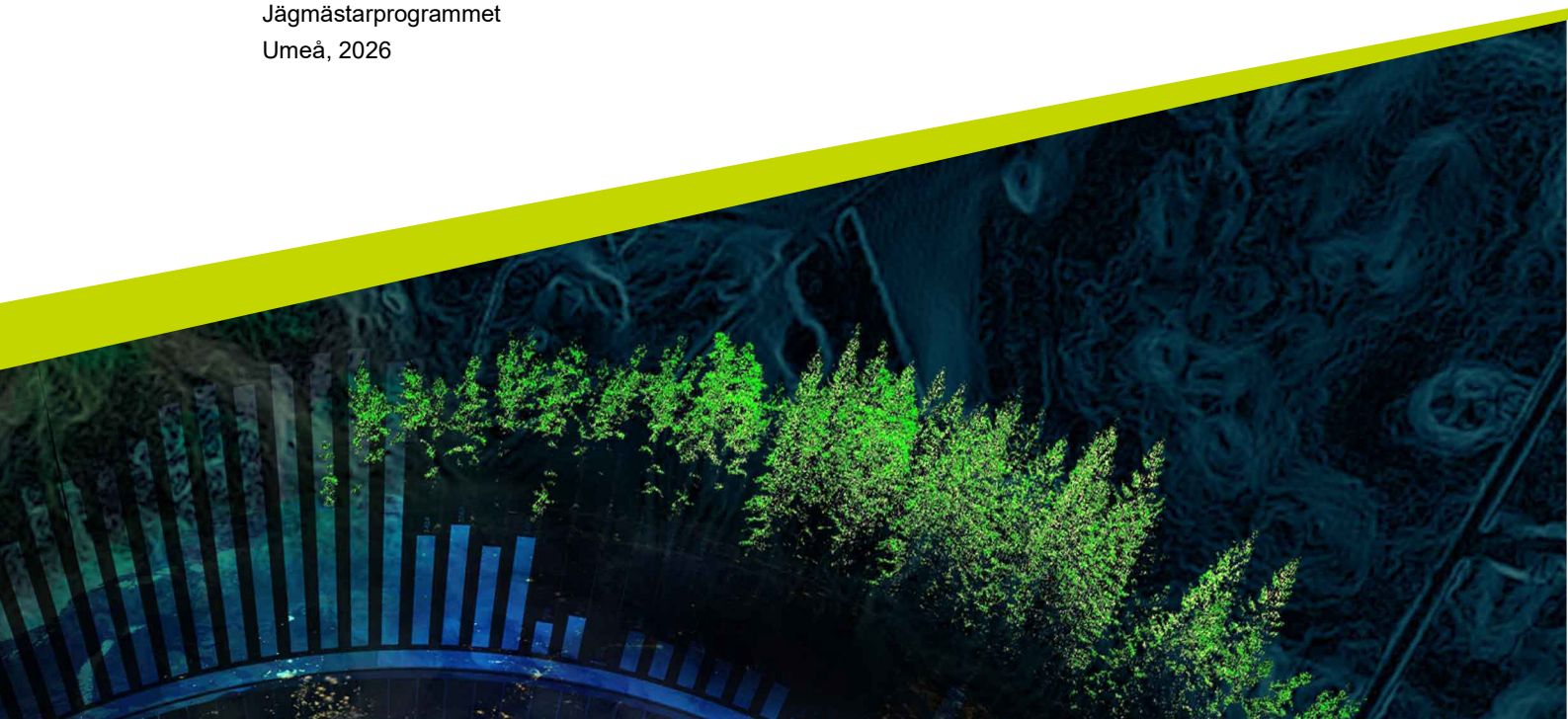
Cluster analysis of historical harvesting sites in Sweden

Spatiotemporal, density-based and non-density-based cluster analysis of forestry-related variables in historical harvesting data

Elliot Eriksson

Degree project • 30 credits
Swedish University of Agricultural Sciences, SLU
Department of Forest Biomaterials and Technology
Jägmästarprogrammet
Umeå, 2026



Cluster analysis of historical harvesting sites in Sweden

Klusteranalys av historiska avverkningar i Sverige

Elliot Eriksson

Supervisor:	**Justin Herdegen, Swedish University of Agricultural Sciences, Department of Forest Biomaterials and Technology**
Examiner:	Dimitris Athanassiadis, Swedish University of Agricultural Sciences, Department of Forest Biomaterials and Technology

Credits:	30 credits
Level:	Advanced level, A2E
Course title:	Master's thesis in Forestry Science
Course code:	EX1033
Programme:	Jägmästarprogrammet
Course coordinating dept:	Department of Forest Biomaterials and Technology
Place of publication:	Umeå
Year of publication:	2026
Keywords:	Cluster analysis, Unsupervised Machine Learning, Knowledge Discovery in Databases, Data mining, Forestry operations, Harvesting, Spatial-temporal data

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

Abstract

Human activity has had a profound effect on the environment and has increased the global average temperature through emissions of greenhouse gases. Although the member states of the United Nations signed Agenda 2030 towards sustainability, harvesting operations in Sweden are still heavily dependent on fossil fuels. Solutions to this are shifting to renewable fuels and electrifying work processes in harvesting operations. These solutions are however challenged by a lack of knowledge on the spatial distribution of worksites, the conditions under which work is done and requirements for the machines. The availability of data from wide scale and everyday data collection of forestry operations enables cluster analysis as a viable option for identifying patterns and help understand the requirements for shifting to renewable energy sources in forestry.

This study aims to identify clusters and patterns of harvesting worksite activity in combination with varying site conditions and landscape infrastructure through clustering analysis of spatial and spatiotemporal data on harvesting operations, soil and weather data.

This study identifies spatiotemporal clusters in historical harvest data through unsupervised cluster analysis, clusters harvesting operations through forestry-related variables and identifies common patterns in harvesting operations. These results can be used to better understand the spatial and temporal distribution of harvesting operations and can be used to further this area of research. This study also identifies the problem of a lack of accurate and wide scale data for cluster analyses, which should be addressed by policymakers to help better understand and mitigate the present challenges in forestry.

Keywords: Cluster analysis, Unsupervised Machine Learning, Knowledge Discovery in Databases, Data mining, Spatial-temporal data, Forestry operations, Harvesting

Sammanfattning

Antropogen aktivitet har genom utsläpp av växthusgaser drastiskt förändrat miljön och ökat den globala medeltemperaturen. Trots att medlemsländerna i Förenta nationerna undertecknade Agenda 2030 mot hållbarhet är skogsavverkningar i Sverige fortfarande ytterst beroende av fossila drivmedel. Lösningar till detta är att nyttja förnybara bränslen och elektrifiera arbetsprocesser inom skogsavverkning. Dessa lösningar utmanas dock av bristande kunskap om skogsavverkningars spatiala distribution, arbetsförhållandena och kraven på arbetsmaskinerna. Tillgängligheten av data från utbred och kontinuerlig datainsamling inom skogsavverkning möjliggör klusteranalys som ett gångbart alternativ för att identifiera mönster och öka förståelsen om kraven för övergången till förnyelsebara energikällor inom skogsavverkningar.

Denna studie syftar till att identifiera kluster och mönster inom skogliga avverkningstrakter i kombination med varierande avverkningförhållanden och landskapsinfrastrukturer genom klusteranalys av spatiala och spatiotemporala data av skogsavverkningar samt jord- och väderdata.

Studien identifierar spatiotemporala kluster i historiska avverkningsdata genom oövervakad klusteranalys, klustrar skogsavverkningar baserat på skogsbruksrelaterade variabler och identifierar vanliga mönster inom skogsavverkningar. Dessa resultat kan användas för att bättre förstå den spatiala och temporala distributionen av skogsavverkningar och för vidare forskning. Studien identifierade även en bristande tillgång till korrekta och storskaliga data for klusteranalyser. Detta problem borde hanteras av beslutsfattare för att öka förståelsen om och motverka utmaningarna inom dagens skogsbruk.

Preface

This 30-credit master's thesis was the final project to graduate my 300-credit programme and receive the title of Jägmästare from the Swedish University of Agricultural Sciences. Choosing this topic was part of my ambition to always learn something new and would not have been possible without my supervisor, Justin Herdegen, and his extensive knowledge on programming and thoughtful support and encouragement. Along with Justin, I would also like to thank Anders Rowell who kindly shared his dataset of proprietary company data.

Table of contents

List of tables

List of figures

Abbreviations

C	Cluster
CRS	Coordinate reference system
D	Database
d_{sij}	Spatial Euclidian distance
d_{tij}	Temporal absolute distance
DBSCAN	Density-based spatial clustering algorithm for applications with noise
ε_1	Spatial threshold
ε_2	Temporal threshold
GPKG	GeoPackage
HDBSCAN	Hierarchical density-based spatial clustering algorithm for applications with noise
k	Number of clusters
MinPts	Minimum number of observations to form clusters
NetCDF	Network common data form
$N_{ST}(p_i)$	Spatiotemporal neighbourhood of p_i
p_i	Observation i
SFA	The Swedish Forest Agency (Skogsstyrelsen)
SGU	Sveriges geologiska undersökning (The Swedish Geological Survey)
SLU	Sveriges lantbruksuniversitet (The Swedish University of Agricultural Sciences)
SMHI	Sveriges meteorologiska och hydrologiska institute (The Swedish Meteorological and Hydrological Institute)
ST-DBSCAN	Spatiotemporal density-based spatial clustering algorithm for applications with noise
t_i	Temporal variable of i
TIF	Tagged image file
VSOP	Värdering Skoglig Operativ Planering
XLSX	Excel open XML spreadsheet
x_i	Spatial variable of i
y_i	Spatial variable of i

1. Introduction

As stated by the International Panel on Climate Change, human activity has profoundly altered the environment and increased the global average temperature through the emission of greenhouse gases (International Panel on Climate Change 2023). To combat the threat of climate change, the member states of the United Nations signed Agenda 2030 towards sustainability (United Nations 2015). The agenda contains several sustainable development goals on clean energy sources, actions for climate change mitigation and protection of terrestrial ecosystems towards 2030 (ibid.). Forestry is commonly linked to these goals but the effects of forestry on climate change are heavily debated throughout contemporary debates and scientific papers (Englund et al. 2025).

Yet, when assessing key parts of the Swedish forestry sector, there is no ambiguity regarding the present dependency on fossil fuels for operating forestry machines. Statistics provided by the Forestry Research Institute of Sweden shows that 1.75 litres of diesel were used per harvested cubic meter, solid under bark, in 2023, with an increasing trend from 2020 to 2023 (Eliasson 2024). In the Swedish Environmental Protection Agency's statistics of national emissions, forestry machines caused approximately 1.1% of Sweden's total emissions of carbon dioxide equivalents in 2024 (Swedish Environmental Protection Agency 2025). This is a staggering increase from the previous year's emissions by approximately 41.7%, mainly due to the reduced legal requirements on mixing in non-fossil diesel with fossil diesel (ibid.). The Swedish Environmental Protection Agency also points out key factors for reducing the emissions of forestry machinery, namely, to increase usage of renewable fuels and electrify machinery (ibid.).

A closely related project on electrifying transports of forestry products called Transition to Efficient, Electrified Forestry Transport pinpoints the complexity of electrifying work processes in the Swedish forestry sector. The challenges of this transport- and forestry-related project are described as understanding the spatial dispersion of worksites, vehicle requirements and the tough outdoor conditions under which work is done (Nimbnet n.d.). Transitioning to renewable fuels for forestry machines is clearly challenged in a similar way as Transition to Efficient, Electrified Forestry Transport.

Given the increasing amount of data collection in most work processes nowadays, cluster analysis could be used to overcome these challenges by establishing patterns in forestry activities. Cluster analysis is a common and powerful tool used in many areas of engineering and scientific applications (Birant & Kut 2006). The common use of cluster analysis falls under the scientific field of

knowledge discovery in databases, where cluster analysis adheres to the group of unsupervised learning processes. This means that clustering algorithms can, without prior knowledge of data, autonomously cluster observations according to similarities and differences therein and thereby establish patterns and discover new knowledge (Ester et al. 1996).

There exist several clustering algorithms adapted to different needs and fields of research. For instance, Density-based Spatial Clustering of Applications with Noise (DBSCAN) and its variants Spatiotemporal DBSCAN (ST-DBSCAN) and Hierarchical DBSCAN (HDBSCAN) utilizes density differences of data distribution to create clusters (Ur Rehman 2014; Birant & Kut 2006; McInnes et al. 2017). This approach is especially useful for data with varying dense distributions as it allows for effective outlier detection.

Another example of clustering algorithms is the K-means algorithm (Ahmed et al. 2020). K-means cluster data by minimizing intra-cluster variance (hence also maximizing inter-cluster variance) to find clusters with as similar properties as possible (Chong 2021). The mentioned algorithms are commonly used in various applications and settings to detect patterns and similarities in data. Hence, they could prove useful in understanding the common traits and patterns of forestry activities.

To understand the spatiotemporal patterns of forestry, it is vital to understand that harvest times of individual sites in Sweden are generally determined by three factors (Lundqvist et al. 2014). These factors are:

- operational planning of wood flow and harvesting resources,
- terrain accessibility and
- road accessibility.

The operational planning of wood flow and harvesting resources is generally done by the harvesting contractor and combines several planning steps to maximize profitability and minimize time and costs of wood procurement. Some examples of decisions are which stands to harvest to meet the delivery goals while minimizing stand entry and road maintenance costs, or which harvesting team to use to minimize movement costs (Frisk et al. 2016, Başkent & Jordan 1991). The conditions under which forestry operations are done are commonly classified for planning purposes. There exist several types of classifications for terrain and forestry roads such as Berg's Terrain Classification System for Forestry Work (1995) and the accessibility classes for forestry roads (Swedish Forestry Research Institute 2014). These classifications will be utilized in this study.

This study aims to identify clusters and patterns of harvesting worksite activity in combination with varying site conditions and landscape infrastructure through clustering analysis of spatial and spatiotemporal data on harvesting operations, soil and weather data. To achieve this aim, the following research questions will be answered:

- Can harvesting operations be spatiotemporally clustered through unsupervised cluster analysis?
- Can harvesting operations be clustered using forestry-related variables?
- Are there patterns in harvesting operations and if so, what patterns are there?

To answer these questions, a database was also generated wherein spatial and spatiotemporal data can be stored and further analysed with statistical and geographical information system tools. The scope of the study was limited by the availability of data. Historical harvest data on regeneration harvests from the Swedish Forest Agency was used for one cluster analysis while company data of a major Swedish forestry company was used for further cluster analysis. The scope of the study was therefore limited to Sweden for periods of time and space where usable data was accessed.

2. Material and methods

2.1 Methodological approach of the thesis project

This study was methodologically split into four parts as seen in figure 1. Firstly, spatial and spatiotemporal data of different file-types were collected. Secondly, a database (Harvest database) was structured from the data by cleaning and filtering it and then merging the different file-types into one combined database. Thirdly, cluster analyses were done on the collected data and the Harvest database through dimensional reduction and clustering algorithms. Dimensional reduction was done through factor analysis of mixed data, which, like principal component analysis, reduces the original variables into components of shared correlation. Lastly, the results were presented in this report.

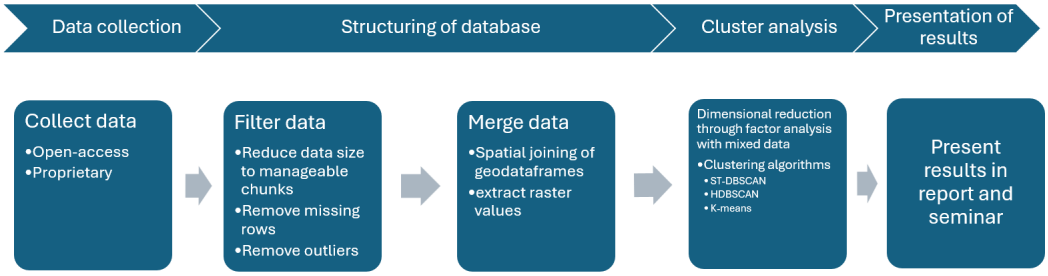


Figure 1. Methodological approach of this thesis project.

2.2 Data

The collected data is comprised of spatial and spatiotemporal data spatially covering Sweden. Data was collected from several sources. As seen in Table 1, mainly open-access data from Swedish agencies involved in the geospatial sector was used. A proprietary dataset comprising of historical harvesting sites from Holmen Skog Aktiebolag, a major Swedish private forestry company, was also used with consent.

Table 1. Overview of all the datasets used in the study. The SFA harvest data is only used for clustering with ST-DBSCAN and not included in the database.

Name	File type	Source	Access	CRS	Temporal coverage
SFA harvest data	**GPKG**	**(Swedish Forest Agency 2025b)**	Open	SWEREF 90 TM	1961-2025
Company data	**XLSX**	**(Rowell 2022)**	Proprietary	SWEREF 90 TM	2013-2020
Soil type data	**GPKG**	**(SGU 2024)**	Open	SWEREF 90 TM	-
Soil moisture data	**TIF**	**(SLU 2020a)**	Open	SWEREF 90 TM	-
Precipitation data	**NetCDF**	**(SMHI 2025)**	Open	RT 90/WGS 84	1961-2025
Temperature data	**NetCDF**	**(ibid.)**	Open	RT 90/WGS 84	1961-2025

2.2.1 SFA harvest data

SFA harvest data is part of the "Skogliga grunddata"-database provided by the Swedish Forest Agency. The database was provided as a GeoPackage containing a polygon layer of historical harvesting sites georeferenced for the whole of Sweden. The polygons also contain attribute data. The only attribute data column used in this cluster analysis was the Harvest date column.

The database is continuously updated with new historical harvesting data by the Swedish Forest Agency, reported to the agency mainly due to the report duty associated with some types of harvests in accordance with the Swedish forestry act (Swedish Forest Agency 2025a). Harvesting worksites are thus mostly reported beforehand by forest owners or forestry contractors if the harvest falls under the report duty. This leads to many harvest sites being excluded from the database, such as most thinnings. The database does therefore not provide a complete picture of Swedish forestry activities. Due to limitations in computational power, only the harvesting type Regeneration harvests (Föryngringsavverkning) was used in the analysis, and the analysis was done for three representative counties instead of the whole of Sweden. Kalmar, Dalarna and Västerbotten counties were used as representatives for southern, middle and northern Sweden, respectively.

