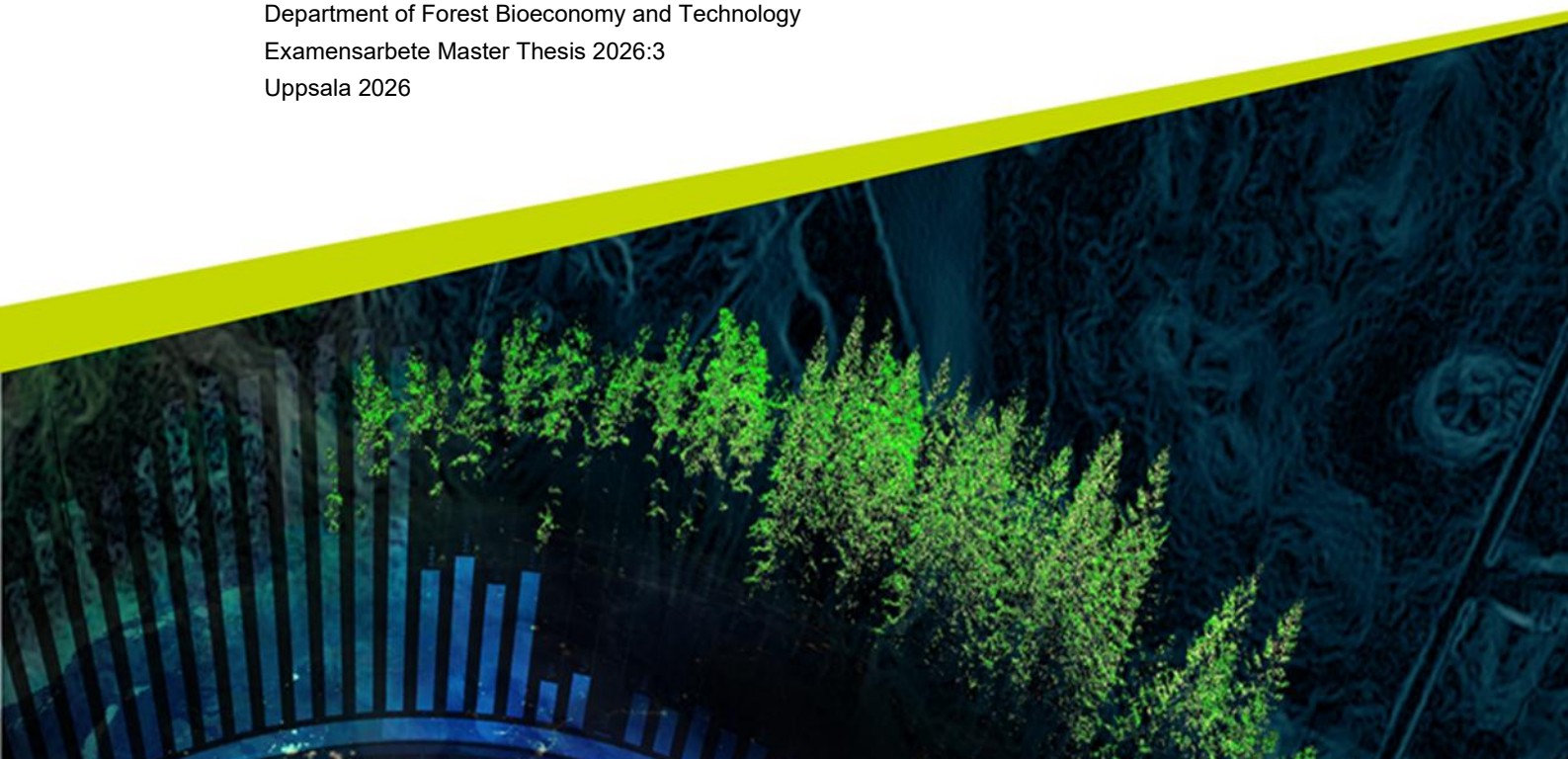
Digitalization in forestry planning

– GIS-based machine learning for assessing pre-harvest clearing needs

Digitalisering i skoglig planering
- GIS-baserad machine learning vid behovsanalys av förrensning inför avverkning

Isak Kautto

Degree project/Independent project • 30 hp
Swedish University of Agricultural Sciences, SLU
Faculty of Forest Sciences
Department of Forest Bioeconomy and Technology
Examensarbete Master Thesis 2026:3
Uppsala 2026



Digitalization in forestry planning
- GIS-based machine learning for assessing pre-harvest clearing needs

Digitalisering i skoglig planering – GIS-baserad machine learning vid behovsanalys av förrensning inför avverkning

Isak Kautto

Supervisor: Cecilia Mark-Herbert, Swedish University of Agricultural Sciences, Department of Forest Bioeconomy and Technology

Examiner: Anders Lindhagen, Swedish University of Agricultural Sciences, Department of Forest Bioeconomy and Technology

Credits: 30 hp
Level: A2E
Course title: Independent project in Sustainable Development
Course code: EX1021
Programme/education: Master's Programme Sustainable Development
Course coordinating dept: Department of Aquatic resources

Place of publication: Uppsala
Year of publication: 2026
Title of series: Examensarbete / Institutionen för skoglig bioekonomi och teknologi
Part number: 2026:3
ISSN:

Keywords: GIS, machine learning, LiDAR, PHC, precision forestry

Swedish University of Agricultural Sciences
Faculty of Forest Sciences
Department of Forest Bioeconomy and Technology

Summary

Pre-harvest clearing is an efficient tool to increase productivity in final harvests in tree stands with dense understory vegetation. As of today however, deeming pre-harvest clearing needs is done manually through forestry professionals field-visiting tracts, and performing an intricate process of assessing whether or not the tract would benefit from being cleared before harvest. This is seen as a very time-consuming process, which is also prone to subjective assessments.

This project explores the potential of digitalizing this process using a non-linear machine learning approach in GIS. The machine learning method is based on Breiman's (2001) Random forest classification model using raster data derived from LiDAR and other stand-growth-related factors as defined by Albrektson et al. (2012).

This thesis is written on a commission from Holmen Forest. Thus, the classification labels in the data used to train the model are based on real-world pre-harvest clearing decisions made by forestry professionals at Holmen Forest.

The model achieved an overall Matthews correlation coefficient of 0.59 and 0.81 accuracy. The model also showed that elevation, the topographic wetness index, slope and LiDAR-based vegetation point clouds between seven to nine meters were the most important factors in the used dataset.

The analysis identified signs of overfitting in the model and certain sources of error in the LiDAR data. However, the model predicted three out of four tracts correctly when tested on unseen data in a more practical setting. This method therefore shows promise as a digital tool for assessing pre-harvest clearing with further development and increased training data.

Keywords: GIS, machine learning, LiDAR, PHC, precision forestry

Sammanfattning

Förrensning är en effektiv åtgärd för att öka produktiviteten i en avverkning i bestånd där undervegetationen är tät. I dagens skogsbruk är dock beslutstagandet bakom förrensning gjort via manuella inventeringar, där traktplanerare fysiskt besöker trakten i fråga och gör en bedömning ifall trakten skulle gynnas av en förrensning. Detta anses vara en tidskrävande process, som även är benägen att göras med subjektiva uppfattningar.

Detta arbete undersöker potentialen i att digitalisera denna process med hjälp av en icke-linjär machine learning metod i GIS. Machine learning-modellen är baserad på Breimans (2001) Random forest classification model, med LiDAR data och ståndortsinformation som inputs.

Uppsatsen är skriven på uppdrag av Holmen skog. Därav är träningsdatan för modellen baserad på verkliga förrensningsbeslut tagna av traktplanerare på Holmen skog.

Modellen uppnådde en Matthews Correlation Coefficient på 0.59 och 0.81 accuracy. Modellen visade att elevation, topographic wetness index, lutning och LiDAR-punktmoln baserade på vegetation i intervallen sju till nio meter var de viktigaste faktorerna i arbetets dataset.

Analysen identifierade tecken på overfitting i modellen samt vissa felkällor i LiDAR-datan. Trots det förutspådde modellen rätt på tre av fyra testtrakter i extern validering. Denna metod visar därför på god potential för att bedöma förrensningsbehov med ytterligare utveckling och träningsdata.

Keywords: GIS, machine learning, LiDAR, förrensning, precision forestry

Table of contents

List of tables

List of figures

Abbreviations

Abbreviation	Meaning	Page
ABA	Area-based approach	31
AI	Artificial Intelligence	12
ALS	Airborne Laser Scanning	34
CAD	Computer-aided Design	16
CEMUS	Centre for Environment and Development studies at Uppsala university and SLU	29
DTM	Digital Terrain model	31
GIS	Geographical Information systems	14
GROT	Harvesting residue used for energy production; i.e branches and tops	13
LAS	LiDAR-data file format in ArcGIS pro	20
LiDAR	Light Detection and Ranging; i.e Airborne laser data	14
ML	Machine learning	23
PF	Precision Forestry	16
PHC	Pre-harvest clearing	13
PTC	Pre-thinning clearing	26
RGP	Relative ground points	31
SLU	Swedish University of Agricultural sciences	18
TWI	Topographic wetness index	22
ULCD	Understory LiDAR Cover Density	30
UP_f	Understory vegetation points	31

1 Introduction

This chapter starts by giving a brief introduction to the concept of pre-harvest clearing, and AI's integration in today's forestry, followed by how these two concepts have inspired this project

1.1 Problem background

Forestry is dependent on mechanized and digitalized systems to keep up with the interchanging conditions for modern forestry (Vatandaslar et al., 2025). This is due to the many advantages it brings to improving efficiency and accuracy, and thus economic viability, profit margins and potential for nature conservation (Sanz et al., 2020; Vatandaslar et al., 2025). However, as Nordqvist (2023, p. 2) proclaims - compared to the history of Swedish forestry, things are at a standstill in some respects today. He continues by declaring that during the 1900s, hundreds of thousands of people in Sweden worked in the forestry sector due to the large need of manpower. Starting from around the 1950s however, on pace with mechanization of forestry practices, the sector as an employer steadily decreased as the manual two-man saw was replaced by motor-driven equipment. All up until the development somewhat stagnated during the 1990s.

Nordqvist (2023) continues by stating that ever since the stagnation in the 1990s, the sector has roughly the same labor needs in many important roles such as harvester and forwarder operators, as well as clearers and planters, due to the practices remaining mostly the same. Seemingly not a very alarming issue other than the lack of innovation itself. However, while somewhat improved over the last few years, there is a major lack of people working in these said roles today (Kindströmer, 2024). Nordqvist (2023) thus raises the question:
Is AI the solution to these challenges? Is AI what is finally going to make forestry faster, better, and more efficient?

Putting AI in the somewhat same context as mechanization in terms of innovation-potential for Swedish forestry is arguably ambitious, since mechanization has had an enormous impact on daily practices and processes. However, there is undeniable potential with AI as a tool in forestry. The growing use of and interest in AI shows several good examples of how it might be used to move towards more automated processes in the forestry sector.

The Swedish Forestry Agency (2026) recently released an AI-driven tool which has been trained to recognize features in the Swedish landscape which indicate particularly high ecological or cultural values, which are both important factors in forestry. This is a process which traditionally has required field observations, and generally still does today.

There is even progress being made on creating three-dimensional digital models of forests. Von Essen (2023) reports on a minor business in Norrbotten county which is doing this with the help of AI and cameras, and hopes that this will be a key to getting detailed information about every individual tree, which will help improve both precision and profitability in forestry.

The future of forestry therefore holds much potential in applying such technology, and the evolution will be exciting to unveil. However, as mentioned, there are still many parts of forestry which are far from the AI-driven levels presented above, and still require manual labor.

While Nordqvist (2023) has high hopes for it, forestry is arguably quite far from being fully automated in the contexts of clearing, forwarding, harvesting and planting. Clearing for example, is still done almost exclusively manually both in terms of planning (Sanz et al., 2020; Holmen Forest, n.d.a) and the clearing in practice (Swedish forest agency, n.d). Clearing as a concept is quite wide however, since it has several different methods and motives (Swedish forest agency, n.d). The main goal of clearing is generally to increase productivity in forestry operations. Clearing is mostly done in young stands to remove undesired vegetation which might reduce the growth of the stems where growth is sought after (Ibid.). However, clearing is also sometimes done in older stands, often before thinning or harvesting. During these types of clearing, the vegetation is instead removed to make machine operations easier by improving vision and making it easier to operate heavy machinery in the terrain (Ibid.), improving **GROT** (Harvesting residue used for energy production (Holmen Forest, n.d.a; n.d.b); i.e branches and tops) quality, or as a measure to improve the regrowth conditions (Holmen Forest, n.d.a).

This project focuses on the type of clearing which is done in connection to harvesting. This is also known as "pre-harvest clearing" (**PHC**). While this concept might seem clear in its goals, it is still a complex process to determine why and where PHC is needed. There are namely different ideas of what determines the need in this context, both on an individual and company basis. Therefore, today's practices leave much to delve deeper into how today's means of determining the need of PHC are done.

As previously presented, PHC is generally undertaken when there is understory vegetation which might obstruct harvester head engagement, vision for machinery when harvesting, GROT quality, or as a regrowth-promoting measure (Holmen Forest, n.d.a). Relating to the operational context in harvester head obstruction and machinery vision, PHC is done to improve the efficiency of the final harvest (Sanz et al., 2020; Haavikko et al. 2019). A study conducted by Kärhä (2006, p. 108) showed that productivity in spruce stand harvest risks a 10-14% percent decrease when understory vegetation over two meters high is concentrated to 2000 stems per hectare and increases to over 30 percent when exceeding 10,000 stems per hectare (Ibid., p. 109). Therefore, PHC stands as an efficiency-wise important action to undertake in certain circumstances.

1.2 Problem

As of today, determining whether a tract is chosen for PHC before harvesting is still done manually. It is based on forest inspectors visiting the tract in question on the field, and making their decision based on on-site observations (Sanz et al., 2020; Holmen Forest, n.d.a). This is not only quite time consuming, and thus expensive in a corporate context (Sanz et al., 2020), but arguably also makes it prone to subjective human assessments of when PHC is considered. With that stated, this project does not intend to invalidate on-site observations in a logical positivist fashion (Kanazawa, 2018, p. 30), since they are in fact performed by trained professionals in corporate contexts. However, there is always reason to have some form of empiric skepticism whenever individual assessments are done, since humans can arguably never reach an objective reality (Ibid., p. 29).

The study by Sanz et al. (2020) presents the results of a questionnaire in which respondents were asked to classify the need of PHC on a scale from 1-5, from high need to no need, in pictures of different tracts. The study showed that the respondents had quite different views of when PHC was needed and at what scale, thus providing an insight into the subjective nature

of the assessment of when PHC is considered needed. This aligns with the ideas of empiric skepticism (Kanazawa, 2018, p. 29), as well as with Vatandaslar et al. (2025) where they argue that forest management and planning on an industry level could benefit from a more streamlined and normalized process. Together with Dash et al. (2016, p. 20) Vatandaslar et al. (2025) continue by arguing that a key to reaching this is through working more with digital methods.

Dash et al. (2016) argue that **LiDAR** data has been one of the biggest contributors to digital method progress in forestry. LiDAR is an airborne laser scanning method, which with the help of an aircraft can create 3D-models of forest from above, and separate different terrain elements based on the intensity of the returned lasers (Ibid.). However, Vatandaslar et al. (2025) still argue that this method is underutilized, and its potential has yet to meet its full potential.

In the same paper as the questionnaire, Sanz et al. (2020) also managed to create a predictive model based on LiDAR data, with roughly 64 percent certainty using linear-regression models. Sanz et al. (2020) therefore demonstrate the potential of LiDAR-data and how it might contribute to PHC planning and decision making. However, the linear-regression model only maps the understory vegetation based on predetermined areas. Put in the context of practical industry led forestry, it might prove difficult since the objective is often to only clear the areas where it is absolutely necessary (Holmen Forest, n.d.a).

The linear-regression model can therefore arguably not replace the manual observations when deeming which specific areas in a tract need PHC and not. However, it still demonstrates the potential of digital tools, either used by themselves or in tandem with manual observations. This project therefore hopes to contribute to the development of digital tools in PHC planning but not based on linear models. Instead, this project intends to follow the proposed AI potential presented by Nordqvist (2023), using machine learning tools integrated in **GIS** (Geographical Information systems), in order to explore whether it is possible to draw remote conclusions on tracts where PHC needs are not predetermined.

