

Data fitness for use in conservation planning

Conservation of crop wild relatives in Norway

Karen Jordal



Thesis submitted for the degree of
Master in ecology and evolution

60 credits

Department of biosciences

Faculty of mathematics and natural sciences

University of Oslo

Spring 2017

Data fitness for use in conservation planning

*Conservation of crop wild relatives in
Norway*

Karen Jordal

© 2017 Karen Jordal

Data fitness for use in conservation planning

<http://www.duo.uio.no/>

Printed: Reprosentralen, University of Oslo

Acknowledgements

This thesis has been carried out at the Natural History Museum as a part of the Geo-Ecology (GEco) research group. My main supervisors were Senior Engineer Dag Endresen and Associate Professor Anders Bryn, and I would like to thank them both for being very supportive and positive throughout the entire master process. Åsmund Asdal from the Nordic Genetic Resource Center was co-supervisor and helped with very useful advice about field work and plant genetic resources in Norway. I would also like to thank the rest of the GEco-group for useful advice and suggestions.

This thesis would not have been finished without the loving support from all my family and friends. Helene Byhring Fosheim has given a lot of time to caring for my kids, while I was busy writing. Anne Kristine Byhring has contributed with very useful advice on the thesis, as well as a lot of emotional and practical support. Thank you!

A big thanks to Sunniva Reitan and Emily Enevoldsen for helping me out with the field work, and for being very supportive all the way.

To my father, thank you for the excellent excel-support and all of the moral support you supply by telephone.

My mother and two sisters deserve a big thank you for listening to all my complaints and helping me see the bigger picture.

The biggest thanks (and apology) goes to Øyvind Byhring, who had to endure a lot of stress and doubts from me, but still kept urging me to finish the thesis. Thank you for all your support and love!

Abstract

Crop Wild Relatives (CWR) are plants that through their close genetic relationship to crop plants have the potential to bring new genetic diversity into crops. Conservation of CWR is therefore an important task both globally and nationally.

A national recommendation of *in situ* and *ex situ* conservation of CWR in Norway has been made by Phillips *et al.* (2016), using occurrence records from the Global Biodiversity Information Facility (GBIF) and species distribution modelling (SDM) with Maxent to find hotspots of CWR diversity. The goal of this study is to explore some of the limitations when using typical GBIF-mediated data, which can be opportunistically and unsystematically sampled presence-only occurrence data. In order to investigate this, SDMs were made with GBIF-mediated presence-only occurrence data from five different CWR plant species from the CWR priority list for Norway made by Phillips *et al.* (2016) namely: *Allium ursinum* L. (ramsons), *Carum carvi* L. (caraway), *Ribes uva-crispa* L. (gooseberry), *Rubus chamaemorus* L. (cloudberry) and *Rubus idaeus* L. (wild raspberries). For each species, occurrence data was sampled to three different time periods: all points from before 1950, from 1950 to 2000 and all after 2000. In addition SDMs were made using smaller and smaller sample sizes. To test SDMs there is a need to gather independent and unbiased test data from field work, and a preliminary work has here been done to investigate possible methods of field validation.

Results indicate that older occurrence data give different models than newer data, and an approach is suggested for identifying the minimum number of presence points needed for stable SDMs. This thesis has highlighted some of the issues with spatial, temporal and species bias in GBIF-data. Being aware that the biases exist is the first step towards finding solutions to deal with it, and many solutions have been suggested by others.

Contents

1	Introduction	1
1.1	Crop Wild Relatives	1
1.2	Global Biodiversity Information Facility	3
1.3	Species Distribution Modelling	4
1.4	Research questions	5
2	Materials and methods	7
2.1	Study design	7
2.2	Investigated species	7
2.3	Study area	9
2.4	Environmental variables	9
2.5	Species Distribution Modelling	14
2.5.1	Age of presence points	14
2.5.2	Number of presence points	15
2.5.3	Model evaluation	16
2.6	Preliminary field test	19
2.6.1	Species Distribution Modelling	19
2.6.2	Field work	20
3	Results	23
3.1	Age of presence points	23

3.2	Number of presence points	31
3.3	Preliminary study	39
4	Discussion	41
4.1	Age of presence points	41
4.2	Number of presence points	43
4.3	Environmental variables	44
4.4	Model evaluation	45
4.5	Species traits	46
4.6	Field work	47
4.7	Conclusion	48

List of Figures

2.1	Outline in grey of the study area on the south-east coast of Norway.	10
2.2	An example of a marginal response curve created by Maxent for the categorical variable land cover in a model of <i>Allium ursinum</i> . On the y-axis is the raw output (relative probabilities of presence), and on the x-axis are the different land cover categories. Category 7 - snow and glacier is missing because there were no cells of this category within the study area.	17
2.3	An example of a marginal response curve created by Maxent for the continuous variable bioclim 1 - annual mean temperature in a model of <i>Allium ursinum</i> . On the y-axis is the raw output (relative probabilities of presence), and on the x-axis is temperature in °C * 10.	18
2.4	Locations of the 100 grid cells (100 X 100 m) selected for field-validation within the study area.	21
2.5	An example of the templates that were used for localization of the correct grid cells and for making registrations in the field. Each grid cell in the map represent the 100 X 100 m resolution of the SDM.	22
3.1	Part of the prediction maps for the three age classes in the <i>Allium ursinum</i> Maxent models, raw output is used.	26

List of Tables

2.1	An explanation of the bioclimatic variables downloaded from worldclim.org together with the codes that will be used as abbreviations for the variables throughout the text.	11
2.2	Environmental variables available for this study.	13
2.3	Settings used in Maxent software version 3.3.3k for main study.	14
2.4	The total number of presence points available for SDM in each year class after data cleaning.	15
2.5	The number of presence points used in each year class together with the variables that were used for SDM. All SDMs of the same species were assigned the same environmental variables.	15
2.6	Number of presence points used in each model for the five species.	16
2.7	Number of presence points, variables used and coordinate precision for the presence points used in the Maxent models.	19
3.1	Results of <i>Allium ursinum</i> models from three year classes. Response curves for the variables bioclim 1 - annual mean temperature, land cover and topographic position index with 1 km diameter. Percent contribution (pct.) for each variable is to the right of the response curve. Training AUC and test AUC for the 20 percent random test sample.	24
3.2	Results of <i>Carum carvi</i> models from three time periods, including response curves for three of the variables used: land cover, bioclim 7 - temperature annual range and bioclim 8 - mean temperature of wettest quarter. Percent contribution (pct.) for each variable is to the right of the response curve for the variable. Training AUC and test AUC for the 20 percent random test sample	27

3.3	Results of <i>Ribes uva-crispa</i> models from three time periods, including response curves for three of the variables used: bioclim 1 - annual mean temperature, land cover and wetness index. Percent contribution (pct.) for each variable is to the right of the response curve for the variable. Training AUC and test AUC for the 20 percent random test sample	28
3.4	Results of <i>Rubus chamaemorus</i> models from three time periods, including response curves for three of the variables used: bioclim 15 - precipitation seasonality, land cover and bioclim 7 - temperature annual range. Percent contribution (pct.) for each variable is to the right of the response curve for the variable. Training AUC and test AUC for the 20 percent random test sample	29
3.5	Results of <i>Rubus idaeus</i> models from three time periods, including response curves for three of the variables used: land cover, bioclim 1 - annual mean temperature and bioclim 8 - mean temperature of wettest quarter. Percent contribution (pct.) for each variable is to the right of the response curve for the variable. Training AUC and test AUC for the 20 percent random test sample	30
3.6	Results of <i>Allium ursinum</i> models with different numbers of presence points. Response curves for the variables bioclim 1 - annual mean temperature, land cover and topographic position index with 1 km diameter. Percent contribution (pct.) for each variable is to the right of the response curve. Training AUC and test AUC for the 20 percent random test sample. Numbers in [] represent different random subsets of the same number.	32
3.7	Results of <i>Carum carvi</i> models with different numbers of presence points. Response curves for the variables land cover, bioclim 7 - temperature annual range and bioclim 8 - mean temperature of wettest quarter. Percent contribution (pct.) for each variable is to the right of the response curve. Training AUC and test AUC for the 20 percent random test sample. Numbers in [] represent different random subsets of the same number.	34
3.8	Results of <i>Ribes uva-crispa</i> models with different number of presence points. Response curves for the variables bioclim 1 - annual mean temperature, land cover and wetness index. Percent contribution (pct.) for each variable is to the right of the response curve. Training AUC and test AUC for the 20 percent random test sample. Numbers in [] represent different random subsets of the same number.	36

3.9	Results of <i>Rubus chamaemorus</i> models with different number of presence points. Response curves for the variables bioclim 15 - precipitation seasonality, land cover and bioclim 7 - temperature annual range. Percent contribution (pct.) for each variable is to the right of the response curve. Training AUC and test AUC for the 20 percent random test sample. Numbers in [] represent different random subsets of the same number.	37
3.10	Results of <i>Rubus idaeus</i> models with different numbers of presence points. Response curves for the variables land cover, bioclim 1 - annual mean temperature and bioclim 8 - mean temperature of wettest quarter. Percent contribution (pct.) for each variable is to the right of the response curve. Training AUC and test AUC for the 20 percent random test sample. Numbers in [] represent different random subsets of the same number.	38
3.11	Number of each study species present in grid cells sorted by prediction value (logistic output) from the Maxent models. Classes are sorted in ascending order of prediction probability, 1 is between 0 and 0.19, 2 is between 0.2 and 0.39 and so forth. A line means that no cells were visited within that prediction interval for the species.	39

Chapter 1

Introduction

1.1 Crop Wild Relatives

Crop wild relatives (CWR) are plants that are genetically related to crop plants (Maxted *et al.*, 2006). But in a sense every life form on earth are related, so there is a need for a more precise definition. Harlan and Wet (1971) proposed the Gene Pool classification for cultivated species and their wild relatives. In this system there is a division into three Gene Pools:

Primary Gene Pool (GP-1) - consisting of the species itself and all its cultivated (GP-1A) and wild (GP-1B) varieties.

Secondary Gene Pool (GP-2) - different species, but which can still be crossed with the target species.

Tertiary Gene Pool (GP-3) - species in which gene transfer is impossible or require sophisticated techniques (Harlan and Wet, 1971; Maxted *et al.*, 2006).

When it comes to CWR in this system, they could be defined as species belonging to GP-1B or GP-2 of a crop species, but that requires information about ease of crossing and genetic relatedness that is often lacking for wild species (Maxted *et al.*, 2006). In cases where genetic information is not available, Maxted *et al.* (2006) recommend the use of the Taxon Group concept.

Taxon Group 1a – crop

Taxon Group 1b – same species as crop

Taxon Group 2 – same series or section as crop

Taxon Group 3 – same subgenus as crop

Taxon Group 4 – same genus

Taxon Group 5 – same tribe but different genus to crop

The definition of CWR that is proposed using both the Gene Pool and Taxon Group concepts is this:

“A crop wild relative is a wild plant taxon that has an indirect use derived from its relatively close genetic relationship to a crop; this relationship is defined in terms of the CWR belonging to Gene Pools 1 or 2, or taxon groups 1 to 4 of the crop.” (Maxted *et al.*, 2006, p. 2680)

Recently, a study by Schröder *et al.* (2015) on European wild grape found resistance to downy mildew, powdery mildew, and black rot in different wild populations. This shows how beneficial traits that are not present, or only present to a lesser degree in domesticated crop species can be found in CWR. There are many other examples of how genetic traits in CWR can be utilized in modern farming other than as a disease prevention tool. These include responding to climate changes or increasing yields of crops to accommodate human population growth (Dempewolf *et al.*, 2017; Redden *et al.*, 2015).

Most nature conservation efforts are constrained by other societal interests. Conserving areas involve legal earmarking which may end up with expropriation and local conflicts. Thus, the societies cannot hope to conserve areas with declared CWR everywhere. An important question will therefore be; How can CWR be conserved in a way that is effective from an ecological standpoint while still maintaining socio-economical viability?

