Mittuniversitetet
MID SWEDEN UNIVERSITY

Classify part of day and snow on the load of timber stacks

A comparative study between partitional clustering and competitive learning

My Nordqvist

Master Thesis – Computer Engineering MA, Final Project
Main field of study: Computer Engineering
Credits: 30hp
Semester/Year: Spring, 2021
Supervisor: Sarder Fakhrul Abedin, sarder.fakhrulabedin@miun.se, Tanja Keisu, Tanja.Keisu@biometria.se
Examiner: Mikael Gidlund, mikael.gidlund@miun.se
Degree program: Master of Science in Engineering: Computer Engineering, 300 credits

Abstract

In today's society, companies are trying to find ways to utilize all the data they have, which considers valuable information and insights to make better decisions. This includes data used to keeping track of timber that flows between forest and industry. The growth of Artificial Intelligence (AI) and Machine Learning (ML) has enabled the development of ML modes to automate the measurements of timber on timber trucks, based on images. However, to improve the results there is a need to be able to get information from unlabeled images in order to decide weather and lighting conditions. The objective of this study is to perform an extensive for classifying unlabeled images in the categories, daylight, darkness, and snow on the load. A comparative study between partitional clustering and competitive learning is conducted to investigate which method gives the best results in terms of different clustering performance metrics. It also examines how dimensionality reduction affects the outcome. The algorithms K-means and Kohonen Self-Organizing Map (SOM) are selected for the clustering. Each model is investigated according to the number of clusters, size of dataset, clustering time, clustering performance, and manual samples from each cluster. The results indicate a noticeable clustering performance discrepancy between the algorithms concerning the number of clusters, dataset size, and manual samples. The use of dimensionality reduction led to shorter clustering time but slightly worse clustering performance. The evaluation results further show that the clustering time of Kohonen SOM is significantly higher than that of K-means.

Keywords: Machine Learning (ML), Unsupervised Learning, Cluster Analysis, Partitional Clustering, Competitive Learning, Dimensionality Reduction, Principal Component Analysis (PCA), K-means, Kohonen Self-Organizing Map (SOM), Timber

Foreword

I would like to thank Tanja Keisu at Biometria, Stefan Stark, and Linda Åstrand at Forefront Consulting, for the opportunity to write this survey. I would also like to thank Sofie Hellmark and Per Fahlén for their help during this time.

Table of Contents

Terminology

Abbreviations

AI	Artificial Intelligence
CSV	Comma-Separated Values
ML	Machine Learning
PCA	Principal Component Analysis
SOM	Self-Organizing Map

1 Introduction

In today’s society, we are living in the “age of data”. Every day, companies try to use approaches to manage all the data they have, considering valuable information and insights, to make better decisions. The opportunities with faster computers, better computation power, cheaper and bigger storage, have enabled the growth of Artificial Intelligence (AI) and Machine Learning (ML) [1].

The Swedish forest is a key enabler in Sweden's transition to a fossil-free and sustainable society. The Swedish forest industry is one of the world's largest wood pulp, paper and sawn timber exporters. Ten percent of Sweden's total export value consists of products from the forest industry. All timber that flows between forest and industry are measured and tracked with data. The methods used could be improved to make the work more efficient, which can make it more time and cost effective.

1.1 Biometria

Biometria [2] is a member-owned, economical association that is a central player in the Swedish forest industry. They carry out measurements of the timber that flows between forest and industry. Biometria has 850 employees spread all over Sweden and 250 member companies that represent both buyers and sellers of timber raw material, and about 90 affiliated sawmill companies.

They offer services for digitization and automation of both, timer flow and trade. Also, they offer a platform with standardized and quality-assured information that can be used to build new modern services, for both Biometria and other companies.

1.2 Forefront Consulting

Forefront Consulting [3] is a business and IT consulting company, which guides companies and organizations in the digital landscape.

Forefront has about 400 employees which guides companies and organizations through the digital landscape. They have three focus areas, Business Agility, Cloud Transformation and AI. According to [4], they want to be in the leading edge of the future with focus on innovations, and how new technologies can help their costumers develop their business value.

1.3 Background and problem motivation

Every day, up to 7000 transports with timber trucks take place at Biometria's 400 measuring places around Sweden. Remote measurement is applied at many of the measuring places, which means that each timber truck is photographed to be able to make measurements of the timber’s volume and quality digitally. This is done manually by timber meters at remote measurement centers. In order to simplify the work ML models exist to automatically assess the volume of timber, which are used as an aid for timber meters. The goal with the measurement systems is that they should be as efficient as possible which is addressed in [5]. Every minute a timber truck is stationary is a cost, and because of this the measurements need to go as fast as possible.

Based on the results of the development there is a lack of knowledge regarding how accurate the models are in different lighting and weather conditions. To be able to improve the results there is a need to be able to classify unlabeled images e.g. if the images are taken in, daylight, headlight, back-light, darkness, or if there is snow on the load.

In order to classify unlabeled imaged, cluster analysis is an important method which is addressed in [6]. The use of cluster analysis can help to discover knowledge in data which includes finding patterns and similar data points, and then grouping them together.

1.4 Overall aim

The overall aim with this thesis is:

- Study the possibility to classify unlabeled images of timber on timber trucks in different weather and lighting conditions.
- Discuss the possibility of using two different cluster analysis algorithms for the classification.
- Investigate whether there is potential that new data and new knowledge can create conditions for training new models for measuring timber volumes.

1.5 Concrete and verifiable goals

The goal with this thesis is:

- Develop cluster analysis models based on unlabeled images of timber on timber trucks.

- The models should be able to classify the images in three categories: daylight, darkness, and snow on the load.

- A comparative study between partitional clustering and competitive learning in order to investigate which of the algorithms gives best result.

1.5.1 Research questions

The main research questions in this thesis are as follows.

- **P1:** Can cluster analysis be used to classify unlabeled images in the categories: daylight, darkness and snow on the load?
- **P2:** Which of the methods, partitional clustering and competitive learning, gives the best result according to size of dataset, data preprocessing, number of clusters, clustering performance, clustering time and manual samples?
- **P3:** How does dimensionality reduction affect the result, according to clustering time and clustering performance?

1.6 Scope

The thesis has the main focus on fulfilling the company's desire to investigate whether it is possible to cluster an unlabeled dataset into daylight, darkness and snow on the load, and at the same time compare two cluster algorithms.

The data is provided by Biometria and different datasets are used. Only the images taken from the side of the timber truck are used. One limitation is the lack of information according to the weather and lighting conditions when the images were taken. To investigate whether the cluster analysis has clustered daylight, darkness, and snow on the load in separate clusters, manual samples must be taken. The samples are taken by the author, who had no experience of examining images of timber before the thesis. Due to that, the result must be limited according to human factors.

The platform used has limited memory, which means that there is a limit to how much data can be used.

1.7 Main contributions

The focus of the thesis is to investigate a solution of the problem described in chapter 1.3, which is that there is no well-known method to automatically determine weather and lighting conditions from images representing timber on timber trucks. In an effort to resolve this problem the main contributions of the thesis are as follows:

- Use unsupervised learning to automate the classification of weather and lighting conditions.
- Apply cluster analysis on an unlabeled dataset of images in order to classify them in darkness, daylight and snow on the load.

1.8 Outline

The first chapter presents an introduction including background, purpose, goal and research questions. Chapter two describes theory and related work for the reader to gain an understanding about the technical parts of the thesis. In the third chapter, the method used in the thesis is shown. Chapter four describes the construction and implementation and in chapter five the result is presented. In chapter six the results are discussed, and future work is presented. The conclusions are drawn in chapter seven and then the references are presented.

2 Theory

This chapter describes relevant theory that will be used in the thesis, to help the reader gain an understanding of the coming chapters.

2.1 Machine learning

Machine learning (ML) is the field for a computer to learn from data with help of a learning algorithm. The authors in [7] explain that ML has developed enormously during the last 20 years. Now it is an important and widespread technology and is used when developing software for among other, computer vision, speech recognition, and robot control.

There exists a wide range of ML algorithms, which has been developed to solve different types of problems, with different types of data. Today it can be easier to use ML algorithms and train systems to get a specific outcome instead of program it manually.

ML algorithms can be divided into three approaches, supervised learning, unsupervised learning and semi-supervised learning [8]. Unlike supervised learning, unsupervised learning does not require labeled data. Semi-supervised learning uses both labeled and unlabeled data.

2.2 Unsupervised learning

Unsupervised learning is one of the three ML approaches. The authors in [9] explain that with this type of learning process the data does not need to be labeled. Instead the algorithm tries to find relationships between the data points in the dataset, which is useful when trying to extract meaningful information from data.

2.3 Artificial neural network

Artificial neural network, or neural network, is computational models with the aim of recognize underlying relationships in data. The authors in [1] describe the architecture of a neural network as a simulation of biological neurons in the human brain. The general structure of neural network is that it consists of layers of interconnected nodes, where the nodes and connections mimic the network of neurons in the human brain. A typical structure of a neural network is shown in *figure 1*, and it consists of an input layer, a hidden layer and an output layer, with inter-connections.

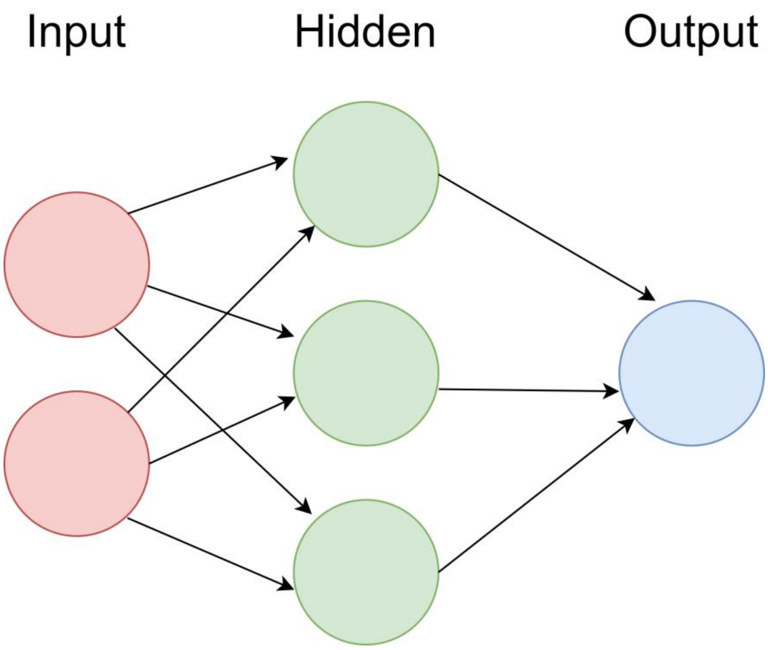

Figure 1: A typical structure of a neural network

2.4 Competitive learning

Competitive learning is a type of unsupervised learning in artificial neural networks. According to [10] different neurons compete regarding who should be allowed to represent the current input. A neuron that is considered to be the best at representing an input becomes the one who is allowed to learn, which leads to different neurons becoming specialized to represent different types of inputs.

2.5 Partitional clustering

Partitional clustering algorithms belong to unsupervised learning methods. According to [11] the data points are split into *k* partitions, where *k* is the amount of clusters. The partition can be calculated with the square error function, where the goal is to minimize the error. Each cluster has two properties which are:

- Each cluster must have at least one data point
- Each data point must belong to exactly one cluster

2.6 Cluster analysis

Cluster analysis methods lies under the unsupervised learning methods. These methods are often used when the data are missing labels, because they are trying to find similarities and relationships in the input data only. The author in [12] describes cluster analysis as a collection of several mathematical methods. It is the mathematical methods that finds similarities in the data, or between objects, and collect them in the same cluster. In that case similar objects will belong to one cluster, and dissimilar objects belong to different clusters, which can be seen in *figure 2*.

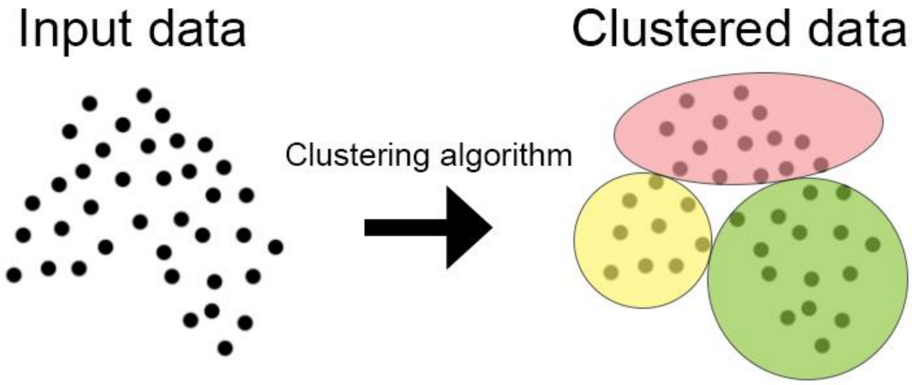

Figure 2: Illustration of cluster analysis

2.6.1 K-means

K-means is a partitional clustering, and cluster analysis, algorithm. According to [13] it is a widely used algorithm because it is easy to implement and understand the results. K-means uses iteration to group a dataset into at least two clusters. The first step in the algorithm is to randomly select cluster centers for each data point in the dataset. The second step is to calculate the Euclidean distance between each data point and the cluster centers, and assign the data point to the closest cluster center. Then new cluster centers, and new Euclidian distances from each data point to the new cluster centers, are calculated. When no data point is assigned to a new cluster, the algorithm is terminated.

2.6.2 Kohonen Self-Organizing Map

Kohonen Self-Organizing Map (SOM) is a common algorithm within the categories, cluster analysis and competitive learning. According to [14] it has only two layers, an input layer and an output layer, which is shown in *figure 3*. The input layer represents the input data, and the output layer represents the output data. During the learning process, all nodes compete for the input signal by finding the best match between the input and the nodes in the output grid. When the input signal finds a winning node, it will be associated to that one.

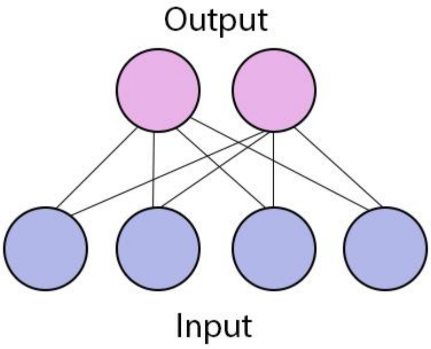

Figure 3: Structure of a Kohonen SOM neural network

2.7 Dimensionality reduction

Dimensionality reduction is a method to transform a dataset from a high-dimensional space to a lower dimensional space. The authors in [15] mention that the number of input variables is often reduced before a machine learning algorithm can be applied, when dealing with high-dimensional data.

2.7.1 Principal component analysis (PCA)

Principal component analysis (PCA) is one technique for dimensionality reduction. The algorithm analyzes information to be able to transform important data into a new data set, usually with fewer dimensions. According to [16] the variables in the transformed data set are called principal components and they are used to find patterns of similarities in data. The variables in the original data set are dependent and after the transformation, the variables, or principal components, are orthogonal. The authors in [17] mean that the orthogonal components enable lack of redundancy of data. PCA’s biggest advantages are the decreased computation requirements and increased efficiency, due to the smaller amount of dimensions. According to [18] PCA is one of the most popular techniques for dimensionality reduction, which are the reason it will be used in this thesis.

2.8 Evaluation

Cluster analysis is used when the ground truth labels are not known. Because of that, the evaluation of the model must be performed by techniques which use the model itself. Three of these techniques are, Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index.

2.8.1 Silhouette Coefficient

The Silhouette Coefficient [19] can be used to evaluate cluster analysis models. The score is between -1 and 1, where higher value represents better defined clusters. The Silhouette Coefficient is calculated by (1), where a_i, is defined as the mean distance of the i-th point to all other points in the same cluster, and b_i is defined as the mean distance of the i-th point to all other points in its nearest neighbor cluster.

$$s = \frac{b_i - a_i}{\max(a_i, b_i)} \qquad (1)$$

2.8.2 Calinski-Harabasz Index

Calinski-Harabasz Index [20] can be used to evaluate cluster analysis models. The score represents the relationship of the spread, the sum of distances squared, between and within clusters. A higher score represents better defined clusters. The index is calculated by (2) where n_E is the size of the dataset, clustered into k clusters. The trace of the spread between the clusters is defined by $tr(B_k)$ and (3), and $tr(W_{k)}$ is the trace of the spread within the clusters, defined by (4).

$$s = \frac{tr(B_{k)}}{tr(W_{k)}} \times \frac{n_E - k}{k-1} \qquad (2)$$

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T \qquad (3)$$

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T \qquad (4)$$

2.8.3 Davies-Bouldin Index

Davies-Bouldin Index [21] can be used to evaluate cluster analysis models. The lowest score is zero, and a lower value represents better defined clusters. This type of method measures the average similarity between clusters. This is done by comparing the distance between clusters and the size of each cluster. Davies-Bouldin Index is calculated by (5), where s_i is defined as the average distance between the data points and the center point in a cluster *i*, s_j is defined as the average distance between the data points and the center point in cluster i most similar cluster *j*, and d_{ij} is defined as the distance between the cluster centers points *i* and *j*.

$$DB = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j} \frac{s_i + s_j}{d_{ij}} \tag{5}$$

2.9 Microsoft Azure

Microsoft Azure [22] is a cloud platform that can be used for building, testing and managing applications through multiple clouds. Biometria uses Microsoft Azure for software developing, and their data is stored in Data Lake, which is the reason it is used.