1.3 Aim and delimitations

This study aims to explore the potential of GIS Machine Learning and LiDar data in PHC planning, and how these methods could contribute to the decision-making within PHC assessments. To address this, the research questions are:

- To what extent can a remote sensing random forest classification Machine learning model predict PHC needs in Swedish production forests?
- Which remote sensing variables contribute most to predicting pre-harvest clearing needs within the frame of this study?
- How well does the developed model predict on data from unseen tracts?

This project is largely based on data acquired from Holmen. I would like to acknowledge the challenges it might bring in generalizability and further research. Since the data is based on decisions made based on Holmen's internal PHC guidelines, the defined PHC needs might differ from other companies (Forsberg & Lodén, 2020) in terms of vegetation intervals, for example. With that said however, this project still hopes to contribute to the PHC remote analysis context in its entirety and demonstrate the potential of GIS and Machine learning-

based tools.

An important matter to point out in this process is that areas with high nature values, i.e “protected areas”, were excluded since this project focuses on the binary division between PHC need and no PHC need. Each tract where a forestry measure is to be undertaken gets a tract directive from a forestry planner. In this directive the aims and areas affected by the measures are briefly outlined so that the individuals doing the clearing or harvesting have written guidelines for the measure. In the directives, the consideration areas and other matters with need of attention are outlined. This project has therefore excluded the consideration areas based on the outlines in the tract directive. More information on the process behind this will be explained in the GIS process part of the methods section.

The geographical extent (See: Figure 6) of this project was determined by the data acquisition dates performed by Lantmäteriet (n.d.b). Therefore, the data used in this project is limited to a relatively small part of Sweden, which emphasizes the inability to generalize this project on a larger scale.

2 Theory

Chapter 2 explains the theoretical basis of the project. This project focuses on remote analyses, GIS and non-linear machine learning methods of PHC analysis. Therefore, this section will explain the remote analysis advantages and potential with grounding in precision forestry, basic GIS concepts, as well as the process, types of, and metrics behind machine learning and the random forest method specifically

2.1 Precision forestry

A central concept behind the core principles in this project is 'precision forestry' (**PF**). As defined by Taylor et al. (2006, p. 399), PF means:
"planning and conducting site-specific forest management, activities and operations to improve wood product quality and utilization, reduce waste, and increase profits, and maintain the quality of the environment."

Taylor et al. (2006, p. 399) define three main pillars to help meet the goals imbedded in the PF definition:

1. Assisting forest management and planning with the help of geospatial-information. This can be seen as the most fundamental part of PF. It refers to helping improve the planning stage of forest management and contributing to improving efficiency and accuracy in decision-making.
2. Place-specific forestry actions. This pillar refers to a more practical context. It focuses on how certain actions are undertaken in relation to where it is and what the place-specific conditions hold.
3. Meeting market demands and maximizing market value. This category arguably also belongs in a more practical context, where it aims to improve efficiency along the whole chain from tree to industry.

Vatandaslar et al. (2025) argue that the concept of PF has evolved over time. They propose that PF has gathered certain additions in terms of how it might contribute to practical contexts in terms of which category, with the help of which tools and where there are gaps between the concept and its practical application. They argue that this evolution is largely due to technological advancements in spatial data and AI, for example, as well as scientific progress within the concept itself. Therefore, they argue that there are more categories which apply to today's means of PF, and they have also gathered which tools it encompasses.

While the concept's goals and potential contributions might seem clear, Vatandaslar et al. (2025) argue that its practical implementations are lacking. It is therefore considered as concerning that the PF driven tools are underutilized, since there are still many ways that they could contribute to improving contemporary practices. Therefore, this project aims to utilize aspects of the framework proposed above, and therethrough hopefully contribute to advancements within the field. Considering the broad scope of the framework however, this project does not aim to incorporate all factors defined by Vatandaslar et al. (2025) due to their varying roles and purposes.

Most relevant for this project are the aspects of *technological advancements* and *workforce transitions* as *drivers*, *forest management* as *category*, *GIS/**CAD** (Computer-aided design)* and *Remote sensing* as *tools,* and *Automation* as a *gap*. Within each of these aspects, there are also different dimensions in terms of their practical contributions. Therefore, the aspects and

dimensions which will be referred to in this project are defined below, in line with their definitions by Vatandaslar et al. (2025, p. 4):

Technological advancements and *workforce transitions* as drivers for PF - Vatandaslar et al. (2025) argue that there is a common expectation that increased utilization of digital means in forestry can contribute to balancing and meeting societal demands and services from the forest. However, they continue by pointing out how this in some ways might over-complicate tasks which are already complicated. Considering that the process behind deeming needs for PHC, for example, is done by trained professionals who have a high level of competence in forestry, the new expectation on this group of professionals would then result in expecting them to apply methods which require a high technical competence. This issue is arguably quite a challenge for forestry professionals. According to Vatandaslar et al. (2025), this is a part of the reason as to why forestry, compared to certain other sectors, applies innovations at a slower rate - and rather focuses on the existing processes due to their contemporary viability and profitability (Weiss et al., 2021; Weiss, 2019).

Vatandaslar et al. (2025) are quite eager to point out that the aging workforce in forestry with a traditional approach is also part of the reason as to why innovation is slow in forestry compared to other industrial sectors, and that the coming generation should have a more technical curriculum. While this might be part of the reason, one could still argue that focusing on shifting the workforce is neither a realistic nor efficient response. Rather, one could argue that it would be a more realistic idea to create more user-friendly tools which can utilize the readily available data and tools such as LiDAR-data and GIS and combine it with the forest competences in the existing workforce. Therefore, this project focuses on the *technological advancements* as the main driver for PF, but also acknowledges the *workforce transition* aspect. However, in this project I do not regard *workforce transitions* as a factor where replacing the workforce with younger and more technically profiled individuals is central, but rather intends to focus on making the methods accessible for all parts of the workforce.

Forest management and planning as a category of PF - This category focuses on contributing to the management and planning stage through using digital means. The goal is therefore to make data collection of tracts more efficient through using LiDAR-data for example, which is quite accessible today. The idea of using digital means is to contribute to normalizing and streamlining the management process. This is highly connected to the approach in this project due to its ambitions to reduce the manual labor required for deeming PHC needs, as well as creating a quantitative approach to deeming the same need.

Remote sensing as a tool for PF - In Vatandaslar et al. (2025) this tool refers to all data which is collected from spaceborne, airborne, or ground-based platforms. The data itself might then come in the form of ortophotos, radar, LiDAR-points. As mentioned throughout the paper, this project will concentrate on the LiDAR aspect of this toolbox.
GIS/CAD as a tool for PF - GIS will be the main tool for this analysis. Thus, the CAD aspect of this tool will not be touched upon. As Vatandaslar et al. (2025) declare, GIS was not widely applied as a tool in forestry until around 2010, and they argue that its potential is still underutilized.

Automation as a gap: Vatandaslar et al. (2025) identify three main gaps for PF to contribute to. *Integration, Equity* and *Automation.* This project hopes to contribute to the *automation* aspect. As Sanz et al. (2020) argue, manual PHC decision-making is a quite labor-intensive

process due to it requiring field-visits and an intricate assessment (Holmen Forest, n.d.a). An automated process could therefore contribute to streamlining PHC assessments. Figure 1 visualizes how drivers and the three PF aspects are connected. The example only represents the aspects seen as most relevant for this project.

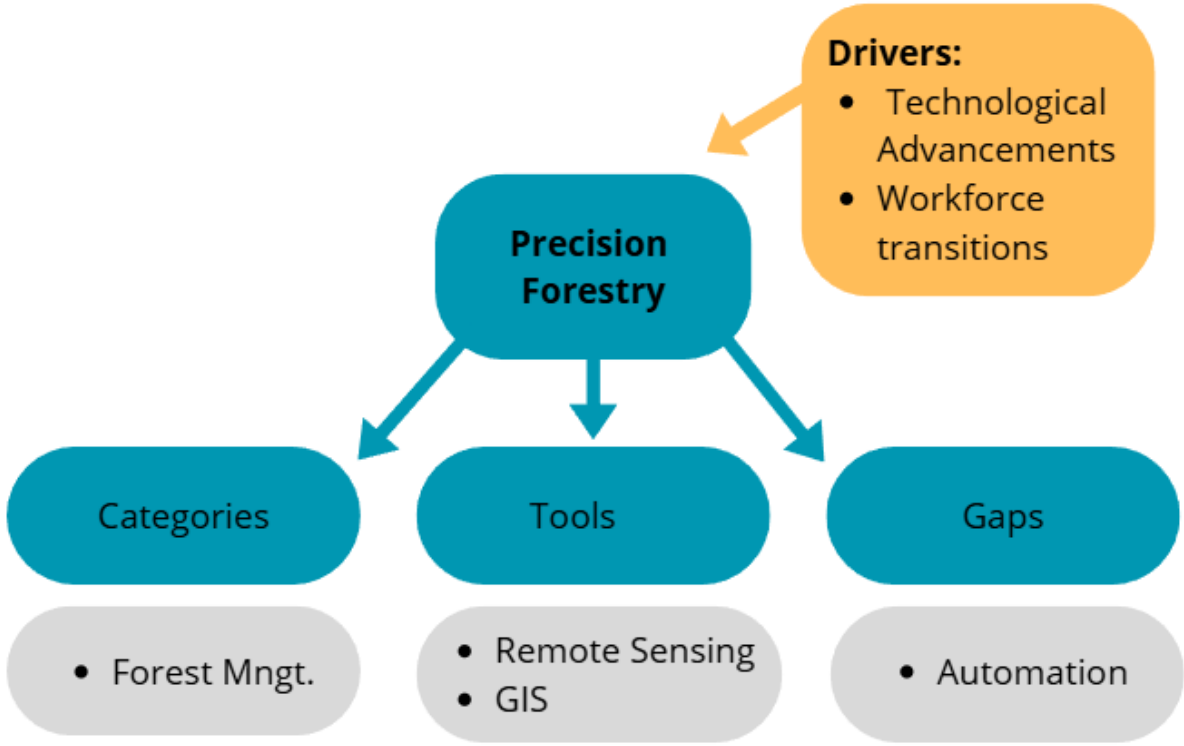


Figure 1. A modified Precision Forestry framework showcasing the drivers and components relevant for this project (Vatandaslar et al., 2025, p. 4).

2.2 Remote sensing and Geographical information systems (GIS)

This section introduces the concepts of remote sensing and Geographic information systems (GIS). The GIS-part explains important fundamentals of GIS, and how it relates to moving between two- and three-dimensional settings.

2.2.1 Remote sensing and GIS

Remote sensing is the concept which is focused on collecting information about an object of interest without physically examining or inspecting it (Lillesand et al., 2015). The data which allows this kind of analysis is gathered through several different kinds of censoring, often airborne through drones, planes or satellites (Merry et al., 2023). Remote sensing is therefore an important concept for precision forestry-related analyses due to the spatial data basis it brings (Ibid.)

2.2.2 Two-dimensional GIS

The main analysis in this project will be done using the GIS-software ArcGIS Pro by ESRI (Licensed through SLU). The specific GIS-tools and workflows will therefore refer to this

software in particular. Information regarding the used tools used in this project can be found on https://pro.arcgis.com/en/pro-app/latest/tool-reference/main/arcgis-pro-tool-reference.htm.

GIS has historically been seen as something separate from the real world and rather bound to computers as a form of interpreting and handling spatial data (Longley et al., 2015). However, Longley et al. (2015) continue by presenting that on pace with digitalization in society, GIS-based tools have become increasingly integrated in daily lives to support everyday spatial decisions as in navigation, for example.

Generally, there are two ways of representing spatial data in GIS, rasters and vectors (Longley et al., 2015). Rasters consist of equally sized cells which form a network of the area in question. Depending on the resolution, the cells represent different sizes which can vary from resolutions of 1 square meter, all the way up to hectares in some cases, depending on how the data was acquired. Within each of the raster cells, a certain value, or attribute, is represented. These values might for example include different land covers, where different colors represent different land types.

Vectors represent objects in more irregular patterns than rasters, and therefore come in the shape of points, lines or polygons. Polygons, somewhat similarly to rasters, represent a certain value within them. However, their shape is not dependent on its surrounding attributes and can therefore be any shape. In relation to rasters, vectors can therefore be advantageous when representing areas of continuous values. Where there might be a need to represent a large continuous forest, vectors can do that with only one object. Meanwhile, a raster could potentially need thousands of cells representing the same object and thus making it much heavier in terms of data. Figure 2 visualizes the differences between vectors and rasters.

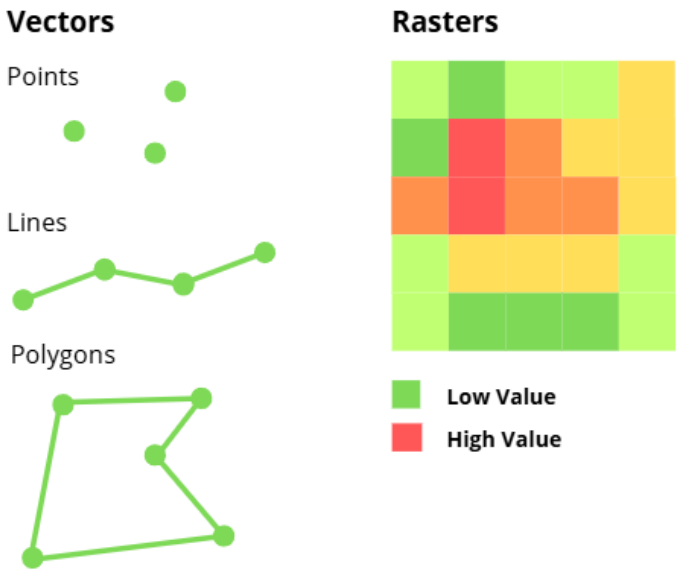


Figure 2 Polygon and raster examples (Longley et al., 2015, p. 69).

2.2.3 Three-dimensional GIS

Areas are two-dimensional (Longley at al., 2015, p. 64), therefore rasters and vectors are good ways of representing the world when areal-based spatial questions are processed. However, considering this project's ambitions to separate different types of vegetation depending on their height, spatial distribution solely on a two-dimensional scale would arguably not suffice. Thus, a three-dimensional aspect must be included. Three-dimensional objects can be represented by so-called point clouds in a **LAS** dataset (LAS is the LiDAR data format in GIS (Esri, n.d.f)). What differs this from the traditional ways of working with rasters and vectors, is that a z-value (i.e height value on the Z-axis) is added, and can therefore let point clouds represent 3d-objects
(Esri, n.d).

For this case in particular, the z-values will be measured through LiDAR-data (light detection and ranging) (Longley et al., 2015). LiDAR-data is acquired through airborne laser scanning, where a range finder scans the ground from above and measures the radiation which is deflected back when it meets an object. When the laser impacts an object, the object produces a so-called point cloud. The point cloud contains not only a longitudinal and latitudinal value (x- and y- values), but also an elevational value (z-value). Figure 3 shows an example of a LiDAR point cloud representing tract 410987952 Råhagaberget from above. The point cloud is symbolized based on each point's elevation above sea level, ranging from roughly 220 to 280 meters.

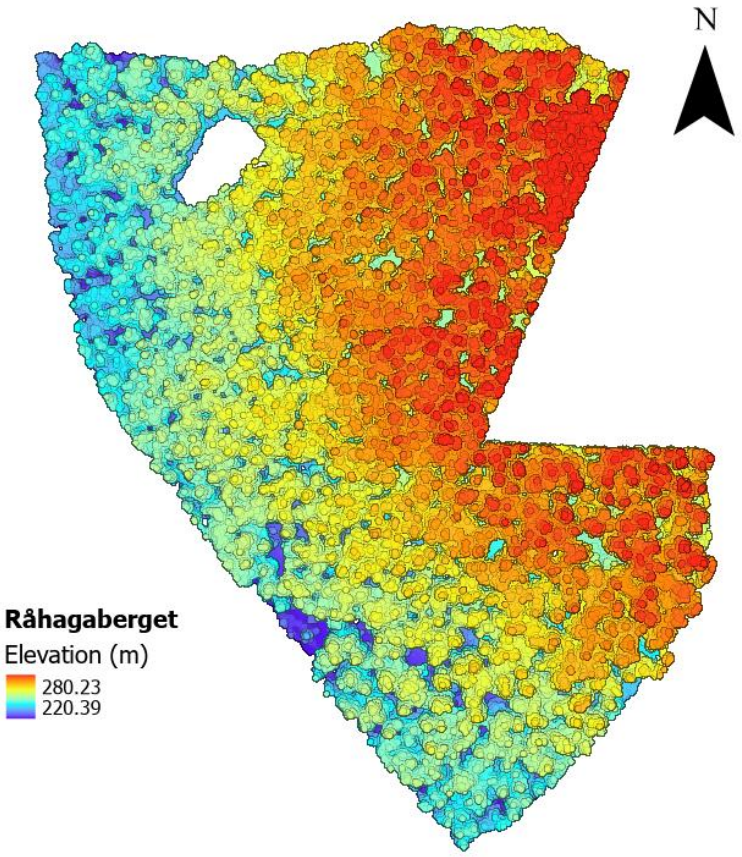


Figure 3. Lidar point cloud model of 410987952 Råhagaberget.