The genetic diversity of CWR are currently inadequately preserved in gene banks. According to Castañeda-Álvarez *et al.* (2016) “over 95 % are insufficiently represented in regard to the full range of geographic and ecological variation in their native distributions”. Although many CWR can be present in protected areas, active management and conservation of these species is still a rarity (Vincent *et al.*, 2013).

A ten year project started by the Millenium Seed Bank of the Royal Botanical Gardens, Kew and the Global Crop Diversity Trust (GCDDT) aims to collect, conserve and initiate the use of CWR globally (Dempewolf *et al.*, 2014). The project is focused on 29 focal crops and the wild relatives in their gene pools. All of these 29 species can be found in Annex 1 of the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA) (FAO, 2009). The aims of the project are to identify CWR missing from gene banks, collect those CWR from the wild, evaluate CWR materials for useful traits and prepare them for use in crop improvement and make the resulting products and information widely available (Dempewolf *et al.*, 2014).

There is currently a plan for a genetic reserve for CWR at Færder national park in Vestfold Norway. This is an archipelago with an especially rich flora, and also a hotspot for CWR plants, with at least 45 CWR taxa (Kell and Maxted, 2015, page 5). Phillips *et al.* (2016) have made a prioritized list of 204 CWR that they recommend protection of in Norway. In this work they have used ecogeographic diversity as a proxy for genetic diversity, to identify complementary *in situ* genetic reserves that can preserve a larger breadth of genetic diversity.

The usefulness of CWR as a resource in agriculture, the increased interest in the subject within ecological academic societies and the difficulty of implementing conservation makes further research on CWR an urgent priority. In order to conserve the genetic diversity of CWR, there is a need for a holistic conservation approach with both genetic reserves in nature (*in situ*) and concentrated collecting efforts into gene banks (*ex situ*) (Maxted *et al.*, 2012). These efforts must have a high degree of efficiency due to limited conservation resources.

1.2 Global Biodiversity Information Facility

A national recommendation of *in situ* and *ex situ* conservation of CWR in Norway has been made by Phillips *et al.* (2016), using occurrence records from the Global Biodiversity Information Facility (GBIF) and Species Distribution Modelling (SDM) with Maxent to find hotspots of CWR diversity. The goal of this study is to explore some of the limitations when using typical GBIF-mediated data, which can be opportunistically, unsystematically sampled presence-only occurrence data.

The GBIF platform contains freely available user generated biodiversity data (Telenius, 2011). One of the challenges with using GBIF data is that you cannot know whether the presence points are representative (in both spatial, temporal and ecological terms) for the species you are interested in. The data is from both amateur collectors and professionals, but you lack information on what criteria they have used for registering the species, and most commonly it is only opportunistically collected when convenient, or in connection with a specific study. In some cases data papers are published, with more detailed explanations of how the data is gathered, but this is the exception. Bryn *et al.* (2015) is an example of such a paper.

Even though GBIF is a global platform, biodiversity data are not equally distributed across the globe. A study by Yesson *et al.* (2007) of GBIF data coverage found that many biodiversity hotspots for legumes in Africa and Asia were data deficient. This kind of spatial bias also occurs on a more local scale, with different sampling intensity in different regions and more data being collected closer to roads.

Isaac *et al.* (2014) recognize four principal forms of bias in opportunistic data: "(i) uneven recording intensity over time, measured as the number of visits per year (a visit is defined as unique combination of site and date in the records data), (ii) uneven spatial coverage, (iii) uneven sampling effort per visit and (iv) uneven detectability" (Isaac *et al.*, 2014, p. 1052).

In this thesis the focus will be on these four types of biases that are listed by Isaac *et al.* (2014), but with (iii) and (iv) as one category (species bias). All of these biases are assumed to be present in GBIF-data, and they will be referred to as:

Temporal bias - there is an increase in data being gathered and entered into GBIF over time, so there is a bias towards newer data and older data is under-represented. This could lead to misinterpretations, for instance: an increase in occurrence records of *Allium ursinum* from 1900 to 2000 does not imply that the species has become more abundant, but could simply be because it has been recorded more often.

Spatial bias - there is not random sampling of species in space, meaning that some locations have higher probability of being visited because they are conveniently close to roads or are more often visited by humans. Therefore we can not assume that the full range of species distributions are represented within GBIF.

Species bias - there is not a direct link between which species are present in an area and which are being reported to GBIF. Some species are more noticeable than others, and some require expert knowledge to correctly identify. This can lead to an over-representation of some species and an under-representation of others.

1.3 Species Distribution Modelling

In order to select possible conservation areas of CWR, Species Distribution Modelling (SDM) can be a powerful tool. The study by Phillips *et al.* (2016) of CWR conservation in Norway has used SDM in order to find potential *in situ* genetic reserves for CWRs.

The goal with SDM is to make a prediction of habitat suitability for a species in a defined area. SDM methods use species presence and sometimes absence data together with wall-to-wall maps of environmental variables for the entire prediction area to make these predictions (Guisan and Zimmermann, 2000). There are many different names for the same concept, such as ecological niche models, habitat models and resource selection functions (Elith and Graham, 2009), but in this thesis i will use the name SDM. Species Distribution Models will be referred to as SDMs.

There are many different methods to make predictions of species

distributions. Some are used when there is access to both presence and absence records, like for instance generalized linear or additive models (GLMs and GAMs), and boosted regression trees (BRTs). However, when using GBIF-data there is (normally) only presence records available, and for presence-only SDM, Maxent (Phillips *et al.*, 2006) is the most widely used and often best performing method (Elith *et al.*, 2011). A further description of the Maxent method and software will follow in the materials and methods section (chapter 2.5). In Phillips *et al.* (2016) Maxent was used, and it will also be used in this thesis.

SDMs are being widely used for many different purposes in fields like conservation planning, ecology and biogeography (Elith *et al.*, 2011). One example in the field of conservation biology and CWRs is the study by Parra-Quijano *et al.* (2012), where they used SDMs together with gap analysis and ecogeographical maps to find prioritized sites for collecting *Lupinus* species into gene banks.

However, the SDM methods have many constraints and uncertainties, a literature review by Beale and Lennon (2012) found that: "uncertainty in SDMs has often been underestimated and a false precision assigned to predictions of geographical distribution" (Beale and Lennon, 2012, p. 247). Of specific importance is the source material, and the models depend on the available species data, as well as the availability of relevant predictor variables (Beale and Lennon, 2012).

The species that is being modelled will also have an effect on the results of SDMs, according to Hanspach *et al.* (2010) the life-history traits of a species will affect the performance of SDMs. The life-history traits used in Hanspach *et al.* (2010) were: dispersal type, lifespan, life form, pollination type, strategy type, number of vegetation units a species is affiliated to and hemerobic level (details can be found in Hanspach *et al.* (2010, table 1)). In this thesis dispersal, life span and reproduction will be discussed in relation to differences in SDMs between species.

1.4 Research questions

The main research questions are these:

Is the available GBIF data from Norway, tested with the most commonly used presence-only DM method (Maxent), useful for detecting potential CWR conservation areas?

Are GBIF in general providing the needed CWR data for regional analyses of such challenges in Norway.

How does distribution models respond to different species with different life-history traits?

More specifically, this thesis addresses the following hypotheses:

Hypothesis 1:

H0 - The Maxent distribution models will be consistent when using presence-only points from different time periods if the number of presence points and all settings are kept the same

H1 - The Maxent distribution models will be inconsistent when using presence-only points from different time periods if the number of presence points and all settings are kept the same

Hypothesis 2:

H0 - The Maxent distribution models will be consistent when different numbers of random presence-only points are used and all settings are kept the same

H1 - The Maxent distribution models will be inconsistent when different numbers of random presence-only points are used and all settings are kept the same

Hypothesis 3

H0 - There will not be species trait specific differences between Maxent models

H1 - There will be species trait specific differences between Maxent models

Hypothesis 4

H0 - There will not be differences between Maxent models made with random samples of presence only points of the same size

H1 - There will be differences between Maxent models made with random samples of presence only points of the same size

Chapter 2

Materials and methods

2.1 Study design

Presence points of the five species were downloaded from GBIF via the Species Map Service (NBIC & GBIF, Download date: 2017-02-16).

After data cleaning, the data was treated in two different ways: sampled into three age classes (before 1950, 1950-2000 and after 2000) with the same number of presence points in each class and pooling of data points into classes consisting of gradually fewer points.

In addition a preliminary field study to collect presence and absence points from the study area was conducted during the summer of 2015. GBIF-data for this study was downloaded on 2014-11-28.

2.2 Investigated species

Five study species were chosen for this study from the national CWR priority list for Norway (Phillips *et al.*, 2016). The species were *Allium ursinum* L. (ramsons), *Carum carvi* L. (caraway), *Ribes uva-crispa* L. (gooseberry), *Rubus chamaemorus* L. (cloudberry) and *Rubus idaeus* L. (wild raspberry). Criteria used for choosing these species were that they should have differing life history traits, such as dispersal, life length and reproduction. Another consideration was that the plants should be recognizable during field work at the same time period.

Ramsons is a perennial plant that grows in nutritious broadleaf and sometimes coniferous forests in the nemoral and boreo-nemoral vegetation zones (Lid and Lid, 2005). Growing season is from spring to early summer. The distribution of ramsons in Norway is southerly and bound to the coast,

the northernmost registrations are from Leksvik in Nord-Trøndelag county (Lid and Lid, 2005). The dispersal modes of ramsons are vegetative growth by bulbs and seeds that are often dispersed by ants (Korsmo, 1954). A study by Herden *et al.* (2012) of genetic diversity in ramsons has shown little genetic variation between populations. Using sequences of the nuclear internal transcribed spacer ITS, and the external transcribed spacer ETS, as well as the plastidic *trnL-rpl32* and the *trnL-trnF* spacer regions, they found no genetic variation between populations in Germany. Not even a population from Ireland differed from the German population. Recently, ramsons has increased in people's awareness and has become a fashionable spice plant in Norway. Traditionally it was used as medicine, but not so much in cooking, and when it is growing in pastures it is considered a weed since it influences the taste of the milk from grazing goats and cows (Høeg, 1976).

Caraway is a wild growing native Norwegian plant that is found mainly in dry places connected to the cultural landscape, like pastures, roadsides and hay fields (Lid and Lid, 2005). It can be found throughout most of the country, but it is rarely seen in mountains, along the west-coast and in Finnmark county (Lid and Lid, 2005). Caraway grows mainly in the nemoral to north-boreal vegetation zones, but sometimes also in the low alpine belt (Lid and Lid, 2005). The seeds of caraway can be dispersed in the fur of animals even though they are smooth with no hooks or other dispersal equipment (Kiviniemi and Eriksson, 1999). Caraway is a biannual plant, it flowers and produce seeds the second year of its life cycle (Høeg, 1976). Wild growing caraway has been extensively used as a spice in foods like cheese, spirits, bread and sausages, but now the spice that you get in shops is predominantly from cultivated plants (Høeg, 1976).

Gooseberry is an old cultivated species, it has been reported grown in Aust-Agder county in 1682, and in France it is supposed to have been grown since the 13th century (Skard, 2007). Even though it is not a native Norwegian species, it is commonly naturalized in nutritious forests, forest edges and on shallow soils (Lid and Lid, 2005). Gooseberry grows in the nemoral to north-boreal vegetation zones, and occasionally in the south-boreal zone. It is quite common in the lowlands of eastern Norway, and further north it can be found in valleys all the way up to Nordland county (Lid and Lid, 2005). In the Species Map Service there are also some observations from Troms county (NBIC & GBIF, accessed: 2017-06-11).

