MASTER THESIS

Uncovering smallholder patterns in agricultural- water-soil characteristics: an unsupervised machine learning method.

Image by (Food+Agri Business, 2021)

Jasper Kouters
S1844024
25-01-2024

UNIVERSITY
OF TWENTE.

Uncovering smallholder patterns in agricultural- water-soil characteristics: an unsupervised machine learning method.

Master Thesis
25-01-2024

Author:
Jasper Kouters
Student - Water Engineering and Management

Supervisors:
Dr. M.S. Krol
Associate Professor – Multidisciplinary Water Management

H. Su MSc
PHD Candidate - Multidisciplinary Water Management

Preface

In front of you lies my Master Thesis ‘Uncovering smallholder patterns in agricultural- water-soil characteristics: an unsupervised machine learning method.’ The thesis marks the completion of my master’s degree in Water Engineering and Management at the University of Twente. The thesis is conducted as internal research within the Multidisciplinary Water Management department.

Throughout my research, I have had the support from many people around me, for which I am incredibly grateful. I would like to thank my supervisors from the University of Twente, Maarten Krol and Han Su for the valuable feedback and support throughout my master thesis. I appreciate the time and effort they both put into my research. The weekly meetings were always pleasant and informative, their mentorship is much appreciated. Furthermore, I would like to thank my fellow students with whom I work together in the past six year of my study.

I would like to express my profound gratitude to my friends and family, especially my parents whose support has been a constant source of strength and assistance throughout my whole school career. A special thanks to my girlfriend, who helped me through the difficult periods I had to deal with and during all the ups and downs of the research.

I hope you enjoy reading this thesis.

Jasper Kouters

Summary

Farmers play a critical role, particularly smallholders, in achieving agricultural sustainability around the world. The impact of water and soil resource management is acknowledged in agricultural behaviour of farmers. By getting a better understanding of the agricultural behaviour of farmers around the world could impact future driven, sustainable policy strategies. The research focuses on the combination of agricultural, water and soil resources to uncover smallholder patterns. An unsupervised machine learning algorithm, namely K-means, is applied to identify patterns creating inside in spatial variation and potential hotspot for change.

The research aims to explore whether smallholder patterns can be identified and what the three most important factors, called dominant factors, are which drive the creation of patterns. The three resources are presented by agricultural data, namely farming class (smallholders or non-smallholders), the crop type (12 different groups) and farming system (irrigation or rainfed), water data, namely water scarcity, ground water level and evapotranspiration, and soil data, namely the nutrient availability, terrain slope and global soil organic carbon. With an addition of temperature and precipitation as climate data, which is required in research with spatial data.

The data is transformed to 30 unique features by means of feature engineering and scaled in order to make them comparable. An unsupervised machine learning algorithm K-means is applied in order to identify the smallholder patterns within the data. Five different clusters are identified: the first in Europe and east part of the USA, the second in central Asia, the third in South America, the fourth in the western part of USA and Mexico and the fifth in India and central parts of Afrika. During the creation of the pattern, three dominant factors are identified. The dominant factors are the water scarcity, ground water level and Global Soil Organic Carbon (GSOC). The dominant factors are used to understand the clusters and in order to describe the differences between the clusters.

In the discussion the robustness for the choice of the number of clusters and the choice of random state used by the algorithm is assessed. The discussion also touches on the possible use of alternative data, which could be used to increase a more detailed description of the patterns. Finally, the suitability of the smallholder definition, used in the research, is discussed.

Table of Contents

Table of Figures

Table of Tables

1. Introduction

1.1 Background

Farmers are at core of sustainability in the agricultural sector. The implementation of a more sustainable agricultural production is in hands of the farmers that produce all the consumer goods. However, farmers are not a homogenous group. In order to make farmers more sustainable in farming, access to knowledge and funding while applying the same methods of production could lead to more sustainability. The literature mostly identified two different type of farm groups, based on their size, namely smallholders and non-smallholders. Unfortunately, there is no worldwide general accepted definitions that distinguished both from each other (Ergin, Conforti, & Khalil, 2019). Every country or region choose their own definition of smallholders considering their own perspective. The United Nations has performed research where the definition of smallholders all over the world are compared (Khalil, Conforti, Ergin, & Gennari, 2017). The research concluded that is no 'right' definition and it will depend on the purposes of the analysis. However, the World Food and Agriculture Organization of the United Nations used the definition of a farm size below 2 hectares while constructing the Sustainable Development Goals (SDG's). When taking this definition into account, around 84% of the total number of farms are considered smallholders and which operate around 12% of the world's agricultural land (Lowder, Skoet, & Raney, 2016). Smallholders are considered important players in the future of agriculture around the world (Giller, Delaune, & Silva, 2021)

Farmers make use of the land they are farming on and water by either rainfall, irrigation or a combination of both. But the allocation and management of water and land resources for the smallholders are different all over the world and mostly regional and national diverse. The water consumption rate differs between planted crops, also the amount of needed nutrients from the soil is different for every crop. It is expected that the characteristics of water and soil in a certain area will affect the crop choice by the farms. The research of (Giordano, Barron, & Unver, 2019) and (Aguilar, Hendrawan, Cai, Roshetko, & Stallmann, 2020) illustrates the water scarcity for smallholder agriculture. The paper states that smallholder agriculture will play key role in the increased food demand the world is facing in 2050. The irrigated agriculture account for almost 70% of the total freshwater withdrawals globally while these withdrawals already exceed planetary boundaries by more than 10%. Water scarcity mostly happen in densely populated area and areas will a dry or semi dry climate where rainfall is rarely happening. The biggest driver of water scarcity is the mismatch between freshwater demand and availability on a geographic and temporal scale (Mekonnen & Hoekstra, 2016). The research of (Chikowo, Zingore, Snapp, & Johnston, 204) shares inside in the soil fertility and nutrient management by smallholder farms. Its states that the current nutrient management across diverse farms is not adequate to intensify sustainable crop production but it needs a farm specific condition. The identification of different soil fertility classes can result in different targeting soil fertility management technologies for the present farms. This approach requires an adequate representation of the combination between smallholders and soil characteristics.

As describe above, smallholders are important for future agriculture. However, not all farmers operate in the same manner but still there are similar patterns of agricultural production behaviour (Guarín, et al., 2020). A pattern can be described as a consistent structure or a regular reoccurring sequence according to the (National Dictionary, 2022). One of the characteristics of a pattern is that the objects within the pattern have similar attributes that separates them from other objects. An example of a pattern can be a group of smallholder, spatially diverse, with the same amount of nutrients in the soil or the similar amount of available water for irrigation. The farms operate with the same natural resources but are located at different parts of the world. In this case it is about the combination of smallholders, land use and water use, in terms of different characteristics. By getting to know what the most important characteristics are, specific changes in policies for farmland can be made. Which opens opportunities for more sustainable farming.

A changing and uncertain future climate, a rapidly growing population that creates an increase in social and economic development will require adequate policies that are future driven and will help managing the natural resource to its full capacity. By doing research on the effect of the natural resources, in terms of water and soil characteristics, and human agriculture, in terms of crop choice and effectiveness, creates an opportunity to develop sustainable agriculture for future generation to come (Cosgrove & Loucks, 2015).

1.2 Problem Context

Research performed in the field of smallholders indicates that relation exists between smallholders and either water characteristics or soil characteristics. However, research to the combination of smallholders characteristics (agriculture), water characteristics (water) and soil characteristics (soil) is limited. Potential patterns in the combinations of agriculture, water and soil, and the dominant factors underlying these patterns can give insight in spatial differences between regions and hotspots where change is possible. Identification of these spatial differences and/or hotspots for change can help in developing specific policies with the identified locations in order to grow in sustainability.

1.3 Research Objective and Questions

The aim of the research is to identify representative patterns of agricultural- water and soil combinations in smallholders' agriculture. Simultaneously getting an insight in the dominant factor of the identified smallholder patterns. The large spatial spread of the data causes the application of an unsupervised machine learning algorithm in order to identify the smallholder patterns.

In order to reach the aim of the research, the following research questions will be answered:

- Can smallholder patterns be identified?
- What are the three dominant factor that are characteristic for a pattern?

1.4 Readers guide

The outline of the report will consist of five chapters, starting with the Introduction where this is a part of. The Methodology is explained in Chapter 2 containing a theoretical elaboration on machine learning, the collected data and the performance of the algorithm. The Results described Chapter 3 will contain the results from the described procedure in the Methodology. A discussion is performed on the limitations and uncertainties of the research in Chapter 4 while a Conclusion and Recommendation for future research is posted in Chapter 5.

2. Methodology

The chapter Methodology will contain the research method which is performed to answer the research questions. The first part will lay the grounds on machine learning, following with the data collection and preparation. The data divided into characteristics in terms of agricultural, water and soil. The agricultural characteristics are chosen based on previous work from the supervisor of the research, while both water and soil characteristics are chosen based on reports from the United Nations which identifies potential relations between the agricultural sector and other subject. The data availability for the characteristics determines the opportunities of implementation in this research. More details are given in the separate section. In the aim of the research, it is stated that a machine learning algorithm will be applied in order to uncover potential patterns. A distinction is made between supervised and unsupervised machine learning, and within the categories are multiple available algorithms. More details about the differences in application and attributes between the two categories are described in the following sections.

2.1 Introduction to machine learning

Before the research process is described, a clarification in terms of machine learning must be made. Machine learning is part of data science and according to (Berry, Mohamed, & Yap, 2020) it is a subset of artificial intelligence, which uses computerized techniques to solve problems.

Machine learning algorithms comes in multiple forms, but the most known types of machine learning are either supervised, labelled datasets, or unsupervised, unlabelled dataset. As mentioned before, the biggest difference between the two different type of algorithms is the existence of labels in the dataset (Berry, Mohamed, & Yap, 2020). Labels or attributes are predetermined results, which satisfy a hypothesis, which can be used to train an algorithm.

Data labelling, in combination with features, function as input for the application in supervised machine learning algorithms. The algorithms learns from the target labels in combination with the input features in order to make an accurate prediction on new, unseen input data (Igual & Sequí, 2017). A benefit of supervised machine learning is it gives explicit feedback on the model predictions because of the labelled data, and it makes high accuracy predictions if the model is trained well. One of the drawbacks of supervised machine learning are the data labelling requirements which can be expensive and time consuming for large datasets. Another drawback is the limitation of the labelled dataset if the new dataset is not in line with the trained data for the model (Berry, Mohamed, & Yap, 2020).

In contrast to supervised machine learning, unsupervised machine learning does not make use of labels or attributes. The algorithms trying to discover hidden patterns, structures or relations within the data. Without making use of explicit guidance in the form of labelled data. A benefit of unsupervised machine learning that it can uncover hidden patterns and exploration of complex dataset by identifying trends or potential interesting areas. While a drawback is the absence of ground truth to evaluate the quality of the model and level of subjectivity cause by assumptions which does not lead to a correct solution.

Both machine learning techniques are quite different from each other, therefor not comparable for the benefits and drawbacks, and serve their own purposes in the realm of artificial intelligence. Depending on the purpose of the research and the layout of the data, both machine learning techniques have their own value. For this research, an unsupervised machine learning algorithm will be applied on the final data set. More details on the different algorithms, the basis and their performance are elaborated on in section 2.4.

2.2 Data collection

2.2.1 Agricultural characteristics

The agricultural characteristics that are used for the research are the farm classes, crop types, harvest area and farming system. The data is extracted from a paper authored by the supervisor of the research H. Su (Su, Willaarts, Luna-Gonzalez, Krol, & Hogeboom, 2022) which is part of his PHD research. A SPAM-based dataset and the supplementary data from the paper will be the core structure of the dataset creation process. The farming classes are different indications of how large a farm is, expressed in hectares. The data is used to distinguish smallholders and non-smallholders from each other. The crop type is an abbreviation of the planted crop on that specific farm described in 42 different crops, which will be grouped in 10 groups. The harvest area is the number of hectares a farm can used for planting the crops, which is different from the farm class which describes the total area of a farm. The farming system is the kind of water system present at the farm, make they use of irrigation technology or are they dependent on rainfall. The data, related to the paper, also contains location-based coordinates, the longitude and latitude are in a 5-arcmin format and form the core structure of the created dataset.

The SPAM-based data contains separate CSV files which are merged into one file, by the means of script in the programming language Python. The dataset will contain contains the columns latitude, longitude, farming class, harvest area, crop type and farming system. The data within the file needs to be grouped in order to give a better representation of the characteristics within the given location and to prepare the data for the feature engineering.

The column of farming class and crop type are grouped, the grouping schema can be found in Appendix A, Table 10 and Table 11. The crop types are grouped based on the supplementary data from the paper (Su, Willaarts, Luna-Gonzalez, Krol, & Hogeboom, 2022), while the farming class is grouped based on the criteria for smallholders. The criteria used for the research is in line with the definition of the Sustainable Development Goals (SDG) of the United Nations, namely: the harvest area is equal or less than two hectares (United Nations, 2017). The agricultural characteristics will consist of different type of data, namely coordinates on a 5arcmin grid, class data and integer data. Class data are the farming system, smallholders and non-smallholders, crop type, ten groups, and farming system, irrigation or rainfed. The characteristics are summed up in the Table 1, including the data type and source.

Table 1 - Summarized agricultural characteristics including data type and source

Data	Type	Source
Longitude & Latitude	Coordinates	(Su, Willaarts, Luna-Gonzalez, Krol, & Hogeboom, 2022)
Farming Class	Class	
Harvest Area	Integer	
Crop Type	Class	
Farming System	Class	

2.2.2 Water characteristics

The water characteristics that are used within the research are water scarcity, evapotranspiration and groundwater. The characteristics are chosen based on the report of the Food and Agriculture Organization of the United Nations (FAO, 2022), which identifies a list of factors that may have an impact on the agricultural sector. The choice of the characteristics is made, in combination with the data availability for the different factors from the report.

The water scarcity data is related to the paper of (Mekonnen & Hoekstra, 2016), which performed research to the freshwater scarcity in a global setting. The water scarcity is the result of the consumption divided by the available water after the environment flow requirement in a specific location. For the interpretation of the water scarcity it is classified in four ranges: low (WS < 1.0), moderate (1.0 < WS < 1.5), significant (1.5 < WS < 2.0) and severe (WS > 2.0) (Mekonnen & Hoekstra, 2016). The groundwater data is related to the paper of (Fan, Li, & Miguez-Macho, 2013), which performed research on global pattern for water table depth. The groundwater level is the depth between the surface and water saturated aquifers. Shallow groundwater level gives more opportunities for the available water for farmers. The actual evapotranspiration data is extracted from the database GAEZv4, Global Agro- Ecological Zones version 4, of the Food and Agriculture Organization of the United Nations, which is accessed by the data portal (GAEZv4 Data Portal, 2023) which comprises of a large volume of spatial natural resource data. The evapotranspiration is the process of transferring moisture from a liquid form on earth to a gas form in the atmosphere by evaporation of water from bare soil or open water bodies or transpiration from plants (Silander, 2001). The soil moisture depletes faster by a high evapotranspiration which can potentially lead into crop water stress (Mishra, 2013).

The water scarcity data and the groundwater data are extracted from an ADF file by the means of the open source, geospatial software QGIS. The software assists in converging the file into a raster TIFF file, the evapotranspiration data is already in a raster TIFF file. The water characteristics are extracted from the raster TIFF files by the means of a script in the programming language Python. The script reads the list of coordinates from a CSV file, created from the dataset of the agricultural data, then reads the raster data from the TIFF file corresponding with the coordinates, the data is saved to a CSV file. The water characteristics will consist of one type of data namely integer data, they are summarized in Table 2.

Table 2 - Summarized water characteristics including data type and source

Data	Type	Source
Water scarcity	Integer	(Mekonnen & Hoekstra, 2016)
Groundwater	Integer	(Su H. , 2023)
Evapotranspiration	Integer	(GAEZv4 Data Portal, 2023)

2.2.3 Soil characteristics

The soil characteristics that are used within the research are the global soil organic carbon (GSOC), the nutrient availability and the terrain slope. A list of potential characteristics is identified in the report of the Food and Agriculture Organization of the United Nations (FAO, 2022). The report identifies factors that have an impact on the agricultural sector. The list of potential characteristics and the data availability results in the choice of characteristics in the research.

