the PPP AUC-value created with 4000 points was higher than the PPP AUC-value created with 2000 points. The values from the 4000-point evaluation might have better predictive value due to larger sample sizes, however the increase in AUC-values could be interpreted as a sign of spatial auto-correlation, making the values artificially high (Halvorsen 2012a). This study does not provide enough information to conclude on which evaluation is best, and this could be a topic for further studies.

## **5.6 Spatial challenges**

A common bias of point distribution is disproportionately many presence observations in easily accessible areas (Halvorsen 2012a; Reddy & Davalos 2003). By creating training points from the polygons of the vegetation map this type of bias was avoided. Each polygon was represented with many points to limit the influence of elements of non-typical vegetation types. A distance of ten meters between the training-points gave enough data to describe the relation to the environmental variables for each type, and portrayed the variation at a fine scale. Due to within-type variation in

most of the vegetation types, the points were not necessarily auto-correlated, and this was supported by the evaluation with independent data. However, polygons for Mud-bottom fens and bogs had small internal variation. The observed reduction in AUC-value for Mud-bottom fens and bogs after evaluation, could be due to auto-correlation in the training dataset causing the RPPP AUC-value to be artificially inflated (Merckx et al. 2011). Future studies modeling vegetation types with differing variances should consider accommodating this by using different densities of training points.

The frame-area for training and the frame-area for projection were in this study located nearby each other, so the results are spatially correlated. This correlation was intended, as the model will work best for areas close to where the model was trained. Barry and Elith (2006) argue that if auto-correlation is included in a model, many samples should be used in order to ensure correct specification of the model, while Halvorsen (2012a) argue that several points from the same polygon will cause validation by non-independent data. Though using fewer points for evaluation than for training in this study, there were still several points from each polygon. Including several points in the same polygon was a way of testing if Maxent managed to predict well for all environmental variables displayed within the same polygon. The projected AUC-values would probably have been lower if the projected area was situated further away, since the vegetation and the factors regulating it are likely to differ more the further from the training area your projection area is situated.

The resolution of the environmental variables might also have affected the results. Guisan et al. (2007) found that a coarser resolution reduced the predictive performance of created DMs. The LiDAR screening provided enough points in each 5 x 5 m square for the resulting environmental variable layers to be robust, while still portraying the variation at a fine scale. The high contribution of the variables derived from LiDAR supports that this was the right resolution for modeling at this scale. The wetness index originally had a resolution of 25 x 25 m, and was created from a DEM with this resolution. This might have been too coarse for modeling at the scale of this study, and could be a reason why the wetness index did not contribute to any of the models.

## **5.7 The effect of model complexity in Maxent**

For modeling with the intention of explaining ecological relationships, the models should be simple in order to be able to understand the output. In DM for projection, the model that best explains the distribution is the best model, and more complexity can be allowed (Halvorsen 2012a). However, complex models are more likely to be overfit, making predictions for other areas less accurate (Merckx et al. 2011). In this study the goal was to predict distribution for a neighboring area, so there was a tradeoff between predictive-value and complexity (Halvorsen 2012a, Fig. 15).

Using default settings have become the standard procedure for Maxent modeling, and recent research has indicated, and this study supports, that these settings gives overfit models (Halvorsen 2012b). Maxent is by many considered an objective method for modeling. This is opposed by Hemsing and Bryn (2012, Table 6) due to the large effects caused by subjectively choosing different settings. Experience with the vegetation type datasets in this study showed that the settings used in the final models hindered the models from becoming unnecessarily complex, while still not restricting them so much that the predictive performance was greatly reduced. Excluding the hinge transformation reduced the number of parameters with between 3 and 34 parameters. This caused a maximum reduction in the AUC-training values of 0.002. Excluding the product transformation removed a maximum of four variables, and did not change the AUC-training value for any model. These two types of transformations were also excluded because they are the hardest to relate to ecological responses, the hinge is an arbitrary function, while product creates interaction variables.

The regularization multiplier was increased to eight to avoid overfitting. This made the response curves smoother and limited the saw-toothed fitting to the values of the environmental variables. This was especially useful where the training-data had few observations for a specific range without any logical environmental reason. The increase in regularization multiplier caused a reduction of between 69 and 18 parameters in the different models, and a maximum decrease in the AUC-training values of 0.011. There was still potential for reducing the number of parameters further by increasing the regularization multiplier even more. However, this would have caused more decrease in the AUC-training value that had already had the largest loss.