The harvest date of each historical harvest is almost exclusively set through computerized picture analysis of satellite pictures[1] by the Swedish Forest Agency. Due to satellite pictures being insufficient for computerized assigning of harvest dates during times with snow cover, harvest dates during winter seasons are unreliable. Therefore, only harvests occurring between the months of June and September were used in the analysis. In correspondence with the Swedish Forest Agency, the accuracy of harvest dates increased significantly after 2022 (ibid.). Therefore, only harvest dates after 2022 were used.

Although only including regeneration harvests for summer months 2023-2025, the database was still sufficient to carry out spatiotemporal analyses. The plan was originally to include this data in the Harvest database. Yet, due to harvesting sites not overlapping correctly with the Company data when combining the datasets, computerized spatial merging of the data was deemed practically impossible. Therefore, this dataset was used for clustering analysis using the ST-DBSCAN algorithm while the Harvest database was used with other cluster analysis methods.

2.2.2 Company data

The Company data is an Excel-sheet data frame where each row is a unique historical harvesting site, and each column describes an attribute of that historical harvesting site. The data originates from Värdering Skoglig Operativ Planering (VSOP), an operative planning system used by Holmen skog Aktiebolag (Rowell 2023). Company-specific data such as names of administrative areas, machine identification numbers, etc. were discarded for the analysis as they posed no value. Instead, non-company-specific spatial and spatiotemporal variables were kept for the analysis. Unfortunately, data in several columns were heavily distorted or missing for unknown reasons and were deemed unusable. The data that were used in the analysis were: Harvested volume, End-date of harvest, Harvest type, X coordinate, Y coordinate, Road accessibility, Ground conditions, Hauling distance and Mean stem volume.

[1] Customer support, Swedish Forest Agency, E-mail 2025-06-17

The Ground conditions in the Company data is classified based on Berg's Terrain Classification System for Forestry Work (1995). This system classifies terrain accessibility of harvesting sites according to Ground conditions, Surface roughness and Slope by ordinal numbers from 1 to 5, where lower numbers indicate better accessibility (ibid.). Forestry roads are normally classified based on accessibility into four ordinal classes, from A to D, based on what time of year they are deemed operational for forestry transports (Swedish Forestry Research Institute 2014). This road classification can also occur in numbers from 1 to 4, where lower numbers indicate better accessibility. In this study, numbers were used.

The X and Y coordinates were used to georeference each historical harvest site. However, it is uncertain whether the coordinates define the centroid of the harvest site, the landing site or are the reference coordinates for emergency aid. Visual inspection of the data suggests the coordinates reference all these types of locations arbitrarily.

2.2.3 Soil type data

The Soil type data is provided by the Swedish Geological Survey (SGU) as a georeferenced GeoPackage. The dataset provides a detailed yet general picture of soil features and should according to the Swedish Geological Survey be used carefully in analyses due to potential errors in the data. The Geopackage contains several layers describing soil features and land formations as categories. Yet, the only layer with complete spatial cover of Sweden and with relevance to forestry is the Soil base layer. Therefore, only that layer was used in the analysis. The Soil base layer provides a complete picture of soil types approximately 0.5m below the soil surface with an estimated mean error of 25m (Swedish Geological Survey 2024).

2.2.4 Soil moisture data

The Soil moisture data is provided by the Swedish University of Agricultural Sciences (SLU) and utilizes a trained machine learning model to classify soil moisture based on the classification used by the Swedish National Forest Inventory (SLU 2020b). The Soil moisture data is provided in a TIF-file with 2x2m raster squares containing soil moisture values. The soil moisture values are ordinal from 1 to three where 1 are dry-mesic, 2 are mesic-wet and 3 are wet soils. The data also contains raster squares with the value 15, which is a NULL-value (SLU 2020a).

2.2.5 Precipitation and temperature data

The Precipitation and Temperature data are both provided by the Swedish Meteorological and Hydrological Institute (SMHI). SMHI utilizes gridded analysis methods and interpolation between weather measuring points to create 4x4 km raster squares with weather data (SMHI 2025). The dataset contains precipitation [mm/day] and temperature [°C] values for every day and raster square. Notably, the data is gridded in RT 90, making reprojection to fit the coordinate reference system of all the other data necessary.

2.3 Data treatment

2.3.1 SFA harvest data

The harvest dataset provided by the Swedish Forest Agency needed to be pretreated before being usable in the given ST-DBSCAN algorithm. Firstly, the georeferenced polygons needed to be reduced into point coordinates. Therefore, a centroid was calculated for each polygon, from which X and Y coordinates were extracted. Secondly, the harvest date timestamps provided in the dataset was transformed from [year-month-day] to [number of days since first observation] to fit the algorithm's input data format. With the new unit of harvest dates, the earliest occurring harvest in the dataset was thus assigned Harvest date = 0 days while, for example, a harvest that occurred exactly one year later was assigned Harvest date = 365 days.

2.3.2 Historical harvest database

The private forestry company's historical data formed the basis of the database but first needed to be treated. It was first manually cleaned in Excel using the CLEAN ()-function to remove formatting residues probably originating from VSOP. String values of numerical variables were converted into number values using the VALUE ()-function in Excel.

Due to the presence of missing values for several rows in columns that are important to the analysis, case deletion was used for rows containing missing data in variables later used in the analysis. As the source of error is unknown for this dataset, there is no other option but to remove rows with missing values (Sharifnia et al. 2026). This was the largest exclusion of data for the whole project, reducing the size of the data from 65,871 to 50,965 rows. To also exclude obvious outliers detected during visualisation of the data (ibid.), the filtering rules in Table 2 were applied on the data. This final exclusion reduced the data size to 50,925 rows.

Table 2. Rules applied to variables to filter out obvious outliers

Variable	Filter	# of rows excluded	$\bar{x}$	σ
Harvested volume	$\bar{x} - 6\sigma \leq x_i \leq \bar{x} + 6\sigma,$ $x_i > 0$	13	$932 m^3 sub$	$1200 m^3 sub$
Mean stem volume	$\bar{x} - 3\sigma \leq x_i \leq \bar{x} + 3\sigma,$ $x_i > 0$	25	$0.275 m^3 sub$	$2.11 m^3 sub$
Hauling distance	$\bar{x} - 3\sigma \leq x_i \leq \bar{x} + 3\sigma$	2	$344 m$	$1560 m$

A modification of the End-date of harvest was also applied to better capture the seasonal changes in the analysis. Instead of describing the time of year as a category or numbers with no linkage to seasonal changes, End-date of harvest was transformed into two variables using the formulas described in Equation 1 and 2. By using this data format, harvest time of year is accurately described with a set of values from Time of $year_1$ and Time of $year_2$ while the gradual change of seasons is captured continuously as numerical variables.

$$\text{Time of year}_1 = \sin \frac{2\pi * Month\ of\ harvest}{12} \quad (1)$$

$$\text{Time of year}_2 = \cos \frac{2\pi * Month\ of\ harvest}{12} \quad (2)$$

With the Company data prepared, the SQL-based database format GeoPackage was used to store the data in a single point layer. The X and Y coordinates of the Company data were used for georeferencing. The resulting GeoPackage is visualized in Figure 2.

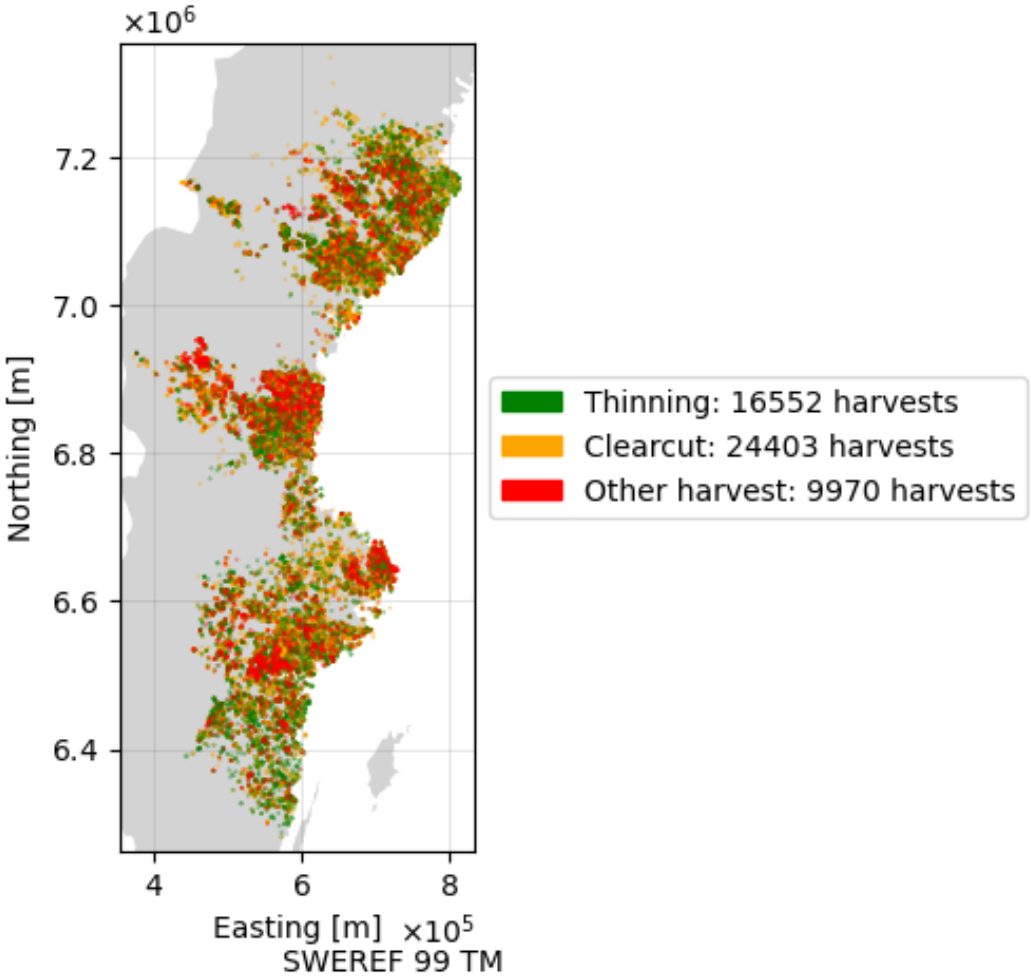


Figure 2. The Company data visualized over a map of Sweden.

Thereafter, the Soil type data from the Swedish Geological Survey was spatially merged to the harvest sites in the database. The Soil base layer contains many soil categories which were generalized into five general categories before analysis. The generalized categories used in this study are Organic, Friction, Cohesion, Bedrock and Other soil types. Thereafter, Soil moisture data from SLU is spatially merged into the database. This merge simply appended the moisture value of the raster square each harvest site is located in.

The Precipitation and Temperature data from SMHI were both georeferenced in the coordinate reference system (CRS) WGS 84 and gridded in RT 90. To use the data in the SWEREF 90 TM-referenced database, both datasets were transformed into SWEREF 90 TM using bilinear resampling (Dorman 2025). Resampling raster data into a new CRS can create a slight tilt in the data but the effects were quite minor, only moving the raster borders approximately 100-200m at most.

With the weather data prepared, Precipitation data were appended to each row by first calculating 14-day precipitation sums for every day and raster square. Each harvesting site in the database was then matched to the correct raster square and date and were thusly assigned a 14-day precipitation sum. Temperature data was merged similarly, although 14-day mean temperatures was calculated and assigned, instead of sums.

2.4 ST-DBSCAN clustering

ST-DBSCAN is a type of algorithm based on DBSCAN (Birant & Kut 2006). ST-DBSCAN distinguishes points into Clusters or as Noise points in a given spatiotemporal database by utilizing a set of definitions and functions (ibid.). The variant of ST-DBSCAN algorithm used in this project utilizes the definition and functions described below (Cakmak et al. 2021, Birant & Kut 2006).

Given a database D (3) containing points p_i (4) with temporal and spatial variables, t_i [number of days since first observation], x_i [m] & y_i [m], respectively (5), the algorithm calculates temporal and spatial distances between each point (6 & 7).

$$D = \{p_i | i = 1,2, \dots, n\} \tag{3}$$

$$where \qquad p_i = (t_i, x_i, y_i) \tag{4}$$

$$where \qquad t_i = temporal\ variable, x_i\ \&\ y_i = spatial\ variables \tag{5}$$

The temporal distance (d_{tij}) is the absolute value of the difference between the temporal variables meanwhile the spatial distance (d_{sij}) in this case is the Euclidian distance between the points. The algorithm utilizes a spatial and temporal threshold, ε_1 [m] & ε_2 [days] (9), a spatiotemporal neighbourhood for each point, $N_{ST}(p_i)$, is defined (8).

$$d_{tij} = |t_i - t_j| \tag{6}$$

$$d_{sij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{7}$$

$$N_{ST}(p_i) = \{\, x_j \in D | d_{sij} \leq \varepsilon_1 \wedge d_{tij} \leq \varepsilon_2 \} \tag{8}$$

$$\text{where} \qquad \varepsilon_1 = spatial\ threshold\ \&\ \varepsilon_2 = temporal\ threshold \tag{9}$$

The spatiotemporal neighbourhood is then used to define Core points (10). For a point to qualify as Core point, the spatiotemporal neighbourhood must equal or exceed MinPts (11), which is a minimum neighbourhood size for clusters to form (12).

$$p_i = Core\ point \tag{10}$$

$$if \quad N_{ST}(p_i) \geq MinPts \tag{11}$$

$$where \quad MinPts = minimum\ neighborhood\ size \tag{12}$$

Core points are the nucleation sites of clusters, but other points can also belong in clusters if they are either directly density reachable from a Core point (13 & 14), density reachable to another Core point (15, 16, 17 & 18) or density connected to another density reachable point (19, 20 & 21). In essence, directly density reachable means that a point is part of a Core points spatiotemporal neighbourhood. A point is density reachable when it is connected to a Core point through a chain of spatiotemporally close points. Likewise, two points are density connected when they are linked by a chain of spatiotemporally close points. However, none of the points are Core points for this to occur. Instead, it is sufficient that one of the points are density reachable.

$$p_i = directly\ density\ reachable\ from\ p_j \tag{13}$$

$$if \quad p_i \in N_{ST}(p_j) \wedge p_j = Core\ point \tag{14}$$

$$p_i = density\ reachable\ from\ p_j \tag{15}$$

$$if \quad \exists \{ p_1, p_2, \dots, p_k \} \subseteq D \tag{16}$$

$$where \quad p_1 = p_j \ \& \ p_k = p_i \tag{17}$$

$$and \quad p_{\{l+1\}} = directly\ density\ reachable\ from\ p_l$$

$$for\ all\ l = 1, \dots, k-1 \tag{18}$$

$$p_i \ \& \ p_j = Density\ connected \tag{19}$$

$$if \quad \exists\ p_k \in D \tag{20}$$

$$such\ that \quad p_i \cap p_j = \ density\ reachable\ from\ \ p_k \tag{21}$$

A cluster, C, is formed from the given database D (22) when a core point is detected (11). The algorithm then processes all neighbouring points within the temporal and spatial thresholds (9). Points are added to the cluster if either a point is density reachable to a cluster (23) or density connected to a cluster (24). This process is repeated until all points have been clustered or labelled as Noise points (25). Noise points are those points that did not match into any clusters (26).

$$C \subseteq D = Cluster \tag{22}$$

$$if \quad \forall\, p_i \in C, if\ p_j = \ density\ reachable\ from\ p_i, \quad then\ p_j \in C \tag{23}$$

$$and \quad \forall\, p_i\ , p_j\ \in\ C, \quad p_i\ \&\ p_j = density\ connected \tag{24}$$

$$p_i = Noise\ point \tag{25}$$

$$if \quad p_i \notin C\ for\ all\ clusters\ C \tag{26}$$

When calling the algorithm on a given database, three parameters are inputted to adjust the clustering. These parameters are the spatial and temporal thresholds, ε_1 & ε_2 (9) and the minimum neighbourhood size, MinPts (12). To simulate the temporal and spatial thresholds relevant to forestry in Sweden, the parameters were set to $\varepsilon_1 = \frac{4000}{1.3}$ m ≈ 3077m, $\varepsilon_2 = 13$ days and MinPts = 5. $\varepsilon_1 = \frac{4000}{1.3}$m was used as 4km is a common, yet arbitrary distance used in Sweden for how long harvesters and forwarders travel by road before it is more profitable to use trailer transport. ε_1 was divided by a winding factor of forestry roads to compensate for the fact that the algorithm uses Euclidian distance while forestry machines travel between harvest sites by road. In this analysis, a winding factor of 1.3 was used, which have been used in other forestry literature (Fjeld & Dahlin 2008). ε_2 = 13 days was used as the mean harvest length of the Company data is 12.4 days. MinPts = 5 was deemed a sufficient minimum cluster size for the clusters to be meaningful.

To validate the naturally clustered structure of the data used in the ST-DBSCAN, Knox tests were calculated together with every ST-DBSCAN cluster. Even though originally being a tool for measuring contagiousness of diseases in the field of epidemiology, the Knox test is regardless a statistical tool for analysing spatiotemporal data structures (Knox 1964). It provides a Knox ratio and a p-value for statistical significance.

The Knox ratio is a descriptive measurement of how spatiotemporally clustered the data is. It is calculated by dividing the observed number of spatiotemporal close-pairs by the expected number of spatiotemporal close-pairs if space and time are independent (27). The p-value is the likelihood of finding the observed number of close-pairs in the dataset compared to the same observations and spatial distribution under temporal randomness (ibid.). When calculating the Knox ratios and p-values, 999 permutations are used and the same temporal and spatial thresholds are used as in the ST-DBSCAN.

$$Knox\ ratio = \frac{Observed\ space - time\ close\ pairs}{Expected\ space - time\ close\ pairs} \quad (27)$$

2.5 Factor analysis with mixed data and cluster analysis

2.5.1 Factor analysis with mixed data

To explore the joint structure of numerical, ordinal and categorical variables in the database, factor analysis with mixed data was used. The method stems from adapting both multiple correspondence analysis and principal component analysis into handling both qualitative and quantitative data in the same analysis (Pagès 2014).

To run the factor analysis with mixed data properly, quantitative variables (numerical and ordinal) were centred and standardised. Qualitative variables (categorical) were transformed into binary indicator variables using one-hot encoding, whereby each category is represented as a binary indicator variable (1 for yes, 0 for no). This increased the dimensionality of the data but allowed for proper incorporation of the data into the analysis. With the expansion of categorical data, each categorical variable was transformed into a variable for each category. In this case, Soil type and Harvest type become five and three dimensional, respectively, increasing the dimensionality of the data from 14 to 20 variables (Table 3).