The GSOC data is extracted from the database of GloSIS, Global Soil Information System, from the Food and Agriculture Organization of the United Nations, which is accessed by the data portal (GloSIS Data Portal, 2023), which provides access to soil resource information and is developed by International Network of Soil Information Institutions (INSII). Soil organic carbon is one of the important indicator for the soil health. The nutrient availability data is extracted from the (GAEZv4 Data Portal, 2023). The nutrient availability is also an indicator of the soils health and would include different nutrients such as nitrogen, phosphorus and the salinity. The terrain slope is extracted from the (GAEZv4 Data Portal, 2023). The slope of the terrain is indicated with a percentage how steep or shallow the terrain is a specific location. The slope of the terrain can give an indication where the farms are located, and which locations are off limits.

The soil characteristics are all extracted in the form of a TIFF file from the mentioned databases. The raster data from the TIFF file is extracted by means of a script in the programming language Python. The same procedure is applied for the soil characteristics as for the water characteristics, the script reads a list of coordinates from a CSV file while reads the raster data form the TIFF file corresponding with the coordinates, the data is saved to a CSV file. The soil characteristics will consist of two type of data, integer data and class data. The nutrient availability and terrain slope are in classes, while the GSOC is an integer. The classes for nutrient availability range between zero and one to indicate how much nutrients are available ([0.1-0.2], [0.2-0.3], … , [0.8-0.9], [0.9-1.0]), the terrain slope classes will consist of values of percentage of slope ([0.0-0.5], [0.5-2], [2-5], [5-8], [8-16], [16-30], [45+]). The characteristics are summarized in Table 3.

Table 3 - Summarized soil characteristics including data type and source

Data	Type	Source
Global Soil Organic Carbon	Integer	(GloSIS Data Portal, 2023)
Nutrient Availability	Class	(GAEZv4 Data Portal, 2023)
Terrain slope	Class	(GAEZv4 Data Portal, 2023)

2.2.4 Climate characteristics

When doing research which contains spatial data, it is required to include some of the climate characteristics from the spatial location that are used with in the research. The climate data that is use are temperature and precipitation. Both data is extracted from the (GAEZv4 Data Portal, 2023) in the form of a TIFF file. The location specific data is extracted from the TIFF file using a script in the programming language Python. The same procedure as for the water and soil characteristics is applied, the script reads coordinates from a CSV file while extracting raster data from the TIFF file and save to a new CSV file. The data type of the climate characteristics are in the form of integer data, they are summarized in Table 4.

Table 4 - Summarized climate characteristics including data type and source

Data	Type	Source
Temperature	Integer	(GAEZv4 Data Portal, 2023)
Precipitation	Integer	(GAEZv4 Data Portal, 2023)

2.3 Data preparation

In the previous sections it is explained which data is used, what type of data is extracted and where the data is coming from. The separate data files from all the characteristics are merged into one CSV file by the means of script in the programming language Python. In the following section it will be explained how the features are constructed and how the data is prepared for the use of the unsupervised machine learning algorithm.

2.3.1 Feature engineering

The development of new data features from the raw data is called feature engineering. A feature is any measurable input that can be used in a predictive model (Patel, 2021). The performance of the model can significantly be increased by the construction of features, which are implemented in the model (Turner, Fugetta, Lavazza, & Wolf, 1999). Feature engineering can be applied with different purposes and depends on the research. In this research, feature engineering is applied to the data in order to compress the dataset and better represent the spatial data per location. In order to accomplish this, the sample is set as the coordinates of every unique location. In this way, the features will represent the data for one location. The location will be the 30arcmin coordinates for the latitude and longitude. The construction of the features is divided into two different parts, bucketizing and one-hot encoding.

- Bucketizing has the goal to map the numerical values of the data into a ‘bucket’ or ‘bin’ and replace the original value with a numerical value that represents the bucket (Kuhn & Johnson, 2019) (Polzer, 2023).
- One-hot encoding has the goal to transform categorical data into numerical data, by doing so the column with the groups is split into several new columns with each describing a unique group (Kuhn & Johnson, 2019) (Polzer, 2023).

2.3.2 Bucketizing

The coordinates for the extraction of the data, as describe in previous sections, are at a 5arcmin grid. In order to compress the dataset, it is chosen to use a 30arcmin grid. The compression of the data from a 5arcmin grid to a 30arcmin grid can be done in combination with bucketizing of the data. The 5arcmin grid is assigned an index of the 30arcmin coordinates, which functionate as the bucket, while the value is replaced by a new value that represents that 30arcmin location. In Table 5 is an example given of a bucket, while Table 6 represents the results of the process.

Table 5 - Example of 'bucket' in dataset

Index	Coordinates 5arcmin (Lon/Lat)	Temperature (°C)	Water Scarcity (-)	Terrain Slope (%)
29950	34,042 / 69.375	0.31	0,001111	8-16
29950	34,125 / 69.375	0.50	0,001111	8-16
29950	34,125 / 69,292	0.30	0,000911	16-30
29950	34,208 / 69,292	0.62	0,000911	16-30
29950	34,292 / 69,292	0.68	0,000911	16-30
29950	34,375 / 69,292	0.98	0,000911	16-30

In order to calculate a representative value for the bucket different methods are used for each data. The calculation are as follows:

- Climate characteristics: for both temperature and precipitation the average of the bucket is taken as new value for the bucket.
- Water characteristics: for the water scarcity, evapotranspiration and groundwater level the average of the bucket is taken as new value for the bucket.
- Soil characteristics: for the GSOC the average of the bucket is taken as new value for the bucket and for the nutrients availability and terrain slope the dominant class of the bucket is taken to represent the new value for the bucket.

Table 6 - Result of bucketizing the 'bucket' in Table 5

Index	Coordinates 30arcmin (Lon/Lat)	Temperature (°C)	Water Scarcity (-)	Terrain Slope (%)
29950	34,25 / 69,25	0.56	0.001	16-30

For the calculation of the bucketizing are no scaling schema or procedures used, each data point is taken as equal importance in the calculation. Especially in the calculation of the soil characteristics this could have an impact on the data. The effects of the assumptions are not further assessed.

2.3.3 One-hot encoding

In the data of the agricultural characteristics, the coordinates of the 5arcmin grid are transformed to a 30arcmin grid by means of an index, the same as used in the bucketizing. For the creation of features for the agricultural characteristics a one-hot encoding procedure is applied. The data still contains the sample coordinates with different combination of variable values for the farming system, crop type and farming system. For one sample, the total harvest area of all combination of data value is calculated and used in the construction of the features. In Appendix B is an example of one sample location illustrated, to give more context in the calculation. In Equation 1, an example of the calculation of the features is given.

$$Variable = \frac{HA\ of\ Variable}{THA\ of\ Sample} * 100 \qquad \text{Equation 1}$$

Where:
Variable is feature to construct
HA is harvest area
THA is total harvest area
Sample is specific combination of coordinates

The variable is the specific feature that will be constructed, for example the percentage of smallholders in the sample, while the harvest area of all smallholders in the sample is placed in the numerator and the constant total harvest area of the sample is placed in the denominator. In order to calculate a percentage, the fraction is multiplied by one hundred. The calculations are performed on all sample locations for the variables: smallholders, non-smallholders, all the 12 different crop types and the 4 different farming systems. The constructed feature will be in the form of a percentage and give an indication of the presence of different variables in a specific location. If there are mostly smallholders or non-smallholders, what the distribution of the harvested crop is and which farming system the farmers mostly apply.

Besides the calculation for the presence of the different agricultural characteristics, the density of farmers in a region could be an interesting feature in order to compare potential clusters. Therefore, the total harvest area of the sample is compared to the total area of the sample. This feature gives an indication of how much of the land is occupied for farmland. In Equation 2, an example of the density calculation for the samples is given.

$$Variable = \frac{HA\ of\ Variable}{TA\ of\ Sample} * 100 \qquad \text{Equation 2}$$

Where:
Variable is specific feature to construct
HA is harvest area
TA is total area of a grid
Sample is specific combination of coordinates

The calculation is performed for the variables total harvest area, total harvest area of smallholders and total harvest area of non-smallholder. This constructed feature will also be in the form of a percentage and indicate the density of farmland in the total grid.

The result of the calculation for the bucketizing and the one-hot encoding will lead to a dataset which contains each 30arcmin grid location with 30 columns that describe the location by the means of features constructed in this section.

2.3.4 Handling outliers

Common practice in the field of data analysis is the handling of outliers because not all data is usable for the analysis in the unsupervised machine learning algorithm. Outliers may result from specific local circumstances and will not be part of a general pattern. There are diverse ways to handle the identified outliers, such as remove outliers, transform outlier, impute outliers or make use of robust statistical methods. For this research, only the removal of outliers will be considered, the other methods will not be discussed. By removing the identified outliers from the dataset, the performance of the model will increase and it is expected that the results will give more accurate identification of the patterns. The method that is used for handling the outliers is the percentile method, the method identifies outliers in a dataset by comparing each data point to the rest of the data using percentiles. By applying the method, the upper and lower bound are determined by the desired percentile for the research (Scaler Topics, 2023). In the research, a maximum of one percent of each feature will be removed in the form of the 1st percentile for the lower bound, the 99th percentile for the upper bound or 0.5th and 99.5th percentile for both lower and upper bound (Firdose, 2023). In order to make an adequate decision for which part will be considered as outlier, the distribution is plotted and analysed. Out of the plotted distributions, the bound with less data density compared to the average will be considered as outlier. Figure 1, Figure 2 and Figure 3 will give an example of the distribution before and after the removal of outliers for the lower bound, upper bound and both bounds. In Table 7, for each feature it is given which part of the data is removed.

Table 7 - Outlier removal for each feature

Features	Lower bound	Upper bound	Both bounds
Temperature	X		
Precipitation		X	
Water Scarcity			X
Ground Water	X		
Evapotranspiration			X
Nutrients	X		
Slope			X
GSOC		X	
% Irrigation		X	
% Low Input Rainfed		X	
% High Input Rainfed			X
% Rainfed in subsistence cond.		X	
% Crop: Stimulates		X	
% Crop: Fruits		X	
% Crop: Vegetables		X	
% Crop: Cereals			X
% Crop: Pulses		X	
% Crop: Roots & Tubers		X	
% Crop: Oil crops		X	
% Crop: Fibres		X	
% Crop: Sugar crops		X	
% Crop: Rest		X	

The identification of the outliers is done for every feature on its own before the samples are remove from the data set. In this way, removing a sample can have multiple identified outliers for distinctive features. In this way, the removal of outliers for the first features will not influence the removal of outliers in later features in the dataset. The total amount of datapoint will be decreased from 15.048 datapoints to 12.933 datapoint, which is a decrease of 14.06 percent compared to the original dataset.

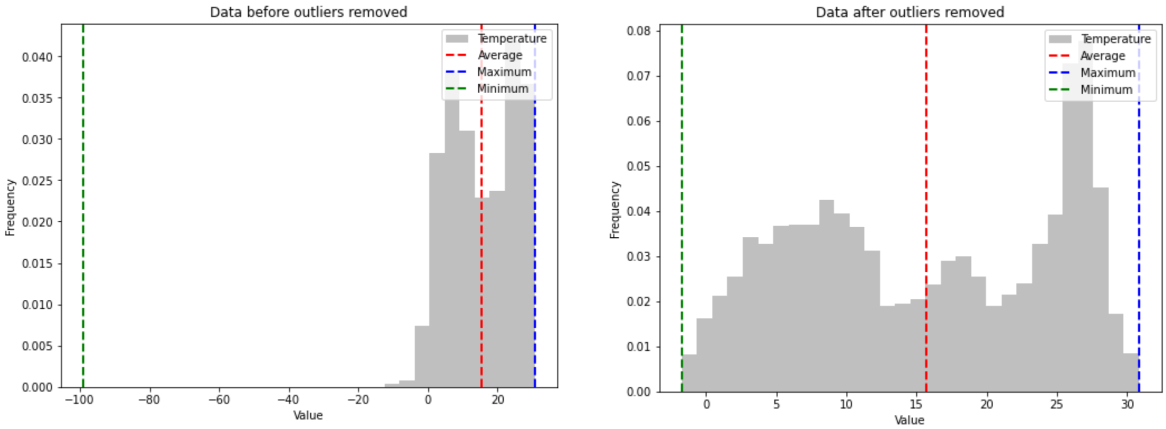

Figure 1 - Outlier removal of 1st percentile lower bound (before and after)

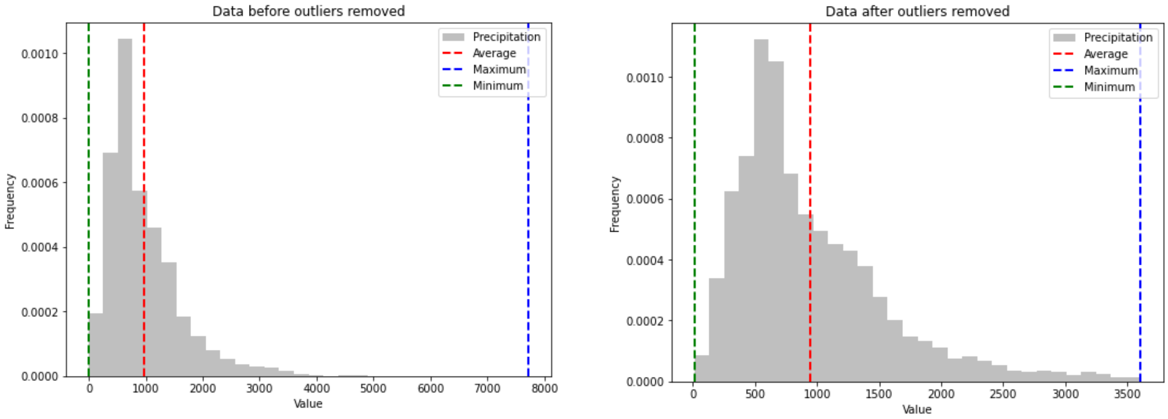

Figure 2 - Outlier removal of 99th percentile upper bound (before and after)

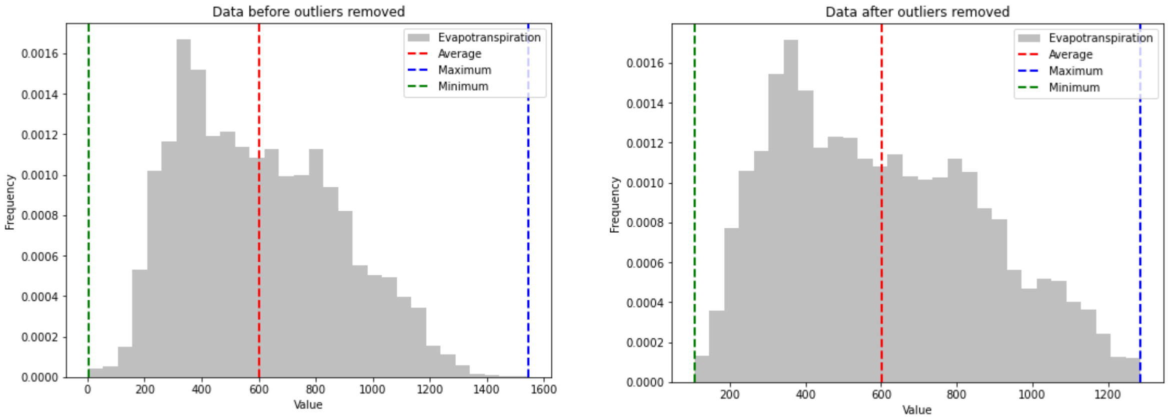

Figure 3 - Outlier removal of both 0.5th and 99.5th percentile upper and lower bound (before and after)

2.3.5 Scaling of data

The different features, such as temperature (°C), precipitation (mm), water scarcity (-) and GSOC (kg/m2), have different units and different magnitudes of the data. In order to make the features comparable and usable for the machine learning algorithm, the data needs to be scaled. Scaling the data can in separate ways namely normalization, standardization or min-max scaling. The goal of the scaling procedures is make all features with a similar range, mostly [0,1], [-1,1] or [0,100] (Rawat & Khemchandani, 2019). In this research the min-max scaling procedure is applied to scale the data. Equation 3 is used for the calculations.

$$\mathrm{x}' = \frac{\mathrm{x} - \min(\mathrm{x})}{\max(\mathrm{x}) - \min(\mathrm{x})} * 100 \qquad \text{Equation 3}$$

Where:
x’ is normalized value
x is original value
min(x) is minimal value of x
max(x) is maximum value of x

The normalized value will be on a scale [0,100], rounded to two decimals places after the decimal point. The scaling is performed with special packages available in the python software.