Cloudberry is a perennial plant that grows in peatlands, nutrient poor swamp forests and moorland, but avoids alkaline soils (Lid and Lid, 2005). It is common in all of Norway both in the lowlands and in the mountains. Mostly the plants are dioecious, that is with male and female parts on separate plants, but sometimes hermaphroditic plants occur and these have been used in cloudberry breeding programs (Rapp *et al.*, 1993). Growing only on peat and in nutrient poor areas, this plant could become an important crop in areas where little else can be grown (Rapp *et al.*, 1993). Two

male and two female cloudberry cultivars have been produced in Norway (Rapp and Martinussen, 2002). Reproduction of cloudberry is mainly by underground production of rhizomes, so large populations can consist of only a few clones spread over large areas. A study by Korpelainen *et al.* (1999) showed that three populations in Finland consisted of 2-4 clones each. They used both 10-base RAPD and 16-base SSR primers. Although sexual reproduction is rare in cloudberry, it is probably important because the seeds can spread over larger distances and recombination will result in more genetic variation (Korpelainen *et al.*, 1999).

Wild raspberries occur in most of Norway, but are uncommon in the northernmost county, Finnmark. It is a perennial berry bush that grows in broadleaf forests, forest edges, thickets, roadsides and abandoned fields, usually on nutrient rich soils. Wild raspberries have vegetative growth by roots, and produce flowers by their second year (Lid and Lid, 2005). A study by Graham *et al.* (2009) analyzed 12 wild and 5 cultivated populations of raspberries in Scotland by 10 simple sequence repeats (SSR). These SSRs showed a much higher diversity in wild raspberries, a total of 80 alleles were found in wild raspberries and only 18 of these alleles were found in the cultivated berries.

2.3 Study area

An area on the south-east coast of Norway was selected (Fig. 2.1). Since the resolution of the environmental variables used were 100 m, modelling on the entire mainland Norway would have been very time consuming and computer intensive. Also, field validation would have been very challenging with respect to distances, topography and the short field-season in Norway. Therefore a smaller area was chosen that still had a lot of environmental variation, from coast to mountains (highest elevation is 1250 meters above sea level). This area is also rich in CWR taxa (Phillips *et al.*, 2016).

2.4 Environmental variables

19 bioclimatic variables were downloaded from worldclim.org (Hijmans *et al.*, 2005). A list of these variables and the abbreviations that will be used for them in this text is included in table 2.1. All 19 variables had a resolution of 30 arc-seconds (approximately 1 km), and were downscaled to 100 m resolution by the interpolation method kriging in the GIS-software ArcMap (ESRI, 2014, Kriging tool). Kriging is a statistical interpolation method that estimates values along a continuous surface based on observed sample values (O'Sullivan and Unwin, 2014, Chapter 10.4)

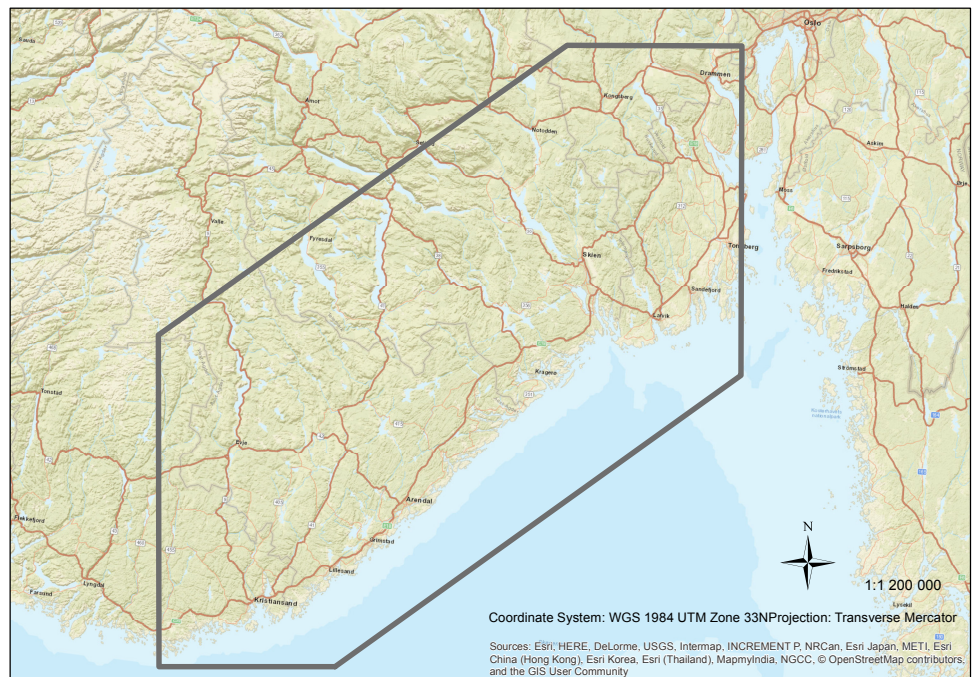


Figure 2.1: Outline in grey of the study area on the south-east coast of Norway.

Table 2.1: An explanation of the bioclimatic variables downloaded from worldclim.org together with the codes that will be used as abbreviations for the variables throughout the text.

Code	Explanation
bio1	Annual Mean Temperature
bio2	Mean Diurnal Range (Mean of monthly (max temp - min temp))
bio3	Isothermality (BIO2/BIO7) (* 100)
bio4	Temperature Seasonality (standard deviation *100)
bio5	Max Temperature of Warmest Month
bio6	Min Temperature of Coldest Month
bio7	Temperature Annual Range (BIO5-BIO6)
bio8	Mean Temperature of Wettest Quarter
bio9	Mean Temperature of Driest Quarter
bio10	Mean Temperature of Warmest Quarter
bio11	Mean Temperature of Coldest Quarter
bio12	Annual Precipitation
bio13	Precipitation of Wettest Month
bio14	Precipitation of Driest Month
bio15	Precipitation Seasonality (Coefficient of Variation)
bio16	Precipitation of Wettest Quarter
bio17	Precipitation of Driest Quarter
bio18	Precipitation of Warmest Quarter
bio19	Precipitation of Coldest Quarter

4 variables with a resolution of 100 m were donated from Lars Erikstad at The Norwegian Institute for Nature Research (NINA). These were a digital elevation model (DEM), land cover and topographic position index (TPI) with a diameter of 1 km (TPI1) and 6 km (TPI6). The TPI is the difference between a cell elevation value and the average elevation of the neighborhood around that cell, so a positive value means that the cell has higher elevation than its surroundings while a negative value means that it has lower elevation (Jennes, 2006).

Solar radiation was derived using the Area Solar Radiation tool from the Spatial Analyst toolbox in ArcMap. This tool uses the DEM raster to calculate insolation. Aspect, curvature, and slope were also calculated from the DEM using the corresponding tools in the Spatial Analyst toolbox. SAGA Wetness index was calculated in the SAGA-GIS software (Boehner and Conrad, 2001). More detail about the variables can be found in table 2.2.

Table 2.2: Environmental variables available for this study.

Variable name	Source	Explanation	Units
Land cover	Lars Erikstad at The Norwegian Institute for Nature Research (NINA)	Divided into 8 classes of land cover	1-Developed land 2-Agricultural land 3 -Fresh water 4-Sea 5-Mires 6-Forest 7-Snow and glacier 8-Open
Aspect	Derived from DEM in ArcMap	Direction of slope	Degrees 0 and 360 is north and 180 is south
BIO1 - BIO19	From worldclim.org	Bioclimatic variables	Temperatures in $^{\circ}\text{C} * 10$ and precipitation data in mm
Curvature	Derived from DEM in ArcMap	Shape of the slope, three types: planform, profile and standard	1/100 of a z-unit positive cell value is convex negative is concave
Digital Elevation Model (DEM)	Lars Erikstad at NINA	Heightmap of Norway	Meters above sea level
Slope	Derived from DEM in ArcMap	Maximum change rate from one cell to its neighbours	Degrees from 0 - 90
Solar radiation	Derived from DEM in ArcMap	Incoming solar radiation divided into four seasons of three months each	Watt hours per square meter (WH / m^2)
Topographic position index 1 (TPI1) and 6 (TPI6)	Lars Erikstad at NINA	height of a cell - mean height of all cells within a 1 km diameter and 6 km diameter	unitless
SAGA Wetness index	Eva Solbjørg Flo Heggem at NIBIO	An index of soil moisture based on climatic data, digital terrain models, land use and satellite data	unitless

Correlation between variables was calculated using the Analysis Tool-Pak in Excel 2010 (Microsoft, 2010). Strongly correlated variables (both positively and negatively), with correlation coefficient < 0.7 were removed. The remaining variables were: Land cover (cover), aspect, BIO1, BIO3, BIO7, BIO8, BIO15, standard curvature (curv), slope, solar radiation summer (solar), TPI1, TPI6 and SAGA wetness index (wetind).

2.5 Species Distribution Modelling

For Species Distribution Modelling (SDM) the software Maxent version 3.3.3k (Phillips *et al.*, 2011) was used. Maxent is short for Maximum entropy, a general statistical method for making predictions from incomplete information (Phillips *et al.*, 2006). In distribution modelling, Maxent is used as a machine learning method, "The idea of Maxent is to estimate a target probability distribution by finding the probability distribution of maximum entropy (i.e., that is most spread out, or closest to uniform), subject to a set of constraints that represent our incomplete information about the target distribution." (Phillips *et al.*, 2006, p. 234). Maxent has many different settings to choose from and the ones used in this study are listed in table 2.3.

Table 2.3: Settings used in Maxent software version 3.3.3k for main study.

<i>Setting type</i>	<i>Chosen setting</i>
Output	raw
Random test percentage	20
Features	Linear, quadratic and product
Create response curves	true
Do jackknife to measure variable importance	true
All else	Default

2.5.1 Age of presence points

Duplicates were removed using Excel (Microsoft, 2010) and the presence points were checked in ArcMap (ESRI, 2014) to see if any points were in the sea. All models of the same species were run with the same number of presence points. Since availability of presence points from before 1950 was the limiting factor, the total number of presences from before 1950 was chosen as the number of points that models were run with (see table 2.4).

The subsets from the other age classes were picked at random from the entire set of presence points in their class. Variables for each species were selected from the 13 available variables after removal of the strongly

Table 2.4: The total number of presence points available for SDM in each year class after data cleaning.

Year class	<i>Allium ursinum</i>	<i>Carum carvi</i>	<i>Ribes uva-crispa</i>	<i>Rubus chamaemorus</i>	<i>Rubus idaeus</i>
Pre 1950	49	24	53	66	163
1950-2000	91	122	539	197	1609
Post 2000	88	581	240	124	1168

correlated variables (see end of chapter 2.4.). A model was run for each species with all the 13 variables and then the ones that had lower than 2 percent contribution were thrown out for the next model run. These preliminary models were made with the after 2000 dataset. Then new models were made with only the variables that had contributed more than 2 percent included. Variables used and number of presence points used in each model is listed in table 2.5.

Table 2.5: The number of presence points used in each year class together with the variables that were used for SDM. All SDMs of the same species were assigned the same environmental variables.

Year class	<i>Allium ursinum</i>	<i>Carum carvi</i>	<i>Ribes uva-crispa</i>	<i>Rubus chamaemorus</i>	<i>Rubus idaeus</i>
Pre 1950	49	24	66	53	163
1950-2000	49	24	66	53	163
Post 2000	49	24	66	53	163
Variables	cover, bio1, tpi1	cover, bio1, bio7, bio8	cover, bio1, bio15, slope, wetind, tpi6	cover, aspect, bio1, bio3, bio7, bio15, slope, tpi6	cover, bio1, bio8, bio15, solar, tpi1

2.5.2 Number of presence points

The total number of presence points varied between species, *Rubus idaeus* had 2953 points and *Allium ursinum* had only 236. The dataset was split in half several times until the number of points was below 10, and Maxent models were run with each subset of points (table 2.6). The subsets were always chosen randomly from the full set of points. Variables used for the different species were the same as in table 2.5.

After all models were run, a sensitivity test was done on some of the

Table 2.6: Number of presence points used in each model for the five species.

