A new method to quantify the accuracy of classification and spatial delineation in land cover maps

Hannah Emma Emilie Strømsrud Haga

Master of Science thesis
Biology: Ecology and Evolution
Department of Biosciences and the Natural History Museum

UNIVERSITY OF OSLO

18.04.2018

A new method to quantify the accuracy of classification and spatial delineation in land cover maps

Hannah Emma Emilie Strømsrud Haga
hannahemmaemilie@gmail.com
+47 47328291

Supervisors:

Anders Bryn
anders.bryn@nhm.uio.no

Heidrun A. Ullerud
h.a.ullerud@nhm.uio.no

GEco-group
Natural History Museum

Master of Science thesis

Biology: Ecology and Evolution

Department of Biosciences and the Natural History Museum

UNIVERSITY OF OSLO

18.04.2018

2018

A new method to quantify the accuracy of classification and spatial delineation in land cover maps

Hannah Emma Emilie Strømsrud Haga

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

Print: Reprosentralen, University of Oslo

Acknowledgements

A special thanks to my supervisors Anders Bryn and Heidrun A. Ullerud for your presence, good mood and guidance during my period as your master student.

I am very grateful for being a part of the social- and professional network at the GEco-group at the Natural History Museum. Thank you for your participation in my field work. A special thanks to Anne-Barbi Nilsen and Eva Lieungh Eriksen for technical support in GIS, R and Excel.

Family and friends, especially Mor and Far. Thank you for always believing in me and encourage me in both good and rough times. A special thanks to my friend Torunn for copy-editing my thesis.

Last, but not least, I thank my fiancée Marius and our son Mikkel. You have filled our home with joy, love and laughter giving me precious moments to unwind and forget the hard work, working with this thesis.

Summary

Land cover maps provide spatial information on the physical cover of the earth. With increasing strain on natural resources, these maps enable knowledge-based nature management. Such maps represent a generalization of nature, and their accuracy is often unknown. There are two main sources that generate inaccuracies in land cover mapping. Classification inaccuracies and spatial delineations inaccuracies. In classification inaccuracies, mappers delineate roughly the same, but assign different ecosystem units. In spatial delineation inaccuracies, mappers assign the same ecosystem unit, but delineate the polygon borders different. The main objective of this thesis is to investigate these main sources of inaccuracies separately, further focusing on ecosystem unit characteristics and biome complexity. To investigate biome complexity, four different biomes were mapped and compared.

A new method presented here, called the ABC-method, can investigate these main sources of inaccuracies separately, by comparing the results obtained by different mappers with a "true map". The study was comprised of two parts, both were executed in Ringsaker municipality in south-east Norway. Part one included the making of the "true map" as a reference. In this study, a consensus map, was used as a reference for a "true map". The consensus map was made when ten mappers came to an agreement on classification and spatial delineation. Thereafter, the study sites were partitioned into three sub-areas and the content was adapted according to the ABC-method. One quarter included polygons without classification (sub-area A), one quarter included classified points without polygon borders (sub-area B) and the last two quarters was without any information (sub-area C). In part two of the study, observers mapped the ABC-partitioned study sites.

Pairwise comparisons showed that the main source of inaccuracy is due to differences in spatial delineation (58.1% accuracy). Classification variation also has inaccuracies (71.8% accuracy), but less prominent. When deviating from consensus, ecologically related units are most frequently chosen. The units that were most frequently mixed varied in lime richness, drought risk, units that relied on estimation of species cover, units that were defined based on type of extensive land use, rather than species compositions and units in late succession states.

There is variation among biomes when it comes to mapping accuracy, both in spatial delineation and classifications. Some biomes are more difficult to map than others and display both inaccuracy in spatial delineation and classification. Ecosystems that are more challenging to map have a greater deviation in mean ecological distance per polygon. System complexity strongly influences the mapping accuracy.

Based on the present findings, further research is needed to completely separate the effects of classification- and spatial delineation. Use of the ABC method will aid in improving the understanding of some of these effects and will probably help us to guide mappers better, and could subsequently lead to lowered inconsistencies and higher accuracy. Thus, enabling more knowledge-based management-decisions.

Keywords: Classification, spatial delineation, accuracy, consistency, land cover mapping, fieldwork, vegetation ecology, wall-to-wall mapping

Table of contents

Acknowledgements V

Summary VI

Table of contents VIII

1 Introduction 1

2 Materials and methods 5

2.1 Study area 5

2.1.1 Physical location of area 5

2.1.2 Climate and nature 5

2.1.3 Historic and current land use 6

2.2 Study design 6

2.2.1 Part one 7

2.2.2 Part two 9

2.2.3 Observers experience level 9

2.2.4 Calibration 11

2.2.5 Classification system 12

2.3 Field method 12

2.4 Data management and corrections 13

2.4.1 Accuracy in classification and spatial delineation 13

2.4.2 Ecosystem units' characteristics 14

2.4.3 Variation among biomes in mapping accuracy 15

3 Results 16

3.1 Classification and spatial delineation accuracy 16

3.2 Ecosystem units' characteristics 20

3.3 Biome complexity 24

3.3.1 Heat maps: sub-area A 25

3.3.2 Heat maps: sub-area B 27

4 Discussion 29

4.1 A new method to separate the main inconsistencies in mapping 29

4.2 Robustness with multiple observers 30

4.3 Classification accuracy 30

4.4 Spatial delineation accuracy 31

4.5 Ecosystem units' characteristics 33
4.6 Biome complexity 36
4.7 Uncertainties in this study 37
4.8 Further studies 39
4.9 Conclusions 40
References 41
Appendix 1: Vegetation maps of study sites 45
Appendix 2: Aerial photos 46

Figure 1: Location of study area and study sites. 5
Figure 2: Study design: Part one - Consensus map and ABC-partitioning 8
Figure 3: Possibilities when partitioning according to the ABC-method 8
Table 1: Experience level of observers from part one. 10
Table 2: Experience level of observers from part two. 11
Figure 4: Major type diagram for Forest biome (T4) adapted to scale 1:5 000 with major type adapted levels (numbers 1-4). Ecosystem units adapted to a scale of 1:5 000 have descriptive names and an added prefix code (capital C). 14
Table 3: Classification accuracy (given in percent), standard deviation and confidence interval 17
Table 4: Paired significance test (students t-test) with α=0.05 and Bonferroni-adjusted α = 0.0083 for all combinations of biomes. 17
Table 5: Spatial delineation accuracy (given in percent), standard deviation and confidence interval. 18
Table 6: Paired significance test (students t-test) with α =0.05 and Bonferroni-adjusted α = 0.0083 for all combinations of biomes. 18
Table 7: Overall accuracy (given in percent). 19
Figure 5: Boxplot displaying variation in polygon size in sub-area B. Lower case letters are added as a suffix where there are multiple points classified as the same ecosystem unit 20
Table 8: Polygon characteristics and mean ecological distance in sub-area A. 21
Figure 6: Frequency distribution of ecological distance units in all four biomes combined. 22
Figure 7: Frequency distribution of ecological distance units in mountain-, agricultural-, wetland- and forest biome. 23
Table 9: The following ecosystem units were least accurately classified, with 0 or 1 observers agreeing with consensus in sub-area A. 24
Figure 8: Comparison of consensus map and heat map from the mountain biome in sub-area A 25
Figure 9: Comparison of consensus map and heat map from the agricultural biome in sub-area A 25
Figure 10: Comparison of consensus map and heat map from the wetland biome in sub-area A 26

Figure 11: Comparison of consensus map and heat map from the forest biome in sub-area A 26
Figure 12: Comparison of consensus map and heat map from the mountain biome in sub-area B. Classified points represented by red stars.. 27
Figure 13: Comparison of consensus map and heat map from the agricultural biome in sub-area B. Classified points represented by red stars. 27
Figure 14: Comparison of consensus map and heat map from the wetland biome in sub-area B. Classified points represented by red stars.. 28
Figure 15: Comparison of consensus map and heat map from the forest biome in sub-area B. Classified points represented by red stars. 28

1 Introduction

The earth's surface is changing rapidly. There is high pressure on resources with new land use, urbanization and climate change (Fuchs *et al.*, 2015). Vegetation loss affects biodiversity, climate, soil protection, water circulation and the refilling of groundwater reservoirs (Biondi *et al.*, 2004). To protect these resources, we need to know the distribution and condition of the present vegetation, as well as the impact of natural or human disturbance. Land cover maps are a good source to retrieve complex ecological information (Bryn *et al.*, 2018), as they show the distribution of different land covers in a geographical area, usually adapted to a predefined scale. Such maps are a result of a relationship between two fields of applied research; botany and geography (Küchler & Zonneveld, 1988). Land cover maps have a wide range of applications, both for management and research purposes (De Cáceres & Wiser, 2012). Land cover mapping is often the starting point for these applications (Cherrill & McClean, 1999b). According to Bryn *et al.* (2018), there are five main applications: Describing nature, documentation of natural variation (the presence, extent and status of land covers), management of areas (basis for decisions on use, development, maintenance etc.), research and documentation (red list assessment, modelling etc.) and monitoring changes in nature. Nature management for conservation purposes is probably the single most used application of detailed land cover maps.

Land cover maps depict the physical cover of the earth, and some classes are usually described by classification of vegetation (Aune-Lundberg & Strand, 2017). Typically, the vegetation is classified according to specific physiognomic features (Ihse, 2007) or characteristic groups of species that are found in locations with similar growing conditions (Box & Fujiwara, 2013). Vegetation is affected by different factors such as climate, environmental influences and disturbances (Box & Fujiwara, 2013). A broad concept of land cover is applied regardless of whether these types are defined by vegetational criteria or not (Franklin, 1995; Xie *et al.*, 2008). Systems that classify land cover are often hierarchical, where similar vegetation, biomes or other kinds of land cover are put into classes on different levels of a hierarchy (Cherrill & McClean, 1999a). Many classification systems of land cover, outside strongly human disturbed systems, capture more or less stable entities of either plant communities or ecosystem units that re-appear in specific parts of ecological complex gradients. These can be characterized by species composition, physiognomy, indicator species or a combination of the three (Bryn, 2006). Other criteria that define land cover types besides vegetation can be types affected by human disturbances (for instance infrastructure, buildings, etc.) or natural disturbances (for instance rock slides). Mapping can be done in the field with field pad and aerial photos, by interpretation of aerial photos or by using a variety of remote sensing techniques. Field based land cover maps are made by identifying areas of homogenous land cover (spatial delineation) and assigning these polygons to predefined types (classification) (Cherrill & McClean, 1999a). The classification systems (units) and map generalizations (delineation) should be pre-adapted to a specific resolution through a defined scale intended for the map series (Hearn *et al.*, 2011). Field mapping is a costly and time-consuming method, which is why methods used to model species distribution are being tested (Ullerud *et al.*, 2016). Historically, it has been most common to map vegetation types in Norway, by using aerial photos in the field (Rekdal & Bryn, 2010; Solheim, 1978). In the recent decade, the focus in Norway has shifted from vegetation to ecosystem types (Bryn *et al.*, 2014). The need to map types not solely defined by plants, e.g. coral reefs, as well as the introduction of the Norwegian Biodiversity Act in 2009, have facilitated this shift. An ecosystem unit is defined as a "Uniform environment, including all living organisms and the environmental factors that operate there, or specific types of natural features such as ponds, field islets or the like, as well as special types of geological features» (Ministry of Climate and Environment, 2009) (The definition of an ecosystem type (according to the NiN 2.0 system) does not exist in English, therefore this is my own translation). In this study the observers have mapped ecosystem units, derived from ecosystem types and adapted to scale 1:5 000.

The production of land cover maps require considerable expertise, because of their many features and the maps' many potential usages (Küchler & Zonneveld, 1988). Land cover mapping is not an objective method, there is subjectivity in the mapping process, both in classification and spatial delineation (Pancer-Koteja *et al.*, 2009). Even if there are strict guidelines for mapping, it is a method associated with inconsistencies. It is likely that observers interpret and apply different rules in practice (Cherrill & McClean, 1999b). To implement land cover maps within a wide range of application in a satisfactory manner, high quality environmental data is essential to make appropriate management decisions (Cherrill, 2016; Hunter, 2016). Some ecosystem units are more highly valued than others, and even closely related units can accommodate different groups of animals (Sutherland & Hill, 1995). In studies evaluating the quality of land cover maps, the term "inconsistencies" is commonly used when comparing observers and assess the inter-observer variation, i.e. when two or more observers obtain different results (Morrison, 2016). According to Ullerud (2018), the term "accuracy" can assess the similarity between a "true" land cover map and a single observers' land cover map. Land cover maps need good quality to be trusted by users (Cherrill, 2016). When evaluating the quality of maps, the important map properties are seen in relation to the intended use of the map. It targets the accuracy/and or consistency of a land cover map and whether it contains the information required for the intended application (Ullerud, 2018). Considerable numbers of land cover maps exist, but there is little information on the reliability and quality of these data (Cherrill & McClean, 1995, 1999b; Hearn *et al.*, 2011). All classification and mapping methods lead to an artificial generalization of nature (Green & Hartley, 2000). An ecosystem unit is an abstract ideal; any ecosystem unit drawn up will be an imperfect representation of reality (Pancer-Koteja *et al.*, 2009). Nature is continuous, changes gradually and has diffuse borders that often have similar species composition, which may lead to that observers having to make arbitrary lines (Hearn *et al.*, 2011). This can give rise to inconsistencies (Cherrill & McClean, 1995; Küchler & Zonneveld, 1988).

The main sources that generate uncertainties in field-based land cover maps can be partitioned into two categories; Inconsistencies in classification of land cover and spatial delineation inconsistencies. Inconsistency can be defined as the difference between land cover maps made by different mappers, when all other factors are kept constant (Ullerud, 2018). In classification inconsistencies, observers delineate roughly the same location, but assign different ecosystem unit. In spatial inconsistencies, the observer assign the same ecosystem unit, but delineate the polygon borders different (Cherrill, 2013). Distinguishing between inconsistencies in land cover mapping can be difficult. Inaccurate georeferencing of field observations can give the impression of classification inconsistency when different maps are compared (Cherrill & McClean, 1999b). Inconsistencies can be introduced at any stage of the mapping process (Hearn *et al.*, 2011). There will always be inconsistencies in maps, but it is important to know the nature and scale of these inconsistencies. This study use a "true map" to quantify the maps accuracy. Accuracy can be defined as the similarity between a land cover map and a "true" land cover map (Ullerud, 2018).