2.4 Choosing the algorithm

By the construction of features, handling the outliers and scaling the data, the data preparation is completed which means the application of the unsupervised machine learning can begin. The first step is to determine what kind of algorithm will be used in the research. It is already determined that an unsupervised machine learning algorithm will be applied, but within the unsupervised domain there are different algorithms available. The choice of algorithms depends on the purpose of the research, the available data and the structure of the data (Berry, Mohamed, & Yap, 2020). First the three main purposes of an unsupervised machine learning algorithm are explained, after that a more detailed elaboration on the performed algorithm will be given.

Within the unsupervised machine learning, three main applications of algorithms can be distinguished namely clustering, dimensionality reduction and association rule mining. Clustering algorithms aim to group similar data points together based on inherent patterns and similarities, commonly applied in customer segmentation or abnormal behaviour is cybersecurity (Mahesh, 2020). Dimensionality reduction algorithms aim to reduce the dimensions of a dataset, they capture the most relevant features of the data which would help visualise high-dimensional data (Mahesh, 2020). Association Rule mining algorithms aim to identify interesting relationships or associations between variables which is mostly applied in recommendation systems or market basket analysis (Kumbhara & Chobe, 2014).

For the purpose of the research a clustering algorithm will be applies in order to uncover patterns between the agricultural, soil and water features. The most common clustering algorithm are K-mean and Mean Shift. The algorithms are an iterative process and make use centroid-based approach, which means that centroids are created while similar datapoint are assigned to its more appropriate centroid (Shumaila, 2021). The main difference between the two algorithms is that K-mean required the determination of clusters in advance while for Mean-Shift that is not necessary. According to (Berry, Mohamed, & Yap, 2020), a K-mean algorithm is the most commonly used algorithm and in collaboration with the supervisor of the research it is decided that this algorithm will be used in the analysis and construction of the clusters and potential patterns.

The goal of a K-mean algorithm is to partition a dataset into a certain number of clusters where each sample point belongs to a cluster with the nearest mean. The algorithm contains of a certain number of steps that need to be taken in order to perform the algorithm properly. First the number of clusters need to be selected by means of different methods, more details are given in section 2.5. The number of clusters will be an input of the algorithm, then the algorithm performs the following tasks:

- Determining random centroids for the number of clusters that is assigned
- For each datapoint, the distance to each cluster centroid is calculated and assigned to the closest centroid. The Euclidean distance metric is used.
- Once all data point are assigned, the mean of all clusters is calculated.
- The mean of the clusters will become the updated centroids of the clusters.
- An iterative process begins where the previous steps are repeated until a convergence is achieved, this means that the clusters do not significantly change or until a certain predefined number of iterations is reached.

The first step of choosing the random centroids can have an influence on the final results, so the algorithm performs the same steps for different initial centroids and choses the best result based on a certain criterion. The criteria used is to minimize the sum of squared distances within the cluster. This criterion is later on used to determine the dominance of the features. The final result of the algorithm is a data set with a number of clusters where the datapoints are grouped based on their similarities and minimization of the clusters variance.

2.5 Number of clusters

In section 2.4 it is mentioned that the number of clusters, also called K, needs to be determined before the K-mean algorithm can be performed. Multiple methods are available to calculate the optimal number of clusters. The most common methods are the Elbow Method and the Silhouette score (Berry, Mohamed, & Yap, 2020). The methods will give an indication of how well the clustering results are formed based on the different number of clusters.

The Elbow Method involves running the K-mean algorithm for eighty percent of the dataset for a range of values of clusters. For each value of clusters, the sum of squared distances of each data point to its assigned cluster centroid (Cui, 2020). The SSD is plotted for every value of clusters, the plot typically shows a decreasing trend. The determination of the optimal number of clusters from the Elbow Methods is subjective but aims to identify the point where the rate of change sharply changes. If an extra cluster identifies an apparent part of the data, the WCSS will reduce significant. While, on the other hand, if an extra cluster will result in dividing or splitting of clusters, the WCSS will marginally reduce. The transition from one to another will indicate the 'elbow' in the graph.

$$WCSS = \Sigma_{i=1}^{n} \Sigma_{j=1}^{k} \left| x_{i-} c_j \right|^2 \qquad \text{Equation 4}$$

Where:
N is number of data points
K is number of clusters
X_i is a single data point
C_j is the centroid of cluster j

The Silhouette Score evaluates the quality of the clusters based on how-well separated they are from each other. For the evaluation it considers both the cohesion within a cluster and the separation between clusters (Naghizadeh & Metaxas, 2020). The score resulting from this method will range from [0,1], with higher score mean better defined clusters. For each value of clusters, the silhouette score is calculated and averaged for the cluster. The averaged values for each number of clusters are plotted, with the maximized average silhouette score is considered the optimal number of clusters.

$$s(i) = \frac{B(i) - A(i)}{max\{A(i), B(i)\}} \qquad \text{Equation 5}$$

Where
A(i) is the average distance from a data point to the next data point in the same cluster
B(i) is the average distance from a data point to a data point in the nearest cluster

The calculation for the optimal number of clusters is performed by the use of a script in the programming language Python. During the calculation different initial settings for the random state and the range of maximum number of clusters to consider are evaluated. The setting will be tested in two tests. The first test, the random state will range from [0-10] while keeping the number of clusters constant. The results are plotted to give an indication if the choice of random state will influence the optimal number of clusters. Similar plots will indicate at minor influence, heavily changing plots will indicate at major influence. The second test, the maximum number of clusters will be set at 100 while keeping the random state constant. The results are plotted, from the plot a maximum number of clusters can be chosen for the analysis. This number will influence the computational time of the script and will be aimed to be minimized.

2.6 Dominant factor

In the domain of unsupervised machine learning, in specific the K-mean algorithm, there is no clear, distinctive way to assess the importance of the individual features. However, there are a lot of validity indexes that can be assessed in order to check the performance of the clustering itself (Sinaga & Yang, 2020). Two of these indexes are already applied in the determination of clusters as input for the K-mean algorithm. For the determination of a dominant factor, the same principles are applied. The within cluster sum of squares (WCSS) aims to minimize the distance between the data point and the cluster centroid by assigning the data point to its nearest cluster centroid. This principle will be used to determine a dominant factor. The algorithm of K-mean aimed to minimize the WCSS of the clusters, with means that the data point is as close as possible to the nearest centroid.

That principle is combined with a permutation test, in order to see the effect of the permutation test on the WCSS. The approach is based on the paper of (Alghofaili, 2021) which describes the approach as a direct analysis of each centroid sub-optimal position. The approach tries to find the feature with is responsible for the highest WCSS minimization for each cluster through finding the maximum absolute centroid dimensional movement. By removing a feature from the dataset, running a K-mean algorithm again where new centroids will be created on the new dataset. The location of the centroid will change for the new dataset, therefor also the Within Cluster Sum of Squares minimization will change. The new centroid and WCSS of a cluster will be compared to its corresponding old centroid and WCSS. To find the corresponding old centroid to the new centroid, the total distance between all the centroids is minimized. When the new centroid and old centroid are linked to each other, the difference in WCSS gives an indication of the feature importance. The highest WCSS, so the most change of the centroid in compared to the data points in that cluster, will have the most influence on the clustering process. The most influential features will be considered as the dominant factor.

The WCSS is the sum of all distances, which can be an extremely high number. In order to make the result more comparable, a scaling procedure is applied. A min-max scaling procedure is applied on all the value of WCSS for the feature importance, similar calculation is done is section 2.3.3 of the data preparation. The scaling is done for a range of [0-1], where the higher the number, the more important the feature.

3. Results

This chapter will contain the outcome of application of the methods described in the previous chapter. First the number of clusters are calculated, then the K-mean algorithms is performed in combination with the determination of the dominant factor for the features. The clusters will be described leading with the dominant factors, with addition description of the clusters. At last, there will be made an analysis on the most standing out results of the clusters in terms of the features.

3.1 Number of clusters

The optimal number of clusters is determined by calculating the Elbow Method and the Silhouette Score for a range of clusters. The initial setting for the random state and the maximum number of optimal clusters to consider are determined. The random state for a range of [0-10] with a constant number of clusters has no influence on the results. All plots for the different random state are similar in form and shape, Appendix C will illustrate the result of the test. The maximum number of clusters to consider will be set to 20, because both methods show no improvement above 20 cluster. In order to limit the computation time, the maximum number of clusters is set at 20, Appendix D will illustrate the results of the test.

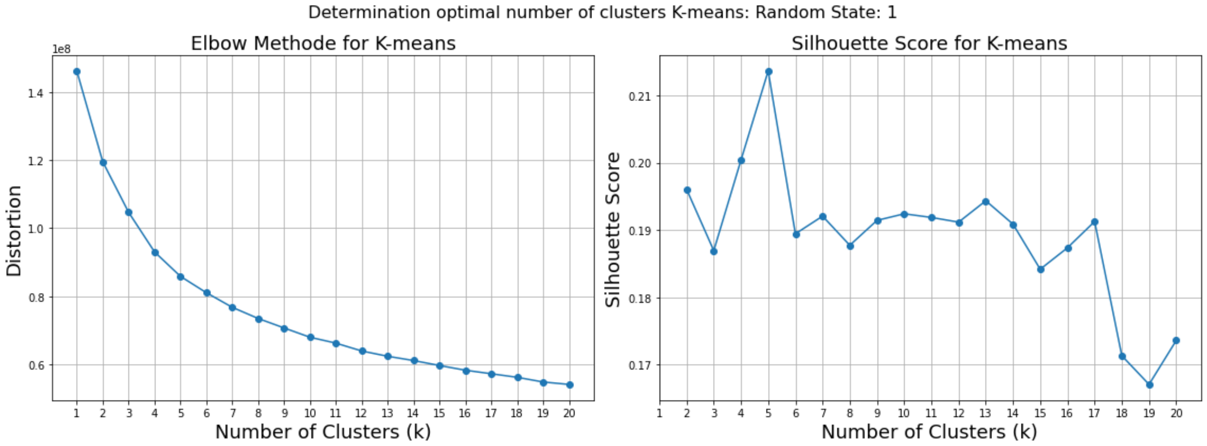

Figure 4 - Results of Elbow Method and Silhouette Score for K-means

The results of the calculations can be found in Figure 4. The Elbow Method indicates a clear decreasing line for an increase in number of clusters, as said in the section 2.5, the method provides a subjective choice for the optimal number of clusters. Results however do not show a clear bend in the graph where a steep decrease in WCSS changes into a shallow decrease. Therefor it is paired with a Silhouette Score, which indicates at the optimal number of clusters equal to 5 clusters. Comparing 5 clusters in the Elbow Method, it is considered as an acceptable number for the optimal number of clusters. Looking at the Elbow Method, it could be considered to look at the number of clusters equal to 6 or 7. That consideration will be part of the discussion section of this research.

3.2 Dominant factor

The permutation test results of the changing WCSS minimizer can be found in Table 8. The values in the table represent a scaled importance of the WCSS, the results are scaled in the range of [0-1] in order to make a better comparison between the features. In the Table 8, all features are listed with the scaled importance.

Table 8 - Dominant factor per feature

Feature	Scaled Importance
Water Scarcity	1,00
Ground Water	0,94
GSOC	0,87
% Crop: Cereals	0,79
% Crop: Oil crops	0,67
% High Input Rainfed	0,60
% Smallholders	0,52
% Non-Smallholders	0,52
% HA non-smallholders of total area	0,50
Slope	0,47
% Irrigation	0,44
% Harvest area of total area	0,44
% Low Input Rainfed	0,43
Nutrient Availability	0,37
Temperature	0,31
% Rainfed in subsistence cond.	0,28
Evapotranspiration	0,23
% Crop: Fruits	0,19
% Crop: Roots & Tubers	0,18
% Crop: Pulses	0,17
% HA smallholders of total area	0,17
Total Harvest Area	0,16
Precipitation	0,13
% Crop: Fibres	0,09
% Crop: Vegetables	0,09
% Crop: Stimulates	0,04
% Crop: Rest	0,01
% Crop: Sugar crops	0,00

Things that stand out are that the influence of smallholders and non-smallholders is the same, so the presence of what type of farm does equally influence the results. The most influential soil characteristic is the Global Soil Organic Carbon and the most influential water characteristics is the Water Scarcity. Most crops have almost none influence on construction of clusters except the crop cereals and oil crops. The influence of the farming system is average, none of the features has a considerable influence.

3.3 Cluster analysis

In order to give a clear representation of the clusters, each cluster will be described individual. The clusters description will be used in order to make an analysis of the differences between the clusters in terms of smallholder and non-smallholder presence, the present farming system, the mostly planted crop type and the dominant factors. The figures illustrate by the cluster description are focused on the spatial distribution and data distribution of the dominant factors. For the details of the data distribution for smallholder presence, farming system and planted crop type, a separate figure for each cluster is presented in Appendix E and are not included in the results section of the cluster description.

Looking at the world map in Figure 5, the spatial distribution of the clusters is well divided over the continents. The spatial distribution, description of a number of features and the statistics of each of the dominant features will be described for each cluster individually. Figure 5 gives a good comparative distribution for the separation of the cluster compared to each other, however a couple of point will overlap into each other. Therefor an individual spatial distribution for each cluster is included in the description of the clusters.

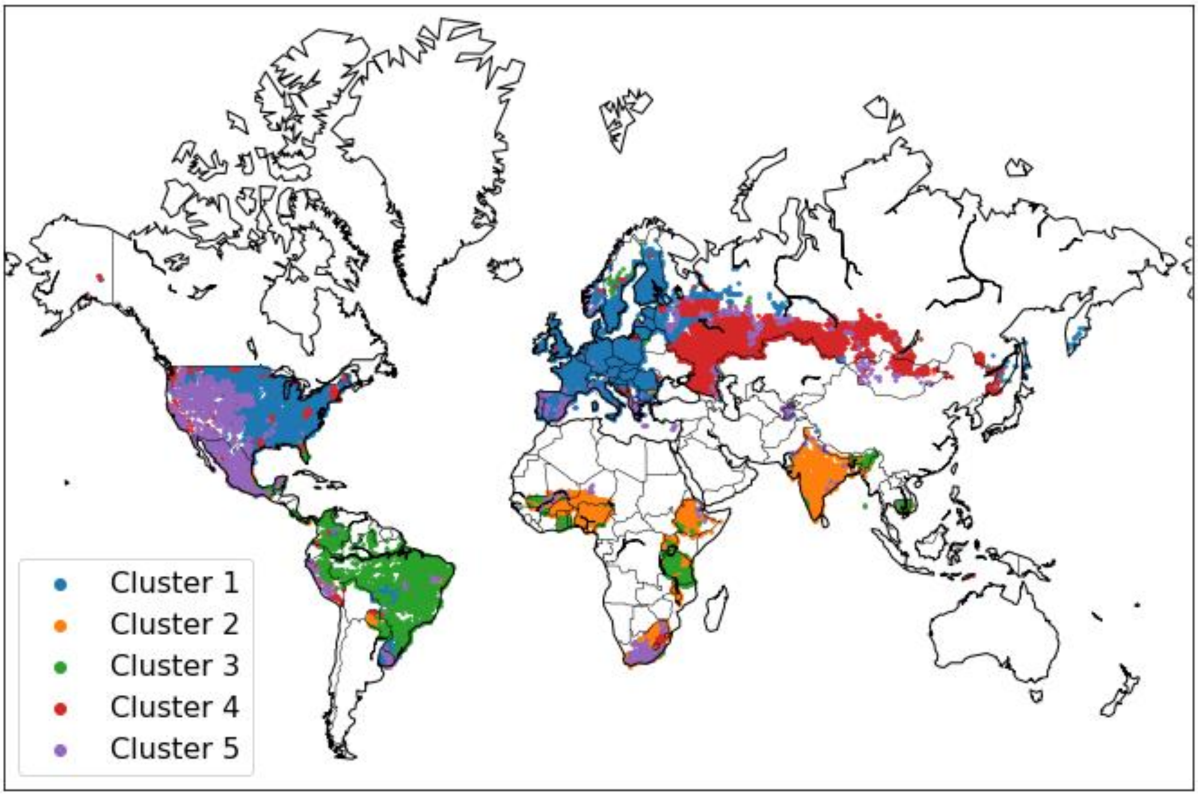

Figure 5 - Cluster distribution on world map (Mercator Projection)

3.3.1 First cluster

The first cluster is located in Europe, the central and east coast of the USA with a couple larger spots in South America in the area of central Brazil and Uruguay. The spatial distribution of the first cluster can be found in Figure 6.