<i>Allium ursinum</i>	<i>Carum carvi</i>	<i>Ribes uva-crispa</i>	<i>Rubus chamaemorus</i>	<i>Rubus idaeus</i>
236	729	434	837	2953
118	365	217	419	1477
59	182	109	209	738
30	91	54	105	369
15	46	27	52	185
7	23	14	26	92
	11	7	13	46
	6		7	23
				12
				6

sample sizes. This means that the same number of points were picked at random from the complete set of points for the species, and models were run with these new subsets. As an example, in *Allium ursinum* three models with 59 points were run and three with 30 points.

2.5.3 Model evaluation

Models were evaluated using the area under the receiver operating characteristic curve (AUC) for the training sample and the 20 % that was used as test samples in all Maxent runs (see table 2.3). The receiver operating characteristic curve (ROC) is created by plotting the fraction of true positives against the fraction of false positives (Hernandez *et al.*, 2006). An AUC value of 0.5 indicates a prediction that is no better than a random prediction. When using presence only data, AUC compares presences with background points in stead of absences (Merow *et al.*, 2013). So with these data high AUC-values indicate that the model can distinguish between presences and background points well (Merow *et al.*, 2013).

In addition the response curves for each variable that Maxent creates were visually inspected for differences between models. There are two types of response curves that Maxent makes. The marginal response curves show how the prediction change as each environmental variable is varied, keeping all other variables at their average sample value (from the HTML-file that is included in the Maxent output folder). In the other type of response curve, each curve represent a different model, a Maxent model created with only the corresponding value. These will be called variable-only response curves in this thesis. For the purpose of model comparisons the marginal response curves were used in this study. An example of a response curve for a categorical variable (land cover) is included in figure

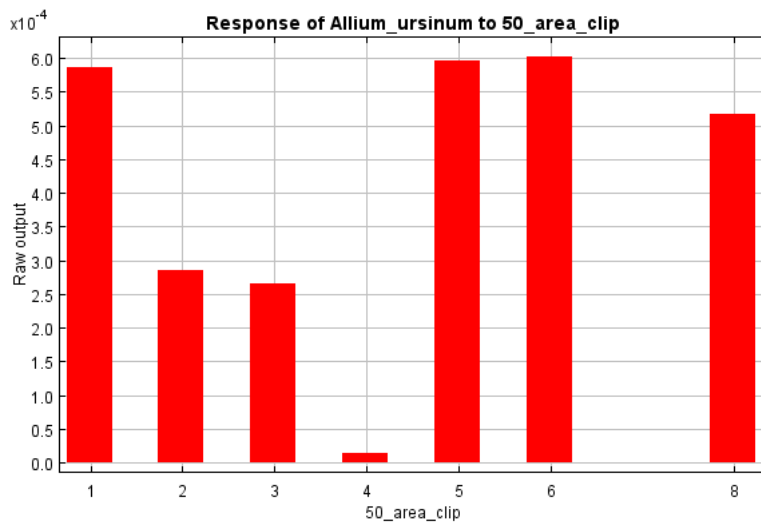


Figure 2.2: An example of a marginal response curve created by Maxent for the categorical variable land cover in a model of *Allium ursinum*. On the y-axis is the raw output (relative probabilities of presence), and on the x-axis are the different land cover categories. Category 7 - snow and glacier is missing because there were no cells of this category within the study area.

2.2, and one for a continuous variable (bio1) in figure 2.3.

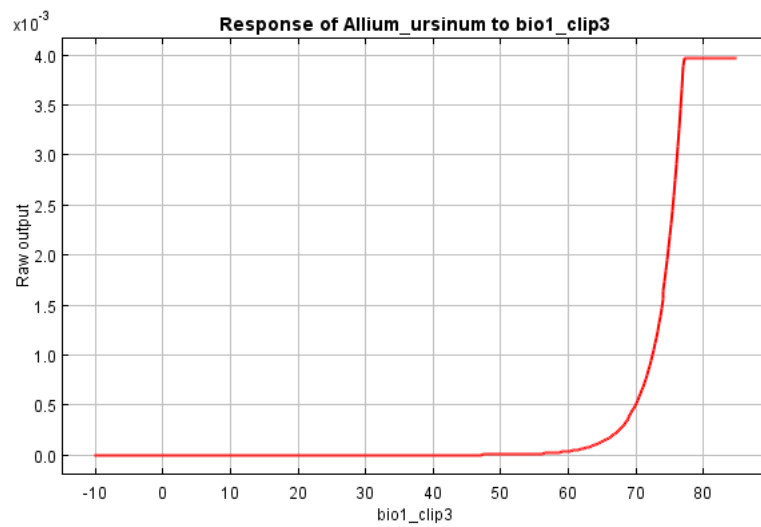


Figure 2.3: An example of a marginal response curve created by Maxent for the continuous variable bioclim 1 - annual mean temperature in a model of *Allium ursinum*. On the y-axis is the raw output (relative probabilities of presence), and on the x-axis is temperature in °C * 10.

2.6 Preliminary field test

A preliminary study to collect presence and absence data from the study area was conducted during the summer of 2015. Before the field work, five SDMs were made. A total of 13 sites were visited.

2.6.1 Species Distribution Modelling

Species presence data was downloaded from the Norwegian Species Map Service (NBIC & GBIF, Download date: 2014-11-28). This was done in separate files for different coordinate uncertainty, better than 10000 m, 1000 m, 100 m and 25 m. The best uncertainty (25 m) was used for all but two of the models. *Allium ursinum* and *Ribes uva-crispa* had so few points that coordinate uncertainty better than 100 m was chosen (see table 2.7). Data points from after year 2000 was used in all models.

A model for each of the study species was chosen based on model selection. First a full model with all variables was run, and then variables that had less than 2 percent contribution were removed. The same Maxent settings were used as in table 2.3, except that logistic output was used in stead of raw output.

Table 2.7: Number of presence points, variables used and coordinate precision for the presence points used in the Maxent models.

Species	<i>Allium ursinum</i>	<i>Carum carvi</i>	<i>Ribes uva-crispa</i>	<i>Rubus chamaemorus</i>	<i>Rubus idaeus</i>
Number of points	54	77	85	100	432
Variables	dem, cover, bio2, slope, tpi6	cover, dem, bio12, tpi1, aspect	dem, cover, bio4	bio12, cover, slope, aspect, tpi1, bio15	cover, bio13, dem
Coordinate precision	100	25	100	25	25

The variables used differed from those used in the main study (table 2.5), because strongly correlated variables were left out after model selection in the preliminary study. In the main study correlated variables were removed before modelling started.

2.6.2 Field work

Using the five models, 100 different locations were chosen including predicted presences from 0 to 1. 10 points from each model representing 10 classes of equal intervals ranging from 0 to 1 in relative prediction probability.

The prediction map, an ascii-file in the Maxent output folder, was used to make the selection of locations. Here is the procedure for selecting locations:

1. In ArcMap: Layer properties - Symbology change from stretched to classified with equal intervals and 10 classes
2. Use Reclassify tool in spatial analyst toolbox to reclassify into 10 classes (1 to 10).
3. Use conversion tool Raster to Polygon
4. Dissolve tool - use gridcode
5. Create random points from the data management toolbox, 6 per polygon class
6. To remove points that are in the sea, extract values to point in the spatial analyst toolbox, input raster is the land cover raster, edit features, select by attributes and delete rastervalue 4 (which is the sea)
7. Merge from data management toolbox all datasets of random points from all 5 species
8. Extract multi values to points in spatial analyst toolbox, using the merged dataset as input point feature and the 5 prediction rasters as input rasters.
9. Delete the superfluous points so there is only two points in each class for each species creating a dataset of 100 random points.

Cells that were found to be in urban areas or in the middle of large fields were excluded, using Google Maps (Google, 2015). Due to time shortage, only 13 of the 100 cells were visited. Figure 2.4 shows a map of the 100 locations that were selected.

For field registrations a field form template was made using ArcMap (ESRI, 2014). An example of these templates is included in figure 2.5. The template contains a number for the grid cell (0-99), the date of the visit, registration form for registering presence or absence of species, area cover (land cover), elevation and aspect of the grid cell. The midpoint coordinate was also included, and this was used to find the grid cell by using a GPS. The map was included in order to find the best driving route to the cell.

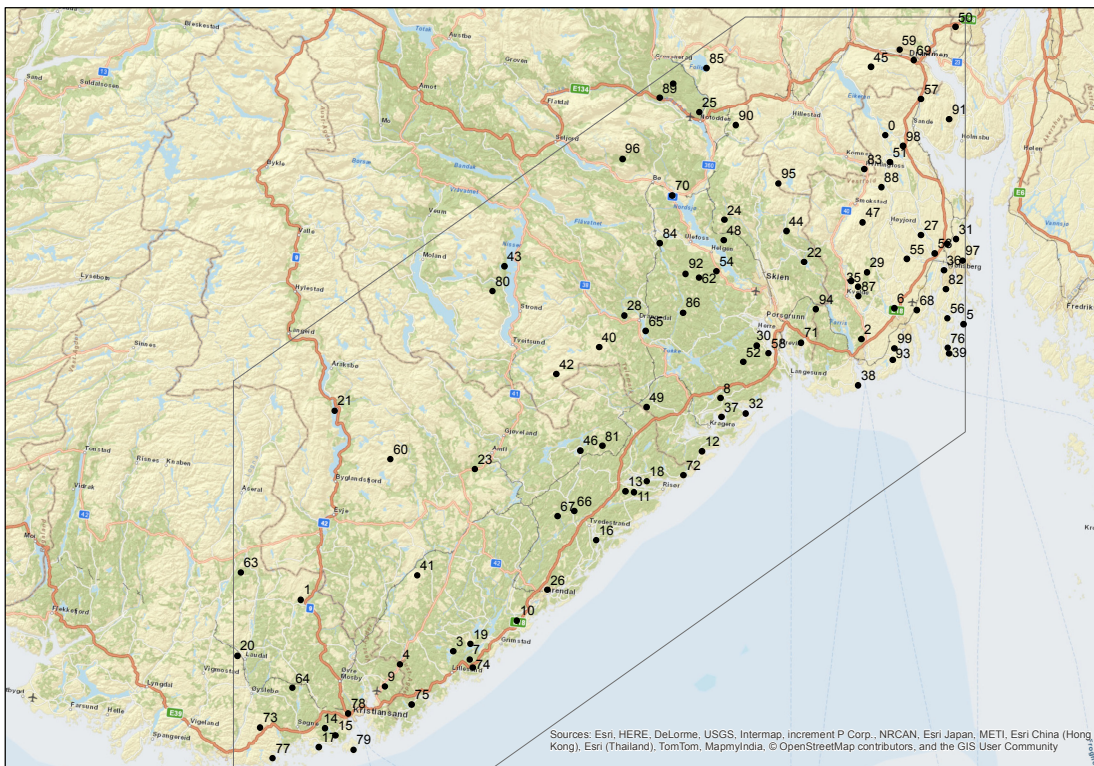


Figure 2.4: Locations of the 100 grid cells (100 X 100 m) selected for field-validation within the study area.