2.3 Machine Learning

This section introduces fundamental machine learning principles, an explanation of the model used in this project and how machine learning model performances are assessed

2.3.1 Machine learning fundamentals

Machine learning relies on building algorithms which are based on different kinds of data from a certain instance or phenomenon (Burkov, 2019), where the algorithms learn the patterns in the data, and can help predict the outcome in similar contexts. There are different ways of practicing machine learning, namely through supervised, semi-supervised, unsupervised and reinforcement learning (Ibid.). The first three are, as the names might reveal, quite similar and mainly differ in level of data labelling, and in other words, supervision. Reinforcement learning leans towards a state where the machine itself can perceive and learn from its environment.

This project focuses on a supervised method. Supervised learning means that every input feature that is processed in the model has a certain value, and is connected to an output (Burkov, 2019). Inputs can generally be in the form of anything that holds any kind of information, and the outputs can be either numbers or labels.

In this context, the inputs in the model are the geodata produced in the methods section, as in land cover, lidar-based vegetation density rasters, topographic wetness index, and so on. In ArcGIS, inputs are referred to as explanatory variables, thus it will be referred to as *variables*.

As mentioned earlier in the methods section, all variables have been fitted into 5x5 raster cells. This resonates with the machine learning fundamentals of creating *feature vectors* (Burkov, 2019). A *feature vector* is namely all variables which describe the unit that is being processed in the machine learning model (Figure 1).

Table 1. An example of two feature vectors

	Elev.	**TWI**	**Veg. Dens. C3**	**Veg. Dens. C4**	**Veg. Dens. C5**	**Veg. Dens. C6**	**Slope**	**Tree cover**	**PHC**
X	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	YES
Y	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	NO

The *feature vector* has therefore allowed us to translate the information about each unit into data which the machine can read. Considering that the study is based on PHC-decisions covering roughly 203 hectares, the data covers a vast amount of 5x5 cells, each of which has a feature vector.

Each raster cell containing these different values has then been assigned an output through the feature classes created, either as in need of PHC or not, based on the data from Holmen Forest. The idea is therefore to learn the machine to see the relationships between the variables to attempt classifying new, unlabeled, data (Burkov, 2019). In this project, the process behind this is called random forest classification.

2.3.2 Random forest classification

As its name might reveal, Random Forest classification is based on combining several decision/classification trees in a "voting" manner through a process called *bagging* (bootstrap aggregation) (Breiman, 2001; James et al., 2021), which will be explained in further detail below.

A decision tree in itself is a quite straight-forward concept. In binary classification the tree consists of an original root node, where the node is split up into two branches based on which variable separates the two classes the best. Thereafter, depending on if the tree has managed to classify or not, more internal nodes are created for further evaluation, or a classification threshold is found and resulting in a leaf (i.e terminal node) (James et al., 2021).

Put in the context of this paper, that could mean that the tree would classify all units of analysis with a vegetation density class 3 over a certain value as in need of PHC. This part is called the *root node*. Meanwhile, there could still be units below that value which are still considered as in need of PHC. Therefore, the tree can help further divide them based on another variable which helps separate the classes. For example, the next node separator, or *internal node*, could be **TWI** (Topographic wetness index), where all units with a lower value than the Class 3 vegetation density node separator, but a higher TWI value than the second node split could be classified as in need of PHC. This process continues until all units have been classified or split up into *terminal nodes* and therefore the classification tree might help deem the PHC need. Figure 4 is an example model of the decision tree process. It shows vegetation density class 3 as best split, and topographic wetness index as the second-best split, with PHC/NO PHC need classifications represented in green and red.

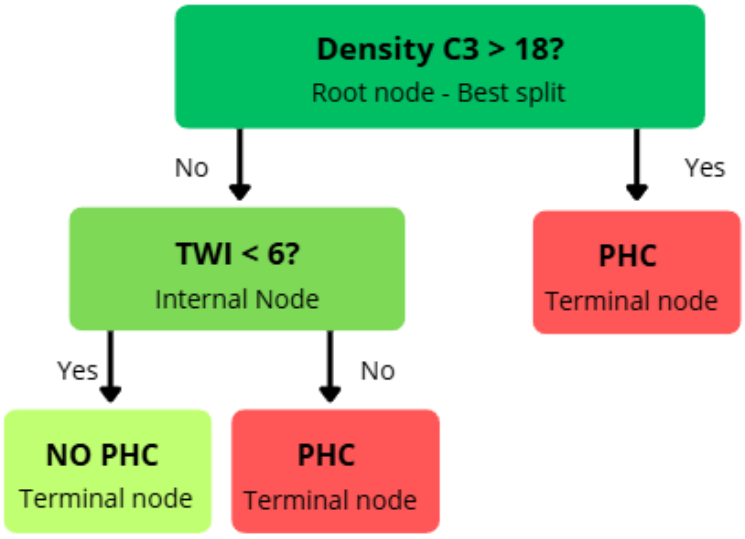


Figure 4. A modified example of the decision tree process (James et al., 2021).

However, while it might bring clarity in certain cases, the decision tree method is still considered as a quite weak predictor. As James et al. (2021) argue, if a tree is trained on two different datasets, the outcomes will probably vary greatly. This is due to the trees learning

the exact pattern which the data shows, which in turn means that when new data is processed in the model it will try to fit it into the same patterns as was in the training data. This consequence is called *overfitting*. Thus, a single decision tree is rarely a good tool for prediction. Therefore, in order to create a model that is less prone to overfitting and has higher prediction potential there is a method called *ensemble modelling.* The idea behind *ensembles* is building a single model based on several simpler models (Ibid.), as a decision tree, for example. This can roughly be understood as creating a model with several perspectives rather than one, which is the basis that connects the random forest method to decision trees - several trees form a forest.

However, the trees in random forest models are not quite the same as in a decision tree explained above. Instead, the trees in a random forest model are based on random samples from the dataset. This is known as resampling (James et al., 2021). The type of resampling used in random forest modelling is called *bootstrapping.* The bootstrapping process picks random parts of the base dataset, and simulates new datasets for the number of trees chosen for the analysis. In addition, the Breiman (2001) random forest method also chooses random features to divide the observation. This means that instead of choosing the most fitting variable to divide the observations with, each tree gets a random set of variables to work with, which further increases the variation between the trees.

For example, if the model has 100 trees, 100 bootstrap samples are taken from random parts of the data and then used to create the trees in the model, similarly to if there would have been 100 different datasets. In addition, the feature randomness also steers the trees to learn the classification connections based on different parameters. Therefore, all trees are somewhat different from each other. This is a very important part of the process: to create diversity in trees and thus the resilience to variation (Breiman, 2001), since the trees have learned several ways to recognise the patterns.

Once the trees have been created with different data and different features, the model continues to the “voting” process mentioned in the start of this section. When attempting to predict the classification value, as the PHC needs in this case, each of the created trees run the new unit through their nodes resulting in a different answer for each tree. A number of 39 of the trees might yield that the unit is not in need of PHC, while 61 yield the opposite. Thus, it becomes a vote of majority, where the majority vote would yield that the unit is probably in need of PHC. This step also demonstrates the importance of variation in the trees, since without the diversity in the trees, all trees would yield the same answer, and thus not differ much from a single traditional decision tree.

2.3.3 Model assessment

When assessing the performance of a predictive ML-model (machine learning model), the most interesting factor is how well it performs on unseen data. Deeming how well a model performs on unseen data can be done in two different ways - either through validation where a share of the training data is excluded for validation, or through testing on external data (Burkov, 2019).

In the chosen ML-tool in ArcGIS Pro, the share of validation data is chosen manually, but the data selection is done randomly by the model (ESRI, n.d.c). A test set however, is a predetermined dataset where the model is run in a more practical context. There is no benchmark for dividing the training, validation and test data. It is rather decided by the size of the dataset. In cases with very large datasets, the validation and test data might attain only 2,5

percent respectively, with the other 95 percent being left for training (Burkov, 2019). If a model has a high success rate in predicting the validation or test data, the common term is that the model *generalizes* well (Burkov, 2019).

These two methods of testing fill quite different purposes though. The validation yields a more detailed result in terms of how the model performs in predicting unseen data, which can help set the model parameters for reaching the models full potential. The test run is rather focused on running the model in a more practical context, which is often the most interesting assessment for a client, if the model is created in a context as such (Burkov, 2019). Whether a model generalizes the validation data well or not can be measured with several different metrics. For classification, the most common metrics are confusion matrix, accuracy, precision/recall and ROC curves (Ibid.).

Confusion matrix metrics are based on two-by-two tables which summarize the validation results (Burkov, 2019). The table is structured on two axes where one axis represents what the model predicted, and the other what the actual label is (Table 2). The results from each prediction are then presented as true positive (TP), true negative (TN), false positive (FP), false negative (FN).

Table 2. Example of a confusion matrix with fictional values

	PHC (Predicted)	**NO PHC (Predicted)**
PHC (Actual)	75 (TP)	6 (FN)
NO PHC (Actual)	20 (FP)	63 (TN)

This example matrix represents a case where 75 PHC need predictions were correct (TP), and 20 PHC need predictions were incorrect (FP). Respectively, 63 NO PHC predictions were correct (TN), and 6 were incorrect (FN).

While the confusion matrix is seen as a performance metric of its own, it is also what other machine learning metrics are based upon. The tool used for analysis in this project yields F1-value, Matthews Correlation Coefficient (MCC), accuracy and sensitivity (also known as *Recall* (James et al., 2013)) (ESRI, n.d.c). All of these metrics are based on the confusion matrix, but with different formulae (Burkov, 2019; Chicco & Jurman, 2020).

The **F1-value** is a product of the following formula

$$F1 = 2\,x\,TP\,/\,((2\,x\,TP)\,+\,FP\,+\,FN)$$

As represented in the formula, the true positive has much influence on the final F1-value. Thus, which category is defined as positive and negative therefore affects the F1 outcome. Chicco & Jurman (2020) argue that this is a potential source of error in binary classification, even though the F1-value is a quite established metric in machine learning. With Table 2 as input, we get the following results

$$F1 = 2 \times 75\,/\,((2 \times 75) + 20 + 6) = 150\,/\,176 = \mathbf{0.852}$$

Matthews Correlation Coefficient (MCC) is a product of the following formula

$$(TP \times TN - FP \times FN) / \sqrt{((TP + FP)(TP + FN)(TN + FP)(TN + FN))}$$

Chicco & Jurman (2020) argue that the most realistic performance depicting metric in a classification setting is the MCC since it requires good performance on both classes to yield a high value. Therefore, it can handle imbalanced datasets better than the other two metrics. The class dependency in its formula also enables the MCC to move between -1 (no correct predictions), 0 (correct on half of the predictions) and 1 (perfect predictions). Chicco & Jurman (2020) continue by stating that the MCC should be the metric of choice in binary classification. With Table 2 as input, we get the following results

$$((75 \times 63) - (20 \times 6)) / \sqrt{(95 \times 81 \times 83 \times 69)} = 4605 / 6641 = \mathbf{0.693}$$

The **accuracy** metric is based on the following formula

$$(TP + TN) / (TP + TN + FP + FN)$$

It depicts the relationship between correct predictions and total number of observations. Chicco & Jurman (2020) argue that this metric works well on balanced datasets. However, when the dataset is imbalanced, the accuracy might yield misleading results if one class has a large majority and performs well in prediction, and the smaller class prediction performs poorly, the poor performance in the smaller class risks being engulfed by the majority. With Table 2 as input, we get the following results

$$(75 + 63) / (75 + 63 + 20 + 6) = 138 / 164 = \mathbf{0.841}$$

Sensitivity is based on the following formula

$$TP / (TP + FN)$$

It is a metric which describes the proportion of true positive predictions compared to total positive observations (Burkov, 2019). In practice, this metric only shows the performance of one of the class predictors. Therefore, similarly to the F1-value, this metric is highly affected by which class is defined as positive or negative. With Table 2 as input, we get the following results

$$75 / (75 + 6) = \mathbf{0.926}$$

In summary, these metrics represent quite different aspects of the confusion matrix. While Chicco & Jurman (2020) argue that some of these metrics are inappropriate for binary classification, the tool in ArcGIS Pro automatically mitigates some of the alleged flaws. As for the F1-value and Sensitivity, ArcGIS Pro helps mitigate the positive/negative class definition impact through providing two separate values where both classes are treated as positives one time each, and then combining the results to a total value (See: 5. Results). Therefore, this project will acknowledge all metrics as valid measurements for model performance.

Another potential source of error pointed out by Chicco & Jurman (2020) is imbalanced datasets. ArcGIS Pro does unfortunately not offer mitigation for this. Since the dataset in question is somewhat imbalanced, the MCC metric might still yield the most accurate results.

Therefore, in line with Chicco & Jurman's (2020) arguments, the MCC will be treated as the main metric in contexts where only a sole metric is presented, but all variables will be presented in more detailed sections.

3 Background for the empirical study

The following chapter gives a brief introduction to both the concept of PHC and clearing in its entirety, and how it might differ between different corporate and forestry contexts

3.1 Clearing as a concept

As mentioned in an earlier section, there are several different types of clearing which have different objectives and purposes (Swedish Forest Agency, 2012). Most commonly, it is done as a form of forest management action in young stands where undesired vegetation is removed in order to promote growth in the trees where growth is desired. The general idea behind it is therefore to generate long term economic advantages. In practice, this is done through concentrating growth resources to fewer stems. In pine stands however, it also serves a purpose to prevent other vegetation from overshadowing the pine trees, since sunlight is an especially important resource in such stands. Since this form of clearing is done at an early stage in the tree stand, the cleared vegetation is quite feeble, and therefore serves no purpose as a resource beyond leaving in where it was cleared in order for it to be naturally broken down and turned into energy for the remaining stand.

In stands where clearing has been undertaken, it has a considerable impact on the growth, both in terms of height and stem diameter (Swedish Forest agency, 2012). Therefore, this is a quite applied practice since it stimulates the growth, which is sought after in a production forest, and can thus be seen as an economic investment in the forest (Ibid.). However, there are still many contexts and reasons as to why certain stands and tracts are not cleared at an early age. For example, in continuous cover forestry, there is seldom a reason to clear the stand (Ibid.) since the idea is to let understory vegetation grow continuously. Therefore, clearing is highly connected to the clear-cutting methods of forestry (Ibid.). Although, even in clear-cutting contexts the forest is not always cleared in what is considered “normal” time-intervals. This can be due to several factors. There can be advantages in letting “undesired” vegetation grow as a tool to mitigate wild-game-caused damages to the “desired” stems, since ungulates often prefer leaf-trees to conifer (Swedish Forest Agency, 2025), for example.

In contexts like this, the forest will naturally grow quite thick, making it difficult to traverse both on foot, but especially in machinery. Therefore, on time with the first thinning in the tract, a clearing known as pre-thinning clearing (PTC) will be required to reach the selected stems. It is however not uncommon that a PTC is needed even if the tract has been cleared on schedule (Swedish Forest Agency, 2012). This type of clearing is therefore quite similar to the one in focus in this project. The practical conditions for PHC will be described in the following section.

3.2 Holmen’s guidelines for pre-harvest clearing

As described in the introduction, PHC is generally done in order to facilitate machinery operations in the final harvest of a fully mature tree stand. However, the terms for when PHC is needed or not differs between different corporate and operative contexts, thus many forestry companies have their own instructions and thresholds to relate to when deeming clearing needs (Forsberg & Lodén, 2020) .

Since this project is written on a commission through Holmen Forest, the analysis will therefore adhere to the conditions related to the operative context at Holmen. In line with said

operative context, there are three main reasons as to why PHC would be conducted (Holmen Forest, n.d.a). Firstly, there is the context where PHC is done in order to favor operative harvesting conditions. As has been briefly explained earlier in the project, this refers to removing undergrowth which is deemed as obstructive in terms of both vision and harvester head engagement. The conditions for this are quite complex though.

