

# Assessing classification performance for sampled remote sensing data

*By*

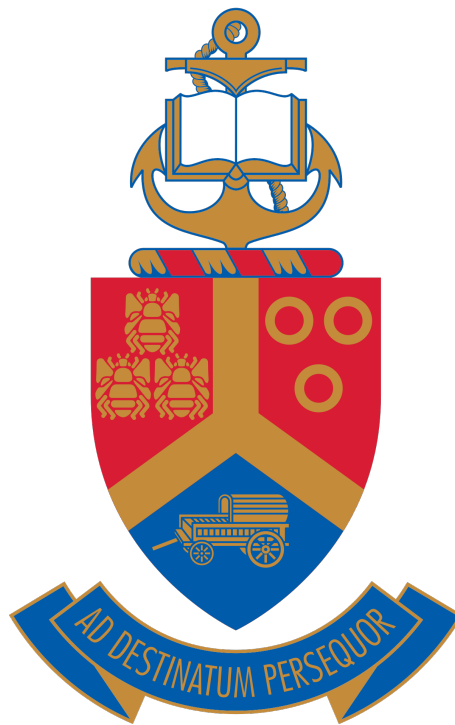
Tshepiso Selaelo Rangongo  
17052395

*Supervisors:*

Renate Thiede  
Inger Fabris-Rotelli

Submitted in partial fulfilment of the requirements for the degree

In the Department of Statistics  
In the Faculty of Natural and Agricultural Sciences  
University of Pretoria



31 October 2022

# Declaration

I declare that the dissertation, which I hereby submit for the degree MSc e-Science at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution"



Tshepiso Selaelo Rangongo

31 October 2022

## *Abstract*

The volume of big data increases daily. Big data poses challenges in storage, management, processing, analysis and visualisation. One technique of handling big data is the use of subset or sample that is good representation of the data. For storage alleviation purposes, a subset of the big data can be obtained from metadata. This paper obtains metadata of a remote sensing image dataset for crop classification. This research proposes a sampling algorithm which makes use of multivariate stratification with the aim of obtaining a sample that best represents the population while minimising the number of images sampled. The proposed sampling algorithm performs effectively on a big spatial image dataset of crop types. The results are assessed by measuring the number of images sampled and as well as matching the proportionality of the population crop percentages. The samples obtained from the proposed algorithm are then used for land cover classification. An ensemble method called random forest is trained on the different samples and accuracy is assessed. Precision, recall and F1-scores per crop type are computed as well as the overall accuracy. The random forest classifier performed best on the proposed sample with the least number of images, followed by the one with the second least. The classifier performed better on the proposed samples than it did on the random samples as the proposed samples contained the most informative data. This research encourages the use of metadata for classification purposes as well as an effective way of sampling big data for crop classification.

# Acknowledgements

I gratefully acknowledge the funding received towards my Masters in e-Science from the NEPTTP Bursary programme and my supervisors for their support. This research is approved under ethics number NAS124/2019.

# Contents

<b>Declaration</b>	<b>i</b>
<b>Abstract</b>	<b>ii</b>
<b>Acknowledgements</b>	<b>iii</b>
<b>List of Figures</b>	<b>vi</b>
<b>List of Tables</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Spatial Big Data</b>	<b>4</b>
2.1 Literature review . . . . .	4
2.2 Database construction . . . . .	8
2.2.1 Data summary . . . . .	8
2.2.2 Metadata construction . . . . .	11
2.2.3 Summary . . . . .	13
<b>3 Sampling</b>	<b>14</b>
3.1 Sampling theory . . . . .	14
3.1.1 Random sampling . . . . .	18
3.1.2 Stratified sampling . . . . .	19
3.1.3 Cluster sampling . . . . .	21
3.2 A multivariate stratified sampling algorithm . . . . .	23
3.3 Simulation . . . . .	26
3.4 Summary . . . . .	32
<b>4 Classification</b>	<b>33</b>
4.1 Feature engineering and selection . . . . .	33

4.1.1	Feature engineering . . . . .	33
4.1.2	Feature selection . . . . .	34
4.1.3	Implementation . . . . .	34
4.2	Land cover classification algorithm . . . . .	36
4.2.1	Random Forest . . . . .	36
4.2.2	Accuracy measures . . . . .	37
4.2.3	Implementation and Results . . . . .	40
	10% sample . . . . .	40
	20% sample . . . . .	45
	30% sample . . . . .	48
4.3	Summary . . . . .	52
<b>5</b>	<b>Discussion</b>	<b>53</b>
	Summary . . . . .	69
<b>6</b>	<b>Conclusion</b>	<b>71</b>
	<b>Bibliography</b>	<b>74</b>

# List of Figures

2.1	Examples of geospatial data before and after adverse events. . . . .	7
2.2	An example of the metadata of a picture of a cat. . . . .	8
2.3	A single image tile from the data shown over 12 different bands . . .	9
2.4	Cloud coverage of the same area over four different dates. . . . .	10
2.5	An illustration of the fields and crop types of the area with tile ID 1114.	10
2.6	Proportions of the crop types using area coverage as well as number of fields. . . . .	11
2.7	General information that applies on all images. . . . .	12
2.8	Location wise information that applies on all images of the same area of land. . . . .	13
3.1	Clear depiction of the difference between bias, accuracy and preci- sion [44]. . . . .	18
3.2	Illustration of the difference between stratification and one-stage clus- tering [9] . . . . .	23
3.3	The difference between one-stage and two-stage cluster sampling [36].	24
3.4	Euclidean norms between (10%-100%) sample and population. . . . .	31
3.5	The bounds of Euclidean norms per sample size as well as averages.	32
4.1	Image bands importances by the mutual information regression method.	35
4.2	Image bands importances by the F-test method. . . . .	35
4.3	Random forest illustration. <sup>8</sup> . . . . .	37
4.4	Interpretations of different precision and recall values. . . . .	39
4.5	The proportion of the crop types using area coverage and number of fields. . . . .	41
4.6	Accuracy per crop type. . . . .	42
4.7	Crop proportions using number of fields. . . . .	42

4.8	Area-wise proportions of the crop types in the proposed sample, random sample and the population. . . . .	43
4.9	Difference between achieved precision and recall values in the proposed sample and the random sample. . . . .	44
4.10	F1-scores between the random sample and the proposed sample. . . . .	44
4.11	Accuracy per crop type. . . . .	46
4.12	Number of fields of different crop types in the two samples. . . . .	47
4.14	Difference between achieved precision and recall values in the proposed sample and the random sample. . . . .	47
4.13	Proportions of the crop types using area coverage. . . . .	48
4.15	F1-scores between the proposed sample and random sample. . . . .	49
4.16	Precision per crop type . . . . .	49
4.17	Number of fields per crop type in the two samples. . . . .	50
4.18	Proportions of the crop types using area coverage. . . . .	51
4.19	Difference between achieved precision and recall values in the proposed sample and the random sample. . . . .	51
4.20	F1-scores between the proposed sample and random sample. . . . .	52
5.1	Visibility of the fields and its labels through the blue (B02), green (B03) and red (B04) bands. . . . .	56
5.2	Average field sizes of each crop type. . . . .	58
5.3	Labelled and unlabelled data between the proposed sample and random sample. . . . .	59
5.4	Differences in precision and recall values in the 10% and 20% proposed samples. . . . .	61
5.5	Labelled and unlabelled data in the 20% proposed and random sample with same number of images. . . . .	62
5.6	Differences in the precision and recall values for the random samples. . . . .	64
5.7	Differences in the precision and recall values in the 20% and 30% proposed sample. . . . .	65
5.8	F1-scores of the crop types in the 10%, 20% and 30% proposed samples. . . . .	67
5.9	Labelled and unlabelled data in the 30% proposed and random sample with same number of images. . . . .	68
5.10	Differences in the precision and recall values for the random samples. . . . .	68



5.11 Accuracy values of the classifier when trained on the different random and proposed samples. . . . . 69

# List of Tables

3.1	Population statistics, description and their estimators in stratified sampling. . . . .	20
3.2	<i>psu</i> level population quantities used in cluster sampling. . . . .	22
3.3	<i>ssu</i> level population quantities used in cluster sampling. . . . .	22
3.4	Achieved area coverage percentage using two different <i>cropAmax</i> parameter. . . . .	27
3.5	Achieved area coverage percentage using two different <i>cropBmax</i> parameter. . . . .	28
3.6	Achieved area coverage percentages over the remaining iterations. . . . .	28
3.7	Achieved Sample size per desired sample size and <i>cropAmax</i> . . . . .	29
3.8	Number of images per desired sample size and <i>cropAmax</i> . . . . .	29
3.9	Achieved sample size per desired sample size and <i>cropBmax</i> . . . . .	30
3.10	Euclidean norm between population and sample proportions per desired sample size and <i>cropAmax</i> . . . . .	30
4.1	An error matrix in land-cover classification . . . . .	38
4.2	Precision, recall and F1-scores per crop type. . . . .	41
4.3	Precision, recall and F1-scores per crop type. . . . .	45
4.4	Precision, recall and F1-scores per crop type . . . . .	48

# Chapter 1

## Introduction

The volume of data worldwide grew from 33 trillion gigabytes in 2018 to 71 trillion gigabytes in 2021 and this is predicted to be 180 trillion gigabytes of data in 2025<sup>1</sup>. This refers to items of information created, captured, copied and consumed. Such large and diverse sets of information are called big data. Big data is characterised by high volumes, high velocity and high variety. Big data arise due to the fact that humans produce close to 2 quintillion gigabytes of data each day using various sources of data such as social media, IoT devices, etc as well as different formats of data including numeric, images, text, etc. The first trace of big data dates back to 1663 but was only recognised by the world in the 1800s.

Big geospatial data is information associated with a location on or near the surface of the earth. Remote sensing is one technique by which geospatial data can be obtained. The increasing amount of satellites orbiting the earth (remote sensors) increase the volume, velocity and variety of geospatial data. Information from remote sensors is used for various purposes including weather and catastrophe forecasting.

The storage, management, analysis and visualisation of big geospatial data is difficult due to the complexity of this data. Although strategies such as parallel programming and distributed programming have been implemented to handle big geospatial data, metadata is a simple useful way of handling big data specifically when classification is to be performed [34]. Metadata is data that gives information about data. It summarises big data which alleviates memory requirement in cases where metadata can be used instead of reading all the big geospatial data. Such

---

<sup>1</sup>The Conversation, Science + Technology, The world's data explained, <https://theconversation.com/the-worlds-data-explained-how-much-were-producing-and-where-its-all-stored-159964>

case is in sampling, as one can get a sample from the metadata instead of reading in all the geospatial data.

An alternative to reading all data into memory is obtaining a representative sample of the data. A sample is considered representative if there are certain characteristics of interest of the population that can be estimated from the sample with known accuracy [12]. Examples of sampling techniques include random, systematic, stratified and cluster sampling. Since this research focuses on crop classification, stratified sampling works best. It requires that each unit must belong to only one stratum and in crop classification one crop can only belong to one crop type. Some applications of stratified sampling in remote sensing include the detection of spatial variability amongst peach orchards. This was in turn used to classify trees into homogenous groups (sampling strata), with the aim of decreasing sampling size [44]. Another application is the estimation of crop area using stratified sampling in remote sensing used in [23] and [64] in China.

Classification uses characteristics or features to distinctively identify categories. Sokal [50] defines classification as the arrangement or ordering of objects into sets or groups based on their relationship. If the data is in image format, then this is called image classification. This process categorises all pixels or groups of pixels (objects) to obtain a set of labels [35]. There are two types of image classification, namely supervised and unsupervised classification. Supervised classification uses labelled input and output data while unsupervised classification does not. Supervised classification trains an algorithm on areas that are similar to areas of interest. These training areas are then used to identify spectral signatures and patterns. Examples of supervised classification include maximum likelihood, decision trees, kNN and random forests. Image classification is commonly used to identify different land cover types. This is referred to as land cover image classification.

The final step of classification is calculating the performance of the classification algorithm of choice. The performance can be quantified using accuracy assessment. This is the process of comparing a classified image to a reference to determine the quality of the classification. Accuracy assessments help to evaluate whether a certain algorithm meets specific requirements of its intended purpose. Accuracy of a classified image is obtained by expressing the correctly classified pixels or objects to all pixels or objects. This can be done per category, by comparing the number of correctly classified pixels or objects of a certain category relative to all pixels or

objects belonging to that category.

This research aims to obtain metadata from a land cover geospatial dataset and proposes an algorithm that makes use of multivariate stratified sampling to obtain a sample that gives the best representation of the population. The multivariate population under consideration consists of a large database of remote sensing images of crop fields, for which each image has a varying number of fields, crop types and field sizes. First, the data summary is obtained in the form of a metadata dataframe. Then the metadata itself is used to obtain a desired sample using the algorithm explained in Chapter 2. The aim of the algorithm is to achieve similar proportionality of crop types between the sample and the population as well as minimize the number of images sampled while maximizing the information obtained in the images. We evaluate the usefulness of the proposed algorithm and the effect of parameter choices. The different resulting samples are then used for land cover classification. Random forest, an ensemble method, is trained on the different samples and performance is assessed via accuracy assessment.

Specifically, this mini-dissertation aims to achieve the following:

- Build a metadata structure for the large database of images of crop types.
- Propose a multivariate stratified algorithm.
- Investigate the efficiency and representativeness of the samples from the proposed algorithm.
- Investigate land cover classification performance on various sample sizes obtained using the proposed sampling algorithm.

Chapter 2 provides the literature review of geospatial big data and metadata construction. Chapter 3 covers sampling theory, the proposed algorithm as well as its implementation. Chapter 4 covers classification and implementation. Chapter 5 discusses results while Chapter 6 concludes and proposes future research.

## Chapter 2

# Spatial Big Data

### 2.1 Literature review

Data is defined as individual facts, items of information or statistics. As much as the terms data and information have been used interchangeably, they are not necessarily the same. Data can be transformed to information when viewed in context or post-analysis<sup>2</sup>. Data is measured, collected, reported and analysed and is used in many different sectors such as health care, education, mining and others. It has been described as the oil of the digital economy [53]. The amount of data worldwide is gradually increasing as for some time now. It has been collected by an increasing number of ways such as surveys e.g. online and telephonic surveys, and devices such as mobile devices, aerial devices, cameras, microphones and wireless sensor networks. Many of these devices and techniques are easily accessible by many people. The continuous increasing collection of data has led to what is known as big data.

Big data can be defined as data sets that are large and complex to be dealt with when using traditional data processing software. This may be as a result of many fields/observations which give greater statistical power or more attributes (columns) which introduce complexity to the data hence higher variance<sup>3</sup>. Xialong [24] has defined big data to be a bond that acts as an integration between human society, the physical world and cyberspace. Big data can be divided into two categories, namely data from the physical world and from human society [33]. Data

---

<sup>2</sup>Data vs Information - Difference and Comparison, 2022, [https://www.diffen.com/difference/Data\\_vs\\_Information\\_google\\_vignette](https://www.diffen.com/difference/Data_vs_Information_google_vignette)

<sup>3</sup>What is Big Data?, Big Data, Oracle South Africa, <https://www.oracle.com/za/big-data/what-is-big-data/put/simply/C/big/data/is,been/able/to/tackle/before>

from the physical world can be obtained through scientific experiments and observations (biological and neural data) or sensors (remote sensing data). Data from human society is acquired through human-computer interfaces and brain-computer interfaces and may fall within numerous sectors such as finance, health care and transportation [10, 59].

Big data is an essential factor in the above-mentioned sectors. Data from human society contributed to modern economic activities amongst other essential factors of production such as human capital. Other advantages include using data to support human decisions with outcomes from automatic algorithms. When made available, it provides transparency which firms can utilize to their advantage to grab the attention of investors or potential stakeholders and it can also be used to help innovate new products, services and business models. Big data is also significant for national development (the Big Data Research and Development Initiative by the US [3]), industrial upgrades (use of cloud computation), scientific research and emerging disciplinary research. It helps people better perceive the present and predict the future. [24]

Geospatial data falls under the category of data from the physical world. Geospatial data is information that describes events, objects or features associated with a location on or near the surface of the earth. Geospatial data can be obtained by remote sensing, ground surveying, laser scanning, mobile mapping, geo-tagged web contents and many more techniques. Geospatial data is continuously growing as machinery used to capture it increases yearly. Remote sensing dates back to the 1840s where pictures of the ground were captured by balloonists using photo cameras. This introduced aerial photography during World War I and became fully effective during World War II [4]. Then the first meteorological satellites called the TIROS-1 were developed in the US in 1960 [22]. In 1972, the Earth Resources Technology Satellites, also called the Landsat satellites, were launched, followed by the Earth Observing System (EOS) being launched in 1999 which provided a higher level of processing, a better global coverage and free and easily accessible data [5]. To date, there are approximately 6000 Earth Observations Satellites orbiting the earth<sup>4</sup>.

Examples of remote sensors are the cameras on aeroplanes and satellites that can

---

<sup>4</sup>World Economic Forum, Who owns our orbit: Just how many satellites are there in space, <https://www.weforum.org/agenda/2020/10/visualizing-earth-satellites-sapce-spacex>

take an image of a large area on the surface of the Earth, sonar systems on ships that can take an image of the ocean floor without having to go to the bottom of the ocean [15] and cameras on satellites which monitor the ocean, land and atmosphere of the earth [30]. Figure 2.1[A] and 2.1[B] show examples of geospatial data captured by satellites before and after adverse events, namely a heat wave<sup>5</sup> and drought<sup>6</sup> respectively. It can be seen in both images that the land has changed from being green to being very dry, for example, the brown spots in Figure 2.1[A] indicate dry land. Specific uses of remote sensing include the tracking of clouds which helps to predict the weather [29]. Upcoming catastrophes can be detected so that people can evacuate or prepare the areas that will be affected [55] and large forest fires can be identified by satellites [6]. The tracking of the growth of population in certain areas [63] and the tracking of changes in farmland and forests over time [52] are also use cases of remote sensing. With the many sensors now available, data increases in volume continuously.