2.9.1 Databricks

Databricks [23] is a platform adapted for the Microsoft Azure cloud services platform. It is a data analytics platform which offers environments for developing applications.

2.9.2 Data Lake

Data Lake [24] is a repository where data is stored and organized in its original, raw format. It is usually configured on a cluster of scalable commodity hardware, which is either locally or in the cloud. An advantage of this is that data can be stored without concerning about storage capacity.

2.10 Programming

In this subchapter, the program language and the most important libraries and modules are presented.

2.10.1 Python

Python is a programming language created by Guido van Rossum. It is a high level language and the authors in [25] mentions that Python is one of the most popular languages for scientific computing. Python offers a big amount of modules such as libraries and implementations of popular algorithms for, among others, ML.

2.10.2 Scikit-Learn

Scikit-learn is a Python library for ML. According to [26] the library offers state-of-the-art implementations of several ML areas, including clustering algorithms e.g. K-means and PCA.

2.10.3 TensorFlow

TensorFlow [27] is an open source platform with it central point in ML. It offers tools, libraries and resources to easy develop and deploy ML applications.

2.10.4 NumPy

NumPy [28] is an open source Python library created in 2005. It enables numerical computing, powerful n-dimensions arrays and is easy to use.

2.10.5 Pandas

Pandas [29] is a Python library containing tools for working with structured data sets. It is common in the fields, statistic, finance and social scientist.

2.11 Related Work

"WeatherNet: Recognising Weather and Visual Conditions from Street-Level Images Using Deep Residual Learning" is an article written by Mohamed R. Ibrahim, James Haworth and Tao Cheng [30]. They introduced a framework, called *WeatherNet*, consisting of four parallel deep CNN models, *NightNet*, *GlareNet*, *PrecipitationNet* and *FogNet*. The models will detect, or recognize, ten different weather and visual conditions, as *dawn* or *dusk*, *day*, *night*, *glare*, *no glare*, *fog*, *no fog*, *rain*, *snow* and *clear*, from street-level images of urban scenes. One limitation is the accuracy of the model which depends on the accuracy of each four models. Despite that, the misclassification error for each category was below 8%.

Tomasz Krzywicki [31] is the author of the article *"Weather and a part of day recognition in the photos using a KNN methodology"*. He investigated methods to analyze images and classify them as day, night, sunny or cloudy. He analyzed images according to colors, RGB, and horizon edges. Different weather conditions have different colors, for example sunny weather has a blue sky, and cloudy weather has not any blue color in the sky. The author used K-nearest neighbors (KNN) for the solution, and the result had accuracy about 94% to daytime classification, and about 96% to weather classification.

In the article *"Comparative Analysis of SOM Neural Network with K-means Clustering Algorithm"* [32] the authors Usha A. Kumar and Yuvnish Dhamija compared Kohonen SOM and K-means. They used a dataset consisting of forest and woody vegetation with 12 variables. The goal was to cluster different forest cover types in different clusters, and explore how the amounts of clusters and observations affect the results. The result shows that K-means performed better than SOM. In the conclusion they pointed out that the number of clusters does not affect the performance of K-means, while SOM performed better with more clusters.

Biometria and Forefront have together developed ML models for defining volume of timber. One of them is AIDA [33] which is a model that automatically produces the height and length of timber stacks on timber trucks. It is built by SDC (now Biometria) and Forefront Consulting. ASTA [34][35] is the other model, which is a developed model of AIDA that automatically produces the volume of timber stacks on timber trucks. One issue with these models is the lack of labels in the images, for example information that the images were taken in daylight, darkness or if there were snow on the load. Due to that, they cannot know if the models perform better or worse in different conditions.

The research shows that there are studies on the automation of weather conditions in street images, but there is no indication that research has been done on timber on timber trucks. The methods for developing the models have been supervised, and what makes this thesis unique is that it is unsupervised with unlabeled images.

3 Methodology

This chapter presents the methodology of the thesis, which includes the workflow and which tools that are used.

3.1 Workflow

The workflow of the thesis can be seen in *figure 4*. It begins with a theoretical pre-study to get necessary knowledge about the research field. The data is investigated regarding technical specifications and metadata which leads to that the data must be modified and preprocessed to be used. After the data preprocessing the number of clusters is chosen, and then the cluster analysis can run over the dataset. The result of the clustering is evaluated and analyzed. Depending on the result, previous steps are iterated or compared with other results.

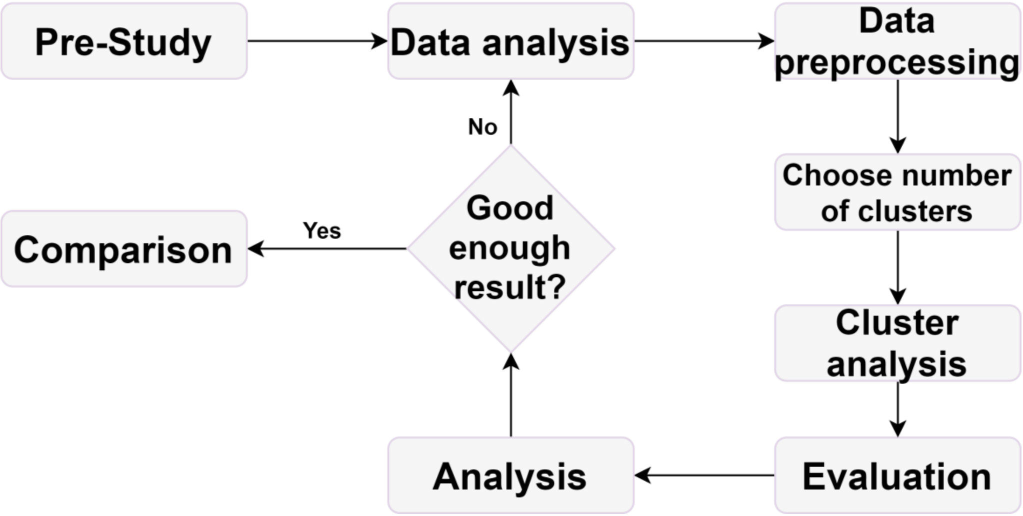

Figure 4: Flowchart of the workflow

3.1.1 Pre-study

The thesis begins by studying the theoretical parts to get knowledge about ML, unsupervised learning, cluster analysis and which algorithms could be used. K-means and Kohonen SOM are selected as algorithms because, as mentioned in chapter 2.6, K-means is a partitional clustering algorithm and Kohonen SOM is a competitive learning algorithm. The programming language Python is selected to be used, because according to [36], it is common when it comes to ML.

3.1.2 Data analysis

The data consists of images with associated metadata. The metadata is examined to be able to select which variables will be relevant in the context. The images are examined with respect to where they are taken, when they are taken and what size they have. This is examined to determine if the thesis has to be delimited. Also, if the size of each image would vary, then some modifications must be done because all data must be in the same size for the results to be correct. Due to the size of the dataset, all images cannot be used. To get a correct distribution throughout the dataset, n is calculated by (6), where *total images* is the whole dataset, and *num of images* is the size of the wanted dataset.

$$n = \frac{total\ images}{num\ of\ images} \tag{6}$$

3.1.3 Data preprocessing

The images are prepared due to size and color. In the beginning of the thesis the original images are used. Later on, in order to improve the results, variables in the metadata is used to crop each image. This is done to only keep the stack of timber in the image, and thus no background. This method is used with the hope that it will be easier for the algorithms to find snow in the images. Also, the time when the timber truck arrives at the measuring point, is used to select about the same distribution of images in darkness as in daylight.

When the images have been loaded to the dataset, another dataset is created. In that dataset, the amount of variables is reduced by dimensionality reduction through PCA [37], in order to answer question P3.

3.1.4 Number of clusters

The numbers of clusters the images will be divided into are changing during the thesis. In the beginning three clusters are used to investigate if the algorithms can cluster the images into "darkness", "daylight" and "snow on the load". Later, two clusters are used to investigate whether the algorithms can cluster the images into "darkness" and "daylight". Then, the given clusters are clustered again to get "snow on the load" or "no snow". This two-step method is used as the final model.

3.1.5 Cluster analysis

The cluster analysis is developed based on one partitional clustering method, K-means, and one competitive learning method, Kohonen SOM. Both algorithms are chosen because they are common and widely used, which is addressed in [38][39].

K-means are running with both the original dataset and the reduced dataset, but Kohonen SOM only runs with the reduced dataset, due to the unreasonable clustering time [40].

Four sizes of datasets are clustered in order to investigate how the size affects the result, including clustering time and clustering performance.

3.1.6 Evaluation

The models are evaluated with three methods, the Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index [41]. The mathematical formulas are explained in chapter 2.8. All three methods are used because they are different from each other, which make it a wider scope of the evaluation.

Due to the lack of ground truth labels, it is not possible to get an exact answer as to whether the clusters are as expected, to answer question P1. To get an understanding of the clusters, manual samples are taken to check which images have been clustered together. Due to the human factor, criteria are set that must be met for an image to belong to one cluster.

3.1.7 Analysis

The results from the evaluation are analyzed to investigate which parameters that could affect the outcome and which changes could be done in order to improve the result.

3.2 Comparative study

The thesis is a comparative study between the algorithms, K-means and Kohonen SOM, which has also been performed by the authors of [32][42-44], where the goal is to answer research question P2. The algorithms are compared considering how the results are affected with different number of clusters, different sizes of dataset, and different data preprocessing. K-means are running with both the original dataset and the reduced dataset, and a comparison of those methods is performed. Kohonen SOM is only running on the reduced dataset, and therefore it is compared with K-means running on the reduced dataset.

3.3 Tools

The tools used in the thesis are presented in this subchapter.

3.3.1 Hardware

The programming is performed on a laptop with the operating system Microsoft Windows 10. In the beginning the cluster analysis is running locally on the computer. Later on, the whole dataset is used and therefore Microsoft Azure, Databricks and Data Lake are used, because it is a platform that Biometria uses and all data are stored in Data Lake. Due to that, the memory capacity is limited and it affects the size of the dataset.

3.3.2 Software

In the beginning of the thesis a small dataset is used, which enables running the cluster analysis locally on the computer. When more data is needed Microsoft Azure and Databricks is used. Biometria stores their data and images in Data Lake, which are the reason Microsoft Azure is used.

3.3.3 Programming

The programming language used is Python together with a few libraries. TensorFlow [45] is used for loading the images to an array. Pandas [46] is used to load the Comma-Separated Values (CSV) file, with metadata to each image. The algorithms K-means [47] and PCA [48] are from the library Scikit-learn. Kohonen SOM is implemented with inspiration from [49].

4 Implementation

Chapter four presents the implementation of the program.

4.1 Structure

Figure 5 shows the structure of the project, where the first step is data analysis, followed by data retrieval, data preprocessing, cluster analysis and evaluation. Depending on the result all steps are repeated.

To know what data to use and how it needs to be prepared it is important with data analysis. Then the data is retrieved from the storage into an array, and it is preprocessed regarding a few parameters. Then the data can be fit into two cluster analysis algorithms, K-means or Kohonen SOM, and the result is evaluated. All steps are iterated with some changes, in order to improve the result.

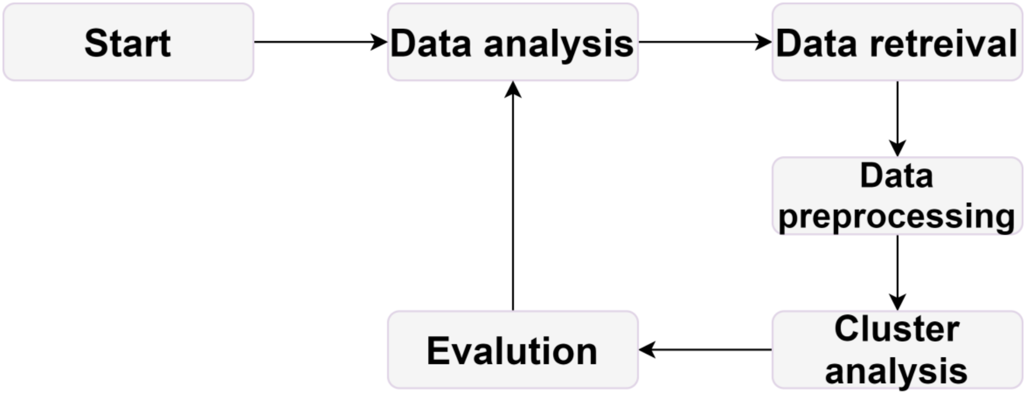

Figure 5: The structure of the project

4.1.1 Data analysis

The data used are images on timber on timber trucks with associated metadata, which have been provided by Biometria. Every truck has three timber stacks and each of them is photographed individually. This means that one image represents a third of a timber truck. Every timber stack is photographed in three angles, one straight from the side and two obliquely from behind which is shown in *figure 6*. Only the images taken straight from the side will be used.

The images are 2952 pixels wide, 1944 pixels high, have 96 dots per inch (dpi), and are in color.

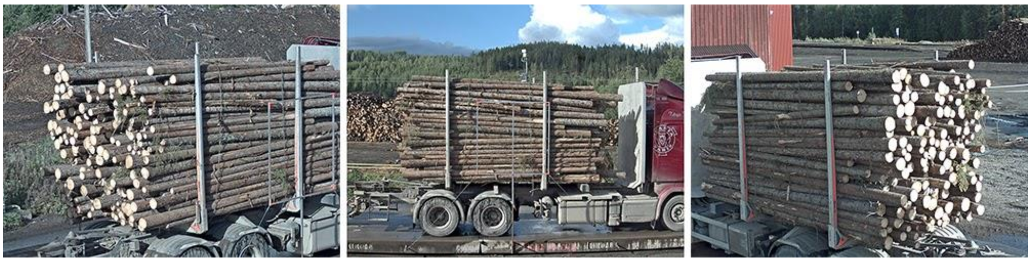

Figure 6: Example of a timber stack photographed in three angles

The metadata is stored in a CSV file with 1 030 162 rows and 32 columns. There are 1 030 161 images and 32 variables with information about each image. The first image in the dataset is from 2019-04-25 and the last is from 2020-09-29. The whole dataset is not used, instead different subsets are used in various tests.

Seven variables are used from the metadata and samples of three of them can be seen in Table 1. *BildTransportIdentitet* combined with *Bildfilnamn* is a unique ID that can be linked to an individual image. Because of that these two variables are used to get the path to each image. *Avlemnad* is date and time when the timber is delivered, which is also the time when the image was taken.

Table 1: Samples from the metadata of the images

BildTransportIdentitet	***Bildfilnamn***	***Avlemnad***
ysj2h8wvr66eee0rughie1xzdl0pwz	*449.jpg*	*2019-04-24 00:02:15:821*
kwm5pbfk3mbckfrlk9ttr7d6zueqs	*584.jpg*	*2019-04-25 00:26:58:158*
dbfiokpeev2ekavx2wo3dz3ev8utxl	*839.jpg*	*2019-10-18 08:03:58:386*
ztzij6klw19v83bics9c0isjcfrnld	*985.jpg*	*2020-09-15 05:09:29:337*

Four variables in the metadata are visualized in *figure 7*, where *TraveMarkeringBredd* is the width and *TraveMarkeringHojd* is the height of the timber stack. *TraveMarkeringX* and *TraveMarkeringY* are the pixels down in the left corner of the stack. These variables are used to crop the images, to be able to do cluster analysis on the timber stack, without the background.