In Holmen Forest's context, this refers to undergrowth which is within 1,5 meters of harvesting stems and over three meters tall. However, undergrowth over one meter tall can also be deemed as obstructive if it can be considered having sprawly, or bush-like, characteristics. Thus, obstructive undergrowth refers to either sprawly undergrowth taller than one meter, or all undergrowth over three meters, within 1,5 meters from harvest stems should be removed when a PHC is conducted. For PHC to be deemed as needed, there must be over 1000 undergrowth stems per hectare which fulfill the requirements, in an area which is at least one hectare. All trees which do not fulfill the stem-specific requirements should be sought to keep.

There are some exceptions within above named conditions though. Since normally, all stems above eight centimeters in diameter at breast height should be kept. In addition, there are separate requirements for the tree-types goat willow, rowan, whitebeam, maple, linden, bird cherry, wild cherry, elm and hawthorn. Trees of this type that are thicker than seven centimeters at breast height are considered as trees with high ecological values and should therefore not be cleared in most cases (Holmen Forest, n.d.a; n.d.b).

Apart from the operative advantages it might bring for machinery, there are two more major reasons for conducting a PHC (Holmen Forest, n.d.a; n.d.b). Firstly, it is done when GROT is being extracted from the stand, since the undergrowth risks mixing into the GROT, which is not desired. Secondly, it is sometimes also done as a regrowth-promoting measure. On pine-tracts where undergrowth in the harvest-stand risks overshadowing the coming pine seedlings, the undergrowth is removed to prevent unfavorable conditions during stand regrowth.

In addition to these two factors, there are also a few minor reasons as to why PHC might be conducted. Mostly due to private landowners demanding it, if the clearing the young stand was not properly conducted, or to promote recreation or reindeer husbandry (Holmen Forest, n.d.a)

4 Method

This chapter will explain which methods were used to reach the results. Firstly, a description of the commission behind this project is presented. Then the process behind the literature review, the data selection and the GIS process with its included tools will be explained.

4.1 Commission description

This project is written on a commission from Holmen Forest. Therefore, much of the conditions and outlines for the project are determined to fit within Holmen Forest's objectives with the project.

Holmen Forest is the forestry focused part of the corporate group Holmen. The corporate group is split in four business areas: forest, energy, wood products and board and paper (Holmen, n.d.c). Holmen is a major actor in the Swedish forest sector. They are Sweden's fourth biggest forest owner (Holmen, n.d.a), as well as one of the largest actors on the Swedish timber market. Their area of operations is nationwide, and stretches from the southern parts of Norrbotten county, all the way down towards Blekinge county (Holmen, n.d.b). Within their area of operations, Holmen's holding is roughly one million hectares of land both in the form of forest and wind farms, hydropower stations, different industry facilities such as paper and sawmills, as well as paper and board refining factories (Ibid.). The dynamics between said forest and industry holdings is what makes the core of Holmen's business strategy - creating climate smart wood and paper products in their own facilities using sustainably extracted timber from their own holding (Holmen, n.d.c).

Holmen (n.d.d) acknowledges the complex interaction of ecological, social and economic factors surrounding forests. Therefore, Holmen practices a future-centered forestry which ensures to balance the different interests of said interaction both today and in the future. Holmen (n.d.d) relies much on external partnerships in order to uphold these ambitions, often linked to the Swedish University of Agriculture (SLU), which is the case for this project as well.

The commission was advertised as "Forestry planning and digitalization/AI", with the goal to evaluate whether it is possible to analyze PHC needs remotely, due to its labor-intensive field-based process as of today. The project extent or geographical focus was not predetermined, thus the project outline was quite free for me to structure on my own based on my own competence and ideas with support from my supervisor at Holmen, who specializes in forestry planning. Therefore, the author holds the responsibility for the method of choice, workflows and conclusions in this project.

The commission also meant that I had access to Holmen's internal database and system for data acquisition, which has acted as base for the analysis itself. How this has affected the data selection will be covered more in detail in section 4.3

4.2 Literature review

Before conducting the analysis, a literature review was conducted. The literature review functioned as a tool to identify and evaluate relevant research to this study topic, and foremost to contribute to the understanding of the issue at hand (Kanazawa, 2018). A thorough literature review does not only contribute to the research quality in terms of understanding the

query, but it also helps avoiding what Kanazawa (2018, p. 49) coins as "*Reinventing the wheel".* Avoiding reinventing the wheel helps promoting the originality of the study through getting an idea of what we already know, and thus knowing how the research might contribute to creating new knowledge.

The first challenge in the literature review is making sure that as much relevant literature is covered as possible. As Kanazawa (2018) proclaims, there are enormous amounts of literature no matter the subject, and many sources where these might be found. Therefore, he continues to explain that internet and library resources are the most efficient. In this project, the starting points for finding literature have been the libraries at the Swedish University of Agricultural Sciences (SLU) and Uppsala University (Access granted through **CEMUS**), Google search engine, Google scholar, Holmen, my supervisor, browsing online database sections and browsing the Swedish Forest Agency's and SLU's webpages.

However, once a set of papers on the subject had been found and reviewed, the literature searching method moved onto a snowballing approach (Wohlin, 2014). The snowballing approach was mostly conducted in a backwards manner. This refers to going through the reference list from the starting set of literature and deeming which literature might be relevant for this project. The first step in this process is excluding literature in the reference list purely based on their titles and purpose in the article at hand (Ibid.). Once the relevant literature had been extracted, the abstract and most relevant parts of each were reviewed to further deem its relevance. If the said parts showed relevance the paper could be included in both the project and as a gateway to further literature snowballing. Considering that two of the main articles from the starting set were quite contemporary (Vatandaslar et al., 2025; Sanz et al., 2020), a backwards snowballing procedure still yielded up-to-date literature.

In order to make sure that more recent literature was not missed however, a forward snowballing procedure was also undertaken (Wohlin, 2014). This was done with the help of Google scholar, where it is possible to see which articles have cited the paper in question. In the cases where the paper had been cited, the same relevance assurance as in the backwards snowballing was conducted.

Once a sufficient number of articles had been found as seemingly relevant through this process, the review of the papers in their entireties could commence (Wohlin, 2014). During this step some articles were still culled, which arguably ensured that only the most relevant papers were included. Figure 5 visualizes the iterative process of backward and forward snowballing (Ibid.).

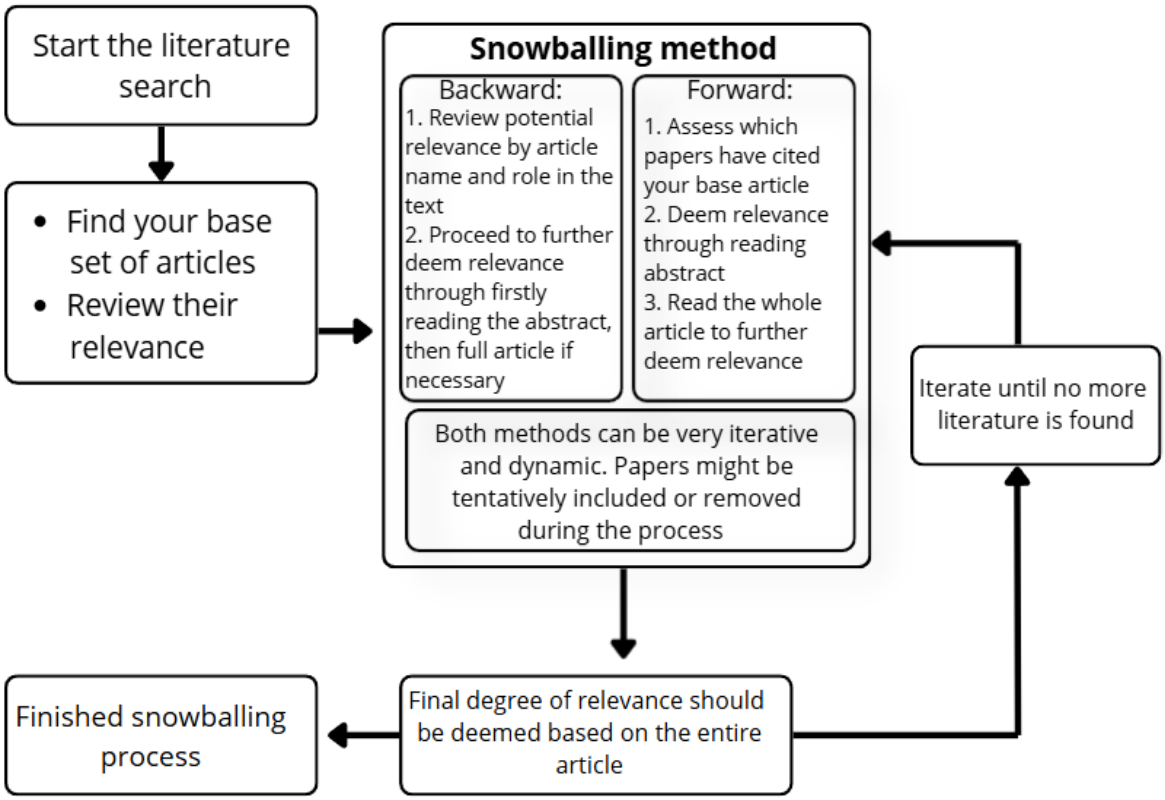


Figure 5. A slightly modified figure showing the literature snowballing procedure by Wohlin (2014, p. 4).

4.2.1 Earlier research

There is a wide variety of literature which touches upon and discusses the concepts of precision forestry and digital tools for forestry. The practical contributions are fewer however, especially in the niche context of PHC. In this sub-section, two examples with similar methods to this project are presented, showing their respective way of working with LiDAR-data and understory vegetation. The first example does not touch upon PHC specifically, and the second uses linear methods unlike this project. However, both papers give valuable insights for the process in this project.

4.2.1.1 The ULCD metric

Wing et al. (2012) conducted a study attempting to predict the understory vegetation on ponderosa pine (*pinus ponderosa*) stands in northeastern California, USA. Their study focus was to contribute to remote analysis of understory vegetation based on its role in forest ecosystems in terms of its influence on wildlife habitats, nutrient cycling and how it relates to fires. The study used LiDAR data with an average point density of around eight pulses per square meter. The experiment produced a metric called *Understory LiDAR Cover Density* (**ULCD**).

The method was based on comparing accurate field measurements from 154 plots, stretching about 40,5m2 each, to LiDAR data using two different linear regression models. The process behind this was to firstly determine a height range for the LiDAR points - in this study the range was 0.2-1.5 meters. Thereafter, the points within this height range were extracted to

serve as base for the model. Before processing the LiDAR points however, the points were filtered based on their intensity. Intensity is the measurement of how intense the reverting laser pulse is (Wing et al., 2012), which the authors found as an important factor since the intensity varied depending on what the laser impacted. Live vegetation, which was in focus in this study, would have intensity values within one standard deviation of the mean. Meanwhile, other objects in the terrain, such as rocks, stumps, etcetera, would have a higher deviation. Thus, the points whose intensity exceeded the standard deviation were removed. Once the unwanted points had been removed, the calculation could be conducted. The ULCD metric was obtained using the following formula

$$ULCD = UPf \ / (UPf + RGP)$$

(Wing et al., 2012, p. 735).

UP_f refers to the amount of understory vegetation points left after intensity filtering, and RGP is the number of *relative ground points* (all points below 0.2m). This model is therefore based on creating a ratio of LiDAR points in the understory height interval to ground points, which provides a normalized way of measuring understory vegetation density.

Using the ULCD metric, they found that LiDAR-based methods can predict understory density with an accuracy of about R2 = 0,7-0,8. While the study by Wing et al. (2012) does not specifically handle PHC metrics, the study still shows promise in LiDAR-based data for remote analyses on understory growth.

4.2.1.2 Voxel based understory density prediction

Sanz et al. (2020) present a model for deeming PHC needs through predicting the number of understory stems based on LiDAR data and a regression model. The model does not aim to identify individual trees, instead it functions as an area-based approach (ABA) which is better suited when the LiDAR data has a relatively low concentration (Ibid.). ABA has been proven to work all the way down to point densities around 1 pulse per square meter (Maltamo et al., 2006).

The first step of their model is creating a digital terrain model (DTM) to accurately distinguish low vegetation from the ground, since the ground's height-value is not linear. Secondly, the height intervals for which vegetation is to be analyzed are set. In Sanz et al. (2020) case it was zero to three meters. Then the intervals were further segmented into 1,5x1,5 meter voxels. Within each of these voxels they analyzed the distribution of point clouds in the vegetation through looking at mean points, maximum points, standard point deviation, point skewness and point kurtosis. These values were then used as variables in their predictive understory vegetation regression model.

The model reached an overall accuracy of roughly 64%, when predicting the PHC need on a scale of 1-5, ranging from no PHC need to compulsory PHC, based on forestry professionals estimations (Sanz et al., 2020)

4.3 Data selection and processing

This chapter will explain which methods were used to reach the results. Firstly, a description of the commission behind this project is presented. Then the process behind the literature review, the data selection and the GIS process with its included tools will be explained.

4.3.1 Geographical selection and LiDAR-data

The most critical part of the PHC-assessment in Holmen's context is based on the density and height of understory vegetation (Holmen Forest, n.d.a). Therefore, a central presumption in this project is that differences in PHC needs might be spotted in the vegetation intervals field-decisions are based upon (See: Section 3.2). Thus, the data selection and acquisition have their starting point in finding the best conditions for finding data which presents the vegetation accurately.

Training machine learning (ML)-models in GIS requires much emphasis on the data selection being as homogeneous as possible, in terms of how, when and where it was collected. Thus, selection of data was highly driven by these factors. The most important factor to relate to in this case was to get as close a match as possible between the collection of LiDAR-data and the PHC in question, so that the LiDAR represents the vegetation which was present when the PHC decision was made. An additional challenge in this was the fact that the LiDAR collections were not conducted regularly, but rather as a mosaic of areas where data was collected (Lantmäteriet, n.d.b). Therefore, a review of when Lantmäteriet (n.d.b) gathered the LiDAR-data was conducted. The hope in this step was to find a larger area where Holmen employees had conducted several PHC decisions and the data collection was relatively close in time between different parts of the mosaic.

Holmen has a lot of operations in southern Västerbotten county and northern Ångermanland county. Therefore, this was one of the first areas which was reviewed in Lantmäteriet's map of data collection. Fortunately, there was a large area between Junsele in the west, Bjurholm in the northeast, and Örnsköldsvik in the southeast, where the data had been collected between 2023-05-01 and 2023-09-31, thus only over the course of a few months. The corresponding area was then evaluated in Holmen's internal system to see if there had been any PHC or plain harvests in the area around the same time. Therefore, the internal search was filtered to look for PHCs and harvests between 2023–05-01 and 2023-09-31. This filtering yielded 43 tracts which could be looked deeper into to deem their relevance for the analysis. Figure 6 shows each tract's location.