Since land cover maps are an artificial generalization of nature, more or less affected by subjective decisions made during field-work, a true land cover map is needed to evaluate the accuracy. To measure consistency among mappers, the same area can be mapped independently by different mappers, and then compared by overlay analysis in GIS (Cherrill

& McClean, 1999a). The degree of similarity between maps can subsequently be calculated (Cherrill & McClean, 1999b). A number of studies have compared maps made by different field-workers and assessed their consistency (Cherrill & McClean, 1995, 1999b; Hearn *et al.*, 2011; Ullerud *et al.*, 2018), but none of the studies have been able to separate the effects of classification from spatial delineation. The main objective of this study is to quantify accuracy in field-based land cover mapping between observers and to develop a new method that enable a separation of the main causes of inaccuracy. The study is designed to answer the following questions: 1) How accurate is the classification? 2) How accurate is the spatial delineation? 3) What characterizes the ecosystem units that are more often inaccurately mapped? 4) Are some biomes more accurately mapped than others, and if so, why?

2 Materials and methods

2.1 Study area

2.1.1 Physical location of area

The study sites are situated at Ringsakerfjellet located in Ringsaker municipality, Hedmark county, south east Norway (**Figure 1**). Ringsakerfjellet is a large mountain biome plateau 700 to 1000 m a.s.l. (Rekdal & Angeloff, 2016). The vegetation is affected by its ecological region and local variation like geology, soil, hydrology and topography (Moen, 1999), and its current and historic usage.

Physical location of study area and study sites

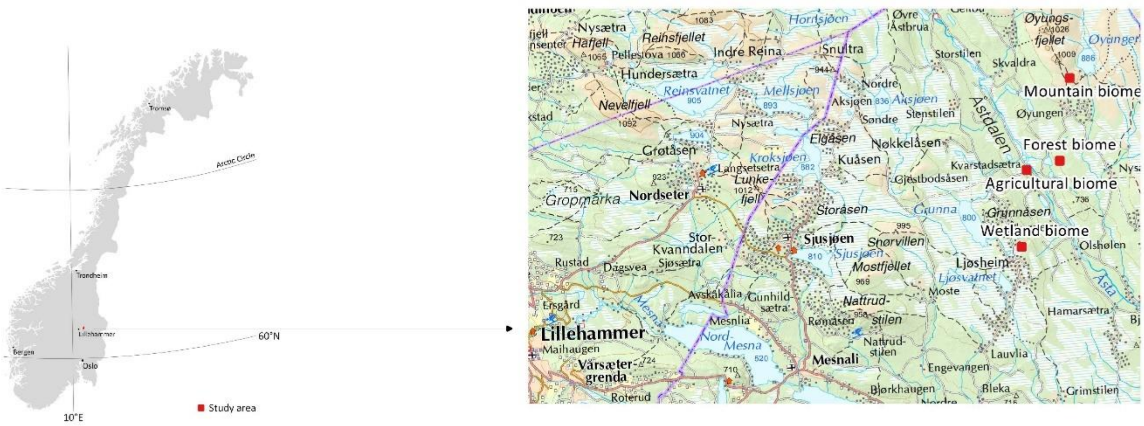


Figure 1: *Location of study area and study sites.*

2.1.2 Climate and nature

Ringsakerfjellet is in the northern boreal vegetation zone and on the border between the indifferent section and the slightly oceanic section. This gives the area low winter temperatures, hot summers and rather little precipitation (Moen, 1999). Wetland biome covers larger areas of this zone than any other. Ringsakerfjellet has 25% wetland, and is also characterized by birch woodland and stunted coniferous woodland (Rekdal & Angeloff, 2016). The growing season is short (150-160 days with temperature equal to or above five degrees Celsius) (Moen, 1999). The area has snow cover approximately 175 to 199 days. Annual precipitation is 1000-1500 mm, and mean annual temperature is 0-2°C (The Norwegian Water Resources and Energy Directorate, 2017). The elevation of the study area is below the climatic forest biome limit. Some of the area is above the empirical forest biome limit, which is at approximately 950 m a.s.l. The empirical forest biome line is lowered because of centuries with summer farming. The bedrock consists of metamorphic sandstone. The bedrock is hard and does not weather easily, thus has a low nutrient level. Areas with higher nutrient levels can occur, because of intrusions with lime-dominated bedrock. The superficial deposits are dominated by till, fluvial deposits with varying grain size and wetland biome. In general, the nutrient levels are low. The soil type is podsoil (Rekdal *et al.*, 2003). The topography of the study sites results in variation in moisture condition, wind exposure and snow cover.

2.1.3 Historic and current land use

The area is influenced by its current and historic land use (Rekdal *et al.*, 2003). There are records of extensive use of the land for summer dairy farming, grazing and hay production since the 16th century. There are records from 1907 of three dairies in the area, but later in the, they were abandoned later in the 20th century (Hasle, 2004). All dairy farms have been abandoned, but grazing is still very common. Sheep and cattle graze mostly in outfield areas (Brodal & Kurud, 2012). The area has a well-developed infrastructure and is accessible by car. The area is extensively used for recreational activities and is the location of approximately 6 000 cabins.

2.2 Study design

Four rectangular study sites, each 100 000 sq.m. and dominated by different biomes, were chosen for land cover mapping. The sites were named after the dominating biome: forest biome, mountain biome, wetland biome and agricultural biome. The location of the four sites were based on an existing vegetation map (Appendix 1) of the area (Rekdal *et al.*, 2003). The vegetation maps displayed several potential sites dominated by the four biomes. The final four sites was determined through a field trip. The criteria for each site were: it should be dominated by one biome but include as much within-biome variation as possible. The size of the sites were based on experience with average 1:5 000 mapping progress (Bryn *et al.*, 2018), but reduced slightly to include the time-lag given by the particular study design. The study consisted of two parts, both mapped according to the same guidelines (Bryn & Halvorsen, 2015).

2.2.1 Part one

In part one (**Figure 2**) the aim was to make a “true” reference map for each study site, to evaluate accuracy in classification and spatial delineation in part two of the study, and partition each site. The field work took place over five days in August 2017. This study used a consensus map as a “true” map. Consensus maps are constructed when several mappers map the same area and come to an agreement on classification and spatial delineation. Maps made by the most experienced mappers were emphasized. The consensus map made in part one was divided into three parts, and the content of each was adapted. One quarter included polygons without classification (sub-area A), one quarter included classified points without polygon borders (sub-area B) and the last two quarters was without any information (sub-area C). This partitioning with its characteristics will henceforth be referred to as the ABC-method. This was executed by ten, mostly experienced mappers. Each mapper was given an equal time-slot for practical mapping. When maps from all the biomes were retrieved, an expert group, comprising the most experienced of the observers, discussed the different maps. Maps made my more experienced mappers were emphasized. A first draft of the consensus map was sent to be assessed by the observers. They had the opportunity to review the draft and provide comments and suggestions on how the maps could be improved further. After their assessment, the expert group reviewed the comments and made changes to the maps. The second draft was sent out for approval by all the participating observers, and the consensus map was completed.

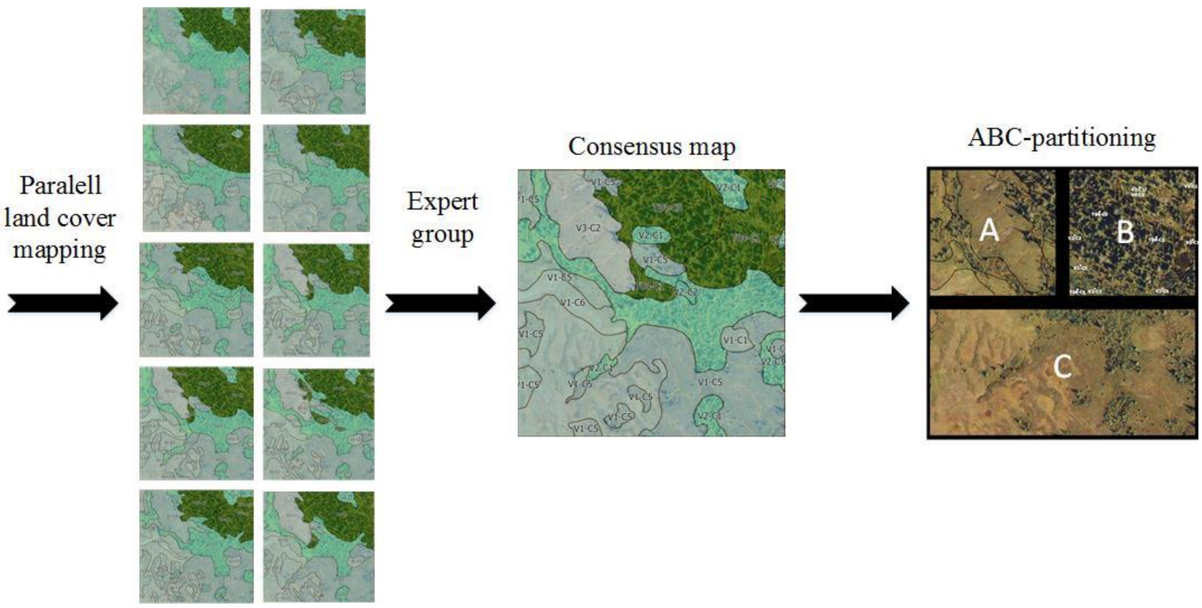


Figure 2: *Study design: Part one - Consensus map and ABC-partitioning*

Partitioning of study sites

When partitioning the study sites into sub-areas according to the ABC-method, there were eight possibilities (**Figure 3**). Before choosing a placement of the different sub-areas, a criterion was given: at least 20% of total amount of different ecosystem units had to be present in each area. Random numbers between 1 and 8 were chosen. The first number was chosen if the criteria were met, if not the next number etc.

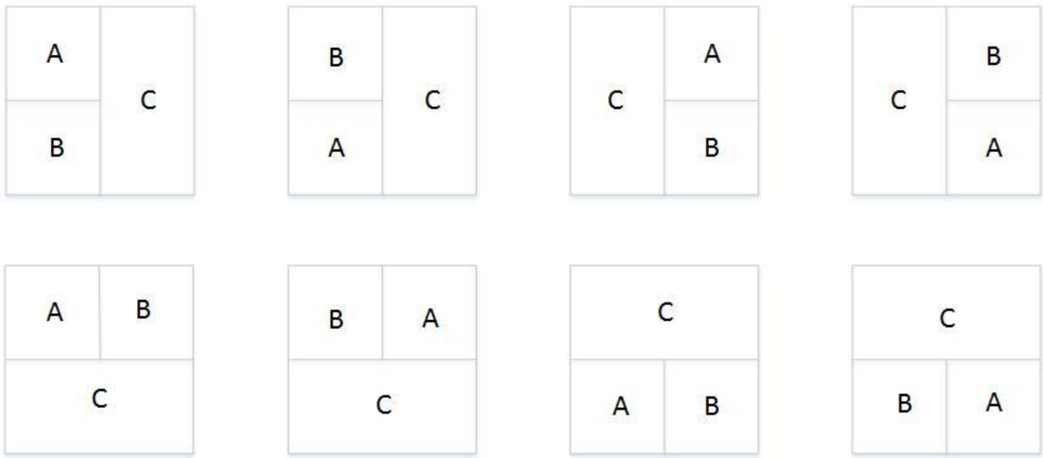


Figure 3: *Possibilities when partitioning according to the ABC-method*

2.2.2 Part two

In part two, the aim was to investigate the accuracy of classification and spatial delineation, using another team of mappers than in part one. The study sites were mapped according to the ABC-method; Sub-area A was mapped first by assigning ecosystem units (classification) to existing polygons, thereafter sub-area B was mapped by delineating a polygon around the given classified points, one point per polygon (spatial delineation). Finally, in sub-area C, the observers made maps the traditional way, carrying out starting off both classification and spatial delineation. Observers received the consensus after completing each sub-area. Part C is not included in the analysis of the study as it is outside the scope of this study. However, this part was important in a training and learning context and was therefore executed by the observers in part two. The observers executing the mapping were 14 Master- and PhD students from a university course in field-based land cover mapping. The observers were divided into seven pairs of observers. In general, the students had limited mapping experience, but some had participated in previous mapping projects. The field work took place over three days in September in 2017. Each pair of observers was assigned with a participant number (corresponding to the pad-id number). They recorded years of mapping experience and amount of time used at each site.

2.2.3 Observers experience level

The observers in this study had varying experience with the NiN mapping system, other mapping systems and botanical surveys. Observers from part one (**Table 1**) had the most experience, with a mean of 2.2 field seasons of mapping experience with NiN. Part two observers (**Table 2**) had mean average of 0.0 field seasons of mapping experience with NiN.

Table 1: *Experience level of observers from part one.*

Part one			
Participant	Number of field seasons with mapping (NiN)	Number of field seasons with mapping (other mapping system)	Number of field seasons with botanical surveys
1	1.0	0.5	0.0
2	5*	0.0	9**
3	2.0	2.0	1.0
4	4.0	20.0	10.0
5	3.0	15.0	27.0
6	1.0	0.0	3.0
7	4.0	1.0	1.0
8	2.0	0.0	45.0
9	2.0	0.0	1.0
10	1.0	0.0	5.0
Mean No. of field seasons	**2.2**	**3.9**	**10.3**

*only in reasearch and only a small part of the field season
**Participant answered 8 to 10 field seasons of experience with botanical surveys, a mean value is used.

Table 2: *Experience level of observers from part two.*

Part two			
Participant	Number of field seasons with mapping (NiN)	Number of field seasons with mapping (other mapping system)	Number of field seasons with botanical surveys
1	0.0	0.0	3.0
2	0.0	1.0	0.0
3	0.0	0.0	1.0
4	0.0	0.0	0.0
5	0.0	0.0	0.0
6	0.0	0.0	0.0
7	0.0	0.0	0.0
8	0.0	4.0	3.0
9	0.0	0.0	1.0
10	0.0	7.0	9.0
11	0.0	0.0	0.0
12	0.0	1.0	0.0
13	0.0	0.0	0.0
14*	NA	NA	NA
Mean No. of field seasons	**0.0**	**1.0**	**1.3**
*Participant did not answer survey			

2.2.4 Calibration

To obtain good quality maps, observers need to be harmonized by calibration (Bryn & Ullerud, 2017). In this study, there were several calibration sessions before both parts of the study. Information on the entire area including bedrock, superficial deposits, ecological region, important species and current and historic usage of area was given. Each field day started with a calibration session in the field, just outside the study sites. The duration was approximately one hour and was led by someone with extensive knowledge of the biome in question, as well as mapping experience. First half of the in-field calibration session covered information about important species (indicators species etc.), ecosystem units and other important factors that aid the distinction of units. The second half covered spatial delineation, including for example how to read the landscape.