The same transformations were used for all models and the effects of selected settings differ between the vegetation types based on factors that determine their distributions. For the final models in this study, the minimum number of parameters was five, and the most complex model had 39 parameters. The number of parameters needed for a model to be considered complex or overfitted is not much investigated. A model of 73 parameters was considered to be overfitted in a study by Auestad (2013). How the model behaves when evaluated with independent data is probably the clue to finding out if a model is overfit or not (Halvorsen 2012b). In this study five of six AUC-values increased when evaluating with independent data. This is a sign that the models were not overfit (Merckx et al. 2011). In future, Maxent might be improved by implementing an internal model assessment criterion that penalizes models with many parameters in the way that for example Akaike Information Criteria (AIC) does.



## 5.8 Limitations posed by proxies and non-existing map layers

The number of available digital variables and maps is increasing rapidly (Franklin 2009), however finding variables that influence the targets' distribution at the right resolution is still a challenge in many DM studies (Franklin 2009). Large-scale DM often includes climatic variables, while on a small scale the proxy-approach is more common (Franklin 2009). Soil nutrients and water availability, as well as heat and radiation, are known to be baseline explanatory variables for vegetation distribution (Guisan & Zimmermann 2000). If these variables could have been included in the study, they might have improved the models, although this is not much studied (Franklin 2009). Newton-Cross et al. (2007) compared models based on digital datasets and variables collected in field, and found that the models based on digital datasets were as good as or better than the other models. This is supported by the results of this study as the models created were based on digital datasets only, and they all had good predictive value.

As in the study by Stokland et al. (2011), the variables' percent contributions varied greatly for different targets. The contributions partly correspond with the ecology and characteristics of the vegetation types as they are described in the survey mapping guidelines (Rekdal & Larson 2005). The variance of the individual environmental variables seems to be an important factor for how much the different variables contribute to the models, and if they were included in the final model. When the amount of variation is a characteristic of the vegetation type, as expected for Blueberry birch forest and Dwarf shrub heath, the variation in vegetation should have been expressed as an environmental variable so that Maxent could use it in the models.

Topographic variables have been found to be highly useful in many modeling studies (Guisan & Zimmermann 2000), and this was supported by the models created in this study. Two main gradients affect the pattern of vegetation in Norway; the first is temperature from north to south and from lowland to highland, the second is precipitation (Bakkestuen et al. 2008). The altitude provided by the DEM was expected to be a proxy for temperature, and it had a fairly high percentage contribution in all models. The high contribution of the DEM in this study could be a result of temperature being one of the main gradients affecting vegetation. The DEM was also one of the environmental variables with the greatest differences in both mean and SD. The SD was largest for Fen, and this was also the model where DEM contributed the least. This makes ecological sense since, as long as the biomass production does not drop too much, the distribution of wetlands is not decided by the altitude. However, in the model for Mud-bottom fens and bogs, the percent contribution of the DEM was high. This was probably because the polygons of this vegetation type in the frame-area for training were found in a very limited range of elevations, creating a model that focused on random variables that were important only in this specific data-set. Maxent did not



manage to find the variables that determine distribution on a general basis. This was supported by the reduction seen in AUC-value for Mud-bottom fens and bogs when projecting the model.

Vegetation height, a variable provided by LiDAR, was expected to separate forests from the non-forest ecosystems in the distribution models. The mean of the vegetation height variable across all training points was 0.89 m for Blueberry birch forest and 1.14 m for Meadow spruce forest, while the forest by definition within the vegetation mapping guidelines in this study is above 2.5 m (Rekdal & Larson 2005). The quality of the data was not good enough to have the intended explanatory effect. The time of year of the shooting might have affected the quality, due to reduced foliage early and late in the season. It could also be because the forests were sparse, causing the LiDAR beams not to hit the top of the trees. The mean value was used when aggregating the vegetation height variable to the resolution used in this study. Probably the maximum value would have given more representative results, especially for forests. In spite of the faults of the vegetation height variable it contributed to three of the models. This could be because it provides a way of separating otherwise similar types. Blueberry birch forest and Dwarf shrub heath are often bordering types. Vegetation height is the main difference between them, and was an important variable in both models.