Table 3. The variables of the database presented based on the variable type, variable analysis name and the number of dimensions the variable adds to the factor analysis with mixed data

Type	Variable	Dimensions
Numerical	**Precipitation**	1
	Temperature	1
	Time of year$_1$	1
	Time of year$_2$	1
	Hauling distance	1
	Harvested volume	1
	Mean stem volume	1
	Y coordinate	1
	X coordinate	1
Ordinal	**Soil moisture**	1
	Road accessibility	1
	Terrain accessibility	1
Categorical	**Soil type**	5
	Harvest type	3
Sum	**14**	**20**

Due to limitations in computational power, no ordinary factor analysis with mixed data library was used (e.g. Prince). Instead, the scikit-learn Incremental Principal Component Analysis was adapted to exactly replicate factor analysis with mixed data mathematically (Pedregosa et al. 2011). This allowed the analysis to be done in batches, allowing the analysis to run smoothly even for huge datasets.

Like a principal component analysis, factor analysis with mixed data produces components which are linear combinations of the original variables to capture maximum amounts of shared variance and association in the data. Each observation is also assigned a component score based on the observation's contribution to each component. The scores can be used for cluster analysis. The method of extracting component scores before clustering allows for using clustering algorithms which might struggle with multidimensional data, one example being K-means (Ahmed, Seraj & Shamsul Islam 2020). Some examples of implementing factor analysis with mixed data before cluster analysis are analyses on patterns and factors in biodiversity (Silva-Souza et al. 2023) and enhancing costumer segmentation techniques (Ufeli et al. 2025).

2.5.2 HDBSCAN

To further the endeavour of data mining the Harvest database, utilizing clustering algorithms on the factor analysis with mixed data outputs were deemed relevant methods. A density-based algorithm like DBSCAN would most likely not perform well due to DBSCAN's known issue with handling high-dimensional data (Raha et al. 2025). Another density-based clustering algorithm that generally works well with high-dimensional data is Hierarchical DBSCAN (HDBSCAN) (McInnes, Healy & Astels 2017). HDBSCAN performs DBSCAN on a given database over varying spatial thresholds (ε) and keeps the clusters with highest stability over ε and integrates the results into clusters with varying densities (ibid.). Thus, HDBSCAN does not suffer from the same "curse of dimensionality" where DBSCAN can struggle to find meaningful threshold values for all dimensions (Raha et al. 2025).

To operate the HDBSCAN algorithm on a database, two parameters are set. The first is the minimum neighbourhood size to form core points and the last is the minimum cluster size. The minimum neighbourhood size for this analysis was set to 20 and the minimum cluster size was set to 200 as it was deemed large enough to create meaningful clusters given the 50,925 observations to cluster. The clustering mode in the analysis is called "leaf" and distance is calculated using Euclidian distance.

Due to unsuccessful clustering results from HDBSCAN, the algorithm is not explained further here. To learn more about the HDBSCAN algorithm used in this study, read the paper describing the code library by McInnes, Healy & Astels (2017).

2.5.3 K-means clustering

K-means is a non-density-based clustering algorithm for unsupervised machine learning (Ahmed, Seraj & Shamsul Islam 2020) and was first described by Stuart Lloyd (1982). The version of the algorithm used in this paper is from the Scikit-learn K-Means library (Pedregosa et al. 2011).

To cluster using the K-means algorithm, an optimal number of clusters, k, must be predetermined. To do this, Silhouette scores, Calinski–Harabasz indexes and Davies–Bouldin indexes for different numbers of k are examined. Higher Silhouette scores indicate more separated clusters while higher Calinski–Harabasz indexes indicate more compact cluster. Lower Davies–Bouldin indexes indicate less inter-cluster similarity (Wang & Xu 2019).

The K-means algorithm used clusters data by utilizing a set of definitions and functions described below. Given a database, D, with n observations, represented by observations of a q-dimensional component score, p_i (28), from the factor analysis with mixed data and a predefined number of clusters, k (29), the K-means algorithm assigns each observation to exactly one cluster (30). For each cluster, j, a cluster centroid is also defined as the mean value of the all the observations assigned to that cluster (31).

$$D = \{p_i | i = 1,2, \dots, n\},\ p_i \in \mathbb{R}^q \tag{28}$$

$$1 < k < n \tag{29}$$

$$c_i \in \{1, \dots, k\} \tag{30}$$

$$\mu_j \in \mathbb{R}^q \tag{31}$$

Given the definitions above, an objective function is defined (32) in which within-cluster sum of squared distances is calculated. The algorithm works by minimizing J, effectively finding the centroids, μ_j, which assignes observation p_i so that the within variance of all clusters are minimized.

The clustering begins by randomly selecting k cluster centroids. Using an assignment step, each observation is assigned to the closest centroid (33). When all observations have been clustered, the centroids are updated by calculating the new mean value for each centroid (34). The assignment and update steps (33 & 34) are repeated until the cluster assignment of observations cease to change. This produces clusters with minimum values of the objective function (32) that might be local. The given algorithm repeats the process several times however with different starting centroids to reduce the risk of getting stuck at local minimums (Pedregosa et al. 2011).

$$J = \sum_{j=1}^{k} \sum_{i:c_i=j} \left\| p_i - \mu_j \right\|^2 \tag{32}$$

$$c_i = \arg \min_{j \in \{1,\dots,k\}} \left\| p_i - \mu_j \right\|^2 \tag{33}$$

$$\mu_j = \frac{1}{|C_j|} \sum_{i \in C_j} x_i \tag{34}$$

By utilizing K-means clustering, each observation is assigned a cluster label (30). The characteristics of each cluster was then visualized using a heatmap of the analysed variables and by spatially mapping of the clustered observations.

3. Results

3.1 Spatiotemporal DBSCAN of SFA harvest data

For all counties and years, spatiotemporal clusters were produced, although to a varying degree given the parameters used when clustering (Table 4). The Knox ratios suggest the harvesting sites are moderately to strongly spatiotemporally clustered. For some counties and years such as Västerbotten 2025, Dalarna 2024 and Kalmar 2024, cluster ratios are relatively low while Knox ratios are high. This suggests that worksites for those counties and years are spatiotemporally clustered, yet that the clusters therein consist of fewer worksites than the minimum cluster neighbourhood size parameter, MinPts, used when conducting the cluster analysis. The p-values produced with the Knox test provide strong evidence against worksites being distributed randomly in time and space.

Table 4. Results of ST-DBSCAN clustering of SFA harvest data from the Swedish Forest Agency. Spatiotemporal clustering results are presented for each county and year. The numbers of clusters found, and ratios of points assigned to each cluster are also presented. The resulting Knox-ratios and p-values are presented for each clustering analysis as well

County	Year	Clusters	Cluster ratio	Knox ratio	P-value
Västerbotten	**2023**	165	0.53	1.231	0.001
	2024	98	0.43	1.356	0.001
	2025	51	0.18	1.487	0.001
Dalarna	**2023**	67	0.41	1.220	0.001
	2024	32	0.17	1.930	0.001
	2025	94	0.45	1.223	0.001
Kalmar	**2023**	54	0.44	1.239	0.001
	2024	35	0.27	1.618	0.001
	2025	43	0.48	1.321	0.001

When visualising the ST-DBSCAN results, clear cluster structures are visible. For Västerbotten County in 2023, 165 clusters were produced (Fig. 4). As the algorithm is density-based, the clusters are harvesting operations close in space and time while outliers are more isolated observations (Fig. 3). The five largest harvesting clusters were in eastern Västerbotten, along the coast, where the largest cluster exceeded 100 harvesting sites (Fig. 5).

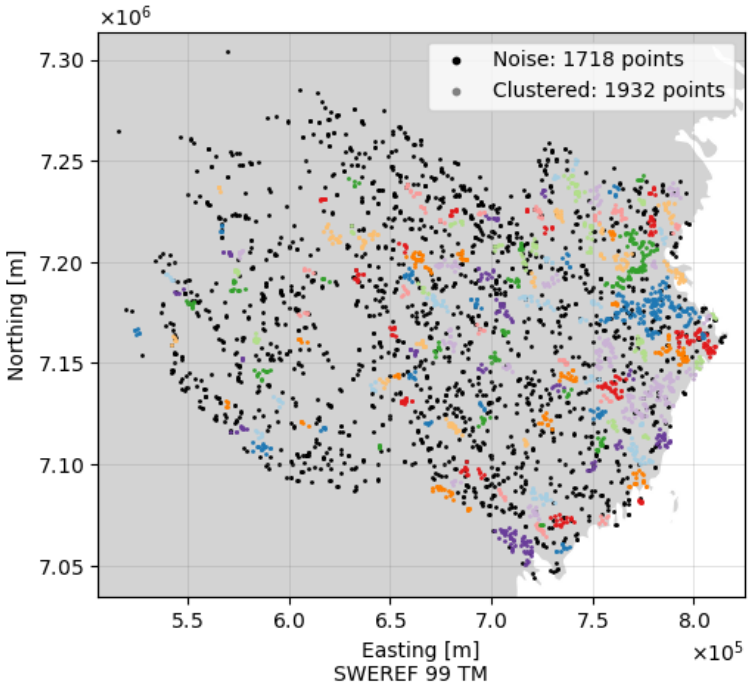


Figure 3. Output from the ST-DBSCAN algorithm on Västerbotten county June-September 2023. Points close to each other with the same colour belong to the same clusters. Outliers are represented with black points.

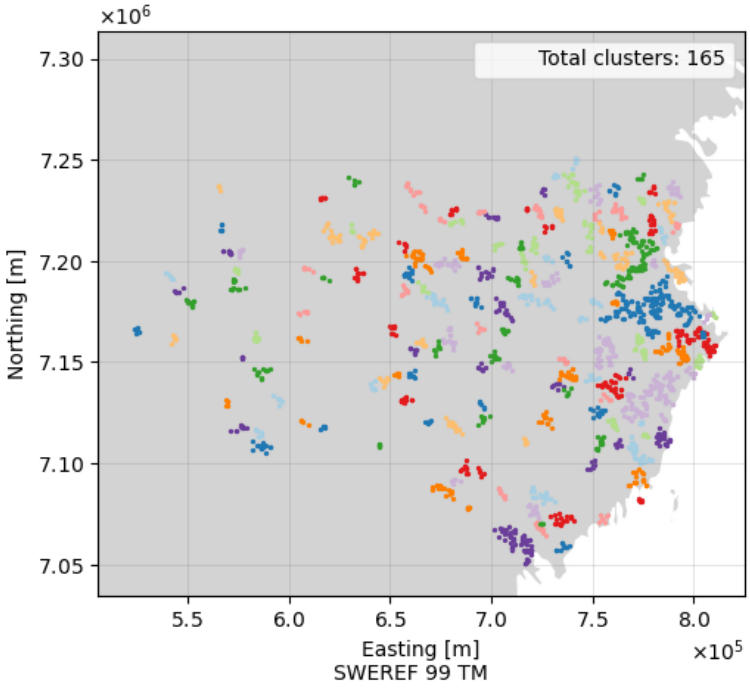


Figure 4. Output from the ST-DBSCAN algorithm on Västerbotten county June-September 2023. Clusters presented without noise points. The total amount of clusters is 165.

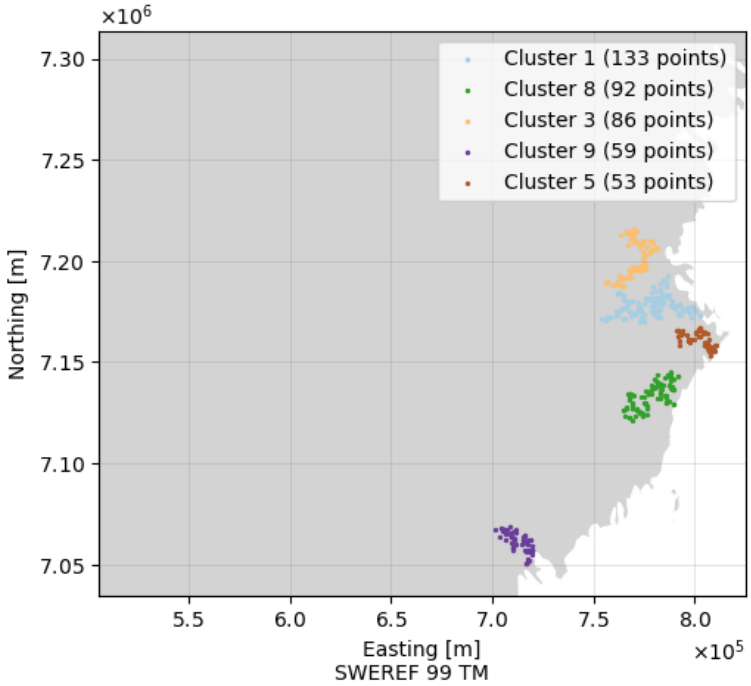


Figure 5. Output from the ST-DBSCAN algorithm on Västerbotten county June-September 2023. Only the five largest clusters are visualized.

In Figures 6-8, further visualizations of the spatial distributions of the clusters are presented. The figures do not visualize the temporal ranges of the clusters but do still prove that harvesting sites for all counties and years contain spatiotemporal clusters. In Figure 6, clusters produced in Västerbotten for 2023, 2024, 2025 are visualized. There are spatiotemporal clusters distributed over the whole county for all years. There is however a slight dominance of clusters along the coast of Västerbotten while the inland is sparser. Notably in 2024, a local spatiotemporal structure of harvesting sites in northern Västerbotten consisting of eight clusters formed a strong linear spatial structure in the harvesting pattern of the county.

The clustering results vary as there are approximately three times more clusters in 2023 than 2025. Yet, the highest Knox ratio for Västerbotten was in 2025 (Table 4). This indicates that the harvesting sites in 2025 are also strongly spatiotemporally clustered but oftentimes smaller than the minimum cluster size used in the cluster analysis.

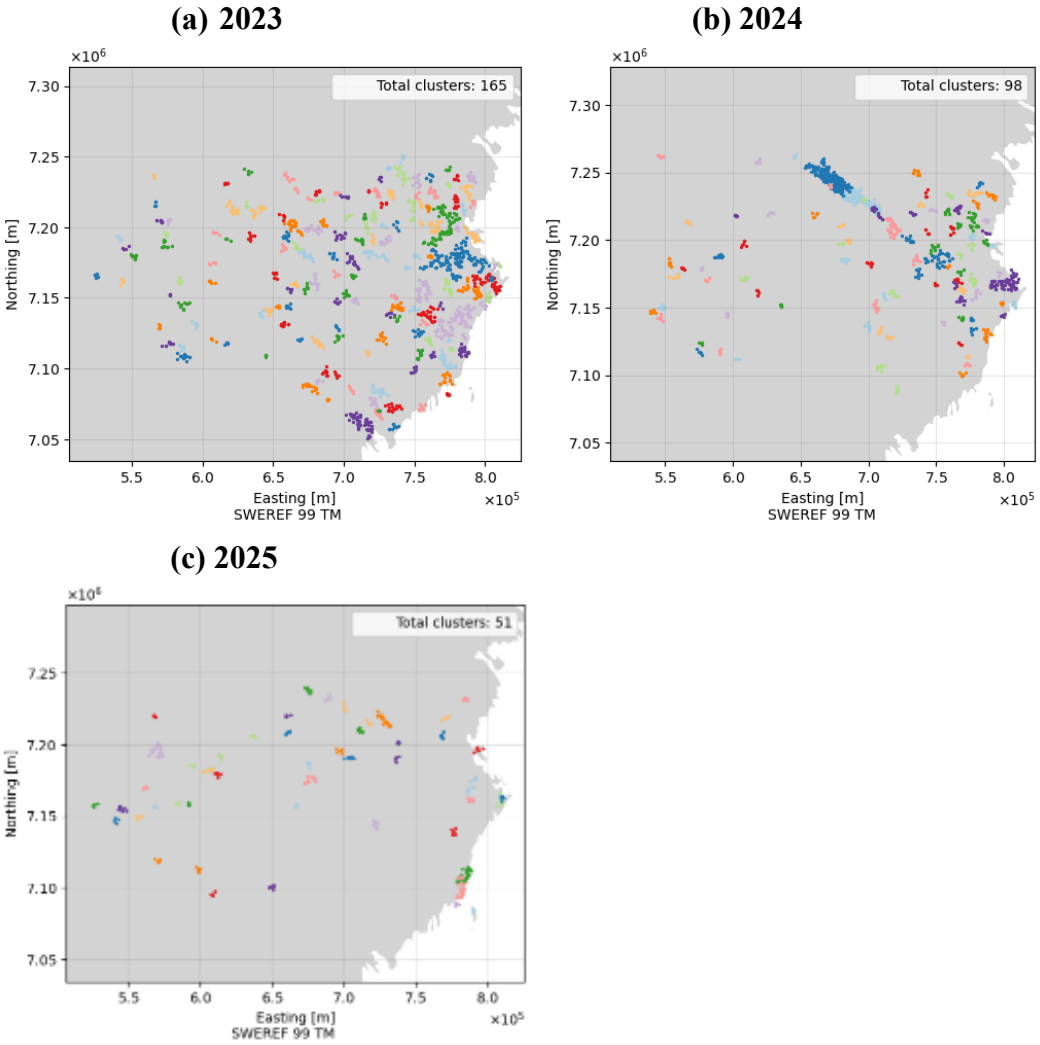


Figure 6. Clusters in Västerbotten County 2023, 2024 and 2025

Dalarna County contains spatiotemporal clusters distributed across the whole county for the analysed time periods (Fig. 7). Like Västerbotten, there is a clear variation of results between years. Although only containing 32 clusters, the Knox ratio of harvest sites in Dalarna 2024 had the highest measured Knox ratio of all counties and years (Table 4). There is therefore strong evidence for spatiotemporal clusters also in Dalarna.