2.2.5 Classification system

This study used the NiN-system (Nature in Norway). NiN comprise three main dimensions; scale, types and attributes. This study use the NiN-system adapted to scale 1:5 000 for land cover mapping (Bryn & Halvorsen, 2015). Division of types in NiN is based on how plants respond to environmental variables, and the units of ecological space they represent. The system is hierarchical and comprises three levels: major type group (7), major type (92) and basic type (741). Basic types are combined differently into units depending on the scale in question. Scale 1:5 000 comprises 277 terrestrial and wetland ecosystem units. The other major types groups were not included, since these were not present in the study area. Freshwater was only mapped at major type group level. 41 ecosystem units are defined by other criteria than species composition, such as land use or natural disturbances like for instance rockslides. Ecosystem units are assigned to polygons by looking at the species composition. Each ecosystem unit is described in the mapping manual for NiN (Bratli *et al.*, 2017), including information about the physiognomy, characteristic species, aerial photo characteristics etc. The description also includes species list, which include information on species placement along important gradients, for instance moisture, lime-richness, and drought-risk. These descriptions aid in recognizing the units in field. The attribute system comprises complementary variables that can be used to add extra information that is not described by the ecosystem units. An example of this is what tree species that dominate, and how many percent tree cover there are. This study has not included any complementary variables from the attribute system.

2.3 Field method

In both part one and two mapping was done in the field using portable field pads with GIS version 2.18.14 (QGIS Developement Team, 2018) and by using aerial photos from 1973 and 2016 (Appendix 2). Observers were equipped with field instructions (Bryn & Halvorsen, 2015), a graphical overview of ecosystem units (Bryn & Ullerud, 2017) and descriptions of the ecosystem units (Bratli *et al.*, 2017). Minimum polygon size was 250 sq.m.. The observers (part one and two) were not allowed to exchange information during the mapping. A discussion session was held after each part of the study after the mapping was completed, giving the observers time to discuss and compare maps.

2.4 Data management and corrections

Data-management and analysis were done in QGIS (version 2.18.14), Microsoft Excel and R (R Core Team, 2018). Maps from part one and two were checked for technical errors in QGIS by using the "Topology checker". The following rules were applied for the maps:

- No gaps
- No invalid geometries
- No overlap

A few of the maps had polygons that were not correctly clipped against the buffer frame. These areas were removed by using the geoprocessing tool "difference" in QGIS.

2.4.1 Accuracy in classification and spatial delineation

Classifications- and spatial delineation from the consensus map were compared with the observers' maps, thereby finding the classification- and spatial delineation accuracy. All polygons in part A and B in the consensus-maps were given a unique ID based on site (biome), sub-area and polygon number. To be able to compare polygons from the consensus maps with polygons from the observer map, the unique IDs in the consensus-polygons where transferred to the corresponding polygons mapped by the observers in part A and B. This task was executed in QGIS. Area statistics for polygons with same unique ID were executed for sub-area A and B to quantify the accuracy. Both overall- and pairwise comparison between each observer and the consensus map were executed. The function "Intersect" in QGIS was used for this purpose. Standard deviation and confidence interval was calculated for the pairwise comparison. A paired significance test with a significance level of 0.95 (α =0.05) and the Bonferroni correction was implemented for multiple testing situations (Bonferroni-adjusted α = 0.0083). The area of polygons with equal ID and ecosystem unit were calculated and added up. The results show the overall accuracy. Boxplots were constructed to view the distribution of the polygon sizes in each biome, which give information on the spatial delineation variation among observers.

2.4.2 Ecosystem units' characteristics

Ecological distance (ED) was developed by Eriksen (2017) to quantify deviations in recorded ecosystem units relative to a reference. NiN uses species turnover along gradients. The ED between two units indicates to what degree they have a shared species pool. ED also indicates differences in which structuring processes that defines the units. A higher ED indicates fewer species in common, when observers have registered the same ecosystem unit as consensus the deviations is zero ED. The ED is calculated by counting the number of major type adapted steps along all relevant local complex gradients, by inspecting major type diagrams (example: **Figure 4**) that separate the two ecosystem units in question. Polygons with the same unique ID in consensus- and each of the observers' maps were compared, and the ED relative to consensus was calculated.

Lime richness	4	T4C4 Lime-rich low-herb forest	T4C8 Lime-rich low-herb heather-bilberry forest	T4C12 Lime-rich low-herb heather forest	T4C16 Lime-rich low-herb lichen forest
	3	T4C3 Low-herb forest	T4C7 Low-herb heather-bilberry forest	T4C11 Low-herb heather forest	T4C15 Low-herb lichen forest
	2	T4C2 Intermediate low-herb forest	T4C6 Intermediate low-herb heather-bilberry forest	T4C10 Intermediate low-herb heather forest	T4C14 Intermediate low-herb lichen forest
	1	T4C1 Bilberry forest	T4C5 Heather-bilberry forest	T4C9 Heather forest	T4C13 Lichen forest
Spring water influence 1		1	2	3	4
Main type Forest (T4) diagram		Drought risk			

Figure 4: *Major type diagram for Forest biome (T4) adapted to scale 1:5 000 with major type adapted levels (numbers 1-4). Ecosystem units adapted to a scale of 1:5 000 have descriptive names and an added prefix code (capital C).*

2.4.3 Variation among biomes in mapping accuracy

Heat maps were constructed for each biome to visually display the mapping accuracy. Heat maps display the frequency of observers that have classified the same ecosystem unit as consensus, represented by points with different colors. Regular points with 3 m spacing was used.

3 Results

A total of 56 maps were generated from part two of the study, 28 from each sub-area, 14 from each biome, seven maps from each sub-area in each site.

3.1 Classification and spatial delineation accuracy

The accuracy was calculated for pairwise comparison between each observer and consensus, and the intersection between all observers and consensus (overall accuracy). The pairwise comparison showed greater accuracy than the overall comparison in both sub-areas (**Table 3**, **Table 5** and **Table 7**).

The mean classification accuracy is 71.8%, the results from each biome range from 54.8% in forest biome to 96.7% in mountain biome. Wetland biome has the largest standard deviation. The mean classification accuracy for mountain biome is significantly different from the mean classification accuracy of the three other biomes. Agricultural-, wetland- and forest biome does not have a significantly different mean classification accuracy. When using the Bonferroni correction mountain biome has a significantly different mean classification accuracy than agricultural- and forest biome (**Table 4**).

Table 3: *Classification accuracy (given in percent), standard deviation and confidence interval*

Classification accuracy					
Observer + consensus	Mountain biome	Agricultural biome	Wetland biome	Forest biome	Mean classification accuracy
1+C	99.7	54.4	48.7	35.7	59.6
2+C	100.0	74.0	53.2	67.3	73.6
3+C	93.4	77.3	79.8	74.1	81.1
4+C	94.4	69.4	90.2	42.4	74.1
5+C	97.4	38.7	37.5	47.8	55.3
6+C	95.6	78.0	100.0	80.7	88.6
7+C	96.3	72.0	78.1	35.7	70.5
Mean classification accuracy	96.7	66.3	69.6	54.8	71.8
Standard deviation (population)	2.3	13.4	21.6	17.4	10.7
Confidence interval (0.05)	2.2	12.4	20.0	16.1	9.9

Table 4: *Paired significance test (students t-test) with α=0.05 and Bonferroni-adjusted α = 0.0083 for all combinations of biomes.*

Significance testing (t-test): Classification accuracy				
	Mountain biome	Agricultural biome	Wetland biome	Forest biome
Mountain biome				
Agricultural biome	0.002			
Wetland biome	0.030	0.573		
Forest biome	0.001	0.115	0.151	

The mean spatial delineation accuracy was 58.1%, ranging from 51.9% in agricultural biome to 63.9% in wetland biome. Agricultural biome has the largest standard deviation. The wetland biome has a significantly different spatial delineation accuracy than mountain- and agricultural biome. Mountain-, agricultural- and forest biome do not differ significantly in spatial delineation accuracy. When using the Bonferroni correction the biomes does not differ significantly in mean spatial delineation accuracy. (**Table 6**).

Table 5: *Spatial delineation accuracy (given in percent), standard deviation and confidence interval.*

Spatial delineation accuracy					
Observer + consensus	Mountain biome	Agricultural biome	Wetland biome	Forest biome	Mean classification accuracy
1+C	46.6	58.1	65.2	61.8	57.9
2+C	54.2	29.4	48.4	56.5	47.1
3+C	57.5	85.7	74.4	69.9	71.9
4+C	42.1	31.7	51.0	65.3	47.5
5+C	58.9	52.5	71.9	68.4	62.9
6+C	53.4	53.1	73.3	59.0	59.7
7+C	64.4	53.1	62.8	56.8	59.3
Mean spatial delineation accuracy	53.9	51.9	63.9	62.5	58.1
Standard deviation (population)	7.0	17.3	9.8	5.0	8.0
Confidence interval (0.05)	6.5	16.0	9.1	4.7	7.4

Table 6: *Paired significance test (students t-test) with α =0.05 and Bonferroni-adjusted α = 0.0083 for all combinations of biomes.*

Significance testing (t-test): Spatial delineation accuracy				
	Mountain biome	Agricultural biome	Wetland biome	Forest biome
Mountain biome				
Agricultural biome	0.775			
Wetland biome	0.040	0.034		
Forest biome	0.060	0.142	0.720	

In the overall comparison, there was a higher accuracy in spatial delineation, with a mean overall classification accuracy of 30.5% and a mean overall spatial delineation accuracy of 33.1% (**Table 7**). The results varied between different biomes in the overall comparison. The mean overall classification accuracy ranging from 0% accuracy in forest biome, to 87.8% in mountain biome. The mean overall spatial delineation accuracy did not display the same large variation in this comparison. Values ranged from 27.2% in mountain biome to 39.8% in wetland biome (**Table 7**).

Table 7: *Overall accuracy (given in percent).*

Overall accuracy					
	Mountain biome	Agricultural biome	Wetland biome	Forest biome	Mean accuracy
Classification accuracy	87.8	0.9	33.2	0.0	30.5
Spatial delineation accuracy	27.2	35.1	39.8	30.4	33.1

The constructed boxplots (Figure 5) display the variation in polygon size in sub-area B. The mountain biome has a small range in area of polygons, apart from three polygons. The polygon with the largest range is classified as T31C2 (lime-poor boreal heathlands) and has an area that range from 391.2 m^2 to 11 334.3 m^2. The wetland biome shows the same pattern as the mountain biome, with small range in polygon sizes. Two polygons differ, with a large range in polygons sizes. Both are major type forest. Agricultural biome display little variation in polygon areas among observers. Forest biome has four polygons with large area range. These polygons are classified as T4C5 (heather bilberry forest) and T4C9 (heather forest).

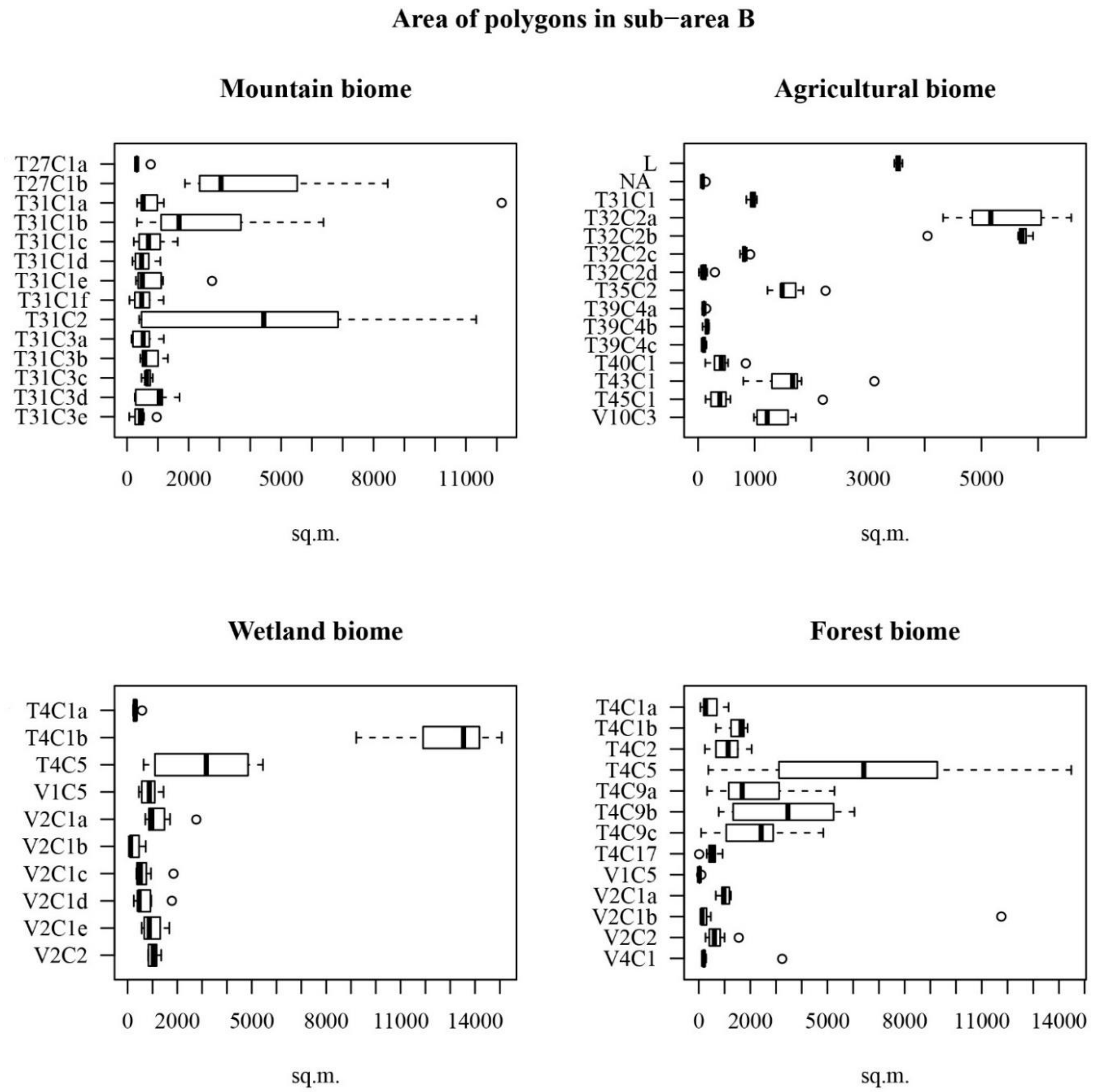


Figure 5: *Boxplot displaying variation in polygon size in sub-area B. Lower case letters are added as a suffix where there are multiple points classified as the same ecosystem unit*

3.2 Ecosystem units' characteristics

To complement the classification accuracy values, the ecological distance (ED) from consensus was calculated for all observers. The results displayed the same trends as the classification accuracy values (**Table 8**), where the forest biome had the lowest accuracy and the mountain biome had the highest. The forest biome had the lowest accuracy with a mean ED of 1.0, whereas mountain biome had an ED of 0.4 (**Table 8**). The Frequency distribution of ED in all biomes showed that the observers chose ecosystem units that were ecologically related to consensus (**Figure 6**). There is a variation between biomes (**Figure** 7). In mountain biome as much as 85.7% of the observations had 0.0 ED, the rest of the observations were spread from 1.0 ED to 6.0 ED. Wetland biome display the same pattern as mountain biome with most of the observations, (71.4%) having 0.0 ED from consensus. Forest biome- and agricultural biome show a more evenly distributed ED than the previous, and fewer observations have 0.0 ED from consensus, respectively 39.8% and 57.1%. Forest biome had the largest amount of registered ecosystem units and number of polygons (**Table 8**).