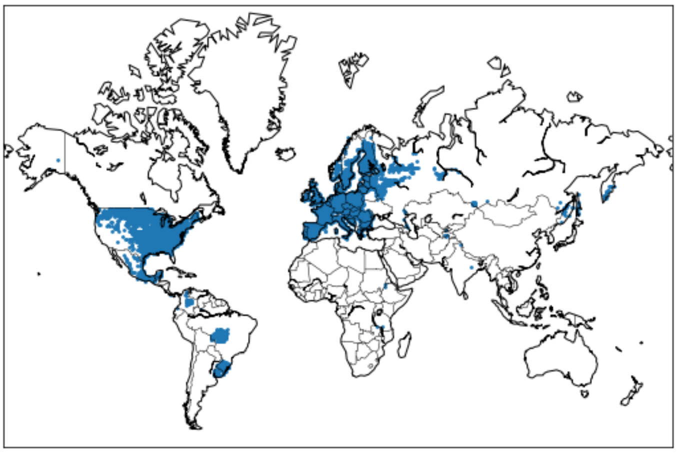

Figure 6 - Spatial distribution of first cluster (Mercator Projection)

The type of farmers in the region consists around nine-tenth of non-smallholders and one-tenth of smallholders. The most common planted crop in the regions are cereals and oil crops while the farming system is mostly high input rainfall. The detail distribution of all different crops, farming systems and farming class are state in Appendix E. The water scarcity has a median of 0.046 [-] with half the data within a range of 0.004 and 0.563 [-], which indicates at low water scarcity rate (Mekonnen & Hoekstra, 2016). Which means that, on average, less water will be used compared to the available water and that the natural flow requirements are met. The ground water has a median of -14.739 [m] with half the data within a range of -7.332 and -27.031 [m], which indicates at a median ground water table which can positively affecting the excess to water in absence of rain (Fan, Li, & Miguez-Macho, 2013). The GSOC has a median of 55.720 [kg/m2] with half of the data within a range of 39.685 and 74.315 [kg/m2], which indicates at a median presence of organic carbon in the soil. A high carbon ration will mean a better soil health (Prăvălie, et al., 2021). The detailed distribution of the water scarcity, groundwater level and global soil organic carbon can be found in Figure 7.

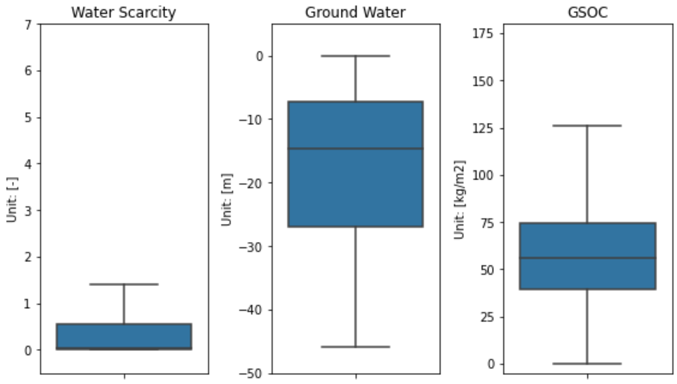

Figure 7 - Data distribution of dominant factor for first cluster

3.3.2 Second cluster

The second cluster is located in India, the East coast of Afrika, the West coast of Afrika and a couple of spots along the coast of South America and Mexico. The spatial distribution of the second cluster can be found in Figure 8.

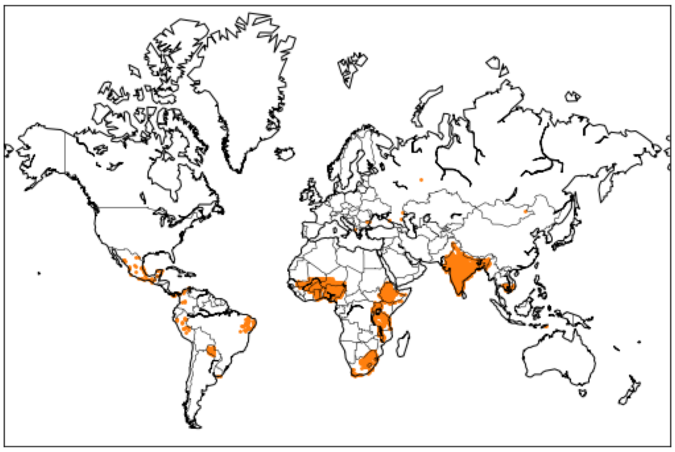

Figure 8 - Spatial distribution of second cluster (Mercator Projection)

The type of farmers in the region are diverse with around half of the farms are smallholders and half of the farms are non-smallholders. The most common planted crop in the regions are cereals with a combination of pulses and oil crops while the farming system is combination of low input rainfall and rainfall in subsistence conditions. The detail distribution of all different crops, farming system and farming class are state in Appendix E. The water scarcity has a median of 2.903 [-] with half the data within a range of 1.972 and 5.275 [-], which indicates at a severe water scarcity rate (Mekonnen & Hoekstra, 2016). Which means that, on average, more water will be used compared to the available water and that the natural flow requirements will probability not be met. The ground water has a median of -21.354 [m] with half the data within a range of -13.462 and -29.945 [m], which indicates at a median ground water table which can positively affecting the excess to water in some regions while in absence of rain (Fan, Li, & Miguez-Macho, 2013). The GSOC has a median of 27.080 [kg/m2] with half of the data within a range of 21.101 and 38.210 [kg/m2] which indicates at a low presence of organic carbon in the soil. A low carbon ration will mean a degraded soil health which is negative for the area (Prăvălie, et al., 2021). The detailed distribution of the water scarcity, groundwater level and global soil organic carbon can be found in Figure 9.

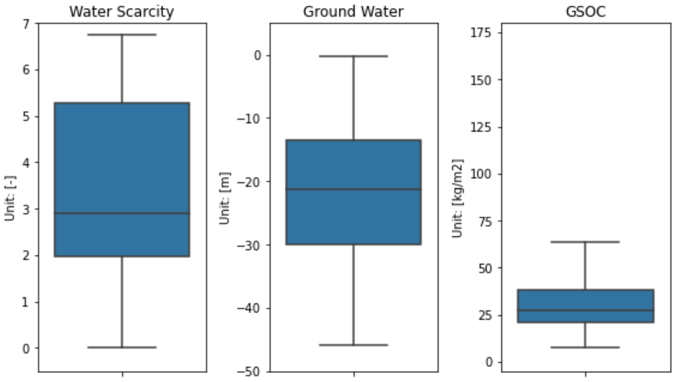

Figure 9 - Data distribution of dominant factor for second cluster

3.3.3 Third cluster

The third cluster is located in the North and East part of South America with a couple of large spots in Tanzania and the Himalaya Mountains. The spatial distribution of the third cluster can be found in Figure 10.

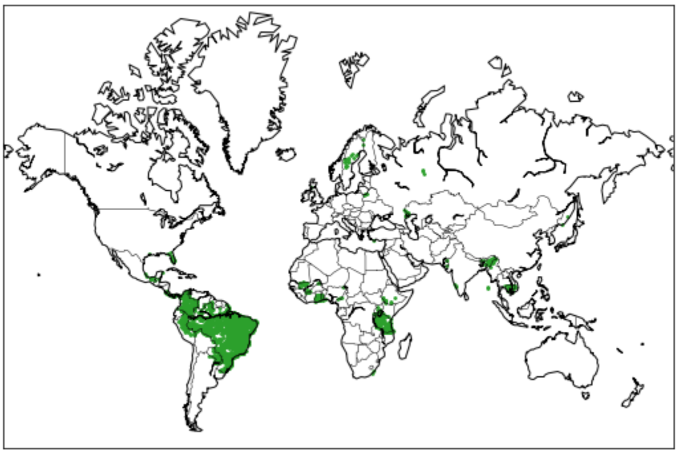

Figure 10 - Spatial distribution of third cluster (Mercator Projection)

The type of farmers in the region are around one-fourth smallholders and three-fourth non-smallholders. The most common planted crop in the regions are cereals, fruits, roots & tubers and oil crops while the farming system is mostly low input rainfall. The detail distribution of all different crops, farming system and farming class are state in Appendix E. The water scarcity has a median of 0.004 [-] with half the data within a range of 0.000 and 0.157 [-], which indicates at low water scarcity rate (Mekonnen & Hoekstra, 2016). Which means that, on average, less water will be used compared to the available water and that the natural flow requirements are met. The ground water has a median of -18.513 [m] with half the data within a range of -9.474 and -28.761 [m] which indicates at a median ground water table which can positively affecting the excess to water in some of the regions in absence of rain (Fan, Li, & Miguez-Macho, 2013). The GSOC has a median of 40.180 [kg/m2] with half of the data within a range of 33.895 and 49.825 [kg/m2], which indicates at a low presence of organic carbon in the soil. A low carbon ration will mean a degraded soil health (Prăvălie, et al., 2021). The detailed distribution of the water scarcity, groundwater level and global soil organic carbon can be found in Figure 11.

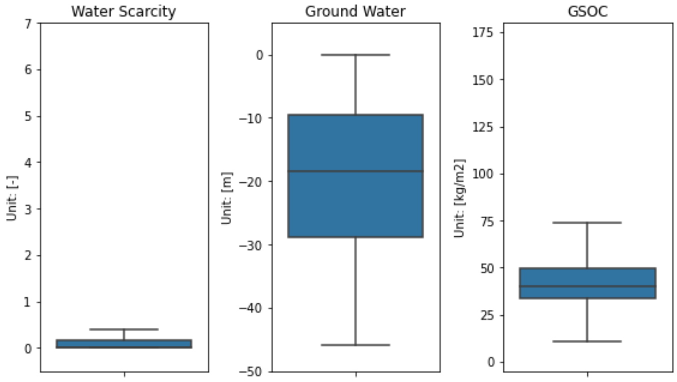

Figure 11 - Data distribution of dominant factor for third cluster

3.3.4 Fourth cluster

The fourth cluster is located in central Asia, going from Russia all the way to China, with a couple of spots along the coast in North and South America. The spatial distribution of the fourth cluster can be found in Figure 12.

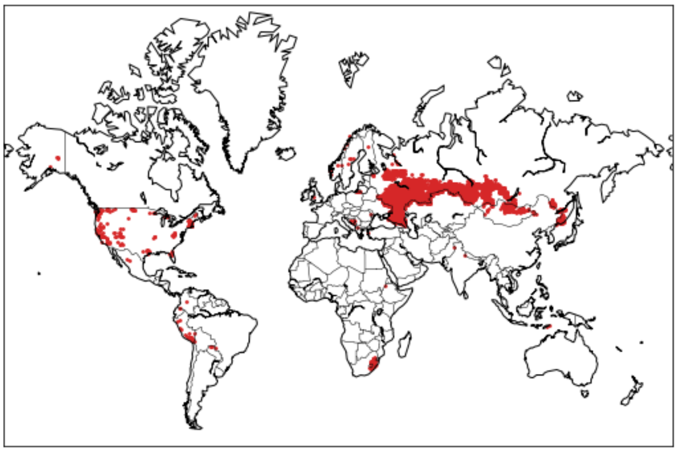

Figure 12 - Spatial distribution of fourth cluster (Mercator Projection)

The type of farmers in the region are around three-fourth smallholders and one-fourth non-smallholders. The most common planted crop in the regions are cereals and oil crops while the farming system is mostly high input rainfall. The detail distribution of all different crops, farming system and farming class are state in Appendix E. The water scarcity has a median of 0.034 [-] with half the data within a range of 0.002 and 0.319 [-], which indicates at low water scarcity rate (Mekonnen & Hoekstra, 2016). Which means that, on average, less water will be used compared to the available water and that the natural flow requirements are met. The ground water has a median of -10.105 [m] with half the data within a range of -4.613 and -16.389 [m] which indicates at a shallow ground water table positively affecting the excess to water in absence of rain (Fan, Li, & Miguez-Macho, 2013). The GSOC has a median of 84.240 [kg/m2] with half of the data within a range of 62.010 and 107.970 [kg/m2], which indicates at a high presence of organic carbon in the soil. A high carbon ration will mean a better soil health (Prăvălie, et al., 2021). The detailed distribution of the water scarcity, groundwater level and global soil organic carbon can be found in Figure 13.

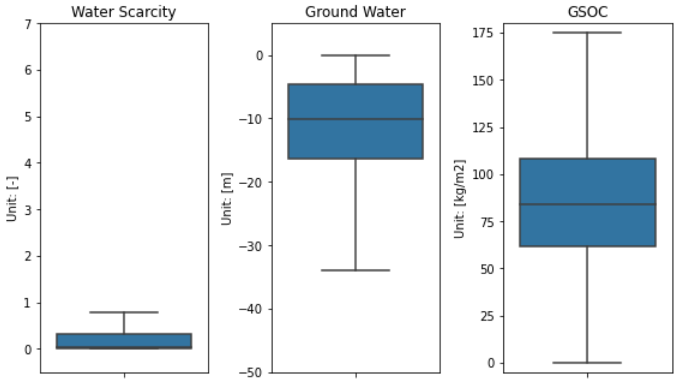

Figure 13 - Data distribution of dominant factor for fourth cluster

3.3.5 Fifth cluster

The fifth cluster is located on the West side of North America, including all of Mexico, with a couple of spots located over various parts of the world namely South Afrika, Russia, Mongolia and parts of Spain. The spatial distribution of the fifth cluster can be found in Figure 14.

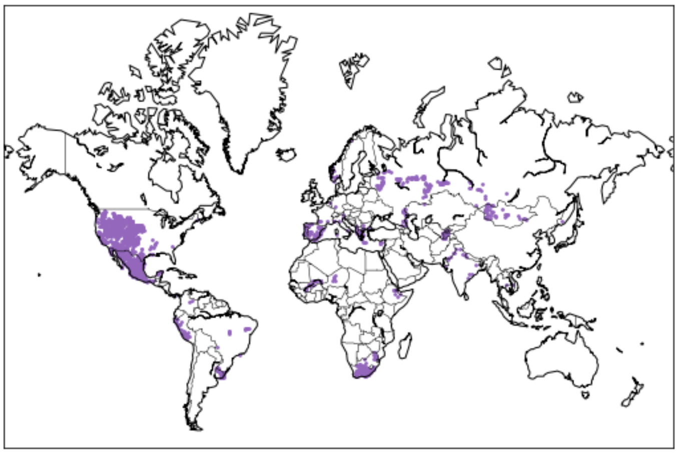

Figure 14 - Spatial distribution of fifth cluster (Mercator Projection)

The type of farmers in the region is consist around one third of smallholders and two third of non-smallholders. The most common planted crop in the regions are cereals, vegetables and fruits while the farming system is mostly in the form of irrigation. The detail distribution of all different crops, farming system and farming class are state in Appendix E. The water scarcity has a median of 3.391 [-] with half the data within a range of 0.795 and 6.194 [-], which indicates at low water scarcity rate (Mekonnen & Hoekstra, 2016). Which means that, on average, more water will be used compared to the available water and that the natural flow requirements will probability not be met. The ground water has a median of -31.722 [m] with half the data within a range of -22.628 and -37.207 [m], which indicates at a deep ground water table negatively affecting the excess to water in absence of rain (Fan, Li, & Miguez-Macho, 2013). The GSOC has a median of 33.120 [kg/m2] with half of the data within a range of 23.180 and 46.050 [kg/m2], which indicates at a low presence of organic carbon in the soil. A low carbon ration will mean a degraded soil health which is negative for crop production (Prăvălie, et al., 2021). The detailed distribution of the water scarcity, groundwater level and global soil organic carbon can be found in Figure 15.