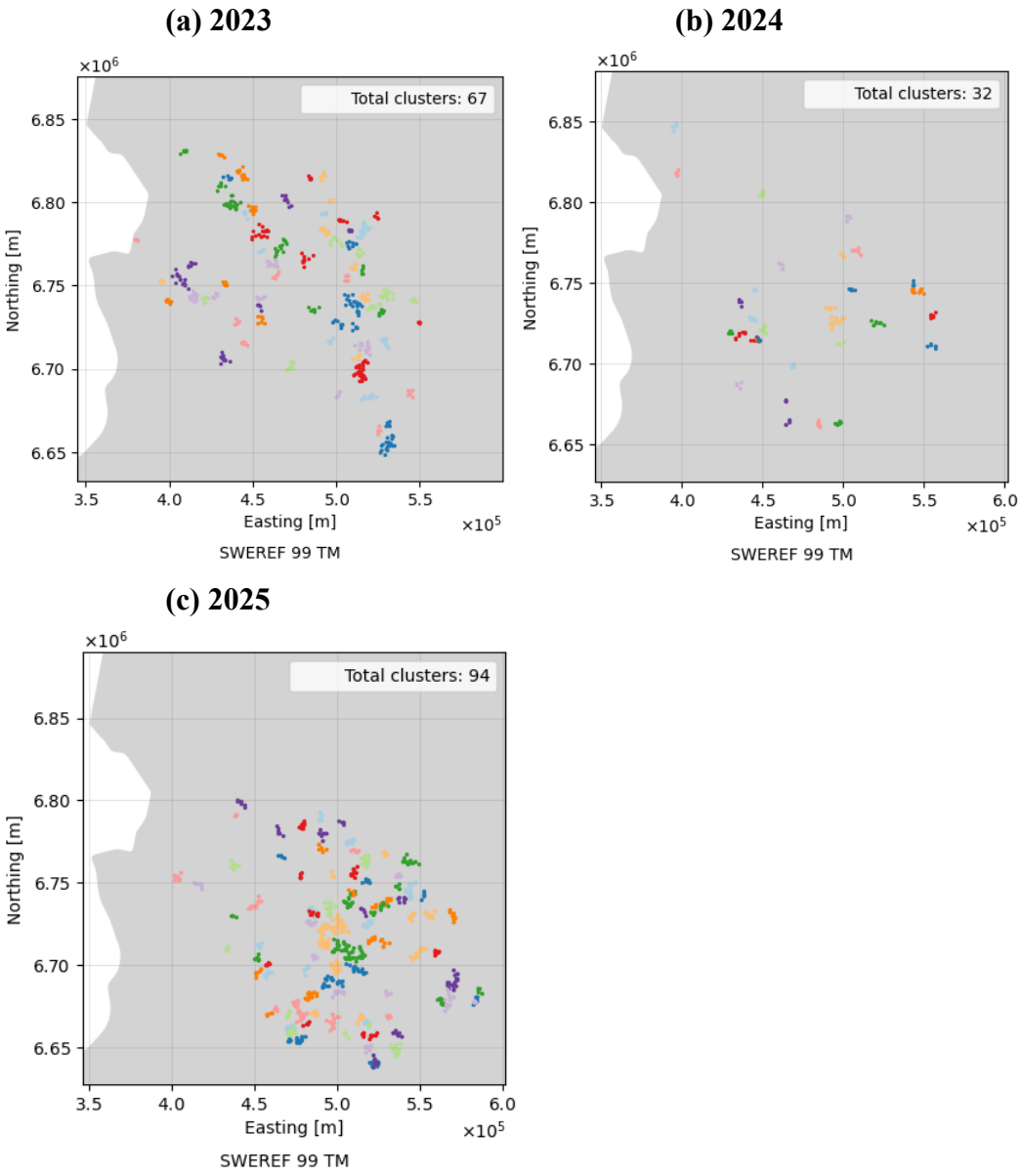


Figure 7. Clusters in Dalarna County 2023, 2024 and 2025

In Kalmar County, clustering results are more even across years than the other counties (Fig. 8). 2023 contained sparse, evenly distributed clusters. The distribution in 2024 is also even but there are more clusters, making the distribution denser. There is a spatiotemporal southern dominance in southern Kalmar in 2025. There, several large clusters are spatially very densely distributed and provide evidence toward a local harvesting hotspot in 2025. For all figures produced with the ST-DBSCAN algorithm, see Appendix 1.

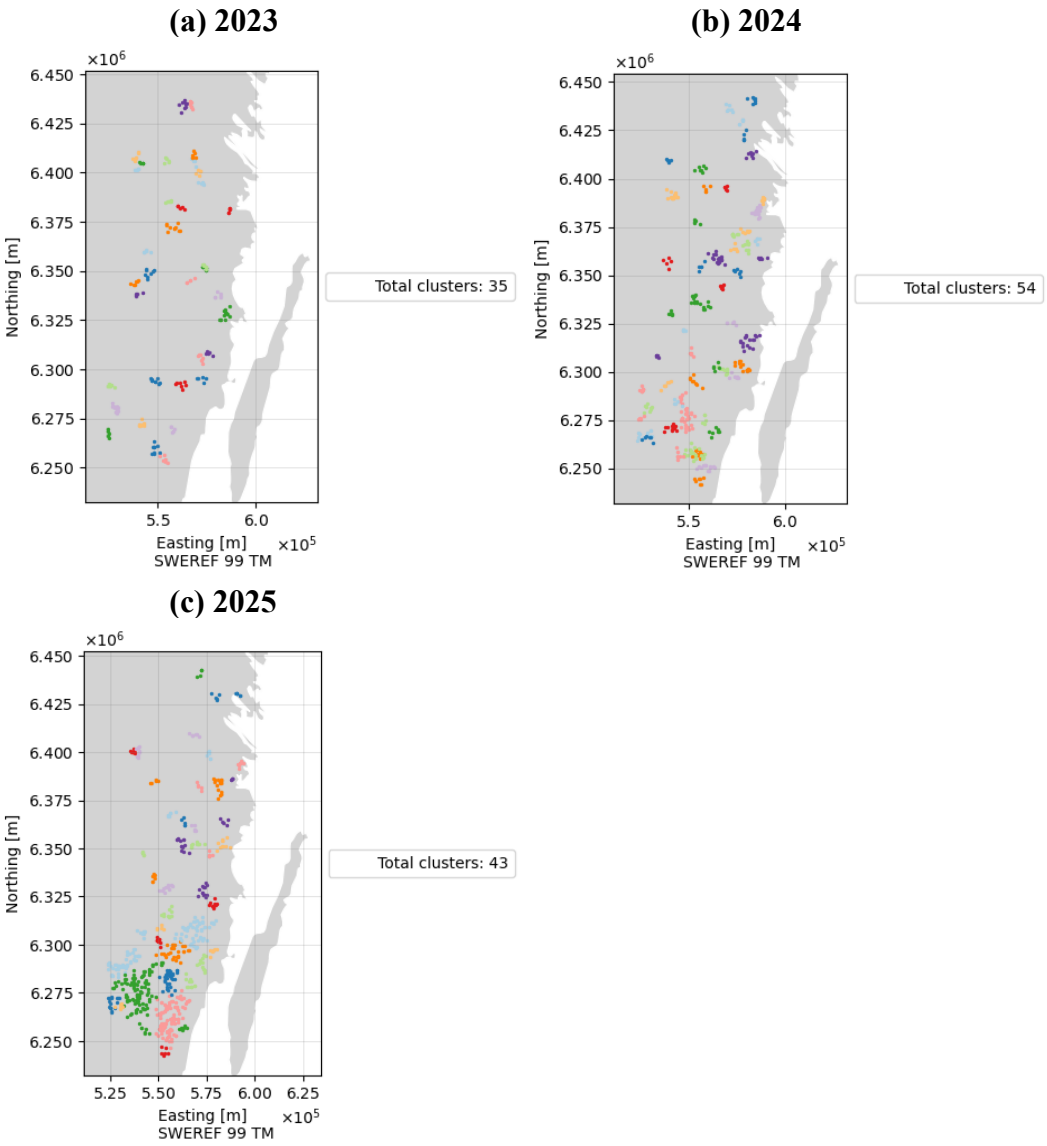


Figure 8. Clusters in Kalmar County 2023, 2024 and 2025

3.2 Factor analysis of mixed data with clustering analysis of Harvest database

3.2.1 Factor analysis with mixed data

The factor analysis of mixed data resulted in seven components, cumulatively explaining 72.3% of the variance of the dataset (Table 5). If applied, common metrics on deciding how many components to keep in principal components analysis such as the Kaiser criterion or parallel analysis would suggest keeping five components. Neither of these metrics are however fully scientifically established, especially not for factor analysis with mixed data (Braeken & van Assen 2017). Components 6 and 7 are kept in the analysis as they are deemed to contain meaningful variance. Dropping 14% of explained variance would also be a significant loss of data structure.

Table 5. Explained and cumulative variance and eigenvalues presented for each principal component.

Component	Explained Variance (%)	Cumulative (%)	Eigenvalues
1	19.8	19.8	2.61
2	12.4	32.2	1.64
3	9.58	41.8	1.26
4	8.68	50.5	1.14
5	7.85	58.3	1.03
6	7.36	65.7	0.970
7	6.57	72.3	0.865

In figure 9, the contributions of the variables to the principal components are visualized. Component 1 mainly explains correlation between temperature, terrain accessibility, Y coordinate, time of year and some correlation with X coordinate, road accessibility, mean stem volume and soil moisture. Component 2 likewise explains correlation between temperature, time of year, X and Y coordinates, mean stem and precipitation. Component 3 explains correlation in harvested volume, time of year, precipitation, mean stem volume and some correlation with harvesting type and road accessibility. Component 4 explains correlation in harvested volume, time of year, precipitation, mean stem volume, harvest type, soil moisture and some correlation in X coordinate and temperature. Component 5 explains correlation in road and terrain accessibility, X and Y coordinates, hauling distance and some correlation in harvested volume, soil moisture and temperature.

Component 6 is dominated by soil moisture which is slightly correlated with precipitation and harvested volume and hauling distance. The correlation with soil type was quite little in Component 6. Component 7 explains correlation mainly in mean stem and harvested volume and in X coordinate, hauling distance, road accessibility, harvest type and some correlation in soil moisture.

Given the dimensional reduction from 14 to seven dimensions resulting from the factor analysis with mixed data, all variables captured by the seven components to some extent. However, Soil types contributed very minorly to the components. This could suggest that using more components could have captured some missing variation. Although, the outputs also suggest the soil variable had very little correlation with the other input variables. Therefore, the Soil type data is a poor descriptor of patterns in this analysis.

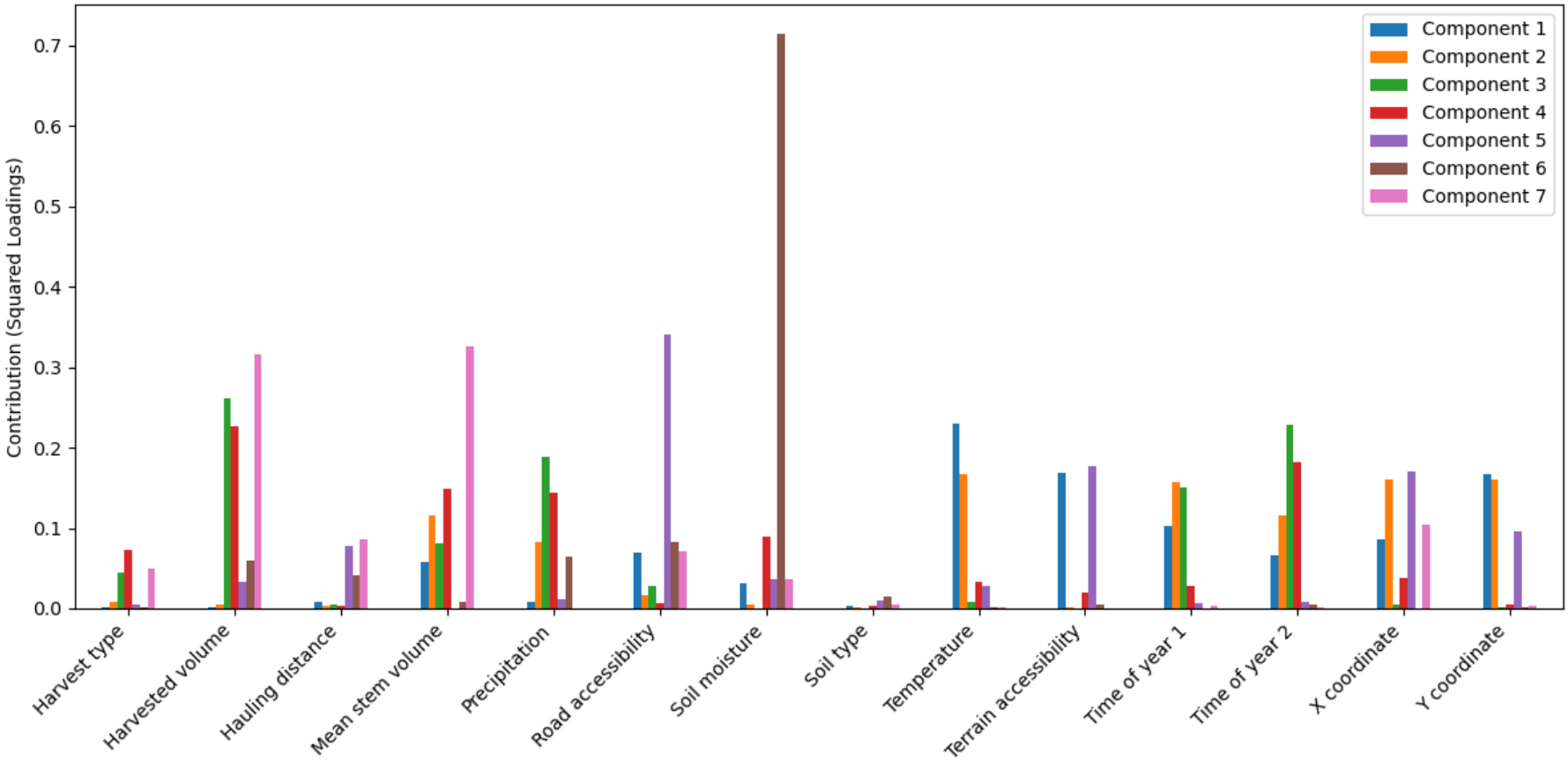


Figure 9. Contribution of each variable to the components visualized as squared loadings per component.

3.2.2 HDBSCAN clustering

The outputs of the factor analysis with mixed data were clustered using HDBSCAN, although the continuous nature of the outputs proved that density-based clustering could not produce meaningful results. Clusters were produced, although with a noise ratio of about 95% and very low cluster stability. For HDBSCAN results, see Appendix 2.

3.2.3 K-means clustering

To choose the appropriate numbers of clusters for k-means, silhouette scores, Calinski–Harabasz index and Davies–Bouldin index were computed for different numbers of clusters (Table 6). Nine clusters are deemed to be a good middle ground, as it has the next highest silhouette score, next highest Calinski-Harabasz index and the lowest Davies-Bouldin index. Overall, the silhouette scores are quite similar and positive but close to zero. This suggest the resulting cluster analysis results are moderate, with somewhat overlapping clusters.

Table 6. Computation of silhouette scores, Calinski–Harabasz index and Davies–Bouldin index for different numbers of clusters. The assessed optimal number of each column in bold text.

Clusters	Silhouette score	Calinski-Harabasz index	Davies-Bouldin index
8	0.1386	**5518**	1.703
9	0.1403	5488	1.**595**
10	0.1371	5219	1.602
11	**0.1437**	5089	1.614

The cluster overlap becomes clear when visualizing the component scores for different components (Fig. 10 & 11). The clusters are clearly grouped in their clusters yet somewhat overlapping. The dense nature of all the points suggests the original data is mostly continuously distributed.

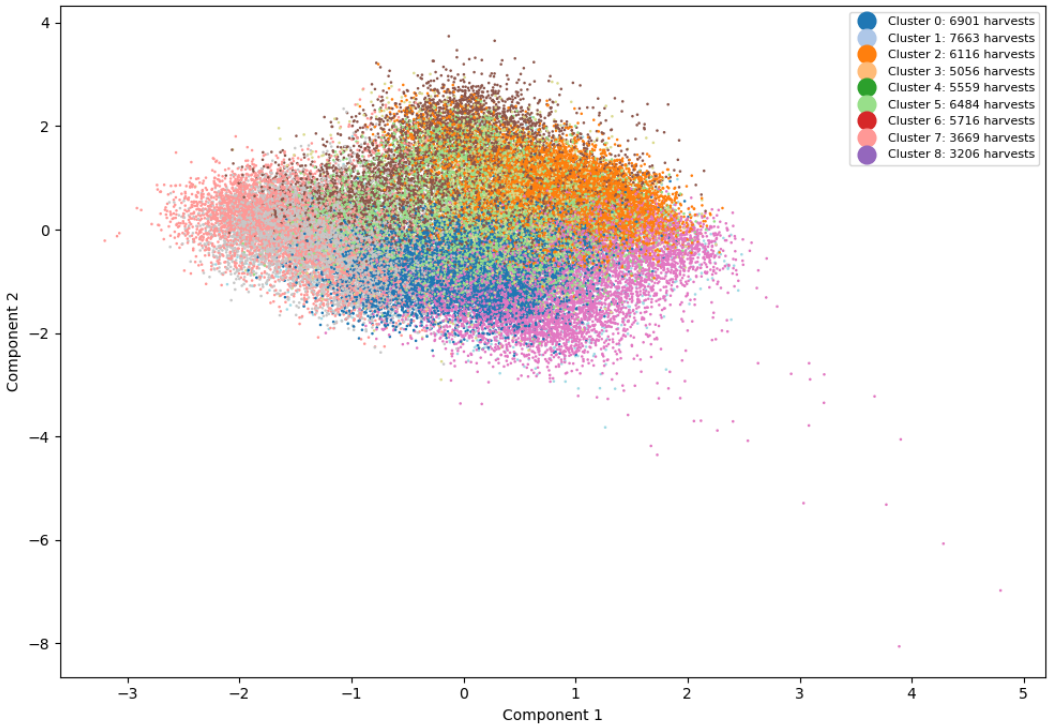


Figure 10. Each harvest site graphed based on Component 1 versus Component 2 scores. Each harvest site is colour-coded based on which cluster it was assigned.