Table 8: *Polygon characteristics and mean ecological distance in sub-area A.*

Sub-area A				
	Mountain biome	Agricultural biome	Wetland biome	Forest biome
No. Of polygons	20.0	20.0	11.0	23.0
Mean area of polygons (sq.m.)	1125.0	1125.0	2045.5	978.3
Mean ecological distance	0.4	0.9	0.5	1.0

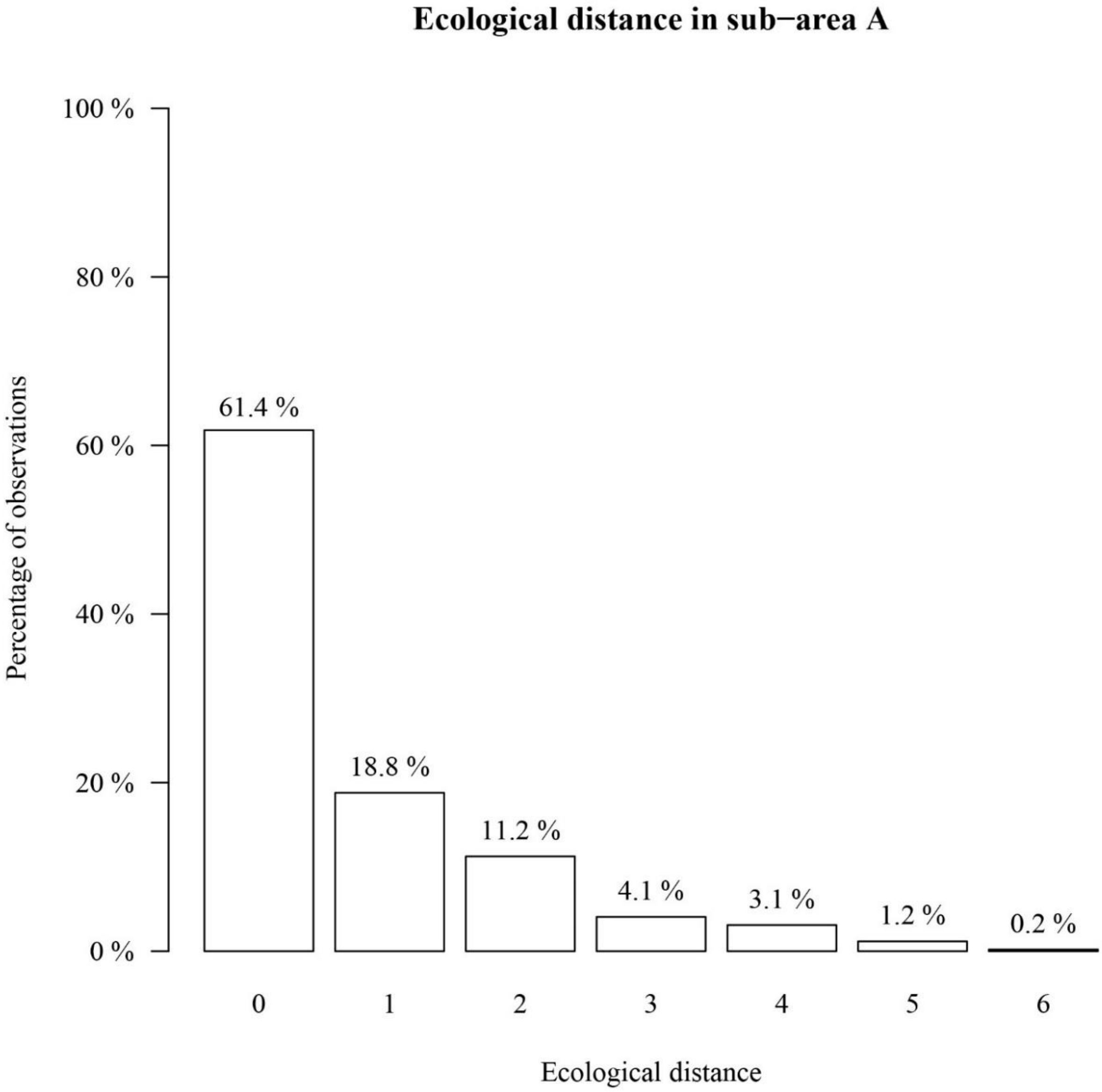


Figure 6: *Frequency distribution of ecological distance units in all four biomes combined.*

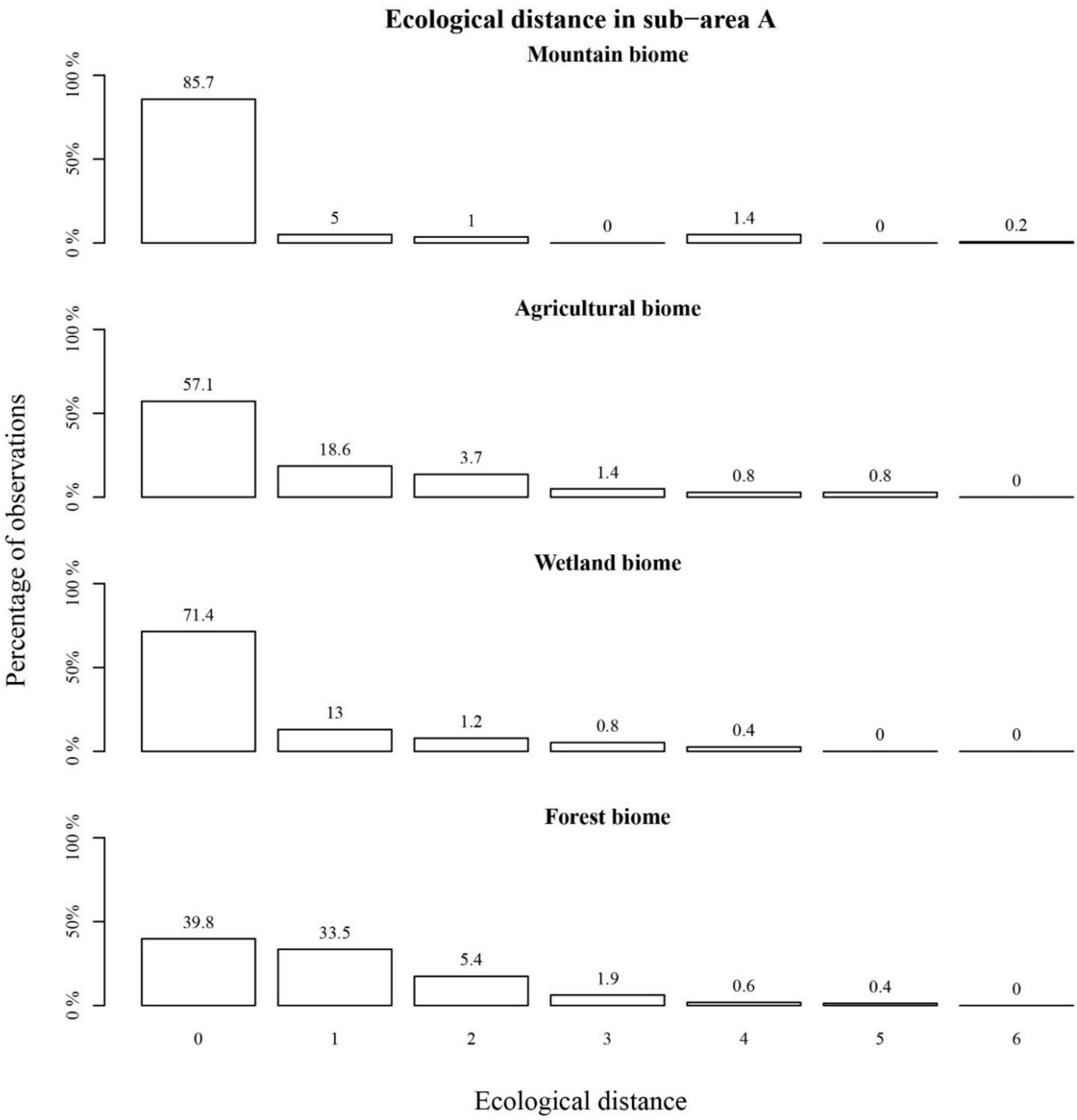


***Figure** 7: Frequency distribution of ecological distance units in mountain-, agricultural-, wetland- and forest biome.*

3.3 Biome complexity

Variation in mapping accuracy varies between biomes. Heat maps visually display this (**Figure 8**, **Figure 9**, **Figure 10**, **Figure 11**, **Figure 12**, **Figure 13**, **Figure 14** and **Figure 15**). The least accurately classified ecosystem units (with 0 or 1 observers agreeing with consensus) can also be examined in **Table 9**.

Table 9: *The following ecosystem units were least accurately classified, with 0 or 1 observers agreeing with consensus in sub-area A.*

Accuracy in classification of ecosystem units		
Biome	Ecosystem unit code	Ecosystem unit name
Mountain biome	V1C5	Very lime-poor wetland edge
Agricultural biome	T31C1	Boreal lee-side
Agricultural biome	T31C2	Lime-poor boreal heathlands
Agricultural biome	T43C1	Lawns, parks, etc.
Agricultural biome	V10C1	Intermediate wetland meadow
Agricultural biome	T32C1	Lime-poor grasslands with moderate management intensity
Wetland biome	V1C6	Lime-poor wetland edge
Forest biome	V1C7	Intermediate wetland edge
Forest biome	V4C4	Highly intermediate and slightly lime rich peatland and spring water
Forest biome	T4C3	Low-herb forest
Forest biome	T4C2	Sparse low-herb forest
Forest biome	V2C1	Lime-poor wetland forest
Forest biome	V1C3	Intermediate wetland

3.3.1 Heat maps: sub-area A

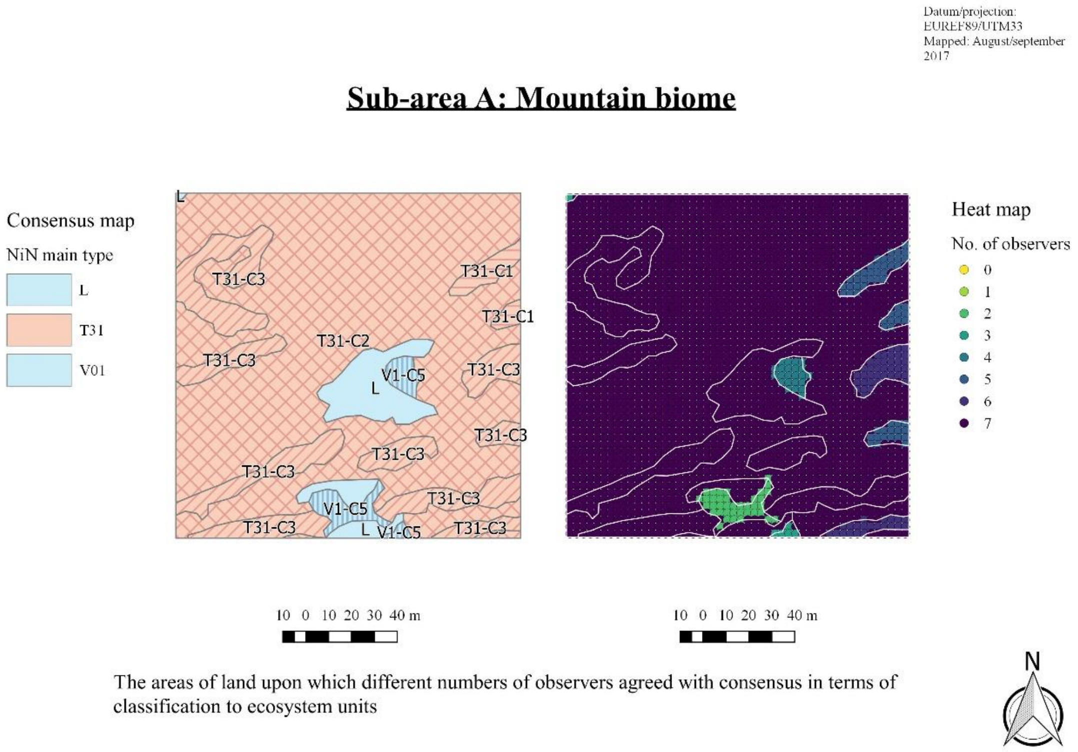


Figure 8: *Comparison of consensus map and heat map from the mountain biome in sub-area A*

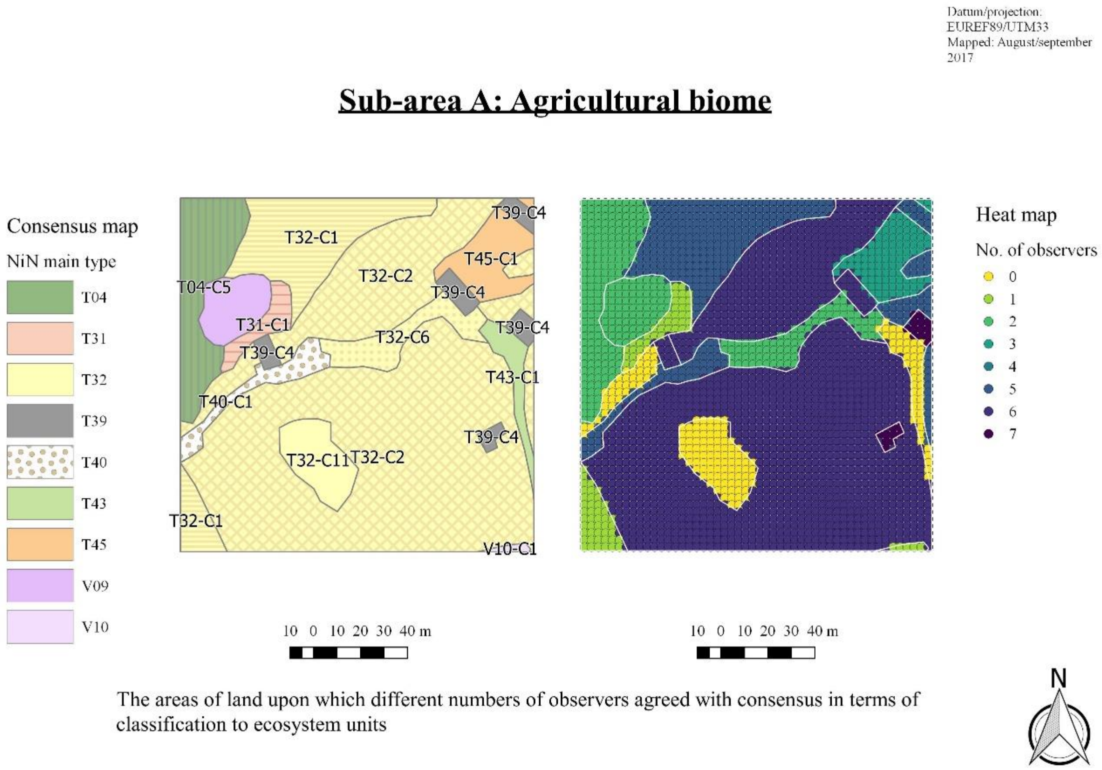


Figure 9: *Comparison of consensus map and heat map from the agricultural biome in sub-area A*