Grid cell 0	Presence	Absence	Amount
Date:	A.urs		
	C.car		
	R.cha		
	R.ida		
	R.ucr		
	Area cover		
	Elevation		
	Aspect		

Midpoint coordinates: 32V 561550 6603020 UTM

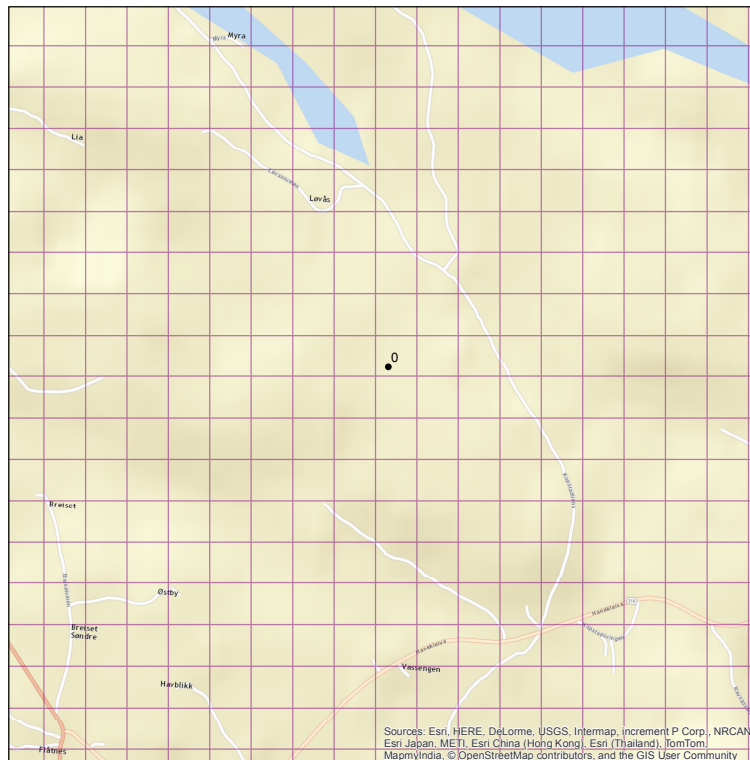


Figure 2.5: An example of the templates that were used for localization of the correct grid cells and for making registrations in the field. Each grid cell in the map represent the 100 X 100 m resolution of the SDM.

Chapter 3


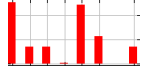
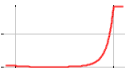




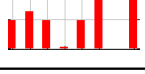

Results

The results of the SDMs based on GBIF-data of different ages and numbers will be presented here, together with the results of the preliminary field work that was conducted during the summer of 2015.

3.1 Age of presence points

The response curves for the variable bioclim 1 - annual mean temperature changes little between models as seen in table 3.1, whilst the response curve for land cover changed quite a bit between the age classes. In the model with presences from before 1950, land cover category 1 - developed land and 5 - mires are high, while in the 1950 - 2000 model 3 - freshwater and 6 - forest are more important. In the after 2000 model 6 - forest and 8 - open are highest (see figure 2.2 for a more detailed view of a land cover response curves and table 2.2 for category names). The topographic position index has the smallest percent contribution in all three models, and it changes a lot as well. The curve is completely flipped around when comparing the before 1950 model with the two others. Both test and training AUC values are above 0.9 in all models, and test AUC is higher than training AUC in the before 1950 and 1950-2000 models.

Table 3.1: Results of *Allium ursinum* models from three year classes. Response curves for the variables bioclim 1 - annual mean temperature, land cover and topographic position index with 1 km diameter. Percent contribution (pct.) for each variable is to the right of the response curve. Training AUC and test AUC for the 20 percent random test sample.

<i>Allium ursinum</i>								
Year class	Bioclim 1	pct.	land cover	pct.	TPI1	pct.	Training AUC	Test AUC
Before 1950		71.3		28.6		0.1	0.905	0.943
1950 - 2000		73.8		23.6		2.7	0.915	0.929
After 2000		73		23		4	0.927	0.924

The prediction maps of *Allium ursinum* in figure 3.1 show that the model with presences from before 1950 have a coarser distinction between areas with lower predicted probabilities of presence (light blue and blue). Some of the high probability areas (red) in the two lower maps are not red in the top one. Overall the 1950 - 2000 and after 2000 models look more similar than the before 1950 model.

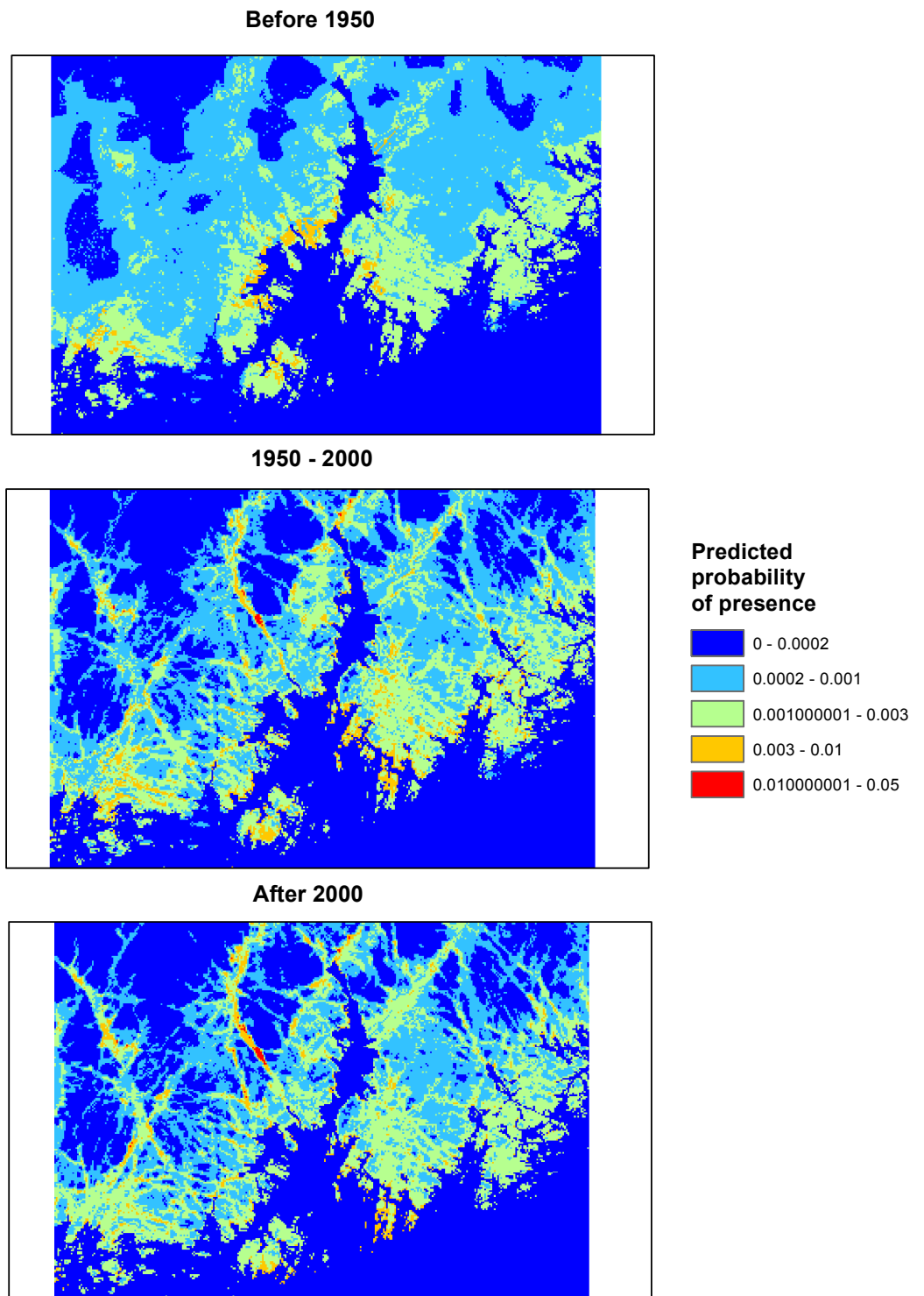
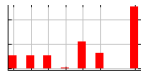


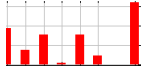


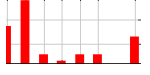




Figure 3.1: Part of the prediction maps for the three age classes in the *Allium ursinum* Maxent models, raw output is used.


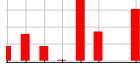


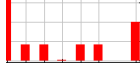


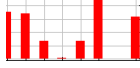

In table 3.2 we see that *Carum carvi* models show the same trend as *Allium ursinum* models, but the training AUC-values are much lower. In addition, the test AUC-values for the before 1950 model is 0.2 lower than the training AUC. The biggest change takes place between the after 2000 model and the two others, here all the response curves changes direction, for instance the highest bar in the land cover variable changes from category 8 - open in the two old models to category 2 - agricultural land in the youngest model. The temperature annual range (bioclim 7), which is the difference between the maximum temperature of the warmest month and the minimum temperature of the coldest month, also changes from higher predicted probability when it is low to higher predicted probability when high.

Table 3.2: Results of *Carum carvi* models from three time periods, including response curves for three of the variables used: land cover, bioclim 7 - temperature annual range and bioclim 8 - mean temperature of wettest quarter. Percent contribution (pct.) for each variable is to the right of the response curve for the variable. Training AUC and test AUC for the 20 percent random test sample

<i>Carum carvi</i>								
Year class	Land cover	pct.	Bioclim 7	pct.	Bioclim 8	pct.	Training AUC	Test AUC
Before 1950		32.3		28.4		3.2	0.862	0.665
1950 - 2000		75.4		9.8		0	0.815	0.848
After 2000		72.8		8.7		17.4	0.799	0.750


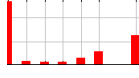

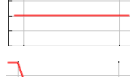
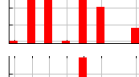




For *Ribes uva-crispa* as well, variation occurs in the response plot of the land cover variable (table 3.3). The wetness index variable has 0 percent contribution in the two older models. Again we see a drop of almost 0.2 from training to test-AUC, this time in the 1950-2000 model.

Table 3.3: Results of *Ribes uva-crispa* models from three time periods, including response curves for three of the variables used: bioclim 1 - annual mean temperature, land cover and wetness index. Percent contribution (pct.) for each variable is to the right of the response curve for the variable. Training AUC and test AUC for the 20 percent random test sample

<i>Ribes uva-crispa</i>								
Year class	Bioclim 1	pct.	land cover	pct.	Wetness index	pct.	Training AUC	Test AUC
Before 1950		63		32.3		0	0.865	0.905
1950 - 2000		55.2		39.8		0	0.819	0.628
After 2000		62.1		30.7		4.2	0.846	0.813

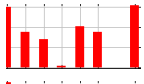
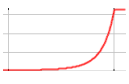

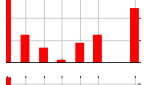

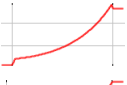
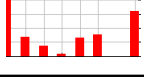
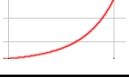
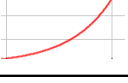
The *Rubus chamaemorus* models shows large variation in both response curves and percent contribution. The percent contribution of Bioclim 15 ranges from 0 to 30.1, whereas land cover ranges from 3.1 to 55.8. (table 3.4). Test AUC in the before 1950 model is especially low, only 0.506 which is approximately the same as a random prediction.

Table 3.4: Results of *Rubus chamaemorus* models from three time periods, including response curves for three of the variables used: bioclim 15 - precipitation seasonality, land cover and bioclim 7 - temperature annual range. Percent contribution (pct.) for each variable is to the right of the response curve for the variable. Training AUC and test AUC for the 20 percent random test sample

<i>Rubus chamaemorus</i>								
Year class	Bioclim 15	pct.	land cover	pct.	Bioclim 7	pct.	Training AUC	Test AUC
Before 1950		0.9		55.8		9.7	0.764	0.506
1950 - 2000		0		3.1		11.3	0.737	0.735
After 2000		30.1		27.8		13.8	0.867	0.749

The general shape of response curves in the *Rubus idaeus* models are similar, except for bioclim 1 - annual mean temperature which has the opposite orientation in the 1950 - 2000 model (table 3.5). However, the percent contribution varies for all variables, bioclim 1 ranges from 12 to 49.7, bioclim 8 from 1.3 to 11 and land cover from 34.4 to 46.6. The AUC values decrease from the oldest to the youngest model, and we see high test AUC-values. Before 1950 and 1950 - 2000 models even have higher test AUC than training AUC.

Table 3.5: Results of *Rubus idaeus* models from three time periods, including response curves for three of the variables used: land cover, bioclim 1 - annual mean temperature and bioclim 8 - mean temperature of wettest quarter. Percent contribution (pct.) for each variable is to the right of the response curve for the variable. Training AUC and test AUC for the 20 percent random test sample

<i>Rubus idaeus</i>								
Year class	land cover	pct.	Bioclim 1	pct.	Bioclim 8	pct.	Training AUC	Test AUC
Before 1950		34.4		49.7		1.3	0.805	0.859
1950 - 2000		56.3		12		7.2	0.731	0.757
After 2000		46.6		29.7		11	0.755	0.731

3.2 Number of presence points

In table 3.6 some of the results of the *Allium ursinum* Maxent models with decreasing number of presence points are shown. AUC values range from 0.841 to 0.878 and there are generally small differences between training and test AUC.