Big data has been characterised by the three initial V's by Laney [28], these are volume, variety and velocity. Others V's have been added as time has gone on such as value, veracity, variability and visualization. Evans et al [56] have shown that geospatial big data exhibits at least one of the 3 initial V's along with the other V's introduced later on. The V's will be explained in the context of geospatial big data. High volume is due to a continuous increase of data captured by satellites. Variety is a result of the many different sensors mentioned earlier such as the different satellites capturing images in different bands, i.e. different segments of the electromagnetic spectrum. Velocity is due to the fact that most satellites make frequent visits to the same location, for example the Sentinel-2 satellite captures each location every 5 days. Veracity depends on the accuracy of the data source, for example how accurate is the Landsat satellite compared to the Sentinel-2. All these factors make the handling of data such as the storing, managing, analysing and visualizing of geospatial data difficult.

Two strategies that have been introduced and implemented for geospatial big data handling include parallel and distributed programming [31, 49, 58]. Others have suggested the use of functional programming concepts or languages such as

---

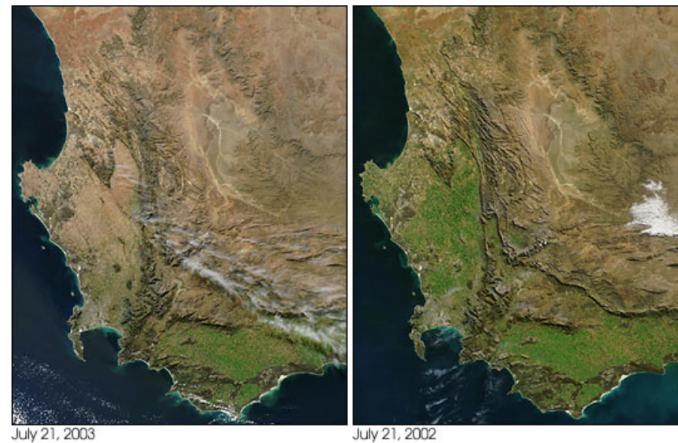
<sup>5</sup>These pictures put northern Europe's heatwave in perspective, euronews, <https://www.euronews.com/2018/07/25/these-pictures-put-northern-europe-s-heatwave-in-perspective>

<sup>6</sup>Drought in Western Cape, South Africa, earth observatory, <https://earthobservatory.nasa.gov/images/11912/drought-in-western-cape-south-africa>





(A) 2018 Eueropean heat wave after effects.



(B) 2003 South African drought implications.

FIGURE 2.1: Examples of geospatial data before and after adverse events.

Haskell Domain-Specific Language [43], Data Flow Graphs [54], Map-Reduce [37, 45] and self-adjusting computational processors [1]. These are useful in handling big data, but specific procedures still have to be developed to, for example, select data from a big data set for a certain model application. One of the suggested ways of dealing with this by Songnian li et al [34] is using metadata which is mentioned to be useful in cases of classification procedures.

Metadata is defined as data that provides information about other data. Metadata provides content, type, quality, creation and also spatial information about another data set. Metadata mostly occurs in one of the following of formats, namely a text file, an Extensible Markup Language (XML) file or a database record. Types

of metadata include descriptive, structural, administrative, statistic, legal and reference metadata. Metadata makes data easier to document, makes data discovery easier and reduces data duplication. Metadata typically consists of the name of the data file, the source agency, the creation date, the type of data, the author of the data, relevant contact information and licensing along with data dictionary and/or restrictions. Geospatial metadata additionally contains a spatial component such as the extent of the surface of the earth that the data covers, for example, the coordinate system and/or spatial extent. [47]

Figure 2.2 is an example of metadata providing information about an image data of a cat. It contains the filename, the author, date captured and location.

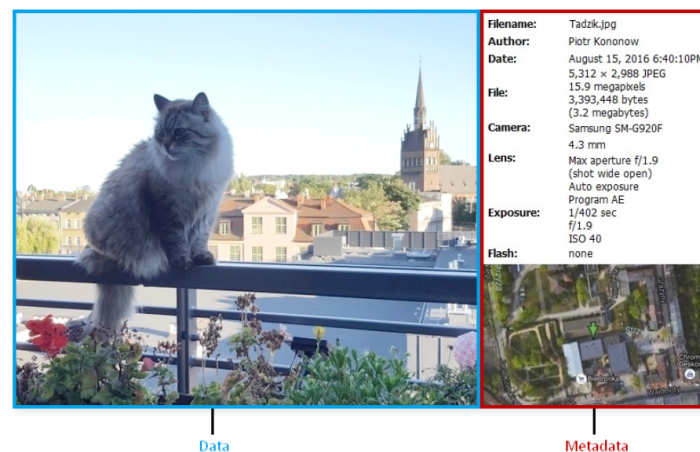


FIGURE 2.2: An example of the metadata of a picture of a cat.

## 2.2 Database construction

### 2.2.1 Data summary

The dataset to be used in this research is the Sentinel-2 time series data for the Western Cape province in South Africa. This dataset is freely accessible on the Radiant MLHub website generated by Radiant Earth Foundation and the Western Cape department of Agriculture in 2021 [2]. Radiant MLHub is a cloud-based open library that has earth observation data including land cover, wildfire, floods, tropical storms, building footprints and crop datasets. The dataset to be used is a crop dataset that has 12 bands in the near infrared, short wave infrared and visible part

of the electromagnetic spectrum and a 13th image type (CLM), which gives the cloud coverage on a tile image. The time series is provided every five days from the 1st of April till the 27th of November (48 dates) . Figure 2.3 shows the 12 bands of one area of land with tile ID 1114 taken by the Sentinel-2 satellite on the 28th of October 2017.

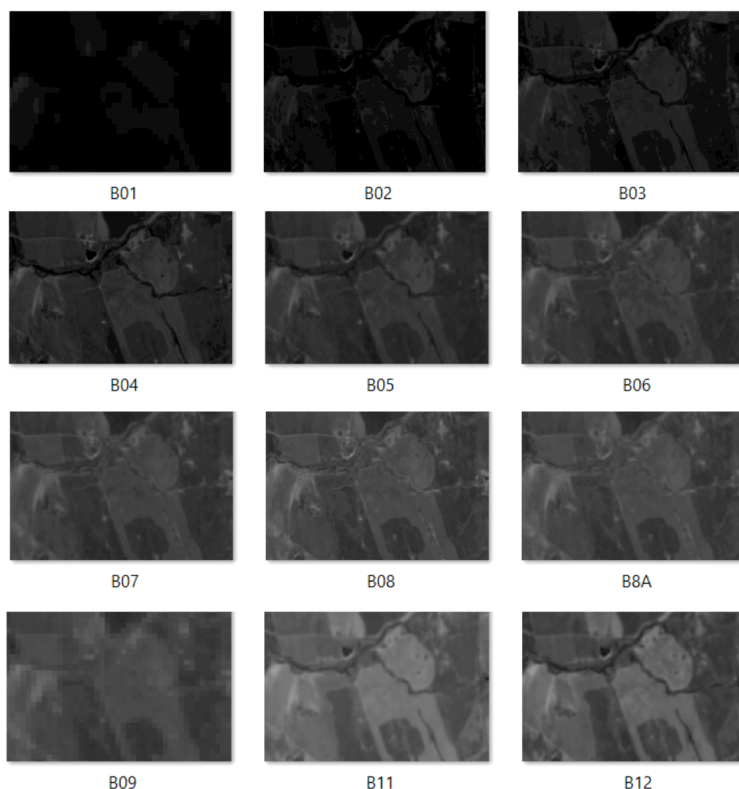


FIGURE 2.3: A single image tile from the data shown over 12 different bands

B01 is the coastal aerosol band with a resolution of 60m. B02, B03 and B04 are the blue, green and red colour bands all with the same resolution of 10m. B05, B06, B07 and B8A are the vegetation red edge bands with the same resolution of 20m. B08 is the near infrared band with resolution of 10m. B09 is the water vapour with resolution of 60m whereas B11 and B12 are the short wave infrared bands with resolution of 20m.

Figure 2.4 shows how the 13th image type presents cloud coverage, with only two colours. Black represents the absence of clouds whereas white represents the presence of clouds. Four different dates are selected to show the difference in cloud

coverage on the same area of land over different dates. The first image has a cloud coverage of 52.25%, the second one has 16.89%, the third has 100% and lastly the fourth one also used in Figure 2.3 has no cloud coverage hence 0%.

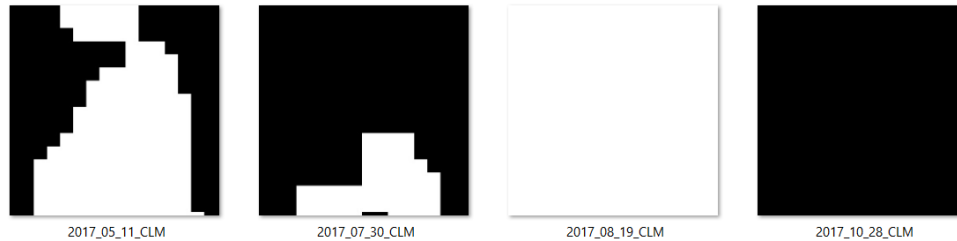


FIGURE 2.4: Cloud coverage of the same area over four different dates.

Each image in the dataset is an area of land made up of crop fields. Each field contains only one crop type. The dataset consists of 9 crop types, namely fallow, canola, wheat, wine grapes, weeds, small grain grazing, lucerne/medics, planted pastures (perennial) and rooibos. Figure 2.5 shows the different fields and labels of the area covered by Figures 2.3 and 2.4. For example, this area is made up of 25 fields that contain 6 different crop types, namely lucerne/medics, planted pastures, fallows, small grain grazing, wheat and canola.

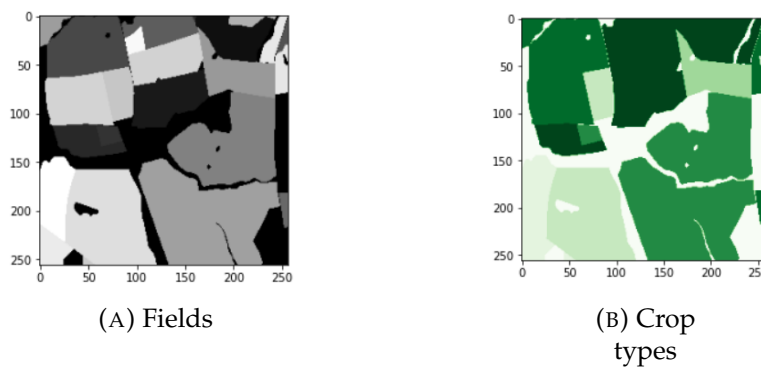


FIGURE 2.5: An illustration of the fields and crop types of the area with tile ID 1114.

Each area of land (2650 locations) was captured every five days ( $\times 48$ ) through 12 bands of the electromagnetic spectrum and the 13th image that shows cloud coverage ( $\times 13$ ), meaning the whole data is made up of 1 653 000 images. The area of interest is 23 850km<sup>2</sup> of land of which 9 063km<sup>2</sup> of it, roughly 38%, has been

labelled. This constitutes the portions that will be considered in assessing the accuracy of sampling. The area coverages of the crop types in each image and field have been calculated and this will help calculate the proportion of the crop types in the population. Summing all the area coverages in each image gives the overall area coverage of each crop type. Figure 2.6 shows the proportions of the crop types using their area coverage as well as number of fields, such that the one with the highest proportion is the one with the highest crop coverage. As can be seen from Figure 2.6[A], wheat contains the highest proportion with 23.08% followed by small grain grazing with 14.146%, with the least being canola with a percentage of 3.405%. Figure 2.6[B] shows the quantity of the fields, size of field is not considered. In this case, the crop type with the highest proportion is now wine grapes followed by wheat with the least still being canola.

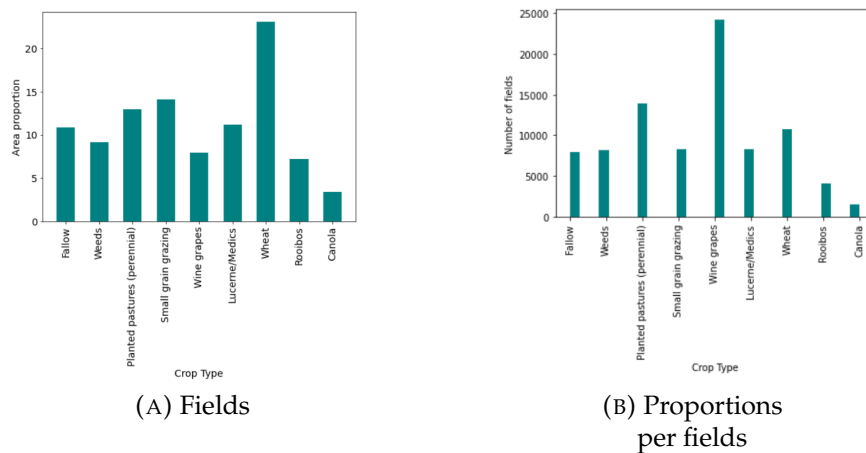
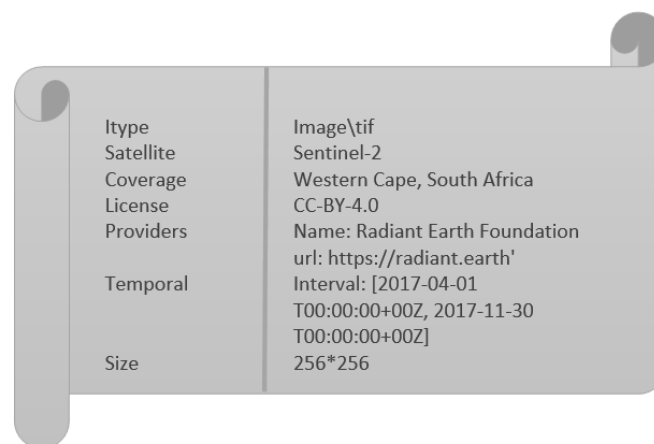


FIGURE 2.6: Proportions of the crop types using area coverage as well as number of fields.

## 2.2.2 Metadata construction

The dataset consists of 1 653 000 images of data which is approximately 45.15GB. One way of avoiding loading this big dataset is using metadata to select only the relevant images of interest to read into memory. Note that some of the metadata was already provided whereas some had to be obtained from the images themselves. The structure of the metadata consists of three categories, namely general

information, information associated with tile ID and information per image. General information includes properties that all images share regardless of location or date captured, namely the satellite used to capture them, the type of image, licence of data, the providers of data and size of images since they are all the same size. These are given in the images STAC (SpatioTemporal Asset Catalogs) files. Figure 2.7 is an illustration of what the general information is for each image regardless of location and date.



Itype	Image\tif
Satellite	Sentinel-2
Coverage	Western Cape, South Africa
License	CC-BY-4.0
Providers	Name: Radiant Earth Foundation url: https://radiant.earth'
Temporal	Interval: [2017-04-01 T00:00:00+00Z, 2017-11-30 T00:00:00+00Z]
Size	256*256

FIGURE 2.7: General information that applies on all images.

Information associated with tile ID is information that has been used to differentiate between the different areas of land/locations such as tile ID, the spatial extent of the area captured, the number of fields along with the crop types they contain. Figure 2.8 shows information associated with tile ID. An image of another area of land thus with a different tile ID will not have the same information as the one in Figure 2.8. The spatial extent also referred to as the bounding box will be different, as will the number of fields as the different areas of land have different fields and crop proportions will also differ.

Information associated with each image is information that is unique for each image, such as the date, time and cloud coverage as it depends on the date. For example, the images with tile ID 1114 will have different cloud coverages as shown in Figure 2.4 due to the different dates on which the area of land was captured. With the three categories brought together, metadata in the form of a database can be created. The database is a Pandas dataframe where the column names are the entries that appear on the left hand side of the Figure 2.7 and 2.8, and the rows

Tile ID	1114
Bounding box	[18.514546656, 33.6848043291, 18.542797031, -33.6611841904]
Number of fields	25
Number of crop types	6
Crop type proportion	['No Data: 17.278%', 'Lucerne/Medics: 4.997%', 'Planted pastures: 12.914%', 'Fallow: 4.218%', 'Small grain grazing: 27.017%', 'Wheat: 13.843%', 'Canola: 19.734%']

FIGURE 2.8: Location wise information that applies on all images of the same area of land.

are indexed by the tile ID and the date on which the images were captured. From the database itself, one can obtain the structure of the data, the description of the data as well as the administration involved in publishing the data. The database is useful because performing procedures such as sampling would not require loading and reading all the images into memory.

### 2.2.3 Summary

The data is made up of 1 653 000 images which requires a lot of memory which not many computers can process all at once. Metadata is obtained and constructed from the images to avoid having to read in all the images. The metadata contains general information such as image type, license and providers which is the same for all images. It also contains information specific to an area of land such as the spatial extent, the date captured, the cloud coverage on the different dates, the number of fields, the crop type as well as their proportions. This is useful when sampling as one can sample from the constructed metadata without having to read in all the images.

## Chapter 3

# Sampling

### 3.1 Sampling theory

The concept of sampling dates back to the 1600s when English merchant John Graunt analysed data about the population of the city of London using partial information. He is mentioned as one of statisticians of the century in the book titled "Statisticians of the century" [19]. This method of estimating population characteristics from partial information later turned into an extensive coverage of theory, method and application covered in a landmark book titled "Sampling techniques" in 1946 and a newer version of it in 1977 [12]. Sampling can be defined as a process in which a certain number of observations (a sample) is selected from a population. The main purpose of sampling is to extract information about the whole population by examining only a selected few observations. The sample should be representative of the population. A sample is representative if there are certain characteristics of interest from the population that can be estimated using the sample with known accuracy [12]. Requirements of a good sample include little to no selection bias. Examples of selection bias include pre-screening leading to a judgement sample, selecting from a certain area because it is easier to collect (also referred to as sample of convenience) as well as self-selection. Other requirements for a good sample are little to no measurement error, which is the difference between a recorded response and the true value, and little to no non-sampling error. Unlike selection bias and measurement error, non-sampling error occurs regardless of the samples used/selected. The method of sampling or biased questions in a survey or questionnaire could cause non-sampling error.