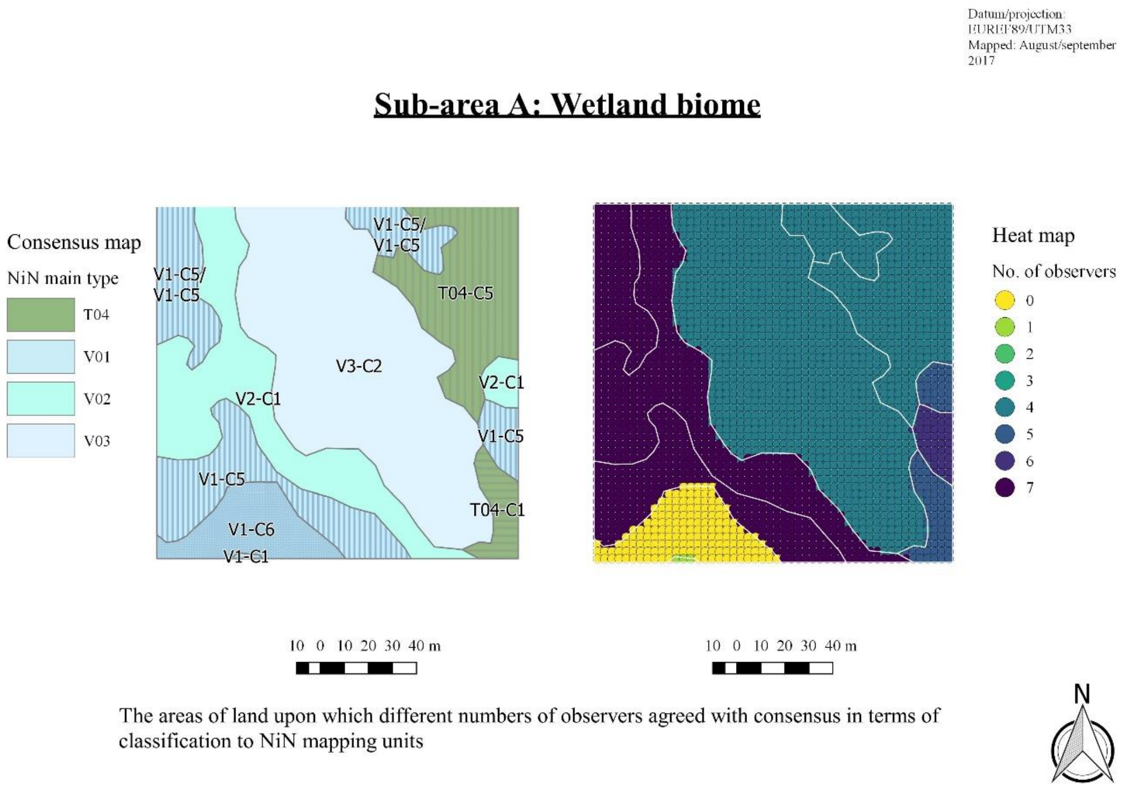


Figure 10: *Comparison of consensus map and heat map from the wetland biome in sub-area A*

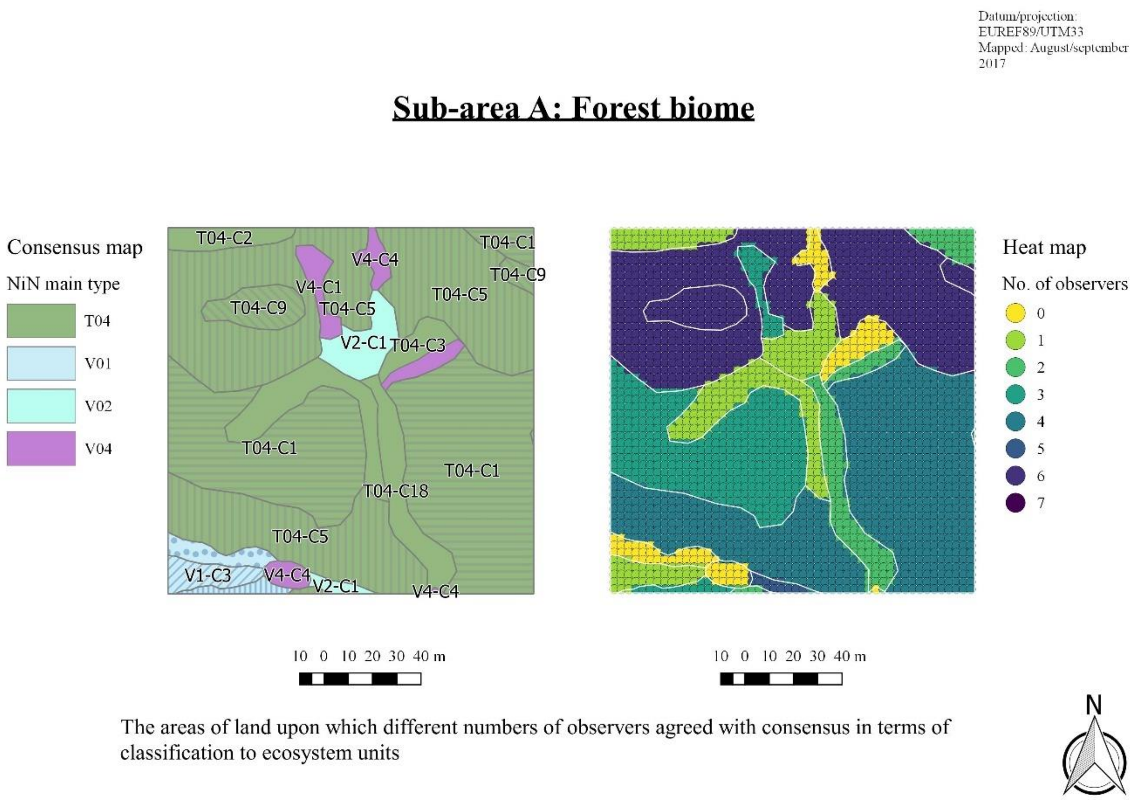


Figure 11: *Comparison of consensus map and heat map from the forest biome in sub-area A*

3.3.2 Heat maps: sub-area B

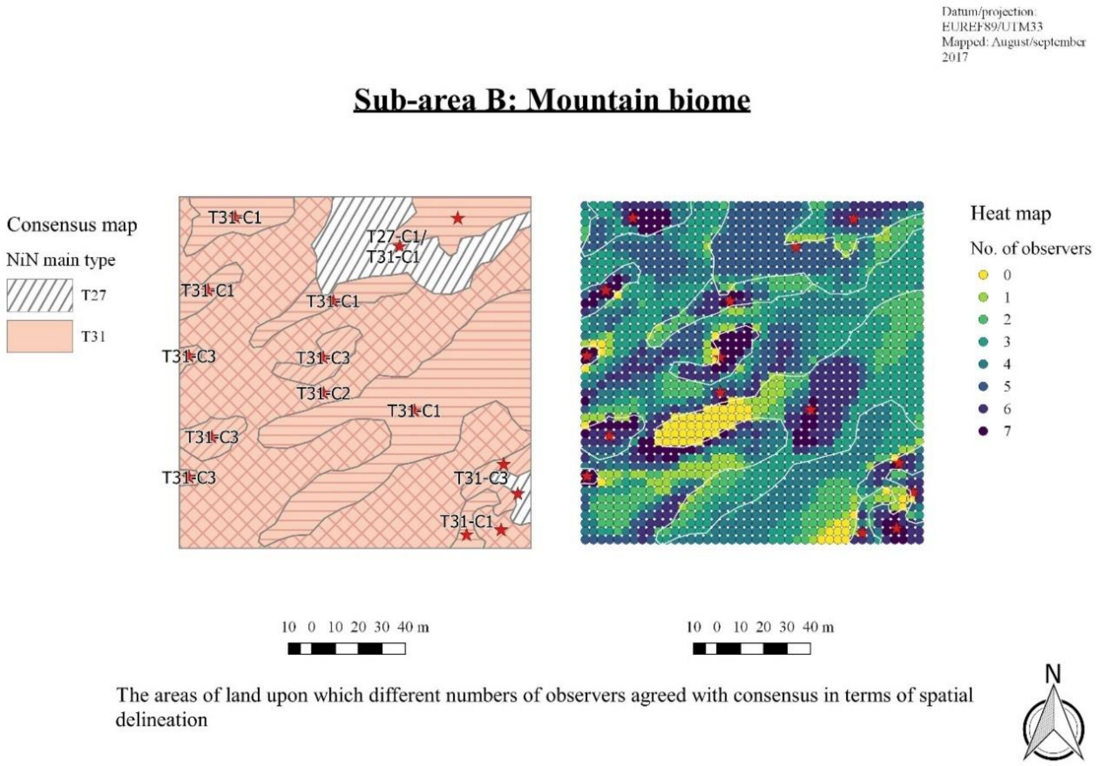


Figure 12: *Comparison of consensus map and heat map from the mountain biome in sub-area B. Classified points represented by red stars.*

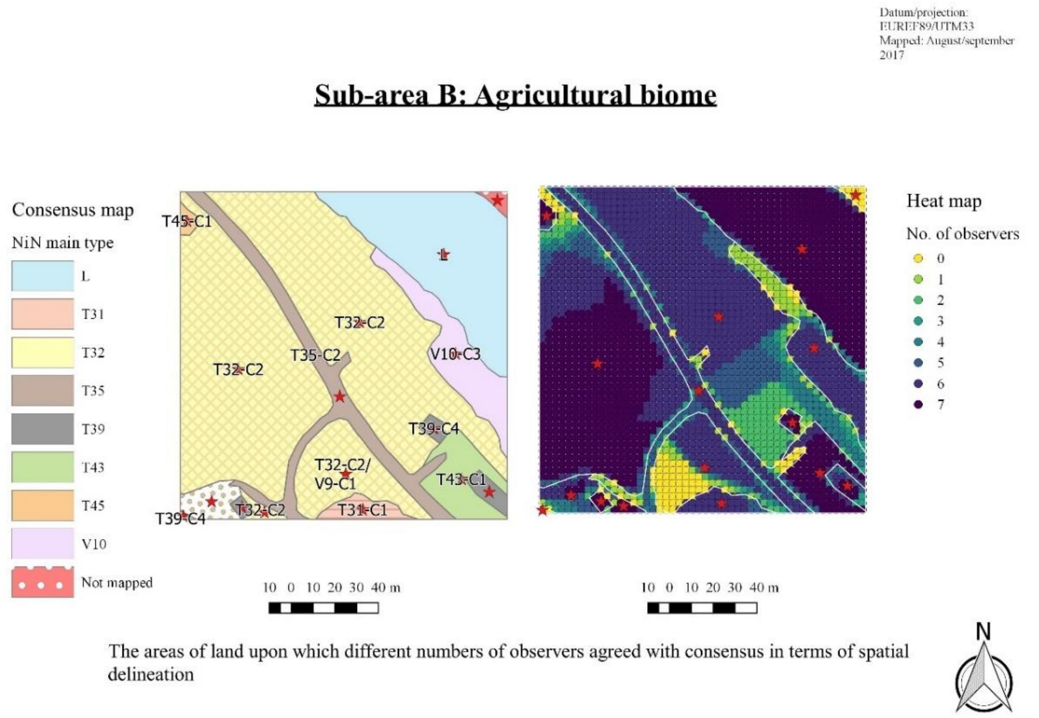


Figure 13: *Comparison of consensus map and heat map from the agricultural biome in sub-area B. Classified points represented by red stars.*

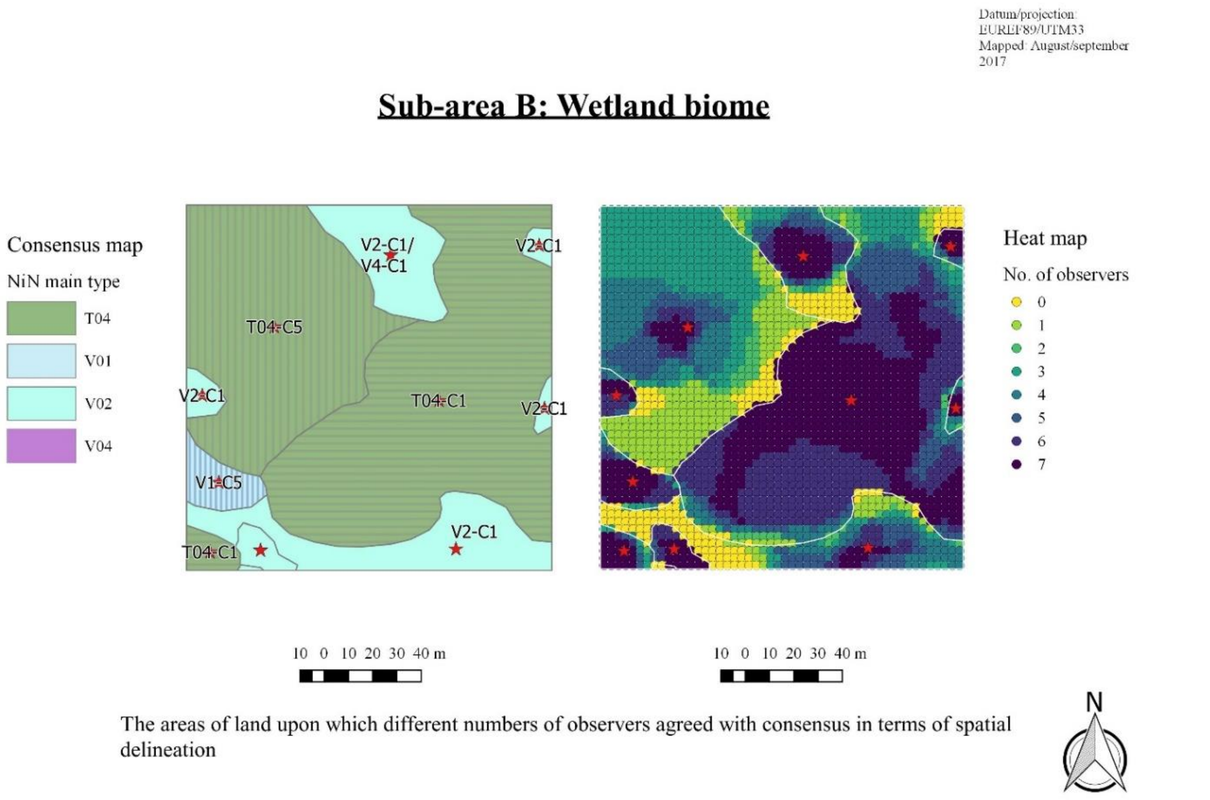


Figure 14: *Comparison of consensus map and heat map from the wetland biome in sub-area B. Classified points represented by red stars.*