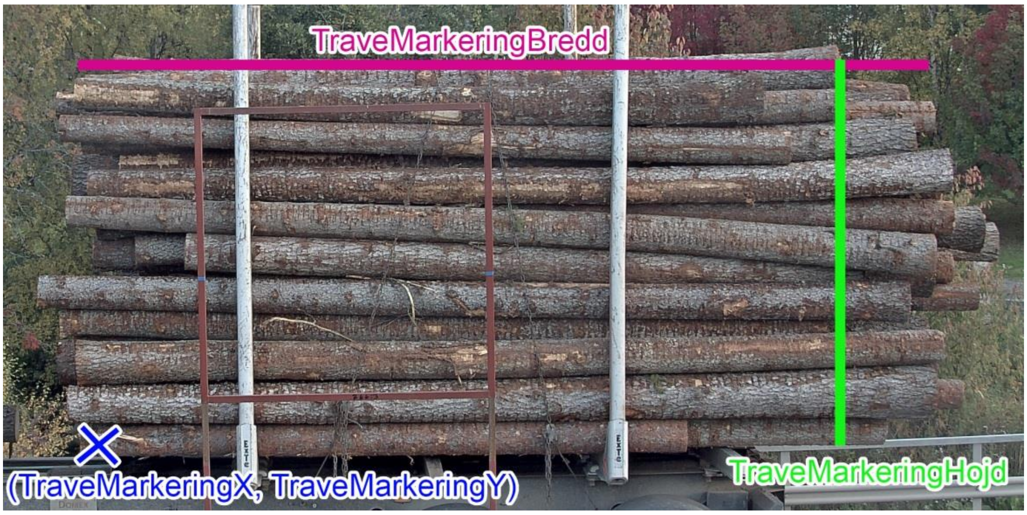

Figure 7: Visualization of four metadata variables

4.1.2 Data retrieval

The images are stored in a folder inside a container in Data Lake, the structure can be seen in *figure 8*. The folder contains subfolders named by the variable *BildTransportIdentitet*, and in each subfolder there are three images named by the variable *Bildfilnamn*.

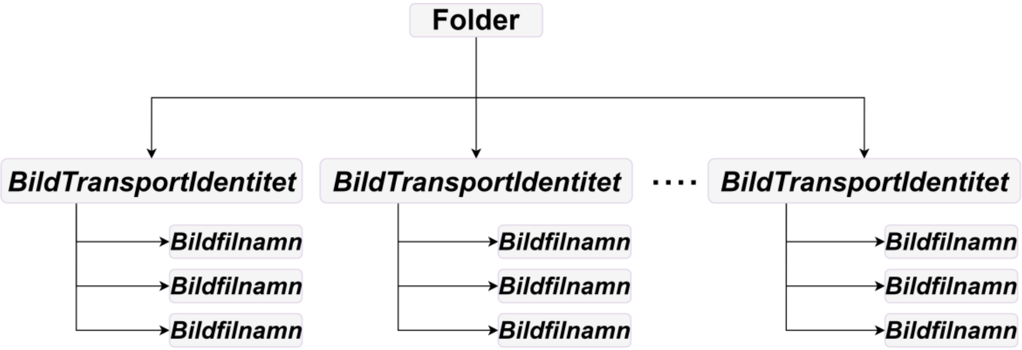

Figure 8: Folder structure in Data Lake

A flow chart for loading the CSV file into the program can be seen in *figure 9*. It begins with reading the CSV file using the library Pandas. A lot of rows are skipped due to the fact that the CSV file has over one million lines. Different types of subsets are used and therefore different functions are used to skip rows. In the first three tests, images from

April 2019 to April 2020 are used, and in the last test only images from January to March 2020 are used. When the CSV file has been read, for each row in the file, the variables *BildTransportIdentitet* and *Bildfilnamn* are merged together into a new variable *Path* which is used to retrieve each image from Data Lake. In the first three tests the program will be terminate when it reaches the end of rows in the CSV file. In the last test another function is added to be sure that the dataset will consists of images taken both in daylight and darkness. The hour in the variable *Avlemnad* is split out into a new variable *Time*. Then all rows with *Time* not equal to *00*, *01*, *02*, *03*, *11*, *12*, *13* and *14* are dropped and will not be used. Then the program will be terminated when it reaches the end of rows in the CSV file.

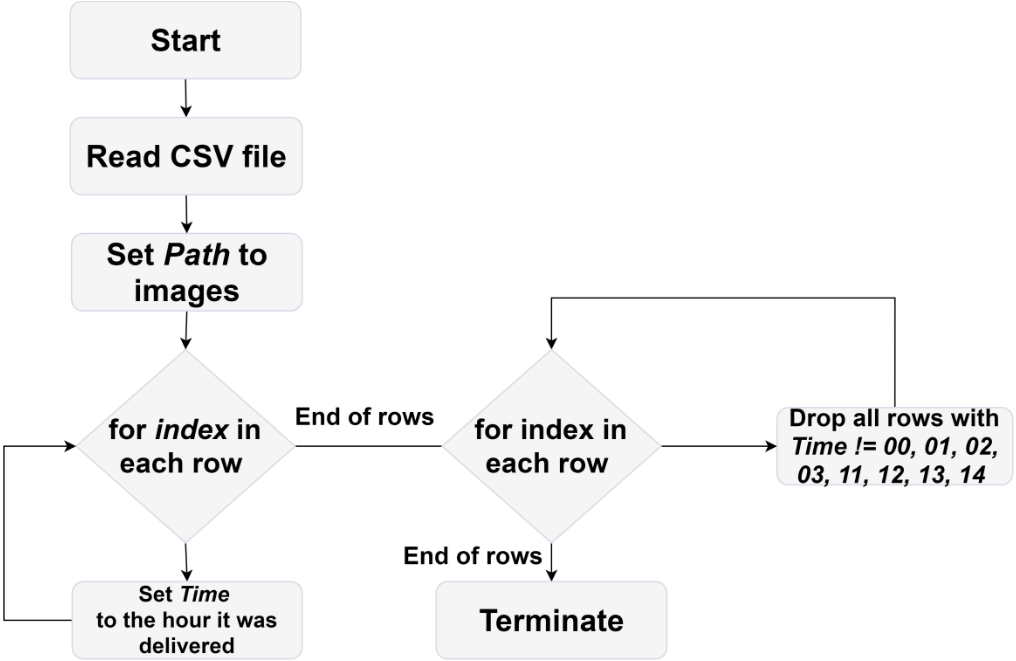

Figure 9: Flowchart for loading data from CSV file

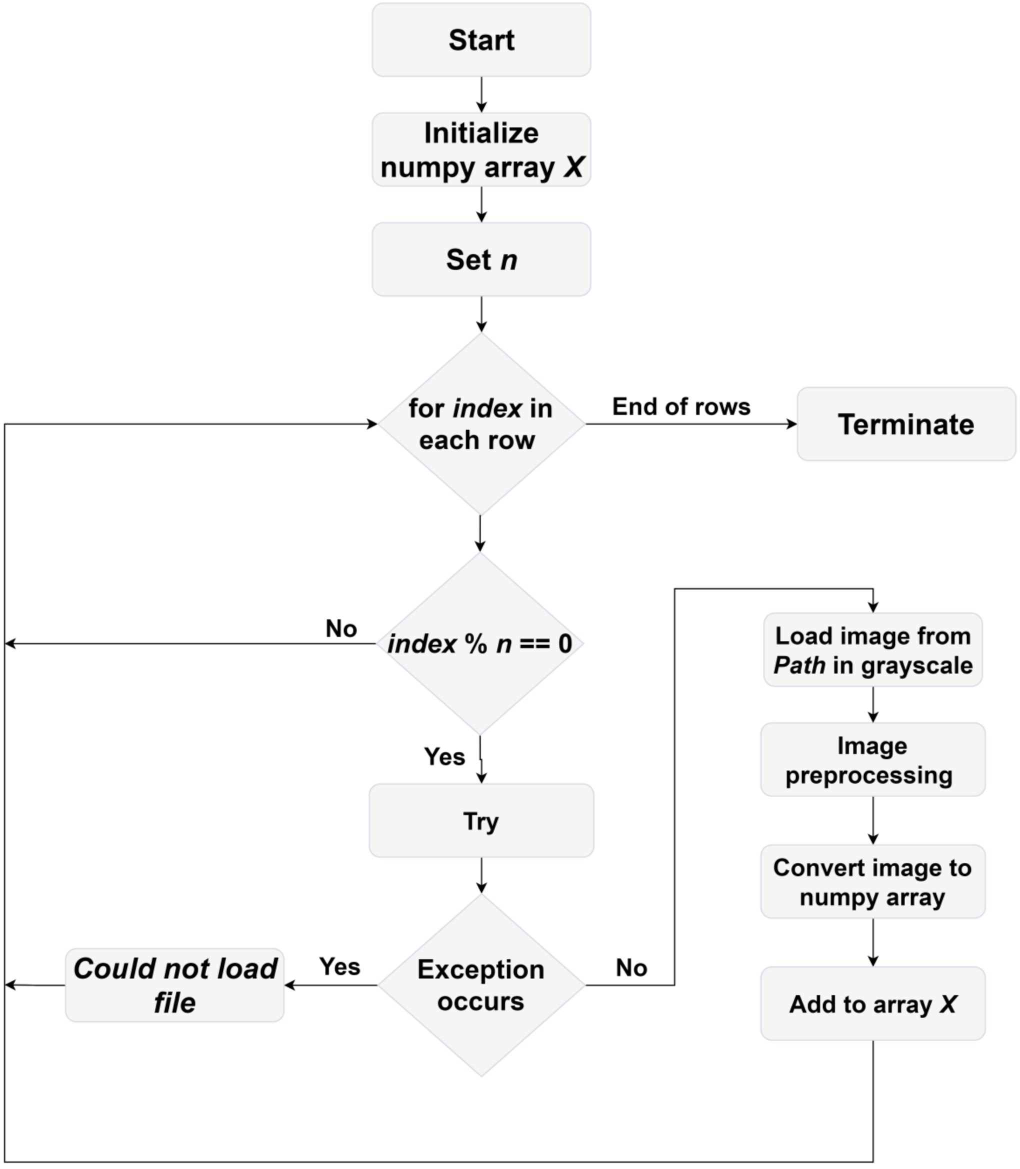

Figure 10: Flow chart for loading the images into a Numpy array

4.1.3 Data preprocessing

To load the images from *Path* into a Numpy array, a function named *Load_images()* is used. A flowchart can be seen in *figure 10.* The variable from column *Path* is used when retrieving the data from Data Lake. A for loop iterates through each row in the file. An if statement is used to check if the image is every *n*:th, where *n* is calculated by (6), to get a spread over all rows loaded from the CSV file. A try block is used to load the images. If no exception occurs the image is loaded in grayscale and preprocessed.

In every test the image is resized to 255x255 pixels to reduce the computation. In the last test the images are preprocessed by cropping with respect to the variables *TraveMarkeringBredd*, *TraveMarkeringHojd*, *TraveMarkeringX* and *TraveMarkeringY,* to only get the timber stack without the background. After the preprocessing the image is converted to a Numpy array and added to another Numpy array, which will represent the dataset, and is called *X*. If an exception occurs while loading the image, the program will print "Could not load file" and iterates to next image. When the loop reaches the end of rows it will be terminated.

Table 2: Example of dimensional reduction with PCA

Images	Before PCA	After PCA
2000	65025	1565
3000	65025	2265
4000	65025	2839
5000	65025	3393

The function *PCA* from the library Scikit-Learn is used to reduce the amount of dimensions in the dataset. Without reduction the components are *n*, *k*, where *n* is the number of images and *k* is the product of the number of components in each image, which are *255x255 = 65025*. After reduction, the components are reduced to approximately

1500-3500 depending on the number of images, which is presented in table 2.

4.1.4 Cluster analysis

A class named *Cluster* is implemented with several functions as shown in *figure 11*. The initializing function takes two parameters, the dataset and the number of clusters, into which the dataset is to be divided. The *K-Means* is implemented with Scikit-Learn's function *KMeans*. The *Kohonen* function is implemented with inspiration from [49]. The function *Cluster_snow* takes the parameters, *dataset*, *algorithm*, *cluster_index* and *cluster*. This function is used to do cluster analysis on sub-clusters. To visualize and evaluate the result, *Plot_graph*, *Plot_samples* and *Evaluate* are used. In the *Evaluate* function the Scikit-Learn functions, *silhouette_score*, *calinski_harabasz_score* and, *davies_bouldin_score* are used.

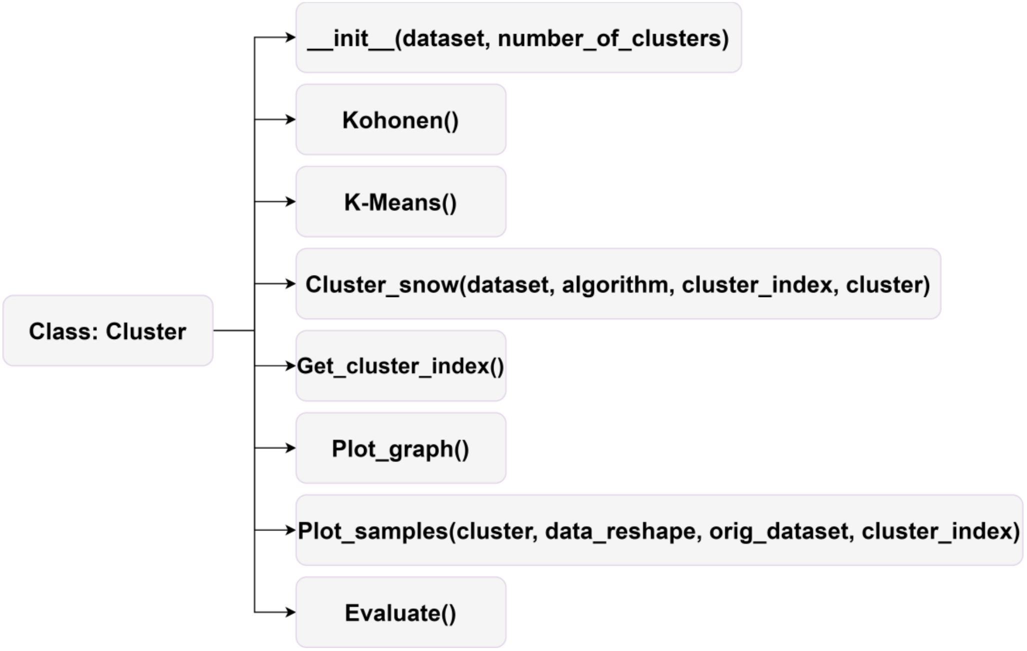
Class: Cluster
__init__(dataset, number_of_clusters)
Kohonen()
K-Means()
Cluster_snow(dataset, algorithm, cluster_index, cluster)
Get_cluster_index()
Plot_graph()
Plot_samples(cluster, data_reshape, orig_dataset, cluster_index)
Evaluate()

Figure 11: Functions in the class Cluster

4.1.5 Evaluation

The models are evaluated with Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index. Due to the images not being labeled, 30 samples are taken from each cluster to check manually which images are clustered together. To define each image, three criteria is set up. *Figure 12* shows three images representing the types of images, where “darkness” is defined by a black sky, “daylight” is defined by a light sky and “snow on the load” is defined as larger white spots on the timber. Peeled bark and birch trees can be seen as white spots, and because of that the evaluation of the samples depends on the human factor and cannot be taken exactly.

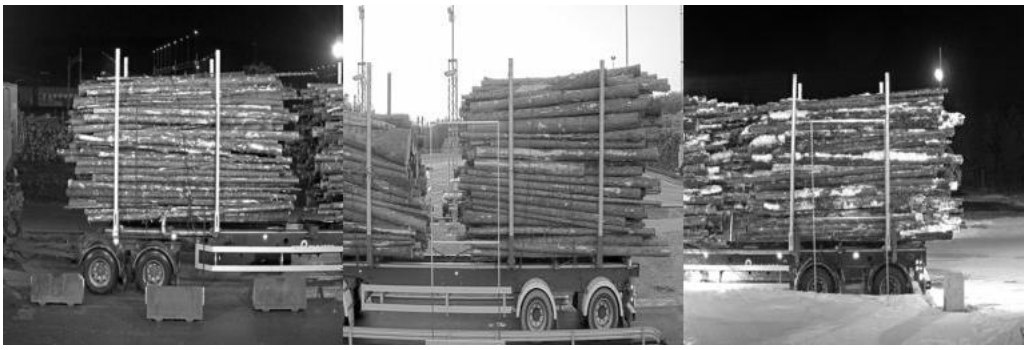

Figure 12: From left: First image is defined as "darkness", second image is defined as “daylight" and third image is defined as “snow on the load"

5 Results

Chapter five presents the results produced from the cluster analysis. The results are presented in several subchapters which are divided into number of clusters, algorithm and graphs. Then the result is presented with respect to size of dataset, dimensional reduction time, clustering time, clustering performance and manual samples. The chapter is presented in text, tables, diagrams and graphs.

5.1 Three clusters

In this subchapter the results of clustering 2000, 3000, 4000 and 5000 images into three clusters with K-means and Kohonen SOM are presented. The images are photographed between April 2019 and April 2020. The performance of the clusters is evaluated with Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index. Also, 30 manual samples have been taken from each cluster to investigate whether the images meet the criteria for daylight, darkness and snow on the load.