Once a good sample is selected, sampling has great advantages in data collection, such as reduced cost and a greater speed. It is a practical method when a



population is infinite or when there are limited available resources to collect information. Examples of sampling are in the healthcare sector, when a doctor draws a few drops of blood (sample) for examination and in retail when for example, a few fruits are collected to assess the quality of all fruits. Another example would be selecting sites or fields of land that grow different types of plants as will be discussed in this dissertation. This is a certain type of sampling called spatial sampling. The purpose of spatial sampling is to draw characteristics of interest to make inferences regarding a population that has observations with a location parameter, such as fields, rivers, buildings and roads. All these are referred to as spatial populations. Spatial sampling, unlike traditional sampling, takes into account partial correlation, which disobeys the assumption of independence that is used in traditional sampling. Partial correlation states that two data points close enough to each other tend to share the same features especially when the population of interest is continuous.

There are three major distinctions of populations in spatial sampling, namely zero-dimensional, one-dimensional and two-dimensional. Zero-dimensional populations are discrete and finite, such as trees in a forest or buildings in a city. One-dimensional populations are linear and continuous by nature but are often sampled as discrete. These include roads and rivers, i.e. linear networks, and two-dimensional features are continuous and often aerial such as air or soil. It is important to keep in mind that sampling, in practice, reduces a continuous space into a discrete space. The two-dimensional population in spatial sampling defies the assumption of independence in traditional sampling. To take this into account, geostatistics was introduced in 1963 by Matheron et al [40] and revised in 1971 [39]. The approach was improved to use spatial autocorrelation to reduce the estimator variance error that arises from sample design and selection bias [17, 11, 51]. Studies of the effects of spatial structure on the error variance, date back to 1959 by Milne et al [42] and 1960 by Matern et al [38].

When drawing samples from geographical data, certain questions need to be answered for one to decide on a sampling design. The first question is the most important because before one uses spatial sampling, one needs to determine if the characteristic of interest constitutes as a spatial property or not [18].

1. What is being estimated?

2. What sample size will achieve the desired level of accuracy and precision?
3. What locations should be included in the sample?
4. What estimator should be used?
5. What measures should the sampling seek to minimise?

There are two approaches to sampling, namely design-based sampling and model-based sampling. These hold advantages over each other depending on the following factors: reason behind sampling, the quality of the model, one's desire of unbiased estimates, sample size and the presence of autocorrelation between observations. Hence depending on one's end goal and the behaviour of the data, one of the two can fulfil one's needs. The fundamental difference between the two approaches is that design-based sampling units are selected by probability sampling and the spatial characteristics are estimated. In model-based sampling, there are no requirements for selecting sampling units/locations and the characteristics of interest are predicted and not estimated. If the values at given locations are fixed but the sample locations are random, then design-based sampling is appropriate and when the values at given location are random with fixed sample locations then model-based sampling is appropriate.

This mini-dissertation only focuses on design-based sampling. Techniques include random sampling, stratified sampling and clustering. Before getting into these sampling techniques, an understanding of the framework for probability sampling is required. Assuming we have  $N$  units that make a finite population, the population can be denoted by the index set

$$U = \{1, 2, 3, \dots, N\}. \quad (3.1)$$

Any sample that can be selected from the population  $U$  is denoted by  $S$ . Each sample has a known probability of  $P(S)$  of being selected and the probabilities of all sample units being selected should sum to 1. The probability of each unit being in a selected sample is denoted by

$$\pi_i = P(\text{unit } i \text{ in a sample}), \quad (3.2)$$

which can be easily calculated by adding all the probabilities of the samples containing that unit. Let  $y_i$  be the characteristic value associated with the  $i^{th}$  unit in the population. Denote the population total by  $t = \sum_{i=1}^N y_i$ . An estimator that might be used to estimate  $t$  is  $\hat{t}_S = N\bar{y}_S$  where  $\bar{y}_S$  is the average of the  $y'_i$ 's in the sample. The sampling distribution of  $\hat{t}$  is said to have an expected value of  $E[\hat{t}]$  and variance of  $V(\hat{t})$ , given by the following formulae:

$$E[\hat{t}] = \sum_S \hat{t}_S P(S) \quad (3.3)$$

$$V(\hat{t}) = E[(\hat{t} - E[\hat{t}])^2] \quad (3.4)$$

$$= \sum_S P(S) [\hat{t}_S - E(\hat{t})]^2. \quad (3.5)$$

The aim is for the estimator to be unbiased, meaning the estimation bias of the estimator  $\hat{t}$  should be 0, where the bias is calculated as follows:

$$Bias[\hat{t}] = E[\hat{t}] - t. \quad (3.6)$$

In practice, an unbiased estimator is not always achieved. Biased estimators are used and, rather than using the variance to measure the accuracy of an estimator, the mean squared error is used to check how close estimates are to actual values.

$$MSE[\hat{t}] = E[(\hat{t} - t)^2] \quad (3.7)$$

$$= V(\hat{t}) + [Bias(\hat{t})]^2. \quad (3.8)$$

Figure 3.1 gives a clear depiction of the three concepts that are mostly used in the assessment namely, bias, precision and accuracy. Accuracy is either the count or area of correctly identified objects, precision is the ability of say an algorithm to identify only relevant points. Precise means the measurements are close to each other, accurate means the measurements are close to the target and biased means the measurements are on one side. Figure 3.1 shows four archers who are different combinations of bias, precision and accuracy. Ideally one aims to be unbiased, precise and accurate.

We introduce the specifics of the different types of sampling methods like simple random sampling, stratified sampling and cluster sampling.

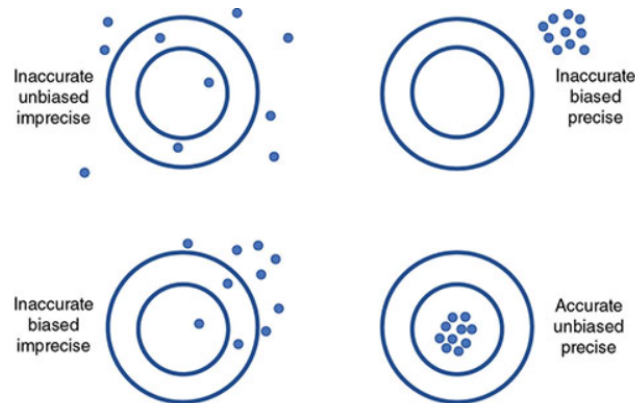


FIGURE 3.1: Clear depiction of the difference between bias, accuracy and precision [44].

### 3.1.1 Random sampling

Random sampling, also referred to as simple random sampling, is the simplest technique in probability sampling. It takes a sample of size  $n$  such that every possible subset of  $n$  units has the same chance of being selected. It is the foundation of most of the more complex sampling techniques. There are two ways of obtaining a random sample, with replacement and without. In simple random sampling with replacement (SRSWR), all the units have the same chance of being picked, as one will select the first unit with probability of  $\frac{1}{N}$ , and it will be replaced and the second unit will be selected with probability of  $\frac{1}{N}$ , and so on. This procedure may include duplicates which is not always ideal, as sampling something twice gives no additional information and the main aim of sampling is to try to extract as much information as one can. An alternative approach is simple random sampling without replacement (SRS) where every possible subset of the population containing  $n$  distinct units has the same chance of being selected. There are  $\binom{N}{n}$  possible samples, so that the probability of selecting any sample is

$$P(S) = \frac{1}{\binom{N}{n}} = \frac{n!(N-n)!}{N!}. \quad (3.9)$$

Let  $w_i$  be a sampling weight of unit  $i$  which are the number of population units that are represented by unit  $i$ . In simple random sampling,  $w_i = \frac{1}{\pi_i} = \frac{N}{n}$  means

$$\sum_{i \in S} w_i = \sum_{i \in S} \frac{N}{n} = N. \quad (3.10)$$

These are referred to as self-weighting sample since all the sampling weights are equal. To estimate the mean of the population  $\bar{y}_i$  using SRS, a sample mean  $\bar{y}_S = \frac{1}{n} \sum_{i \in S} y_i$  with variance  $V(\bar{y}) = \frac{S^2}{n} (1 - \frac{n}{N})$  is used.

The advantages of simple random sampling include the allowance for statistical inferences of a population and that it mostly gives a sample that highly represents the population. The disadvantages are that it is possible only when the whole population is complete and that an adequate proportion of the sample should participate which can be time-consuming. [12]

### 3.1.2 Stratified sampling

In stratified sampling, a population of size  $N$  is divided into  $H$  homogeneous sub-populations also referred to as strata. The strata denoted by  $N_1, N_2, \dots, N_{H-1}, N_H$  do not overlap, such that  $N_1 + N_2 + \dots + N_{H-1} + N_H = N$ . Independent sampling units are drawn from each stratum so that overall population estimates can be obtained from pooling the information. Stratification gives an assurance of not obtaining a really bad sample. It is ideal when there is some desired level of precision that is dependent on subgroups. It is easier to administer and in many cases results in a lower cost than SRS. Lastly, it generally gives lower variance with high precision when the characteristics of interest are population means and totals.

The process of sampling starts by identifying what your desired subgroups should be, then dividing the population into those  $H$  strata. A sample size  $n$  is then chosen and after taking into consideration the proportions of the variables present in the population, the size of each sample from each stratum  $n_h$  is determined such that  $n_1 + n_2 + \dots + n_h = n$ . The sampling units are then drawn using simple random sampling (SRS) as described in Section 3.1.1 or another method called systematic sampling where every  $k^{th}$  unit is selected in each sample.

The notation used for stratified sampling estimates differs from that used for SRS and is given in Table 3.1. The sampling weights for the units in stratified sam-

Notation	Description	Estimator
$\bar{y}_{hU}$	Population mean in stratum h	$\bar{y}_h = \sum_{j \in S_h} \frac{y_{hj}}{n_h}$
$t_h$	Population total in stratum h	$\hat{t}_h = N_h \bar{y}_h$
<b>t</b>	Population total	$\hat{t}_{str} = \sum_{h=1}^H \hat{t}_h$
$\bar{y}_U$	Population mean	$\bar{y}_{str} = \frac{\hat{t}_{str}}{N}$

TABLE 3.1: Population statistics, description and their estimators in stratified sampling.

pling are not equal as the  $\pi_{hj}$  are not equal, but the weights assist estimate the population mean, namely,

$$\bar{y}_{str} = \frac{\sum_{h=1}^H \sum_{j \in S_h} w_{hj} y_{hj}}{\sum_{h=1}^H \sum_{j \in S_h} w_{hj}} \quad (3.11)$$

Stratified sampling is self-weighting if the sampling proportions are the same for each stratum i.e.  $\frac{n_h}{N_h}$ . This is called the method of proportional allocation, and it falls amongst three methods of allocating observations to strata. The other two are optimal allocation and Neyman allocation. Proportional allocation is used when one wants to ensure that the sample reflects the population with respect to the stratification variable. The number of sampled units in each stratum should be proportional to the size of the stratum in the whole population meaning the inclusion probability is  $\pi_{hj} = \frac{n_h}{N_h} = \frac{n}{N}$ . Proportional allocation is advised if the variances in each strata are more or less equal. If the variances is significantly different, optimal allocation gives smaller costs. In practice, optimal allocation is the most stable as the sample sizes vary significantly and high samples tend to have more variability as opposed to small ones. The objective of optimal allocation is to get the most information for the lowest cost. Let  $C$  represent total cost, then

$$C = c_0 + \sum_{h=1}^H c_h n_h, \quad (3.12)$$

where  $c_0$  is overhead costs and is independent of the strata. The aim is to minimize the cost for a fixed variance, so the optimal sample size in stratum  $h$  should be

$$n_h = \left( \frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{l=1}^H \frac{N_l S_l}{\sqrt{c_l}}} \right) \times n. \quad (3.13)$$

Neyman allocation is a special case of optimal allocation, It is when the variances vary greatly but the costs in the strata ( $c_h$ ) are approximately equal.[7]

Advantages of stratified sampling include a better representation of measurements, reduced variance when dividing the population into subgroups, increased precision and effectiveness for populations with extreme observations. Disadvantages include that a complete list of the population is required and that complexity is added to the sampling procedure resulting from the assumption that each unit must belong to only one stratum. [36]

### 3.1.3 Cluster sampling

The previously discussed sampling techniques assume that units are well defined which is not always the case. Cluster sampling is an alternative when units are not well defined. Cluster sampling divides the population into mutually homogeneous groups but internally heterogeneous groups which are evident [7]. This may appear similar to stratification but unlike in stratification, not all clusters are sampled from. Cluster sampling can be performed as either one-stage, two-stage or multi-stage sampling. One-stage sampling occurs when every element in a chosen cluster is selected. In two-stage clustering, after the clusters are selected, the elements in the clusters are also subsampled, so that not all of the elements in a chosen cluster are included in the sample. The clusters are referred to as primary sampling units (*psu*) while the elements within the clusters are referred to as secondary sampling units (*ssu*). The notation for cluster sampling is different from the other sampling techniques. For instance, the universe  $U$  is a population of  $N$  *psus*, while  $S$  is the sample of the chosen *psus*.  $S_i$  is the sample of the chosen *ssus* from the  $i^{th}$  *psu*. The value of a unit is given by  $y_{ij}$  which is the measurement for the  $j^{th}$  element in the  $i^{th}$  *psu*. In the stratification sampling notation, there were two different population

quantities. The same applies here as there are *psu* level population quantities and *ssu* level population quantities. Table 3.2 gives the *psu* level quantities.

Notation	Description
$N$	Number of <i>psus</i> in a population
$M_i$	number of <i>ssus</i> in <i>psu i</i>
$M_0 = \sum_{i=1}^N M_i$	Total number of <i>ssus</i> in a population
$t_i = \sum_{j=1}^{M_i} y_{ij}$	Total in <i>psu i</i>
$t = \sum_{i=1}^N t_i = \sum_{i=1}^N \sum_{j=1}^{M_i} y_{ij}$	Population total

TABLE 3.2: *psu* level population quantities used in cluster sampling.

Table 3.3 gives the *ssu* level population quantities.

Notation	Description
$\bar{y}_{iU} = \sum_{j=1}^{M_i} \frac{y_{ij}}{M_i}$	Population mean in <i>psu i</i>
$\bar{y}_U = \sum_{i=1}^N t_i = \sum_{i=1}^N \sum_{j=1}^{M_i} \frac{y_{ij}}{M_0}$	Population mean

TABLE 3.3: *ssu* level population quantities used in cluster sampling.

One-stage cluster sampling is closely related to stratified sampling, since all elements in selected clusters are selected. Figure 3.2 shows the difference between one-stage clustering and stratified sampling techniques. Figure 3.2(A) shows that every stratum is randomly sampled from while in Figure 3.2(B), all observations are selected from the chosen clusters.

In stratified sampling, the variance of  $\bar{y}_U$  depends on the within stratum variance, while in cluster sampling, the variance of  $\bar{y}_U$  depends on the variability between clusters [49]. The *psus* in cluster sampling can be of equal or unequal size, with the latter being more common in practice.

Two-stage cluster sampling involves selecting a random sample of  $n$  *psus* from the population, and then selecting a sample of *ssus* from each selected *psu*. Figure



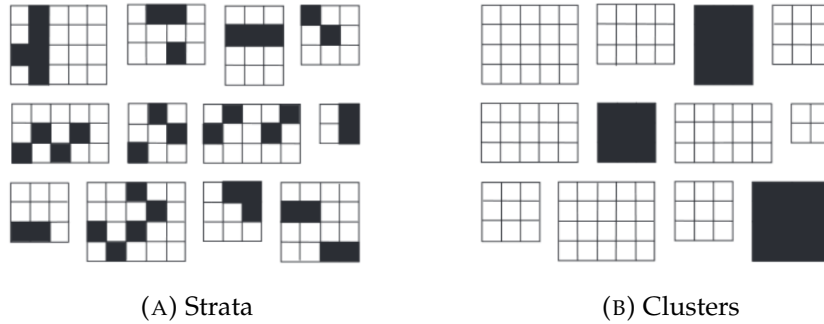


FIGURE 3.2: Illustration of the difference between stratification and one-stage clustering [9]

3.3 gives an illustration of the difference between one-stage and two-stage clustering.

The selection of the *ssus* from the *psus* does not necessarily have to be random. An individual may decide to use systematic sampling, for example, every 3<sup>rd</sup> element in all chosen *psus* is selected.

Advantages of cluster sampling are that it provides more information with less cost and that it is administratively convenient than SRS and stratified sampling. Its shortfalls are that it gives less precise estimates when compared to the other two sampling techniques and that there is a high probability of sampling error because some clusters are not sampled from at all. [14]

## 3.2 A multivariate stratified sampling algorithm

This section develops the proposed sampling algorithm that makes use of multivariate stratification.