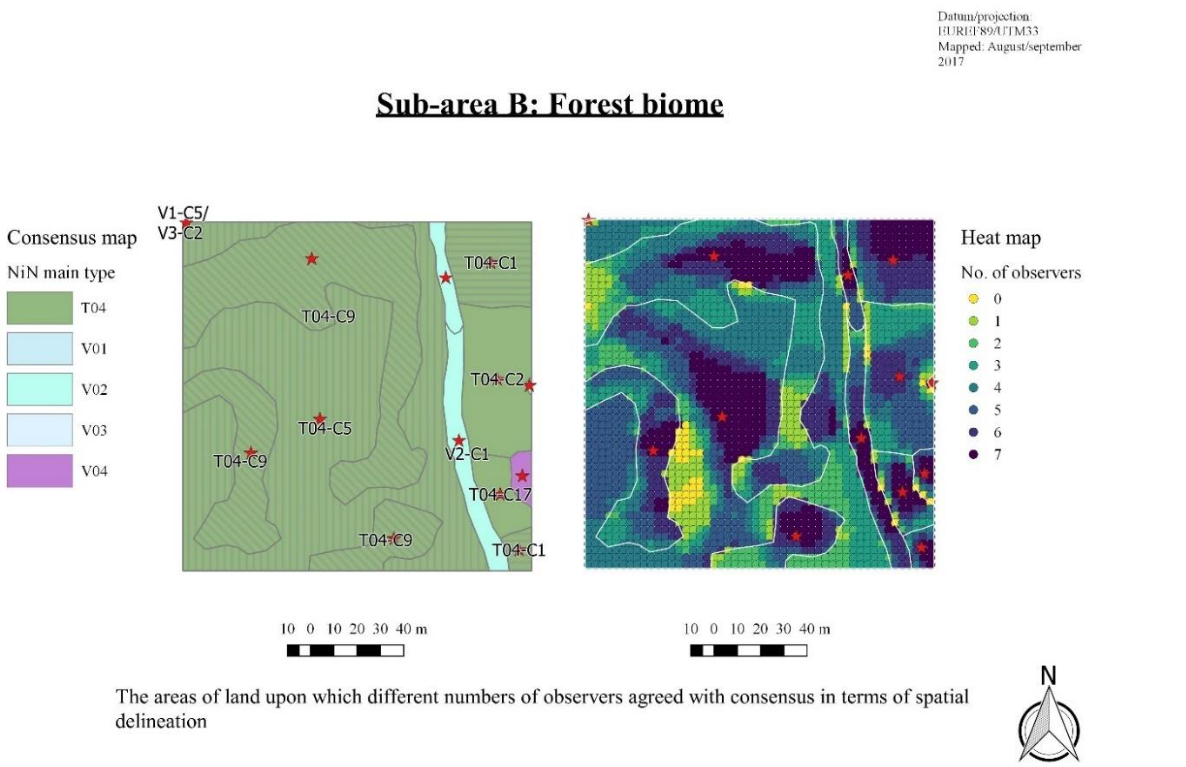


Figure 15: *Comparison of consensus map and heat map from the forest biome in sub-area B. Classified points represented by red stars.*

4 Discussion

4.1 A new method to separate the main inconsistencies in mapping

Numerous studies have investigated the quality of land cover maps and aimed to investigate the main sources of inaccuracy; classification and spatial delineation. This study has developed a new method that can investigate these main sources of inaccuracies separately. The results of implementing the ABC-partitioning show that in pairwise comparison between observers and a consensus map, there was a higher accuracy in classification than in spatial delineation. The mean classification accuracy was 71.8%, whereas the mean spatial delineation accuracy was 58.1%. This is in direct contrast to other comparable studies (Cherrill & McClean, 1995, 1999a; Eriksen *et al.*, Subm. ms.; Hearn *et al.*, 2011; Ullerud *et al.*, 2018), that have concluded that classification is the main source of map inconsistencies. Their findings were based on a buffer method, this is not an independent evaluation of classification versus spatial delineation inconsistencies, but rather a measure of delineation precision. Cherrill and McClean (1995, 1999a) and Hearn *et al.* (2011) improved the consistency by an average of only 4-5% when removing a buffer around the polygon delineations, thus concluding that classification is the main source of inconsistency. This method however, is an assessment of the variation in delineation precision, rather than a full analysis of the complexity in spatial delineation in land cover maps. This is especially challenging in maps with low consistency, since this makes it even more difficult to separate classification and spatial delineation inconsistencies (Alexander & Millington, 2000). Since the ABC-method captures more of the complexities in spatial delineation than in previous studies, the results show that the mean spatial delineation accuracy is lower than in previous studies. The presented results indicate that the inconsistencies emerging from spatial delineation is larger than the inconsistencies emerging from classification alone. Consequently, field-based mapping programs should put more efforts into training and harmonizing of spatial delineation.

The level of overall inconsistencies among this and comparable studies are approximately equal. Cherrill and McClean (1995, 1999a) and Hearn et al. (2011) found an overall consistency among mappers ranging from 25.6% to 34.2%, whereas the mean overall accuracy in this study was 30.5% for classification and 33.1% for spatial delineation.

4.2 Robustness with multiple observers

Comparing observer maps with a consensus map, provides the possibility of estimating the accuracy. This presupposes that one accepts the consensus map as a true map. In this study ten observers' interpretation of the area is included in the consensus map. This is not a perfect solution, but gives a more robust "true" map than using only one observers map. Several vegetation studies recommend the use of multiple observers, because working in teams has the effect of avoiding extreme estimates and detecting more species (Archaux, 2009; Archaux *et al.*, 2009; Gorrod & Keith, 2009; Klimeš *et al.*, 2001; Symstad *et al.*, 2008; Vittoz *et al.*, 2010). Ideally, one would use a map that depicts the true land cover types. However, this kind of map is not available, as land cover types are a simplification of reality (Pancer-Koteja *et al.*, 2009), influenced by subjective judgements of the mappers (Hearn *et al.*, 2011).

4.3 Classification accuracy

This study found a mean classification accuracy of 71.8%, ranging from 54.8% in forest biome to 96.7% in mountain biome. Mountain biome have very little variation in classification accuracy, whereas the other biomes have substantially larger variation. Wetland biome had the largest spread in classification accuracy. Wetland biome had a higher classification accuracy than both the forest- and agricultural biome, but a larger spread in classification accuracy. The mountain biome has a significantly different mean than wetland-, agricultural-, and forest biome. When implementing the Bonferroni correction, mountain biome was significantly different than agricultural- and forest biome. Studies using the same mapping system have found approximately the same or a lower accuracy. Eriksen (2017) found a mean classification accuracy of 65% in a point study, and Ullerud et al. (2018) found a consistency of 43.8% between pairs of observers (not a comparison with consensus). The latter study was a wall to wall mapping study, which can explain the lower consistency, since spatial delineation was included in the mapping.

There are many possible reasons for inaccuracies in classification. All classification methods result in maps with a degree of inaccuracy due to artificial simplification and generalization of natural features (Hearn et al., 2011). Different observers have varying degrees of botanical knowledge. Sufficient species knowledge is crucial in order to be able to recognize important indicator species needed to distinguish between ecosystem units. Differences in the ability to detect and identify species is a known cause of inconsistencies between observers (Bacaro et al., 2009; Hearn et al., 2011; Kirby, 2003). Units that are characterized by abundance of species that indicate a specific part of the gradient can also be difficult (Symstad et al., 2008). Morrison (2016) found that species can be overlooked and/or misidentified, where overlooking is a more prominent problem (Archaux et al., 2009). Morrison's study is about vegetation plots, but similar challenges are likely to occur in mapping as well. Subjectivity can play an important role when making decisions in mapping, which can lead to varying classifications among observers. In addition, a widespread challenge in mapping is that mappers have the tendency to find different units when mapping the same area (Cherrill & McClean, 1999b; Ullerud et al., 2018).

The NiN mapping system is still new and it has only been tested for a few mapping seasons. NiN is based on theoretical ecology rather than specialized for practical mapping. This can lead to a number of challenges. Units can be practically impossible to identify in the field, or observers may find that some units are "missing" from the system because there are no units that can describe the observed nature. Ihse (2007) recommends that mapping systems should be adapted to aim, scale and the nature of the collected data. Furthermore, a feedback loop to modify the mapping system, is suggested when the accuracy is low.

4.4 Spatial delineation accuracy

The mean spatial delineation accuracy is 58.1% and it varies among biomes. The spatial accuracy varies from 51.9% in agricultural biome to 63.9% in wetland biome. There is little variation in the mean spatial delineation accuracy between biomes (51.9 – 63.9%). The agricultural biome has a larger spread than the other biomes. Wetland has a significantly different mean spatial delineation accuracy than the mountain- and agricultural biome. When implementing the Bonferroni correction, none of the biomes were significantly different mean spatial delineation accuracy than any other biome.

The lowest accuracy in spatial delineation is between ecologically related units and between strongly modified units that resemble semi-natural units. Inaccuracy in spatial delineation can be treated as a function of the ability to distinguish adjacent units (Aspinall & Pearson, 1995). Also, natural phenomenon, such as general fuzzy boundaries and more or less continuous vegetation (Couclelis, 1992), make it difficult to delineate polygons. Previous studies have found that some observers stand out in comparison to others and consequently over- or under-estimate gradient positions (Eriksen, 2017). In map generalization adapted to the scale in question, these fuzzy boundaries are drawn as simplified lines. Even when the borders between units are sharp, the level of detail may be too complex for the map (Aune-Lundberg & Strand, 2017).

This study found that units that include estimates of species- and tree-cover are difficult to spatially delineate. This is especially prominent in the forest- and mountain biome. In the mountain biome spatial delineation between the ecologically related lime-poor mountain biome heath units is inconsistent. The only factor that separate these units is drought risk, since the lime richness is very low. Species groups, like bryophytes and lichens, occur in different amounts in these units. Separating them therefore includes estimation of cover of these species groups. The clearest patterns of inaccuracy are displayed in the wetland biome study site, on the border between wetland and forest, where both areas are tree covered, but differ in moisture. Estimation of coverage is known to be difficult (Kennedy & Addison, 1987; Tonteri, 1990; Willem & Mead, 2000). Shadows from trees can be problematic in aerial photos, making it harder to estimate borders and tree cover (Ihse, 2007).

Since it is not always possible or time to examine all borders in the field, the mapper needs to rely on aerial photos to spatially delineate (Cherrill & McClean, 1999b). Species cover, species types, moisture, soil nutrients, management level and succession state are considered the most difficult tasks to interpret from aerial photos, while separating open land from tree covered is considered the easiest (Ihse, 2007). Previous studies have found large variation in estimation of tree and species cover, with a low accuracy (Kennedy & Addison, 1987; Tonteri, 1990; Willem & Mead, 2000). The quality of the aerial photos is very important when spatially delineating. Time of photography and scale are mentioned as some of the important factors to consider when using aerial photos (Ihse, 2007). Different types of vegetation, for instance deciduous forest, wetland, semi-natural biomes etc. have their own optimal time of photography (Ihse, 1978). The main period suitable for vegetation mapping is between May and September. Deciduous forest has an early optimal photography time (May) and wetland late (August). This can be one of the reasons for the high spatial delineation accuracy in the wetland biome. The aerial photos from this study are taken in October, which is considered late. Ihse (2007) found a decreasing accuracy when deviating from optimal photography time.

Regional and local variation of abundance can vary. The meaning of vague formulations like "much" can mean different things in different regions (Eriksen, 2017). There were also difficulties distinguishing units with the same lime-richness, but differences in drought-risk. The areas with the same lime richness (in this case poor lime richness) have mostly species groups like bryophytes and lichens in different amounts like in lime-poor forest ecosystem units and poor mountain biome heath. Relative abundance of species can also be hard to estimate in field (Cherrill & McClean, 1999b). Gallegos Torrell and Glimskär (2009) recommend calibration with feedback to improve the accuracy of visual estimates.

Lime richness indicators are mostly not possible to see in aerial photos, making it difficult to distinguish between units with different lime-richness (Ihse, 2007). This was especially prominent in the forest biome. Also, species respond differently to gradients in different settings. Local- and regional variation can make an indicator species misleading or even obsolete (Halvorsen *et al.*, 2016). The relative abundance of species indicative to lime-richness has been found not to be consistent (Cherrill & McClean, 1999b).

Succession state is also mentioned as a challenging attribute to recognize in aerial photos (Ihse, 2007). This could be one of the main causes for low spatial delineation accuracy in the agricultural biome, where the observers had difficulties distinguishing between especially strongly modified units that resemble semi-natural units.

In sub-area B observers were given the identity of the points, therefore, as expected, the observers agree on these areas and they agree less when the distance from the points increase. If the points had been spatially randomized among the observers, it would most likely have led to lowered spatial accuracy. The reported 58.1% spatial delineation accuracy is therefore a modest estimate.

4.5 Ecosystem units' characteristics

Accuracy varied with the units that were mapped. This is supported by previous studies on mapping accuracy (Eriksen *et al.*, Subm. ms.; Ullerud *et al.*, 2018). Previous studies have found that ecologically related units are most often confused (Cherrill & McClean, 1999a;

Eriksen, 2017; Hearn *et al.*, 2011). These claims are supported by this study. The ecosystem units that were most often confused were mostly ecologically related, and were within the same major type, but there were also units classified as different major types that were confused.