5.1.1 Kohonen SOM

Dimensional reduction time, clustering time, Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index are presented in table 3, where each row represents each dataset. The dimensional reduction with PCA took between 1.33 to 8.65 minutes, and the average is 4.85 minutes. The clustering time increases from 15.18 to 87.60 minutes, with average 48.62 minutes. The Silhouette Coefficient varies between 0.139 and 0.149, with average 0.144, the Calinski-Harabasz Index increases from 476 to 1149, with average 828, and the Davies-Bouldin Index varies between 2.304 and 2.509, with average 2.467.

30 manual samples from each cluster are shown in *figure 14*. Each staple represents each cluster, and the green part represents daylight, the red part represents darkness, and the blue part represents snow on the load. The images in each clusters varies between 25 to 30 daylight images, 9 to 22 darkness images, and 0 to 8 snow images.

Table 3: Dimensional reduction time, clustering time and performance evaluation for Kohonen SOM, with three clusters

Images	*PCA time (minutes)*	*Clustering time (minutes)*	*Silhouette Coefficient*	*Calinski-Harabasz Index*	*Davies-Bouldin Index*
2000	1.33	15.18	0.149	476	2.550
3000	2.67	33.45	0.142	706	2.506
4000	6.78	58.28	0.139	982	2.304
5000	8.65	87.60	0.149	1149	2.509
Average	4.85	48.62	0.144	828	2.467

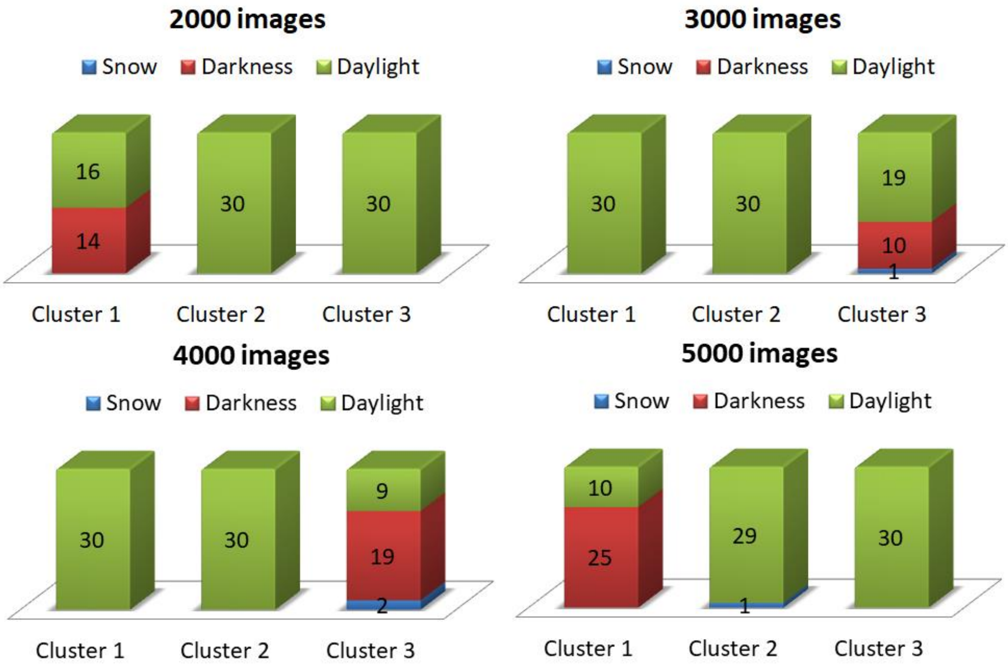

Figure 14: Manual samples taken from each cluster with Kohonen SOM

5.1.2 K-means

Clustering time, Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index are presented in table 4, where each row represents each dataset. The clustering time increases from 1.46 to 3.12 minutes, with average 2.19. The Silhouette Coefficient varies between 0.135 and 0.140, with average 0.137, the Calinski-Harabasz Index increases from 534 to 1301, with average 908, and the Davies-Bouldin Index varies between 2.211 and 2.255, with average 2.229.

Table 4: Clustering time and performance evaluation for K-means, with three clusters

Images	*Clustering time (minutes)*	*Silhouette Coefficient*	*Calinski-Harabasz Index*	*Davies-Bouldin Index*
2000	1.46	0.140	534	2.211
3000	1.5	0.135	774	2,255
4000	2.71	0.137	1024	2.239
5000	3.12	0.138	1301	2.212
Average	2.19	0.137	908	2.229

30 manual samples from each cluster are shown in *figure 15*. Each staple represents each cluster, and the green part represents daylight, the red part represents darkness, and the blue part represents snow on the load. The images in each clusters varies between 24 to 29 daylight images, 15 to 26 darkness images, and 1 to 15 snow images

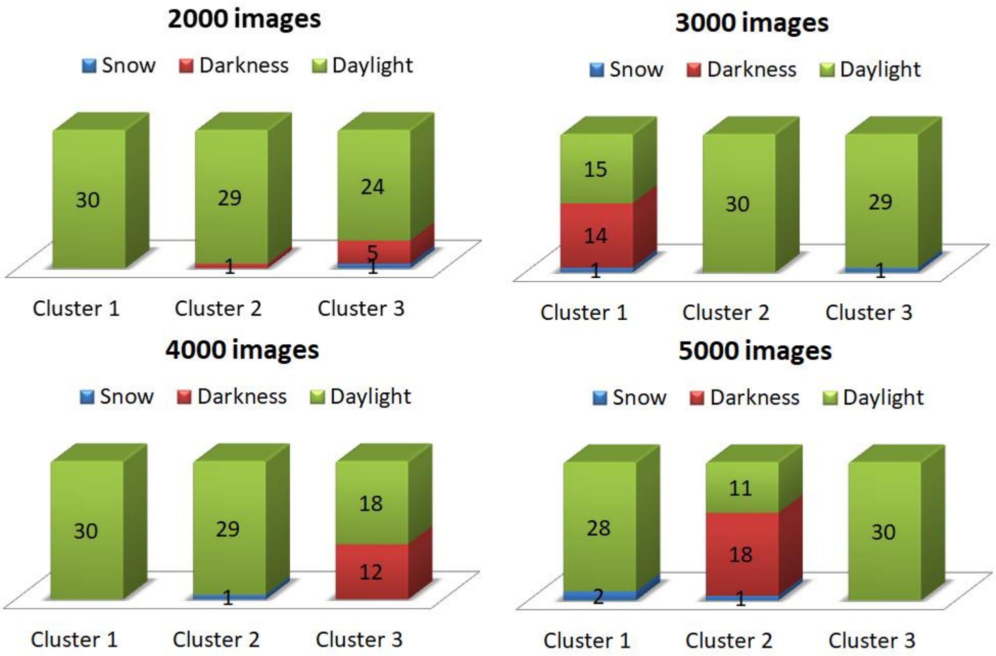

Figure 15: Manual samples taken from each cluster with K-means

5.1.3 K-means with PCA

Dimensional reduction time, clustering time, Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index are presented in table 5, where each row represents each dataset. The dimensional reduction with PCA took between 1.33 to 8.65 minutes, with average 4.85 minutes. The clustering time increases from 4.89 to 12.62 seconds, with average 8.74 seconds. The Silhouette Coefficient varies between 0.138 and 0.144, with average 0.140, the Calinski-Harabasz Index increases from 484 to 1149, with average 837, and the Davies-Bouldin Index varies between 2.424 and 2.526, with average 2.462.

30 manual samples from each cluster are shown in *figure 16*. Each staple represents each cluster, and the green part represents daylight, the red part represents darkness, and the blue part represents snow on the load. The images in each clusters varies between 24 to 30 daylight images, 20 to 22 darkness images, and 0 to 9 snow images.

Table 5: Dimensional reduction time, clustering time and performance evaluation for K-means with PCA, and three clusters

Images	*PCA time (minutes)*	*Clustering time (seconds)*	*Silhouette Coefficient*	*Calinski-Harabasz Index*	*Davies-Bouldin Index*
2000	1.33	4.89	0.144	484	2.526
3000	2.67	9.43	0.138	716	2,467
4000	6.78	8.04	0.139	949	2,424
5000	8.65	12.62	0.140	1198	2.432
Average	4.85	8.74	0.140	837	2.462

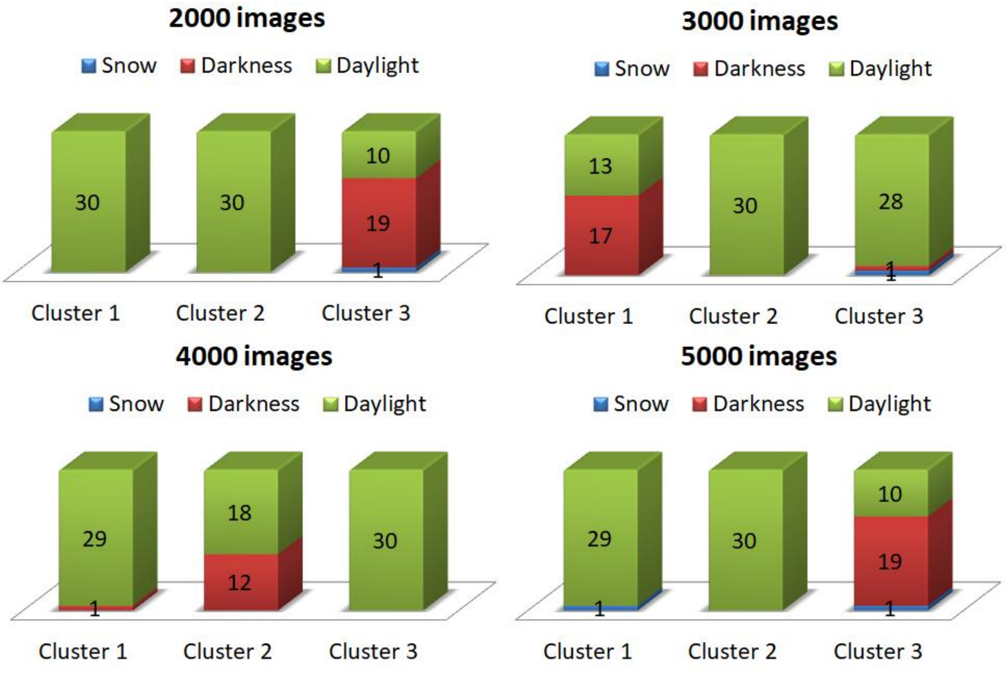

Figure 16: Manual samples taken from each cluster with K-means and PCA

5.2 Two clusters

In this subchapter, the results of clustering 2000, 3000, 4000 and 5000 images into two clusters with K-means and Kohonen SOM are presented. The images are photographed between April 2019 and April 2020. The performance of the clusters is evaluated with Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index. Also, 30 manual samples have been taken from each cluster to investigate whether the images meet the criteria for daylight and darkness.

5.2.1 Kohonen SOM

Dimensional reduction time, clustering time, Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index are presented in table 6, where each row represents each dataset. The dimensional reduction with PCA took between 1.33 to 8.65 minutes, with average 4.85 minutes. The clustering time increases from 11.16 to 62.4 minutes, with average 35.06 minutes. The Silhouette Coefficient varies between 0.235 and 0.242, with average 0.238, the Calinski-Harabasz Index increases from 814 to 1989, with average 1390, and the Davies-Bouldin Index varies between 1.548 and 1.584, with average 1.567.

Table 6: Dimensional reduction time, clustering time and performance evaluation for Kohonen SOM, with two clusters

Images	*PCA time (minutes)*	*Clustering time (minutes)*	*Silhouette Coefficient*	*Calinski-Harabasz Index*	*Davies-Bouldin Index*
2000	1.33	11.67	0.242	814	1.548
3000	2.67	23.87	0.238	1194	1.567
4000	6.78	42.31	0.235	1562	1.584
5000	8.65	62.40	0.238	1989	1.570
Average	4.85	35.06	0.238	1390	1.567

30 manual samples from each cluster are shown in *figure 17*. Each staple represents each cluster, the blue part represents daylight, and the red part represents darkness. One cluster in each dataset has 30 samples of daylight images, and in the second cluster there are between 8 to 10 daylight, and 20 to 22 darkness images.

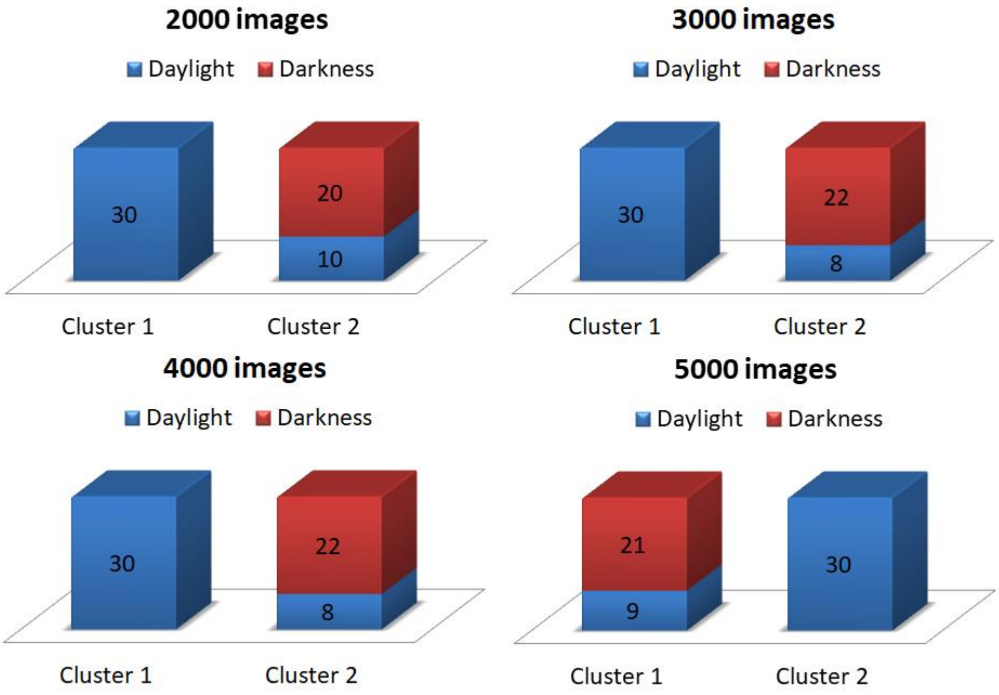

Figure 17: Manual samples taken from each cluster with Kohonen SOM

5.2.2 K-means

Clustering time, Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index are presented in table 7, where each row represents each dataset. The clustering time increases from 1.01 to 1.94 minutes, with average 1.42 minutes. The Silhouette Coefficient varies between 0.237 and 0.240, with average 0.241, the Calinski-Harabasz Index increases from 835 to 2010, with average 1406, and the Davies-Bouldin Index varies between 1.528 and 1.572, with average 1.554.

Table 7: Clustering time and performance evaluation for K-means and two clusters

Images	*Clustering time (minutes)*	*Silhouette Coefficient*	*Calinski-Harabasz Index*	*Davies-Bouldin Index*
2000	1.01	0.247	835	1.528
3000	1.03	0.240	1202	1.557
4000	1.73	0.237	1576	1.571
5000	1.94	0.240	2010	1.560
Average	1.42	0.241	1406	1.554

30 manual samples from each cluster are shown in *figure 18*. Each staple represents each cluster, the blue part represents daylight, and the red part represents darkness. One cluster in each dataset has 30 samples of daylight images, and in the second cluster there are between 4 to 10 daylight, and 20 to 30 darkness images.