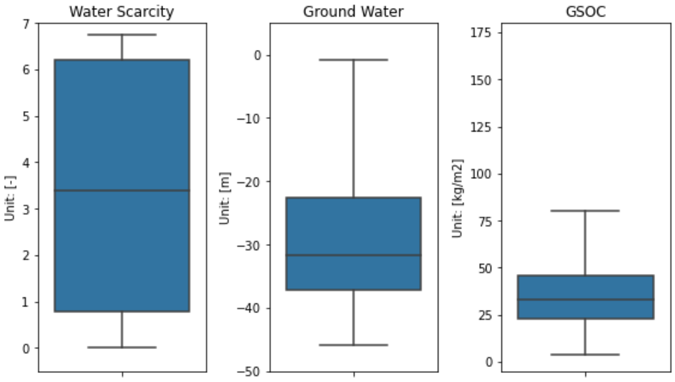

Figure 15 - Data distribution of dominant factor for fifth cluster

3.4 Cluster differences

The clusters have differences and similarities between them, this chapter will highlight in terms of the presence of smallholders, the water scarcity, ground water and GSOC on the basis of crop dominance, presence of the farming system and interesting features that stand out. The cross referencing between the different subjects is not included in order to avoid repetition of what is already mentioned. At the end of the section, a summarising table is included.

3.4.1 Smallholders

The first and third cluster are clearly dominated by non- smallholders while smallholders clearly dominate the fourth cluster. The mostly planted crop in the non-smallholders cluster are cereals for both, but the first cluster the second mostly planted crop are oil crops while for the third cluster the second mostly planted crop are roots and tubers. In the third cluster, there is more variety of planted crops while the first cluster is mostly only cereals and oil crops. The presence of farming system is in both non-smallholder clusters concentrated in rainfed production where the first cluster is mostly high input rainfed and third cluster is mostly low input rainfed. The average temperature is more than twice as high in the third cluster as in the first cluster, same applies for precipitation.

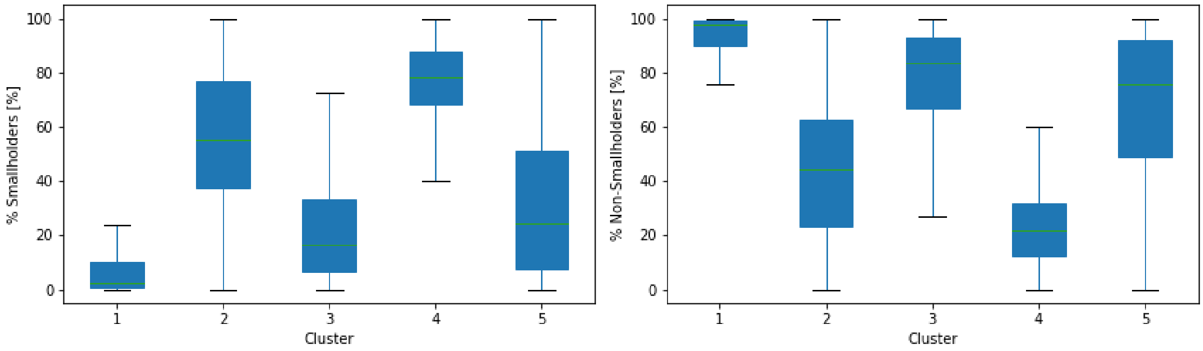

Figure 16 - Data distribution of all cluster for smallholders and non-smallholders

3.4.2 Water scarcity

The first, third and fourth cluster have a low water scarcity while cluster two and five have a severe water scarcity. Looking at the low water scarcity regions, one of the major differences is the presence of smallholders. The first and third are non-smallholder dominated while the fourth is smallholder dominated. In all clusters the most dominant crop is cereals, so looking at the seconds and third most planted crop is more interesting. The first and fourth cluster have a high presence of oil crops, while the second cluster has more variety in the presence of crops. Still oil crops have a relatively high presence but also roots & tubers, fruits and pulses have a more prominent presence in the second cluster than in the first and fourth cluster. All the clusters are dominate by rainfed farming system, while the first and fourth are high input rainfed and the third is low input rainfed. The first and fourth cluster are cold regions while the third cluster has a high average temperature. The precipitation and evapotranspiration are in the first and fourth cluster much lower than in the third cluster. All the clusters have similar average slopes while the third cluster has much lower nutrient availability than the other clusters.

Looking at the severe water scarcity regions, they have both not a dominant presence of smallholders and non-smallholders, it is a mix of both farm types. The dominant crop type in both clusters are cereals but the second mostly planted crop is different. The second cluster is more focused on oil crops and pulses while the fifth cluster is more focused on fruits and vegetables. The farming system for both clusters is different, the second cluster has a variety of different forms of rainfall while the fifth cluster is mostly dependent on irrigation for the excess to water. For the most part of the features they have similar values, but the temperature differs a lot between them.

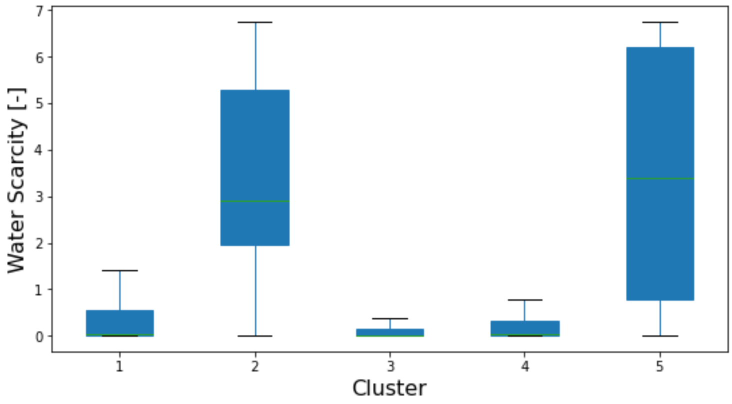

Figure 17 - Data distribution of all cluster for water scarcity

3.4.3 Ground water

A clear shallow ground water level can be seen in the fourth cluster while the fifth cluster shows a clear deep ground water level, the ground water level of the first, second and third cluster are in a moderate range. The mostly planted crops in the shallow area is cereals with a small part of production in oil crops while for the deep area the mostly planted crops are also cereals with more focus on fruits and vegetables. The fourth cluster is highly dependent on rainfed production while the fifth cluster irrigation is the commonly used method for access to water. Considering other features, the average temperature of the fourth cluster and the fifth cluster is low while the precipitation on both areas is also low. For the other features in terms of soil characteristics are similar between the clusters.

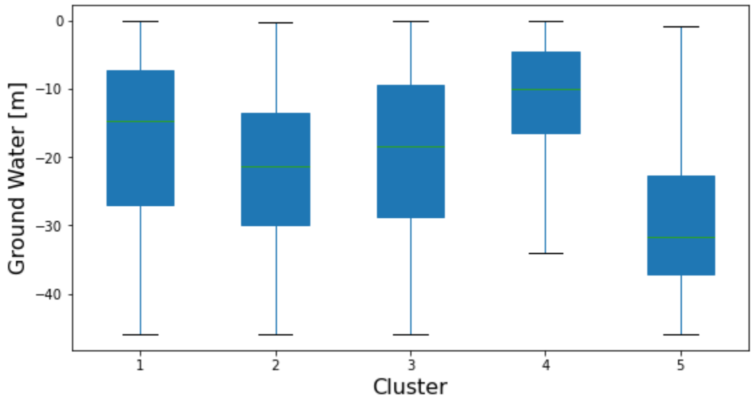

Figure 18 - Data distribution of all clusters for ground water

3.4.4 GSOC

The levels of global soil organic carbon are different for the clusters, there are three different ranges determined where the first cluster has a moderate level, the fourth cluster has a high level and the second, third and fifth cluster have a low level of soil organic carbon. The low level of soil organic carbon is used for the comparison. The dominant crop type in all clusters is cereals, the second cluster has besides the dominant crop also a presence of pulses and oil crops, the third cluster has a presence of roots & tubers and oil crops while the fifth cluster has a presence of fruits and vegetables. The farming system is quite different over the clusters, the second cluster has a combination of all farming system, the third cluster has a dominant presence in low input rainfed while the fifth cluster is mostly irrigation. For the other features, the temperature is different for the clusters. The second and third cluster have a high average temperature while the fifth cluster has a much lower average temperature. The precipitation is for the third cluster high while for the other clusters much lower. The evapotranspiration of the second and third cluster is much higher than the evapotranspiration of the fifth cluster. Also, the nutrients availability between the clusters is different. The second and fifth clusters have a higher nutrient content in the soil compared to a much lower nutrient content in third cluster.

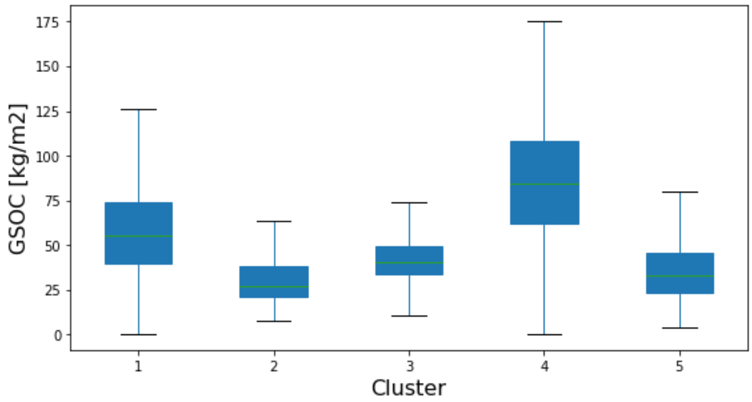

Figure 19 - Data distribution of all clusters for GSOC

In Table 9, the cluster differences are summarised for the farm size, water scarcity, groundwater level, GSOC, the most common planted crop and farming system.

Table 9 - Summarized cluster differences

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Farm size	Large scale	Mixed	Large scale	Smallholder	Mixed
Crops	Cereals and oil crops	Cereals, pulses and oil crops	Cereals, fruits and roots & tubers	Cereals and oil crops	Cereals, fruits and vegetables
Farming system	High input rainfed	Rainfed	Low input rainfed	High input rainfed	Irrigated
Water scarcity	Low	Severe	Low	Low	Severe
Groundwater	Moderate	Moderate	Moderate	Shallow	Deep
GSOC	Moderate	Low	Low	High	Low

4. Discussion

This chapter will discuss a couple decision, choices and procedures which were taken during the research. The limitations and constraints about the different decisions, choices and procedures will be given in order to put the results in context and draw a better conclusion from the research and recommendations for future research.

4.1 Influence of number of clusters

In section 2.5 of the methodology and section 3.1 of the results, the determination of the optimal number of clusters is explained in both context as outcome. In the section is mentioned that the determination of the optimal number of clusters is highly subjective, the literature supports this. Because of the subjective nature of choices, it is interesting to see what the different choice would look like in a spatial perspective. How is the spatial distribution over the world map made by the use of 6 or 7 clusters. These values are mentioned in the methodology as potential alternatives for the chosen optimal number of clusters in the research. During the analysis of 6 and 7 clusters, the order of determined clusters will become different, therefor the analysis is made on the location and not on the numbering.

The locations for 5 clusters, Figure 5 in the results section, were Europe/Central and East of USA, India/East and West coast of Afrika, South America, Central Asia and West coast USA/Mexico. These locations are compared in order to see if there is a shift of location and a split or merging of locations.

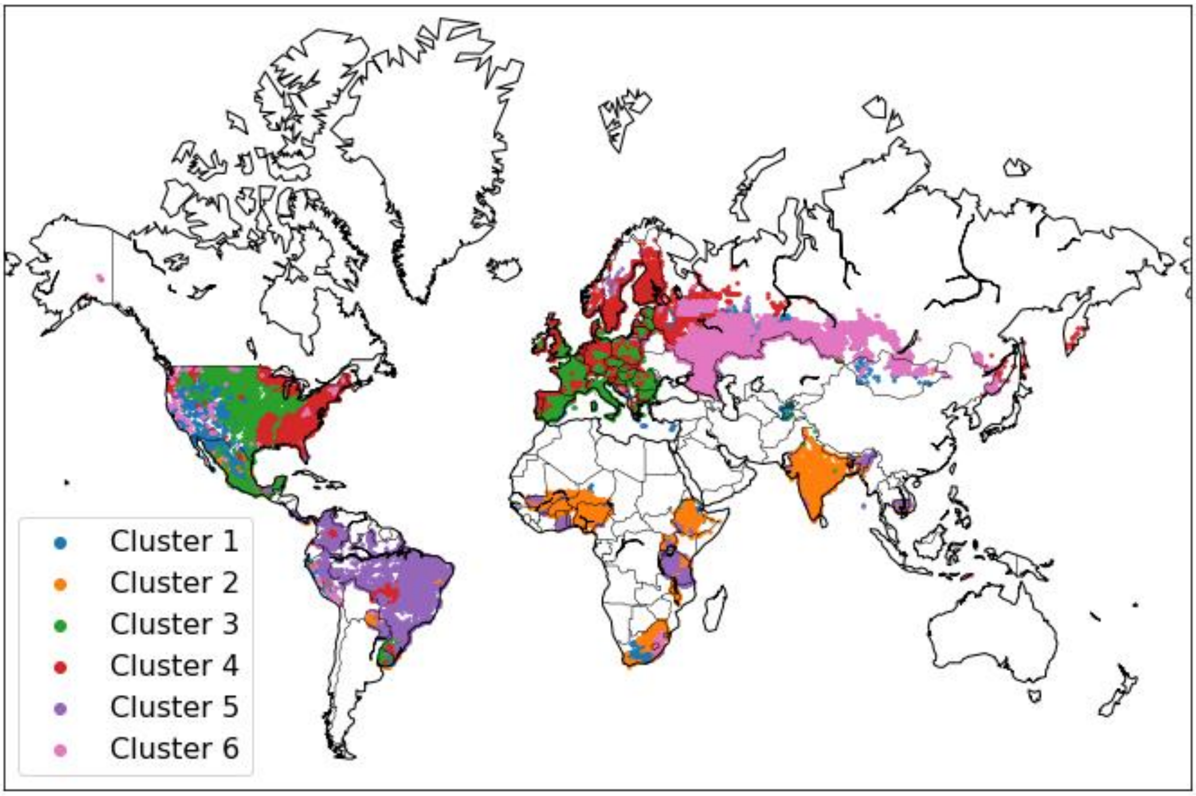

Figure 20 - Spatial distribution for 6 clusters (Mercator Projection)

The spatial distribution of 6 clusters is illustrated in Figure 20, the figure has overlap due to the way of plotting by the software. For the comparison, separate plots of each cluster is used. It can be seen that the spatial distribution is mostly the same for all the clusters. There is still a cluster on the West coast USA/Mexico, South America, Central Asia and India/East and West coast of Afrika. However, the main difference can be seen in the cluster of Europe/Central and East of USA. This cluster is split into two different clusters divided over the two different continents. A separation of the clusters over the continents, so each cluster in one continent, would be expected. However, looking at the data distributions of the features, most data ranges stay similar to each other. The features that change the most between the clusters are the dominant factor water scarcity and the nutrient availability.