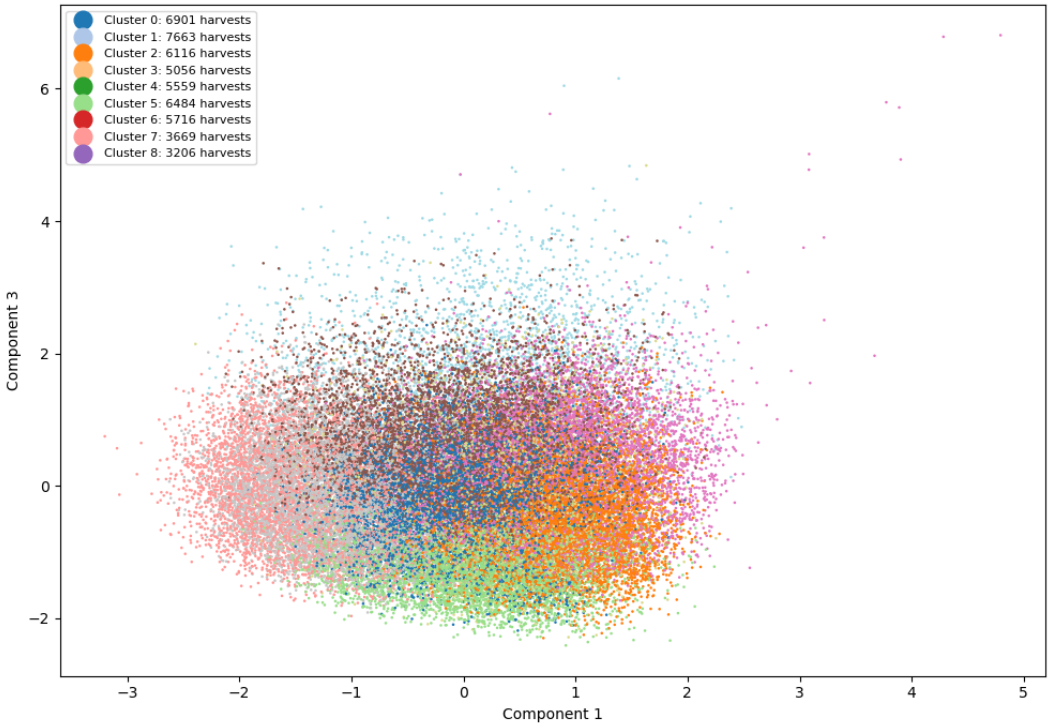


Figure 11. Each harvest site graphed based on component 1 versus component 3 scores. Each harvest site is colour-coded based on which cluster it was assigned.

The k-means clusters were visualized both spatially and per variable using a map over Sweden (fig. 12) and a heatmap (fig. 13). The heat map shows more unique values in darker colours. Some clusters were clearly more regionally defined while others were spatially dispersed yet defined by other variables such as temperature, time of year, mean stem volume etc.

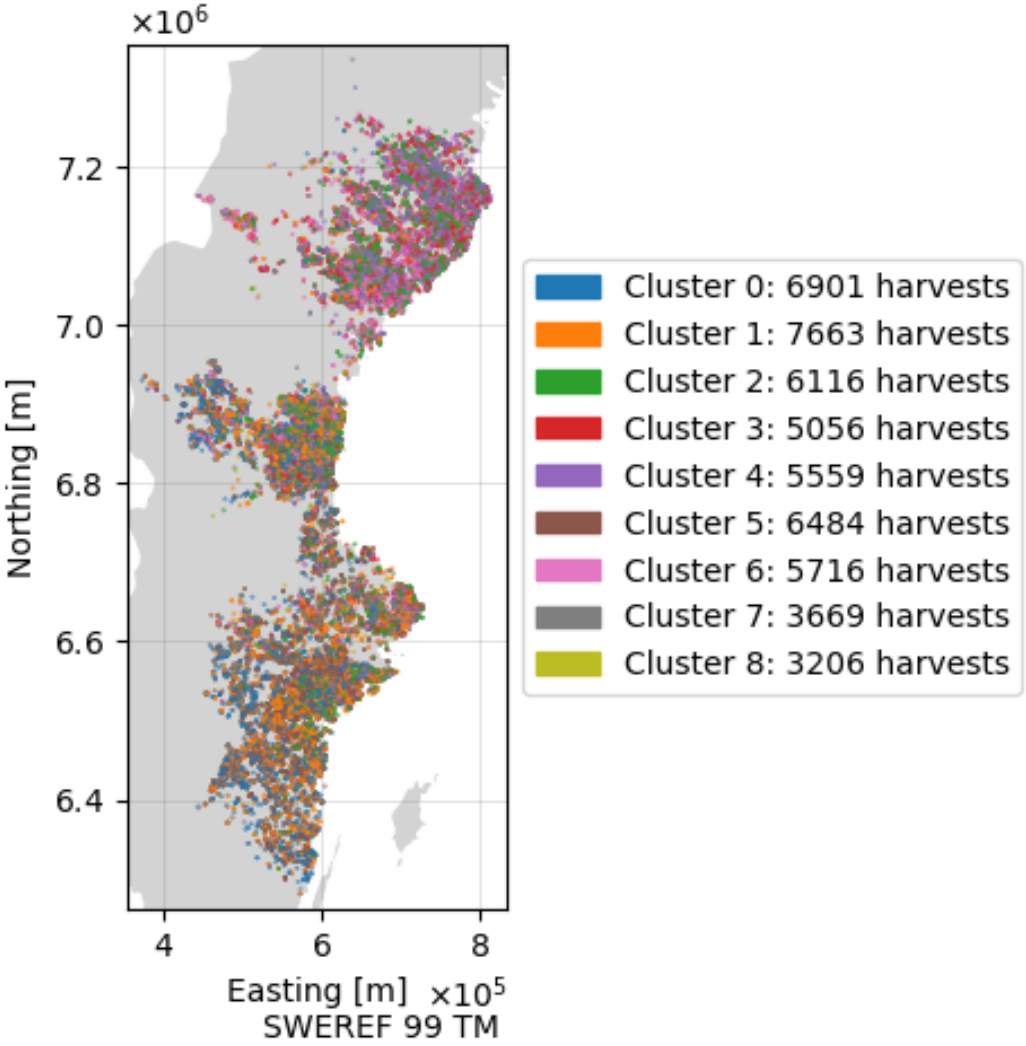


Figure 12. The k-means clusters spatially visualized. each cluster is colour-coded

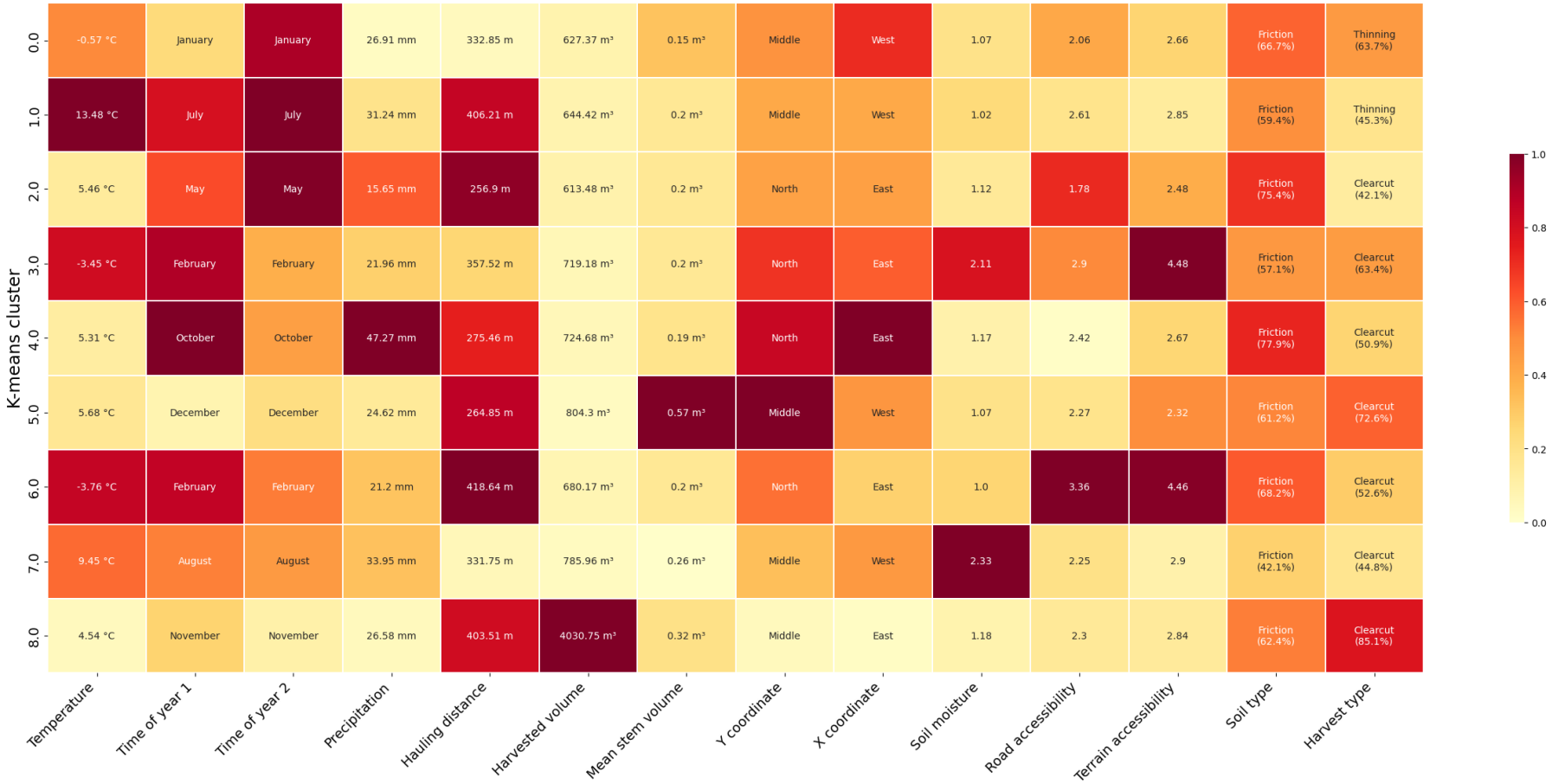


Figure 13. A heatmap for each k-means cluster where clusters are coloured in dark based on their mean values' departure from global mean for numerical variables and prevalence of the mode value relative to number of categories for categorical variables.

Given the characteristics of the clusters in Figure 12 and 13, the patterns in the data resulted in clusters with the following characteristics:

- Cluster 0 is a regional winter season cluster in mostly western Sweden with temperatures close to 0°C. The terrain and road accessibilities are relatively high in this cluster as well as the share of thinnings. The soils are generally dry.

- Cluster 1 is a warm summer season cluster in middle Sweden with longer hauling distances, relatively high amounts of precipitation, road and terrain accessibilities in the middle ranges and generally dry soils.

- Cluster 2 is spatially a rather dispersed thawing season cluster with the least precipitations, shortest hauling distances and highest road accessibilities of all clusters. The soils consist of unusually much friction soils and are generally dry. The terrains are the second most accessible of all clusters.

- Cluster 3 is a regional northeastern and winter season cluster. The temperatures are unusually cold, the soils are unusually wet, and the terrain accessibilities are low. The road accessibilities are the lowest of all clusters.

- Cluster 4 is a regional northeastern cluster in the autumn season. This cluster has the most precipitations of all clusters, very short hauling distances, generally dry friction soils and higher terrain and road accessibilities.

- Cluster 5 is a regional warm winter season cluster of harvesting sites in middle western Sweden. The hauling distances are very short, the accessibility classes are generally high, and the soils are mostly dry friction soils. The mean stem of these harvesting sites are the highest by far. The share of clear-cuts is high while the harvested volumes are somewhat high.

- Cluster 6 is very similar to cluster 3 as it is a regional winter season cluster in northeastern Sweden. This cluster have the coldest temperatures, longest hauling distances, least terrain accessibility but also the driest soils. The road accessibilities are also very low in this cluster.

- Cluster 7 is a spatially dispersed summer-autumn season cluster. The amounts of precipitation are high, and the road accessibilities are somewhat high, relatively while the terrain accessibilities are in the middle-range. The soil moistures are the wettest while the share of friction soils is the lowest of all clusters.

- Cluster 8 is a spatially dispersed cluster of all the large harvesting sites. Both the share of clear-cuts and harvested volumes are the highest of all clusters. The mean stem volumes and hauling distances are also relatively high.

4. Discussion

4.1.1 Spatiotemporal cluster analysis

The results indicate harvesting operations are moderate to strongly clustered in space and time. This finding supports the economical aspects of harvesting in clusters discussed by Başkent and Jordan (1991). The results also visualize the clusters, which are mostly distributed evenly across the counties. There are also some spatially local harvesting hotspots identified by the analysis. These local hotspots could be random but could also be patterns of forest management decisions such as clearing of stands damaged by local storms or pests.

Harvest resources are commonly specialized on some specific type of harvest due to specialization of harvesting equipment (Frisk et al. 2016). Due to the data in the spatiotemporal cluster analysis only including regeneration harvests, the results may mostly be patterns of only a share of the harvesting resources operating in the analysed areas. The analysis most likely missed many spatiotemporal patterns due to unavailability of data for extensive periods and harvesting operations.

Future studies could, if data is available, make more complete cluster analyses of harvesting operations to better explore the extent and dispersion of harvesting operations. Other clustering algorithms could also be used as the results of this study clearly were affected by the choice of parameters. Other algorithms, adapted to dealing with the difficulty of finding meaningful threshold values, could cluster harvesting operations in a more meaningful way for different regions and areas. The algorithms used in this study was used due to their accessibility and relative ease of application. More advanced researchers can surely utilize and/or adapt other algorithms successfully in similar settings.

4.1.2 Cluster analysis with forestry-related variables

The cluster analysis of the Harvest database indicate forestry-related variables can successfully be used to cluster harvesting operations. The main limitations to using forestry-related variables are the availability of accurate data and format requirements of clustering algorithms. Further studies could utilize new data and improved algorithms to test more variables in robust analyses to better understand how patterns in forestry operations can be identified.

This study could not produce meaningful results through factor analysis of mixed data followed by HDBSCAN on the output components. The reason for this is most likely the continuous nature of the data used. Further studies with access to more computational power could however test density-based clustering algorithms further by skipping the factor analysis of mixed data and directly use the harvest data in the cluster analysis. Algorithms as HDBSCAN does not require dimensional reduction to find meaningful clusters but higher dimensionality does require more computational power (McInnes, Healy & Astels 2017).

The implications of poor data quality, uncertainties and potential errors in the data led to exclusion of several relevant forestry-related variables and historical harvesting sites from the analysis. It may also have caused several historical harvest sites to be paired with faulty spatial data from other sources in the data treatment. Another potential source of error regarding the weather data is the resampling warp from transforming the CRS of the weather data rasters. This potentially misplaced some historical harvesting sites close to the original raster borders and thus led to faulty weather data (Dorman et al. 2025). The original data has a low resolution of 4x4 km however and can still provide insight into general weather conditions for when each harvest took place. All these implications means that the result of this study only identifies general patterns and phenomena that can vary if the same methods are applied to similar data.

4.1.3 Patterns in harvesting operations produced by cluster analysis

Cluster analysis groups data based on similarities and differences through parameters and algorithms. This study produced clusters of good quality with the ST-DBSCAN algorithm. These clusters show harvesting operations are oftentimes close in time and space. This study also produced clusters of moderate quality through the K-means algorithm. The K-means clusters are groups of historical harvesting sites of similar properties that can be mainly spatial, temporal, weather-based or based on any of the other variables used. In all clusters, time of year correlates with the temperature data as temperatures are higher during summer seasons and vice versa.

The road accessibilities also share a pattern with the temperature and time of year as the lowest temperatures during winter season also had the least road accessibility, the summer season had somewhat more road accessibility while autumn/spring season had better accessibility. The highest road accessibility was found in spring. These findings correlate with the forestry road classes described by the Swedish Forestry Research institute (2014) where the highest accessibility is required during thawing season and rainy periods while the least accessibility is required during winter season when the soil is frozen. Interestingly, the results show hauling distances are generally shorter for periods of less accessibility as well. This indicates harvesting sites are closer to roads during periods of less accessibility.

The terrain accessibility correlated in a similar pattern with time of year and temperature as road accessibility. The coldest clusters also had the least terrain accessibility while the thawing season cluster have the highest. These findings are conclusive with how terrain accessibility is described in Berg's Terrain Classification System for Forestry Work (1995).

Patterns in mean stem volumes, harvested volumes and share of clear-cuts were also captured by the cluster analysis. These variables share a positive correlation.

There are clearly more patterns in the data stemming from similarities in the data. Although, the impact of errors in the data potentially contribute to deficiencies in the results. Further studies could utilize data of higher quality and potentially of a greater extent than only one private forestry company to better capture spatial dispersion of harvesting operations on a landscape level.

5. Conclusion

This study utilized cluster analysis on data of historical harvesting operations, weather and soil to identify clusters and patterns in forestry operations. The collected data was used to generate a historical harvest database. The collected data and the database were then applied to different cluster analysis algorithms.

Harvesting operations can be clustered spatiotemporally into clusters through unsupervised cluster analysis. The analysed areas and periods were moderately to strongly clustered with mostly an even dispersion of harvesting operations. There were also some local spatial hotspots identified where harvesting operations were spatially dense.

Forestry-related variables can be used in cluster analysis of harvesting operations to group data based on similarities and differences. The main limitations to this method are the availability of usable data and relevant algorithms for cluster analysis. No open-access database provides a full and accurate picture of forestry operations and cluster algorithms are oftentimes designed for different settings and data types than the study aims at.

Patterns were identified using cluster analysis. Harvesting operations are moderately to strongly clustered in space and time when clustering with forestry-relevant parameters. Data on weather conditions and time of year during harvesting generally aligns with requirements on road and terrain accessibility. Hauling distances are also shorter for higher accessibility classes. Volumes of harvested volumes and mean stems and share of clear-cuts also share a positive correlation.

This study provides some valuable and important real-world applications and implications for future research. Firstly, this study can be used to better understand distribution and patterns of harvesting operations. It can also be used to inform further research of machine learning in forestry. This study also identified the problem of a lack of accurate and wide scale data for conducting cluster analyses. To further research spatiotemporal distribution of forestry operations and meet the demand for knowledge in shifting to renewable energy sources, more and better data is required. Policymakers should ensure data availability by centralizing and standardizing data collection from practitioners in forestry. This data can then be analysed to better understand and mitigate the present challenges in forestry.

References

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics*, 9(8), 1295. https://doi.org/10.3390/electronics9081295

Başkent, E. & Jordan, G. (1991). Spatial wood supply simulation modelling. *The Forestry Chronicle*, 67, 610-621. 10.5558/tfc67610-6.