Let  $N$  be the number of images in the population and  $M$  be the number of crop types. Let  $n$  be the number of images in a sample and  $N_i$  be the number of images that contain crop type  $i$  in the population. We notate  $A_{pop}^i$  and  $A_{samp}^i$  as the area coverages of crop type  $i$  in the population and sample respectively.  $\mathbf{A}_{pop}$  and  $\mathbf{A}_{samp}$  are vectors of area coverages of the  $M$  crop types in the population and the sample respectively, and  $\mathbf{V}_{pop}$  and  $\mathbf{V}_{samp}$  are vectors containing the proportions of the  $M$  crop types in terms of area coverage in the population and sample respectively. We propose an algorithm to obtain a sample from the population which ensures that the

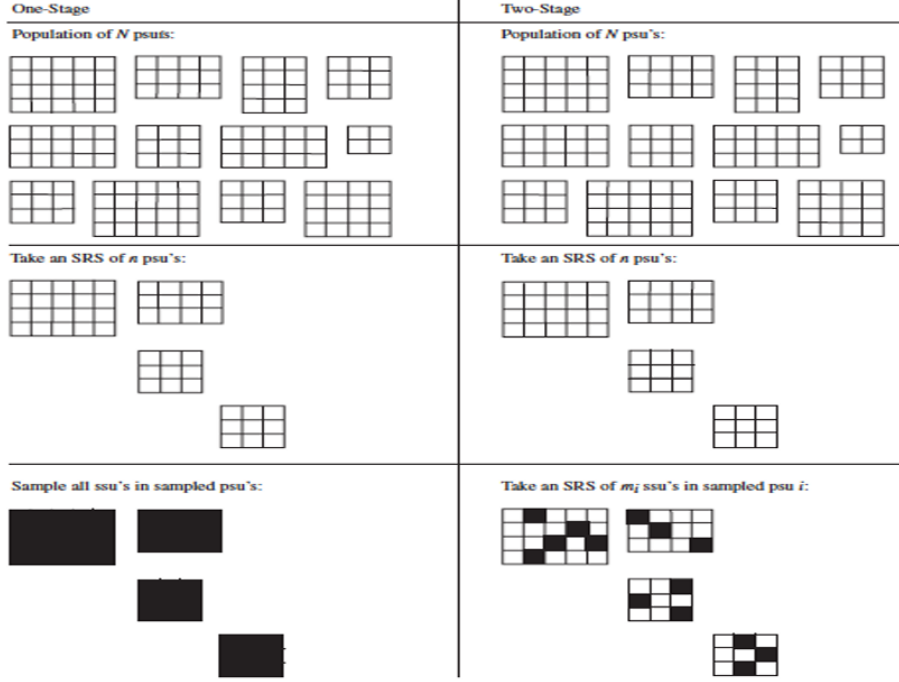


FIGURE 3.3: The difference between one-stage and two-stage cluster sampling [36].

proportion between the population and the sample are similar, while minimizing the number of images sampled. The proportions are calculated in terms of area coverage. The area coverages of the  $M$  crop types in the population ( $\mathbf{A}_{pop}$ ) should be directly proportional to the area coverages of the crop types in the sample ( $\mathbf{A}_{samp}$ ).

$$\mathbf{V}_{pop} = \begin{bmatrix} V_{pop}^1 \\ V_{pop}^2 \\ \vdots \\ V_{pop}^M \end{bmatrix}, \mathbf{V}_{samp} = \begin{bmatrix} V_{samp}^1 \\ V_{samp}^2 \\ \vdots \\ V_{samp}^M \end{bmatrix}, \mathbf{A}_{pop} = \begin{bmatrix} A_{pop}^1 \\ A_{pop}^2 \\ \vdots \\ A_{pop}^M \end{bmatrix} \text{ and } \mathbf{A}_{samp} = \begin{bmatrix} A_{samp}^1 \\ A_{samp}^2 \\ \vdots \\ A_{samp}^M \end{bmatrix}.$$

Ideally the desired area coverages in the sample,  $\mathbf{A}_{samp}$ , should be  $\frac{n}{N} \times \mathbf{A}_{pop}$ . Mathematically, the aim is to show the following equation holds for some small  $\epsilon$ :

$$\|\mathbf{V}_{pop} - \mathbf{V}_{samp}\| \leq \epsilon. \quad (3.14)$$

The algorithm is separated into two main steps where the first samples by considering the most represented crop type in the population. The second uses the partial sample from the first step and focuses on the least represented crop type.

This is done iteratively until all crop types are represented, while satisfying equation (1). The steps of the algorithm are provided in detail below.

1. Calculate  $\mathbf{A}_{pop}$ , the area coverages of the  $M$  crop types in the population. From this, compute  $\mathbf{V}_{pop}$  the proportions of the crop types in the population.
2. Let  $cropA$  be the crop type in  $\mathbf{V}_{pop}$  with the highest proportion, such that  $cropA = \underset{i}{\operatorname{argmax}}(V_{pop}^i)$ .
3. Extract a sub-dataframe  $newA\_df$  from the dataframe containing the meta-data such that  $newA\_df$  only contains images that have  $cropA$  such that  $N_A$  is the length of  $newA\_df$ .
4. Order the images  $I_{(A,1)}, I_{(A,2)}, \dots, I_{(A,N_A)}$  in  $newA\_df$  in descending order according to the area coverage of  $cropA$  in each image. The new order will now be  $I'_{(A,1)}, I'_{(A,2)}, \dots, I'_{(A,N_A)}$ .
5. Introduce parameter  $cropAmax$ , this is a percentage of the area coverage for a particular crop type. This ensures that when other crop types are considered, the desired area coverage  $A_{samp}^i$  of the previously considered crop type is not exceeded.
6. Include images  $I'_{(A,1)}, I'_{(A,2)}, \dots, I'_{(A,n_A)}$  such that the area coverage of  $cropA$  in the  $n_A \leq N_A$  images is  $cropAmax\%$  of the desired  $cropA$  sample area coverage. These  $n_A$  images are included in the sample.
7. From the thus far  $n_A$  sampled images, the area coverages of the other crop types are also captured and stored in  $\mathbf{A}_{samp}$  as the corresponding  $\mathbf{A}_{samp}^i$ .
8. Considering the current  $\mathbf{V}_{samp}$ , let  $cropB$  be the crop type currently least represented by the sample, such that  $cropB = \underset{i}{\operatorname{argmin}}(V_{samp}^i)$ . Extract another sub-dataframe  $newB\_df$  that contains images with  $cropB$  in them but excluding the  $n_A$  already in the current sample. Let  $N_B^*$  denote the number of these images which may be less or equal to  $N_B$  depending on whether or not the  $n_A$  sampled images contain  $cropB$ .
9. Rearrange the images  $I_{(B,1)}, I_{(B,2)}, \dots, I_{(B,N_B^*)}$  in  $new\_df$  in descending order according to the area coverage of  $cropB$  in each image, such that the new order is  $I'_{(B,1)}, I'_{(B,2)}, \dots, I'_{(B,N_B^*)}$ .

10. Introduce another parameter  $cropBmax$  which works similar to  $cropAmax$ , except that it is now imposed on the desired  $cropB$  area coverage in the sample. Denote by  $n_B$  the number of images that make up  $cropBmax\%$  of the remaining desired  $cropB$  area coverage in the sample.
11. Capture the area coverages of all the crop types in the  $n_B$  images and add to the ones from the previously sampled images, in  $\mathbf{A}_{samp}$ . The total  $n_A + n_B$  now becomes the updated sample size with images  $I'_{(A,1)}, I'_{(A,2)}, \dots, I'_{(A,n_A)}, I'_{(B,1)}, I'_{(B,2)}, \dots, I'_{(B,n_B)}$  being the sample.
12. Repeat Step (8)-(9) for the next least represented crop type. Step (10) is modified to (10\*) such that the  $cropAmax$  and  $cropBmax$  parameters are not included any longer i.e. we want to make up the remaining desired area coverage. Iterate step (8), (9) and (10\*)  $M - 2$  times to account for the remaining crop types. Each time an iteration occurs, the previously  $n_i$  selected images are not considered in the next iteration as they have already been added to the sample.
13. After the iterations, the final sample will now be the images  $I'_{(A,1)}, I'_{(A,2)}, \dots, I'_{(A,n_A)}, I'_{(B,1)}, I'_{(B,2)}, \dots, I'_{(B,n_B)}, \dots, I'_{(M,1)}, I'_{(M,2)}, \dots, I'_{(M,n_M)}$ . From the final  $\mathbf{A}_{samp}$ , compute  $\mathbf{V}_{samp}$ , the proportions of the crop types in the sample.

### 3.3 Simulation

This section investigates the effect of the different values of  $cropAmax$  and  $cropBmax$  on the sample area coverages, the number of images sampled and the Euclidean norm. The implementation of the sampling algorithm is done in Python and the notebook containing the code for the algorithm is available on figshare<sup>7</sup>.

We investigate first the role of the parameters  $cropAmax$  and  $cropBmax$  on the values of  $n_A$  and  $n_B$ . In our dataset,  $cropA$  is wheat and the length of  $newA\_df$  is  $N_A=106$ . The values of  $cropAmax$  that were used are [0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1]. To understand the effect of  $cropAmax$  on  $n_A$ , we compare the different values e.g when  $cropAmax=0.4$ , then  $n_A$  is 12 and when  $cropAmax=1$ , then  $n_A$  is 33. Table 3.4

<sup>7</sup>Sampling algorithm, Figshare, Python code, <https://doi.org/10.25403/UPresearchdata.20444061>

gives an example of the percentages of the desired  $A_{samp}$  area coverages that have been already achieved from the  $n_A$  image, i.e. the sample at the end of step 7 of the algorithm.

Crop Type	Achieved area coverage	
	$cropAmax=40\%$	$cropAmax=100\%$
Wheat	4.06%	10.1%
Weeds	0.0%	0.14%
Canola	1.71%	4.15%
Wine grapes	0.12%	0.15%
Fallow	0.11%	0.26%
Rooibos	0.01%	0.01%
Planted Pastured (perennial)	0.057%	0.15%
Lucerne/Medics	0.24%	1.99%
Small grain grazing	0.34%	1.54%

TABLE 3.4: Achieved area coverage percentage using two different  $cropAmax$  parameter.

Using a  $cropAmax$  value of 0.4, the partial sample is made up of  $n_A=12$  images. This is the selected number of images containing  $cropA$  in the first iteration.  $cropB$  which is the least represented crop in the partial sample is weeds. The number of images that contain  $cropB$  in the population is  $N_B=N_B^*=1428$  images because the previously sampled images do not include weeds. Comparing two values of  $cropBmax$ , 0.4 and 1, in combination with the  $cropAmax$  of 0.4, we found  $n_B$  to be 16 and 38 respectively. This is the selected number of images with  $cropB$  added to the partial sample. Table 3.5 shows how different values of  $cropBmax$  increases the area coverages of each respective crop.

Using  $cropAmax=cropBmax=0.4$ , it can be seen from Table 3.5 that the next crop to be considered is wine grapes, followed by planted pastures, with the last iteration ( $M^{th}$ ) focusing on canola. From Table 3.5, the overall sample size relative to the population is 10.277%. Table 3.6 shows how the area coverages increase over the iterations.

Table 3.7 consists of the sample size achieved given different values of  $cropAmax$  and desired sample sizes. Note that the sample sizes are calculated using area coverages and not number of images, and this is before introducing  $cropBmax$  which is the same as taking  $cropBmax=1$ .

Crop Type	Achieved area coverage			
	<i>cropAmax</i> =40%		<i>cropAmax</i> =100%	
	<i>cropBmax</i> =40%	<i>cropBmax</i> =100%	<i>cropBmax</i> =40%	<i>cropBmax</i> =100%
Wheat	4.11%	4.159%	10.1%	10.164%
Weeds	4.286%	10.3789%	0.246%	0.893%
Canola	1.71%	1.714%	4.15%	4.15%
Wine grapes	0.12%	0.174%	0.15%	0.152%
Fallow	0.53%	2.187%	0.597%	1.05%
Rooibos	0.325%	0.524%	4.391%	10.64
Planted Pastured (perennial)	0.135%	0.290%	0.231%	0.277%
Lucerne/Medics	0.24%	0.248%	1.99%	1.99%
Small grain grazing	0.446%	0.599%	1.54%	1.639%

TABLE 3.5: Achieved area coverage percentage using two different *cropBmax* parameter.

Crop Type	Achieved area coverage						
	3rd it- eration (Wine grapes)	4th it- eration (Planted Pas- tures)	5th it- eration (Rooi- bos)	6th it- eration (Lucerne Medics)	7th it- eration (Fallow)	8th it- eration (Small grain)	9th it- eration (Canola)
Wheat	4.11%	4.54%	4.60%	7.73%	7.8%	8.79%	8.83%
Weeds	4.34%	4.6%	5.20%	5.20%	6.22%	6.63%	6.63%
Canola	1.71%	2.68%	2.68%	4.4%	4.42%	4.80%	10.92%
Wine grapes	10.73%	11.43%	11.43%	11.44%	11.44%	11.66%	11.66%
Fallow	0.85%	0.96%	1.75%	1.8%	10.26%	10.55%	10.61%
Rooibos	0.33%	0.33%	10.31%	10.31%	10.1%	10.34%	10.79%
Planted pastures (peren- nial)	0.19%	10.24%	10.37%	10.66%	10.80%	11.65%	11.9%
Lucerne Medics	0.38%	1.20%	1.2%	10.2%	10.25%	10.53%	11.02%
Small grain grazing	0.45%	1.49%	1.59%	2.18%	2.2%	10.32%	10.38%

TABLE 3.6: Achieved area coverage percentages over the remaining iterations.

Table 3.8 contains the number of images sampled given different values of *cropAmax* and the different desired sample size (area-wise). This is similar to Table 3.7 where the parameter imposed on the second considered crop type is not included.

Sample size	<i>cropAmax</i>									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
10%	10.6	10.7	10.7	10.8	10.9	10.9	11.1	11.2	11.3	11.3
20%	21.0	21.2	21.0	21.2	21.3	21.6	21.7	21.8	22.0	22.2
30%	31.9	32.3	32.2	32.3	32.4	32.3	32.3	32.8	32.9	32.9
40%	43.1	43.3	43.0	43.1	43.3	43.5	43.4	43.3	43.4	43.4
50%	54.0	53.4	53.3	54.4	54.6	54.5	53.9	53.8	53.9	53.6
60%	64.4	64.7	63.6	64.9	64.4	64.0	64.0	63.9	63.6	63.8
70%	74.6	74.9	75.8	75.4	74.0	74.2	73.8	73.5	73.4	73.6
80%	83.9	84.0	84.1	83.8	83.2	83.2	82.9	82.8	82.8	83.3
90%	91.9	92.0	92.0	92.1	92.1	91.5	91.5	91.4	92.0	91.9
100%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

TABLE 3.7: Achieved Sample size per desired sample size and *cropAmax*.

Sample size	<i>cropAmax</i>									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
10%	129	131	132	134	137	138	140	143	145	146
20%	269	273	271	275	278	284	286	290	294	297
30%	430	435	436	440	444	447	450	458	462	463
40%	620	617	613	616	621	627	629	630	634	635
50%	813	798	800	810	814	822	820	822	828	839
60%	1026	1033	1006	1009	1020	1019	1022	1045	1040	1037
70%	1248	1253	1238	1240	1258	1259	1263	1251	1249	1247
80%	1484	1486	1489	1496	1490	1513	1508	1501	1496	1471
90%	1799	1801	1803	1800	1816	1826	1824	1814	1785	1798
100%	2646	2646	2646	2646	2646	2646	2646	2646	2646	2650

TABLE 3.8: Number of images per desired sample size and *cropAmax*.

Table 3.9 contains the achieved sample sizes given different desired sample sizes and values of *cropBmax*. The parameter imposed on *cropAmax* is now kept constant, *cropAmax*=1, to illustrate the effect of *cropBmax*.

The Euclidean norm is computed to quantify the difference between the area coverages of the crop types in the sample to those in the population. Table 3.10 gives the Euclidean norm between the different samples for different values of *cropAmax* not considering the effect of *cropBmax* (i.e. *cropBmax*=1).

Figure 3.4 gives the effect of the different values of *cropAmax* and *cropBmax* using the Euclidean norm calculated from the different desired sample sizes and

Sample size	<i>cropBmax</i>									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
10%	10.3	10.4	10.6	10.7	10.8	10.9	11.0	11.1	11.1	11.3
20%	20.3	20.6	20.8	20.9	21.2	21.4	21.6	21.8	22.0	22.2
30%	30.1	30.5	30.9	31.3	31.7	31.9	32.1	32.5	32.7	32.9
40%	39.9	40.3	40.7	41.2	41.5	41.3	42.2	42.6	42.9	43.4
50%	50.7	51.1	51.6	52.0	52.4	52.8	52.9	53.2	53.4	53.6
60%	61.1	61.5	61.9	62.5	62.7	62.8	63.0	63.4	63.6	63.8
70%	70.6	71.1	71.6	72.0	72.2	72.5	72.6	72.9	73.3	73.6
80%	78.5	78.3	78.9	79.6	80.2	81.0	81.5	82.2	82.7	83.3
90%	90.0	90.4	90.2	90.6	90.9	91.1	91.3	91.5	91.8	91.9
100%	99.2	99.2	99.2	99.3	99.4	99.4	99.5	99.6	99.8	100.0

TABLE 3.9: Achieved sample size per desired sample size and *cropBmax*.

Sample size	<i>cropAmax</i>									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
10%	5.18	4.62	4.16	3.83	3.38	3.23	3.61	4.04	4.68	4.99
20%	9.33	8.22	6.39	5.67	4.86	5.28	5.6	6.44	7.44	8.72
30%	13.33	12.01	10.24	9.25	8.72	8.06	8.05	9.42	10.32	11.44
40%	15.99	14.35	12.47	11.81	11.24	11.55	11.41	11.8	12.54	13.88
50%	18.3	17.45	16.45	15.88	15.8	15.24	13.86	14.21	15.17	15.49
60%	20.02	19.32	17.28	17.84	16.58	15.67	15.41	14.08	14.77	16.64
70%	20.44	20.42	21.15	19.46	16.02	16.26	13.71	13.57	14.59	16.52
80%	17.93	17.83	17.25	15.42	14.28	12.36	11.75	12.14	12.99	16.68
90%	10.43	10.42	10.11	10.48	10.04	5.73	6.0	6.11	10.1	9.81
100%	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.02

TABLE 3.10: Euclidean norm between population and sample proportions per desired sample size and *cropAmax*.

the population. The lighter and larger the dots, the higher the Euclidean norm and the darker and smaller the dots, the smaller the Euclidean norm.

Figure 3.5 illustrates how the lowest and highest Euclidean norms change with sample size. The averages of the Euclidean norms in each sample are also plotted.