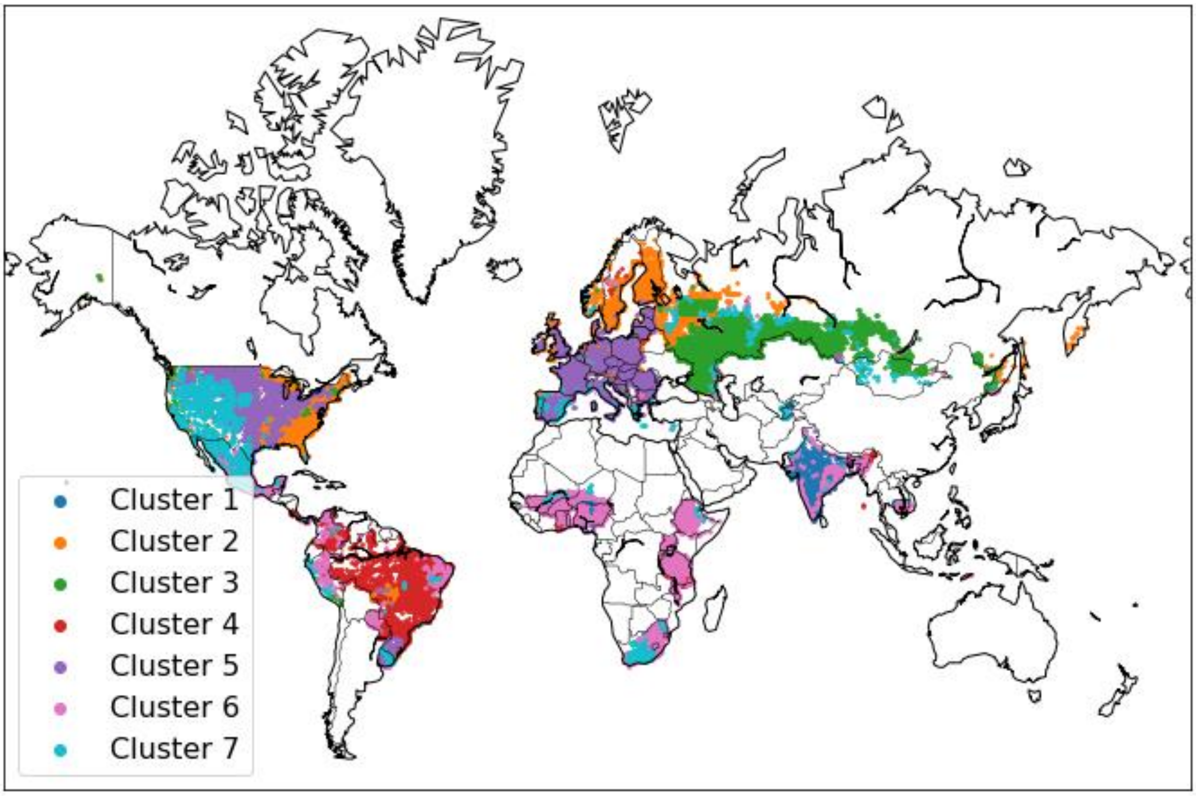

Figure 21 - Spatial distribution for 7 clusters (Mercator Projection)

The spatial distribution of 7 clusters is illustrated in Figure 21, the figure has overlap due to the way of plotting by the software. For the comparison, separate plots of each cluster is used. The spatial distribution of the new clusters will be compared to the distribution of 5 clusters and is changed more than the distribution of 6 clusters. The clusters at locations West coast of USA/Mexico, Central Asia and South America are almost not changed. However, the other two locations show major changes. The Europe/Central and East coast of USA is again split into two clusters shattering the data point over both continents, in the same way as by 6 clusters. The same data ranges between the two split clusters can be identified as in the clustering for 6 clusters. The cluster located in India/East and West coast of Afrika is also split into two clusters. A contraction of the datapoint is made in India, splitting themselves from Afrika. The range of data point is for most features the same however for the dominant factors water scarcity and GSOC the data is change in more distinctive differences.

The overall trend in increasing the number of clusters from 5 clusters to 6 or 7 clusters would mean that the clusters will split up. However, clear and centred split is preferred and would increase the clustering performance, however the opposite can be seen. It could be expected that the other clusters will split up when increasing the number of clusters beyond 7 clusters.

4.2 Influence of random state

In section 2.5 of the methodology and section 3.1 of the results, the influence of the random state is assessed on the determination of the optimal number of clusters. In the results of the test, it could be seen that the choice of random state does not influence the optimal number of clusters. However, it could be that the random state has an influence on the clustering performance of the K-mean algorithm itself. The algorithm is performed for a random state of 2 and the spatial distribution is plotted on a world map in order to give an indication of the changes. The spatial distribution is used in order to uncover major changes, no individual data point are compared.

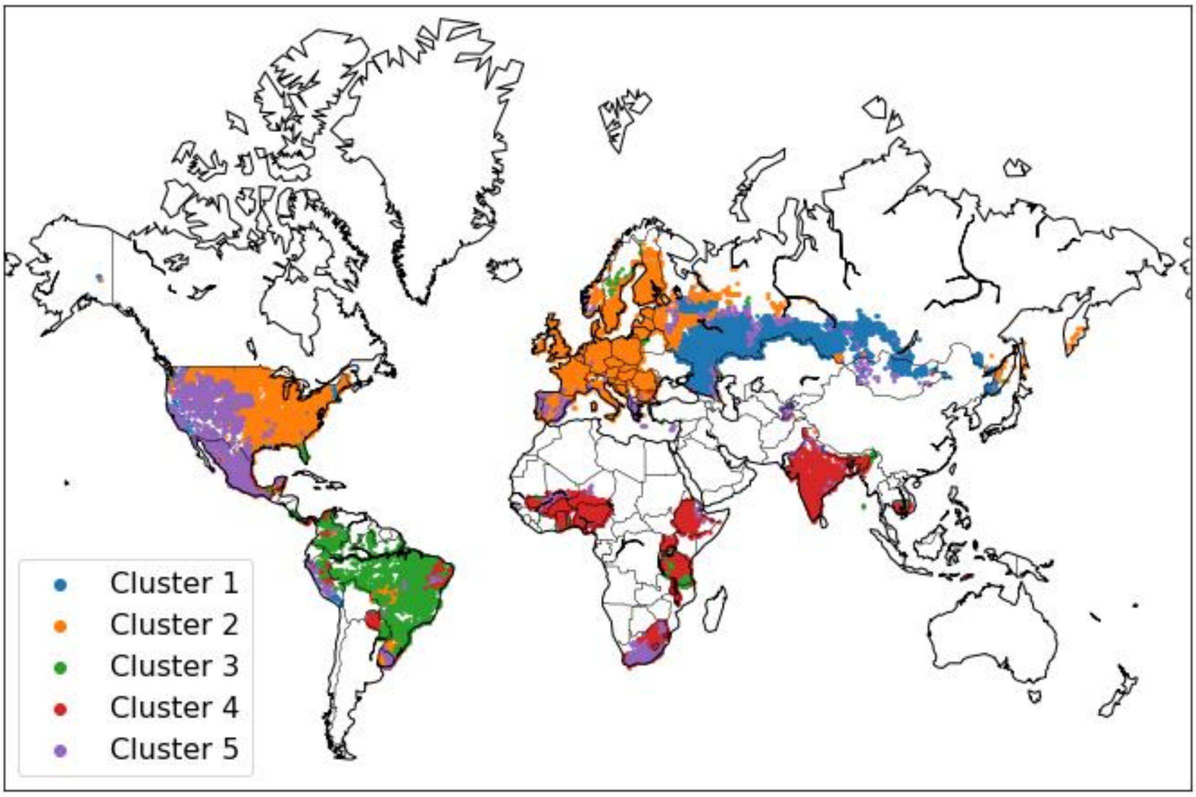

Figure 22 - Spatial distribution for 5 clusters with Random State 2 (Mercator Projection)

In the comparison, Figure 5 in section 3.3 and Figure 22 are used. A clear visual difference to Figure 22 is that the colour pattern is different. This is due to the arbitrary ordering of the clusters identified, it has no influence on the cluster performance. The order of cluster identification changes, but the clusters itself will remain the same. There are still clusters in Europe/Central and East of USA, India/East and West coast of Afrika, South America, Central Asia and West coast USA/Mexico. The major spatial locations do not change by changing the random state, which is in line with the influence of random state on the determination of optimal number of clusters. Minor changes can be identified in spots where in previous analysis also spots where identified, the previous identified spots are enlarged or reduced.

4.3 Construction of features

The construction of the features will be discussing in two ways, the first point is the construction of the current features and the second point is the construction of potential alternative features.

During the preparation of the agricultural data, the classes of the farming class, crop type and farming system were transformed by a one-hot encoding procedure to give a better presentation of the presence of smallholder, the different crop types and farming system for the locations. That is a common way of representing class data, however for the soil characteristics a bucketing procedure was used by averaging or choice of dominance in the bucket. A similar encoding procedure could have been used in order to give a better presentation of the nutrient availability classes and the slope classes instead of the choice of dominance. By choosing the dominant value of the bucket, data which is not dominant is neglected in the representative value. By linking the slope classes to the harvest area, in a comparable way as the agricultural features are constructed, would give a better representation of the location and slope of agricultural area. It is hard to say how the adjustment would change the clustering procedure and performance, but it could give more context to a cluster.

The total regional water footprint of a 30arcmin grid cell could give more context to the water usage of that region or farm, because on average 70% of the blue water is uses for agricultural purposes. The green water footprint can be linked to the precipitation and evapotranspiration, while blue water footprint give an indication of access to water subsurface aquifers and the grey water footprint which represents the pollutant load caused by the household, industry and agricultural processes based on water quality standard (Lovarelli, Bacenetti, & Fiala, 2016).

In order to accomplish a more sustainable agriculture and food security the water productivity is an essential element in improvement. The water productivity defines the amount of agricultural output per unit of water used (Descheemaeker, et al., 2013). By looking at patterns with diverging water productivity, interesting locations could be identified which similar water conditions but different productivities. Improving water productivity can be done in many ways such as choice of crop variety or adapted crop type choice (Descheemaeker, et al., 2013). The patterns could help identify the difference between regions.

The use of fertilizer is different in all parts of the world, in combination with the applied policies for the use of fertilizer. The paper of (Marenya & Barrett, 2009) makes some suggestion on the productivity enhancing role played by fertilizing and the limited use in parts of Afrika compared to different parts of Asia and Latin America. The use of fertilizer can give an indication on soil quality of the agricultural land. The paper of (Rousseau, Fonte, Téllez, Van der Hoek, & Lavelle, 2013) suggest more indicator of the soil quality such as macrofauna, chemical fertility and physical properties of the soil.

For all features that are discusses, the data availability and feasibility of implementation in an unsupervised machine learning algorithms is not assessed. A combination of existing data and newly added data in the feature construction could give a better representation of an area. The existing data, newly added data and the feature construction would all depend on the aim and purpose of the research itself.

4.4 Smallholder criteria

The criterion: ‘A smallholder is a farm with an harvest area of equal or less than two hectares’, used for the research which is in line with the definition of the Sustainable Development Goals (SDG) of the (United Nations, 2017), may not be the representative criteria for all different farms spread over the world. According to (Lowder, Skoet, & Raney, 2016) one criterion for all farms in the world is not suitable. The criteria would highly depend on the scale and the purpose of the research itself. In order to make a consistent interpretation of the results, one criteria of smallholders is necessary. For smaller, continental size areas other criteria would be more suitable. Low-income regions such as Asia, Latin America and Afrika have a significantly large quantities of farm below the criteria while high income regions such as Europe and parts of North America have a very slim quantity of farms below the criteria. Counties as China and India have more than half of the total amount of farms in the world while they have less than a quarter of the total agricultural area of the world. Looking at the distribution of agricultural land and number of farmers in the area, another choice could be made for the criterion of smallholders.

5. Conclusion & Recommendations

This chapter gives answer to the main research question posed in the research objective and questions in section 1.3 of the Introduction. Additionally, recommendations are offered for further research to smallholder pattern in the world.

5.1 Conclusion

It can be concluded that the research shows 5 clear smallholder patterns of agricultural- water-soil combination on a global scale. One located in Europe and east part of the USA, one located in central Asia, one located in South America, one located in western part of USA and Mexico and one located in India and central parts of Afrika. The clusters result from an unsupervised machine learning algorithm K-mean by the use of 30 features which describe a spatial location. The K-mean algorithm is a suitable tool to identify smallholder patterns for different characteristics. The performance of the algorithm is acceptable for the research, which does not mean that the smallholder patterns are generic interpretable. The result of the algorithm highly depends on the choices made in research, which functions as the input. A clear relationship of smallholder to one of the water characteristics and soil characteristic is absent in the patterns because of the wide range the data is assigned to the clusters. That can also be caused by the high number of features and small number of clusters. However, a higher number of clusters would not indicate at a better performance of the unsupervised machine learning algorithm.

The research shows a clear distinction between the influence of the features by the creation of the smallholder patterns, also called the dominant factor. The most influential factors are the Water Scarcity, Groundwater level and the Global Soil Organic Carbon content. The dominant factor does not indicate a direct relationship between the feature and smallholders but only the influence of creation of the patterns. The dominant factors are specific for this research and not directly transferable to other research, because the dominant factor is relative to the data that is included in this research.

The robustness of the choices for the optimal number of clusters and the considered random state are well determined. The choice of optimal number of clusters has definitely influence on the formation of clusters while it does not increase the performance of the clustering itself. While the choice of random state does not influence the formation of clusters and no performance increase.

In it can be concluded that the locations of the patterns are well spread over the continents which have similar attributes. That information could be used in future research.

5.2 Recommendations

The recommendations give both advise on practical aspects as well of future research suggestions.

A practical recommendation is to start with what kind of features will be included in the analysis and try to find data that accomplish the features. Doing it the other way around than this research did, could create another interesting description of the locations and the patterns. A more details investigation on the possible interesting feature would give a better understanding of smallholder patterns in combination with water characteristics and soil characteristics.

On other practical recommendation is to look into an alternative crop grouping schema, which was used to compress the datapoint in terms of crop type from 42 crop types to 10 crop group types, to use to group the data during data preparation. The recommendation is related to the domination of one group, namely cereals, compared to the other groups in the current grouping schema. In order to prevent the domination of one specific crop group a more equal representation would be preferred, even if groups have to be merged or split.

The significance of both smallholders and non-smallholders in determining the dominant factor is evident from Table 8 in section 3.2, where their contributions are equal. The balanced influence of smallholders and non-smallholders presence in patterns formation suggests that, under the research conditions, the smallholder criteria is valid. Nevertheless, I would recommend reducing the global scale of the research and adopting a more generalized definition of smallholders tailored to the specific area. Analysing the cluster outcome and spatial locations can guide the reduction in global scale. This adjustment enables the exploration of relationships between smallholder patterns and potential policies, offering a value addition to the research.

In future research it is advised to include multiple unsupervised machine learning algorithms in order to compare the differences in results of the smallholder pattern creation. Although the current research indicates at a well performing clustering algorithm, other algorithms could both support current results and identify alternative patterns, which both can be helpful for uncovering relationships between smallholders, water characteristics and soil characteristic.

6. References

Aguilar, F., Hendrawan, D., Cai, Z., Roshetko, J., & Stallmann, J. (2020). *Smallholder farmer resilience to water scarcity.* Environment, Development and Sustainability. doi:https://doi.org/10.1007/s10668-021-01545-3

Alghofaili, Y. (2021). *Interpretable K-means: Cluster Feature Importance.* Towards Data Science.

Berry, M. W., Mohamed, A., & Yap, B. W. (2020). *Supervised and Unsupervised Learning for Data Science.* Switzerland: Springer Nature. doi:https://doi.org/10.1007/978-3-030-22475-2

Bishop, C. (2006). *Pattern Recognition and Machine Learning.* Springer Science.

Chikowo, R., Zingore, S., Snapp, S., & Johnston, A. (204). *Farm typologies, soil fertility veriability and nutrient management in smallholder farming in Sub-Saharan Africa.* Dordrecht: Springer Science+Business media. doi:DOI 10.1007/s10705-014-9632-y

Cosgrove, W., & Loucks, D. (2015). *Water management: Current and future challanges and research directions.* New York, USA: AGU Publications. doi:https://doi.org/10.1002/2014WR016869

Cui, M. (2020). *Introduction to the K-Means Clustering Algoritm based on the Elbow Method.* Canada: Clausius Scientific Press. doi:DOI: 10.23977/accaf.2020.010102

Descheemaeker, K., Bunting, S., Bindrahan, P., Muthuri, C., Molden, D., & Beveridge, M. (2013). *Increasing water productivity in agriculture.* Managing water and agroecosystems for food security. doi:https://doi.org/10.1079/9781780640884.0104

Ergin, I., Conforti, P., & Khalil, C. (2019). *Methodology for computing and monitoring the sustainable development goal indicators 2.3.1 and 2.3.2.* Food and Agriculture Organization of the United Nations. Rome: FAO Statistics Division (ESS/18-14).