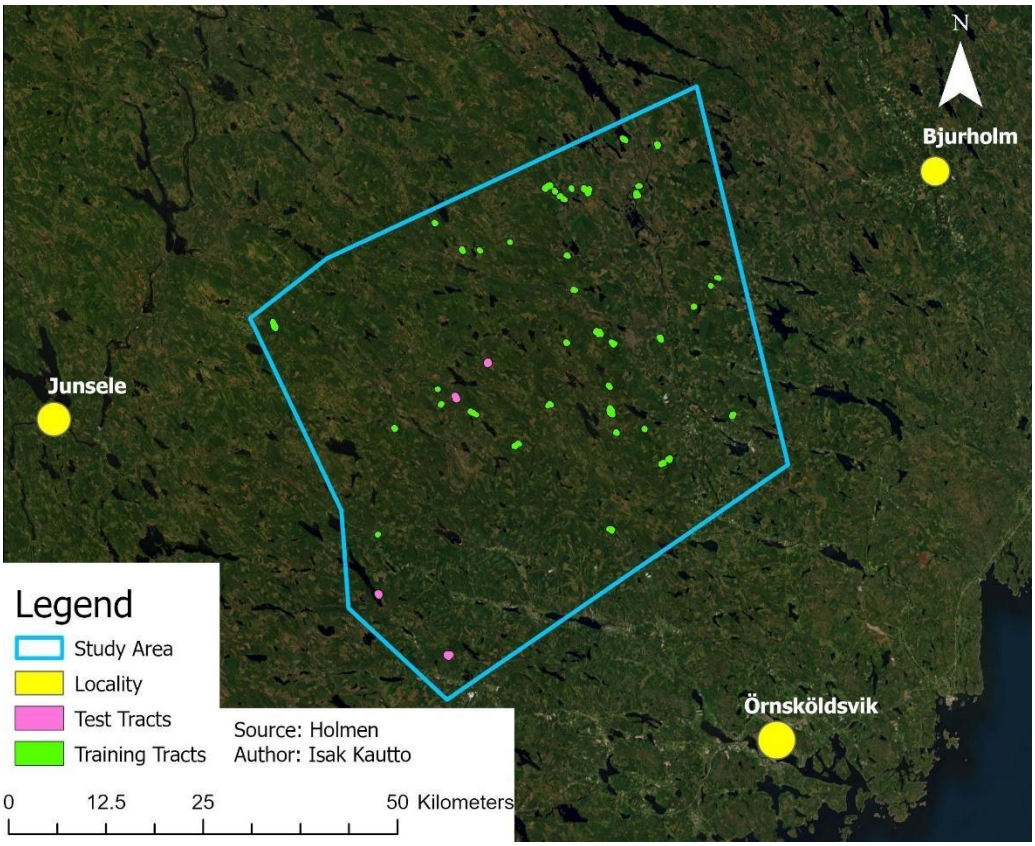


Figure 6. A map of all included tracts. All tracts are represented as either training data or external validation (test) data. The map also outlines the study area with nearby localities as reference.

When deeming the tracts' relevance for the analysis, each tract was assessed to make sure it was eligible both in the lines of this project's focus, but also in line with Holmen's guidelines for PHC (See: section 3.2). This project focuses on the harvesting operational aspects of PHC, thus tracts where PHC is undertaken only to make the harvest easier. Therefore, all tracts with other reasons for PHC than that were excluded. Another important aspect from Holmen's perspective is that PHC must cover an area larger than one hectare, therefore tracts smaller than such were also excluded.

After filtering the search results to PHCs that were done for operative harvesting reasons (See: 4.1), and finding a corresponding set of random tracts where PHC need was not deemed, 43 tracts were found in total. The gathered tracts were quite different in terms of size, shape, and shares of tree types, but they all consisted of either pine, spruce or common leaf trees, such as birch or aspen. These tracts were then randomly divided into training/internal validation data, and external validation (Table 3).

Table 3. Number of tracts and area per PHC need

	Definition	Number of tracts	Area (ha)
Training data	**PHC need**	20	90,2
	No PHC need	19	95
External Validation data	**PHC need**	2	8,4
	No PHC need	2	9,3
Total		43	202,9

4.3.2 Stand data

A GIS machine learning model benefits from having several different types of input data in order to map how much each variable contributes (ESRI, n.d.c). Thus, a few additional variables which all serve a significant purpose in vegetation growth were added.

Albrektson et al. (2012) identify several important factors in stand growth rates. **Stand closure** – refers to the density of stems and low vegetation in the tracts, which impacts the competition for nutrients and water as well as how sunlight might penetrate through the stand. **Topography** – refers to both the elevation over sea and how the ground slope is shaped in the tract. The slope affects how water flows through the soil and sometimes how slopes might increase sun penetration. **The texture of the mineral soil** – This factor also affects the movement of water, as well as the grounds' ability to hold water. **The weathering propensity of the mineral soil** – how the rock minerals are broken down into nutrition available for the trees. This affects the soil's nutritional values. **Temperatures** – This affects which type of tree is chosen for the tract, how the ground is managed and which regrowth method is used. **Stand history** – How the tract has been managed before, i.e what damages might have been caused, and which trees have been chosen before.

While it would be preferable to include all these factors in the analysis, some aspects are arguably difficult to translate into comprehensible map-based information with the limited time frame of the project. For example, compiling stand history of 75 tracts in itself could be worthy of a master's thesis. Other factors such as temperatures and soil properties as weathering propensity and textures are also difficult to capture in the same high resolution as above-ground characteristics, thus they have also been excluded. The remaining factors however, are all included in different forms, as presented below.

Elevation data - Similarly to the vegetation density data, the elevation data, or Digital Terrain model (DTM), is also created with the help of LiDAR data (Lantmäteriet, 2019). When LiDAR–data is collected through **ALS** (Airborne Laser Scanning), the laser can sort ground points from vegetation thanks to its difference in intensity when the laser is reflected (Dash et al., 2016). Thus, an elevation model can help show the elevation above sea level in meters over the whole study area, which is an important factor identified by Albrektson et al. (2012).

Slope - As mentioned by Albrektson et al. (2012) topography in terms of slopes might impact how sunlight penetrates the tract due to how the canopy is divided on different levels, thus a

slope raster was also included. The slope raster was derived from the DTM mentioned above. The slope raster used for the analysis was in radians rather than the degrees format yielded from the slope tool. Thus, the values spanning from 0-2 in this project reflects the radian formula
$1\,radian = \pi / 180$ (Ermold, 2019).

Soil moisture - As presented by Albrektson et al. (2012) a major factor for forest growth rate is water. Therefore a soil moisture layer was included as part of the analysis.
This data was collected from SLU-GET. The data is in the form of a raster which classifies the soil moisture in 2x2 cells between: 1. Dry-healthy, 2. Moist-healthy, 3. Moist-wet and 4. Water bodies; i.e lakes and streams. The data itself consists of a flow accumulation raster (SLU, 2020). A flow accumulation raster is derived from a height model, based on the assumption that the water follows the downstream patterns in the topography. With this information, the flow accumulation for each specific cell is calculated through how many upstream "contributing cells" there are to the cell in question, thus showing how much water flows through each specific place.

While the flow accumulation raster shows how the water flows, it is also of interest to know where the water might accumulate (Albrektson et al., 2012). Therefore, a TWI is also included in the analysis. A TWI is largely based on the same principles as the water accumulation, but takes the process one step further by including slope values. The slope shows how likely it is that the water keeps flowing from that cell in particular, thus showing how wet the cell is expected to be (Ermold, 2019). High index cells are therefore often concentrated in flat areas where the water is probable to stay.

Tree type ground cover raster - Another variable chosen for inclusion is tree type ground cover. As Lantmäteriet (2022) presents, vegetation type often has a high impact on LiDAR point density. Therefore, this raster was also included with the aim to nuance the model's calculations in terms of variation in vegetation.

The ground cover reflects the approximate share of tree types in each cell, based on canopy cover percentage (Table 4). The data is derived from a machine learning model based on a time series learning method from Sentinel-2 satellite pictures (The Swedish Environmental Protection Agency, 2023). There are nine different ground cover rasters produced by the Swedish Environmental Protection Agency, each covering a different tree type. The layers chosen for this project relate to the dominant tree types found in the data selection step (See: 4.3.1), which were pine, spruce and common leaf trees as in birch or aspen. These tree types relate to the ground cover layers with the same definitions, where the common leaf tree layer includes both birch and aspen (Swedish Environmental Protection Agency, 2023)

Table 4. Variables included in the analysis with their name in ArcGIS, definition and description

Name in ArcGIS Pro	Variable Definition	Description
HODDATAGRID5	Elevation	Digital Terrain model in meters
TWIFINAL	TWI	Topographic Wetness Index
RADIANSLOPE	Slope	Slope in Radians
GRANGRID5	Spruce Cover	Share of spruce in each cell
TALLGRID5	Pine Cover	Share of pine in each cell
TRIVIALLOVGRID5	Leaf Tree cover	Share of common leaf trees in each cell
MARKFGRID5KLASSAD	Soil Moisture	Categorical Soil Moisture in each cell
DENSITET_KLASS_3	Vegetation density C3	LiDAR-derived point density raster 0-3m
DENSITET_KLASS_4	Vegetation density C4	LiDAR-derived point density raster 3-5m
DENSITET_KLASS_5	Vegetation density C5	LiDAR-derived point density raster 5-7m
DENSITET_KLASS_6	Vegetation density C6	LiDAR-derived point density raster 7-9m

4.4 GIS Process

Information regarding the used tools used in this project can be found on https://pro.arcgis.com/en/pro-app/latest/tool-reference/main/arcgis-pro-tool-reference.htm. The tools used are referred to with their names in cursive writing.

4.4.1 Data vectors

The first step in GIS was to extract the vectors for each tract found in the data selection process. The data in Holmen's internal database was not transferable to units not registered as company property and, thus, the vectors had to be extracted manually from Holmen's internal database through georeferencing and creating vector features. Georeferencing refers to assigning an object its geographical elements, such as its location (Longley et al., 2015). Assigning location was the most critical georeferencing part of this project, due to transferring the data from Holmen's database, including taking map screenshots of them in the internal program VSOP where tract data is kept. Fortunately, the tracts in VSOP had one set of coordinates included which could be used as the first point of reference to the world imagery. However, to work with the screenshot, scale and projection also had to be fitted in order to integrate it into GIS. Thus, manual georeferencing points were made through fitting features in the screenshotted map to their real location. In this context, where the main characteristics of all imported screenshots was forest, roads or lakes served as the main georeferencing points.

In the example of Uddersjöliden (See: Appendix 1) a road was used for matching the points. Therefore, protruding features in the roads such as sharp turns or crossroads were matched with their position in the screenshot to their actual position in GIS. This required quite high attention to details since much of the data quality rested on the georeferencing being correct, otherwise the data would show areas which were not affected in reality.

Once the imported screenshot had been georeferenced, the feature vectors could be created. This meant manually "redrawing" the lines from the imported screenshot to create features in the new GIS setting. Once the features had been drawn, the imported screenshot could be removed, and the features from the real-world forestry setting had been imported to GIS (See: Appendix 1). This process was done for all 43 tracts in the analysis.

4.4.2 LiDAR-data preparation

Once the boundaries for the areas of interest had been imported in the step above, the data preparation could commence. The first data that was prepared was the LiDAR. The data was collected from Lantmäteriet (n.d.a). LiDAR-data is a very heavy data type containing much information, thus it was not possible to collect data for the entire study area in one download due to it simply taking up too much disk space. Instead, relevant tiles of geographic data were downloaded. The created vectors from the previous step could be uploaded as reference files to Lantmäteriet's page, and therethrough only the relevant map tiles were included in the download.

Once all the data was collected, the data formatting started by creating a LAS-dataset with the *Create LAS Dataset* tool, which combined all the separate tiles downloaded from Lantmäteriet. To extract the LiDAR-data from the vectors created earlier the *Extract LAS* tool was used. After extracting the LAS, the data was split up into separate datasets for each tract with the help of the tract vectors. Once that was done, the vegetation intervals could be set through the *Classify LAS by Height.* This tool creates new classifications based on the inputs of choice, which in this case were 1: Above threshold (excluded data) 2: Ground, 3: Low-vegetation (0-3m), 4: Medium-Low vegetation (3-5m), 5: Medium-high vegetation (5-7m), 6: High Vegetation (7-9m). The classification tool thus creates the final three-dimensional product before turning it into a two-dimensional raster. In Figure 7 below, the different levels of vegetation can namely be seen in a three-dimensional setting. The vegetation levels are set with the ground (brown) points as reference, differences in elevation do therefore not affect the vegetation intervals. Another important thing to point out in reference to the ground points is that even though the low-vegetation points are set to heights 0-3, the ground points are distinguished through the point intensity.

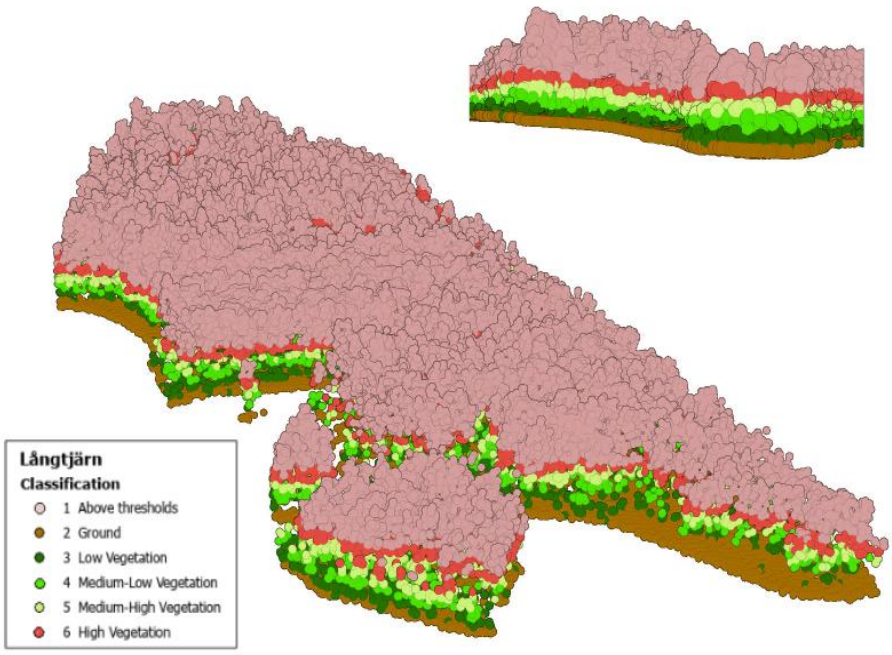


Figure 7. A LiDAR forest model with height interval classification.

The different classification also enables the next step in the process, namely turning this three dimensional data to two dimensional. Each classification group can be processed into its own layer using the *Make LAS dataset layer* tool. Furthermore, each of the new LAS dataset layers can be used to create the vegetation density raster layers using the *LAS Point Statistics as raster* tool (See: Appendix 3). While the study area was relatively small on a national scale, its size impacted the raster resolution in the analysis. The fundamental idea was to have as high a resolution as possible. However, data volumes are highly connected to raster resolutions (Longley et al., 2015), bigger areas and higher resolutions thus equal more intense and time-consuming data processing and analysis. The highest viable resolution was found at 5x5m cells. Therefore, the raster conversion tool calculates how many LiDAR points are in each 5x5m cell. Once the point density rasters for all four intervals, for all tracts, had been extracted the layers were once again unified per interval to a single layer using the *Mosaic to Raster* tool.

4.4.3 Additional stand data

Elevation data - The elevation data for this analysis consisted of a Digital Terrain Model (DTM) in 2x2m resolution produced by Lantmäteriet (2019). The data is created with LiDAR data where all the ground points are processed into a raster in the vertical coordinate system RH2000. The DTM was fetched from SLU–GET.

Due to the relatively high resolution in the data, four separate downloads were required. Thus, the separate rasters had to be combined using the *Mosaic to raster* tool. Once combined, this data had to be fitted into the same resolution as the other data in order for the analysis to run as smoothly as possible. This was done using the resample tool with a bilinear technique,

which means that the new larger cells were calculated based on averaging the value of the surrounding pixels. The bilinear technique was chosen due to its suitability for continuous data.

Soil moisture - The SLU flow accumulation required little preparation due to it being downloaded directly from SLU-GET, SLU's own geodata platform. However, since the data covers a large part of southern Västerbotten and northern Ångermanland in 2x2 meter cells, this data was quite heavy. Thus, it needed four separate downloads. These four different raster files therefore needed to be merged into one separate, which was done using the *Mosaic to New Raster* tool. After merging the rasters, the cell size also needed to be matched to the rasters yielded from the LiDAR data in order to have the same resolution in the Random Forest analysis. Thus, the *Resample* tool was run, with a bilinear interpolation. This technique is based on determining the most probable value of each new pixel based on the values of the surrounding pixels.

After this, pixels without data and water bodies had to be removed. This was done using the following Python script in the *Raster calculator tool,* where the numbers represent the classes presented above.

$$Con((\text{"}mergedresample.tif\text{"} \geq 1) \,\&\, (\text{"}mergedresample.tif\text{"} < = 3), \text{"}mergedresample.tif\text{"})$$

Topographic wetness index (TWI) and slope - The TWI layer was derived directly from the DTM, thus the flow accumulation layer in the TWI is not the same as the one used for the analysis. The TWI was done with the same method as presented by Ermold (2019). The workflow behind building the TWI has two branches stemming from the DTM and ends in combining the TanSlope raster with the flow accumulation. Firstly, the *Slope* tool is applied to the DTM to calculate the steepness in each cell in degrees. However, to produce the TanSlope, the slope is first needed in radian format. Therefore, the *Slope* product is run through *Raster Calculator* with the following function

$$Slope * \pi / 180$$

This function yields the slope in radians, which was the layer used as slope in the analysis. Furthermore, the slope radian could be run in the *Tan* tool, producing the TanSlope raster, thus completing the first branch of the TWI workflow.