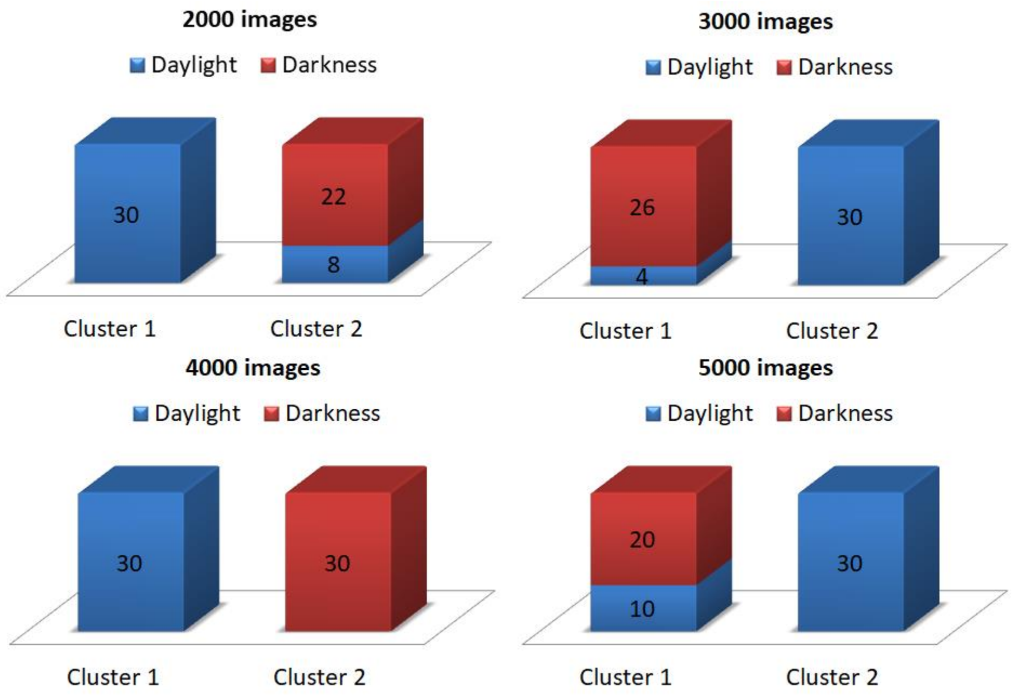

Figure 18: Manual samples taken from each cluster with K-means

5.2.3 K-means with PCA

Dimensional reduction time, clustering time, Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index are presented in table 8, where each row represents each dataset. The dimensional reduction with PCA took between 1.33 to 8.65 minutes, with average 4.85 minutes. The clustering time increases from 3.45 to 6.79 seconds, with average 5.11 seconds. The Silhouette Coefficient varies between 0.236 and 0.245, with average 0.239, the Calinski-Harabasz Index increases from 829 to 1999, with average 1399, and the Davies-Bouldin Index varies between 1.535 and 1.580, with average 1.560.

30 manual samples from each cluster are shown in *figure 19*. Each staple represents each cluster, the blue part represents daylight and the red part represents darkness. One cluster in each dataset has 30 samples of daylight images, and in the second cluster there are between 7 to 0 daylight, and 21 to 23 darkness images.

Table 8: Dimensional reduction time, clustering time and performance evaluation for K-means with PCA and two clusters

Images	*PCA time (minutes)*	*Clustering time (seconds)*	*Silhouette Coefficient*	*Calinski-Harabasz Index*	*Davies-Bouldin Index*
2000	1.33	3.45	0.245	829	1.535
3000	2.67	4.26	0.239	1198	1.562
4000	6.78	5.96	0.236	1569	1.580
5000	8.65	6.79	0.239	1999	1.566
Average	4.85	5.11	0.239	1399	1.560

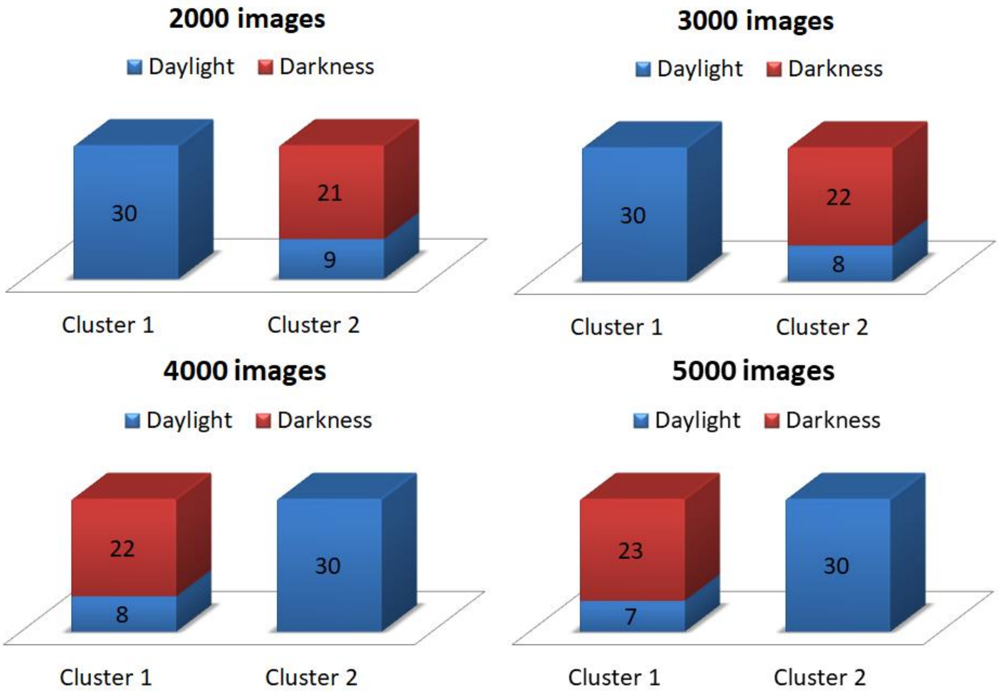

Figure 19: Manual samples taken from each cluster with K-means, with PCA

5.3 Clustering sub-cluster

In this subchapter the result of chapter 5.2 is clustered again into two sub-clusters. The performance of the clusters is evaluated with Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index. Also, 30 manual samples have been taken from each cluster to investigate whether the images meet the criteria for snow on the load.

5.3.1 Kohonen SOM

Dimensional reduction time, clustering time, Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index for each sub-cluster are presented in table 9. The clustering time varies between 5.40 and 28.45 minutes, with average 15.58 minutes. The Silhouette Coefficient varies between 0.080 and 0.123, with average 0.096, the Calinski-Harabasz Index increases from 63 to 195, with average 127, and the Davies-Bouldin Index varies between 3.024 and 3.540, with average 3.334.

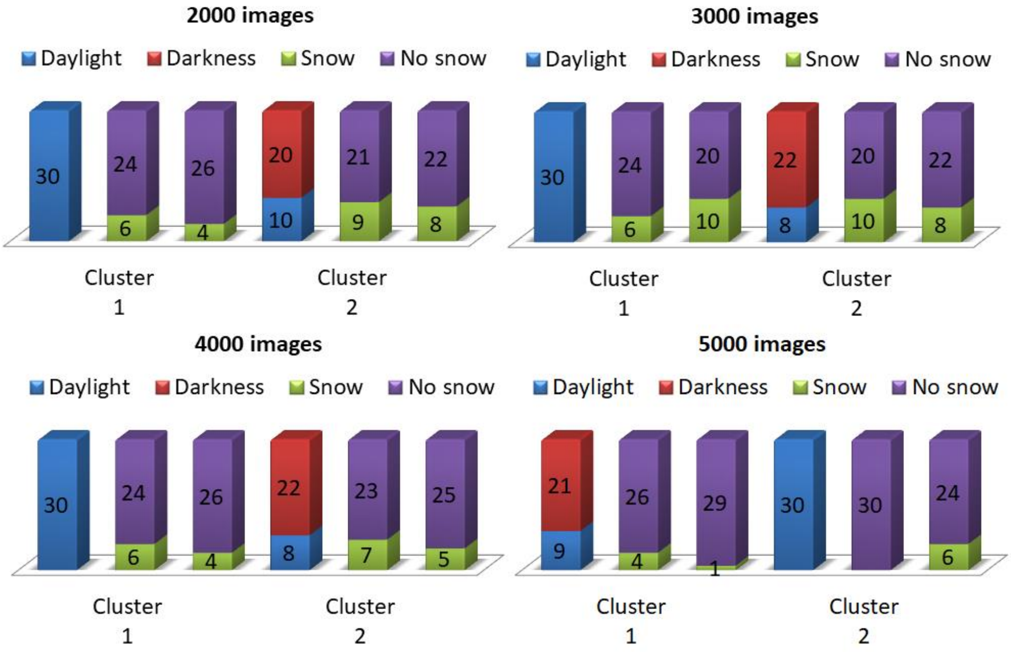

Figure 20: Manual samples taken from each cluster with Kohonen SOM

30 manual samples from each cluster are shown in *figure 20*. Each staple represents each cluster. The blue part represents daylight and the red part represents darkness which is the same result as in chapter 5.2.1. The green parts represent snow and the purple part represents no snow. The images in each sub-cluster vary between 0 to 10 snow images, and 20 to 30 images with no snow.

Table 9: Dimensional reduction time, clustering time and performance evaluation for Kohonen SOM, with PCA and two clusters

Images	*Sub-cluster*	*Clustering time (minutes)*	*Silhouette Coefficient*	*Calinski-Harabasz Index*	*Davies-Bouldin Index*
2000	Cluster 1	5.97	0.108	90	3.178
	Cluster 2	5.40	0.080	63	3.540
3000	Cluster 1	9.31	0.076	66	3.499
	Cluster 2	14.29	0.113	166	3.024
4000	Cluster 1	15.67	0.080	89	3.531
	Cluster 2	27.21	0.103	195	3.284
5000	Cluster 1	18.34	0.123	167	3.274
	Cluster 2	28.45	0.091	185	3.345
Average		15.58	0.096	127	3.334

5.3.2 K-means

Clustering time, Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index for each sub-cluster are presented in table 10. The shortest clustering time is 22.08 seconds, the longest is 79.80 seconds, and the average time is 41.40 seconds. The Silhouette Coefficient varies between 0.086 and 0.123, with average 0.105, the Calinski-Harabasz Index increases from 112 to 276, with average 184, and the Davies-Bouldin Index varies between 2.194 and 2.992, with average 2.597.

Table 10: Clustering time and performance evaluation for K-means and two clusters

Images	*Sub-cluster*	*Clustering time (seconds)*	*Silhouette Coefficient*	*Calinski-Harabasz Index*	*Davies-Bouldin Index*
2000	Cluster 1	22.08	0.108	123	2.760
	Cluster 2	35.57	0.097	112	2.984
3000	Cluster 1	30.42	0.112	204	2.677
	Cluster 2	33.71	0.096	163	2.992
4000	Cluster 1	40.90	0.122	276	2.502
	Cluster 2	79.80	0.086	212	2.194
5000	Cluster 1	54.45	0.123	180	2.431
	Cluster 2	34.34	0.096	203	2.241
Average		41.40	0.105	184	2.597

30 manual samples from each cluster are shown in *figure 21*. Each staple represents each cluster. The blue part represents daylight and the red part represents darkness which is the same result as in chapter 5.2.2. The green parts represent snow and the purple part represents no snow. The images in each sub-cluster vary between 2 to 10 snow images, and 20 to 28 images with no snow.

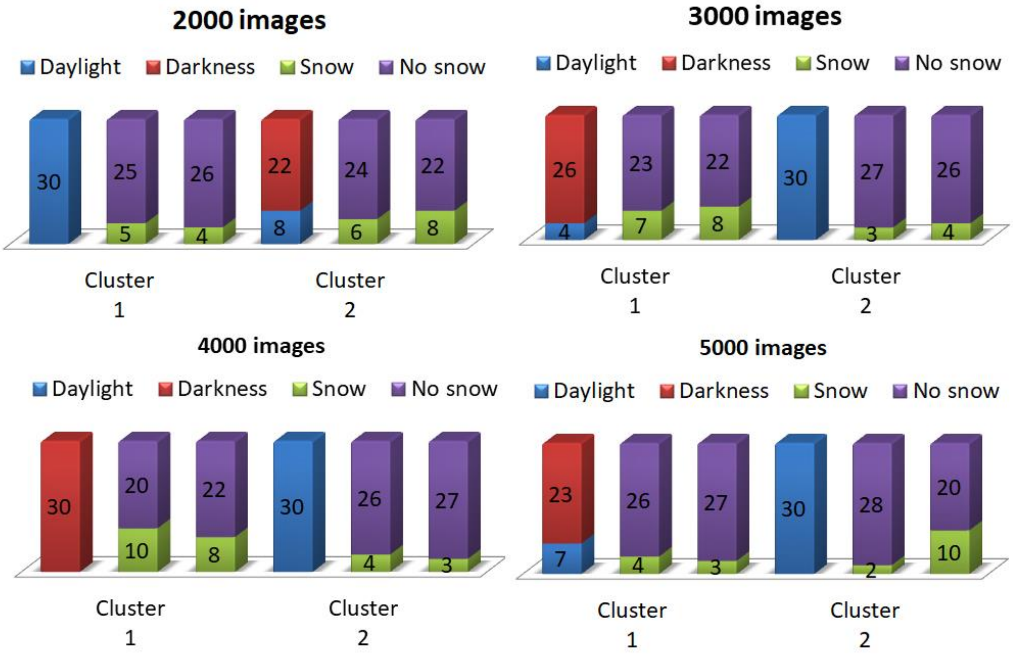

Figure 21: Manual samples taken from each cluster with K-means

5.3.3 K-means with PCA

Dimensional reduction time, clustering time, Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index for each sub-cluster are presented in table 11. The shortest clustering time is 2.32 seconds, the longest is 12.94 seconds, and the average time is 4.67 seconds. The Silhouette Coefficient varies between 0.023 and 0.112, with average 0.076, the Calinski-Harabasz Index increases from 73 to 178, with average 132, and the Davies-Bouldin Index varies between 3.214 and 3.609, with average 3.419.

Table 11: Dimensional reduction time, clustering time and performance evaluation for K-means, with PCA and two clusters

Images	*Sub-cluster*	*Clustering time (seconds)*	*Silhouette Coefficient*	*Calinski-Harabasz Index*	*Davies-Bouldin Index*
2000	Cluster 1	2.52	0.071	73	3.609
	Cluster 2	2.32	0.093	85	3.325
3000	Cluster 1	3.34	0.067	115	3.635
	Cluster 2	3.22	0.082	120	3.413
4000	Cluster 1	12.94	0.094	158	3.442
	Cluster 2	4.25	0.073	165	3.338
5000	Cluster 1	5.45	0.023	165	3.382
	Cluster 2	3.34	0.112	178	3.214
Average		4.67	0.076	132	3.419

30 manual samples from each cluster are shown in *figure 22*. Each staple represents each cluster. The blue part represents daylight and the red part represents darkness which is the same result as in chapter 5.2.3. The green parts represent snow and the purple part represents no snow. The images in each sub-cluster vary between 2 to 10 snow images, and 20 to 28 images with no snow.

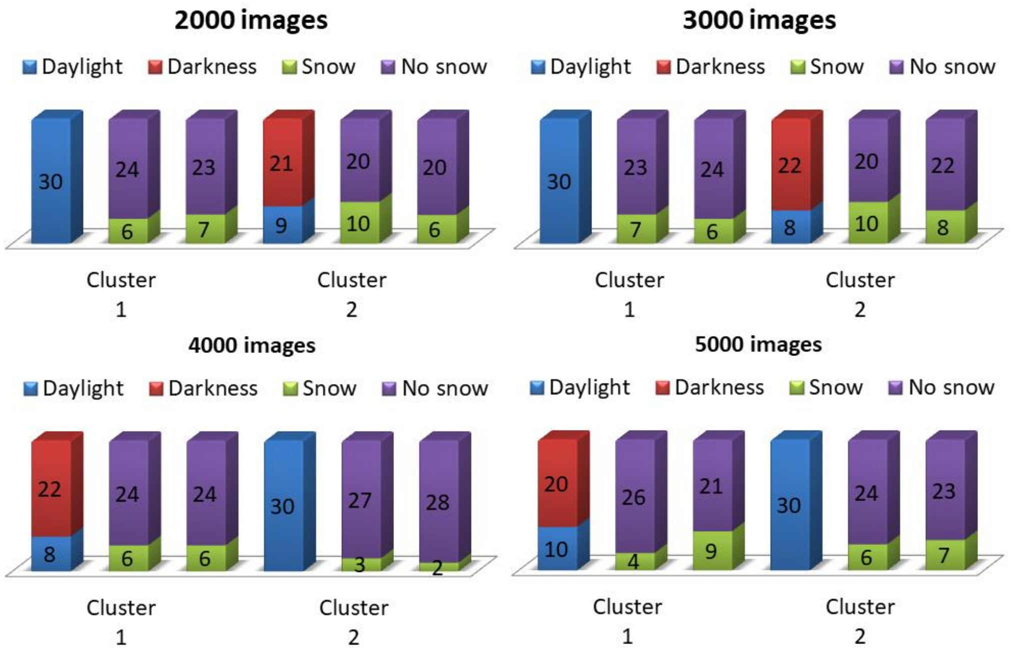

Figure 22: Manual samples taken from each cluster with K-means and PCA

5.4 Clustering sub-cluster with cropped images

In this subchapter, the results of clustering 2000 and 3000 images into two clusters, and cluster them again into two sub-clusters with K-means and Kohonen SOM, are presented. The images are photographed between January and Mars 2020, and the images clustered in the sub-clusters are cropped.

The performance of the clusters is evaluated with Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index. Also, 30 manual samples have been taken from each cluster to investigate whether the images meet the criteria for daylight, darkness and snow on the load.