Curvature contributed to the models, but only for the wetland types. This was probably because curvature represents a proxy for landscape formations where water drainage is mirrored across the contour intervals, giving wet areas, rich in nutrients, suitable for fens in landscape depressions. Curvature was also expected to represent a proxy for the gradient of wind exposure. As Lichen heath often forms the dominating vegetation type on wind exposed ridges where snow is absent and the soils are dry during summer time (Fremstad 1997), curvature was expected to have a large percent contribution to the model for Lichen heath. When curvature does not contribute to this model, it could be a sign that the variable as used in this study was a better representation for concave than convex land forms, or that the resolution is too coarse for representing the narrow ridges where Lichen heath is found. Another possible reason for curvature not to be included in the model for Lichen heath could be that there was a large variation in this variable. The SD's in curvature for the two wetland types were small in comparison.

Slope was expected to be a proxy for soil moisture and soil characteristics. This would make it an important variable for vegetation types that demand soils rich in nutrients, such as Meadow spruce forest. Slope contributed but, like curvature, only to the two wetland type-models. This indicates that slope only worked as a topographic variable. Fens occur in flat to gently sloping areas, whereas Mud-bottom fens and bogs are found in flat areas (Rekdal & Larson 2005). This is confirmed by the low means and small SDs in the slope variable for wetlands. For Meadow spruce forest the slope variable

had, as expected since it is found at the bottom of slopes, a high mean. However, the variable also had a high SD, making the influence less clear and probably also more difficult to base a model on.

The IR-environmental variables only contributed to the models for Lichen heath and Fen, and they did not contribute by very much in either of the models. The reason for this could be technical errors in splitting the photos into 3 bands, or there being a smaller color variation in the photos than expected. What time of the year the photos were taken affects how the vegetation is portrayed (Ihse 2007). The IR-photos used in this study were from late September. Summer is optimal, and photos from late September are likely to have low interpretability both for deciduous trees and general vegetation mapping (Ihse 2007). This could be an important reason for why the IR-layers did not contribute more to the final models.

LiDAR intensity contributed in only one model, by two percent in the model for Meadow spruce forest. The poor results from intensity could be due to small differences in feedback from different vegetation types. The wetness index did not contribute to any of the models, although it was expected to be an important variable for the wetland types. The variable had large variations for the wetland types and this could be an explanation of why it did not contribute. The wetness index could also be correlated to the curvature variable as they showed similar trends, though the values were opposite; negative values for dry areas in the wetness index and positive values for dry areas in curvature. If correlated, including both would cause disturbance in the data. The curvature variable had lower variation, and this could be a reason it was chosen instead of the wetness index. A possible cause of the large variation is that the values of the wetness index were multiplied by 100 by the creator in order to avoid used of float-values in a large raster. This should have been reversed before using the index for modeling, as it might affect the results. Lastly, the wetness index might not have been made in a way that makes it a suitable proxy for groundwater availability in modeling of vegetation types.

## 6 Conclusion

This study has shown that distribution modeling of survey vegetation types based on presence-only data is possible. The models can be projected locally, and the model predictions are good for all the six vegetation types in this study. The results of this study confirm that the training AUC-value is negatively related to the number of points used to train the model, making evaluation by independent data necessary for the AUC-value to be a valid measure of the models predictive value.

Occurrence was not found to affect the prediction performance of the models, and there was no relationship between the AUC-values and the variation in the environmental variables. The measure for variation in environmental variables used in this study was based on the variables included in the final models. High variance in an environmental variable for a vegetation type seemed to make the variable less likely to contribute to the final model, and this could explain why the measure of variance in the environmental variables was not found to affect the predictive performance of the models. The modeling ability was not found to be affected by what ecosystem the vegetation type belongs to.

The most important environmental variable when modeling vegetation was found to be the DEM. Other valuable variables depended on the vegetation type to be modeled. Four of the six vegetation types in this study were modeled using only environmental variables generated from LiDAR-data. This study should be repeated in other areas of Norway and in different vegetation zones to examine if the same factors determine vegetation elsewhere. LiDAR-data should be made available as wall-to-wall maps in order to facilitate increased use of modeling. Further studies should be conducted to find how far the models for survey vegetation types can be projected and still be valid, as well as to find what other vegetation types can be modeled.