For the other branch, the first step is filling the DTM with the *Fill* tool. Filling the DTM means removing minor flaws and thus creating a smoother surface, which makes the layer easier to work with. Then the *Flow direction* tool was utilized. This tool creates a raster which shows which direction the water flows from the cell in focus, based on the elevation. Then comes the *Flow Accumulation* tool which, as explained in data selection, calculates the number of contributing cells to the cell in focus, thus accumulating the flows.

Once the Flow accumulation and TanSlope rasters were finished. Therefore, the last step was calculating the natural logarithm of Flow accumulation divided by TanSlope using the following function in *Raster Calculator*

$$Ln((\text{"\%}FlowAcc1\text{\%"} + 1) / (\text{"\%}Tan\_Raster\text{\%"} + 0.001))$$

Tree type ground cover raster - This data was acquired from The Swedish Environmental Protection Agency. The data is produced in separate sets based on tree type in a 10x10m resolution (The Swedish Environmental Protection Agency, 2023). Therefore, three separate downloads were needed to acquire the pine, spruce and common leaf trees. The processing for

these layers needed only a *Resample* to 5x5 to fit the other data in the analysis. The resampling was done using a majority technique due to its suitability for discrete data. In a resolution-increasing context, the majority technique means dividing a cell into more cells with the same data in practice.

4.4.4 AutoML

Once all the data was prepared for running a Machine learning model a suitability tool was run in GIS, called *Train using AutoML*. The tool tests a set of different ML-models and ranks their results based on prediction errors, where a low error represents a good model fit (ESRI, n.d.e). A major advantage to being able to test model suitability is the consequences of choosing the wrong model for the assignment. As Burkov (2019) argues, having for example a too complex model might risk overfitting (See: 2.3.2). Meanwhile a too simple model might also struggle in finding patterns if the issue is too complex for it. Thus, to decrease the risk for these issues, the suitability model was run on the prepared data (Table 5).

Table 5. Train using AutoML derived model ranking

Model type	Ensemble	Random Forest	LightGBM	CatBoost	Xgboost	Linear	Decision Tree
Prediction error	0.602	0.624	0.664	0.674	0.906	1.188	2.545

The AutoML tool ran Decision tree, Linear, LightGBM, Xgboost, CatBoost, Random Trees, extra trees and Ensemble models. Where a simple decision tree was least fitting, and an ensemble model was most fitting. An ensemble model refers to building a single model through combining several others (James et al., 2021). Unfortunately, ArcGIS does not have a tool which automatically combines several models. Due to the limited time for the project, combining several methods was deemed too labor and time intensive. Fortunately however, ArcGIS pro does have a tool for the close second best - random forests.

4.4.5 Random forest analysis

As deemed most appropriate within the scope of this project, *The Forest-based and Boosted Classification and Regression* tool was chosen in ArcGIS, focusing on the forest-based option within the tool.

The tool offers several options both in terms of model types, but also for which type of data is used and which prediction type is sought after. The model can for example be built on either traditional regression, or on either forest-based or gradient boost machine learning. The data types can be either rasters or vectors in continuous, discrete or categorical values. Prediction can be done either based on vectors or rasters. Thus, it is a multifunctional tool.

As identified in the AutoML run, a forest-based analysis was chosen. This random forest model is based on Leo Breiman's (2001) ditto. The data for the analysis is as explained in the earlier section quite mixed in terms of type, but since all data was in form of rasters, they were input under the Explanatory training rasters section.

The tool creates models and builds predictions using two different machine learning methods – either the random forest algorithm or Extreme gradient boosting algorithm. For this project,

the random forest algorithm was chosen due to its proficiency according to the ArcGIS Pro tool *Train using AutoML*. This random forest model is based on Leo Breiman's (2001). What differs this method from a traditional regression analysis is that this model helps identify which variables are explanatory in practice. The idea is therefore to provide the model with several different types of data in different forms, and therethrough let the algorithm calculate and identify the variables which are most important (ESRI, n.d.c)

5 Results

This chapter will present the parameters and results from each run in order, presenting the included variables, model characteristics, variable importance table and validation data.

5.1 First training run

During the first run, the default number of 100 trees were chosen. As explained in the methods section, this means that 100 bagging processes were undertaken. The *leaf size* represents a type of "isolation" metric, where the model wants to isolate each observation into a leaf of its own. This means that even though observations with the same classification might be over the same threshold, the model still wants to separate the two *tree depth* metrics shows how many times the trees split. A mean tree depth of 739 shows that several hundreds of splits are needed in order to isolate each observation. The two respective percentages of data available show how much data was available in different instances of the training procedure. The latter, *percentage of data excluded for validation*, shows that 20 percent of the data was excluded for validation, or in other words to test the trained model. However, as shown in the first metric data availability metric, all of the remaining data after validation data exclusion was available for the bagging procedure in each tree. The *number of randomly sampled variables* represents the additional diversity emulation as covered by Breiman (2001), where each tree was assigned three random variables which their splitting was to be based upon.

Table 6. First training run model Characteristics

No. of trees	Leaf size	Mean Tree Depth	No. of randomly sampled variables	Percentage of data excluded for validation
100	1	739	3	20

The Top Variable importance table shows an aggregation of how important each variable is in the analysis in its entirety (ESRI, n.d.d). This run showed that the elevation data was the most important variable at 13 percent, while the categorical soil moisture variable had the lowest importance. This result differed from the real-world context where the vegetation densities are of the highest importance in a PHC assessment (Holmen Forest, n.d.a). Variable importance is presented in Table 7.

Table 7. First training run Variable importance

Variable	Importance Percentage
Elevation	13
TWI	12
Vegetation density class 6	12
Slope	11
Spruce cover	10
Pine cover	10
Vegetation Density class 5	8
Vegetation Density class 3	8
Vegetation Density class 4	8
Common leaf tree cover	7
Soil moisture	0

The validation data table shows how the model performed when predicting the unseen data excluded for validation as shown in the model characteristics table (Table 8). As presented in section 2.3, MCC is a strong metric for contexts such as in this project, therefore it is considered the most important metric. A 0.58 MCC shows that the validation predictions are somewhat closer to perfect (MCC=1) than equal to randomness (MCC=0).

Table 8. First training run classification diagnostics

Category	F1-score	MCC	Sensitivity	Accuracy
NO PHC	0.74	**0.58**	0.77	0.80
PHC	0.84	**0.58**	0.82	0.80
Combined	0.79	**0.58**	NA	0.80

5.2 Second training run

Seeing as the importance for soil moisture had no importance, it was removed for the second run to improve the model performance (ESRI, n.d.d). The model characteristics for the second run did not differ much from the first, except the removal of the categorical soil moisture variable. However, as shown in the model characteristics table, the trees ran deeper during this run (Table 9).

Table 9. Second training run Model Characteristics

No. of trees	Leaf size	Mean Tree Depth	No. of randomly sampled variables	Percentage of data excluded for validation
100	1	923	3	20

The variable importance in the second run showed little change from the first run. The biggest relative difference was in the slope variable, where its importance grew one percent, and surpassed the 7-9 meter vegetation density slightly. Thus, this run showed a result differing even more from the vegetation focus in the PHC real-world assessments (Table 10).

Table 10. Second training run variable importance

Variable	Importance Percentage
Elevation	13
TWI	12
Slope	12
Vegetation Density class 6	12
Pine cover	10
Spruce cover	10
Vegetation Density class 5	8
Vegetation Density class 3	8
Vegetation Density class 4	8
Common leaf tree cover	7

In the second run all validation metrics decreased slightly, with the Matthews correlation coefficient (MCC) decreasing the most (Table 11)

Table 11. Second training run Classification Diagnostics

Category	F1-score	MCC	Sensitivity	Accuracy
NO PHC	0.70	**0.51**	0.74	0.77
PHC	0.81	**0.51**	0.88	0.77
Combined	0.76	**0.51**	NA	0.77

5.3 Optimization run

The first two runs show signs that the model was somewhat overfit. Thus, in order to tune the parameters for the model, an optimization run was done. During an optimization run, the model tests how several different parameters fit. In ArcGIS, the optimization can be set to focus on improving a certain metric. Thus, due to its importance in binary classification (Chicco & Jurman, 2020), the MCC parameter was chosen as the main focus. Further, the model was instructed to test the performance with a number of trees between 100 and 300, with an interval of 50. Same was done for the number of leaves between 1 and 5, with the interval of 1. The tree depth was also tested in the ranges of 100 to 1000, with an interval of 50.

This run yielded some other characteristics (Table 12). The number of trees were increased to 300, leaf size increased to two, and depth range also increased to a mean 955. The percentage of data excluded for validation was also decreased to 10 percent for this run as part of the iterative process of optimizing a ML-model. As Burkov (2019) argues, there is no norm of how much data is to be excluded for validation, thus different levels might yield different outcomes.

Table 12. Optimization run model characteristics

No. of trees	Leaf size	Mean Tree Depth	No. of randomly sampled variables	Percentage of data excluded for validation
300	2	955	3	10

The shares of variable importance remained unchanged in this run compared to the second training run (Table 13).

Table 13. Optimization variable importance

Variable	Importance Percentage
Elevation	13
TWI	12
Slope	12
Vegetation density class 6	12
Pine cover	10
Spruce cover	10
Vegetation density class 5	8
Vegetation density class 3	8
Vegetation density class 4	8
Common leaf tree cover	7

The optimization run yielded a slight improvement in the metric where improvement was sought after, with an MCC increase of 0.02. All other metrics also improved slightly, except the NO PHC sensitivity which decreased by 0.01 (Table 14).

Table 14. Optimization run classification diagnostics

Category	F1-score	MCC	Sensitivity	Accuracy
NO PHC	0.71	**0.53**	0.73	0.78
PHC	0.82	**0.53**	0.80	0.78
Combined	0.76	**0.53**	NA	0.78

5.4 Third training run

While the optimization run yielded an improvement in MCC, the first run was still the best in those terms. Even though the soil moisture variable had an importance of zero percent, it still shows a meaningful contribution in terms of complexity to the trees (Table 15). Thus, the following run includes the soil moisture variable, combined with the improved parameters from the optimization run. All values from the optimization run were kept for this one, except the leaf size which was again preferred as one. However, the mean tree depth still shrunk to 793.

Table 15. Third training run model characteristics

No. of trees	Leaf size	Mean Tree Depth	No. of randomly sampled variables	Percentage of data excluded for validation
300	1	793	3	5

The variable importances remained relatively unchanged in the third training run, apart from the Vegetation density class 7-9 which overtook the slope variable (Table 16).

Table 16. Third training run variable importance

Variable	Importance Percentage

Elevation	13
TWI	12
Vegetation density class 6	12
Slope	12
Spruce cover	11
Pine cover	10
Vegetation Density class 5	8
Vegetation Density class 3	8
Vegetation Density class 4	8
Common leaf tree cover	7
Soil moisture	0

Taking insights from the previous training runs and the optimization run the MCC increased to 0.59, thus helping reach the highest metric score throughout the analysis (Table 17).

Table 17. Third training run classification diagnostics

Category	F1-score	MCC	Sensitivity	Accuracy
NO PHC	0.74	**0.59**	0.77	0.81
PHC	0.85	**0.59**	0.83	0.81
Combined	0.79	**0.59**	NA	0.81

5.5 Prediction run

Combined with the internal validation, external validation was also undertaken. Model characteristics were similar to earlier runs (Table 18). The main difference in this case is that the model was run without any training data excluded for internal validation. The mean tree depth was marginally increased from the previous run, showing that the model is still quite complex and deep.

Table 18. Prediction run model characteristics

No. of trees	Leaf size	Mean Tree Depth	No. of randomly sampled variables	Percentage of data excluded for validation
300	1	824	3	0

The variable importance shares remained unchanged from the previous run. Elevation was still considered as the most important factor, and soil moisture as least important. Throughout all runs the variable importances remained relatively the same except a few changes of place in the midtable (Table 19).

Table 19. Prediction run variable importance

Variable	Importance Percentage
Elevation	13
TWI	12
Vegetation density class 6	12
Slope	12
Spruce cover	11
Pine cover	10
Vegetation Density class 5	8
Vegetation Density class 3	8
Vegetation Density class 4	8
Common leaf tree cover	7
Soil moisture	0

Since no validation data was excluded from the training data, this run did not yield classification diagnostics. The validation step was instead done on unseen training data, i.e test data. Since the test data does not have a predetermined label the model could not reassure its prediction accuracy. The test results were instead validated manually. The test data consisted of two tracts of a total 8.4ha deemed in need of PHC, and two tracts of a total of 9.3ha without PHC need. Figure 8 represents a raster prediction.

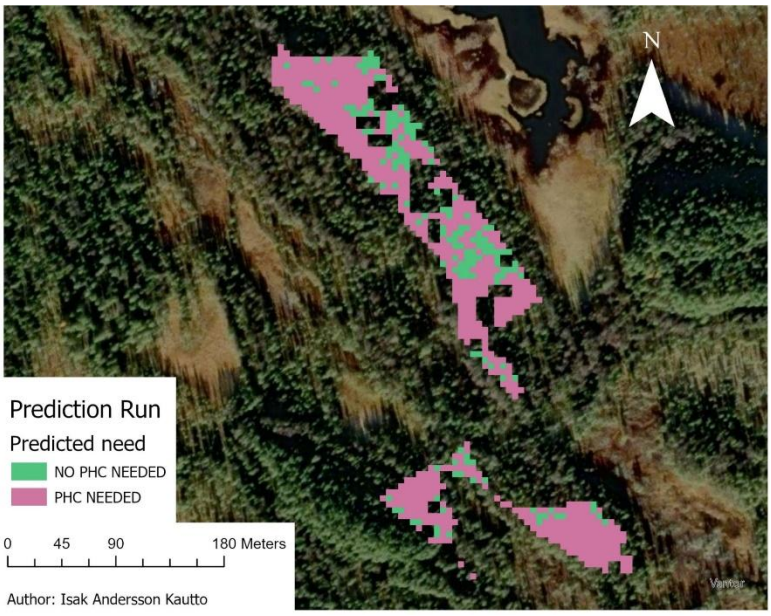


Figure 8. A prediction raster from tract "400786878 Plågan". The green and pink represent 5x5 cells which have been predicted as either in need of PHC or not.

In the figure above, it is arguably visible by eye that the PHC need is in majority. However, to get an exact majority calculation, the tract vectors were connected to the rasters using the *Zonal Statistics as Table,* which helped yield the tract majority. The final majority result acted as the final prediction verdict. The results for majority calculations is presented in Figure 9.