Berg, S. (1992). *Terrain classification system for forestry work*. Forskningsstiftelsen Skogsarbeten

Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 60(1), 208-221. https://doi.org/10.1016/j.datak.2006.01.013

Braeken, J., & van Assen, M. A. L. M. (2017). An empirical Kaiser criterion..*Psychological Methods*, 22(3), 450–466. https://doi.org/10.1037/met0000074

Cakmak, E., Plank, M., Calovi, D. S., Jordan, A., & Keim, D. (2021). Spatio-temporal clustering benchmark for collective animal behavior. 10.1145/3486637.3489487.

Chong, B. (2021). K-means clustering algorithm: a brief review. *Academic Journal of Computing & Information Science*, 4(5), 37-40.

Dorman, M., Graser, A., Nowosad, J., & Lovelace, R. (2025). *Geocomputation with Python (1st ed.)*. Chapman and Hall/CRC. https://doi.org/10.1201/9781003379911

Eliasson, L. (2024) *Bränsleförbrukning i drivning*. Swedish Forestry Research institute, Kunskapsbanken. https://www.skogforsk.se/kunskapsbanken/kunskapsartiklar/2024/bransleforbrukning-i-drivning-2023/ [2025-08-24]

Englund, G., Eggers, J., Jonsson, BG., Schulte, M. & Skytt, T. (2025). Why We Disagree about the Climate Impact of Forestry – A Quantitative Analysis of Swedish Research. *Environmental Management*. 75, 1923–1937. https://doi.org/10.1007/s00267-025-02208-z

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd, 96(34), 226-231.

Fjeld, D. & Dahlin, B. (2008). *Nordic logistics handbook: forest operations in wood supply (V 17-11-20)*. Swedish University of Agricultural Sciences, University of Helsinki

Frisk, M., Flisberg, P., Rönnqvist, M., & Andersson, G. (2016). Detailed scheduling of harvest teams and robust use of harvest and transportation resources. *Scandinavian Journal of Forest Research*, 31(7), 681–690. https://doi.org/10.1080/02827581.2016.1206144

International Panel on Climate Change (2023). *Climate Change 2023*. International Panel of Climate Change.

https://www.ipcc.ch/report/ar6/syr/downloads/report/IPCC_AR6_SYR_FullVolume.pdf

Knox, E. G. (1964). The Detection of Space-Time Interactions. *Journal of The Royal Statistical Society Series C-applied Statistics*, 13(1), 25–29. https://doi.org/10.2307/2985220

Lloyd, S.P. (1982). Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28, 129-136.

McInnes, L., Healy, J. & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, *The Open Journal*, 2(11).

Nimbnet (n.d.). *Hållbar Omställning: Sveriges Skogsindustris Pionjärerer vill göra Elektrifiering av Vägtransporter möjlig*. https://nimbnet.com/sv/nyheter/hallbar-omstallning-sveriges-skogsindustris-pionjarerer-vill-gora-elektrifiering-av-vagtransporter-mojlig/ [2025-11-20]

Pagès, J. (2014). *Multiple Factor Analysis by Example Using R (1st ed.)*. Chapman and Hall/CRC. https://doi.org/10.1201/b17700

Raha, Z. S., Khandaker, M. A. A., Hossain, M. Z., & Akhter, S. (2025). Navigating High-Dimensional Data with Advanced Clustering Algorithms. *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. 1-6.

Rowell, A. (2022). *Suitable tract bank size: exploration & estimation. Swedish University of Agricultural Sciences*. Department of Forest Biomaterials and Technology. https://stud.epsilon.slu.se/18564/

United Nations (2015). *Resolution 70/1. Transforming our world: the 2030 Agenda for Sustainable Development.* United nations.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12, 2825-2830.

Swedish Geological Survey (2024). *Jordarter 1:25 000–1:100 000*. Jordartsdata. https://www.SGS.se/produkter-och-tjanster/geologiska-data/jordarter--geologiska-data/jordartsdata/ [2025-08-20]

Sharifnia, A. M., Kpormegbey, D. E., Thapa, D. K., & Cleary, M. (2026). A Primer of Data Cleaning in Quantitative Research: Handling Missing Values and Outliers. *Journal of advanced nursing*, 82(1), 970–975. https://doi.org/10.1111/jan.16908

Silva-Souza, K. J. P., Pivato, M. G., Silva, V. C., Haidar, R. F., & Souza, A. F. (2022). New patterns of the tree beta diversity and its determinants in the largest savanna and wetland biomes of South America. *Plant diversity*, 45(4), 369–384. https://doi.org/10.1016/j.pld.2022.09.006

SLU (2020a). *SLU Soil moisture map*. *Department of Forest Ecology and Management*, Swedish University of Agricultural Sciences.

SLU (2020b). *Dokumentation nya hydrografiska kartor – vattendrag och SLU Markfuktighetskartor*. Swedish University of Agricultural Sciences.

SMHI (2025). *Griddad nederbörd- och temperaturdata (PTHBV).* Swedish Meteorological and Hydrological Institute. https://www.smhi.se/data/nederbord-och-fuktighet/nederbord/griddad-nederbord--och-temperaturdata [2025-09-01]

Swedish Environmental Protection Agency (2025). *Arbetsmaskiner, utsläpp av växthusgaser*. https://www.naturvardsverket.se/data-och-statistik/klimat/vaxthusgaser-utslapp-fran-arbetsmaskiner/ [2025-12-01]

Swedish Forest Agency (2025a). *Skogliga åtgärder du behöver anmäla eller ansöka om*. https://www.skogsstyrelsen.se/lag-och-tillsyn/atgarder-du-behover-anmala-eller-ansoka-om/ [2025-12-01]

Swedish Forest Agency (2025b). *Utförda avverkningar*. Skogliga grunddata. https://www.skogsstyrelsen.se/e-tjanster-och-kartor/karttjanster/geodatatjanster/ladda-ner-geodata/ [2025-09-01]

Swedish Research Institute (2014). *Skogsbilvägar - service, underhåll och upprustning*. Gävle offset. ISBN: 978-91-979694-4-4

Ufeli, C. P., Sattar, M. U., Hasan, R. & Mahmood, S. Enhancing Customer Segmentation Through Factor Analysis of Mixed Data (FAMD)-Based Approach Using K-Means and Hierarchical Clustering Algorithms. *Information 2025*, 16, 441. https://doi.org/10.3390/info16060441

Ur Rehman, S. (2014). DBSCAN: Past, present and future. *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*.

Wang, X. & Xu, Y.S. (2019). An Improved Index for Clustering Validation Based on Silhouette Index and Calinski-Harabasz Index. *IOP Conference Series: Materials Science and Engineering*, 569. https://doi.org/10.1088/1757-899x/569/5/052024

Appendix 1: All figures from the ST-DBSCAN cluster analysis

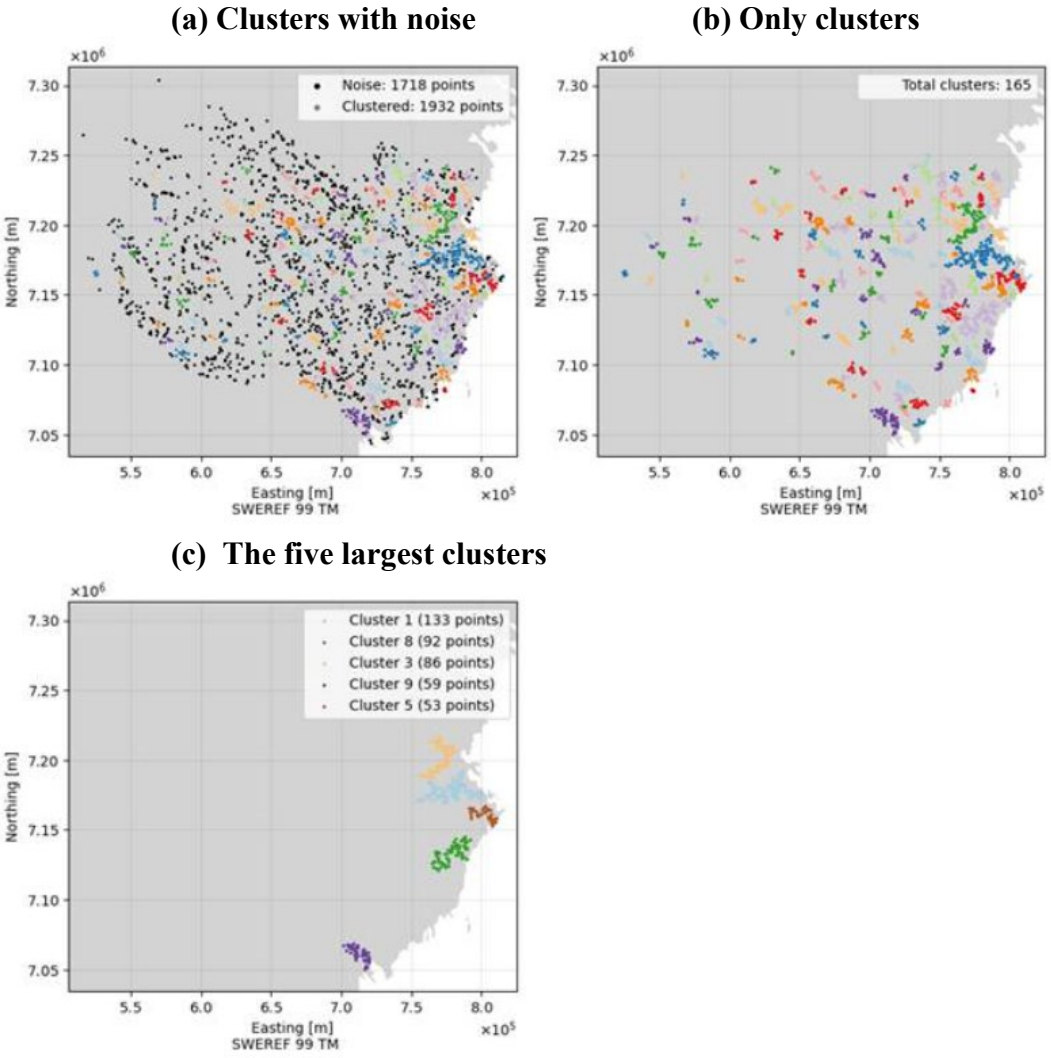


Figure 14. ST-DBSCAN clusters identified in Västerbotten County 2023. Noise is represented as black points while clusters are similar-coloured points close to each other.

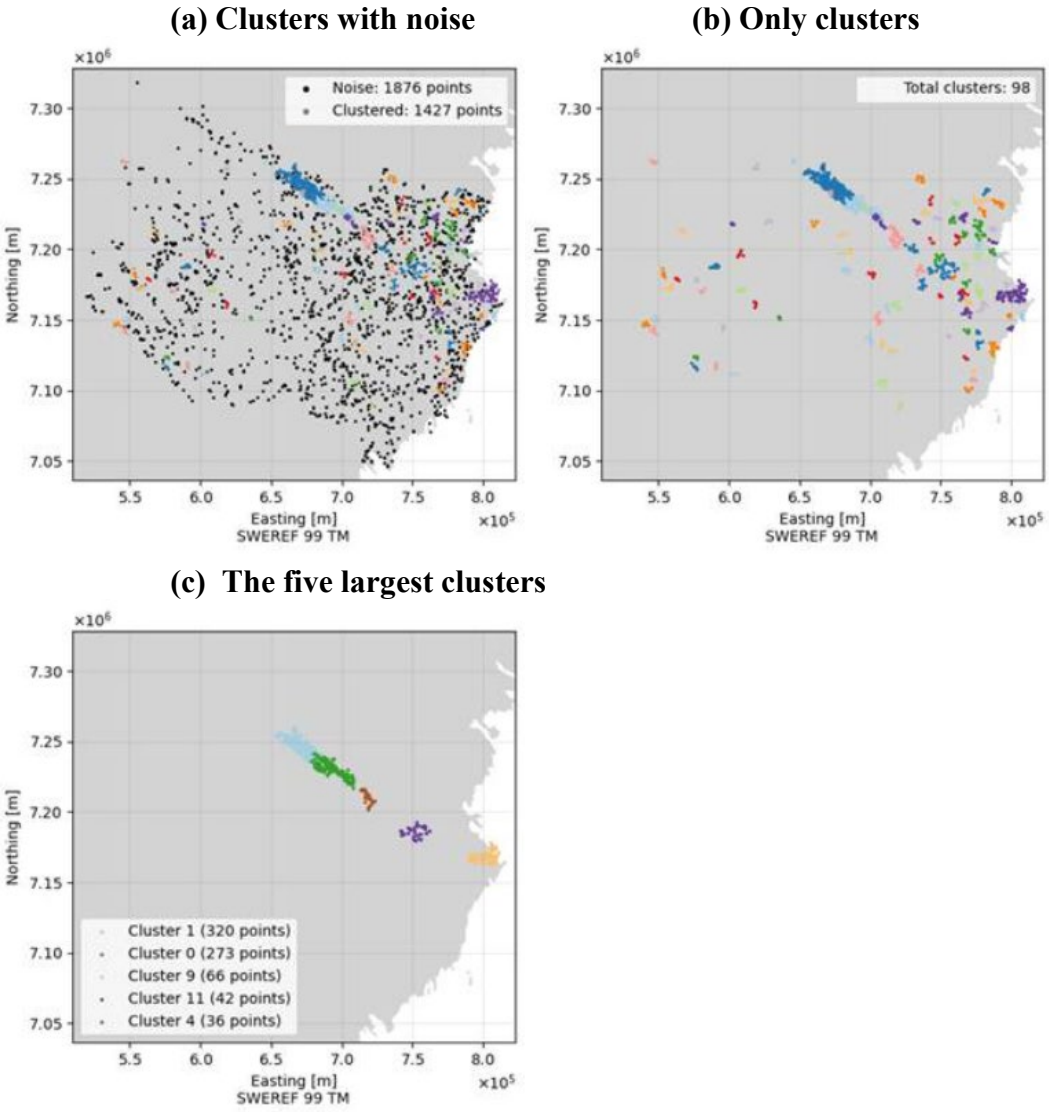


Figure 15. ST-DBSCAN clusters identified in Västerbotten County 2024. Noise is represented as black points while clusters are similar-coloured points close to each other.

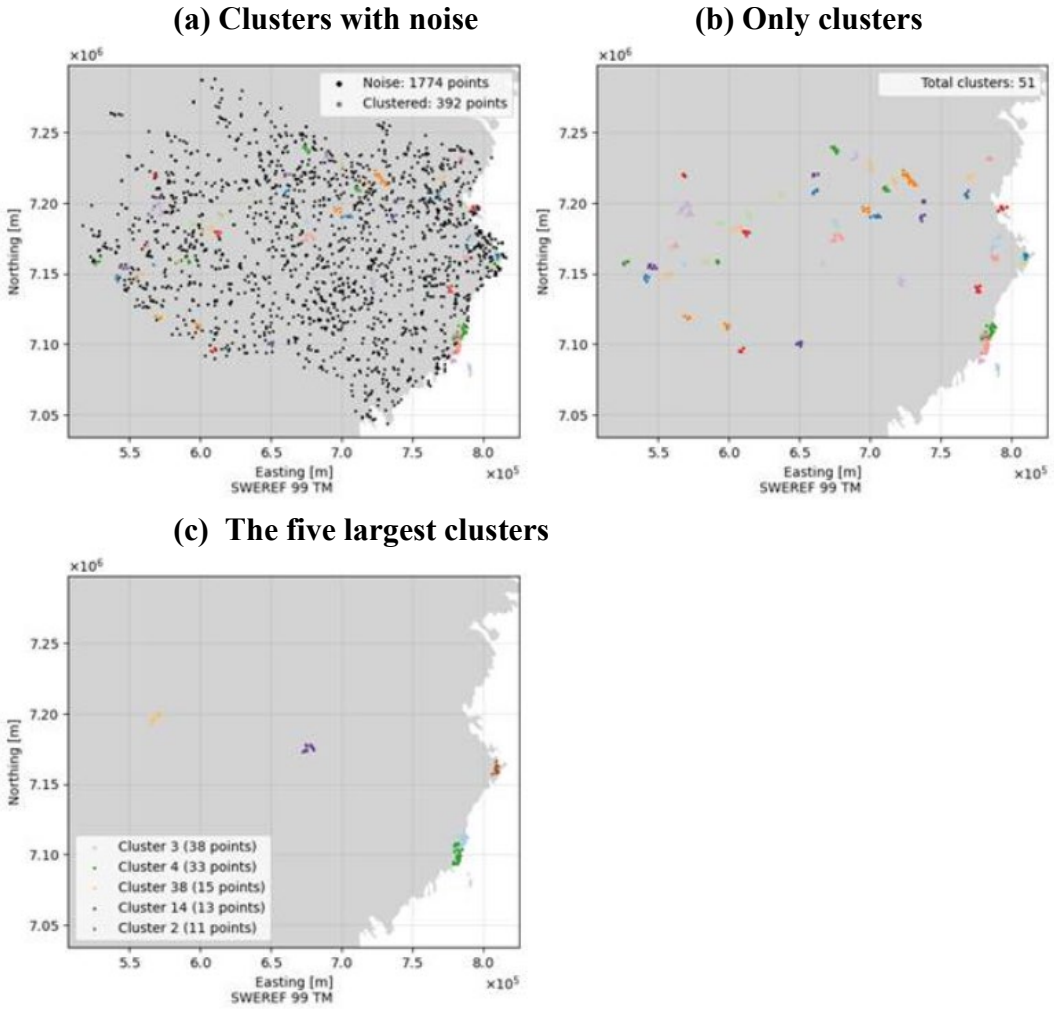


Figure 16. ST-DBSCAN clusters identified in Västerbotten County 2025. Noise is represented as black points while clusters are similar-coloured points close to each other.