## 7 References

- Almås, R., Gjerdåker, B., Lunden, K., Myhre, B. & Øye, I. (2004). *Norwegian agricultural history*. Trondheim: Tapir Academic Press. 365 p.
- Andersson, L. (2010). Geographical vegetation data of Lantmäteriet in Sweden. In Bryn, A., Dramstad, W. & Fjellstad, W. (eds) *Mapping and monitoring of nordic vegetation and landscapes, Hveragerði, Iceland, September 16-18, 2009*, p. 9-12: Viten fra Skog og Landskap 1/10.
- Araujo, M. B., Pearson, R. G., Thuiller, W. & Erhard, M. (2005). Validation of species-climate impact models under climate change. *Global Change Biology*, 11 (9): 1504-1513.
- Auestad, I. (2013, March 13-15). *Invasive species in road verges - distribution modelling as a tool for early detection*. Biodiversity and ecosystem services in a changing environment: Challenges for science and conservation, Trondheim: Norsk Økologisk Forening.
- Axelsen, B. J. (1975). *Ressursutnyttelse i et fjellområde. Seterfjellet i Øystre Slidre - Utnyttelse av og påvirkning på naturgrunnlaget*. Hovedoppgave. Oslo: Universitetet i Oslo, Det matematisk-naturvitenskapelige fakultet - Geografisk institutt. 111 p.
- Bakkestuen, V., Erikstad, L. & Halvorsen, R. (2008). Step-less models for regional environmental variation in Norway. *Journal of Biogeography*, 35 (10): 1906-1922.
- Barry, S. & Elith, J. (2006). Error and uncertainty in habitat models. *Journal of Applied Ecology*, 43 (3): 413-423.
- Beitelag - seter. Kilden - til arealinformasjon: Skog og landskap. Available at: <http://kilden.skogoglandskap.no> (accessed: 01.03.2013).
- Bekkby, T., Erikstad, L., Bakkestuen, V. & Bjørge, A. (2002). A landscape ecological approach to coastal zone applications. *Sarsia*, 87 (5): 396-408.
- Biondi, E. (2011). Phytosociology today: Methodological and conceptual evolution. *Plant Biosystems*, 145 (Supplement): 19-29.
- Bryn, A. & Daugstad, K. (2001). Summer farming in the subalpine birch forest. In Wielgolaski, F. E. (ed.) *UNESCO Man and Biosphere Series 27, Nordic mountain birch ecosystems*, p. 307-315. New York: Parthenon Publishing Group.
- Bryn, A. (2006). Vegetation mapping in Norway and a scenario for vegetation changes in a mountain district. *Geographia Polonica*, 79 (1): 41-64.
- Bryn, A. & Hemsing, L. Ø. (2012). Impacts of land use on the vegetation in three rural landscapes of Norway. *International Journal of Biodiversity Science, Ecosystem Services & Management*, 8 (4): 360-371.
- Bryn, A., Dourojeanni, P., Hemsing, L. O. & O'Donnell, S. (2013). A high-resolution GIS null model of potential forest expansion following land use changes in Norway. *Scandinavian Journal of Forest Research*, 28 (1): 81-98.
- Cawsey, E. M., Austin, M. P. & Baker, B. L. (2002). Regional vegetation mapping in Australia: a case study in the practical use of statistical modelling. *Biodiversity and Conservation*, 11 (12): 2239-2274.
- Dobrowski, S. Z., Safford, H. D., Cheng, Y. B. & Ustin, S. L. (2008). Mapping mountain vegetation using species distribution modeling, image-based texture analysis, and object-based classification. *Applied Vegetation Science*, 11 (4): 499-508.
- Edwardsen, A., Bakkestuen, V. & Halvorsen, R. (2011). A fine-grained spatial prediction model for the red-listed vascular plant *Scorzonera humilis*. *Nordic Journal of Botany*, 29 (4): 495-504.
- eKlima. eKlima: Meteorologisk institutt. Available at: [www.eklima.no](http://www.eklima.no) (accessed: 28.02.2013).
- Elith, J., Graham, C. H., Anderson, R. P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., et al. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29 (2): 129-151.
- Elith, J., Phillips, S. J., Hastie, T., Dudik, M., Chee, Y. E. & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, 17 (1): 43-57.
- EU Habitat Directive. (1992). *Council Directive 92/43/EEC of 21 May 1992 on the conservation of natural habitats and of wild fauna and flora*.