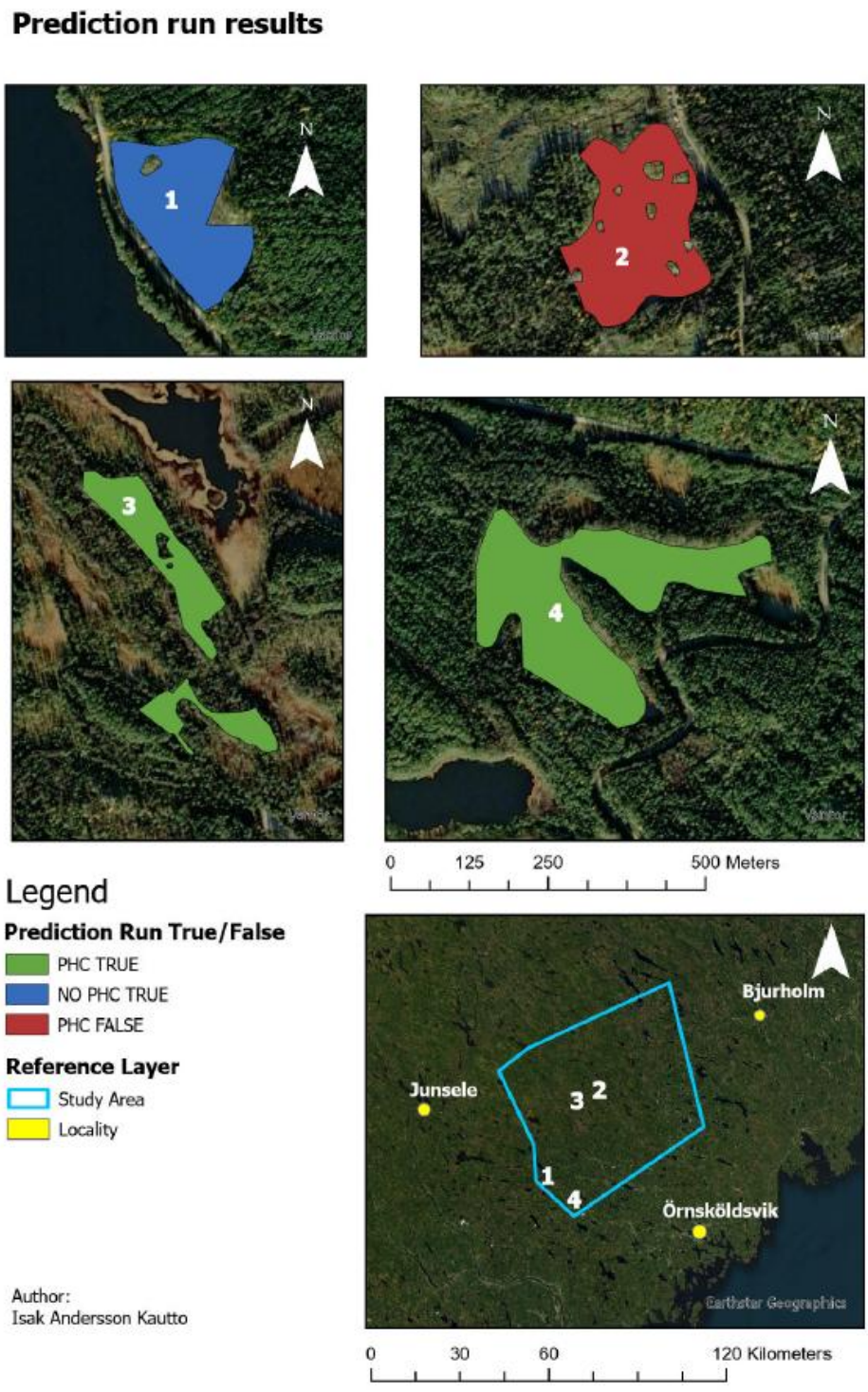


Figure 9. The final prediction results after calculating the majority on the four test tracts. Results are presented based on a confusion matrix layout, where the one faulty prediction is represented as "PHC FALSE". The bottom right shows the tracts' geographic locations

6 Discussion

This section presents the discussion points of the project. It starts with relating model performance to earlier research and potential key points to improve the model, followed by an overview of data quality. Lastly, the project's contribution and relation to precision forestry is discussed.

6.1 Model accuracy

One of the main ambitions of this project was to test non-linear machine learning methods as an alternative to earlier linear approaches to the same problem. Sanz et al. (2020) has acted as a major source of inspiration and comparison due to their linear approach. Seeing that they reached an overall accuracy of 64%, and this model has an MCC of 0.59 and accuracy of 0.81, it might give the impression that this method performs better. However, it must be pointed out that the research contexts of this project and Sanz et al. (2020) are quite different even though the topic is the same.

This project takes its starting point in a binary classification due to that being the state of the data, where each tract either needs PHC, or not. Meanwhile, Sanz et al. (2020) attempts predicting PHC needs on a scale from 1 to 5, which arguably makes the analysis much more complex in certain aspects. In addition, Sanz et al. (2020) had a quite different way of working with the LiDAR point clouds where they looked closer on the distribution of points in their segments, rather than raw amount points as was done in this project. Due to these reasons, determining whether linear methods or non-linear machine learning perform better on predicting PHC needs still remains unclear. A rigorous comparison would require both methods being applied to identical datasets and in the same conditions. While the *AutoML* run (See: Table 3) simulated linear runs on this dataset, and suggests that linear methods were a poor fit, this dataset is arguably still too small to present a generalizable result.

6.2 Possible data error sources

The model showed that elevation, TWI, slope and 7-9m vegetation point density were the most important variables. As Albrektson et al. (2012) present, the first three are important factors for vegetation growth. However, considering that vegetation density is pointed out as the most important factor in practical contexts (Holmen Forest, n.d.a; Sanz et al., 2020), it was an unexpected result to see class 6 vegetation density at 12 percent importance, and the remaining three at eight percent and thus considerably lower than many other variables..

Considering that the model yielded quite positive metrics, there is a meaningful connection between elevation, TWI, slope, and PHC needs in this particular dataset. However, it might be caused by a simple geographical bias, where the PHC need happens to be common on a certain elevation, slope and TWI interval. Thus, additional data could prove beneficial to yield a more realistic result in terms of each variable's actual importance.

Another possible reason for the unexpectedly low vegetation importance is the data quality. When processing the LiDAR-data, manual sight inspections of the three dimensional forest models were done. During an inspection, the point density in tract “400786975 Brännanvägen” caught my attention - it had a visibly higher point density than many of the other tracts in the analysis. Upon closer inspection, there were roughly 56700 ground points,

and 67700 points classified as vegetation (0–9m) in a 1,97ha tract, correlating to roughly 29000 ground points and 34400 vegetation points per hectare.

Another tract which also caught my attention for the opposite reason was “400785321 Backevägen”. In this tract the number of points per hectare was only roughly 6000 ground points and 16100 vegetation (0-9m) points. The point density therefore varies greatly between some tracts, and in this case it differs almost 500 percent in ground points, and 420 percent for the vegetation. While both these tracts are rare extremities in their respects, it still has a major impact on the point cloud variable validity, since areas with higher point density are naturally going to yield more points per cell when represented in rasters.

The definite reasons for these extreme cases in particular remain unclear. Lantmäteriet (2022, p. 6) acknowledges that there might be ground point density variation between certain areas caused by several different reasons. One error source pointed out is vegetation thickness. In areas where the vegetation is very thick, ground points risk “disappearing” due to the laser not getting through the vegetation. However, considering that Brännanvägen has almost twice the amount of vegetation points per hectare than Backevägen, it is not likely that the vegetation is the perpetrator.

While the point density is advertised as one or two points per square meter, Lantmäteriet (2022, p. 4) deems that approximately five percent of the data does not exceed one point per square meter. The likely situation with “Backevägen” is therefore that it is included in those five percent, since one point per square meter would equal 10000 points per hectare.

The issue with varying point densities could possibly be partially resolved with the ULCD metric proposed by Wing et al. (2012), where the vegetation density is based on relating the ground and vegetation points (See: 3.1 Literature review). Using a vegetation density variable in the same manner as the ULCD, the raster processing would be less sensitive to fluctuations in point cloud density. This would however not resolve the issue with ground points “disappearing” in thick vegetation, which in itself could potentially inflate a ULCD-based variable. Although, the inflated ULCD-value would arguably still represent the extreme vegetation density if the amount of vegetation points per ground point was due to “disappearing” ground points.

In cases such as “Backevägen”, another issue is raised. Maltamo et al. (2006) argue that area-based approaches need at least one pulse per square meter to yield a meaningful result. In that case, a relative metric would still not help improving the quality to a useful level, and the issue might simply only be resolved during the data collection.

In summary, including more data and a relative point density metric could potentially help improve and generalize the model further. The issue remains however with the data points where the point density does not exceed one point per square meter, and future research should therefore be aware of said risk.

6.3 Model overfitting

Overfitting is a common issue in Random Forest models (Burkov, 2019). Overfitting refers to the model being overfit to the training data, where it in other words performs well on the seen data, but whenever it is met with variation beyond the diversity in the training data, it does not yield as good results. In statistics terms, this is also called high variance (Ibid.)

Looking at the results from the model developed in this project, while it performed reasonably well with an MCC of 0.59, there are still some factors pointing to there being a somewhat overfit. According to Burkov (2019) two of the most common reasons for this is either that there are too few training examples compared to features, or the model is overtly complex in relation to the data, which in a random forest context could mean that the decision trees are too deep.

Both of these common factors thus seem quite probable in this case due to the limited amount of data and tree depth reaching towards one thousand. Burkov (2019) further proposes a select set of solutions to this problem. First of which, is trying a simpler model. In this case, this would mean changing major parts of the methods and trying for example a linear model. This would go against the idea of this project though, since there have been attempts on PHC-prediction using linear models already. Although, within ArcGIS Pro there is another tool called AutoML (ESRI, n.d.e)

Secondly, Burkov (2019) recommends reducing dimensionality. Dimensionality can roughly be explained as simplifying the model through for example removing variables with low importance. This is often done on features which are highly correlated to each other. Thus, this model would seemingly be prone to dimensionality issues due to the TWI and Soil Moisture features, as well as the three tree ground cover rasters, being closely connected. Reducing dimensionality is also encouraged by ESRI (n.d.d) based on the variable importance table from each run. However, as presented in the results section, an attempt on dimensionality reduction was done in the second run through removing the Soil moisture feature based on its low importance. This proved no improvement due to this change decreasing the MCC from 0.58 to 0.51. For the ground cover rasters, dimensionality reduction was never attempted due to them having a certain level of importance in each run.

Thirdly, Burkov (2019) proposes simply adding more training data. Due to the projects' limited time this solution could unfortunately not be tested.

Fourth and finally, Burkov (2019) points out regularization as a key to reducing overfitting. Regularization means forcing the algorithm to build simpler models. In this project, attempts were made to regularize the random forest model through the optimization run. However, as presented in the results section, testing lower tree depths and fewer trees did not show signs of improving the model, but rather the opposite. Thus, regularization within the random forest method does arguably not help the overfitting issue since the classification process still requires much complexity (Breiman, 2001).

Over the course of different runs, different shares of data saved for validation were tried. Burkov (2019) argues that there is no optimal proportion for training to validation dataset sizes, thus the validation dataset size was decreased as a trial process. In the past, a 70 to 30 percent division was the norm. Today however, the training to validation shares of data span between 30 percent to five, depending on how large the dataset is. Thus, the share of validation data was gradually decreased in order to provide the model with more observations since it tends to improve the model performance (James et al., 2021).

6.4 Precision forestry contribution

While this model is argued to not be quite ready for a practical context, it arguably still contributes to further understanding the challenges and dynamics behind closing in on the

automation gap in *forest management* with the help of *remote sensing and GIS,* as defined by Vatandaslar et al.'s (2025) framework for precision forestry. It managed to reach three correct predictions out of four, showing the *automation* potential which LiDAR and other geodata might bring. It focused on the PHC aspect of *forest management and planning* and it used data collected with airborne platforms in GIS, fulfilling the *remote sensing* and *GIS* tools (Ibid.). The proposed framework therefore offers a sound base to work from in terms of building methods. However, relating this to the data quality challenges in this project, an important aspect for reaching precision forestry potential beyond simply data and software accessibility, is the quality. In this project, the LiDAR quality was identified as a probable source of error (Lantmäteriet, 2022), and thus one might propose including quality aspects as an important factor for precision forestry.

Vatandaslar et al. (2025) identified *technological advancements* and *workforce transitions* as important drivers for precision forestry. This project has made use of several different aspects within GIS, laser-based data and machine learning. Thus, many modern technological tools have been applied to reach the outcome, correlating to how important the *technological advancements driver* identified by Vatandaslar et al. (2025) is.

As argued in section 2.1, the *workforce transitions* among Vatandaslar et al. (2025) drivers can be understood as quite critical towards today's forestry professionals. They argue that today's approach to forestry is too traditional and thus slows down innovation within the frames of precision forestry (Ibid.). While that might be true in certain cases, another factor which needs consideration is the time aspect of applying new methods. As in this case, creating a ML model for PHC has above all required putting much time into learning technical concepts, such as how ML functions, rather than utilizing forestry competence. Although much time was put into learning ML and developing the model, it is still not ready for practical implementation. Therefore, the barrier to adopting innovative methods for contemporary practices is arguably more complex than the workforce attitude itself. This aligns with the argument (See: 2.1) that the workforce transition aspect should rather be focused on making user-friendly and accessible tools that help propel existing forestry knowledge, rather than shifting the workforce’s competence to a more technical profile.

In summary, and in similar contexts to this project, the results propose including data quality and method accessibility aspect if a predictive ML model for PHC is to become reality.

7 Conclusion

This final chapter summarizes the project's results and relates them to their respective research question. Finally, suggestions for further research on the subject are presented.

This thesis contributes to further understanding of digitalizing forestry planning practices within the frames of precision forestry. In relation to the research aim, GIS Machine learning and LiDAR-data shows great promise as remote tools for PHC assessments.

While this study could not produce a practically ready model, building on the precision forestry framework (Vatandaslar et al., 2025) contributes to creating digital forestry practices, which might help automate certain processes. Another major conclusion from this study is the technical aspect of creating predictive models. Vatandaslar et al. (2025) argue that workforce attitude is a hinder for the progress of precision forestry. This study recognized that creating predictive models puts much more emphasis on technical competence rather than forestry competence. That resonates with Vatandaslar et al. (2025) insight that digital means in forestry might overcomplicate things technically, which are already demanding high forestry competence.

7.1 Contributions

The model reached internal validation scores of MCC 0.59, F1-score 0.79, sensitivity 0.77 (NO PHC) & 0.83 (PHC), and Accuracy 0.81. While these metrics are difficult to translate into an accuracy percentage, the internal predictions still showed that the model is considerably closer to perfect predictions rather than plain chance.

The most important variables identified were elevation, TWI, slope and vegetation point density in the 7-9m interval. This result went against the presumption that the most important variables would be found in the different vegetation density height intervals. However, since the included data in this project is quite limited, the results cannot be further generalized than this dataset itself.

When tested in a more practical setting, in external validation, the model predicted three out of four test tracts correctly. This result is neither generalizable beyond the test tracts, and rather needs extended testing on more tracts in order to reach an idea of how it would perform were it put into practical use.

7.2 Suggestions for future research

For future research on this subject, including more data would be encouraged. Including more data could help mitigate the model overfit, and possibly also yield a more realistic result due to the increased variety in the data. While the model metric scores showed promising predictions, the variable importance still showed unexpected results where the point density intervals had relatively low importance in contrast to its importance in today's manual PHC analysis. This might be due to the data quality pointed out in the discussions section, and thus a relative point density variable, such as the ULCD metric by Wing et al. (2012) is recommended.

Another challenge in this context is the temporal aspect. This project has only analyzed data which was quite on time with the PHC decisions. In a more practical setting, having perfectly

up-to-date data is arguably not possible since Lantmäteriet (n.d.b) only collects LiDAR-data every few years, for example. Therefore, a temporal aspect must also be included before a practical application is realistic.

Returning to Nordqvist's (2023) question on whether AI will make forestry faster, better and more efficient – This study suggests we are on the right track, but there is still much to be done in terms of data quality, model tuning and methodological accessibility before fully digitalized processes become reality. Whether or not AI will become the innovation equivalent of mechanization, and random forest models the equivalent of chainsaws, remains unclear. But the potential is undoubtedly there.

References

Breiman, L. (2001) 'Random forests', *Machine Learning*, 45(1), pp. 5–32. doi:10.1023/A:1010933404324.

Burkov, A. (2019) *The hundred-page machine learning book*. Quebec City: Andriy Burkov.

Chicco, D. and Jurman, G. (2020) 'The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation', *BMC Genomics*, 21, p. 6. doi:10.1186/s12864-019-6413-7.

Dash, J., Pont, D., Brownlie, R., Dunningham, A., Watt, M. and Pearse, G. (2016) 'Remote sensing for precision forestry'. *New Zealand Journal of Forestry*, 60, pp. 15–24.

Forsberg, M. and Lodén, H., 2020. *Förröjning i Sverige: en granskning av skogsföretagens strategier för underväxtröjning inför förstagallring*. Bachelor's thesis. Umeå: Sveriges lantbruksuniversitet, Institutionen för skogens biomaterial och teknologi (SBT). Available at: https://stud.epsilon.slu.se/15935/ (Accessed: 25 February 2026)

Haavikko, H., Kärhä, K., Hourula, M. and Palander, T. (2019) 'Attitudes of small and medium-sized enterprises towards energy efficiency in wood procurement: a case study of Stora Enso in Finland', *Croatian Journal of Forest Engineering*, 40(1), pp. 107–123.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) *An introduction to statistical learning: with applications in R*. New York: Springer.

Kanazawa, M. (2023). Research Methods for Environmental Studies: A Social Science Approach (2nd ed.). Routledge. https://doi.org/10.4324/9781003261117

Kärhä, K. (2006) 'Effect of undergrowth on the harvesting of first-thinning wood', *Forest Studies*, 45, pp. 101–117.

Lillesand, T., Kiefer, R. W., & Chipman, J. (2015). *Remote sensing and image interpretation*. John Wiley & Sons.

Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (2015) *Geographic Information Science and Systems*. 4th edn. Chichester: John Wiley & Sons.

Maltamo, M., Eerikäinen, K., Packalén, P., & Hyyppä, J. (2006). Estimation of stem volume using laser scanning-based canopy height metrics. *Forestry*, *79*(2), 217-229.

Merry, K., Bettinger, P., Crosby, M. and Boston, K. (2023) 'Remote sensing', in Merry, K., Bettinger, P., Crosby, M. and Boston, K. (eds.) *Geographic Information System Skills for Foresters and Natural Resource Managers*. Elsevier, pp. 269–301. doi:10.1016/B978-0-323-90519-0.00001-7.

Sanz, B., Malinen, J., Heiskanen, J., & Tokola, T. (2020). 'Need for Pre-Harvest Clearing of Understory Vegetation Determined by Airborne Laser Scanning'. *Forests*, 11(3), 294. https://doi.org/10.3390/f11030294

Taylor, S.E., McDonald, T.P., Fulton, J.P., Shaw, J.N., Corley, F.W. and Brodbeck, C.J. (2006) 'Precision forestry in the southeast US', *Proceedings of the 1st International Precision Forestry Symposium*, pp. 397–414.

Taylor, S. E., Veal, M. W., Grift, T. E., McDonald, T. P., & Corley, F. W. (2002). 'Precision forestry: operational tactics for today and tomorrow'. In: *25th annual Meeting of the council of Forest Engineers* (Vol. 6).

Vatandaslar, C., Boston, K., Ucar, Z., Narine, L. L., Madden, M., & Akay, A. E. (2025). 'Precision Forestry Revisited'. *Remote Sensing*, 17(20), 3465. https://doi.org/10.3390/rs17203465

Weiss, G. (2019). Innovation in Forestry: New Values and Challenges for a Traditional Sector. In: Encyclopedia of Creativity, Invention, Innovation and Entrepreneurship. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-6616-1_441-2

Weiss, G., Hansen, E., Ludvig, A., Nybakk, E. and Toppinen, A. (2021) 'Innovation governance in the forest sector: Reviewing concepts, trends and gaps', *Forest Policy and Economics*, 130, 102506. doi:10.1016/j.forpol.2021.102506.

Wing, B.M., Ritchie, M.W., Boston, K., Cohen, W.B., Gitelman, A. and Olsen, M.J., (2012). 'Prediction of understory vegetation cover with airborne lidar in an interior ponderosa pine forest'. *Remote Sensing of Environment*, 124, pp.730–741. https://doi.org/10.1016/j.rse.2012.06.024

Wohlin, Claes. (2014). 'Guidelines for snowballing in systematic literature studies and a replication in software engineering'. ACM International Conference Proceeding Series. 10.1145/2601248.2601268.

Internet sources

Albrektson, A., Elfving, B., Lundqvist, L. & Valinger, E. (2012). *Skogsskötselns grunder och samband*. Skogsskötselserien, kapitel 1. Available at: https://www.skogsstyrelsen.se/globalassets/mer-om-skog/skogsskotselserien/skogsskotsel-serien-1-skogsskotselns-grunder-och-samband.pdf (Accessed: 2026-04-10).

Ermold, M. (2019). *Potentiella våtmarkslägen – en avvägning av olika GIS-metoder.* Available at: https://www.skogsstyrelsen.se/globalassets/projektwebbplatser/grip-on-life-ip/rapporter-grip-on-life/2019.01-potentiella-vatmarkslagen-en-avvagning-av-olika-gis-metoder.pdf (Accessed: 2026-01-07).

Esri (n.d.a) LAS datasets. Available at: https://pro.arcgis.com/en/pro-app/latest/help/data/las-dataset/what-is-a-las-dataset-.htm (Accessed: 9 February 2026)

Esri (n.d.b) *Deep learning in ArcGIS Pro*. Available at: https://pro.arcgis.com/en/pro-app/3.4/help/analysis/deep-learning/deep-learning-in-arcgis-pro.htm (Accessed: 19 February 2026).

Esri (n.d.c) *Forest-based and boosted classification and regression (Spatial Statistics)*. Available at: https://pro.arcgis.com/en/pro-app/3.4/tool-reference/spatial-statistics/forestbasedclassificationregression.htm (Accessed: 8 April 2026).

Esri (n.d.d) *Forest-Based Classification and Regression using ArcGIS Pro*. Available at: https://www.esri.com/training/ (Accessed: 9 april 2026).

Esri (n.d.e) *How AutoML works*. ArcGIS Pro Documentation. Available at: https://doc.esri.com/en/arcgis-pro/latest/tool-reference/geoai/how-automl-works.html (Accessed: 11 april 2026)

Esri (n.d.f) *Use LiDAR in ArcGIS Pro*. Available at: https://doc.esri.com/en/arcgis-pro/latest/help/data/las-dataset/use-lidar-in-arcgis-pro.html (Accessed: 19 May 2026).

Holmen (n.d.a) This is Holmen Skog. Available at: https://www.holmen.com/sv/skog/om-oss/det-har-ar-holmen-skog/ (Accessed: 28 January 2026).

Holmen (n.d.b) *Vår organisation*. Available at: https://www.holmen.com/sv/skog/om-oss/var-organisation/ (Accessed: 28 January 2026).
Holmen (n.d.c) *Affärsidé, strategi och mål*. Available at: https://www.holmen.com/sv/om-holmen/Holmen-i-korthet/affarside-strategi-mal/ (Accessed: 19 February 2026)
Holmen (n.d.d) *Hållbart och ansvarsfullt skogsbruk*. Available at: https://www.holmen.com/sv/skog/om-oss/vart-skogsbruk/hallbart-och-ansvarsfullt-skogsbruk/ (Accessed: 19 February 2026)
Holmen (n.d.e) *Holmen satsar mer på GROT.* Available at: https://www.holmen.com/sv/skog/nyheter/holmen-satsar-pa-mer-grot/ (Accessed: 4 March 2026)
Kindström, K. (2024) *Så får vi fler att jobba i skogen*. *Land Lantbruk och Skogsbruk.* Available at: https://www.landlantbruk.se/sa-far-vi-fler-att-jobba-i-skogen (Accessed: 3 March 2026).
Lantmäteriet (2019a) *Produktbeskrivning: GSD-Höjddata, grid 2+* (Version 2.6). Available at: https://www.lantmateriet.se/ (Accessed: 13 April 2026).
Lantmäteriet (2022) *Kvalitetsbeskrivning laserdata*. Available at: https://www.lantmateriet.se/ (Accessed: 19 April 2026).
Lantmäteriet (n.d.a) *Laserdata – nedladdning, skog*. Available at: https://www.lantmateriet.se/sv/geodata/vara-produkter/produktlista/laserdata-nedladdning-skog/ (Accessed: 29 January 2026).
Lantmäteriet (n.d.b) *Skoglig laserskanning, SLS, produktionsstatus*. Web GIS portal. Available at: https://webgisportal2.lantmateriet.se/portal/apps/experiencebuilder/experience/?id=25679ae1c6604a04949fa49460329f3d (Accessed: 11 March 2026).
Lantmäteriet (n.d.c) *Om Lantmäteriet*. Available at: https://www.lantmateriet.se/sv/om-lantmateriet/ (Accessed: 19 May 2026).
The Swedish Environmental Protection Agency (2023) *NMD2023 Produktbeskrivning tilläggsskikt – Trädslag*. Available at: https://geodata.naturvardsverket.se/nedladdning/marktacke/NMD2023/Tillaggsskikt/NMD2023_Produktbeskrivning_till%C3%A4ggsskikt_Tr%C3%A4dslag.pdf (Accessed: 13 April 2026).
Nordqvist, C (2023) *Skogsvärden*, In: Skogssällskapet no. 4. Available at: https://www.skogssallskapet.se/download/18.18aec07e18c629dc702a7e/1702467335741/Skogsv%C3%A4rden-2023-nr4-web.pdf (Accessed: 9 March 2026).
Swedish Forest Agency (2025) *Viltskador*. Available at: https://www.skogsstyrelsen.se/bruka-skog/skogsskador/viltskador/ (Accessed: 13 March 2026)
Swedish Forest Agency (2012) *Skogsskötselserien 6: Röjning*. Available at: https://www.skogsstyrelsen.se/globalassets/mer-om-skog/skogsskotselserien/skogsskotsel-serien-6-rojning.pdf (Accessed: 1 February 2026).
Swedish Forest Agency (2026) *Digitala AI-kartor underlättar för skogsägare att ta hänsyn*, 18 February. Available at: https://www.skogsstyrelsen.se/nyhetslista/digitala-ai-kartor-underlattar-for-skogsagare-att-ta-hansyn/ (Accessed: 3 March 2026).

Von Essen, M. (2023) *Ny teknik kan ge bättre betalt för virket*. In: Skogssällskapet no. 4. Available at: https://www.skogssallskapet.se/kunskapsbank/artiklar/2023-12-13-ny-teknik-kan-ge-battre-betalt-for-virket.html (Accessed: 9 March 2026).

Non-published material

Holmen Forest (n.d.a) Förrensningsrutin: HS-0084. Unpublished document, Holmen Skog

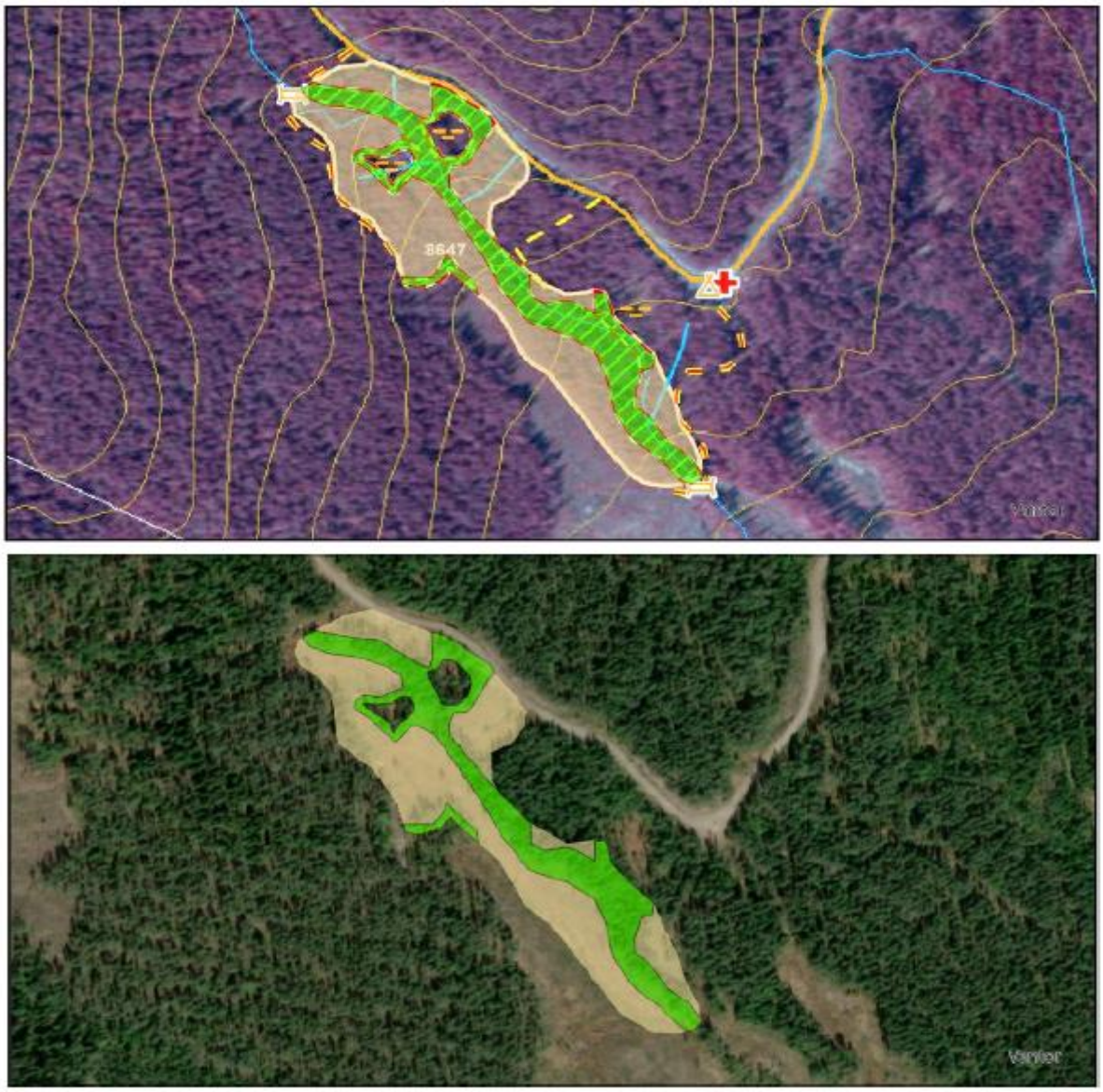
Holmen Forest (n.d.b) Fältinstruktion förrensning och hyggesrensning: HS-0123, Holmen Skog

Swedish University of Agricultural Sciences (SLU) (2020). Produktbeskrivning: SLU Markfuktighetskarta version 1.0 – markfuktighetsskattningar från laserdata

Appendix 1 Georeferencing and feature extraction

Georeferencing and manual feature extraction

Example of 400784925 Uddersjöliden



0 87.5 175 350 Meters

Legend

Protected areas

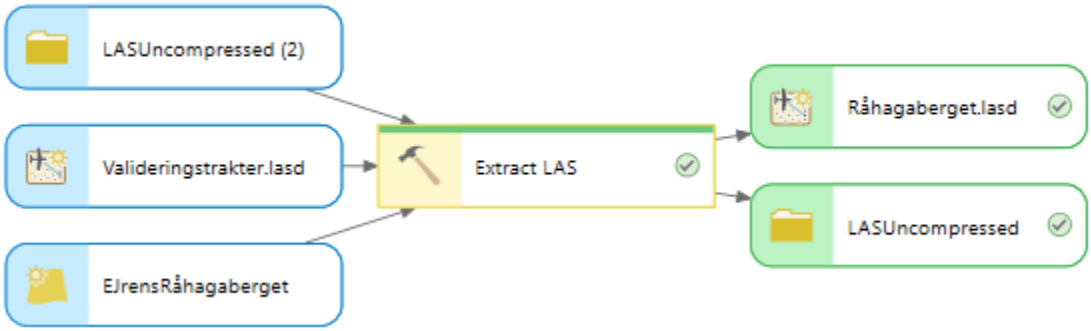
Identified PHC need

Source:
Holmen Forest

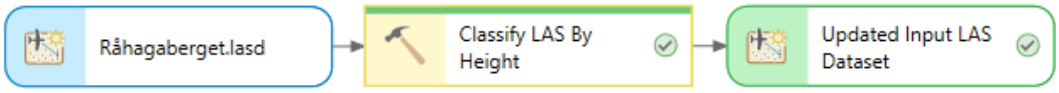
Author:
Isak Andersson Kautto

Appendix 2 LAS workflow

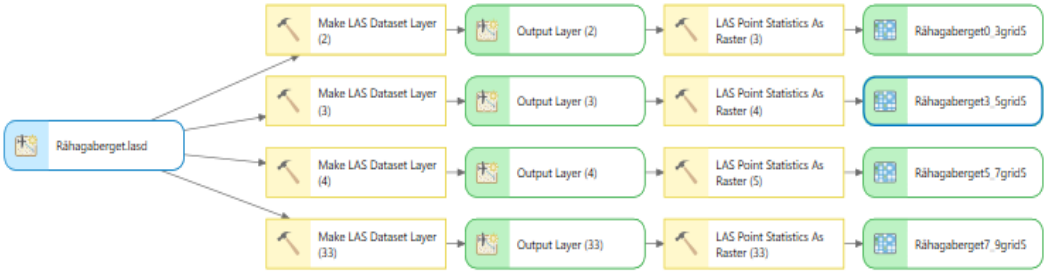
Extract LAS



Reclassification



Raster conversion (From 3D to 2D)



Appendix 3 LAS Point statistics as raster

LAS Point statistics to Raster

Point Cloud density in LAS and Raster - Height interval 3-5 meters

Example of 400787430 Tellmo City

0 20 40 80 Meters

LAS Point Cloud

Classification

- Vegetation 3-5 meters

Point Density Raster

Point Count Per 2x2 Cell

9

1

Source:
Holmen Forest
Lantmäteriet

Author:
Isak Andersson Kautto

Appendix 4 TWI Workflow