Bioclim 1 - annual mean temperature remain the most contributing variable in all models, and the shape of the response curve varies little, except for a less steep curve in the 7 points model. The land cover variable change little between the 236 and 118 points models, both percent contribution and shape of response is similar. There is more variation both in shape and in percent contribution of land cover from 59 points and down to 7. The topographic position index is the least contributing variable in all models, and the shape of its response curve starts to change from 30 points.

Regarding sensitivity to different samples of same size, there is more variation in variables response curves among the three 30 presence points models than among the three 59 presence points models.

Table 3.6: Results of *Allium ursinum* models with different numbers of presence points. Response curves for the variables bioclim 1 - annual mean temperature, land cover and topographic position index with 1 km diameter. Percent contribution (pct.) for each variable is to the right of the response curve. Training AUC and test AUC for the 20 percent random test sample. Numbers in [] represent different random subsets of the same number.

<i>Allium ursinum</i>								
# of points	Bioclim 1	pct.	land cover	pct.	TpI1	pct.	Training AUC	Test AUC
236		73.6		24.3		2.1	0.908	0.948
118		74.2		24.2		1.6	0.935	0.949
59		74.8		24.8		0.3	0.923	0.941
59 [2]		71.7		27.7		0.6	0.934	0.874
59 [3]		73.2		25.3		1.5	0.933	0.882
30		67.6		30.5		1.9	0.895	0.888
30 [2]		74.6		25		0.4	0.942	0.962
30 [3]		72.7		27.3		0	0.962	0.891
15		66.7		32.7		0.6	0.969	0.978
7		68.7		31.3		0	0.841	0.961

Training AUC values become higher with smaller sample sizes in the *Carum carvi* models, highest is 0.953 and lowest 0.764 (tabel 3.7). There is less of a general trend in the test AUC, these range from 0.660 to 0.996. The biggest gap between test and training AUC occurs at model 46 [2], with a difference of 0.204.

Land cover has a more or less similar shape down to 91 points, then it starts to differ more. This is also true for percent contribution, which ranges between 39 % and 42.2 % above 91 points and from 32.2 % to 100 % contribution from 91 points and down to 6. Bioclim 7 - temperature annual range is the most stable in shape of all the response curves, but percent contribution ranges between 0 % and 30.5 %. Bioclim 8 - mean temperature of wettest quarter response curves varies a lot and percent contribution lies between 0 % and 19.1 %.

There is more variation between models at 91 and 46 points than at 182 points.

Table 3.7: Results of *Carum carvi* models with different numbers of presence points. Response curves for the variables land cover, bioclim 7 - temperature annual range and bioclim 8 - mean temperature of wettest quarter. Percent contribution (pct.) for each variable is to the right of the response curve. Training AUC and test AUC for the 20 percent random test sample. Numbers in [] represent different random subsets of the same number.


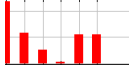


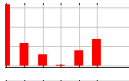








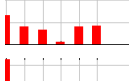


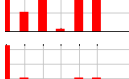
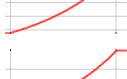

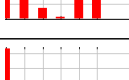
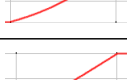
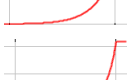
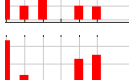
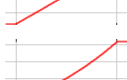

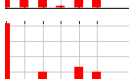


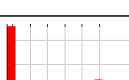
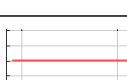

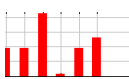







<i>Carum carvi</i>								
# of points	land cover	pct.	Bioclim 7	pct.	Bioclim 8	pct.	Training AUC	Test AUC
729		42.2		28.9		0.9	0.771	0.742
365		41		26.8		4.2	0.787	0.769
182		41.6		27		5.6	0.805	0.845
182 [2]		40.3		9.3		19.1	0.822	0.768
182 [3]		39		26.1		12	0.792	0.862
91		32.2		4.6		8.2	0.784	0.660
91 [2]		41.6		30.5		0.1	0.764	0.845
91 [3]		39.1		7.1		13.8	0.818	0.777
46		53.6		0.1		6.8	0.813	0.693
46 [2]		37.5		24.4		3.1	0.845	0.660
46 [3]		36.8		31.6		0	0.822	0.811
23		55.2		14.2		0	0.862	0.796
11		58.1		20.9		0	0.949	0.846
6		100		0		0	0.953	0.996

As seen in table 3.8 the *Ribes uva-crispa* models have smallest AUC values at the lowest number of presence points, with the exception of training AUC in the 14 point model. This is also the largest gap between training and test AUC, 0.254.

The bioclim 1 response curves keeps the same direction for all models, but the shape changes from a logarithmic to a linear curve in the 7 points model (table 3.8). Lowest percent contribution occurs at 27 points (27.4 %) and highest at 14 points (67 %). Land cover starts to act differently at 109 points with a drop in class 1 - developed land, and after this point there is much variation in most of the classes. Percent contribution range between 8.4 % (7 points model) and 65.9 % (27 points model). Wetness index retain its shape until 27 points, at which point it becomes flat and has 0 % contribution to the models.

The models are more stable at 217 points than at 109 points, at least when it comes to percent contribution of variables.


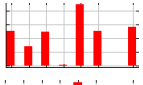
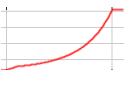

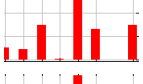





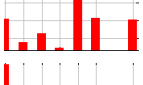


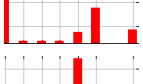






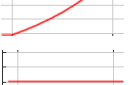
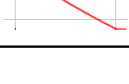
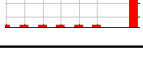
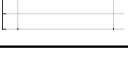
Table 3.8: Results of *Ribes uva-crispa* models with different number of presence points. Response curves for the variables bioclim 1 - annual mean temperature, land cover and wetness index. Percent contribution (pct.) for each variable is to the right of the response curve. Training AUC and test AUC for the 20 percent random test sample. Numbers in [] represent different random subsets of the same number.

<i>Ribes uva-crispa</i>								
# of points	Bioclim 1	pct.	land cover	pct.	Wetness index	pct.	Training AUC	Test AUC
434		59.6		36.5		1.4	0.822	0.787
217		58.2		36.6		3.3	0.833	0.738
217 [2]		56.2		39.1		2	0.827	0.834
217 [3]		59.4		36.3		0.9	0.799	0.787
109		55.1		40.1		0.2	0.778	0.821
109 [2]		63.7		27.2		1.3	0.831	0.839
109 [3]		59.3		37		1	0.796	0.864
54		43		41.8		0.3	0.832	0.800
54 [2]		60.8		36.8		0	0.826	0.586
54 [3]		51.1		45.8		1.1	0.834	0.753
27		27.4		65.9		0	0.766	0.667
14		67		29.7		0	0.911	0.657
7		45.8		8.4		0	0.662	0.674

In the *Rubus chamaemorus* models in table 3.9 training AUC values become higher with fewer presence points, while test AUC are lower except at 7 points. The two extreme lows are test AUCs for 26 and 13 points which are only slightly better than random (0.585) and worse than a random prediction (0.413). The largest gap between AUCs occur at 13 points with 0.386 in difference between training and test AUC.

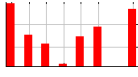

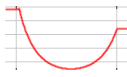
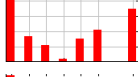


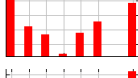


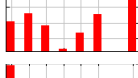




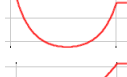








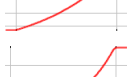
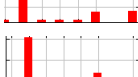


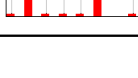
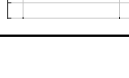
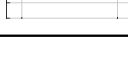
The two bioclim variables are quite constant in shape except for bioclim 7 in the 7 points model, which becomes flat and has 0 % contribution. Land cover vary both in shape and percent contribution, with the biggest change taking place at 52 points. Here class 5 - mires shrink in size, while class 1 - developed land becomes one with highest prediction values. Models with less than 52 points vary even more in this variables response and percent contribution.

Table 3.9: Results of *Rubus chamaemorus* models with different number of presence points. Response curves for the variables bioclim 15 - precipitation seasonality, land cover and bioclim 7 - temperature annual range. Percent contribution (pct.) for each variable is to the right of the response curve. Training AUC and test AUC for the 20 percent random test sample. Numbers in [] represent different random subsets of the same number.

<i>Rubus chamaemorus</i>								
# of points	Bioclim 15	pct.	land cover	pct.	Bioclim 7	pct.	Training AUC	Test AUC
837		6.1		51.8		18.5	0.746	0.737
419		3.2		54.1		23.7	0.756	0.711
209		4.9		47		22.7	0.759	0.702
105		4.8		35.9		18.7	0.769	0.629
52		1.1		44.9		15.4	0.773	0.684
26		5.3		43.6		12.1	0.847	0.585
13		14		59.6		11.8	0.799	0.413
7		1		93.7		0	0.936	0.859

Training AUC values for the *Rubus idaeus* models range from 0.701 in the 2953 model to 0.867 in the 12 points model (table 3.10). There is one very low test AUC value, namely 0.145 in the 23 points model. There is generally larger differences between training and test AUC in the lower points models. The most radical changes in response curves of variables take place in the 92 points model, although bioclim 8 changes shape already at 738 points, from unimodal to one sided. Percent contribution of land cover range from 37.7 % to 100 %, Bioclim 1 range from 0 % to 40.9 % and Bioclim 8 range from 0 % to 16.1 % contribution.

Table 3.10: Results of *Rubus idaeus* models with different numbers of presence points. Response curves for the variables land cover, bioclim 1 - annual mean temperature and bioclim 8 - mean temperature of wettest quarter. Percent contribution (pct.) for each variable is to the right of the response curve. Training AUC and test AUC for the 20 percent random test sample. Numbers in [] represent different random subsets of the same number.

<i>Rubus idaeus</i>								
# of points	land cover	pct.	Bioclim 1	pct.	Bioclim 8	pct.	Training AUC	Test AUC
2953		49.8		20.5		13.2	0.701	0.692
1477		52.4		25.2		8.9	0.718	0.744
738		42.5		27.5		17	0.729	0.720
369		37.7		30.6		13.3	0.729	0.752
185		46.4		27.9		8.4	0.729	0.718
92		56.2		28.2		12.5	0.777	0.729
46		55.9		27.1		16.1	0.780	0.643
23		46.7		40.9		3.9	0.817	0.145
12		70.3		0		11	0.867	0.785
6		100		0		0	0.732	0.670

3.3 Preliminary study

There was *Rubus idaeus* present in 11 of 13 visited sites and *Carum carvi* present in 1. Total number of days in the field was 5, and between 2 and 4 sites were visited per day. The original plan was to visit 100 sites, which would have taken approximately 40 days at the same speed. The total driving distance for the 5 days was 856 km. When a cell was visited and the species was not found, this was registered as an absence of the species. This means that there were 2 absences of *Rubus idaeus*, 12 absences of *Carum carvi* and 13 absences of the three remaining species. Table 3.11 shows the number of presences in all the grid cells that were visited, and which predicted probability class the cells were in.

Table 3.11: Number of each study species present in grid cells sorted by prediction value (logistic output) from the Maxent models. Classes are sorted in ascending order of prediction probability, 1 is between 0 and 0.19, 2 is between 0.2 and 0.39 and so forth. A line means that no cells were visited within that prediction interval for the species.