- Ferrier, S., Drielsma, M., Manion, G. & Watson, G. (2002). Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. Community-level modelling. *Biodiversity and Conservation*, 11 (12): 2309-2338.
- Fielding, A. H. & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24 (1): 38-49.
- Franklin, J. (2009). *Mapping species distributions: spatial inference and prediction*. Cambridge: Cambridge University Press. 320 p.
- Fremstad, E. (1997). *Vegetasjonstyper i Norge*. NINA Temahefte 12. Trondheim: Norsk institutt for naturforskning. 279 p.
- Gudjonsson, G. (2010). Large-scale vegetation mapping in Iceland. In Bryn, A., Dramstad, W. & Fjellstad, W. (eds) *Mapping and monitoring of nordic vegetation and landscapes, Hveragerði, Iceland, September 16-18, 2009*, p. 52-54: Viten fra Skog og Landskap 1/10.
- Guisan, A. & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, 135 (2-3): 147-186.
- Guisan, A., Graham, C. H., Elith, J., Huettmann, F. & the NCEAS Species Distribution Modelling Group. (2007). Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions*, 13 (3): 332-340.
- Halvorsen, R., Andersen, T., Blom, H. H., Elvebakk, A., Elven, R., Erikstad, L., Gaarder, G., Moen, A., Mortensen, P. B., Norderhaug, A., et al. (2009). Naturtyper i Norge - Teoretisk grunnlag, prinsipper for inndeling og definisjoner. *Naturtyper i Norge versjon 1.0*, Artikkel 1: 1-210.
- Halvorsen, R. (2012a). A gradient analytic perspective on distribution modelling. *Sommerfeltia*, 35: 1-165.
- Halvorsen, R. (2012b). A maximum likelihood explanation of MaxEnt, and some implications for distribution modelling. *Sommerfeltia*, 36: 1-132.
- Hemings, L. Ø. & Bryn, A. (2012). Three methods for modelling potential natural vegetation (PNV) compared: A metodological case study from south-central Norway. *Norsk Geografisk Tidsskrift-Norwegian Journal of Geography*, 66 (1): 11-29.
- Ihse, M. (2007). Colour infrared aerial photography as a tool for vegetation mapping and change detection in environmental studies of Nordic ecosystems: A review. *Norsk Geografisk Tidsskrift-Norwegian Journal of Geography*, 61 (4): 170-191.
- Jimenez-Valverde, A., Lobo, J. M. & Hortal, J. (2009). The effect of prevalence and its interaction with sample size on the reliability of species distribution models. *Community Ecology*, 10 (2): 196-205.
- Lobo, J. M., Jimenez-Valverde, A. & Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17 (2): 145-151.
- MacDonald, D., Crabtree, J. R., Wiesinger, G., Dax, T., Stamou, N., Fleury, P., Gutierrez Lazpita, J. & Gibon, A. (2000). Agricultural abandonment in mountain areas of Europe: Environmental consequences and policy response. *Journal of Environmental Management*, 59 (1): 47-69.
- Marino, J., Bennett, M., Cossios, D., Iriarte, A., Lucherini, M., Plissock, P., Sillero-Zubiri, C., Villalba, L. & Walker, S. (2011). Bioclimatic constraints to Andean cat distribution: a modelling application for rare species. *Diversity and Distributions*, 17 (2): 311-322.
- Merckx, B., Steyaert, M., Vanreusel, A., Vincx, M. & Vanaverbeke, J. (2011). Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. *Ecological Modelling*, 222 (3): 588-597.
- Moen, A. (1999). *Vegetation*. National Atlas of Norway. Hønefoss: Norwegian Mapping Authority. 200 p.
- Mork, E. & Heiberg, H. H. (1937). Om vegetasjonen i Hirkjølen forsøksområde. *Meddelelser fra Det norske Skogforsøksvesen*, 19: 617-668.
- Newton-Cross, G., White, P. C. L. & Harris, S. (2007). Modelling the distribution of badgers *Meles meles*: comparing predictions from field-based and remotely derived habitat data. *Mammal Review*, 37 (1): 54-70.
- NOU. (1983: 46). *Norsk Kartplan 2*. Norges Offentlige Utredning. Oslo.