There are some ecosystem unit characteristics that are more prominent when it comes to mapping accuracy. Units within major types that were most often confused where often characterized by lime richness, drought risk or rarity. A recent study (Eriksen, 2017) that used points to quantify the consistency and accuracy in classification, found similar results. Ecosystem units that were most difficult to map were units with similar lime richness and drought risk. Ullerud (2018) and Eriksen (2017) found a lower accuracy in units with high lime-levels in soil. Units that differ in lime richness are often confused, especially among the forest biome units, but also among wetland biome units. Mostly ecosystem units with a lower lime-richness than consensus was chosen. Eriksen (2017) found the opposite results. Observers chose a higher mean gradient level for lime-richness. Classification inaccuracy in these units can indicate lack of botanical abilities in detecting and recognizing indicator species of lime-richness.

Rare units like V4C4 (moderate to slightly lime-rich peat spring water) were often confused with more common mapping units. Observers can lack experience in recognition of the characteristics of the unit, thus classifying it as a more common unit. The observers were primarily inexperienced with both the mapping and the mapping system and may not have had a complete overview of the units. However, Eriksen (2017) found an overestimate of rare units. Ullerud *et al. (2018)* found shortcomings both with the most common units and the rare units.

Semi-natural units were often confused with strongly modified units that resemble semi-natural units. The strongly modified units are not defined by vegetation, but by the type of extensive intervention rather than species compositions (Bratli *et al.*, 2017). Stevens *et al.* (2004) found the lowest accuracy in semi-natural units. Eriksen (2017) and Ullerud (2018) who also used the NiN mapping system, found a low accuracy in units defined by human disturbances or other structuring processes. Many NiN-types are defined by land use, for instance grazing, ploughing or other long-term agricultural biome management or leveling, drifting of lakes or other strongly modified changes in addition to or instead of indicator species (Bratli *et al.*, 2017; Halvorsen *et al.*, 2016). In ecosystem units defined by land use, extensive local knowledge is needed to make informed and correct classifications. Both current and historic land use, especially the latter, are almost never available, and the observer will have to guess and make assumptions, which can make classification difficult.

Semi-natural major types' end succession state is natural major types. When the semi-natural major type is close to its end succession state, it can be difficult to know which major type the polygon should be classified as. In agricultural biome, semi-natural grassland was mostly confused with either T31 (boreal heath) or T4 (forest biome). These two major types can be very similar, since forest biome is the end succession state of boreal heath and species typical of the semi-natural unit can gradually be replaced by species characteristic of the end succession state (Eriksen, 2017).

Major types can in some cases be very similar, with similar species composition, and mostly separated by tree cover (Bratli *et al.*, 2017) and without distinct plant composition (Aune-Lundberg & Strand, 2017). Units with these attributes were frequently confused. In the agricultural biome, T32 (semi-natural grassland) was mostly confused with either T31 (boreal heath) or T4 (forest). Major type forest is defined by a 10% tree cover in NiN (Bratli *et al.*, 2017). Regrowth, late succession state and tree cover close to 10% can be the cause of this. Estimation of tree cover is challenging (Gallegos Torell & Glimskär, 2009) and the estimation is more difficult with smaller estimates (Morrison, 2016).

Confusion between units is an important factor in classification inaccuracies, but can also play an important role in spatial delineation. Most ecosystem units, especially those that are defined by vegetation cover, change gradually rather than having crisp borders (Faliński, 1994; Hearn *et al.*, 2011; Küchler & Zonneveld, 1988). Different perception of units can lead to inconsistencies and calibration is therefore essential (Cherrill & McClean, 1995, 1999a, 1999b).

By aggregating ecosystem units and using a coarser scale when mapping, consistency among observers can probably be increased (Cherrill & McClean, 1999a; Eriksen *et al.*, Subm. ms.; Ullerud *et al.*, 2018). By aggregating units into broader classes, the ecological variation within units is larger. Total agreement among observers will likely only be achieved when land cover types have been combined into a small number of categories, so that all ecological resolution is lost (Cherrill & McClean, 1999b). Even though inaccurate classifications are often ecologically similar to the reference, closely related units can accommodate different species that can be of interest (Newsome et al., 1995).

Although many land cover mapping projects/programs have detailed mapping instructions, it is likely that observers apply different rules and/or interpret them differently (Cherrill & McClean, 1999b). There are many reasons for this. Observers can have varying knowledge of mapping instructions and they can estimate things differently in the field (for instance minimum polygon size, width etc.) and overemphasize interesting areas. Available ecosystem units may not always be suitable for the nature that is being observed. Combinations of different factors can be uncommon, or it is not captured by the mapping scale that is used, and is therefore not included as ecosystem units. In these cases, observers have to choose the "least wrong" unit.

4.6 Biome complexity

The present results indicate that some biomes are more difficult to map consistently than others. Biomes with the most possible units to choose from had the lowest accuracy (forest- and agricultural biome). Mountain- and wetland biome had the highest accuracy and the fewest units to choose from. A higher amount of available mapping units, with similar species composition is associated with lower accuracy (Cherrill & McClean, 1995, 1999a; Halvorsen *et al.*, 2011; Hearn *et al.*, 2011; Ullerud *et al.*, 2018).

The forest biome site had great variation in topography, which can impact drought risk and lime richness (Ihse, 2007) which again lead to many possible mapping units to choose from and, thus contribute to confusion when classifying. Forest biome had the largest deviation in ED (1.0) from consensus, and the most evenly distributed ED frequency. Ullerud *et al.* (2018) also found low accuracy in forest biome when using the same mapping system (NiN). However, the same study found the lowest accuracy in wetland biome when using another mapping system (NIBIO) (Rekdal & Larsson, 2005). This is contrast to the results from this study where wetland biome had the highest classification accuracy. These results support Ullerud *et al.*'s (2018) hypothesis that the mapping system may be more important for the resulting maps than which biome that is mapped.

Agricultural biome was also difficult to map. The results show low accuracy in both classification and spatial delineation, respectively 66.3% and 53.1%, which is the lowest accuracy in spatial delineation of all investigated biomes. The ecological distance was also quite evenly distributed and had a mean deviance from consensus of 0.9 ED. Agricultural biome also had many possible units, due to variation in human intervention intensity and succession state. Several studies support the findings on the difficulties mapping semi-natural units (Eriksen *et al.*, Subm. ms.; Stevens *et al.*, 2004; Ullerud, 2018).

Mountain- and wetland biomes had the highest classification accuracy, respectively 96.7% and 69.6% and low ED from consensus. Wetland biome had similar accuracy in both classification and spatial delineation (63.9%), but mountain biome had a large difference in accuracy with a mean a mean spatial delineation accuracy of 53.9%. Although the wetland biome had a high mean classification accuracy, there was great variation among observers (37.5 – 100.0%). The mountain biome did not display this large variation (93.4 – 100.0%). Mountain biome had few possible units to choose from, but the units were ecologically very similar, mostly variating in drought risk. The mountain biome was mainly lime-poor with few indicator species. The observers had to rely on the visual estimates of species cover and moisture, which can explain the that the mean spatial delineation accuracy was lower than the mean classification accuracy.

In all biomes except the forest biome, the mean classification accuracy (54.8%) is higher than the mean spatial delineation accuracy (62.5%). The forest biome, had in general low mean accuracy. The forest biome site was mostly tree covered. This can be one of the causes for the low spatial delineation accuracy in this biome, since delineation of tree covered units often are based on interpretation of aerial photos.

4.7 Uncertainties in this study

This study takes a step further towards the separation of the main causes of inaccuracies in land cover mapping, but there is a degree of uncertainty. Closely related units are often confused when classifying, but also when delineating. This can mean that differences in perception of units can be the cause of inaccuracies in both classification and spatial delineation. Spatial delineation inconsistency can be treated as a function of the ability to distinguish adjacent units (Aspinall & Pearson, 1995), implying that if the perception of these units are different, the border between them may be placed differently. Large differences in spatial delineation may indicate that observers do not perceive the same gradients. This can be a secondary symptom of varying perceptions of types of land cover present, rather than inconsistencies in spatial delineation (Cherrill & McClean, 1999a). Furthermore, this can indicate that the two factors are not completely separated. According to Ullerud (2018), the delineation can be affected by factors like differences in perception of units due to differences in border placement. by placing borders differently, the content of the polygons will differ.

This study has a small sample size. Therefore, the results from statistical tests, respectively standard deviation, confidence interval and t-test, has a high degree of uncertainty. The Bonferroni-correction counteract some of the problems of multiple comparisons, with a lowered significance level. Nevertheless, these results can give an indication.

These are results from one area in Norway with one mapping system, and may therefore have limited transferability. In order to draw more certain conclusions, the study should be repeated in other parts of the country or other countries, and with other mapping systems.

Part one and two of this study is performed by different composition of observers with varying experience level. Mapping was done individually in part one and in pairs in part two. In part one, the observers were mainly experienced, but in part two of the study, there were mainly inexperienced observers. Some participants had some land cover mapping experience, but none with experience in this mapping system. Similar to the other group of observers they had some experience in other botanical fields, though mostly limited experience in general. These differences in mapping experience is an important factor to consider. Eriksen (2017) found a positive correlation between experience and accuracy. The most experienced observers deviated least from consensus. Hearn *et al.* (2011) also found a higher agreement in habitat type between experienced observers. However, the observers from part two worked in pairs, which has shown to increase accuracy and is recommended in many studies (Archaux, 2009; Archaux et al., 2009; Gorrod & Keith, 2009; Klimeš et al., 2001; Symstad et al., 2008; Vittoz et al., 2010). Although the two parts of the study were executed by observers with different experience levels, the difference between observers in part one and part two may not be that prominent as they worked together in pairs. Archaux (2009) found a significant reduction in bias when the observers worked in pairs, compared to observing individually. This can indicate that although the observers in part one and part two of the study had different experience levels, they may be comparable when it comes to robustness. To recieve more comparable results, the groups could have been made up by observers with more similar experience levels, both with land cover mapping and other botanical fields, and that both parts of the study were performed either by pairs or individuals.

Even though the observers in part two got clear instructions, some errors were made. For example the inclusion of more than one point in one polygon in part B. In sub-area A, some of the smallest polygons were overlooked and thus not classified. Results obtained when deviating from the given field instructions were not included in the study. These types of errors were not a widespread problem, but is important to consider, especially for the purpose of repeating the study design. Clearer instructions and more control in the field could eliminate these errors.

The ED measure is a useful tool to get a more balanced view on inaccurate classifications (Eriksen, 2017). This method has some weaknesses. In some cases, the ED between units can be exaggerated and the difference can seem more prominent than what it actually is. This method can also underestimate differences. In most calculations the ED seemed appropriate, but in a few cases it did not. This was especially the case in units that were close to their successional end level.

4.8 Further studies

There are several measures that can be made to improve the quality of land cover maps, but further studies are needed to understand the causes of inconsistencies and varying accuracy. This study leads us a step closer to the understanding of the proximate causes of inaccuracy in mapping, but further research is needed to understand the ultimate causes. Through this study, we now know that e.g. spatial delineation strongly contributes to low accuracy, but we have not yet tested why the mappers delineate differently. Is it because the mappers lack experience with spatial delineation, or because they lack harmonization? Is it because the mappers lack knowledge on how to use aerial photos properly? Is it the order of processing varying, i.e. classifying first and then delineation or the other way around? Is it the way in which they move around from place to place when they map? Is it the lack of species knowledge, leading them provide a low classification accuracy, which subsequently leads to low accuracy in spatial delineation? To improve our understanding, we suggest establishing in-depth studies that focus on getting the complete overview of the possible causes of inconsistencies.

Improving the understanding of these causes, may help us to guide mappers better and could subsequently lead to lowered inconsistencies and higher accuracy.

4.9 Conclusions

Pairwise comparisons show that the main source of inaccuracy is differences in spatial delineation. Classification also has inaccuracies, but less prominent. When deviating from consensus, ecologically similar units are more frequently chosen. The units that most frequently mixed varied in lime richness and drought risk, but units that were defined based on type of extensive land use, rather than species compositions and units in late succession states, were more often the cause of confusion between major types. Units that relied on estimation of species cover was a source of inaccuracy within and between major types. There is variation among biomes when it comes to mapping accuracy, both in spatial delineation and classifications. Some biomes are more difficult to map than others, and display inaccuracy in spatial delineation and classification. Ecosystems that are more challenging to map have a greater deviation in mean ecological distance per polygon. System complexity strongly influences the mapping accuracy. Complex ecosystems with many ecosystem units are therefore more difficult to map concisely than simple ecosystems with few ecosystem units.

References

Alexander, R., & Millington, A. C. (Eds.). (2000). *Vegetation mapping: From patch to planet*.

Archaux, F. (2009). Could we obtain better estimates of plot species richness from multiple-observer plant censuses? *Journal of Vegetation Science, 20*(4), 603-611.

Archaux, F., Camaret, S., Dupouey, J.-L., Ulrich, E., Corcket, E., Bourjot, L., . . . Dumas, Y. (2009). Can we reliably estimate species richness with large plots? An assessment through calibration training. *Plant ecology, 203*(2), 303-315.

Aspinall, R. J., & Pearson, D. M. (1995). Describing and managing uncertainty of categorical maps in GIS. In P. Fischer (Ed.), *Innovations in GIS* (2 ed., pp. 71-83): Taylor & Francis.

Aune-Lundberg, L., & Strand, G.-H. (2017). Composition and spatial structure of dwarf shrub heath in Norway. *Norsk Geografisk Tidsskrift-Norwegian Journal of Geography, 71*(1), 1-11.

Bacaro, G., Baragatti, E., & Chiarucci, A. (2009). Using taxonomic data to assess and monitor biodiversity: are the tribes still fighting? *Journal of Environmental Monitoring, 11*(4), 798-801.

Biondi, E., Feoli, E., & Zuccarello, V. (2004). Modelling environmental responses of plant associations: a review of some critical concepts in vegetation study. *Critical Reviews in Plant Sciences, 23*(2), 149-156.

Box, E. O., & Fujiwara, K. (2013). Vegetation types and their broad-scale distribution. In J. Franklin & E. Maarel (Eds.), *Vegetation ecology* (2 ed., pp. 106-128): Wiley & Blackwell.

Bratli, H., Halvorsen, R., Bryn, A., Arnesen, G., Bendiksen, E., Jordal, J. B., . . . Aarrestad, P. A. (2017). *Dokumentasjon av NiN versjon 2.1 tilrettelagt for praktisk naturkartlegging i målestokk 1:5000* Norwegian Biodiversity Information Centre, Trondheim, Norway.