Fan, Y., Li, H., & Miguez-Macho, G. (2013). *Global Patterns of Groundwater Table Depth.* Science 339. doi:https://doi.org/10.1126/science.1229881

FAO. (2022). *The State of the World's Land and Water Resources for Food and Agriculture - Systems at breaking point. Main report.* Rome. doi:https://doi.org/10.4060/cb9910en

Firdose, T. (2023). *Treating Outliers using IQR and Percentile approach*. Opgehaald van https://tahera-firdose.medium.com/treating-outliers-using-iqr-and-percentile-approach-part-2-9d8c4ec55af7#:~:text=Commonly%20used%20thresholds%20are%2095th,desired%20level%20of%20outlier%20removal.

Food+Agri Business. (2021, February 24). *Sustainable Food*. Opgehaald van Foto by Peter Roek: https://www.foodagribusiness.nl/lto-wil-strokenteelt-niet-opgedrongen-krijgen/

GAEZv4 Data Portal. (2023). *Global Agro-Ecological Zones*. Opgehaald van Food and Agriculture Organization of the United Nations: https://gaez.fao.org/

Giller, K., Delaune, T., & Silva, J. (2021). The future of farming: Who will produce our food? In *Food Security* (pp. 1073-1099). doi:https://doi.org/10.1007/s12571-021-01184-6

Giordano, M., Barron, J., & Unver, O. (2019). *Water scarcity and challenges for smallholder agriculture.* Elsevier Inc. doi:https://doi.org/10.1016/B978-0-12-812134-4.00005-4

GloSIS Data Portal. (2023). *Global Soil Organic Carbon map*. Opgehaald van Food and Agriculture Organization of the United Nations: https://data.apps.fao.org/glosis/

Guarín, A., Rivera, M., Pinto Correia, T., Guiomar, N., Sumane, S., & Moreno-Pérez, O. (2020). *A new typology of small farms in Europe.* London, UK: Elsevier. doi:https://doi.org/10.1016/j.gfs.2020.100389

Igual, L., & Sequí, S. (2017). *Introduction to Data Science.* Switzerland: Springer Nature. doi:DOI 10.1007/978-3-319-50017-1

Khalil, C., Conforti, P., Ergin, I., & Gennari, P. (2017). *Defining small-scale food producers to monitor target 2.3. of the agenda for sustainable development.* Food and Agriculture Organization of the United Nations. Rome: FAO Statistics Division (ESS 17-12).

Kuhn, M., & Johnson, K. (2019). *Feature Engineering and Selection - A Practical Approach for Predictive Models* (1st Edition ed.). doi:https://doi-org.ezproxy2.utwente.nl/10.1201/9781315108230

Kumbhara, T. A., & Chobe, S. V. (2014). An overview of association rule mining algorithms. In *International Journal of Computer Science and Information Technologies* (pp. 927-930).

Lovarelli, D., Bacenetti, J., & Fiala, M. (2016). *Water Footprint of crop productions: A review.* Department of Agricultural and Environmental Sciences. doi:https://doi.org/10.1016/j.scitotenv.2016.01.022

Lowder, S., Skoet, J., & Raney, T. (2016). *The number, size and distribution of farms, smallholder farms and family farms worldwide.* Food and agriculture organization of the United Nations. Rome, Italy: The Elsevier. doi:http://dx.doi.org/10.1016/j.worlddev.2015.10.041

Mahesh, B. (2020). *Machine Learning Algorithms - A review.* Research Gate Impact Factor. doi:DOI: 10.21275/ART20203995

Marenya, P., & Barrett, C. (2009). *Soil quality and fertilizer use rates among smallholder farmers in western Kenya.* doi:https://doi.org/10.1111/j.1574-0862.2009.00398.x

Mekonnen, M., & Hoekstra, A. (2016). *Four billion people facing severe water scarcity.* Science Advance 2 (e1500323). doi:https://doi.org/10.1126/sciadv.1500323

Mishra, V. (2013). Food Security Implications of Climate Variability and Climate Change. In *Climate Vulnerability.* doi:https://doi.org/10.1016/B978-0-12-384703-4.00223-9

Naghizadeh, A., & Metaxas, D. S. (2020). *Condensed Silhouette: An Optimized Filtering Process for Cluster Selection in K-Means.* Elsevier. doi:https://doi.org/10.1016/j.procs.2020.08.022

National Dictionary. (2022). *Groot woordenboek van de Nederlandse taal* (16th Edition ed.). VBK Media.

Patel, H. (2021). *What is Feature Engineering - Importance, Tools and Techniques for Machine Learning.* Towards Data Science. Opgehaald van https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10

Polzer, D. (2023). *7 of the Most Used Feature Engineering Techniques.* Towards Data Science. Opgehaald van https://towardsdatascience.com/7-of-the-most-used-feature-engineering-techniques-bcc50f48474d

Prăvălie, R., Nita, I.-A., Patriche, C., Niculiță, M., Birsan, M.-V., Roșca, B., & Bandoc, G. (2021). *Global changes in soil organic carbon and implications for land degradation neutrality and climate stability.* Environmental Research. doi:https://doi.org/10.1016/j.envres.2021.111580

Rawat, T., & Khemchandani, V. (2019). *Feature Engineering (FE) Tools and Techniques for Better Classification Performance.* doi:DOI: 10.21172/ijiet.82.024

Rousseau, L., Fonte, S., Téllez, O., Van der Hoek, R., & Lavelle, P. (2013). *Soil macrofauna as indicators of soil quality and land use impacts in smallholder agroecosystems of western Nicaragua.* Ecological Indicators. Elsevier. doi:https://doi.org/10.1016/j.ecolind.2012.11.020

Scaler Topics. (2023). *Handling outliers in data science*. Opgehaald van https://www.scaler.com/topics/data-science/handling-outliers-in-data-science/

Shumaila, M. N. (2021). *A comparison of K-means and Mean Shift Algorithms.* Preprints. doi:https://doi.org/10.20944/preprints202108.0140.v1

Silander, J. J. (2001). Temperate Forests. In *Encyclopedia of Biodiversity* (pp. 112-127). doi:https://doi.org/10.1016/B978-0-12-384719-5.00142-8

Sinaga, K. P., & Yang, M.-S. (2020). *Unsupervised K-mean Clustering Algorithms.* DOI: 10.1109/ACCESS.2020.2988796.: IEEE Access.

Su, H. (2023). *Internal Database UT.*

Su, H., Willaarts, B., Luna-Gonzalez, D., Krol, M. S., & Hogeboom, R. J. (2022). Gridded 5 arcmin datasets for simultaneously farm-size-specific and crop-specific harvested areas in 56 countries. In *Earth Syst. Sci. Data* (pp. 14, 4397- 4418). doi:https://doi.org/10.5194/essd-14-4397-2022

Turner, R., Fugetta, A., Lavazza, L., & Wolf, A. (1999). *A conceptual basis for feature engineering.* Elsevier. doi:https://doi.org/10.1016/S0164-1212(99)00062-X

United Nations. (2017). *Defining small scale food producers to monitor target 2.3. of the 2030 agenda for sustainable development.* Food and Agriculture Organization, Statistics Division, Rome. Opgehaald van SDG indicators - Metadata repository: https://www.fao.org/3/a-i6858e.pdf

7. Appendices

Appendix A: Grouping schemas

Table 10 - Grouping schema for farming class

Farming Class group	Criteria
Smallholders	Harvest area below or equal to 2 hectares
Non-smallholders	Harvest area more than 2 hectare

Table 11 - Grouping schema for crop type

Crop group	SPAM-based crop	Crop full name
Stimulates	Acof	Arabica Coffee
	Rcof	Robusta Coffee
	Coco	Cocoa
	Teas	Tea
	Toba	Tobacco
Fruits	Bana	Banana
	Plnt	Plantain
	Temf	Temperate Fruit
	Trof	Tropical Fruit
Vegetables	Vege	Vegetables
Cereals	Barl	Barley
	Maiz	Maize
	Pmil	Pearl Millet
	Rice	Rice
	Smil	Small Millet
	Sorg	Sorghum
	Whea	Wheat
	Ocer	Other Cereals
Pulses	Bean	Bean
	Chic	Chickpea
	Cowp	Cowpea
	Lent	Lentil
	Pige	Pigeon Pea
	Opul	Other Pulses
Roots & Tubers	Cass	Cassava
	Pota	Potato
	Swpo	Sweet Potato
	Yams	Yams
	Orts	Other Roots
Oil crops	Cnut	Coconut
	Grou	Groundnut
	Oilp	Oilpalm
	Rape	Rapeseed
	Sesa	Sesame Seed
	Soyb	Soybean
	Sunf	Sunflower
	Ooil	Other Oil Crops
Fibres	Cott	Cotton
	Ofib	Other Fibre Crops
Sugar crops	Sugb	Sugarbeet
	Sugc	Sugarcane
Rest	Rest	Rest of Crops

Appendix B: Sample location in Feature Engineering

Table 12 - Dataset example for one-hot encoding procedure

30 arcmin Coordinate (X/Y)	Index 30arcmin	Farming Class	Crop Type	Farming System	Harvest Area (Ha)
9,25 / 55.75	49340	Non-smallholders	Cereals	H	35431,29
9,25 / 55.75	49340	Non-smallholders	Cereals	I	12258,73
9,25 / 55.75	49340	Non-smallholders	Cereals	L	6253,16
9,25 / 55.75	49340	Non-smallholders	Oil crops	H	6086,83
9,25 / 55.75	49340	Non-smallholders	Pulses	I	234,7
9,25 / 55.75	49340	Non-smallholders	Roots & tubers	I	2791,87
9,25 / 55.75	49340	Non-smallholders	Vegetables	I	821,27
9,25 / 55.75	49340	Non-smallholders	Fruits	H	112,2
9,25 / 55.75	49340	Non-smallholders	Fruits	L	118,33
9,25 / 55.75	49340	Non-smallholders	Fruits	S	44,97
9,25 / 55.75	49340	Non-smallholders	Sugar crops	I	5407,62
9,25 / 55.75	49340	Smallholders	Pulses	I	1,06
9,25 / 55.75	49340	Smallholders	Sugar crops	I	56,82
9,25 / 55.75	49340	Smallholders	Oil crops	H	7,24
				Total Harvest Area	69626,09

Appendix C: Random state test – Range [0-10]

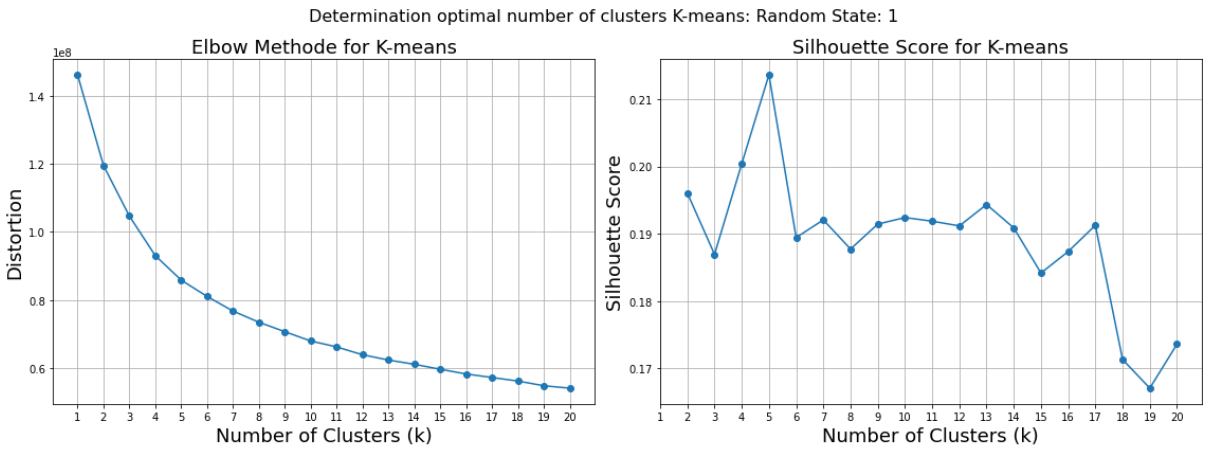

Figure 23 - Results of Elbow Method and Silhouette Score for K-means (Random State 1)

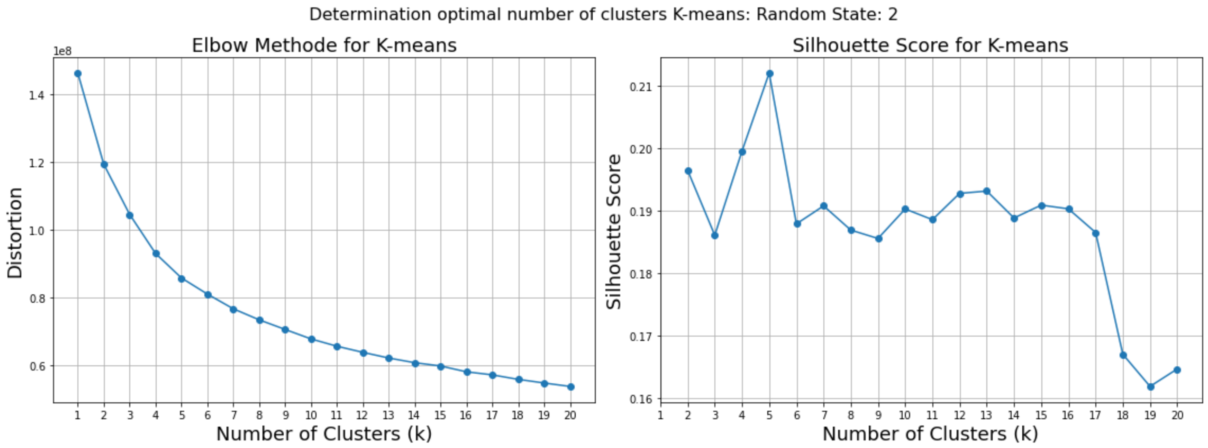

Figure 24 - Results of Elbow Method and Silhouette Score for K-means (Random State 2)

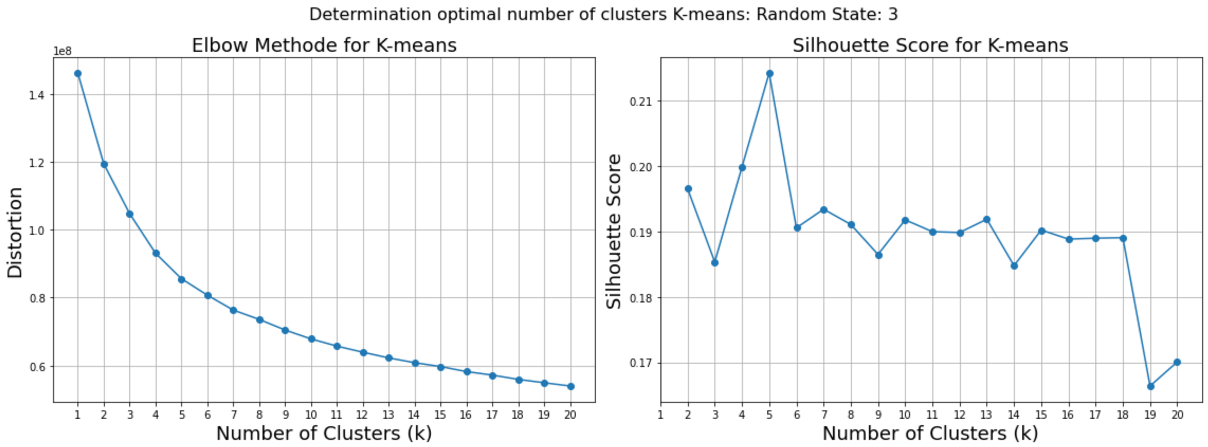

Figure 25 - Results of Elbow Method and Silhouette Score for K-means (Random State 3)