- Ortega-Huerta, M. A. & Peterson, A. T. (2008). Modeling ecological niches and predicting geographic distributions: a test of six presence-only methods. *Revista Mexicana De Biodiversidad*, 79 (1): 205-216.
- Parolo, G., Rossi, G. & Ferrarini, A. (2008). Toward improved species niche modelling: Arnica montana in the Alps as a case study. *Journal of Applied Ecology*, 45 (5): 1410-1418.
- Pearce, J. & Ferrier, S. (2000). Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, 133 (3): 225-245.
- Phillips, S. J., Dudik, M. & Schapire, R. E. (2005). A brief tutorial on Maxent. *AT&T Research*: 38 p.
- Phillips, S. J., Anderson, R. P. & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190 (3-4): 231-259.
- Phillips, S. J. & Dudik, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, 31 (2): 161-175.
- Påhlsson, L. (1998). Vegetationstyper i Norden. *TemaNord 1998:510*. København: Nordiska Ministerrådet. 710 p.
- Reddy, S. & Davalos, L. M. (2003). Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, 30 (11): 1719-1727.
- Rekdal, Y. & Larson, J. Y. (2005). Veiledning i vegetasjonskartlegging M 1:20 000 - 50 000. *NIJOS-rapport 5/05*. Ås: Norwegian Forest and Landscape Institute. 108 p.
- Rekdal, Y. & Bryn, A. (2010). Mapping of vegetation in Norway. In Bryn, A., Dramstad, W. & Fjellstad, W. (eds) *Mapping and monitoring of nordic vegetation and landscapes, Hveragerði, Iceland, September 16-18, 2009*, p. 93-96: Viten fra Skog og Landskap 1/10.
- Roy, P. S., Joshi, P. K., Singh, S., Agarwal, S., Yadav, D. & Jegannathan, C. (2006). Biome mapping in India using vegetation type map derived using temporal satellite data and environmental parameters. *Ecological Modelling*, 197: 148-158.
- Stokland, J. N., Halvorsen, R. & Støa, B. (2011). Species distribution modelling-Effect of design and sample size of pseudo-absence observations. *Ecological Modelling*, 222 (11): 1800-1809.
- Strand, G. H. & Rekdal, Y. (2010). The norwegian area frame survey of land resources: AR18x18. In *Mapping and monitoring of nordic vegetation and landscapes, Hveragerði, Iceland, September 16-18, 2009*, p. 113-116: Viten fra Skog og Landskap 1/10.
- Strand, G. H. (2013). The Norwegian area frame survey of land cover and outfield land resources. *Norsk Geografisk Tidsskrift-Norwegian Journal of Geography*, 67 (1): 24-35.
- Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, 240 (4857): 1285-1293.
- Topografisk rasterkart 2WMS*. geoNorge: Norge digitalt. Available at: [www.geonorge.no](http://www.geonorge.no) (accessed: 10.4.2013).
- Vegetasjonskart - dekning*. Kilden - til arealinformasjon: Skog og landskap. Available at: <http://kilden.skogoglandskap.no> (accessed: 01.03.2013).
- Weber, T. C. (2011). Maximum entropy modeling of mature hardwood forest distribution in four U.S. states. *Forest Ecology and Management*, 261 (3): 779-788.
- Whittaker, R. H. (1960). Vegetation of the Siskiyou mountains, Oregon and California. *Ecological Monographs*, 30 (3): 280-338.
- Wollan, A. K., Bakkestuen, V., Kausrud, H., Gulden, G. & Halvorsen, R. (2008). Modelling and predicting fungal distribution patterns using herbarium data. *Journal of Biogeography*, 35 (12): 2298-2310.



## Appendices

**Appendix 1** – List of all 16 vegetation types that were mapped, with the total area covered, as well as area covered in each of the two frame-areas of the study area. The total percentage coverage is also provided.

Vegetation type	Total area covered (m <sup>2</sup> )	Frame-area training (m <sup>2</sup> )	Frame-area projection (m <sup>2</sup> )	Proportion of study area (%)
2c - Lichen heath	296444	152007	144437	4
2e - Dwarf shrub heath	2155866	1199038	956828	27
3b - Tall forb meadow	50196	-	50196	1
4b - Blueberry birch forest	1934303	1031210	903093	24
4c - Meadow birch forest	735535	320311	415224	9
7b - Blueberry spruce forest	485481	135262	350219	6
7c - Meadow spruce forest	318731	215904	102827	4
8c - Poor swamp forest	15933	-	15933	0
8d - Rich swamp forest	2021	2021	-	0
9a – Bog	723315	337408	385907	9
9b - Deer-grass fen	46710	-	46710	1
9c – Fen	726133	275664	450469	9
9d - Mud-bottom fens and bogs	161276	118825	42451	2
9e - Sedge marsh	18062	7209	10853	0
11a - Cultivated land	36845	36845	-	0
11b - Pastures	164433	74373	90060	2
Water	90658	50749	39909	1
<b>Total</b>	<b>7961942</b>	<b>3956826</b>	<b>4005116</b>	<b>100</b>