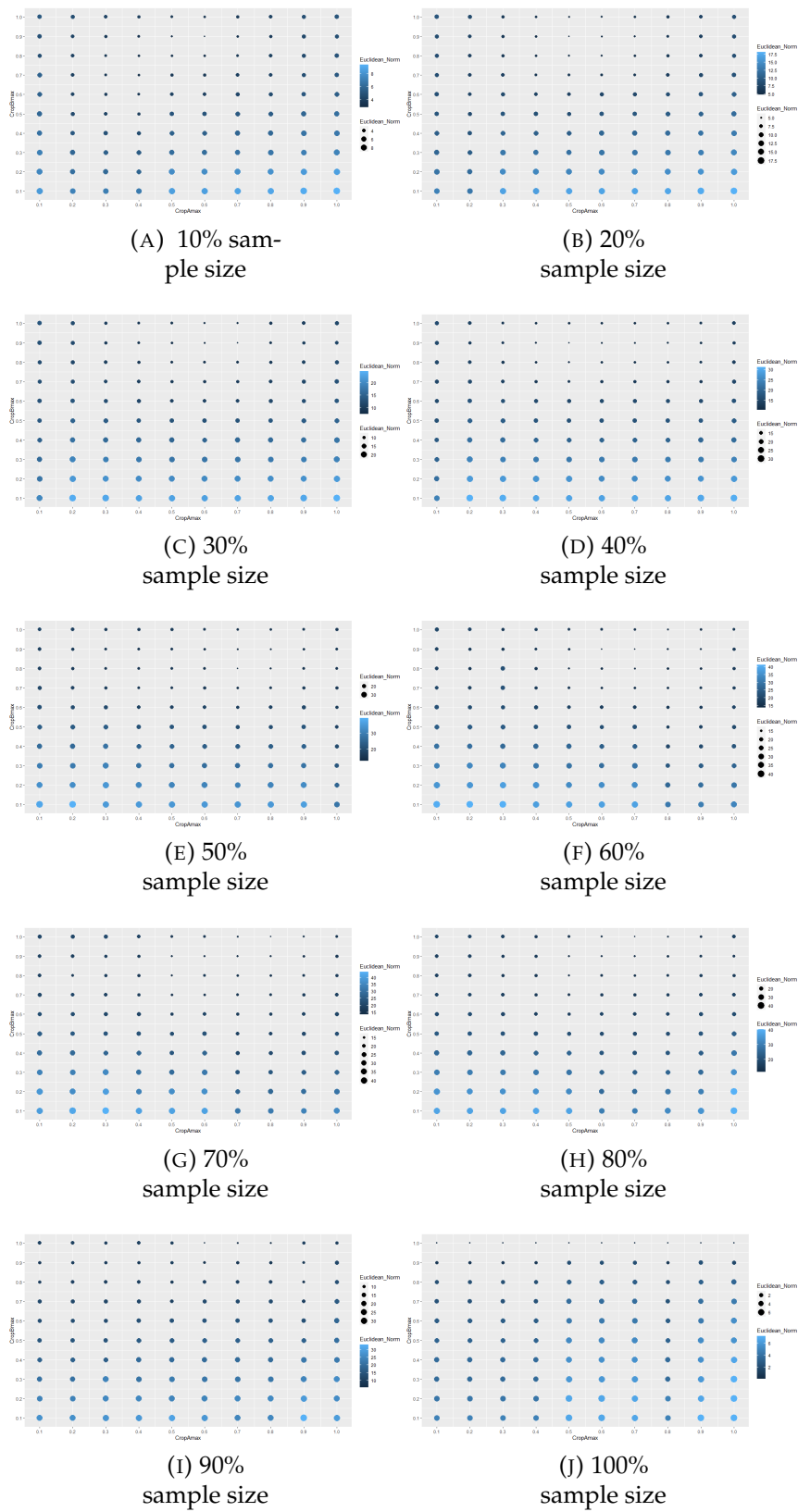


FIGURE 3.4: Euclidean norms between (10%-100%) sample and population.

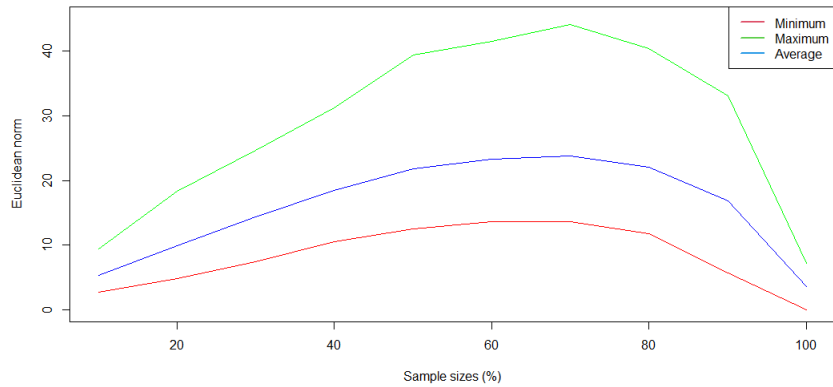


FIGURE 3.5: The bounds of Euclidean norms per sample size as well as averages.

### 3.4 Summary

This chapter covers the theory of three sampling techniques, namely random sampling, stratified sampling as well as cluster sampling. It covers advantages as well as disadvantages of these sampling techniques. A multivariate stratified sampling algorithm is developed. This algorithm minimises the number of images sampled, aims to keep the proportions in the sample and the population similar while also maximising the information obtained in samples. The sampling algorithm is assessed for efficiency and representativeness by computing Euclidean norms as well as looking at the sample sizes in terms of area coverage and number of images.

## Chapter 4

# Classification

### 4.1 Feature engineering and selection

Feature engineering as well as feature selection are both performed on the resulting samples from Chapter 3.

#### 4.1.1 Feature engineering

Feature engineering is the process of calculating additional features from the raw data. Additional features increase the predictive power of a final model and also help capture extra information that is not clear in the original data. Additional features considered in crop type classification are vegetation indices and water indices. The type of vegetation index used herein is the normalised difference vegetation index (NDVI) [62]. The NDVI is an indicator used to assess whether or not vegetation is observed. It takes on values from -1 and 1, where values approaching -1 correspond to water, those close to 0 are barren land and the ones approaching 1 correspond to high vegetation. The NDVI is calculated as follows:

$$\text{NDVI} = \left( \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}} \right) \quad (4.1)$$

where NIR is the near-infrared image band and Red is the red visible image band.

Two normalised difference water indices (NDWI) were also calculated, which are known to be strongly related to water content in plants [16]. Hence they are used in vegetation related applications. The two NDWI used are the NDWI<sub>green</sub>

and the NDWI\_blue with the following formulas:

$$\text{NDWI}_{\text{green}} = \left( \frac{\text{NIR} - \text{Green}}{\text{NIR} + \text{Green}} \right) \quad (4.2)$$

$$\text{NDWI}_{\text{blue}} = \left( \frac{\text{NIR} - \text{Blue}}{\text{NIR} + \text{Blue}} \right) \quad (4.3)$$

The Blue is the blue visible image band and the Green is the green visible image band.

### 4.1.2 Feature selection

Feature selection is the process of selecting the most informative features. It reduces the dimensionality of data, making the data easier to store and analyse. Feature selection, which is synonymous with feature importance, eliminates irrelevant and highly correlated features resulting in a more easily interpretable data. The feature selection techniques used in this research are mutual information regression [26], minimum-redundancy-maximum-relevance (mRMR) [8] and the F-test [13]. The mRMR selects features that reduce their redundancy in the presence of other features while simultaneously increasing their own relevance. The F-test, a correlation-based method, calculates a correlation coefficient which is then converted into a F-statistic. An F-test is performed and the statistically significant features with the highest F-statistics are chosen. The mutual information regression works to identify any sort of dependence between features and eliminates those with high dependency. The mutual information regression has the following formula:

$$I(X;Y) = H(X) - H(X|Y) \quad (4.4)$$

where  $I(X;Y)$  represents mutual information between variables X and Y,  $H(X)$  is the entropy of X and  $H(X|Y)$  is the conditional entropy of X given Y.

### 4.1.3 Implementation

After exploring the proportions of the data, feature selection was conducted, where the mRMR, mutual information regression method and the F-test were used to find the most informative features. According to the mRMR, the selected features are

the NDWI\_green, NDVI, B04, B8A, B06 and B07. Figure 4.1 shows the importances as determined by the mutual information regression method.

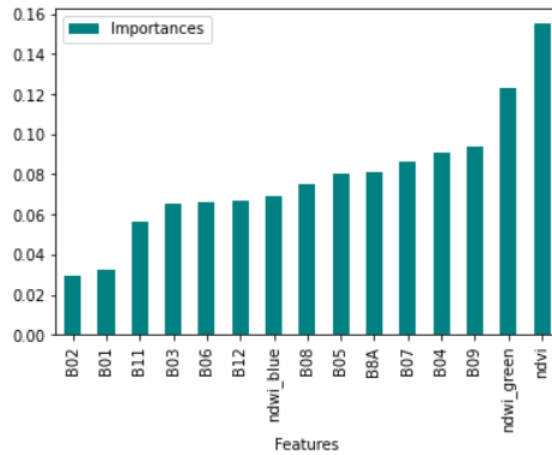


FIGURE 4.1: Image bands importances by the mutual information regression method.

Figure 4.2 gives the importances as determined by the F-test method.

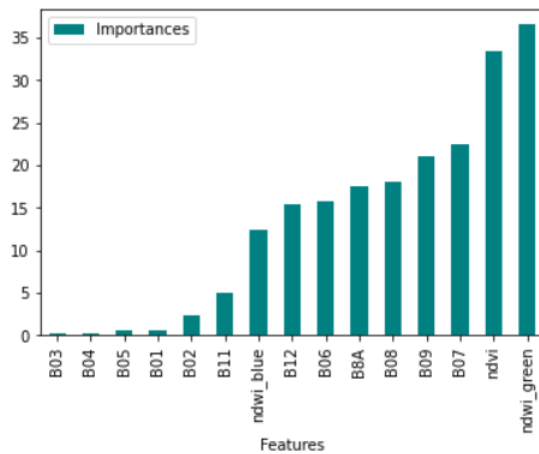


FIGURE 4.2: Image bands importances by the F-test method.

Taking all three feature selection techniques into account, the following bands were determined to be the most significant: NDWI\_green, NDWI\_blue, NDVI, B04, B8A, B06 and B07. These are the bands that will therefore be considered for training.

## 4.2 Land cover classification algorithm

The selected machine learning algorithm is random forest as not only is it easier to implement, but is widely used for crop classifications.

### 4.2.1 Random Forest

The machine learning technique to be considered consists of certain base models combined to produce an optimal predictive model. These are called ensemble methods. Ensemble methods usually produce much more accurate predictions than a single model [46]. Ensemble methods can be used either for classification or regression. In classification, some of the widely used ensemble methods include voting, stacking, bagging and boosting [61]. Voting is the simplest method in terms of implementation. It involves creating multiple classification models from the same training data. Different base models can be created using different datasets but using the same algorithm, or using different algorithms on the same dataset. The decision is taken either by using majority vote (plurality vote) or weighted voting [60]. Stacking, also referred to as stacked generalization, involves combining models using another algorithm. Stacking learns how to best combine predictions from many base models that perform well, so not all base models are considered. Boosting converts the weak base models into strong models [25]. The weights of instances are adjusted using the error measured from previous predictions. It forces base models to learn from hard instances by returning new datasets with only those instances. The most widely known boosting algorithm is the AdaBoost [27]. Last but not least, bootstrap aggregation also referred to as bagging makes use of the same machine learning algorithm with smaller random samples drawn using the bootstrap sampling method (repeatedly drawing smaller samples independently from the population with replacement) [32]. Its biggest advantage is that variance is reduced during the bootstrap sampling.

The machine learning algorithm used herein is random forest, which is an ensemble method that is an extension over bagging [32]. The extension is a result of not just taking a subset of data but also taking a random selection of features to make a decision tree. A decision tree has decision nodes, i.e. two or more branches and a leaf node that represents the classification. The branches of a decision tree depend on a number of factors, it splits the data into branches until it achieves a

certain threshold value. A random forest is roughly defined as a combination of decision trees. Figure 4.3 shows how a random forest randomly chooses features and make observations, builds a forest of decision trees, and then takes the majority or the average of the results<sup>8</sup>.

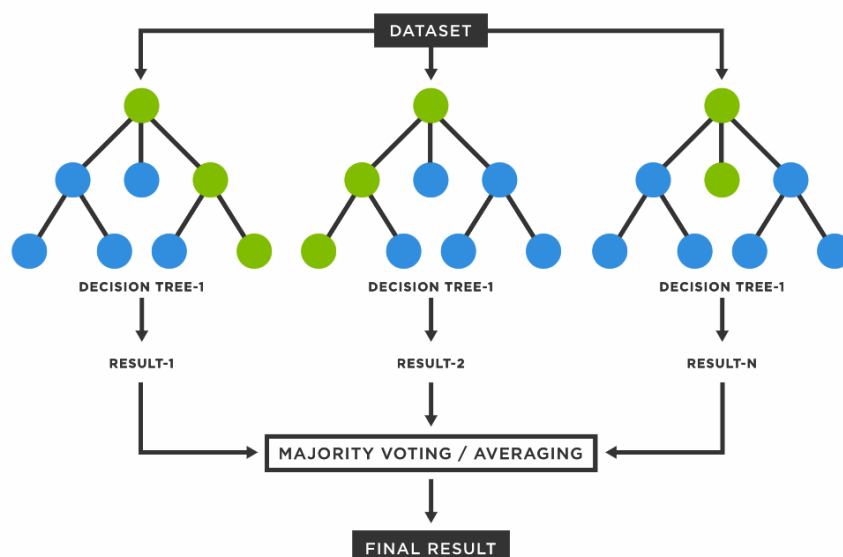


FIGURE 4.3: Random forest illustration.<sup>8</sup>

## 4.2.2 Accuracy measures

The performance of any classification algorithm is quantified using accuracy assessment. Accuracy assessment is the process of comparing a classified image to a reference image to determine the quality and performance of the classifier. Accuracy assessment is used to estimate the accuracy of the extraction of information, which helps to evaluate whether an algorithm meets the requirement of its intended purpose. Accuracy assessment is important as the derivation of maps from remote sensing images inevitably results in errors [57]. The most common way of showing the accuracy of a classified image is by expressing the percentage of the image that has been correctly classified when being compared to the reference image. The expression is in the form of an error matrix, sometimes referred to as a confusion

<sup>8</sup>What is a random forest, TIBCO, <https://www.tibco.com/reference-center/what-is-a-random-forest>

matrix or a contingency table [41, 20]. The columns of the error matrix normally represent the reference while the rows are the classification categories. This form of expression helps a user to assess the performance of individual categories [21]. Table 4.1 is an example of an error matrix where 3 land cover categories are considered. The numbers in Table 4.1 may represent either pixels or objects.

		Reference data			
		River	Land	Trees	Total
Classification data	River	21	6	0	27
	Land	5	31	1	37
	Trees	7	2	22	31
	Total	33	39	23	95

TABLE 4.1: An error matrix in land-cover classification

The overall accuracy of a classification algorithm can be calculated by taking the sum of diagonal entries over the sum of all entries in a confusion matrix. A confusion matrix contains true positives, false positives, true negatives and false negatives. These can be used to calculate quantitative measures such as recall and precision. Recall, also referred to as sensitivity, is the fraction of relevant instances (correctly classified pixels/objects or true positives) that were returned (note that this does not account for the irrelevant instances returned). Precision also called positive predictive value is the fraction of relevant instances among the retrieved instances (true and false positives). The F1-score is another measure which is defined as a harmonic mean between precision and recall. It ranges from 0 to 1. The closer it is to 1, the better the model fitted. The formulae for precision, recall and F1-score are as follows: [48]

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4.5)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4.6)$$

$$\text{F1-score} = \frac{(2 \times \text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (4.7)$$

Figure 4.4 gives good explanation on how to interpret both recall and precision values. High precision and low recall values mean out of the many true positives,



only few of these instances were returned. Low precision and high recall values mean a lot of instances were returned but only a few of those instances are true positives.

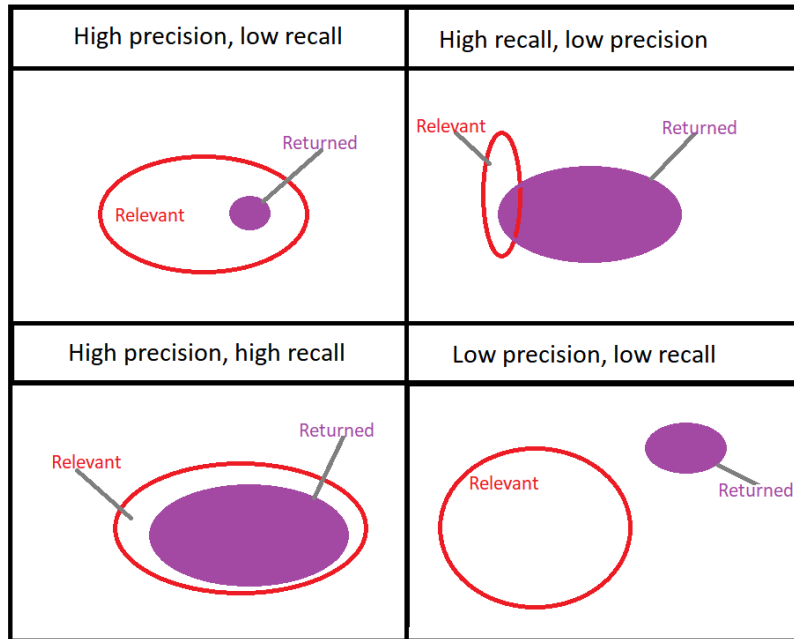


FIGURE 4.4: Interpretations of different precision and recall values.