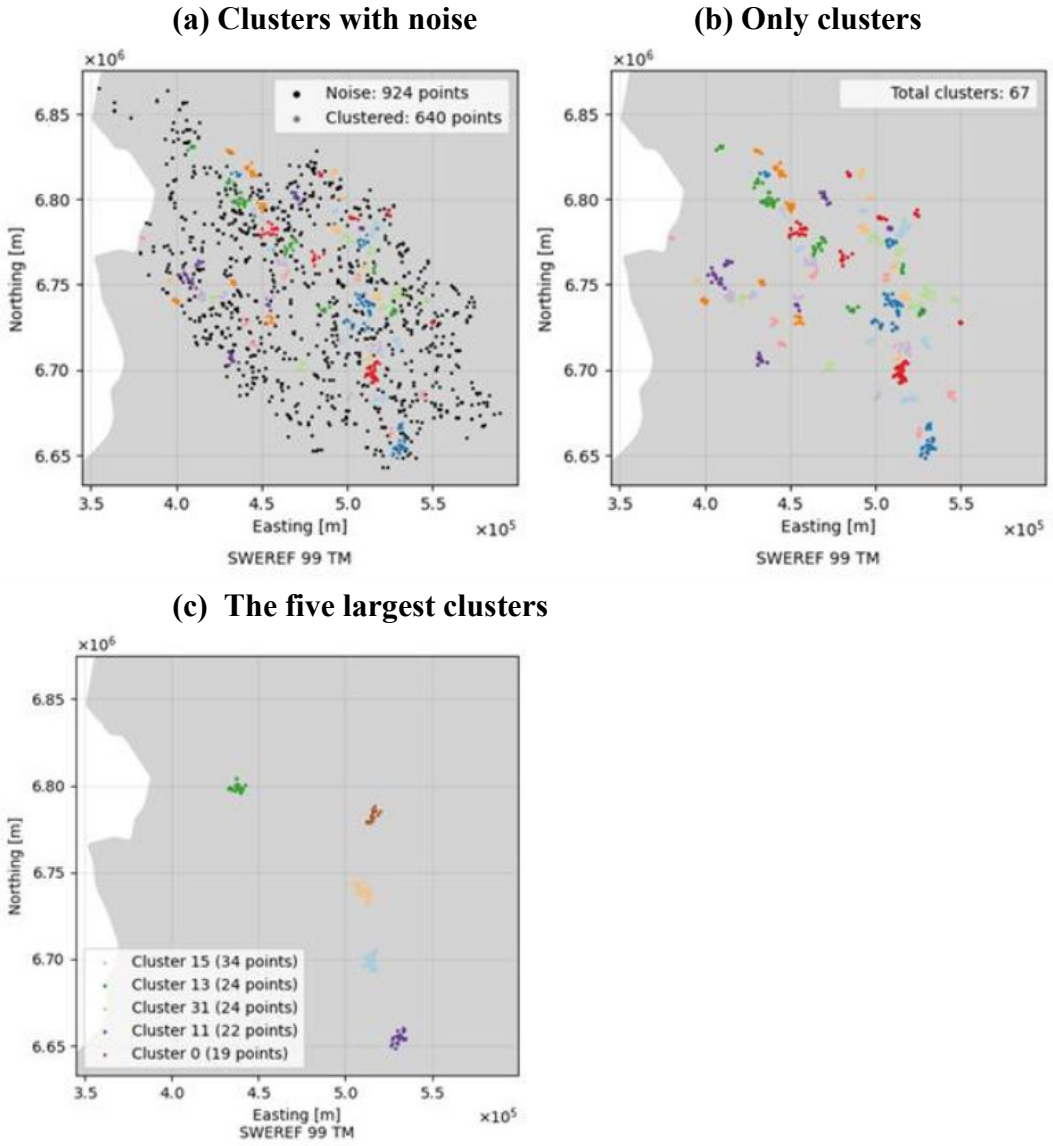


Figure 17. ST-DBSCAN clusters identified in Dalarna County 2023. Noise is represented as black points while clusters are similar-coloured points close to each other.

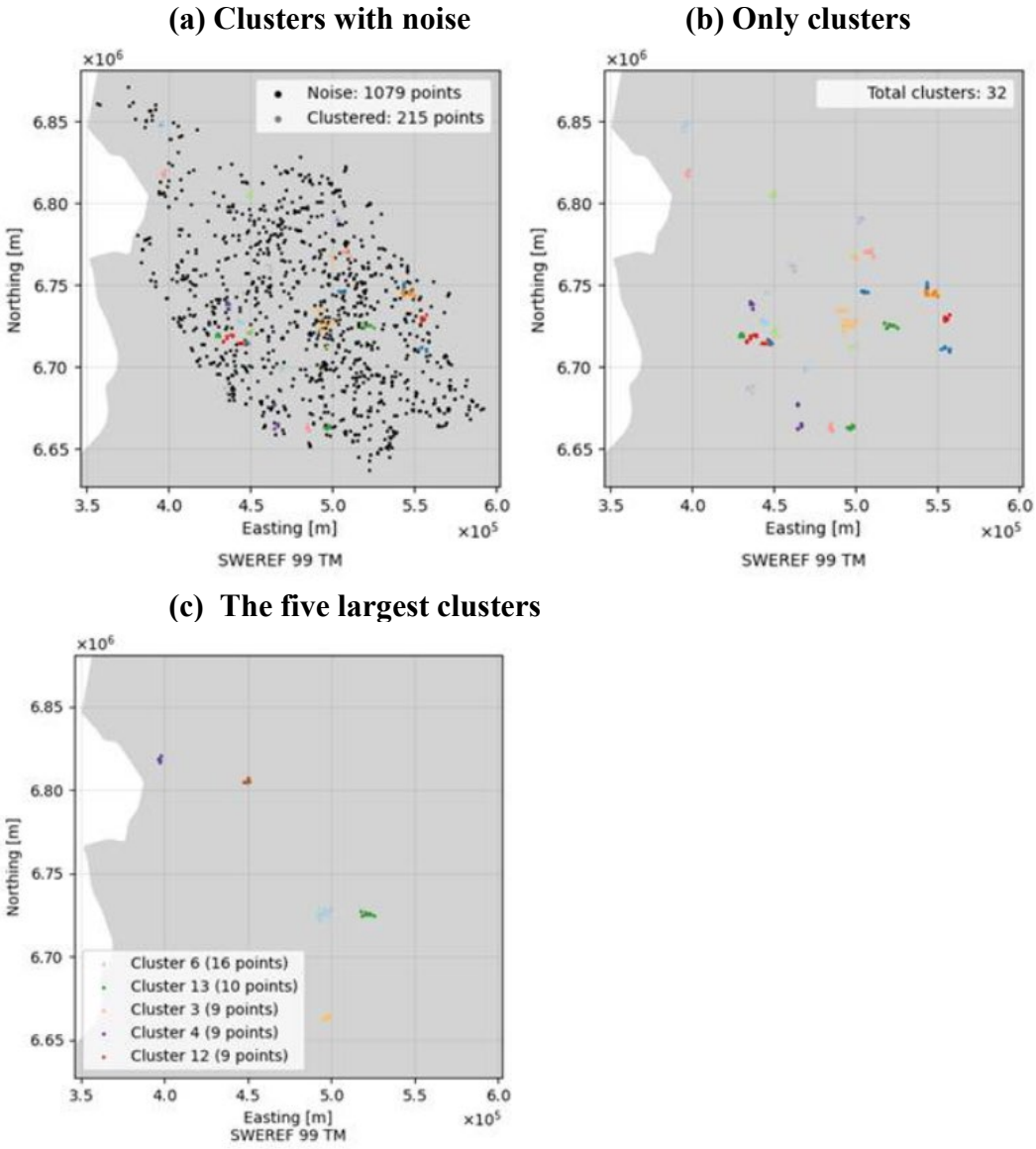


Figure 18. ST-DBSCAN clusters identified in Dalarna County 2024. Noise is represented as black points while clusters are similar-coloured points close to each other.

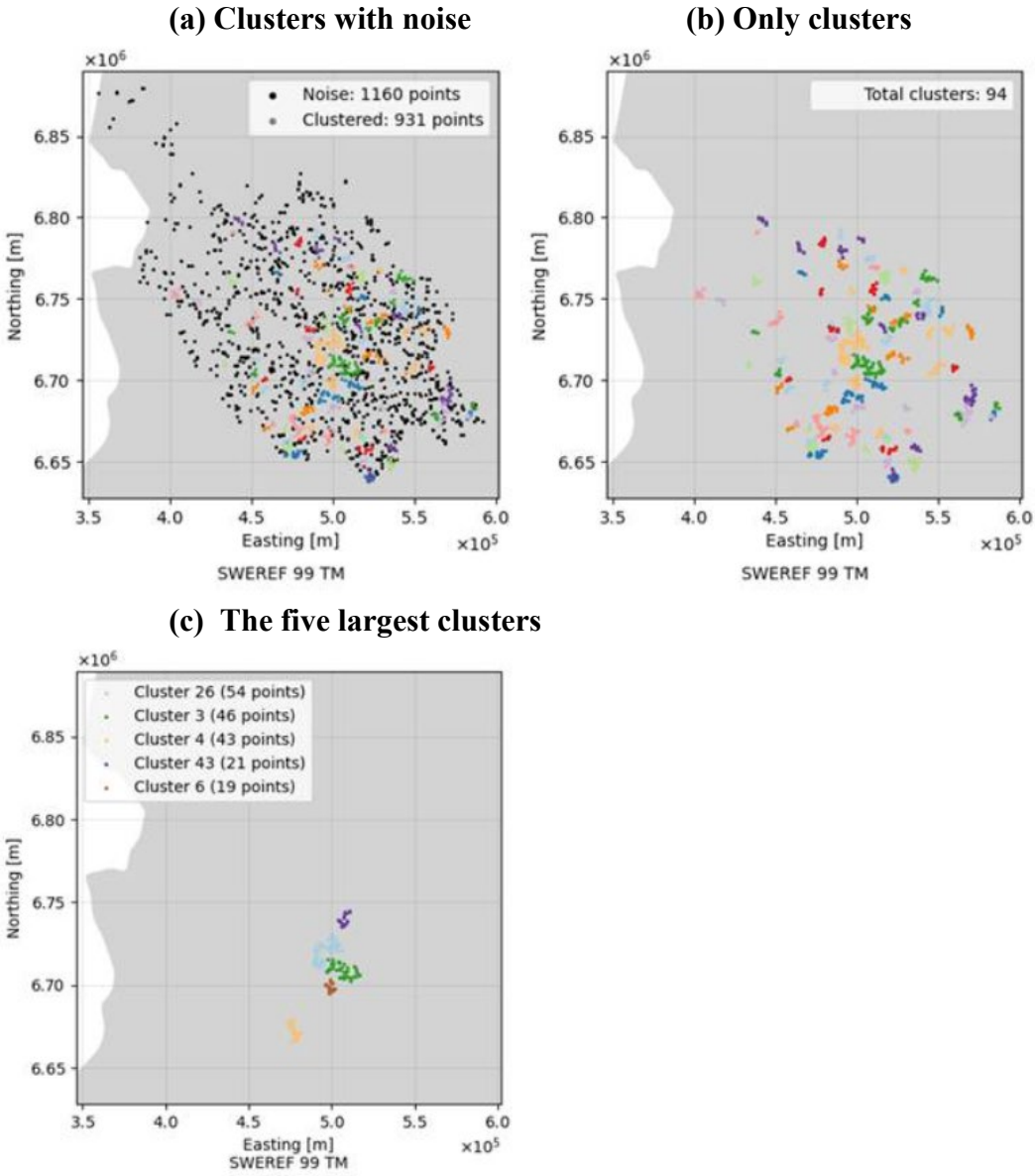


Figure 19. ST-DBSCAN clusters identified in Dalarna County 2025. Noise is represented as black points while clusters are similar-coloured points close to each other.

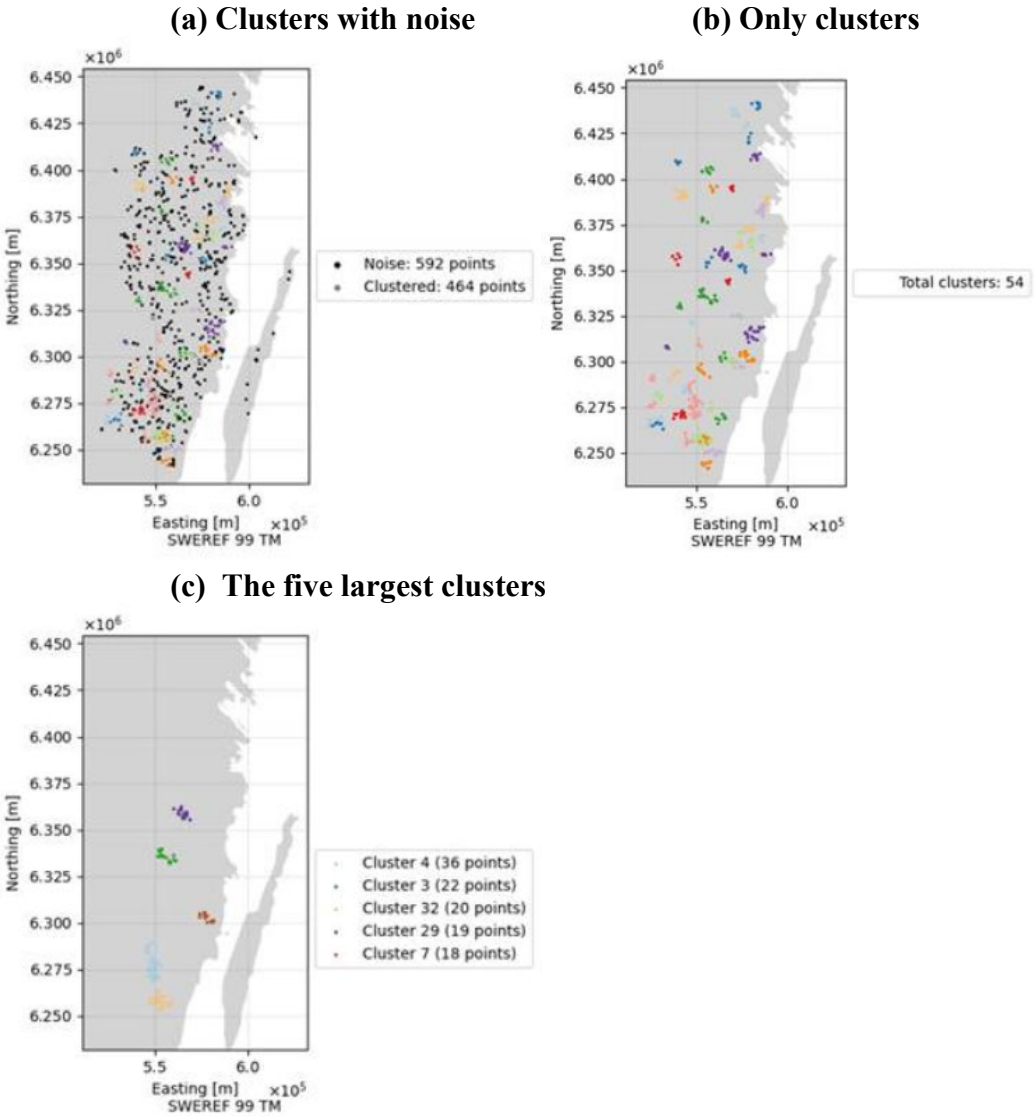


Figure 20. ST-DBSCAN clusters identified in Kalmar County 2023. Noise is represented as black points while clusters are similar-coloured points close to each other.

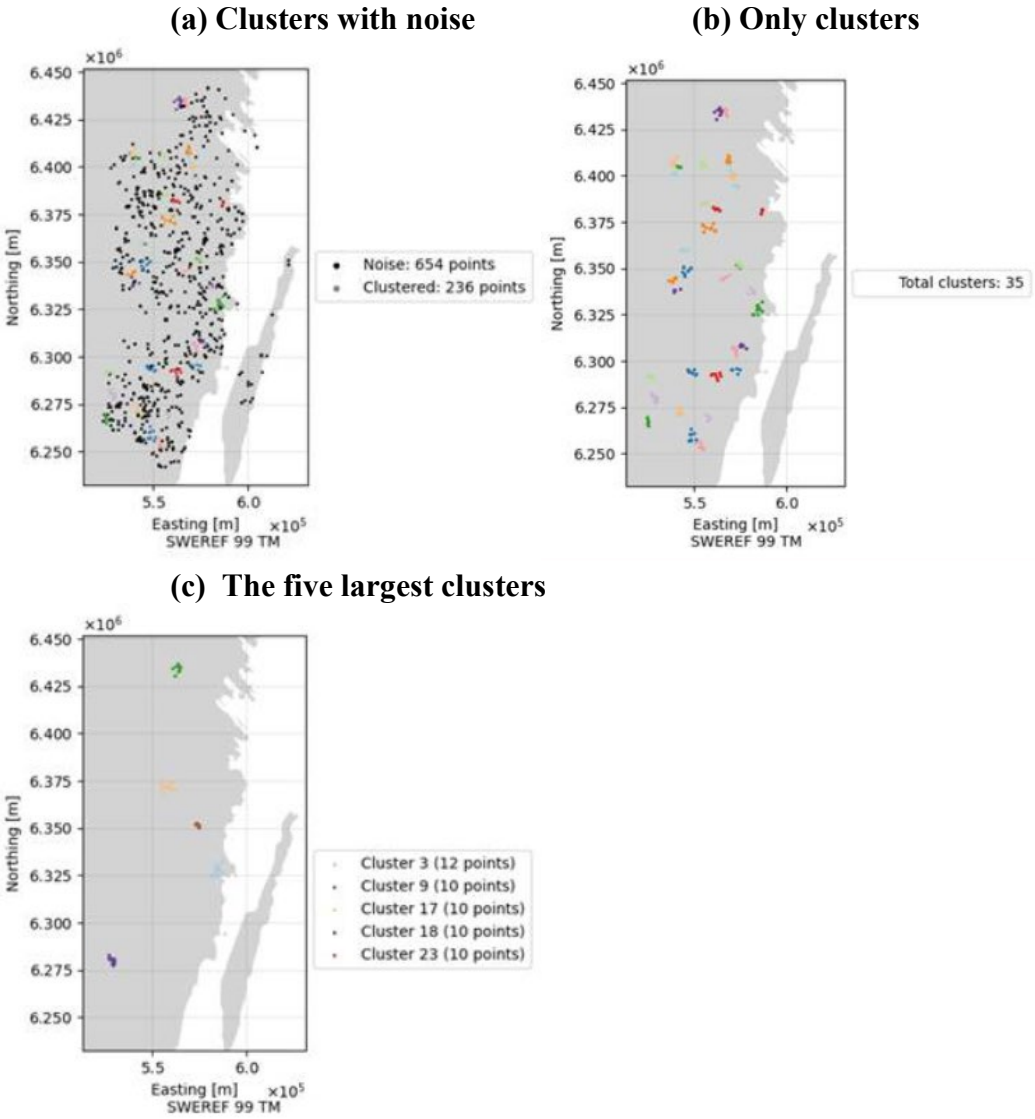


Figure 21. ST-DBSCAN clusters identified in Kalmar County 2024. Noise is represented as black points while clusters are similar-coloured points close to each other.