Class	<i>A. ursinum</i>	<i>C. carvi</i>	<i>R. uva-crispa</i>	<i>R. chamaemorus</i>	<i>R. idaeus</i>
1	0	0	0	0	-
2	0	0	0	0	-
3	-	1	0	0	3
4	0	0	-	0	8
5	0	-	-	0	-

Chapter 4

Discussion

Crop Wild Relative (CWR) populations in Norway could be at the northern extremity of the species range. This means that potentially unique traits could be present in these populations, making them especially important to conserve. In this part there will be a discussion of the results of Species Distribution Modelling (SDM) with presence only GBIF-mediated data sampled into categories of different age and different numbers of points. There will also be a discussion of the results from the preliminary field test to collect presences and absences from the study area.

4.1 Age of presence points

The tests indicate that the age of records have an effect on the resulting SDMs. Models differ in AUC, response curves and percent contributions among the three age classes. The hypothesis that is relevant for this part of the thesis is this:

Hypothesis 1:

H0 - The Maxent distribution models will be consistent when using presence-only points from different time periods if the number of presence points and all settings are kept the same.

H1 - The Maxent distribution models will be inconsistent when using presence-only points from different time periods if the number of presence points and all settings are kept the same.

The SDMs of ramsons have high AUC values in all year classes (table 3.1). A large number of background points that are easily distinguished from the presence points will yield high AUC-values (Merow *et al.*, 2013). Since ramsons is a plant with quite distinct habitat needs (moist, nutritious

forests along the coast) the presences will probably be easy for the model to distinguish from the 10 000 background points since they will have a large probability of being different from these requirements. Land cover categories change between year classes, in the before 1950-model developed land and mires have higher relative predicted probabilities, in the 1950 - 2000-model it is fresh water and forest and in the after 2000-model it is forest and open (table 3.1). This might reflect a change in land use over time, since the land cover variable show recent conditions, what is now developed land might have been a forest before 1950. A study by Sang *et al.* (2014) found that agricultural land in Norway change continuously. That would explain some of the changes, but the large prediction value for the open category is a bit surprising. This could simply be a random trend that comes from the low number of presences used in this model (49).

To illustrate some visual differences among models, figure 3.1 was included in the results. Here we see that the oldest model yield a less distinct model, with more areas being lumped together in the same class.

When modelling with presence points from different time periods, the models with data from after 2000 were used to select which variables that were used in all models for that species. This could explain why the same variables are not equally important in the different models, like we see with caraway in table 3.2. The land cover variable has the largest percent contribution in the after 2000-model, but not in the before 1950-model, and it is not the same land cover types that are important. In the before 1950 and the 1950 - 2000 models, category 8 - open has the highest relative probabilities of presence (table 3.2). After 2000 the highest category changes to 2 - agricultural land. There are several possible explanations for this: the land use might have changed, so that there is more developed and agricultural land in areas that used to be open. It is also possible that there is a spatial bias in where the caraway has been gathered, this is highly likely since the sample size was particularly low for caraway, only 24 points were used in each model (table 2.5). Bio7 - temperature annual range and bio8 - mean temperature of wettest quarter also change shape between models, and it is not likely that the caraway changes how it responds to different temperatures during the last 100 years (Wiens *et al.*, 2010). It is far more likely that the small data sets used for each model means that different responses to climate has been picked up by the model as important for caraway.

The choice of year class will affect the models, and pooling more data together would perhaps yield different results. For instance in the *Carum carvi* models there were only 24 presence points available in the before 1950 model (table 2.5), and this makes it hard to say whether age of points or the small number of points had most influence on the differences observed between year classes.

What seems clear is that in most cases the older data tell a different

story than the newer data, which could mean that the species distribution has changed over time. However, this could be an example of sampling bias, that the different data sets had different sampling design. Phillips *et al.* (2016) aim to find suitable areas for conservation, so using more recent data would be advisable. This is further supported by the recent development in GPS technology and the availability of such equipment, which probably makes newer data more reliable when it comes to sampling precision.

Based on the results from the SDMs of different age classes I would reject the H0 in hypothesis 1, and conclude that age of presence points affect the models. However, there are reasons to be skeptical of these results regarding some of the species as there were few presence points available. Also the variables that were used were chosen based on the newest models, so with relevant variables for the older models results might have been different.

4.2 Number of presence points

Results of the SDM on decreasing number of presence points clearly show that all the parameters included differ when presence points decrease. There are also clear differences between the models made with different subsets of the same number of presences. The two hypothesis that will be relevant in this section are these:

Hypothesis 2:

H0 - The Maxent distribution models will be consistent when different numbers of random presence-only points are used and all settings are kept the same.

H1 - The Maxent distribution models will be inconsistent when different numbers of random presence-only points are used and all settings are kept the same.

Hypothesis 4

H0 - There will not be differences between Maxent models made with random samples of presence only points of the same size.

H1 - There will be differences between Maxent models made with random samples of presence only points of the same size.

When sample size is decreased, the models generally become less similar in regards to the variables response curves.

Some of the models were tested with different random subsets of the same number of points, and this could be a useful method to detect when the models become unstable. Once an "unstable" number of points is found, it is possible to check the interval between the last stable point and the first unstable one to get a more fine tuned result. One could for instance do several runs with the midpoint between the stable and unstable numbers. In the case of *Allium ursinum* (table 3.6) this would be 89 presence points, which is between 118 (last stable) and 59 (first unstable). An even better approach would be to do a larger bootstrapping (random sampling with replacement) study of some of the sample sizes to get a better picture of the stability within a sample size. This however, was beyond the scope of this thesis.

In a study by Hernandez *et al.* (2006), they found that model accuracy increased with larger sample sizes. Still, Maxent was the method that performed best at even very small sample sizes (5 and 10). They also found that species that are more widespread in both in geographical and environmental space are more difficult to model than more restricted species (Hernandez *et al.*, 2006). This can also be seen in this study, where both of the widespread and common species *Rubus idaeus* and *Rubus chamaemorus* generally had lower AUC values than *Allium ursinum* which is more restricted to nutritious forests along the coast (Lid and Lid, 2005).

With lower sample sizes there is more variation between models made with different random subsets of the same number. The number of points required to make stable models is not the same for all the species. For ramsons, 59 points can produce quite stable models, while at 30 points there is clearly more variation in AUC, response curves and percent contribution of variables (table 3.6). Caraway models are more stable at 182 points than at 91, and even more unstable at 46 points (table 3.7). Gooseberry can produce quite stable models at 109 points, but at 59 points AUC, response curves and percent contribution vary a lot, lowest test AUC is 0.586 (table 3.8).

Based on the results of the SDM on decreasing numbers of presences, the H0 of hypothesis 2 should be rejected. Models clearly change when numbers are decreased, and there is more variation among the models the lower the sample size gets. There is also variation within a sample size, so H0 of hypothesis 4 should also be rejected.

4.3 Environmental variables

All environmental variables will have uncertainties and errors, that can affect the the quality of models. The variables used in this study have many uncertainties connected to them. Firstly, all the Bioclim variables from the WorldClim database were made by interpolating monthly precipitation,

and mean, minimum and maximum temperature from a large number of weather stations globally (Hijmans *et al.*, 2005)¹. There was not a high density of the weather stations they used for the interpolation in Norway, so we do not know if the data for Norway will be accurate. In addition another interpolation was made to obtain the 100 m grid size (kriging) so this adds even more uncertainty to the data.

Another important questions is: Do we have the relevant variables for the species we want to model?

For *Allium ursinum*, a variable like soil quality could perhaps give more accurate models, since it is known to grow in nutritious soils.

Cloudberry SDMs would probably have benefited from a distinction between different mires in the land cover variable, so that alkaline fens with little cloudberry could be separated from more cloudberry-rich types of mires.

The land cover category 8 - open can represent a variety of different land cover types, like exposed bedrock, outfield pastures and meadows, shrub land and gravel dominated beaches. This variable should have had more classes, so that a better distinction of the species niches could be modelled. Caraway SDMs would probably have been better, since the open category is important for this species.

Since gooseberry is a naturalized species that has spread from gardens, a variable connected to where people live could be useful. The land cover variable includes category 1 - developed land, but this signal is not apparent in the before 1950 model (table 3.3), only in the 1950 -2000 model is this the category with the highest predictions.

4.4 Model evaluation

In order to evaluate the models, test and training AUC was used. When dealing with presence-only data, the AUC value only reflects how well the models distinguish between presence and background data (Merow *et al.*, 2013). This may not be a good reflection of model quality, since the background locations are chosen randomly from the entire study area and thus contain both presences and absences. In this study, default maximum number of background points were used, which is 10 000. When the sample size is low, training AUC will often become higher, because it is easy for the model to distinguish between the background points and the few presences. This does not mean that the model is better at making predictions, instead it is probably over-fitted to the training set (Halvorsen

¹WorldClim version2 is now available at <http://worldclim.org/version2>, but it was not used in this thesis because it was not available at the time of variable preparations.

et al., 2015). When the test AUC is much lower than the training AUC, this is also an indication of over-fitting.

There are other inbuilt evaluation methods in Maxent that could be used, like jackknife of variable importance and permutation importance.

Another way to evaluate Maxent models is to set a threshold value and let everything above it be categorized as a predicted presence and everything below as an absence. Then you can examine if two models predict the same grid cells as presences or not. This is another way to quantify differences between models. Merow *et al.* (2013) argue against the use of threshold values, "Thresholding is problematic because choosing biologically meaningful thresholds may depend on prevalence or population density, which is typically unknown" (Merow *et al.*, 2013, p. 1067). Also Guillera-Aroita *et al.* (2015) advice against the use of conversion into predicted presence or absence based on a threshold, because continuous outputs provide richer information.

Another approach to model evaluation is to do a correlation test, such as Kendall, between predictions from two models. The problem with this approach is that equal weight is given to all different predictions, and there will be many very low prediction values and fewer high values (R.Halvorsen, personal communication, 2017-05-18).

4.5 Species traits

Species traits can have an effect on SDMs, for instance, Hanspach *et al.* (2010) found that species with a short life span and living in human disturbed habitats had lowest model performance. The relevant hypothesis in this section will be this:

Hypothesis 3

H0 - There will not be species trait specific differences between Maxent models.

H1 - There will be species trait specific differences between Maxent models.

Ramsons has a more narrow distribution than the other species, has clear ecological requirements like nutrient rich forest floors, and it is restricted to the coast. Earlier studies have shown that species with small geographic ranges and narrow environmental tolerance give more accurate models than generalist species (Hernandez *et al.*, 2006).

Both cloudberry and wild raspberries are widespread and common

species in Norway, and according to Hernandez *et al.* (2006) should therefore be harder to model accurately.

Caraway is a biannual plant, while all the other species are perennial. This could lead to it being harder to model, since it will move around more and not necessarily be found in the same place year after year. The study by (Hanspach *et al.*, 2010) also found that short lived species had lower model performance.

To test this hypothesis of species traits a more rigorous testing regime would have to be applied. Since this was outside the scope of this thesis, there can only be some general observations about species life-history traits. The most clear trend that is seen, is that the widespread and common species cloudberry and wild raspberry have lower AUC-values than ramsons which has more narrow habitat demands. This is in line with previous studies.

4.6 Field work

The preliminary field test that was conducted during the summer of 2015 show that gathering independent, unbiased evaluation data for SDM on a regional scale is a challenge.

Modelling species distributions can also be seen as modelling a species fundamental niche (the set of all conditions that allow for its long term survival) from its realized niche (the subset of the fundamental niche that it occupies) (Phillips *et al.*, 2006). In this respect it is important to have enough variation in environmental conditions in order to capture the breadth of the species fundamental niche, and thus a diverse and large study area is advisable. The problem with a large study area is that it is very challenging to make a good field evaluation of the models. Since Phillips *et al.* (2016) has modelled all of Norway with a grain size of 10 X 10 km in their Maxent-model, ideally we would want to check these models with test data on this scale, and from the entire range of environmental conditions for all the 204 priority CWR in all of Norway. This is of course an impossible task. Therefore it was decided to use a smaller geographical area, that still contained a lot of environmental variation, and also to use a smaller grain size of 100 X 100 m.