In statistical modelling, one way of measuring quality of model fit is by calculating the Root Mean Square Deviation (RMSE). This measures the differences between the observed values ( $y$ ) to the predicted values ( $\hat{y}$ ). The formula is as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n}} \quad (4.8)$$

Since different datasets of different sizes will be considered, the RMSE is normalized, for fair comparison of how good a fit a model is. RMSE can be normalized by either the mean ( $\bar{y}$ ), the difference between the maximum and minimum ( $y_{max} - y_{min}$ ), by the standard deviation or by the interquartile range ( $Q_1 - Q_3$ ) of the observations.

### 4.2.3 Implementation and Results

The implementation of the random forest is done using the Python package called `sklearn` with the classifier called *RandomForestClassifier*. The algorithm is trained on the smaller samples (i.e. 10%, 20% and 30%) generated using the proposed multivariate sampling algorithm covered in Chapter 3. A random forest classifier with 100 estimators is defined and trained on the samples with the lowest Euclidean norms. The *cropAmax* values that resulted in the lowest Euclidean norm for the 10%, 20% and 30% sample are 0.6, 0.5 and 0.7 respectively, whereas the *cropBmax* value for all the samples is 0.9.

Random forest classifier is trained on the seven features. The arguments used for the classifier are default parameter values, for example, number of trees in the forest is 100 (`n_estimators`), the Criterion is Gini impurity ("`gini`"), there is no specified maximum depth of a tree, the minimum number of samples required for spitting a node is 2 and the maximum number of features required is the square root of the number of features.

#### 10% sample

The classifier is trained on the 136 images sampled (when a 10% sample was targeted). Note that this sample will be referred to as the 10% proposed sample. Proportions of the crop types using area as well as number of fields are shown in Figure 4.5. Wheat has the highest proportions area-wise but second highest to wine grapes fields-wise with canola being the least represented according to both criteria.

When fitted on the 10% proposed sample, the random forest classifier achieved an overall accuracy of 80.977% with a RMSE of 1.442. To understand how accurate the classifier was per category, precision, recall as well as F1-score are shown in Table 4.2

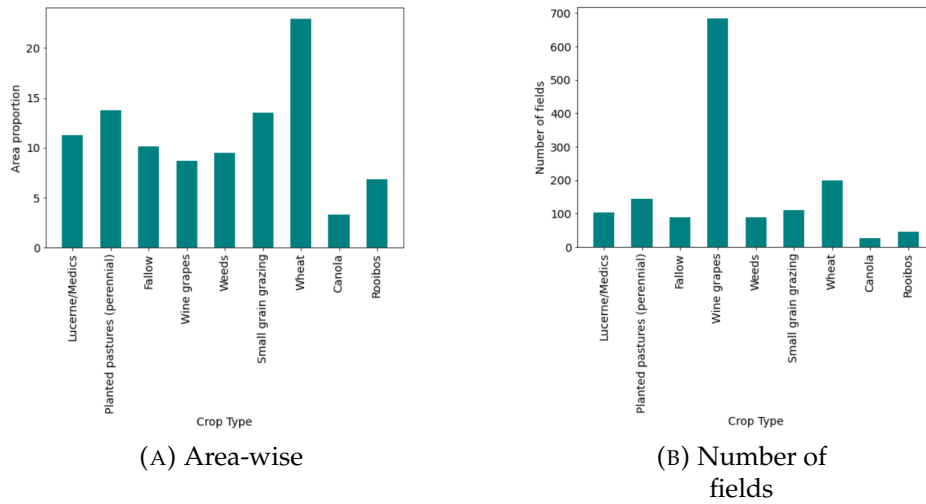


FIGURE 4.5: The proportion of the crop types using area coverage and number of fields.

Crop Type	Accuracy assessments		
	Precision	Recall	F1-score
Lucerne/Medics	48.077%	66.667%	0.559
Planted pastures	75.172%	68.553%	0.717
Fallow	56.18%	64.103 %	0.599
Wine grapes	98.243%	93.194%	0.957
Weeds	59.551%	70.667%	0.646
Small grain grazing	54.545%	63.83%	0.588
Wheat	85.0%	73.913%	0.791
Canola	25.926%	77.778%	0.389
Rooibos	84.783%	73.585%	0.788

TABLE 4.2: Precision, recall and F1-scores per crop type.

Figure 4.6 illustrates how the classifier was able to correctly classify a certain field to its true crop type. This is the quotient of true positives (fields correctly classified to a certain crop type) to all true and false positives (all fields classified as a certain crop type). Wine grapes have the highest correct classifications at 98.243% followed by wheat with 85%. The least crop type with the lowest correct classifications is canola with 25.926%.

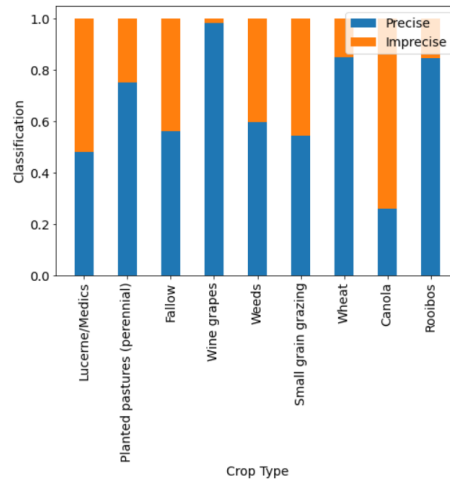


FIGURE 4.6: Accuracy per crop type.

A further comparison is necessary between samples from the proposed sampling algorithm and from a simple random sampling algorithm. Since the 10% sample using the proposed sampling algorithm was achieved at only 136 images, the same number of images are sampled randomly for fair comparison. Looking at the number of fields from the randomly sampled images, Figure 4.7 shows wine grapes to be the crop type with the most fields in the sample followed by planted pastures then wheat with the least being canola. Figure 4.8 shows the representations of the crop types in the two samples compared to the population.

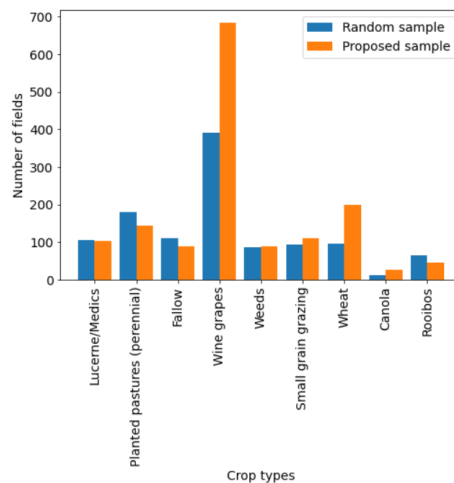


FIGURE 4.7: Crop proportions using number of fields.

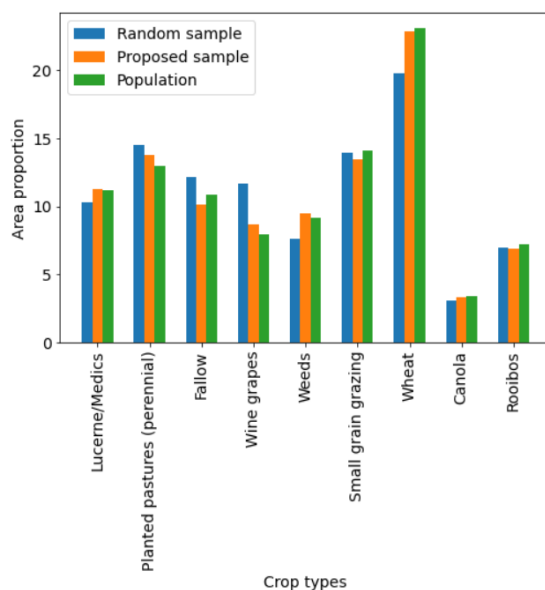


FIGURE 4.8: Area-wise proportions of the crop types in the proposed sample, random sample and the population.

Training the same random forest classifier on the random sample, an overall accuracy of 69.062% is achieved with a RMSE of 1.788. Looking at the accuracy measures per category, the highest precision measure is 96.675% for wine grapes followed by 78.351% for wheat, with the least being canola with a precision value of 25.0%. Figure 4.9 gives an illustration of how the precision and recall values between the random sample and proposed sample differ. Positive values mean the precision and/or recall measures in the proposed sample are higher than those achieved in the random sample and negative values vice versa.

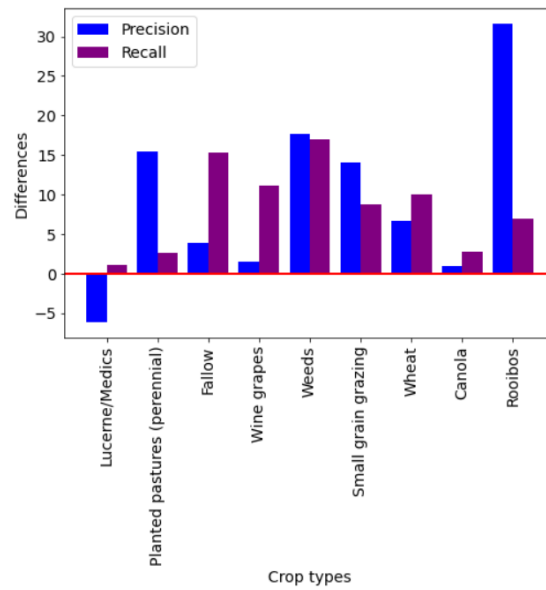


FIGURE 4.9: Difference between achieved precision and recall values in the proposed sample and the random sample.

Figure 4.10 shows a graph of F1-scores for each crop type for both the random sample and 10% proposed sample .

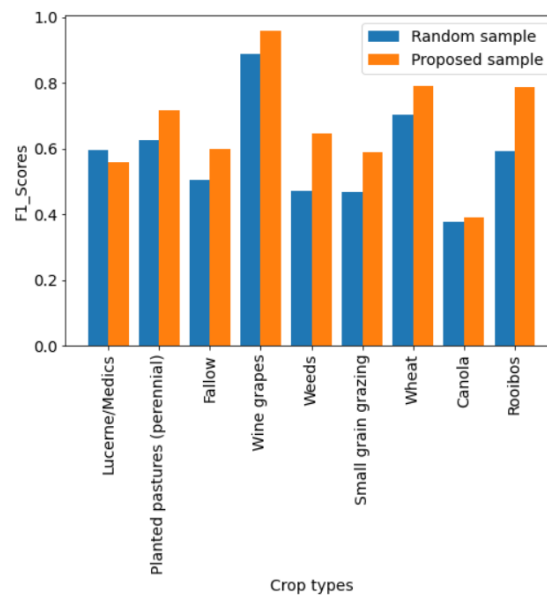


FIGURE 4.10: F1-scores between the random sample and the proposed sample.

## 20% sample

The random forest classifier was also trained on the 20% sample (278 images) achieved from the proposed sampling algorithm. An overall accuracy of 76.479% was achieved with a RMSE of 1.573. The same per category accuracy assessments were measured. Table 4.3 contains the precision, recall and F1-scores per crop type.

Crop Type	Achieved area coverage		
	Precision	Recall	F1-score
Lucerne/Medics	59.848%	68.996%	0.641
Planted pastures	71.92%	66.755%	0.692
Fallow	51.813%	59.88%	0.556
Wine grapes	96.845%	90.574%	0.936
Weeds	52.151%	55.747%	0.539
Small grain grazing	50.385%	59.817%	0.547
Wheat	83.081%	71.522%	0.769
Canola	28.846%	100.0%	0.448
Rooibos	63.366%	78.049%	0.699

TABLE 4.3: Precision, recall and F1-scores per crop type.

A better look at how accurate the classifier was per crop type is by using the same measure as in Figure 4.6 but for the 20% proposed sample. This is illustrated in Figure 4.11.

Similar to the 10% sample from the proposed sampling algorithm, the same number of images as from the 20% proposed sample is randomly selected for further comparison. Hence 278 images are sampled randomly and the difference in the number of fields per crop is plotted. Figure 4.12 shows the number of fields of different crop types in each sample.

Since the proposed stratified sampling algorithm aims to keep the area-wise proportions in the population the same as in the sample, a graph that shows the comparison between the samples as well as the population is plotted. Figure 4.13 gives the proportions of the crop types area-wise between the random sample, the proposed sample as well as the population.

The random forest classifier is then trained on the random sample with 278 images. The classifier achieved an overall accuracy of 66.798% with a RMSE of 1.921.

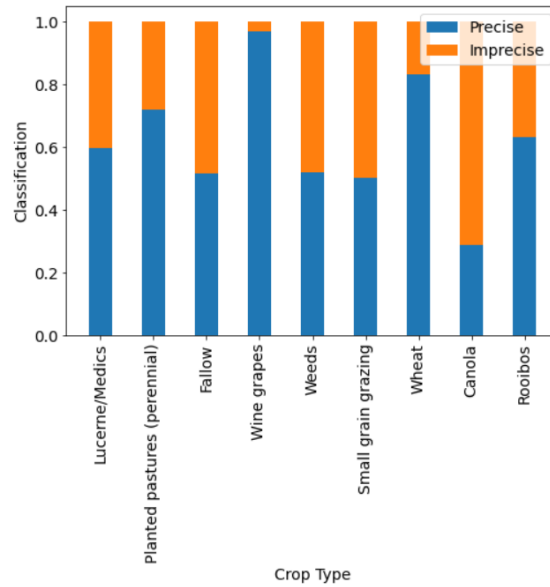


FIGURE 4.11: Accuracy per crop type.

The highest precision measure is 94.397% achieved for wine grapes followed by wheat with 73.828%, with canola being the least with a measure of 20.0%. Figure 4.14 gives an illustration of how the precision and recall values between this random sample and 20% proposed sample differ.



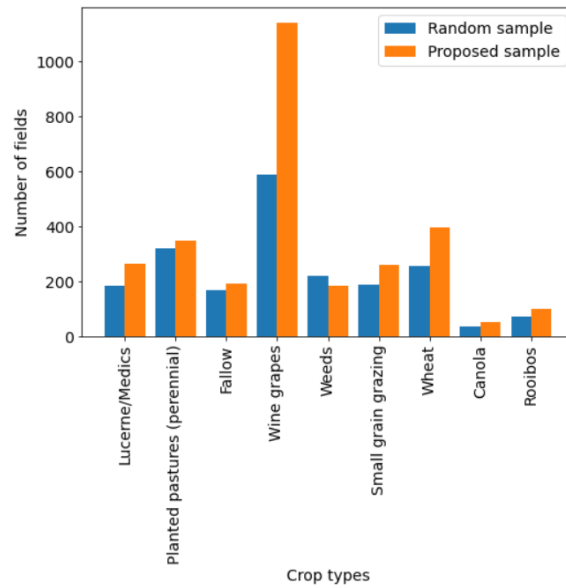


FIGURE 4.12: Number of fields of different crop types in the two samples.

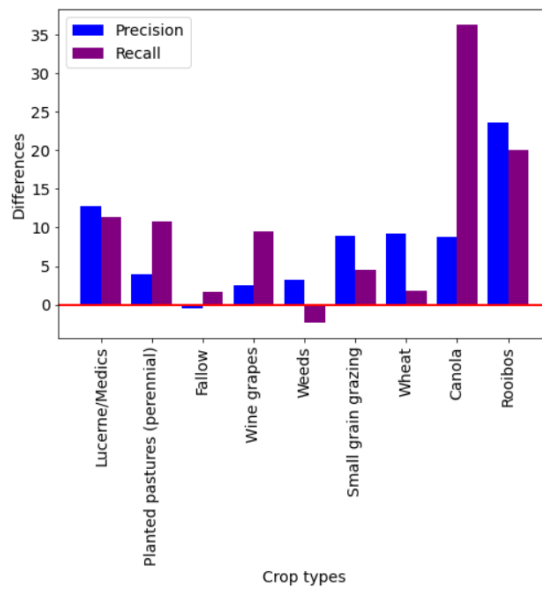


FIGURE 4.14: Difference between achieved precision and recall values in the proposed sample and the random sample.

Figure 4.15 gives a comparison of the F1-scores for the two samples.

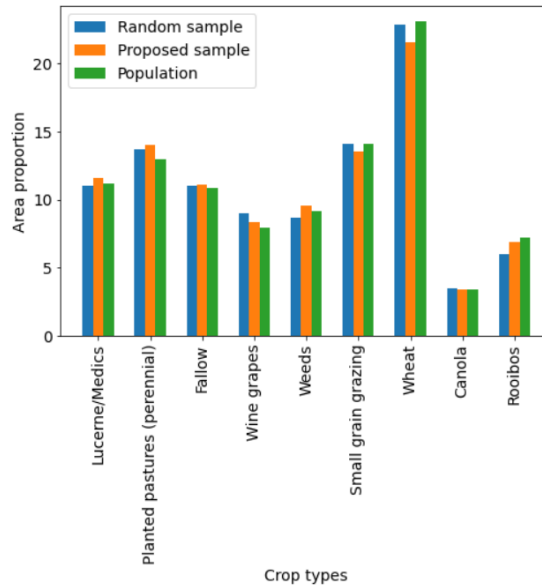


FIGURE 4.13: Proportions of the crop types using area coverage.

### 30% sample

When trained on the 30% proposed sample (445 images), the overall accuracy of the random forest classification algorithm is 74.260% with a RMSE of 1.695. To illustrate how accurate the classifier is per category, precision, recall as well as F1-scores are computed for each crop type in Table 4.4.

Crop Type	Achieved area coverage		
	Precision	Recall	F1-Score
Lucerne/Medics	51.105%	66.071%	0.576
Planted pastures	68.503%	60.763%	0.644
Fallow	43.046%	63.725%	0.514
Wine grapes	95.113%	89.835%	0.924
Weeds	58.02%	51.672%	0.547
Small grain grazing	47.156%	58.017%	0.520
Wheat	85.169%	73.511%	0.789
Canola	29.167%	67.742%	0.408
Rooibos	57.927%	65.517%	0.615

TABLE 4.4: Precision, recall and F1-scores per crop type

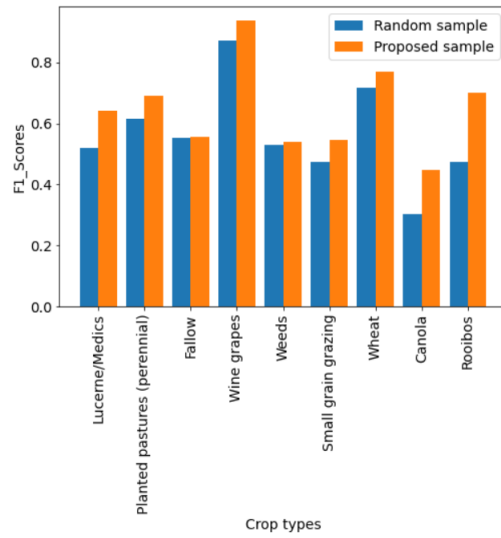


FIGURE 4.15: F1-scores between the proposed sample and random sample.