**Appendix 2** – Additional symbols used to describe variation within the same vegetation type.

Sign	Described variation
&	Deciduous trees, unspecified
+	Scots pine
*	Norway spruce
j	More than 50 % cover of <i>Juniperus communis</i>
!	25-50 % cover of <i>Salix</i> sp.
s	More than 50 % cover of <i>Salix</i> sp.
g	More than 50 % cover of grass
v	25 – 50 % cover of lichens
x	More than 50% cover of lichens
<	25 – 50 % of area < 30 cm soil thickness/ 25-75 % cover of bare ground
>	> 50 % of area < 30 soil thickness / 25-75 % cover of bare ground
{	25-50 % cover of stone block
}	50-75 % cover of stone block
]	25-50 % cover of trees

**Appendix 3** – The number of presence-points and the prevalence (number of presences divided by the sum of all presence and absence points) used in training and projection for each vegetation type. Explanations of the codes for the vegetation types are given in Appendix 1.

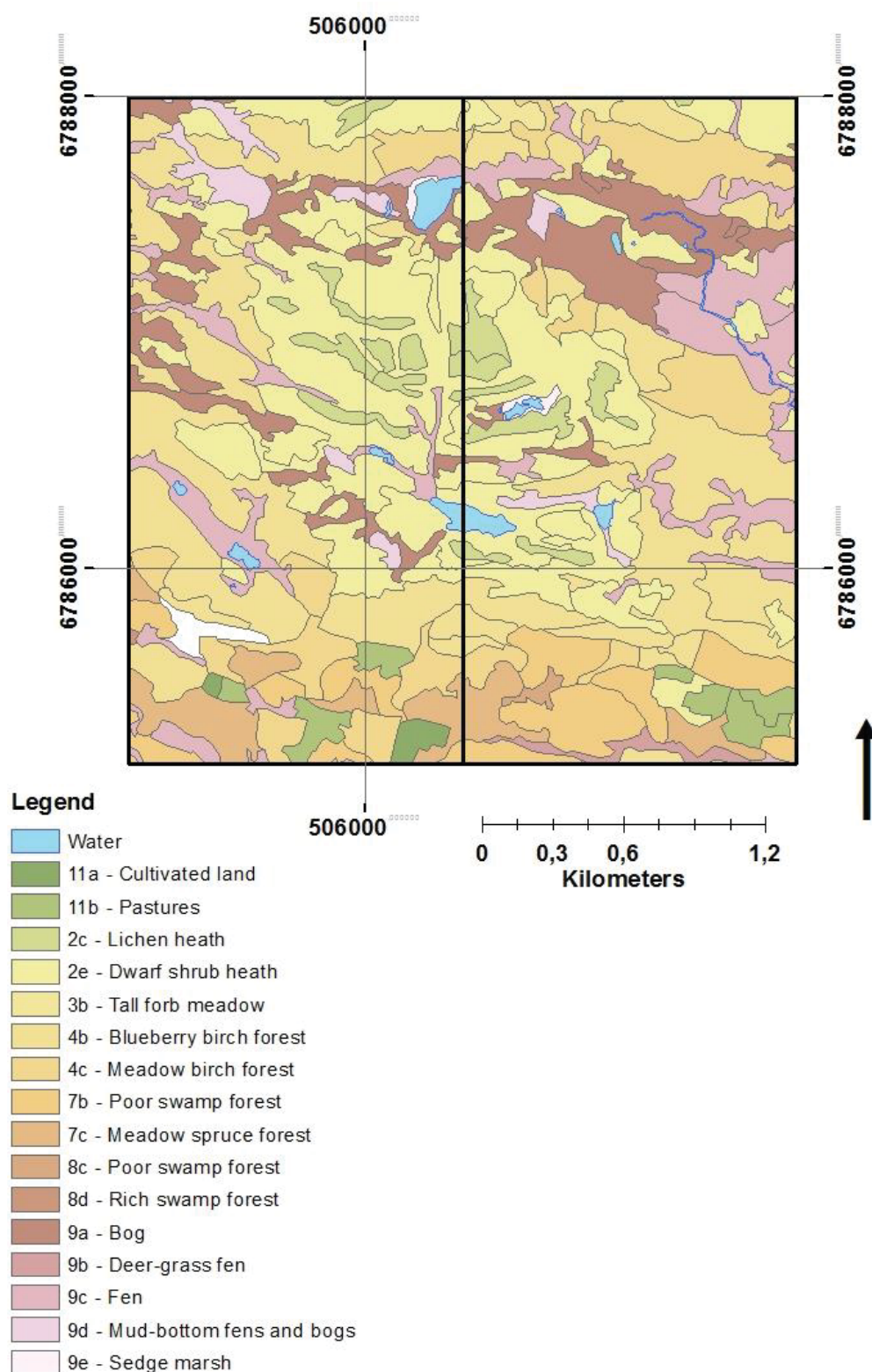
Vegetation type	Training			Projection and evaluation by independent data			
	Presences	Background points	Prevalence	2000 points		4000 points	
				Presences	Prevalence	Presences	Prevalence
<b>2c</b>	1286	11206	0.11	61	0.03	110	0.03
<b>2e</b>	11614	20906	0.56	455	0.23	912	0.23
<b>4b</b>	9728	19128	0.51	383	0.19	806	0.20
<b>7c</b>	2122	11977	0.18	51	0.03	91	0.02
<b>9c</b>	2708	12542	0.22	176	0.09	313	0.08
<b>9d</b>	1170	11094	0.11	21	0.01	39	0.01