There were some challenges with gathering randomized data from field work. First of all the study area was large and there were lots of cells. In order to not recreate the same biases that the GBIF-dataset probably has, visited cells should be selected at random. But we also wanted to visit sites that represented the entire range of environmental variation in the area. A compromise was made by using the first round of models to select cells with different predictions from low to high. We then expected to find the

species in cells with high predictions and not in cells with low. The result was that all the cells that had *Rubus idaeus* in them had medium to high predictions (between 0.40 and 0.79) and the one cell with *Carum carvi* had a medium high prediction (0.44) in the *Carum carvi* model.

Using the models made with GBIF-data to select where to collect a new sample is of course not optimal since the new data then will be semi-independent on the GBIF-data, but it was seen as a necessary compromise because randomly chosen cells from the whole ecological space would probably yield a lot of absences and very few presences.

Another possible approach to field validation on this scale, could be some kind of block-design where smaller areas within the region was sampled more thoroughly, as travel distances were long in this study. If smaller areas with a representative gradient of natural variation could be studied you could get a larger test set with less effort.

Based on my experiences, I would say that gathering a good test set for SDM require careful planning of the study design. When it comes to modelling all of mainland Norway, it is likely that it is not economically viable to gather this kind of data. What is important is to take all the inherent weaknesses of the data and modelling method into account when making conservation plans, to avoid making overoptimistic assumptions based on SDM.

4.7 Conclusion

The GBIF-platform is an important factor in addressing the issue of conservation of plant genetic resources like CWR. Together with different SDM methods it is possible to make good conservation plans on both the national and global level. Keeping in mind the inherent biases of GBIF-data is therefore necessary to make good predictions of species distributions and thereby ensuring effective conservation of CWR. This thesis has highlighted some of the issues with spatial, temporal and species bias in GBIF-data. The age of presence records had an effect on the SDMs, and number of presences also changed model outputs. A proposed solution for finding the minimal number of points required to make stable models was presented. Being aware that the biases exist is the first step towards finding solutions to deal with it, and many solutions have been suggested by others (some suggestions for Maxent modelling is here: (Merow *et al.*, 2013)).

Bibliography

- Beale, C. M. and J. J. Lennon (2012). Incorporating uncertainty in predictive species distribution modelling. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 367 (1586), pp. 247–258. DOI: 10.1098/rstb.2011.0178.
- Boehner, J. and O. Conrad (2001). *SAGA-GIS Module Library Documentation (v2.1.4)*. URL: http://www.saga-gis.org/saga_tool_doc/2.1.4/ta_hydrology_15.html (visited on 2017-05-23).
- Bryn, A., H.-P. Kristoffersen, M. Angeloff, I. Nystuen, L. Aune-Lundberg, D. Endresen, C. Svindseth, and Y. Rekdal (2015). Location of plant species in Norway gathered as a part of a survey vegetation mapping programme. *Data in Brief* 5, pp. 589–594. DOI: doi.org/10.1016/j.dib.2015.10.014.
- Castañeda-Álvarez, N. P., C. K. Khoury, H. A. Achicanoy, V. Bernau, H. Dempewolf, R. J. Eastwood, L. Guarino, R. H. Harker, A. Jarvis, N. Maxted, *et al.* (2016). Global conservation priorities for crop wild relatives. *Nature plants* 2, p. 16022.
- Dempewolf, H., G. Baute, J. Anderson, B. Kilian, C. Smith, and L. Guarino (2017). Past and Future Use of Wild Relatives in Crop Breeding. *Crop Science*. DOI: 10.2135/cropsci2016.10.0885.
- Dempewolf, H., R. J. Eastwood, L. Guarino, C. K. Khoury, J. V. Müller, and J. Toll (2014). Adapting Agriculture to Climate Change: A Global Initiative to Collect, Conserve, and Use Crop Wild Relatives. *Agroecology and Sustainable Food Systems* 38 (4), pp. 369–377. DOI: 10.1080/21683565.2013.870629.
- Elith, J. and C. H. Graham (2009). Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography* 32 (1), pp. 66–77.

- Elith, J., S. J. Phillips, T. Hastie, M. Dudík, Y. E. Chee, and C. J. Yates (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and distributions* 17 (1), pp. 43–57.
- ESRI (2014). *ArcMap 10.3.1*.
- Food and Agriculture Organization of the United Nations (FAO) (2009). *Annex 1 of the International Treaty on Plant Genetic Resources for Food and Agriculture*.
- Google (2015). *Google Maps*. URL: <https://www.google.no/maps/> (visited on 2015-05).
- Graham, J., M. Woodhead, K. Smith, J. Russell, B. Marshall, G. Ramsay, and G. Squire (2009). New insight into wild red raspberry populations using simple sequence repeat markers. *Journal of the American Society for Horticultural Science* 134 (1), pp. 109–119.
- Guillera-Aroita, G., J. J. Lahoz-Monfort, J. Elith, A. Gordon, H. Kujala, P. E. Lentini, M. A. McCarthy, R. Tingley, and B. A. Wintle (2015). Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography* 24 (3), pp. 276–292. DOI: 10.1111/geb.12268.
- Guisan, A. and N. E. Zimmermann (2000). Predictive habitat distribution models in ecology. *Ecological modelling* 135 (2), pp. 147–186.
- Halvorsen, R., S. Mazzoni, A. Bryn, and V. Bakkestuen (2015). Opportunities for improved distribution modelling practice via a strict maximum likelihood interpretation of MaxEnt. *Ecography* 38 (2), pp. 172–183.
- Hanspach, J., I. Kühn, S. Pompe, and S. Klotz (2010). Predictive performance of plant species distribution models depends on species traits. *Perspectives in Plant Ecology, Evolution and Systematics* 12 (3), pp. 219–225.
- Harlan, J. R. and J. M. J. de Wet (1971). Toward a Rational Classification of Cultivated Plants. *Taxon* 20 (4), pp. 509–517.
- Herden, T., B. Neuffer, and N. Friesen (2012). *Allium ursinum* L. in Germany – surprisingly low genetic variability. *Feddes Repertorium* 123 (1), pp. 81–95. DOI: 10.1002/fedr.201200019.
- Hernandez, P. A., C. H. Graham, L. L. Master, and D. L. Albert (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography* 29 (5), pp. 773–785.

- Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* 25 (15), pp. 1965–1978. DOI: 10.1002/joc.1276.
- Høeg, O. A. (1976). *Planter og tradisjon: Floraen i levende tale og tradisjon i Norge 1925-1973*. Universitetsforlaget, Oslo.
- Isaac, N. J. B., A. J. van Strien, T. A. August, M. P. de Zeeuw, and D. B. Roy (2014). Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution* 5 (10), pp. 1052–1060. DOI: 10.1111/2041-210X.12254.
- Jennes, J. (2006). *Topographic Position Index extension for ArcView 3.x*. URL: www.jennessent.com/arcview/tpi.htm (visited on 2017-10-03).
- Kell, S. and N. Maxted (2015). Crop Wild Relative Issue 10.
- Kiviniemi, K. and O. Eriksson (1999). Dispersal, Recruitment and Site Occupancy of Grassland Plants in Fragmented Habitats. *Oikos* 86 (2), pp. 241–253.
- Korpelainen, H., K. Antonius-Klemola, and G. Werlemark (1999). Clonal structure of *Rubus chamaemorus* populations: comparison of different molecular methods. *Plant Ecology* 143 (1), pp. 123–128. DOI: 10.1023/A:1009898209220.
- Korsmo, E. (1954). *Ugras i nåtidens jordbruk*. Oslo, AS Norsk Landbruks Forlag.
- Lid, J. and D. T. Lid (2005). *Norsk Flora*. Det Norske Samlaget.
- Maxted, N., B. V. Ford-Lloyd, S. Jury, S. Kell, and M. Scholten (2006). Towards a definition of a crop wild relative. *Biodiversity & Conservation* 15 (8), pp. 2673–2685.
- Maxted, N., S. Kell, B. Ford-Lloyd, E. Dulloo, and Á. Toledo (2012). Toward the systematic conservation of global crop wild relative diversity. *Crop Science* 52 (2), pp. 774–785.
- Merow, C., M. J. Smith, and J. A. Silander (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography* 36 (10), pp. 1058–1069.
- Microsoft (2010). *Analysis ToolPak for Excel*.

- Norwegian Biodiversity Information Center (NBIC) and GBIF (2007-2017). *Species Map Service 1.6*. URL: artskart.artsdatabanken.no/Default.aspx.
- O'Sullivan, D. and D. Unwin (2014). *Geographic information analysis*. John Wiley & Sons.
- Parra-Quijano, M., J. Iriondo, and E. Torres (2012). Improving representativeness of genebank collections through species distribution models, gap analysis and ecogeographical maps. *Biodiversity and Conservation* 21 (1), pp. 79–96.
- Phillips, J., Å. Asdal, J. Magos Brehm, M. Rasmussen, and N. Maxted (2016). In situ and ex situ diversity analysis of priority crop wild relatives in Norway. *Diversity and Distributions* 22 (11), pp. 1112–1126. DOI: 10.1111/ddi.12470.
- Phillips, S. J., R. P. Anderson, and R. E. Schapire (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling* 190 (3–4), pp. 231–259. DOI: 10.1016/j.ecolmodel.2005.03.026.
- Phillips, S. J., M. Dudík, and R. E. Schapire (2011). *Maxent software for modeling species niches and distributions (Version 3.3.3k)*. URL: www.cs.princeton.edu/~schapire/maxent/ (visited on 2014-03-24).
- Rapp, K. and I. Martinussen (2002). Breeding cloudberry (*Rubus chamaemorus* L.) for commercial use. *Acta horticultrurae*.
- Rapp, K., S. K. Næss, and H. J. Swartz (1993). Commercialization of the Cloudberry (*Rubus chamaemorus* L.) in Norway.
- Redden, R., S. S. Yadav, N. Maxted, M. E. Dulloo, L. Guarino, and P. Smith (2015). *Crop wild relatives and climate change*. John Wiley & Sons.
- Sang, N., W. Dramstad, and A. Bryn (2014). Regionality in Norwegian farmland abandonment: Inferences from production data. *Applied Geography* 55, pp. 238–247.
- Schröder, S., A. Kortekamp, E. Heene, J. Daumann, I. Valea, and P. Nick (2015). Crop wild relatives as genetic resources — the case of the European wild grape. *Canadian Journal of Plant Science* 95 (5), pp. 905–912.
- Skard, O. (2007). *Jord- og hagebruksvekster - røtter i kulturhistorien*. Tun Forlag AS.

- Telenius, A. (2011). Biodiversity information goes public: GBIF at your service. *Nordic Journal of Botany* 29 (3), pp. 378–381. DOI: 10.1111/j.1756-1051.2011.01167.x.
- Vincent, H., J. Wiersema, S. Kell, H. Fielder, S. Dobbie, N. P. Castañeda-Álvarez, L. Guarino, R. Eastwood, B. León, and N. Maxted (2013). A prioritized crop wild relative inventory to help underpin global food security. *Biological conservation* 167, pp. 265–275. DOI: 10.1016/j.biocon.2013.08.011.
- Wiens, J. J., D. D. Ackerly, A. P. Allen, B. L. Anacker, L. B. Buckley, H. V. Cornell, E. I. Damschen, T. Jonathan Davies, J.-A. Grytnes, S. P. Harrison, B. A. Hawkins, R. D. Holt, C. M. McCain, and P. R. Stephens (2010). Niche conservatism as an emerging principle in ecology and conservation biology. *Ecology Letters* 13 (10), pp. 1310–1324. DOI: 10.1111/j.1461-0248.2010.01515.x.
- Yesson, C., P. W. Brewer, T. Sutton, N. Caithness, J. S. Pahwa, M. Burgess, W. A. Gray, R. J. White, A. C. Jones, F. A. Bisby, and A. Culham (2007). How Global Is the Global Biodiversity Information Facility? *PLOS ONE* 2 (11), pp. 1–10. DOI: 10.1371/journal.pone.0001124.