Figure 4.16 illustrates how the classifier was able to correctly classify each field to its true crop type.

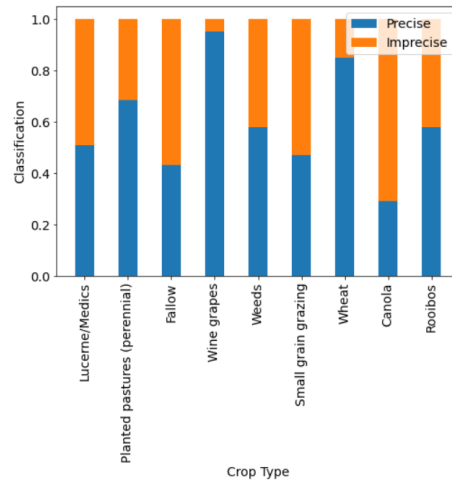


FIGURE 4.16: Precision per crop type

Next, a random sample with the same number of images as the 30% proposed sample was drawn. Figure 4.17 shows the difference between the number of fields belonging to each crop type in the two samples.

Figure 4.18 gives the area-wise proportions of the crop types between the random sample, the proposed sample as well as the population.

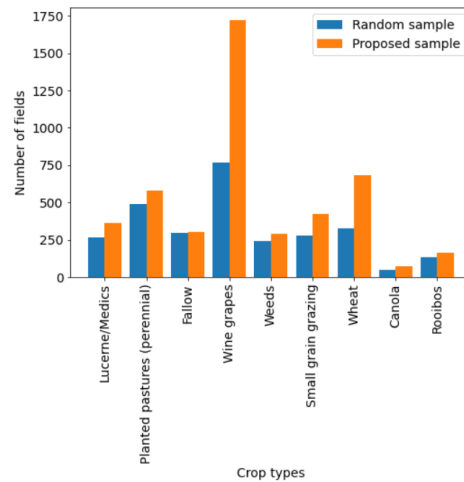


FIGURE 4.17: Number of fields per crop type in the two samples.

The random forest classifier is then trained on the random sample with 445 images. The overall accuracy of the classifier when trained on this random sample is 64.429% with a root mean square error of 1.905. The highest precision measure is 91.123% by wine grapes followed by wheat with 71.429% with canola being the least with a measure of 30.435%. Figure 4.19 gives an illustration of how the precision and recall values between this random sample and the 30% proposed sample differ.

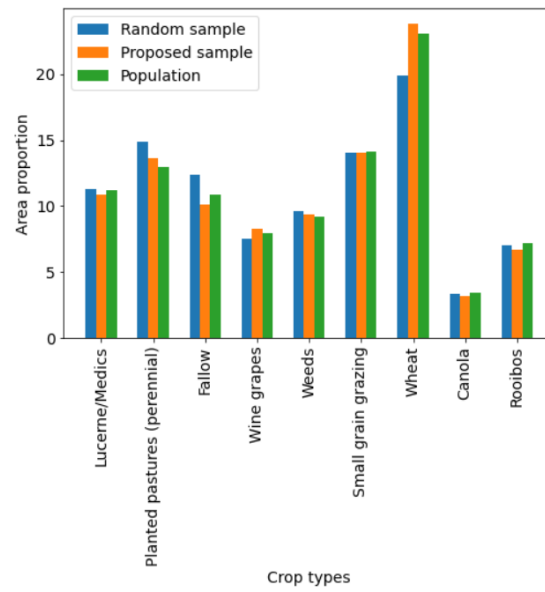


FIGURE 4.18: Proportions of the crop types using area coverage.

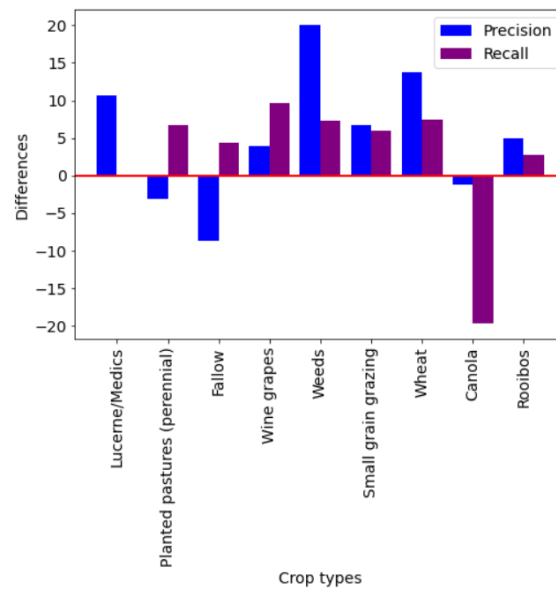


FIGURE 4.19: Difference between achieved precision and recall values in the proposed sample and the random sample.

Figure 4.20 gives a comparison of the F1-scores for the two samples.

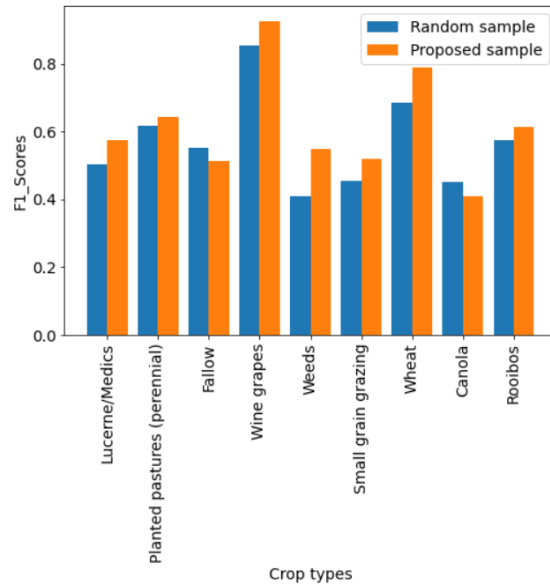


FIGURE 4.20: F1-scores between the proposed sample and random sample.

### 4.3 Summary

This chapter covers the pre-classification procedures which are feature engineering and feature importance. Three additional features were added to the already existing 12 image bands, and 7 of these were found to be the most informative features. A random forest classifier is then trained on the three samples resulting from the proposed stratified sampling algorithm. Random samples with the same number of images are drawn and also trained on for comparison. The classifier is assessed using overall accuracy, normalised RMSE, precision, recall and F1-scores.

## Chapter 5

# Discussion

The research aims to develop a stratified sampling algorithm for crop datasets. The algorithm works in such a way that the proportions of the crop types in the sample are representative to that in the population, while minimising the number of images sampled.

From Table 3.2, focusing on wheat, we see how the parameter *cropAmax* works. If *cropAmax* is set to 100%, then the achieved area coverage is already at the desired level of 10% in the first iteration which will increase when considering other crop types. The same is observed for the parameter *cropBmax* in Table 3.5. The desired sample size of weeds if *cropBmax* is set to 100% is already achieved at the second iteration. Using this information, one can conclude that using high values of *cropAmax* and *cropBmax* will lead to over-sampling. From Table 3.5, we can see that if *cropAmax* as 40% is chosen, then the least represented crop type in the sample is weeds but if *cropAmax* as 100% is chosen, rooibos is the least represented crop type in the sample and will be considered in the second iteration.

Table 3.6 shows how the area coverages increase as other crop types are considered throughout the remaining iterations. Most of the crop types achieved the desired 10% area coverage. Only wheat and weeds were under-sampled. The reason for this might be the choice of *cropAmax* (imposed on wheat) and *cropBmax* (imposed on weeds). One might argue that 40% is small and a higher value such as 60% might lead to the first two crops considered to not be undersampled. Table 3.7 shows the effect of the parameter *cropAmax* on the achieved sample area coverage given certain sample sizes. For smaller sample sizes, the achieved sample area coverages seem to increase with higher values of *cropAmax*. Note that the effect of *cropBmax* is not included in this instance. This seems to support the conclusion drawn above. If one assessed the level of accuracy by comparing the desired sample

sizes to the achieved sample area coverages, then the algorithm would be accurate. This is because the highest difference between the two is 5.8, and this is for a sample size of 70%. Looking at the smaller samples, which is ideally what we want to work with, the difference can be considered trivial. The achieved sample area coverages are always higher than the desired sample sizes in this instance because as much as we are considering area coverages, we consider a whole image and not just segments of images (fields). Thus if an image selected contains 1.6% of area coverage for a specific crop type during an iteration to meet the desired 10% sample size, and already from previously selected images we have a sample area coverage of 9.1%, then all the area will be considered and not just the required (10%-9.1%). We will therefore have 10.5% area coverage.

Table 3.8 gives a summary of how the number of images sampled changes according to desired sample size and the parameter *cropAmax*. For lower values of the sample size, the number of images sampled increases as the *cropAmax* parameter increases, which corresponds to the result from Table 3.7. For a 10% desired sample size, an average of 138 images are sampled, which makes up 5% of the total number of images. So instead of using simple random sample (SRS) which will result in 10% of all the images, one can use half the number of images using the proposed algorithm and get the same information area-wise. The algorithm ensures that the most information is obtained with fewer images. For lower sample sizes (10%-40%), roughly 48%-60% of the images that would be selected using SRS are obtained using the proposed algorithm. Even for a sample size of 90%, roughly 75% of images selected using a SRS approach are selected using this algorithm. This is due to the fact that some of the fields in the images were not labelled (62% of area is not labelled), so having more images does not necessarily mean having more information.

Table 3.9 shows how the area coverages achieved change with *cropBmax*. Note the effect of *cropAmax* is nullified by setting it to 100%. In this instance, the achieved area coverages increase with increasing values of *cropBmax* for all different sample sizes. This is different to when we were considering the effect of *cropAmax* only, where this was true only for smaller sample sizes. We see that the differences in this instance are lower compared to in Table 3.8. The highest difference decreased from 5.8 to 3.6 with the lowest being 0. The highest difference similar to when only considering *cropAmax* (Table 3.8) is from a high sample size of 70%.



The Euclidean norm is considered to compute the difference between the proportions of the crop types between a sample and the population. Only the *cropAmax* parameter is considered in this instance. As the sample size increases, the Euclidean norm increases as well and it is at its smallest at 100% sample size, as we expect. For lower values of the sample size, the Euclidean norm is at its lowest when *cropAmax* is at 50% and 60%. This might be because choosing lower values of *cropAmax* leads to undersampling, while higher values of *cropAmax* lead to oversampling. Looking at the 10% sample size in Figure 3.2, higher values of *cropAmax* and lower values of *cropBmax* give high Euclidean norm values, whereas low and high values of *cropAmax* give high Euclidean norms. As *cropBmax* increases, the Euclidean norm decreases. The lowest Euclidean norms are obtained when *cropAmax* is contained in (0.4-0.7) and *cropBmax* in (0.6-0.9). Higher values of *cropBmax* and middle values of *cropAmax* tend to give the smallest Euclidean values in the smaller sample sizes. The Euclidean norm is at its lowest when *cropBmax* is 0.9 and *cropAmax* is between 0.5 and 0.7.

From Figure 3.5, the lowest Euclidean norm is achieved when the sample size is 100%, with the second lowest being at a 10% sample size followed by 20%. As the sample size increases, so does the range of Euclidean norms (this is true until after sample size of 70%). For smaller sample sizes, the Euclidean norm range is quite small. The smallest range is [2.79,8.4] for sample size of 10% and is largest for sample size of 70% with [13.57,44.06]. The sample size of 70% does not only give the highest Euclidean norms, but also the highest difference in terms of sample size and achieved area coverage as already discussed.

Feature engineering and feature selection are two pre-processing techniques performed on the image data where the image bands are the features. One vegetation index and two water indices were added. These indices range between -1 and 1 and are constructed from 4 of the 12 bands, namely the NIR, the red, the green and the blue. After the 3 features were added to the existing 12 bands, the most important features were selected. Three feature selection techniques were used, namely mutual information regression, minimum-redundancy-maximum-relevance and F-test. The seven bands that had consistent high importance ratings were NDWI\_blue, NDWI\_green, NDVI, B04, B06, B07, B8A. The engineered features alone gave a lot of information when considered for vegetation detection purposes. B04 which is the red band has high reflectance for soils regions compared

to the other two colour bands and it's normally used to discriminate soil and vegetation. Figure 5.1 shows the level of reflectance (visibility) of the crop fields in the different colour bands. The red band has a higher visibility that clearly outlines the different fields of crop types when compared to the other two colour bands. This can also be seen when comparing the field labels in the colour bands to the label image in Figure 5.1[D].

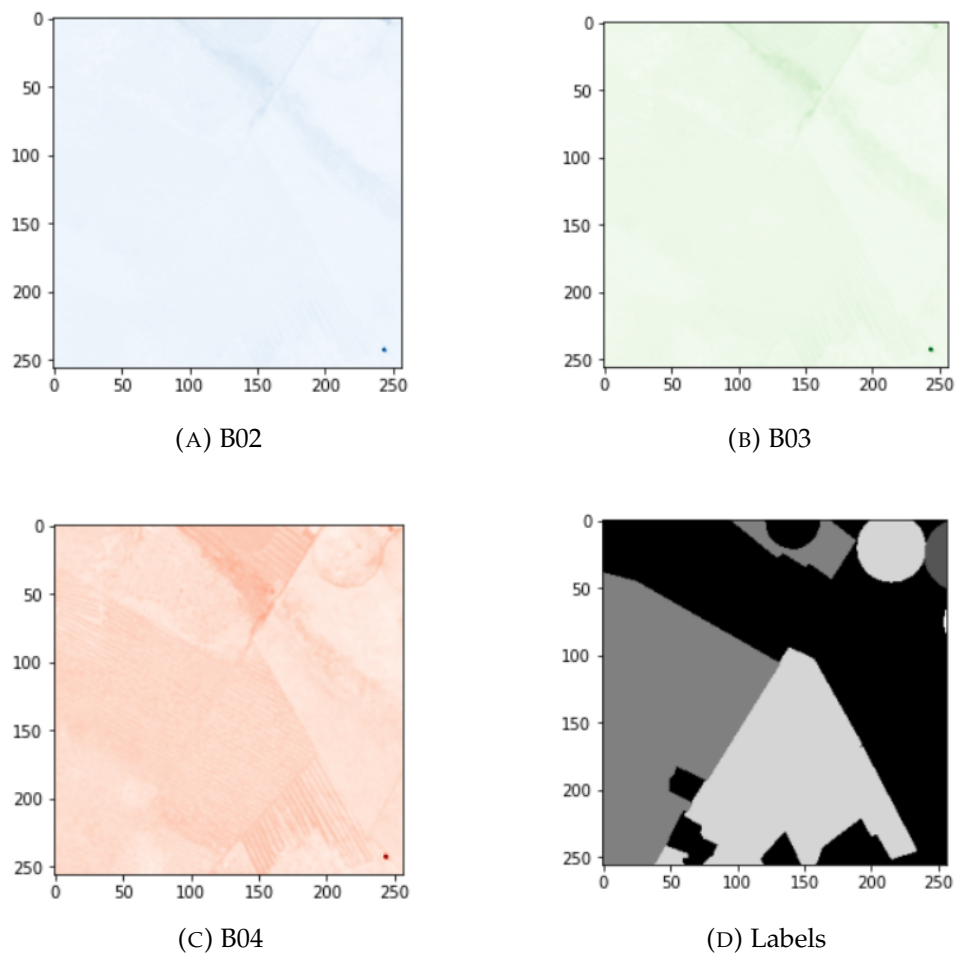


FIGURE 5.1: Visibility of the fields and its labels through the blue (B02), green (B03) and red (B04) bands.

The B06, B07 and B8A are the near-infrared NIR bands. These are useful in vegetation management as they can be used to measure vegetation density differences unlike other image bands because the light can transmit through the upper leaves and reflect off the bottom ones.

Considering the samples that were selected for training, the ones with the lowest Euclidean norms are chosen. The 10% sample chosen contains 136 images. Note that it does not have the lowest number of images achieved from the 10% samples. This is because the lowest number of images is achieved when *cropAmax* is the lowest as already discussed. However, the lowest Euclidean norm means the proportions within the sample are closest to those in the population. A lower Euclidean norm obtained equates to some degree as a good representation of the population by the sample. Hence the samples with the lowest Euclidean norms were considered and not necessarily the ones with the lowest number of images sampled. The 10% sample chosen is a result of setting up *cropAmax* to 0.6 and *cropBmax* to 0.9. For the 20% sample, the *cropBmax* remains the same but the *cropAmax* was set at a lower value of 0.5. For the 30% sample, the *cropBmax* value is still set to 0.9 but with a higher *cropAmax* value of 0.7.

The 10% sample that consists of 136 images has the lowest Euclidean norm not just compared to other 10% samples, but compared to all the other samples (10% to 90%). This should mean the area-wise proportions of the crop types should be closest to those in the population. The biggest difference between the proportion in the population and the sample is 0.7894 in planted pastures followed by 0.727 in wine grapes with the lowest being lucerne/medics with 0.0553. These are small differences, such that one can say the proportions are close to each other. The most represented crop type in both the population and sample is wheat (this is measured area-wise), and the least represented in both the sample and population is canola. However, the second most represented crop type in the population is the third most represented crop type in the sample and vice-versa. Apart from these two crop types, the order of the proportions of the crop types in the population and the sample are the same.