**Appendix 4** – Details about the aerial photos used in the mapping in field. The IR-photos are from the same flight.

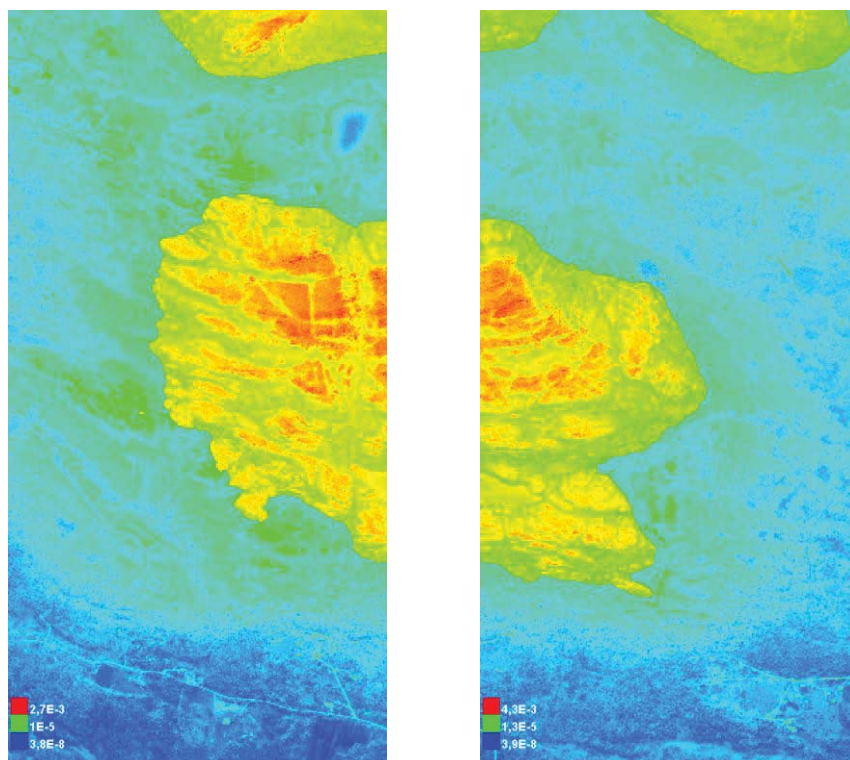
<b>Company</b>	Blom Geomatics AS
<b>Date of photography</b>	28th September 2010
<b>Series name</b>	Sogn 2010
<b>Coverage number</b>	BNO10044_01_04
<b>Resolution</b>	0.5 m
<b>Pictures</b>	Band 64, pictures 1521 - 1523 and 1821 Band 65, pictures 1390 - 1392



**Appendix 5** – Vegetation map for the study area generated during field-work, the frame-area for training to the left and the frame-area for projection and evaluation by independent data to the right. The vegetation map is simplified to improve readability; each vegetation type is represented by one color, and mosaics and additional signs are not included.



**Appendix 6** – The predictions created by Maxent for Lichen heath; the frame-area for training below to the left and the frame-area for projection to the right. Warmer colors show higher probability of presence.



**Appendix 7** – The predictions created by Maxent for Blueberry birch forest; the frame-area for training below to the left and the frame-area for projection to the right. Warmer colors show higher probability of presence.