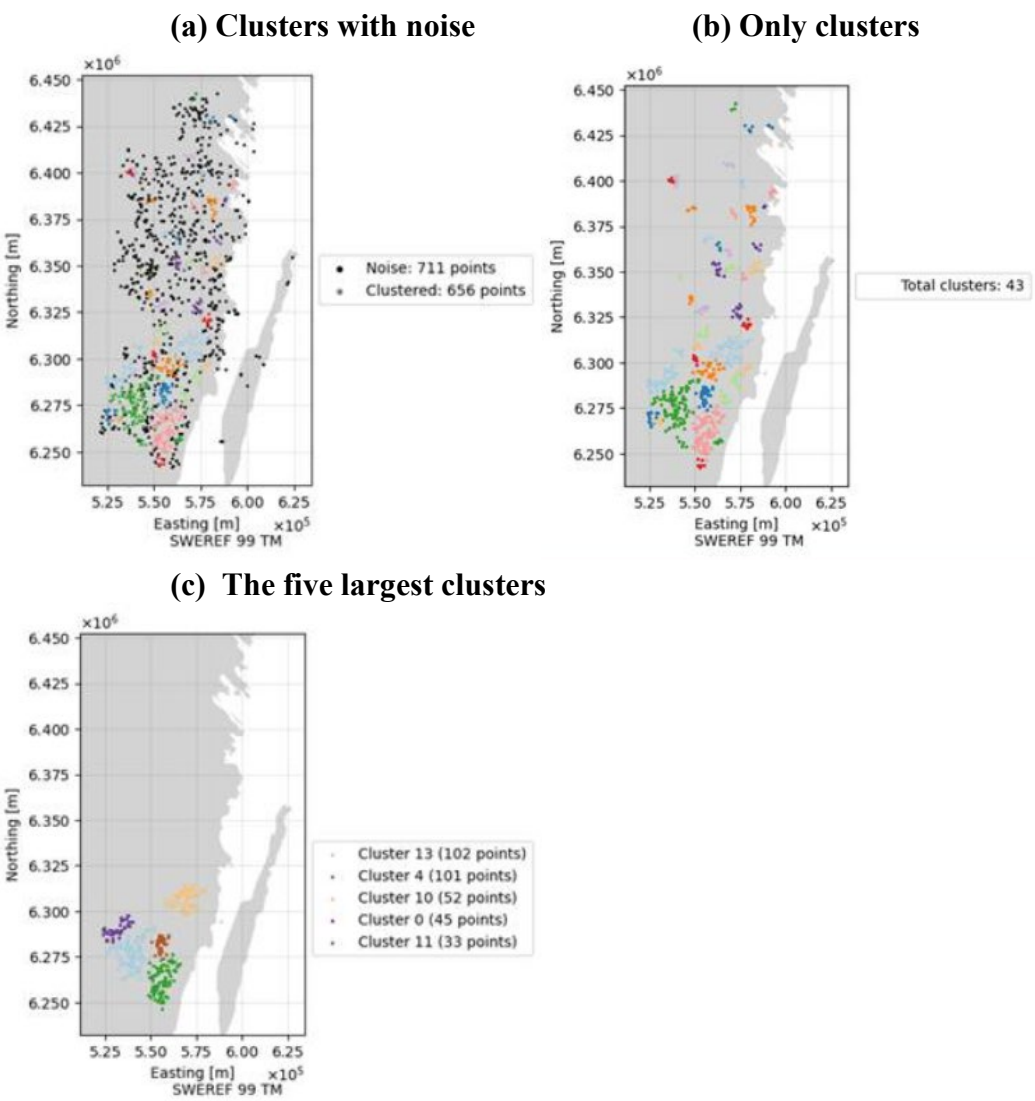


Figure 22. ST-DBSCAN clusters identified in Kalmar County 2025. Noise is represented as black points while clusters are similar-coloured points close to each other.

Appendix 2: HDBSCAN results

The Silhouette score of the HDBSCAN cluster analysis indicate the results are less overlapping than the K-means clusters (Table 7). This is also visually apparent in Figure 23 and 24. Yet, the clustering only resulted in a small share of the harvesting sites being grouped in clusters. The Noise ratio indicate these results are not meaningful (Table 7).

Table 7. Computation of silhouette score, Calinski–Harabasz index and Davies–Bouldin index and Noise ratio of the HDBSCAN clustering results.

Metric	Result
Silhouette score	0.3355
Calinski–Harabasz Index	907.5
Davies–Bouldin Index	1.163
Noise ratio	0.9600

The continuous nature of the data in the Harvest database becomes clear when visualized based on component scores (Fig. 23 & 24). The density-based clustering of HDBSCAN did not produce meaningful results given the continuous nature of the data.

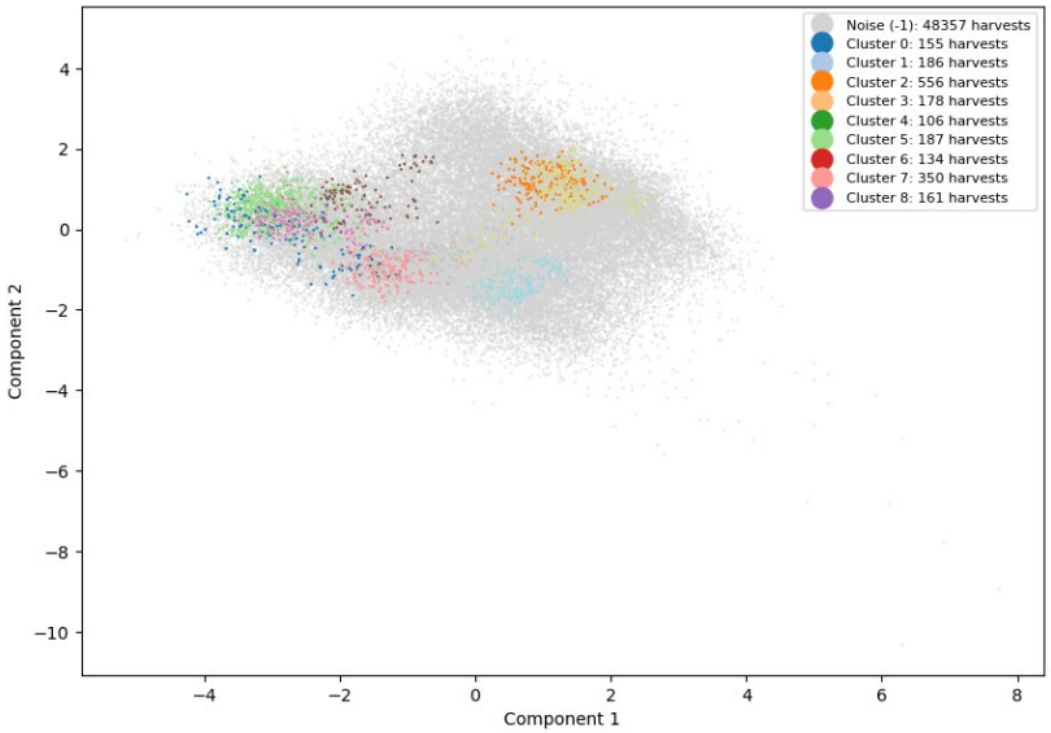


Figure 23. Each harvest site graphed based on Component 1 versus Component 2 scores. Each harvest site is colour-coded based on which cluster it was assigned.

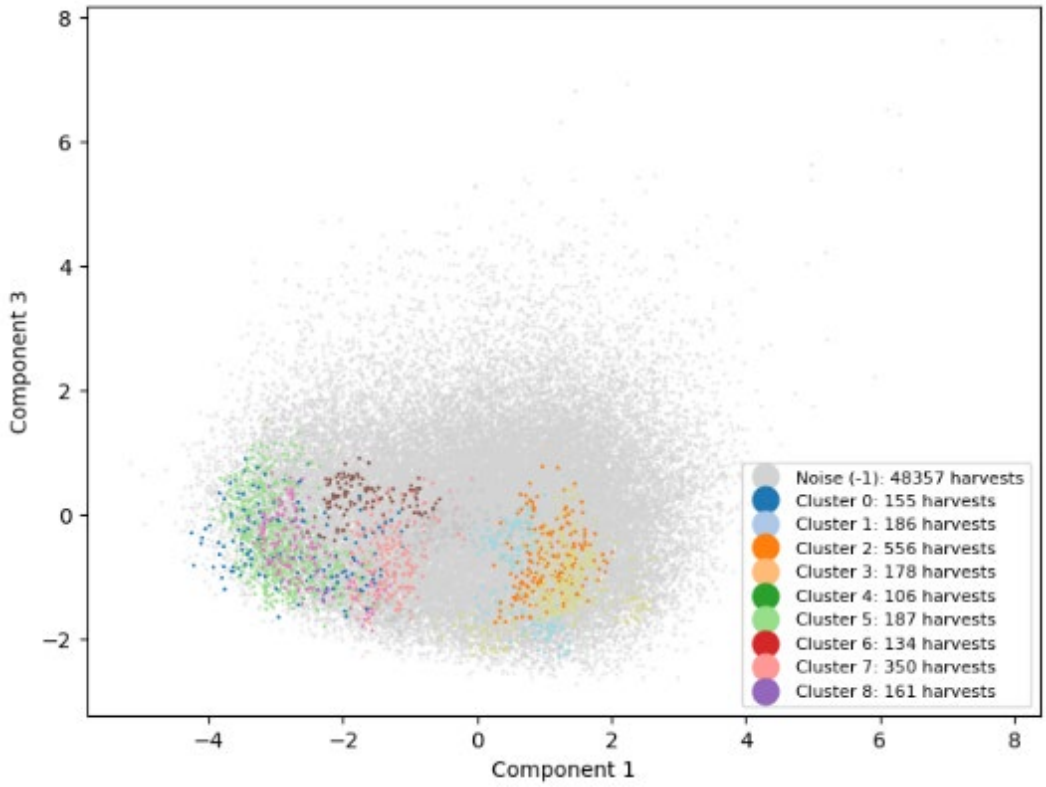


Figure 24. Each harvest site graphed based on Component 1 versus Component 3 scores. Each harvest site is colour-coded based on which cluster it was assigned.

The HDBSCAN clusters are spatially regional and not dispersed across the full spatial ranges (Fig. 25). The clusters identify pattern in the data of harvesting operations regarding time of year, weather conditions, road and terrain accessibilities, soil conditions and harvested and mean stem volumes and harvest type. Although showing clear patterns between these variables, the high noise ratio indicate density-based clustering algorithms are not appropriate for this data.

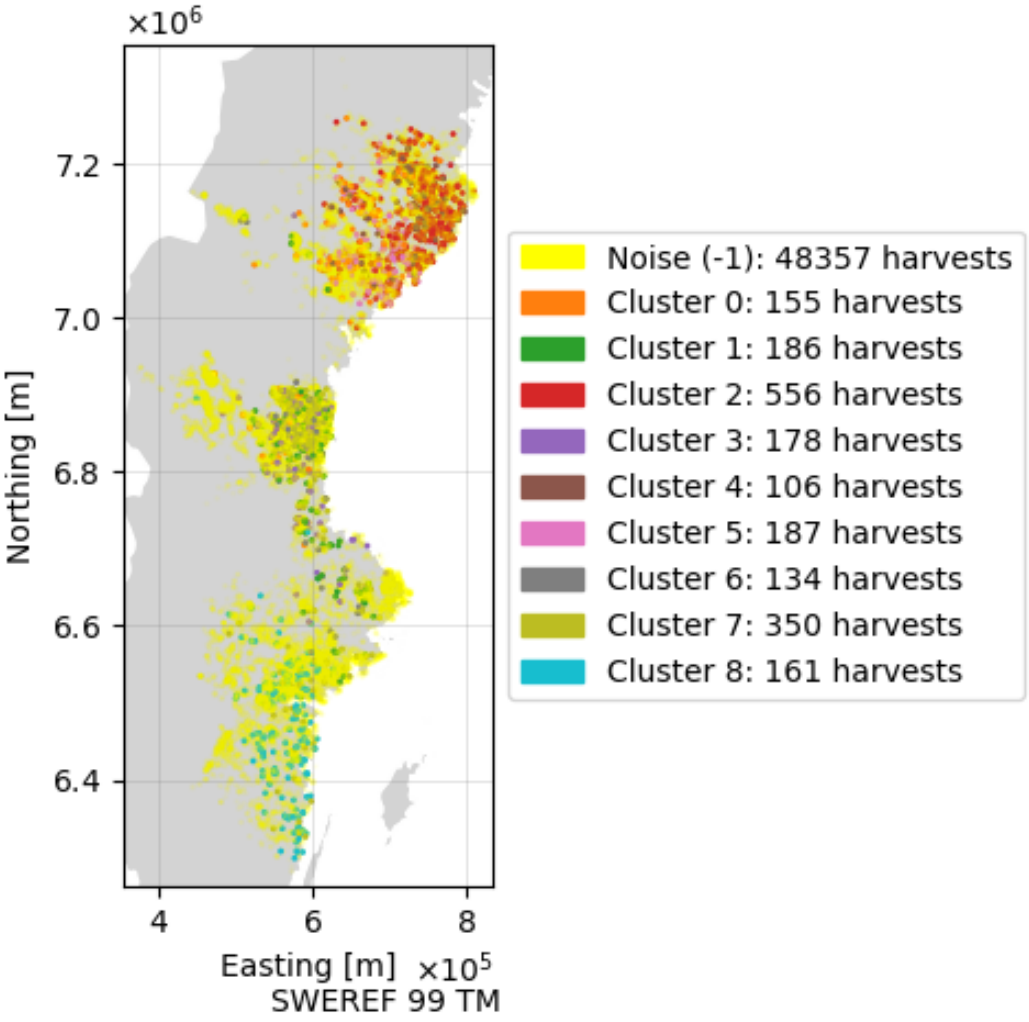


Figure 25. The HDBSCAN clusters spatially visualized. each cluster is colour-coded

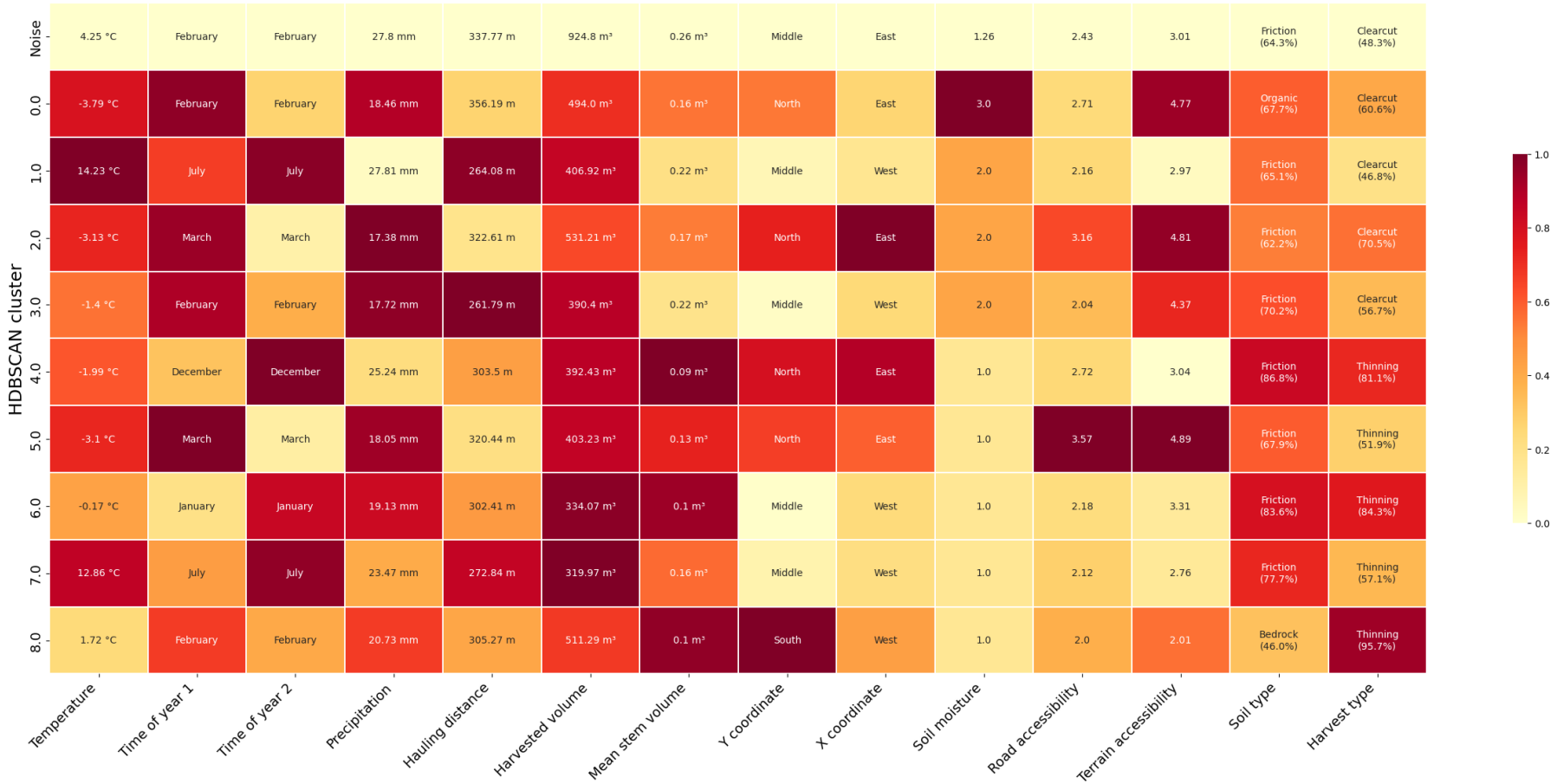


Figure 26. A heatmap for each HDBSCAN cluster where clusters are coloured in dark based on their mean values' departure from global mean for numerical variables and prevalence of the mode value relative to number of categories for categorical variables.

Publishing and archiving

☒ YES, I, Elliot Eriksson, have read and agree to the agreement for publication and the personal data processing that takes place in connection with this.