5.4.1 Kohonen SOM

Dimensional reduction time, clustering time, Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index for each cluster are presented in table 12. The time for dimensional reduction with PCA is 1.14 minutes and 1.98 minutes, with average 1.56 minutes. The first clustering time is 9.37, the second is 21.98 minutes, and the average is 15.67 minutes. The Silhouette Coefficient varies between 0.292 and 0.227, with average 0.259, the Calinski-Harabasz Index increases from 568 to

888, with average 728, and the Davies-Bouldin Index varies between 1.351 and 1.584, with average 1.467.

Table 12: Dimensional reduction time, clustering time and clustering performance for Kohonen SOM and two clusters

Images	*PCA time (minutes)*	*Time (minutes)*	*Silhouette Coefficient*	*Calinski-Harabasz Index*	*Davies-Bouldin Index*
2000	1.14	9.37	0.292	568	1.351
3000	1.98	21.98	0.227	888	1.584
Average	1.56	15.67	0.259	728	1.467

Table 13: Clustering time and clustering performance for Kohonen SOM and sub-clusters

Images	*Sub-cluster*	*Clustering time (minutes)*	*Silhouette Coefficient*	*Calinski-Harabasz Index*	*Davies-Bouldin Index*
2000	Cluster 1	9.41	0.022	49	5.287
	Cluster 2	1.73	0.119	37	2.786
3000	Cluster 1	6.58	0.093	88	2.958
	Cluster 2	18.43	0.043	92	3.627
Average		9.03	0.069	67	3.664

Clustering time, Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index for each sub-cluster are presented in table 13. The shortest clustering time is 1.73 minutes, the longest is 18.43 minutes, and the average is 9.03 minutes. The Silhouette Coefficient varies between 0.043 and 0.119, with average 0.069, the Calinski-Harabasz Index increases from 37 to 92, with average 67, and the Davies-Bouldin Index varies between 2.786 and 5.287, with average 3.664.

30 manual samples from each cluster are shown in *figure 23*. Each staple represents each cluster and corresponding sub-cluster, the blue part represents daylight, the red part represents darkness, the green part represents snow and the purple represents no snow. The clusters in both datasets have 30 samples of daylight, and 30 samples of darkness images, each. In the sub-clusters there are between 4 and 12 snow images and the rest represents no snow.

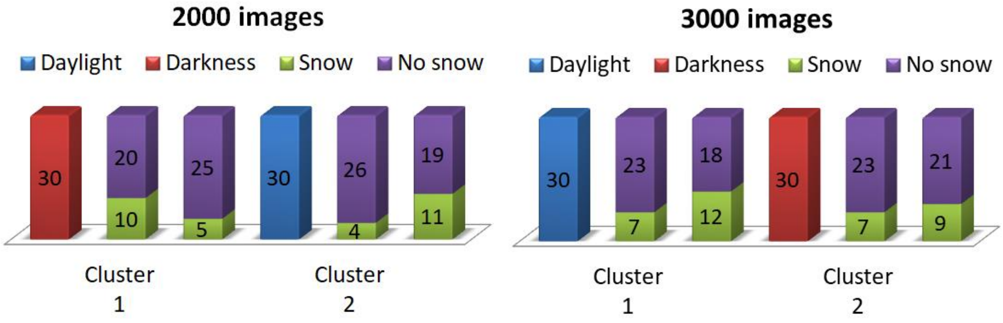

Figure 23: Manual samples taken from each cluster with Kohonen SOM

5.4.2 K-means

Clustering time, Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index for each cluster are presented in table 14. The first clustering time is 30.92 minutes, the second is 59.58 seconds, and the average is 45.25. The Silhouette Coefficients are 0.292 and 0.247, with average 0.269, the Calinski-Harabasz Index is 726 and 1176, with average 951, and the Davies-Bouldin Index has the values 1.465 and 1.578, with average 1.521.

Table 14: Clustering time and clustering performance for K-means and two clusters

Images	*Clustering time (seconds)*	*Silhouette Coefficient*	*Calinski-Harabasz Index*	*Davies-Bouldin Index*
2000	30.92	0.292	726	1.465
3000	59.58	0.247	1176	1.578
Average	45.25	0.269	951	1.521

Table 15: Clustering time and clustering performance for K-means and sub-clusters

Images	*Sub-cluster*	*Clustering time (seconds)*	*Silhouette Coefficient*	*Calinski-Harabasz Index*	*Davies-Bouldin Index*
2000	Cluster 1	65.79	0.022	40	6.015
	Cluster 2	11.98	0.110	69	2.639
3000	Cluster 1	50.45	0.027	56	5.207
	Cluster 2	34.22	0.092	151	2.883
Average		40.61	0.062	79	4.186

Clustering time, Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index for each sub-cluster are presented in table 15. The shortest clustering time is 11.08 seconds, the longest is 65.34 seconds, and the average is 40.61 seconds. The Silhouette Coefficient varies between 0.027 and 0.110, with average 0.062, the Calinski-Harabasz Index increases from 40 to 151, with average 79, and the Davies-Bouldin Index varies between 2.639 and 6.015, with average 4.186.

30 manual samples from each cluster are shown in *figure 24*. Each staple represents each cluster and corresponding sub-cluster, the blue part represents daylight, the red part represents darkness, the green part represents snow and the purple represents no snow. The clusters in both datasets have 30 samples of daylight and 30 samples of darkness images each. In the sub-clusters there are between 3 and 9 snow images and the rest represents no snow.

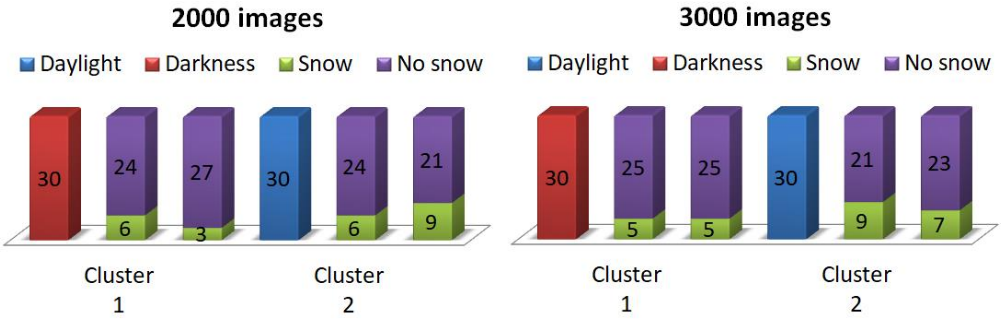

Figure 24: Manual samples taken from each cluster with K-means

5.4.3 K-means with PCA

Dimensional reduction time, clustering time, Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index for each cluster are presented in table 16. The time for dimensional reduction with PCA is 1.14 minutes and 1.98 minutes, with average 1.56 minutes. The first clustering time is 2.23 seconds, the second is 3.76 seconds, and the average is 2.99 seconds. The Silhouette Coefficients are 0.291 and 0.245, with average 0.268, the Calinski-Harabasz Index are 725 and 1170, with average 947, and the Davies-Bouldin Index has the values 1.471 and 1.586, with average 1.528.

Table 16: Clustering time and clustering performance for K-means with PCA, and two clusters

Images	*PCA time (minutes)*	*Time (seconds)*	*Silhouette Coefficient*	*Calinski-Harabasz Index*	*Davies-Bouldin Index*
2000	1.14	2.23	0.291	725	1.471
3000	1.98	3.76	0.245	1170	1.586
Average	1.56	2.99	0.268	947	1.528

Table 17: Clustering time and clustering performance for K-means with PCA and sub-clusters

Images	*Sub-cluster*	*Clustering time (seconds)*	*Silhouette Coefficient*	*Calinski-Harabasz Index*	*Davies-Bouldin Index*
2000	Cluster 1	2.34	0.022	41	5.865
	Cluster 2	0.73	0.113	71	2.600
3000	Cluster 1	5.09	0.026	55	5.189
	Cluster 2	3.55	0.095	158	2.851
Average		2.92	0.064	81	4.126

Clustering time, Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index for each sub-cluster are presented in table 17. The shortest clustering time is 0.73 seconds, the longest is 5.09 seconds, and the average time is 2.92 seconds. The Silhouette Coefficient varies between 0.022 and 0.113, with average 0.064, the Calinski-Harabasz Index increases from 41 to 158, with average 81, and the Davies-Bouldin Index varies between 2.600 and 5.865, with average 4.126.

30 manual samples from each cluster are shown in *figure 25*. Each staple represents each cluster and corresponding sub-cluster, the blue part represents daylight, the red part represents darkness, the green part represents snow and the purple represents no snow. The clusters in both datasets have 30 samples of daylight and 30 samples of darkness images each. In the sub-clusters there are between 7 and 13 snow images and the rest represents no snow.

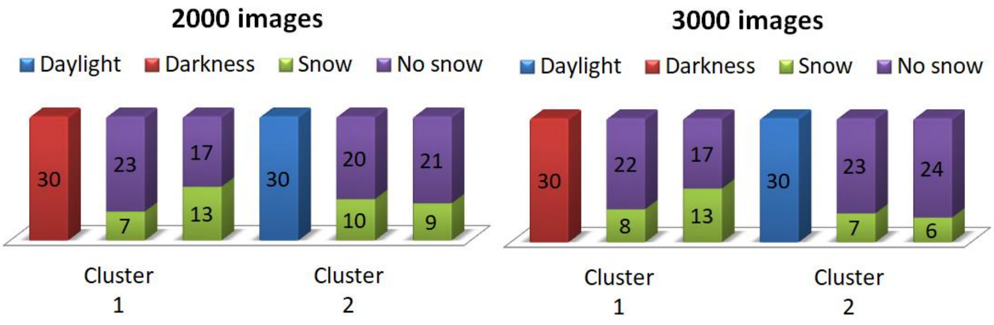

Figure 25: Manual samples taken from each cluster with K-means and PCA

5.5 Cluster graphs

In this subchapter, graphs of the clusters in chapter 5.1, 5.2 and 5.4 are presented.

5.5.1 Three clusters

Graphs of the three clusters from chapter 5.1 can be seen in *figure 26, 27* and *28.* There are four different sizes of datasets with 2000, 3000, 4000 and 5000 images. *Figure 26* represents Kohonen SOM and *figure 27* represents K-means, where the components in the data have been reduced by PCA before running the algorithms. *Figure 28* represents K-means with the original amount of components in the dataset. The blue dots in the graph symbolize cluster 1, the red dots symbolize cluster 2 and the yellow dots symbolize cluster 3.

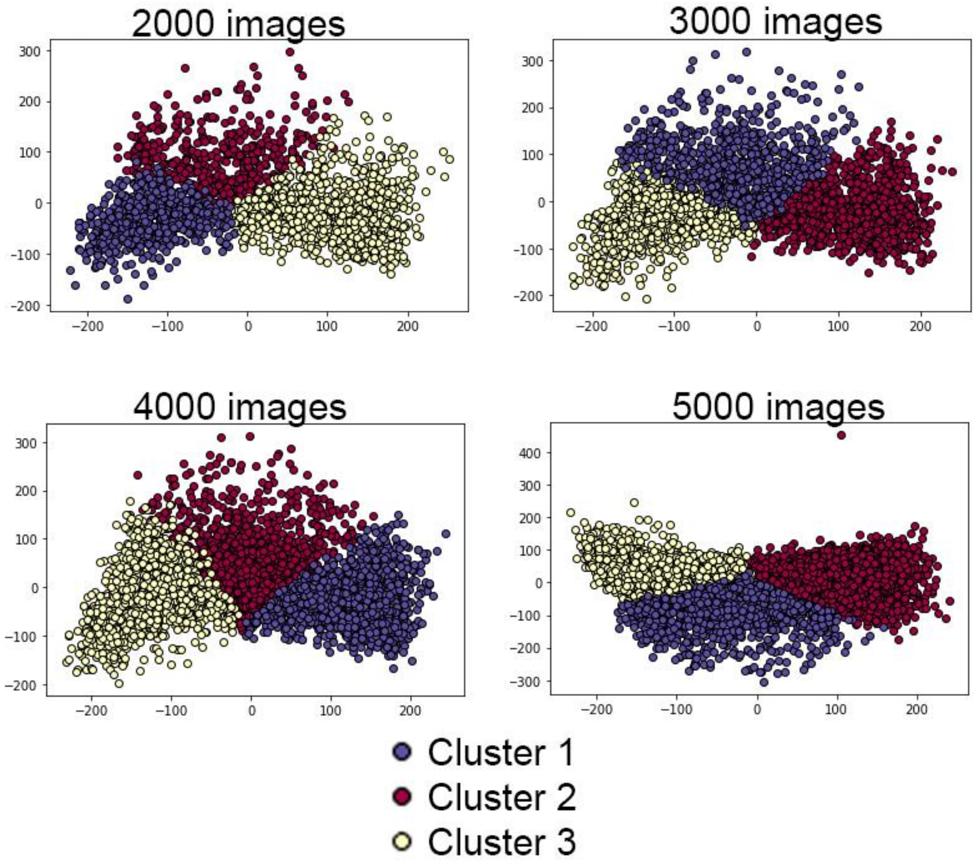

Figure 26: Graphs of three clusters with Kohonen SOM

]

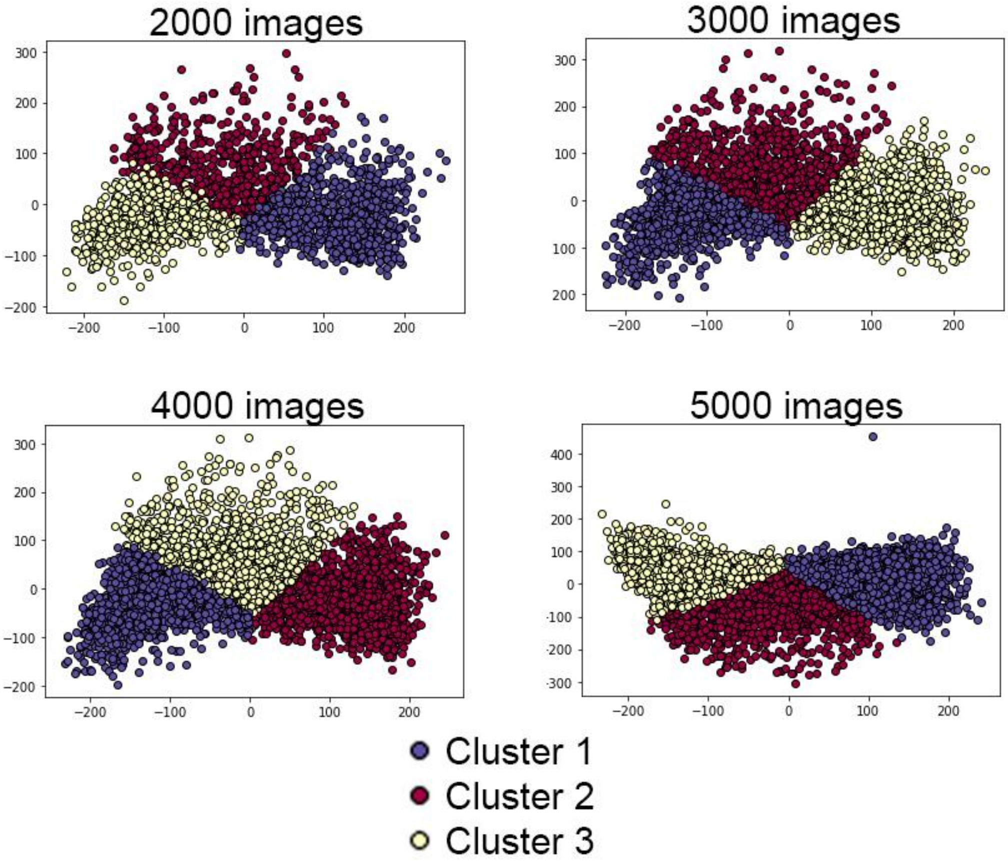

Figure 27: Graphs of three clusters with the algorithm K-means and dimension reduced dataset

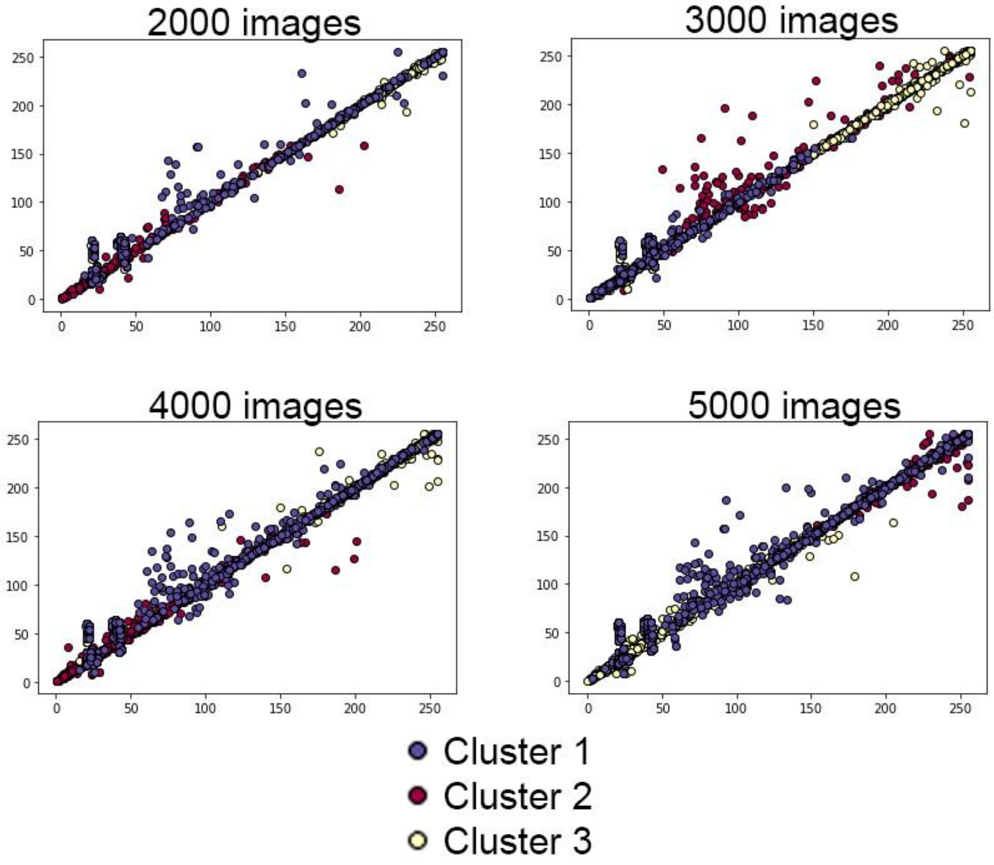

Figure 28: Graphs of three clusters with K-means

5.5.2 Two clusters

Graphs of the two clusters from chapter 5.2 can be seen in *figure 29, 30* and *31.* There are four different sizes of datasets with 2000, 3000, 4000 and 5000 images. *Figure 29* represents Kohonen SOM and *figure 30* represents K-means, where the components in the data have been reduced by PCA before running the algorithms. *Figure 31* represents K-means with the original amount of components in the dataset. The blue dots in the graph symbolize cluster 1 and the red dots symbolize cluster 2.