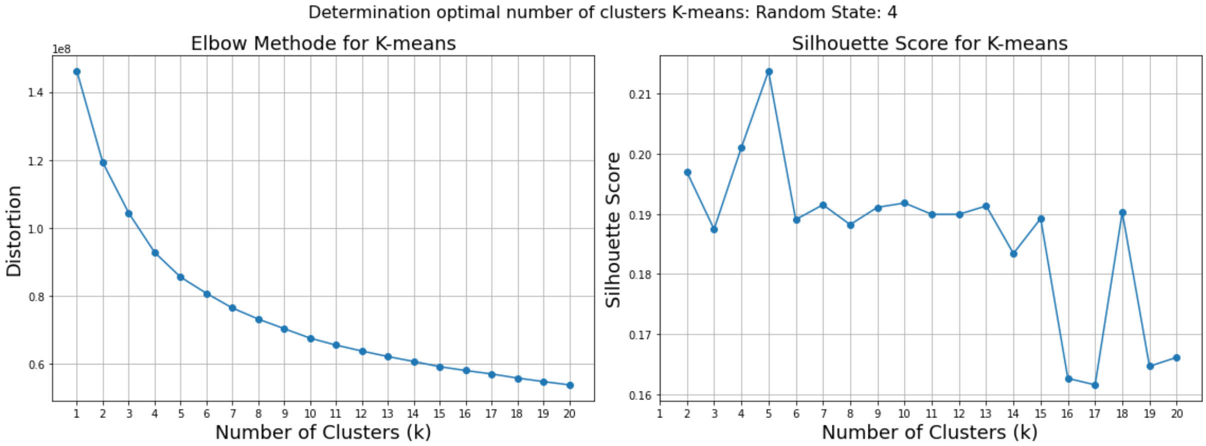

Figure 26 - Results of Elbow Method and Silhouette Score for K-means (Random State 4)

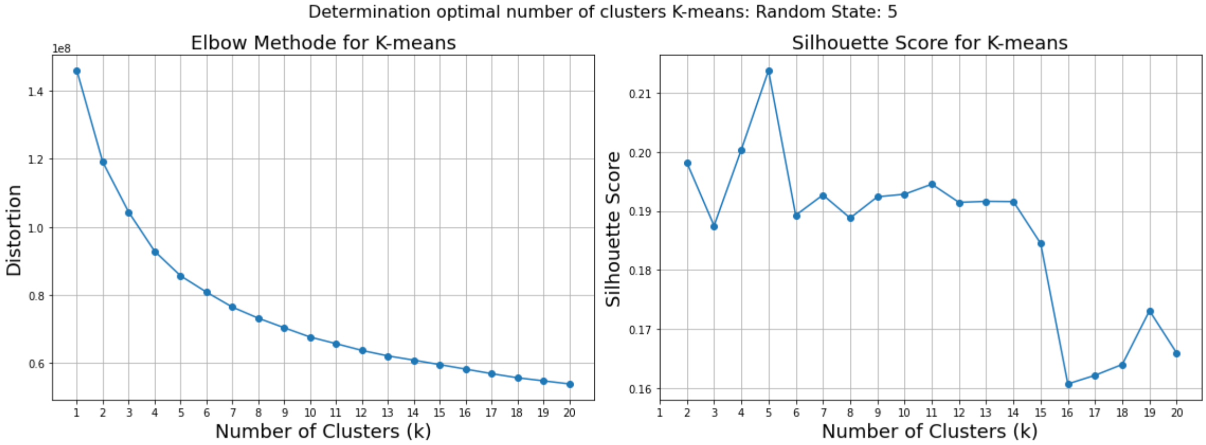

Figure 27 - Results of Elbow Method and Silhouette Score for K-means (Random State 5)

Appendix D: Maximum number of cluster to consider

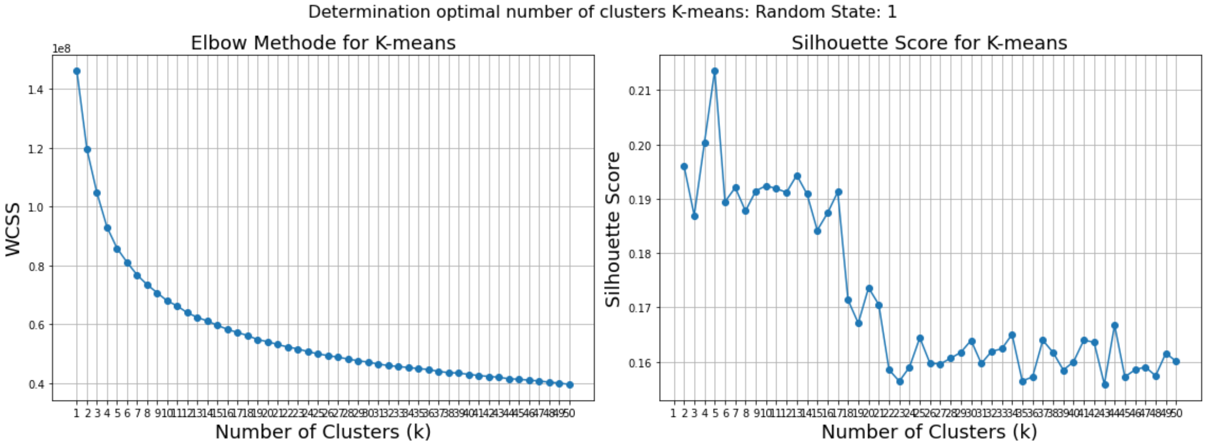

Figure 28 - Results of determination of maximum number of cluster for Elbow Method and Silhouette Score for K-means (Random State 1)

Appendix E: Data distribution all features per cluster (normalized [0-100])

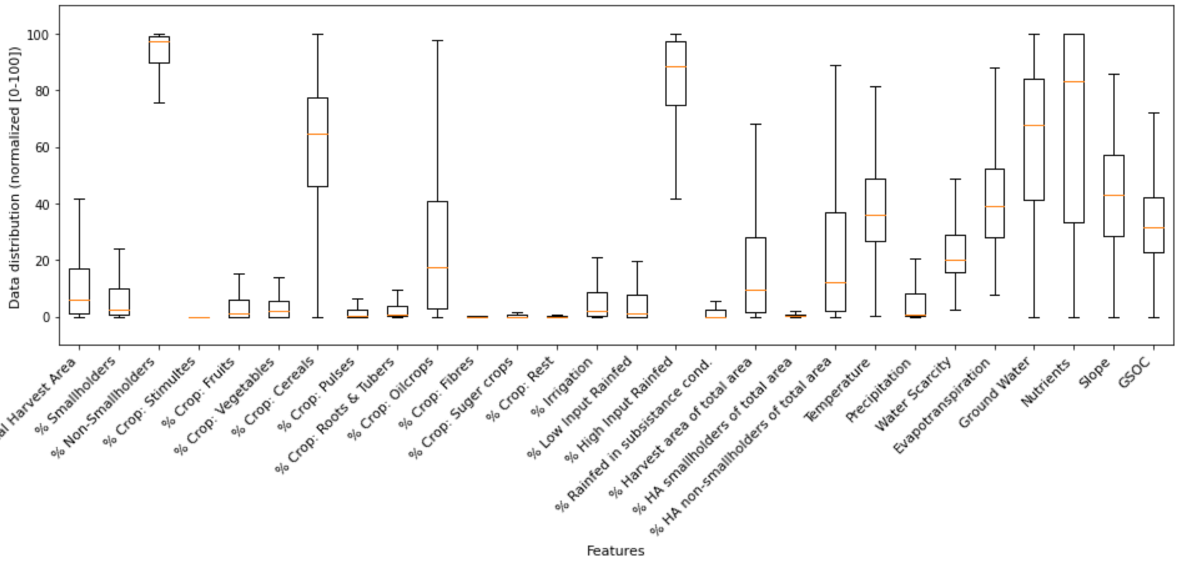

Figure 29 - Data distribution of all features for first cluster (normalized)

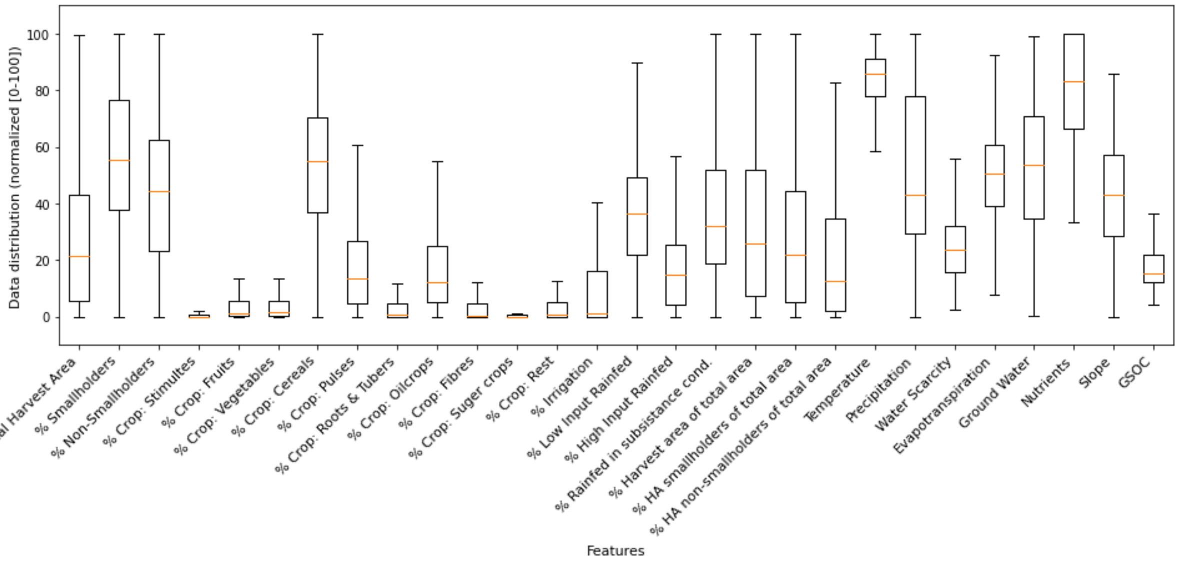

Figure 30 - Data distribution of all features for second cluster (normalized)

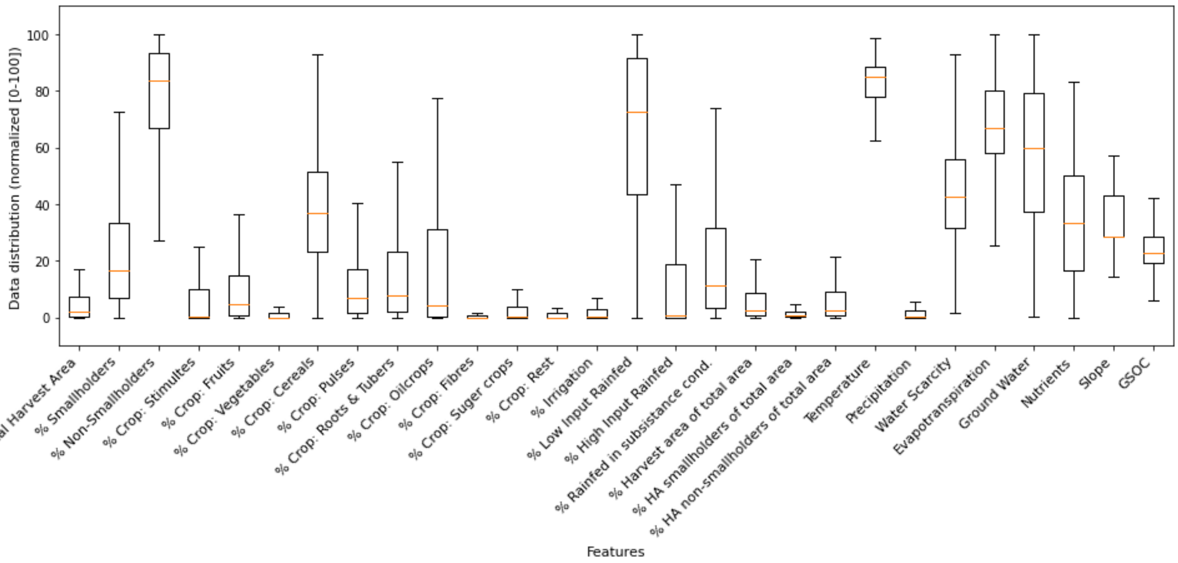

Figure 31 - Data distribution of all features for third cluster (normalized)

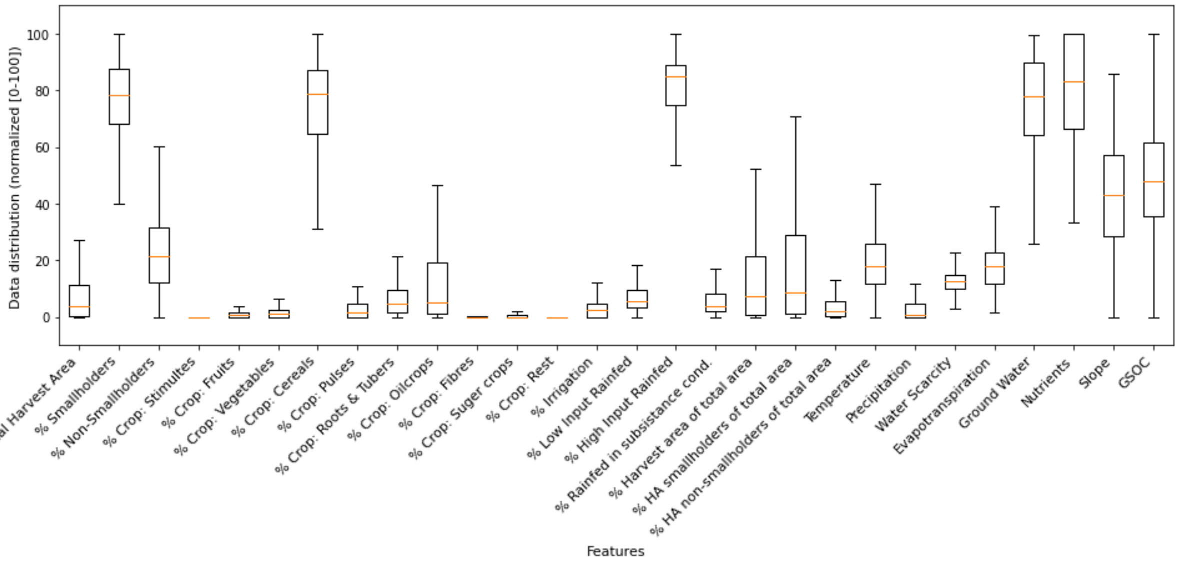

Figure 32 - Data distribution of all features for fourth cluster (normalized)

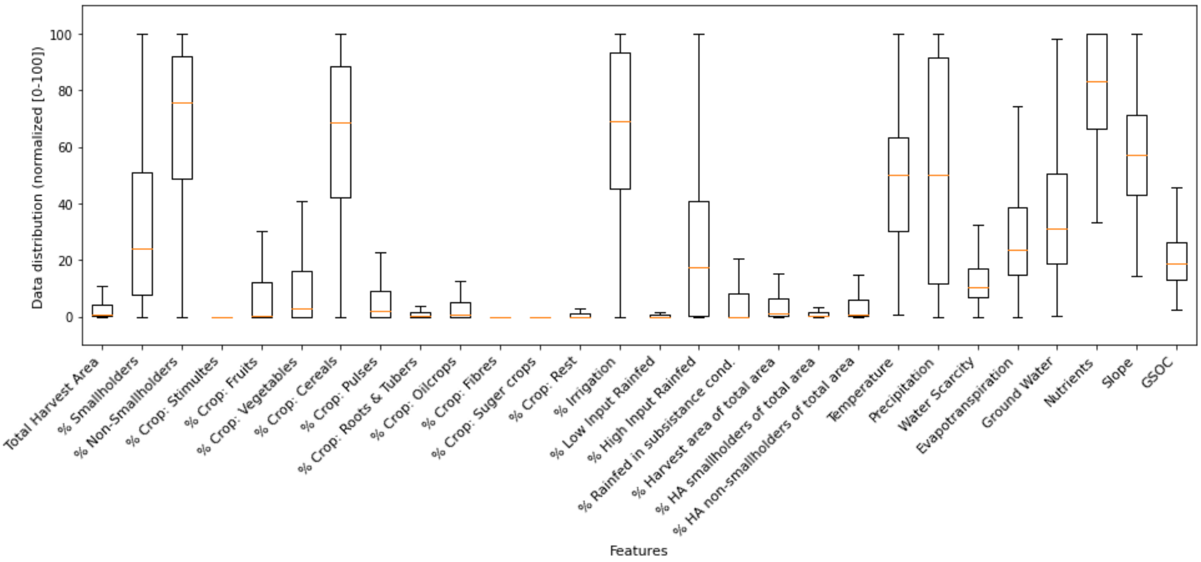

Figure 33 - Data distribution of all features for fifth cluster (normalized)