Brodal, G., & Kurud, P. (2012). *Bjønnåsen - Fra seterliv til aktiv hyttegrend*. Bjønnåsen Vel.

Bryn, A. (2006). Vegetation mapping in Norway and a scenario for vegetation changes in a mountain district. *Geographia Polonica, 79*(1), 41-64.

Bryn, A., Gudjonsson, G., Skånes, H., & Käyhkö, N. (2014). The Nordic region: Denmark, Finland, Iceland, Norway and Sweden. . In J. Ichter, D. Evans, & D. Richards (Eds.), *Terrestrial habitat mapping in Europe: an overview* (Vol. Technical report 1/2014, pp. 97-98). Copenhagen: European Environment Agency.

Bryn, A., Halvorsen, R., & Ullerud, H. A. (2018). *Hovedveileder for kartlegging av terrestrisk naturvariasjon etter NiN (2.2.0) - Utgave 1*. Norwegian Biodiversity Information Centre, Trondheim, Norway.

Bryn, A., & Ullerud, H. A. (2017). *Veileder for arealdekkende terrestrisk naturvariasjon etter NiN i målestokk 1:500 og 1:20000. Kort versjon (2.0.1.b)*. Norwegian Biodiversity Information Centre, Trondheim, Norway.

Cherrill, A. (2013). Repeatability of vegetation mapping using Phase 1 and NVC approaches: Implications for professional practice and surveyors' training requirements. *InPractice, 81*, 41-45.

Cherrill, A. (2016). Inter-observer variation in habitat survey data: investigating the consequences for professional practice. *Journal of environmental planning and management, 59*(10), 1813-1832.

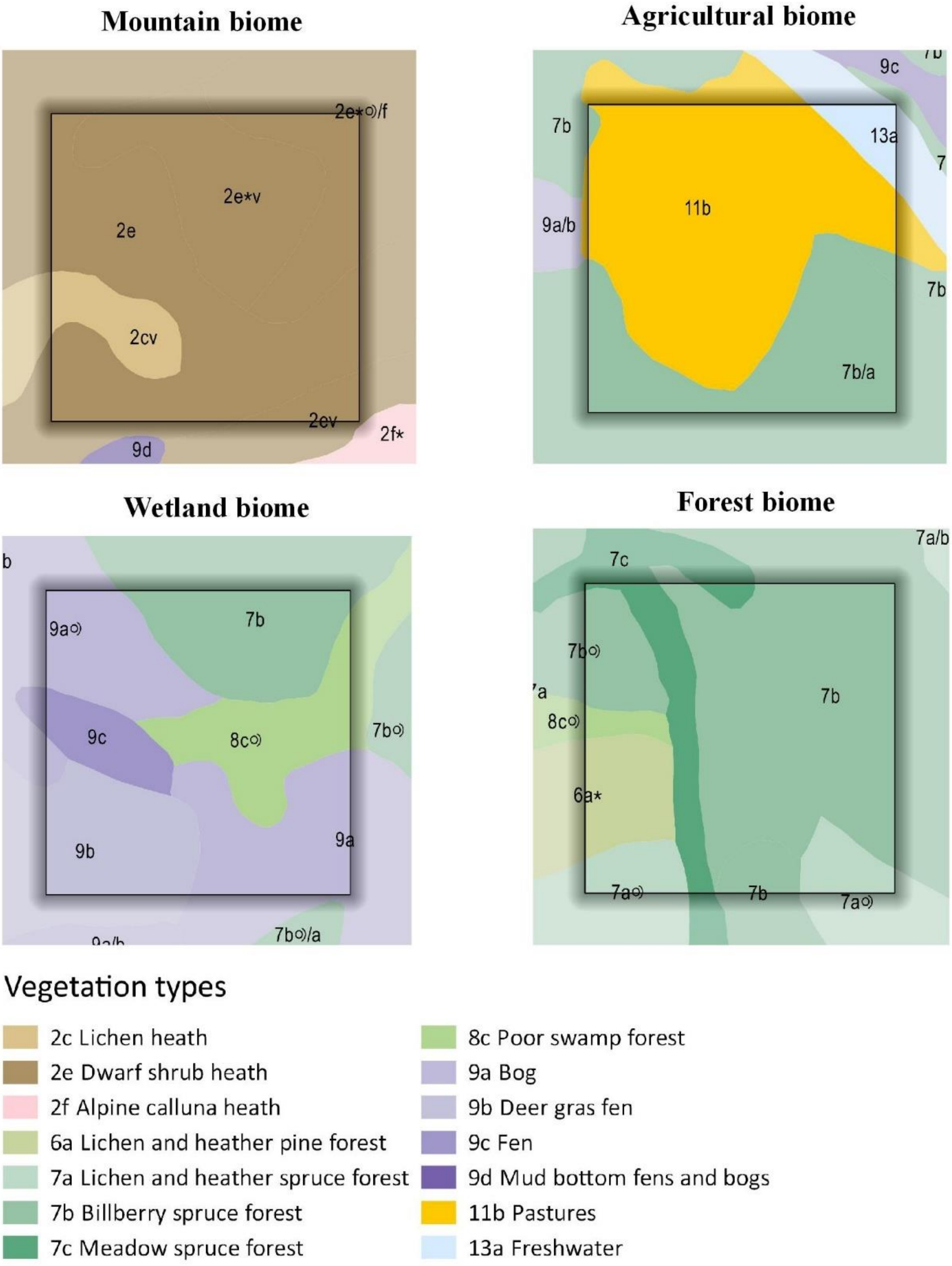
Cherrill, A., & McClean, C. (1995). An investigation of uncertainty in field habitat mapping and the implications for detecting land cover change. *Landscape Ecology, 10*(1), 5-21.

Cherrill, A., & McClean, C. (1999a). Between-observer variation in the application of a standard method of habitat mapping by environmental consultants in the UK. *Journal of Applied Ecology, 36*(6), 989-1008.
Cherrill, A., & McClean, C. (1999b). The reliability of 'Phase 1' habitat mapping in the UK: the extent and types of observer bias. *Landscape and urban planning, 45*(2-3), 131-143.
Couclelis, H. (1992). People manipulate objects (but cultivate fields): beyond the raster-vector debate in GIS. In A. U. Frank, i. Campari, & U. Formentini (Eds.), *Theories and methods of spatio-temporal reasoning in geographic space. Lecture notes in computer science* (Vol. 639, pp. 65-77): Springer.
De Cáceres, M., & Wiser, S. (2012). Towards consistency in vegetation classification. *Journal of Vegetation Science, 23*(2), 387-393.
Eriksen, E. L. (2017). *Point of View: Uncertainty in Nature Type Identification.* (MSc), University of Oslo.
Eriksen, E. L., Ullerud, H. A., Halvorsen, R., Aune, S., Bratli, H., Horvath, P., . . . Bryn, A. (Subm. ms.). *Point of view: Error and inter-observer variation in field assignment of mapping units*.
Faliński, J. B. (1994). Comparison of the maps and general experiences in the vegetation cartography. *Phytocoenosis NS, Suppl. Cartogr. Geobot, 4*, 91-104.
Franklin, J. (1995). Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in physical geography: Earth and environment, 19*(4), 474-499.
Fuchs, R., Herold, M., Verburg, P. H., Clevers, J. G. P. W., & Eberle, J. (2015). Gross changes in reconstructions of historic land cover/use for Europe between 1900 and 2010. *Global change biology, 21*(1), 299-313.
Gallegos Torell, Å., & Glimskär, A. (2009). Computer-aided calibration for visual estimation of vegetation cover. *Journal of Vegetation Science, 20*(6), 973-983.
Gorrod, E. J., & Keith, D. A. (2009). Observer variation in field assessments of vegetation condition: implications for biodiversity conservation. *Ecological Management & Restoration, 10*(1), 31-40.
Green, D. R., & Hartley, S. (2000). Integrating photointerpretation and GIS for vegetation mapping: some issues of error. In R. Alexander & A. C. Millington (Eds.), *Vegetation Mapping: From Patch to Planet* (pp. 103-134). West Sussex, England: John Wiley & Sons Ltd.
Halvorsen, R., Bryn, A., & Erikstad, L. (2016). *NiNs systemkjerne – teori, prinsipper og inndelingskriterier* (Vol. 2.1.0). Norwegian Biodiversity Information Centre, Trondheim, Norway.
Halvorsen, R., Mazzoni, S., Bratli, H., Engan, G., Fjeldstad, H., Gaarder, G., & … Nordbakken, J.-F. (2011). *Utprøving av NiN versjon 1.0 som naturtypekartleggingssystem. NHM rapport 11*. Universitetet i Oslo, Naturhistorisk museum.
Hasle, L. (Ed.) (2004). *Et melkeeventyr - Tine Meieriet Øst Brumunddal 50 år*. Brumunddal
Hearn, S. M., Healey, J. R., McDonald, M. A., Turner, A. J., Wong, J. L. G., & Stewart, G. B. (2011). The repeatability of vegetation classification and mapping. *Journal of environmental management, 92*(4), 1174-1184.
Hunter, J. T. (2016). Validation of the Greater Hunter native vegetation mapping as it pertains to the Upper Hunter region of New South Wales. *Ecological Management & Restoration, 17*(1), 40-46.
Ihse, M. (1978). *Flygbildstolkningar av vegetation i syd-och mellansvensk terräng:-en metodstudie för översiktlig kartering*. Naturvårdsverket, Stockholm, Sweden.

Ihse, M. (2007). Colour infrared aerial photography as a tool for vegetation mapping and change detection in environmental studies of Nordic ecosystems: A review. *Norsk Geografisk Tidsskrift-Norwegian Journal of Geography, 61*(4), 170-191.
Kennedy, K. A., & Addison, P. A. (1987). Some considerations for the use of visual estimates of plant cover in biomonitoring. *The Journal of Ecology, 75*(1), 151-157.
Kirby, K. J. (2003). *NVC in semi-natural woodland*. Retrieved from
Klimeš, L., Dančák, M., Hájek, M., Jongepierová, I., & Kučera, T. (2001). Scale-dependent biases in species counts in a grassland. *Journal of Vegetation Science, 12*(5), 699-704.
Küchler, A. W., & Zonneveld, I. S. (Eds.). (1988). *Vegetation mapping. Handbook of vegetation science* (Vol. 10): Kluwer Academic Publisher, Dordrecht.
Ministry of Climate and Environment. (2009, 2018.01.01). Norwegian Biodiversity Act. Retrieved from https://lovdata.no/dokument/NL/lov/2009-06-19-100
Moen, A. (1999). *National atlas of Norway: vegetation*. Hønefoss, Norway: ISBN 82-7945-000-9.
Morrison, L. W. (2016). Observer error in vegetation surveys: a review. *Journal of Plant Ecology, 9*(4), 367-379.
Newsome, T. H., Wikeem, B. M., & Sutherland, D. C. (1995). *Sheep grazing guidelines for managing vegetation on forest plantations in British Columbia*: Province of British Columbia, Ministry of Forests Research Program.
Pancer-Koteja, E., Szwagrzyk, J., & Guzik, M. (2009). Quantitative estimation of vegetation changes by comparing two vegetation maps. *Plant ecology, 205*(1), 139-154.
QGIS Developement Team. (2018). QGIS Geographic Information System (Version 2.18.14). Retrieved from http://qgis.osgeo.org
R Core Team. (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.r-project.org/
Rekdal, Y., & Angeloff, M. (2016). *Beiteressursar i Hedmark, ressursgrunnlag i utmark og areal av innmarksbeite i kommunar og fylke* (Vol. 2/59): Norsk institutt for Bioøkonomi, Ås, Norway.
Rekdal, Y., & Bryn, A. (2010). Mapping of vegetation in Norway. *Viten, 1*, 93-96.
Rekdal, Y., Bryn, A., & Hofsten, J. (2003). *Vegetasjon og beite på Ringsakerfjellet. Rapport fra vegetasjonskartlegging* (Vol. 8/03): NIJOS-rapport, Ås, Norway.
Rekdal, Y., & Larsson, J. Y. (2005). *Veiledning i vegetasjonskartlegging M 1: 20 000–50 000*. Norsk institutt for jord- og skogkartlegging, Ås, Norway.
Solheim, E. (1978). Vegetation Mapping Using Colour Infrared Air Photos. *Kart og Plan*(4), 246-256.
Stevens, J. P., Blackstock, T. H., Howe, E. A., & Stevens, D. P. (2004). Repeatability of Phase 1 habitat survey. *Journal of environmental management, 73*(1), 53-59.
Sutherland, W. J., & Hill, D. A. (Eds.). (1995). *Managing habitats for conservation*. Cambridge, England: Cambridge University Press.
Symstad, A. J., Wienk, C. L., & Thorstenson, A. D. (2008). Precision, repeatability, and efficiency of two canopy-cover estimate methods in northern Great Plains vegetation. *Rangeland Ecology & Management, 61*(4), 419-429.
The Norwegian Water Resources and Energy Directorate. (2017). Se Norge. Retrieved from www.senorge.no
Tonteri, T. (1990). Inter-observer variation in forest vegetation cover assessments. *Silva Fennica, 24*(2), 189-196.
Ullerud, H. A. (2018). *Land cover mapping: Methods and Quality.* (PhD), University of Oslo.
Ullerud, H. A., Bryn, A., Halvorsen, R., & Hemsing, L. Ø. (2018). Consistency in land cover mapping: influence of fieldworkers, spatial scale and classification system. *Applied vegetation science*, n/a-n/a. doi:10.1111/avsc.12368

Ullerud, H. A., Bryn, A., & Klanderud, K. (2016). Distribution modelling of vegetation types in the boreal–alpine ecotone. *Applied vegetation science, 19*(3), 528-540.
Vittoz, P., Bayfield, N., Brooker, R., Elston, D. A., Duff, E. L., Theurillat, J. P., & Guisan, A. (2010). Reproducibility of species lists, visual cover estimates and frequency methods for recording high-mountain vegetation. *Journal of Vegetation Science, 21*(6), 1035-1047.
Willem, W. S., & Mead, B. R. (2000). Ocular estimates of understory vegetation structure in a closed Picea glauca/Betula papyrifera forest. *Journal of Vegetation Science*, *11*(2), 195-200.
Xie, Y., Sha, Z., & Yu, M. (2008). Remote sensing imagery in vegetation mapping: a review. *Journal of Plant Ecology, 1*(1), 9-23.

Appendix 1: Vegetation maps of study sites



Appendix 2: Aerial photos

Orthophotos downloaded from www.norgeibilder.no.

Ringsaker 1973: photographed 16. June 1973, downloaded in a 0.20m resolution, datum: EUREF89, projection: UTM32

Østlandet 2016: photographed 3. October 2016, downloaded in a 0.25m resolution, datum: EUREF89, projection: UTM32