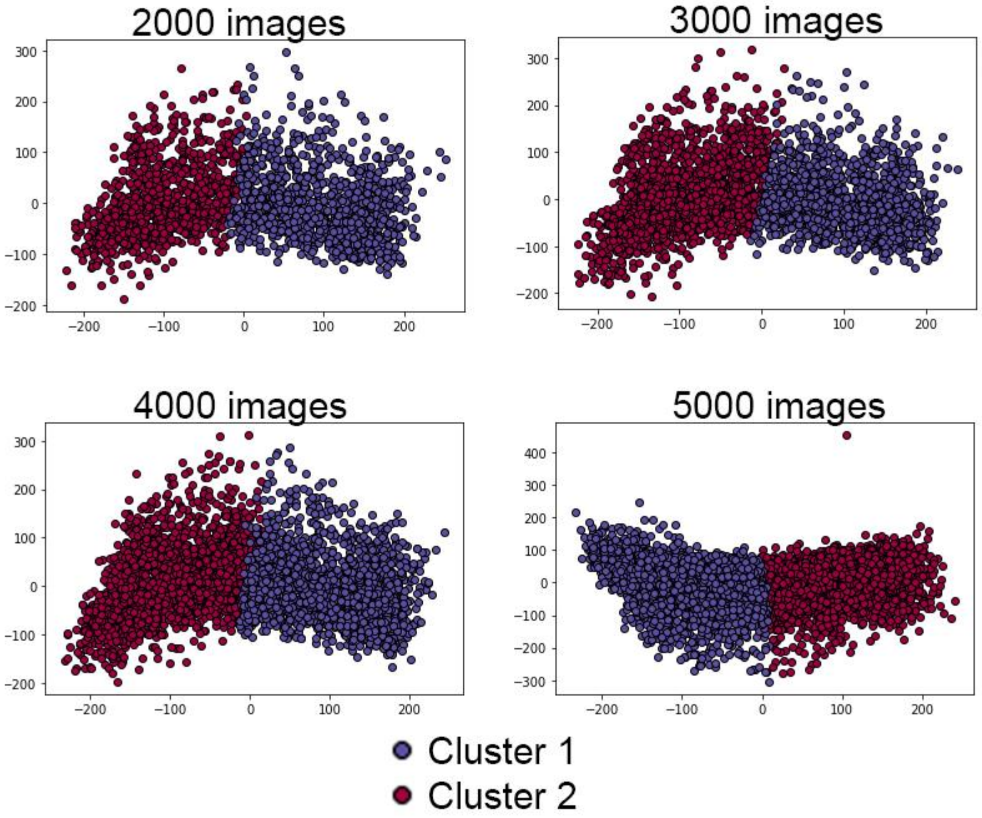

Figure 29: Graphs of two clusters with the Kohonen SOM

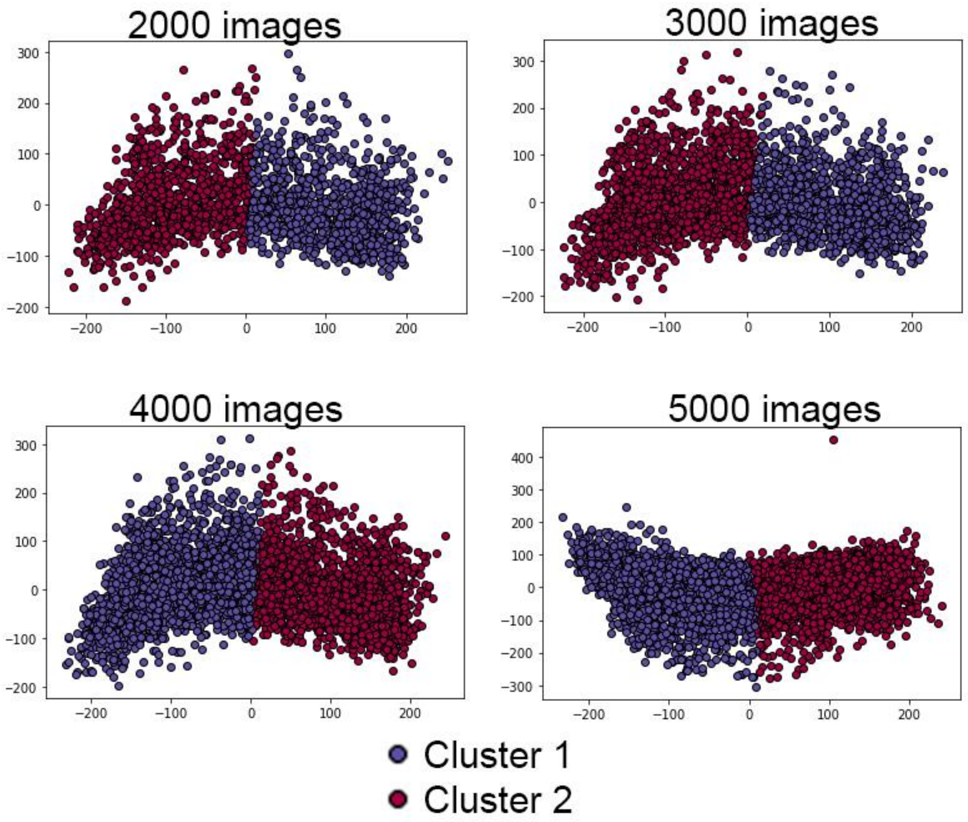

Figure 30: Graphs of two clusters with K-means and PCA

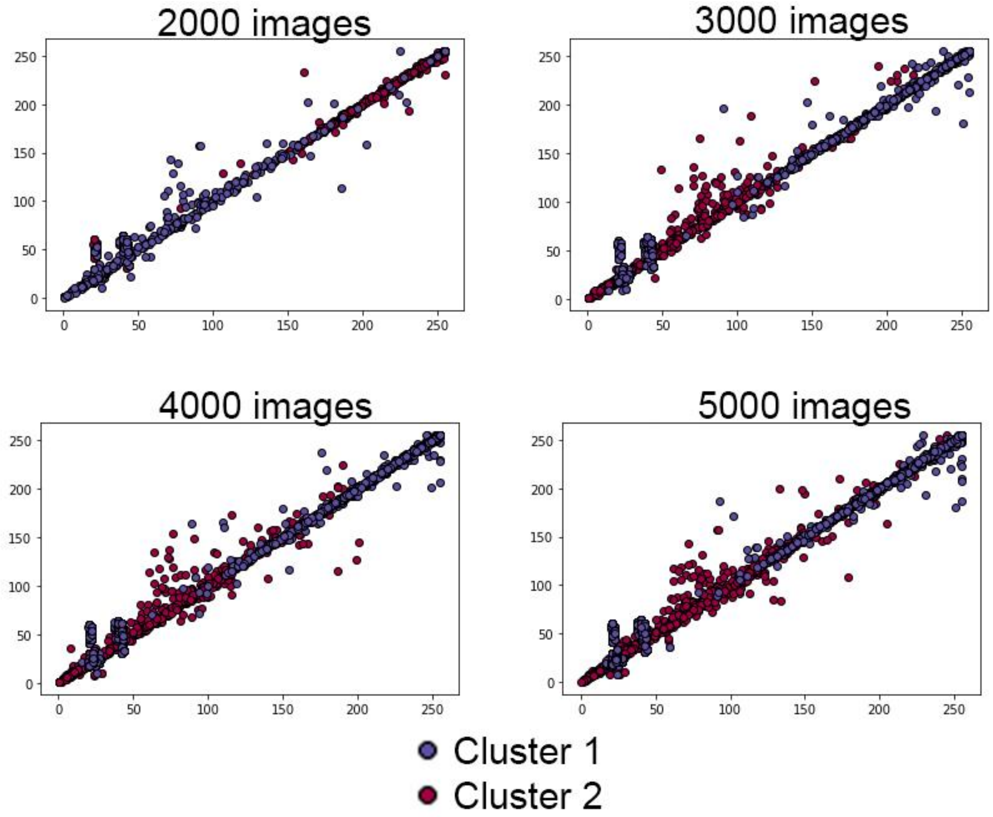

Figure 31: Graphs of two clusters with K-means

5.5.3 Sub-clusters with cropped images

Graphs of two sub-clusters from chapter 5.4 can be seen in *figure 32, 33* and *34.* The size of the dataset is 2000 images, before the first clustering, and in each cluster there are between 200-1000 images. *Figure 32* represents Kohonen SOM and *figure 33* represents K-means, where the components in the data have been reduced by PCA before running the algorithms. *Figure 34* represents K-means with the original amount of components in the dataset. The blue dots in the graph symbolize cluster 1 and the red dots symbolize cluster 2.

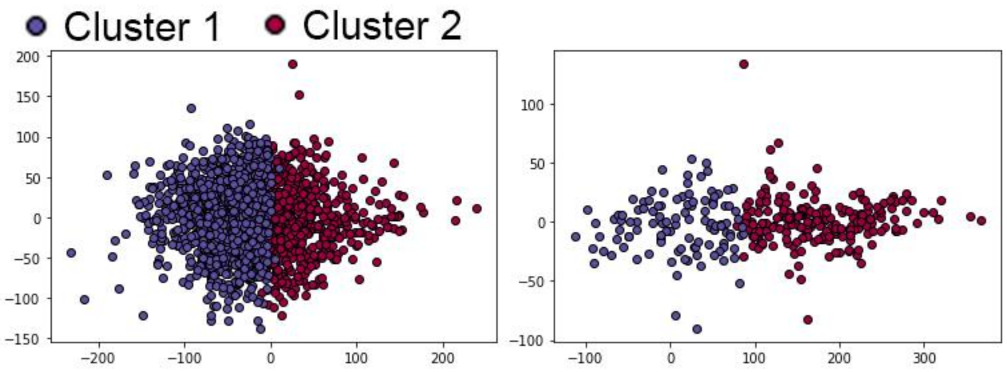

Figure 32: Graphs of two sub-clusters with Kohonen SOM

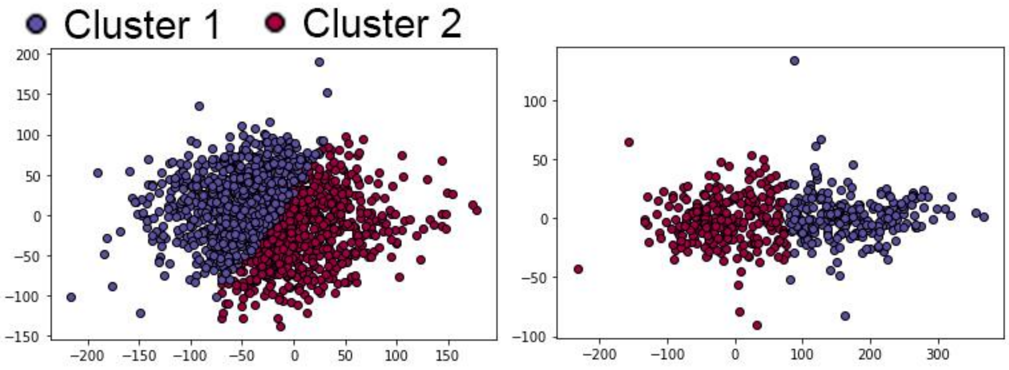

Figure 33: Graphs of two sub-clusters with K-means and PCA

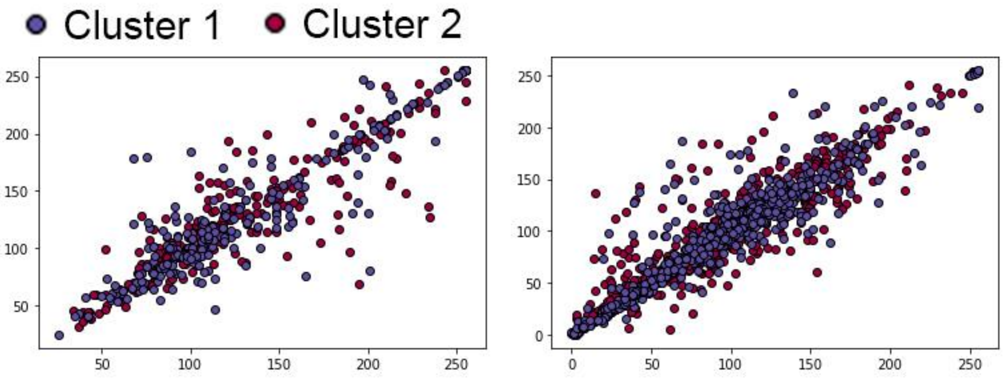

Figure 34: Graphs of two sub-clusters with the K-means

6 Discussion

In this chapter the result are discussed.

6.1 Evaluation of number of clusters

The averages of Calinski-Harabasz Index and Davies-Bouldin Index are better for all algorithms with two clusters in comparison with three clusters, which presented in Table 3-8. The Silhouette Coefficient is about 0.1 greater for the evaluation of two clusters, which means that it is a slightly worse value compared to three clusters. Based on the manual samples in *figure 14*, *15*, *16*, *17*, *18* and *19*, it can be assumed that two clusters fits the dataset better, to get the desired clusters, daylight and darkness, and then cluster two sub-clusters to get snow or no snow.

6.2 Evaluation of the dataset's size and distribution

According to Table 3-8, the Calinski-Harabasz Index increases as the number of images increases, for two and three clusters. For both the Silhouette Coefficient and the Davies-Bouldin Index, there is a less noticeable difference between each dataset, at the thousandth, and hundredth level. Table 3 and 5, each have a Davies-Bouldin Index that differs from the other data sets at the tenth level. As a result, based on the Calinski-Harabasz Index, the size of the dataset matters, while for the Silhouette Coefficient and Davies-Bouldin Index, the size of the dataset does not matter.

Based on the manual samples in *figure 14*, *15* and *16*, it can be read that, when the dataset increases in size, more darkness images are taken as samples. Since more images are retrieved from the total dataset with over one million images, there is a greater chance of including more varied images, and therefore the result of manual samples can be seen as better.

For *figures 17*, *18* and *19*, there is a less noticeable difference for the manual samples taken from K-means and Kohonen SOM, with dimensional reduced dataset, with two clusters. From the two algorithms, it distinguishes two and one darkness images for the 2000 dataset and the 5000 dataset, respectively. For K-means, the result is best for the 4000 dataset but worst for the 5000 dataset. Based on this, it is not possible to say that the larger the dataset, the better.

According to *figure 20*, *21* and *22*, the samples vary markedly, and what can be read is that there are few images with snow in all clusters. From this, it can be concluded that the distribution of daylight, darkness and snow is more important than the size of the dataset.

Figure 23, *24* and *25* shows manual samples from the dataset that only contains images from January to March, and about as many images taken during the day as at night. That could be the reason why the samples show the best of all models according to daylight and darkness. The samples for the sub-clusters, showing snow or no snow, do not indicate on a good desired result. According to the high amount of samples with no snow it can be suggested that there are not enough snow images for the algorithms to be able to cluster the desired categories. Although the images are taken between January and Mars, there can have been lack of snow during the winter 2019. Also, the images are taken on places throughout Sweden, and there is less snow in the south comparing to the north.

6.3 Evaluation of clustering time

Based on the results in Table 3-8, it is clear that Kohonen SOM with PCA takes significantly longer than K-means with and without PCA. The clustering time with Kohonen SOM increases twice as much when increasing the dataset, which corresponds to an exponential time distribution. K-means increases by approximately a few seconds up to one minute, which can correspond to a polynomial time distrubution. For K-means with PCA, there is only a few single seconds difference in the clustering time.

When comparing Kohonen SOM and K-means with the reduced dataset, the latter is much more time efficient. In the case of K-means with and without PCA, the clustering time is shorter with PCA, but the time to reduce the dimensions leads to that the total time is still longer than without PCA.

6.4 Evaluation of manual samples

The manual samples have been taken by the author who had no experience of examining images of timber, before the thesis. Because of this, it cannot be assumed that the results show 100% reliability, but rather it is an indication of how approximately the images have been clustered for the desired categories.

6.5 Evaluation of graphs

What can be read from *figures 26, 27, 28, 29, 30* and *31*, is that it looks like there are slightly smaller outliers with larger datasets. This is most evident with the reduced dataset because the components are, in that case, orthogonal compared to the original dataset, which is linear. Based on the graphs, it appears that there are more overlaps with the original dataset, while with reduction, the clusters are more well defined.

What can also be seen is that K-means and Kohonen SOM choose to cluster the data points in a different order and the boundary between the clusters differs slightly, which in turn leads to the differences in the clustering performance methods.

When examining *figure 32, 33* and *34*, there are significantly more outliers in those graphs, which also contributes to Silhouette Coefficient, Davies-Bouldin Index and Calinski-Harabasz Index giving a worse result when clustering sub-clusters. This may be largely due to the fact that when 2000 images were clustered in two rounds, there were only a few 100 images in the sub-clusters, which could make it more difficult for the algorithms to find similarities in the data. To cluster sub-clusters, a larger dataset would probably have been needed to get a larger amount of images to cluster in the sub-clusters.

When there are a large number of outliers, the manual samples are not as reliable as it may give an incorrect insight in the cluster. Outliers can have properties other than the data points that have been clustered together in a more well-defined cluster.

6.6 Comparison of K-means and Kohonen SOM with PCA

The clustering time for Kohonen SOM is significantly longer compared to K-means. The average clustering times with Kohonen SOM are between 9-48 minutes, compared to K-means which have average times between 2-9 seconds. Because of this, K-means is considered to be a more time-efficient algorithm than Kohonen SOM.

When comparing the average values in table 3, 5, 6, 8, 9, 11, 12, 13, 16 and 17, there are small differences between the two algorithms. The average Silhouette Coefficient and Davies-Bouldin Index is sometimes a few thousandths or hundredths better for any of the algorithms. The average Calinski-Harabasz Index is about 1-2% better with K-means. Due to that, it is difficult to decide which algorithm is best, based on these methods. Based on related work, K-means has performed better than Kohonen SOM. This may have been due to the fact that they have a different type of data set and that they have been able to perform more extensive tests.

Considering the manual samples in *figure 23* and *25* there are a few more snow images which have been taken as samples. In addition, there is very little difference in all manual samples for the two algorithms, which makes it not possible to use manual sampling as a method to determine which algorithm is most accurate, because it is too little difference.

6.7 Comparison of K-means with and without PCA

The clustering time for K-means, with dimensionality reduction, by PCA, are shorter than without. The average clustering time for the dataset with reduced dimensions is between 2-9 seconds, compared to the original dataset where the average time is around 1-3 minutes. Something that must not be forgotten to take into account is the dimensional reduction time on the dataset, which is between 1-9 minutes depending on the size of the dataset. If many different clusters are to be created with the same dataset, time can be saved with PCA, but if it is only to be clustered once, the total time is longer with PCA.

When studying the table 4, 5, 7, 8, 10, 11, 14, 15, 16 and 17 it can be seen that K-means with the original data delivers slightly better results than with the reduced data. A few average values are a little bit worse with the original data, but based on all results, it can be said that K-means without PCA generates slightly better results according to Silhouette Coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index.

From the *figure 15*, *16*, *18*, *19*, *21*, *22*, *24* and *25* it can be concluded that there is too little difference between the manual samples to be able to decide that one is better than the other.

6.8 Tools and dataset

The cluster analysis was running on the platform, Microsoft Azure, which Biometria uses. There was limited memory, which meant that it was only possible to cluster 5000 images at one and the same time. When clustering the sub-clusters in chapter 5.4, two datasets were loaded simultaneously, one with the original images and one with cropped images. This meant that 3000 images in each dataset was the maximum limit before the platform crashed. Due to that, it became a limitation in the work, as it became more difficult to determine how important the size of the dataset was for that type of model.

The outcome depends largely on the data. If it is only a small proportion of e.g. snow images, the algorithms choose to cluster based on other parameters. In this case, the total dataset extends from April 2019 to September 2020. During the winter 2019-2020, there was not much snow, which may contribute to the fact that there were not enough images depicting snow on the load. Therefore, the hard part is to give clear data as an input, thus the computer can get the best conditions to cluster in desired categories.

6.9 Ethical and Societal Discussion

This study has aimed to investigate whether it is possible to classify images of daylight, darkness, and snow on a load of timber trucks. If it were possible, the new data could have been used to re-train existing ML models in order to improve the results. The results of the existing models are used today as support in remote measurement when assessing the volume of timber. The models have not replaced human labor, but have existed as a support in the assessment. If the models could be developed and become even better, there is a risk that the physical measurements will be reduced in number, and the measurement will be handled even more remotely, via remote measurement. Nevertheless, it will not replace the human-in-the-loop in the near future because it still requires remote timber meters, which assesses more than just the volume. If the models were good enough and could become independent, random sampling would still be required both at a distance but also on site, which means that humans would still be needed to control the system. Jobs will also be created for system developers and other IT jobs, to maintain the intelligent systems.

With the help of better ML models, efficiency can increase because the measurements are made faster than with physical and manual work. This means that the timber trucks would not need to be stationary for the same length of time, or even better, may not even have to stop. If it were possible, it would be both more time-and cost-effective.

6.10 Future Work

To continue the investigation, if it is possible to classify unlabeled images, the next step could be to clear out all the images that represent birch trees, to sort out white spots in images that could be mixed with snow. Another dataset could be used, with images from a winter that is known to have had a lot of snow to investigate if it gives better results. If it gives good enough results, the images would be labeled with information. Then supervised learning can be used to train a model in daylight, darkness, and snow on the load, to really show the computer what is desired. Hopefully an accurate model can be created, which can then be used to determine daylight, darkness and snow on the load for all existing and upcoming images.

In order to continue the study of whether partitional clustering or competitive learning is better, the size of the dataset and the distribution, need to vary more. This is to be able to investigate how the two different methods behave with more varying conditions and to get a broader basis from which to draw conclusions.

7 Conclusion

In the beginning of the thesis, three research questions were mentioned. To answer P1, cluster analysis can be used to classify unlabelled images in the categories daylight and darkness. According to the result and discussion, it could be possible to classify snow on the load. For this to be possible, a cleaner dataset with less birch trees and other white spots that can be mistaken for snow, is required.

The second question, P2, addresses the comparative study between partitional clustering and competitive learning. K-means was chosen as the partitional clustering algorithm, and Kohonen SOM was chosen as the competitive learning algorithm. What can be determined is Kohonen SOM’s clustering times, which are significantly longer than K-means. According to the three clustering performance methods, Silhouette Coefficient, Davies-Bouldin Index and Calinski-Harabasz Index, there a small difference between the two algorithms. In some cases, Kohonen SOM generates better outcomes, and in other cases K-means generates better outcomes. This may be because Kohonen SOM may be better at clustering complex datasets, and K-means is worse at clustering when the data is more varied in size and distribution. Because of this, more research needs to be done with more varied data sets, in both size and distribution, to gain a greater understanding of how both the algorithms behave.

To answer the question P3, K-means without dimensionality reduction has a longer clustering time, but generates slightly better result than with the reduction, according to the three clustering performance methods, Silhouette Coefficient, Davies-Bouldin Index and Calinski-Harabasz Index. When it comes to clustering time, dimensionality reduction can be an advantage, but if clustering performance is the most important, it is better to use the original dataset.

References

[1] S. Dipanjan, B. Raghav and S. Tushar, “Practical Machine Learning with Python”, Springer Science + Business Media, New York, 2018

[2] Biometria, “Om Biometria”, https://www.biometria.se/om-biometria/, Retreived 2021-01-21

[3] Forefront Consulting, “Det här är Forefront”, https://www.ffcg.se/om-oss/ , Retreived 2021-01-21

[4] Linda Åstrand, Forefront Consulting

[5] Kari Hyll and Maria Nordström, “Review of technology, methods, and information flows for measuring products from forestry”, Uppsala Science Park, Uppsala, 2020

[6] Alboukadel Kassambara, “ Practical Guide To Cluster Analysis in R, Unsupervised Machine Learning”, STHDA, 2017

[7] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects”, Science vol. 349, 2015, pp. 255-260

[8] R. Saravanan and Pothula Sujatha, “A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification” Second International Conference on Intelligent Computing and Control Systems, 2018, pp. 945-949

[9] Abass Olaode, Golshah Naghdy and Catherine Todd, “Unsupervised Classification of Images: A Review”, International Journal of Image Processing (IJIP), vol. 8, 2014, pp. 325-342

[10] Sammut C., Webb G.I, “Competitive Learning” Encyclopedia of Machine Learning and Data Mining, Springer Science+Business Media, New York, 2017

[11] S. Anitha Elavarasi, Dr. J. Akilandeswari and Dr. B. Sathiyabhama, ”A survey on partition clustering algorithms”,

International Journal of Enterprise Computing and Business Systems (Online), vol. 1, 2011, pp. 1-13

[12] Romesburg, H. Charles, “Cluster analysis for researchers”, Lulu Press, North Carolina, 2004

[13] Joaquín Pérez-Ortega et al,” The K-Means Algorithm Evolution, Introduction to Data Science and Machine Learning” Keshav Sud, Pakize Erdogmus and Seifedine Kadry, IntechOpen, 2019

[14] Zhang J, “Kohonen Self-Organizing Map–An Artificial Neural Network”, Visualization for Information Retrieval. The Information Retrieval Series, vol. 23, Springer, Berlin, Heidelberg, 2008

[15] C.O.S. Sorzano, J. Vargas, A. Pascual-Montano, “A survey of dimensionality reduction techniques”, Natl. Centre for Biotechnology (CSIC), Madrid, 2014, 35 pages

[16] Hervé Abdi and Lynne J. Williams, “Principal Component Analysis”, Wiley Interdisciplinary Reviews: Computational Statistics, vol. 2, 2010

[17] Sasan Karamizadeh et al., “An Overview of Principal Component Analysis”, Journal of Signal and Information Processing, vol. 4, 2013, pp. 173-175

[18] Laurens van der Maaten, Eric Postma and Jaap van den Herik, ” Dimensionality Reduction: A Comparative Review”, Tilburg centre for Creative Computing, Tilburg University, The Netherlands, 2009

[19] Shuhan Luan et al.,”Silhouette coefficient based approach on cell-phone classification for unknown source images”, IEEE International Conference on Communications (ICC), 2012

[20] Xu Wang and Yusheng Xu, “An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index”, Materials Science and Engineering, vol. 569, 2019

[21] Slobodan Petrovic, “A Comparison Between the Silhouette Index and the Davies-Bouldin Index in Labelling IDS Clusters”, Gjøvik University College, Norway

[22] Microsoft Azure, “Popular questions about Azure”, https://azure.microsoft.com/en-us/overview/what-is-azure/#most-popular-questions, Retrieved 2021-02-03

[23] Microsoft, “What is Azure Databricks?”, Published 2020-10-04, https://docs.microsoft.com/en-us/azure/databricks/scenarios/what-is-azure-databricks Retrieved 2021-02-0

[24] Databricks, “Data Lake”, https://databricks.com/glossary/data-lake, Retrieved 2021-02-03

[25] Sebastian Raschka, “Python Machine Learning Equation Reference”, Packt Publishing, 2015

[26] Fabian Pedregosa et al., “Scikit-learn: Machine Learning in Python”, Journal of Machine Learning Research vol. 12, 2011

[27] Martin Abadi et al., “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems “, 2015-11-09

[28] Travis E. Oliphant, “Guide to NumPy”, 2006-12-07

[29] Wes McKinney, “Pandas: a Foundational Python Library for Data Analysis and Statistics”, 2011

[30] Mohamed R. Ibrahim, James Haworth and Tao Cheng, “WeatherNet: Recognising Weather and Visual Conditions from Street-Level Images Using Deep Residual Learning”, ISPRS Int. J. Geo-Inf, vol. 8, nr. 12, page 549, 2019

[31] Tomasz Krzywicki, “Weather and a part of day recognition in the photos using a KNN methodology”, Technical Sciences, vol. 21 nr. 4, 2018,pp. 291–290

[32] Usha A. Kumar and Yuvnish Dhamija, “Comparative Analysis of SOM Neural Network with K-means Clustering Algorithm “, IEEE International Conference on Management of Innovation & Technology, 2010, pp. 55-59

[33] Calle Finnström, Sven Jägbrant and Peter Lenaers, ”Mätning av travmått med hjälp av objektdetektering”, 2018, 25 pages

[34] Lars Björklund, Sven Jägbrant, Tanja Keisu and Peter Lenaers, ”AI stödd travmätning baserad på travbilder, vikt och skördardata”, 2020, 18 pages

[35] Sofie Hellmark, Sven Jägbrant, ”ASTA 2.0 - Volymmätning av virkestravar”, 2020, 15 pages

[36] F. Pedregosa, et al., ”Scikit-learn: Machine Learning in Python”, Journal of Machine Learning Research 12, 2011

[37] Markus Ringnér, “What is principal component analysis?”, Nature Biotechnology, vol. 26, nr. 3, 2008

[38] Renato Pelessoni Liviana Picech, “Simple competitive learning. Kohonen S.O.M and Schmitter-Straub’s method: A proposal in finding “good” tariff classes”, Dipartimento di Matematica Applicata alle

[39] Ujjwal Maulik and Sanghamitra Bandyopadhyay, “Performance Evaluation of Some Clustering Algorithms and Validity Indices”, IEEE transactions on pattern analysis and machine intelligence, vol. 24, nr.12, 2002

[40] Michael J. Watts and S.P. Worner, “Estimating the risk of insect species invasion: Kohonen self-organising maps versus k-means clustering”, Ecological Modelling, vol. 220, 2009, pp. 821–829

[41] Fedor Krasnov and Anastasiia Sen, “The Number of Topics Optimization: Clustering Approach”, Machine learning & knowledge extraction, 2009, pp. 416-426

[42] Usha A. Kumar and Yuvnish Dhamija, “Comparative Analysis of SOM Neural Network with K-means Clustering Algorithm”, IEEE ICMIT, 2010

[43] Alex Tay Leng Phuan and Sandeep Prakash, “K-means fast learning artificial neural network, an alternative network for classification”, International Conference on Neural Information Processing, vol. 2

[44] Yiheng Chen et al., "The Comparison of SOM and K-means for Text Clustering", Computer and Information Science, vol. 3, nr. 2, 2010

[45] Isaiah Hull, "Machine Learning for Economics and Finance in TensorFlow 2", Apress, Berkeley, CA, 2021

[46] Polina Lemenkova, "Processing oceanographic data by Python libraries Numpy, Scipy and Pandas", Aquatic Research, vol. 2, nr. 2, 2019, pp. 73-91

[47] Scikit-learn, "sklearn.cluster.KMeans", https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html Retrieved 2021-05-31

[48] Scikit-learn, "sklearn.decomposition.PCA", https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html#sklearn-decomposition-pca, Retrieved 2021-05-31

[49] Github, "Clust", Published 2017-10-21, https://github.com/ArtemKovera/clust/blob/master/Kohonen%2Bnetwork%2B.ipynb, Retrieved 2021-03-20