Figure 4.5 gives the difference of the proportions of the crop types in the sample using number of fields and area coverage. Wheat crops is the most contained crop type with a proportion of 22.9% followed by planted pastures with 13.8% followed closely by small grain grazing with 13.5%. The crop type that is the least contained is canola with a proportion of 3.3% followed by rooibos with 6.9%. However, when considering number of fields, wine grapes is highly represented followed by wheat and the least represented is still canola. Comparing wheat and wine grapes, area-wise, wheat is represented twice as much as wine grapes but when using number

of fields, wine grapes is represented more than 3 times as much as wheat. To better illustrate the difference between the area-wise proportions and number of fields, the averages of the field size of each crop type are computed. Figure 5.2 gives the average field sizes of the crop types. Canola has the highest average field size with an area of  $543.52 \text{ m}^2$  followed by wheat with an area of  $513.88 \text{ m}^2$ . Wine grapes has the highest number of fields but has a lower overall area coverage as it has the lowest average field size of  $78.31 \text{ m}^2$ .

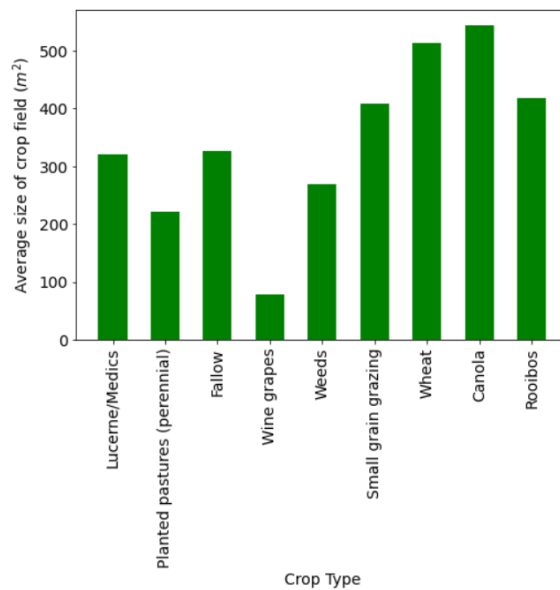


FIGURE 5.2: Average field sizes of each crop type.

Canola and rooibos are in the top three highest field sizes category but in the bottom two overall area coverage because they have the least number of fields.

When fitted on the 136 images, the classifier has a good accuracy score of 80.977% with a low RMSE of 1.4. Normalising the RMSE using the minmax method gives 0.18. Note the minmax is used throughout for normalising the RMSE. This is fairly close to 0 which means the model is a good fit for the data. Looking at the per category accuracy measures, wine grapes had the highest precision, recall and F1-score. This is due to it having the highest number of fields. Although wheat had the second highest precision value, its recall value is surpassed by that of canola. For canola, the classifier gives high recall with low precision. Recall that recall is defined as the quotient of correctly predicted positive field and total positive instances in a dataset (sample). Lucerne/medics, fallow, weeds, small grain grazing as well

as canola have higher recall values than precision values, i.e. the classifier returned more irrelevant instances than relevant instances. Rooibos, wheat and wine grapes both have high precision and recall values. The classifier gave F1-scores that are higher than 0.55 for all crop types except for canola. This means that overall, the model is a good fit for the crop data. Wine grapes has the highest correct classification followed by wheat and rooibos. Canola has the least correct classification and lucerne/medics with the second least correct classifications.

A random sample with the same number of images as the 10% proposed sample is drawn. Figure 4.7 shows the number of fields of each crop type. Wine grapes has the highest number of fields followed by planted pastures, then wheat, with the least still being canola and the second least being rooibos. The total number of fields in the random sample is 1141 whereas the ones in the proposed sample is 1493. Since most of the images contain unlabelled data (roughly 62% of the images is not labelled), ideally we want to sample images with the most labelled data, i.e. informative images. Comparing the two samples with the same number of images, we have that the sample coming from the proposed sampling algorithm gives around 2 times more information than from the random sample. Figure 5.3 shows the difference between the labelled and unlabelled data in the two samples. The proposed sampling algorithm resulted in a sample with 20.648% uninformative data while the random sample has over 59% uninformative data.

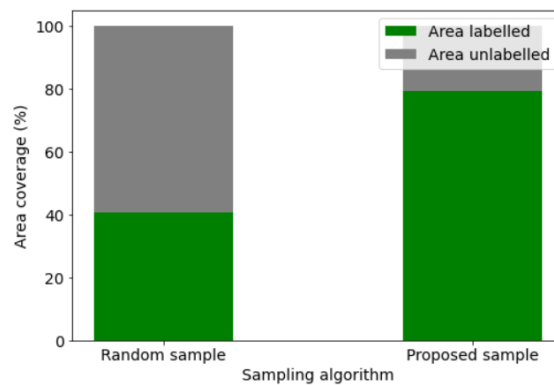


FIGURE 5.3: Labelled and unlabelled data between the proposed sample and random sample.

Looking at the area-wise proportions in Figure 4.8, we have already established that the biggest difference in the crop proportions is 0.7894 between the proposed sample and the population. With the random sample, the biggest difference is in

wine grapes with a difference of 3.702 followed by wheat with 3.32 and planted pastures with 1.327. The smallest difference in proportion is in small grain grazing with 0.239. Looking at Figure 4.8, the proportions of the population are much closer to the proposed sample than the random one.

When trained on the random sample, the classifier gave a lower accuracy of 69.062% which is lower than when trained on the proposed sample. The normalised RMSE is 0.2235 which is greater than that of the proposed sample. Figure 4.9 gives the differences between the precision and recall values in the proposed and random sample. The precision returned for the crops wine grapes is at 96.675% which is less by 1.568 from the precision returned when trained on the proposed sample. All crop types except for lucerne/medics has higher precision in the proposed sample than in the random sample. The classifier also gave higher recall values when trained on the proposed sample than when on trained on the random sample for all crop types. The highest difference in precision values between the two samples is for the crop type rooibos.

On average the proposed sample has better F1-scores than the random sample as shown in Figure 4.10. Only lucerne/medics has higher scores in the random sample than in the proposed sample. This is due to the higher precision value in the random sample than in the proposed sample. The biggest difference between F1-scores is in the crop type rooibos. This is due to the difference in the precision values as shown in Figure 4.9.

The 20% samples obtained from the proposed stratified algorithm has 278 images. This is about 10.49% of the total number of images. The classifier when trained on the 278 images has an overall accuracy of 76.479% which has declined by 4.498 from the 80.977% accuracy in the 10% proposed sample. The normalised RMSE has increased from 0.18 to 0.197 which is still quite low. Considering only these two measures, one may say the model is still good at predicting observed data. When trained on the 20% proposed sample, the classifier detected more noise than when trained on the 10% proposed sample. This is because of how the proposed sampling algorithm is setup: it takes the images with the most information and least noise first, so the higher the sample size, the higher the noise added. The images in the 10% sample from proposed algorithm are included in the 20% sampled from the proposed algorithm. Hence the noise in the 20% sample is higher than the one in the 10% proposed sample.

Table 4.3 gives the per category measures, these are the precision, recall as well as F1-scores. The highest precision and F1-score values are for wine grapes at 96.845% and 0.936 respectively. Unlike in the 10% sample where wine grapes had the highest recall, we now have canola at the highest value of 100%, which means the classifier was able to return all relevant instances belonging to the crop type canola. However, the classifier has the lowest precision value for canola which means as much as all relevant instances were detected, it detected way more irrelevant instances hence the overall lowest F1-score. Wheat has the second highest precision value but once again is surpassed by rooibos when looking at recall. Figure 5.4 illustrates how the precision and recall values have changed from the 10% proposed to the 20% proposed sample. The positive values indicate that the values in the 20% proposed sample are higher than those in the 10% proposed sample.

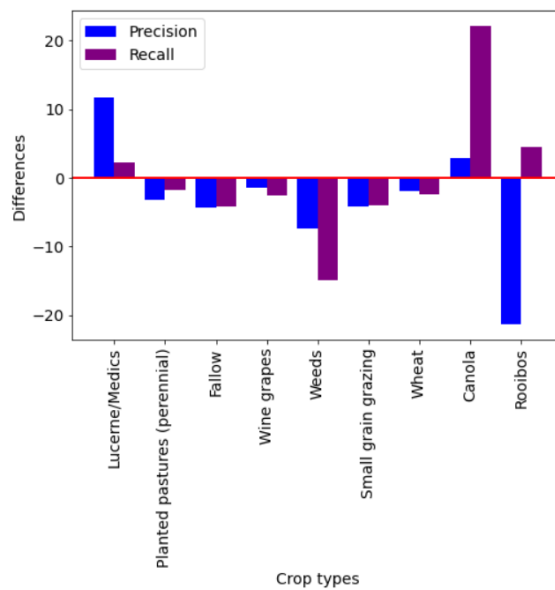


FIGURE 5.4: Differences in precision and recall values in the 10% and 20% proposed samples.

From Figure 5.4, we have two biggest differences. The recall value of canola increased from 77.778% to 100% and the precision value for rooibos which decreased from 84.783% to 63.366%. The many negative differences mean the 10% gave better precision and recall values than the 20% proposed sample. Overall, the classifier performed better on the 10% proposed sample than when trained on the 20%

proposed sample considering precision, recall, F1-scores values as well as overall accuracy and normalised RMSE.

A random sample of 278 images is drawn to be compared to the 20% proposed sample. Figure 4.12 gives the differences in the number of fields in each crop type. More fields are in the proposed sample than in the random sample for each crop type. The crop type wine grapes has twice as many fields in the proposed sample than in the random sample. Comparing Figure 4.12 to Figure 4.7, we have that the number of fields in the proposed samples (10% and 20%) are generally higher than the number of fields in their corresponding random samples. The total number of fields trained in the 20% proposed sample is 2942 which is more than the 2036 fields in the corresponding random sample containing the same number of images. Comparing the two samples with the same number of images, we have that the sample coming from the proposed sampling algorithm gives more information than the random sample. Figure 5.5 shows the difference between the labelled and unlabelled data in the two samples. The proposed algorithm resulted in a sample with 23.86% uninformative data while the random sample has over 60% uninformative data.

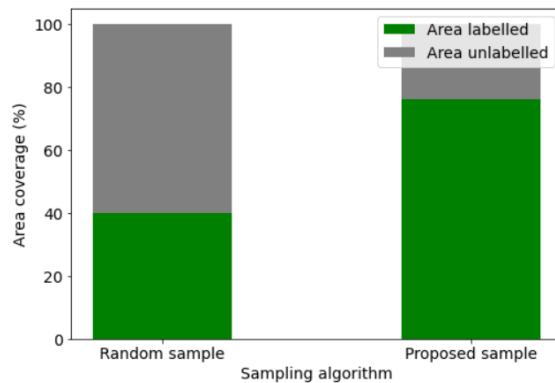


FIGURE 5.5: Labelled and unlabelled data in the 20% proposed and random sample with same number of images.

Figure 4.13 gives the area-wise proportions of the crop types in the two samples and the population. Five of the crop types had a higher area coverage in the proposed sample than in the random sample. Comparing this random sample with 278 images with the random sample with 136 images, the one with 278 images seems to have proportions that are a bit closer to those in the population, e.g. wine grapes are



no longer as oversampled and wheat as undersampled as they were in the sample with 136 images.

When trained on the random sample with 278 images, the classifier has an overall accuracy of 66.798% which is lower than what is achieved when trained on 20% proposed sample and also the random sample with 136 images. The normalised RMSE is 0.240125 which is higher than all previous obtained normalised RMSE values. A higher RMSE makes sense as increasing the training data increases noise. However, what is interesting is the decline in overall accuracy in both the proposed and the random samples. The accuracy has decreased by 4.9% in the proposed samples and by 2.264% in the random samples.

Looking at the per category measures in Figure 4.14, there are only two negative differences. These are the precision value for fallow and recall value for weeds. This means the precision and recall values are higher in the 20% proposed sample than they are in the corresponding random sample. The biggest difference is the recall measure for canola of 36.36%. This is followed by the precision values of rooibos of 23.04% which is consistent with the difference in the 10% samples in Figure 4.9. Figure 5.6 further illustrates how the precision and recall values differ for the different random samples. A positive difference means a precision or recall value achieved in the random sample with 278 images is higher than in the random sample with 136 images.

From Figure 5.6, the random sample with 136 images has higher precision and recall values than the other random sample with more images. The biggest difference is the precision value for rooibos with a decline of 13.4% followed by a decline in the recall value for canola from 75.0% to 63.636% in the random sample with 278 images. One can conclude based on the precision, recall, overall accuracy as well as normalised RMSE that the model performs better when trained on the random sample with 136 images as opposed to the one with 278 images. So far, the classifier performs best when trained on the 10% proposed sample, followed by the 20% proposed followed by random sample with 136 images and last being when trained on the random sample with 278 images.

Another sample obtained from the proposed sampling algorithm is considered, this sample has 445 images, this is 30% proposed sample. Note that 445 images is 16.79% of all 2650 images. This sample had a Euclidean norm of 7.47 which is higher

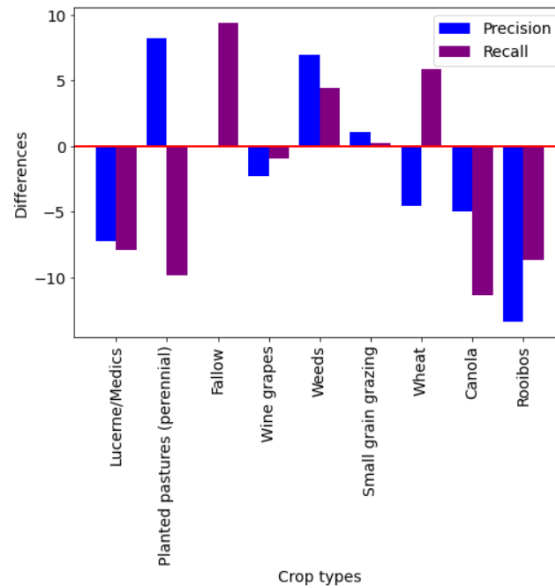


FIGURE 5.6: Differences in the precision and recall values for the random samples.

than the norm for the 20% proposed sample of 4.82 and that of 10% proposed sample of 2.79. This means that the proportionality between the population and the 30% proposed sample is close to each other, but not as close as the proportions between the 10% and 20% proposed sample to the population. The most represented crop type in both this 30% proposed sample and the population is wheat and the least also being canola. The second most represented crop type in both the sample and population is small grain grazing. The order of the area proportions of the crop types in the 30% proposed sample is the same as in the population. Looking at the proportions using number of fields in the 30% proposed sample, wine grapes is still the most represented followed by wheat with the least being canola followed by rooibos. This is similar to the order of crop types using number of fields in both the 10% and 20% proposed sample. Remember that the difference between the order of crop types using number of fields and also using area coverage is because of the different average field sizes of the crop types as shown in Figure 5.2.

When fitted on the 30% proposed sample, the classifier has a good accuracy of 74.26%. Note that this is lower than the accuracies obtained from the 10% and 20% samples but is higher than the accuracies for the previously considered random samples. The RMSE value is 1.695 which when normalised, gives a value

of 0.211875. This is higher than the normalised RMSE for the 10% and 20% proposed samples. This supports the statement that adding more data adds more noise. Also looking at how the accuracy values has decreased as the sample increases, this means that the model performs best when trained on the sample with the least noise, which is the 10% proposed sample. Diving deeper into the accuracy measures per category, precision, recall as well as F1-scores are computed for each crop type in the 30% proposed sample. The highest precision values is for wine grapes with a value of 95.113%, which is fairly high. This is followed by wheat with 85.169% and the least still being canola but now with an even lower value of 29.167%. The highest recall value is for wine grapes followed by wheat, with the lowest value of recall being 51.672% for weeds. The biggest difference between the precision and recall values is that of canola. The recall is far higher than the precision value. This means the classifier was able to retrieve a good amount of relevant instances, but more irrelevant were returned as well, hence the overall lower F1-score. Table 4.4 shows that planted pastures, wine grapes, weeds and wheat had higher precision values than recall values. Figure 5.7 compares these values to the ones achieved in 20% proposed sample. As before, positive values mean that the precision or recall values in the 30% proposed sample are higher than those in the 20% proposed sample.

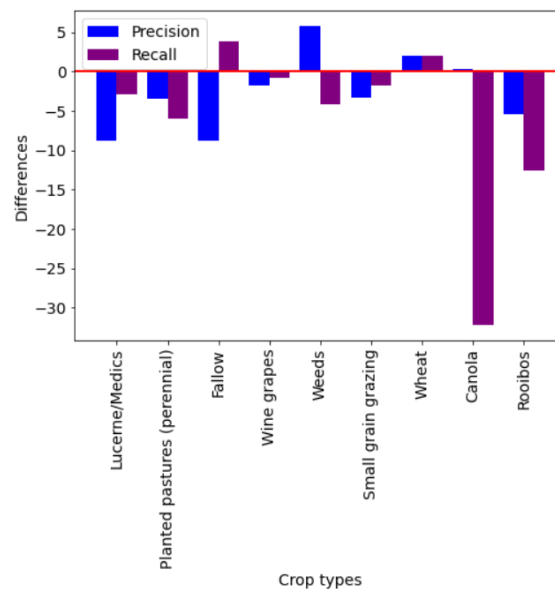


FIGURE 5.7: Differences in the precision and recall values in the 20% and 30% proposed sample.

Figure 5.7 shows that the 20% proposed sample has higher precision and recall values than the 30% proposed sample. Crop types such as wheat seem to have been better predicted in the 30% sample than in the 20% sample. There is a substantial decline in the recall value of canola as it decreased by 32.258% from 100% in the 20% proposed sample to 87.5% which is still a high value.

Wine grapes has the highest correct classifications followed by wheat which is closely followed by planted pastures and rooibos. The crop type with the least correct classifications is canola followed by fallow. Note that this order of crop types by correct classifications is different compared to when training on the 10% proposed sample because then we had rooibos as the third best classified crop type, planted pastures as fourth and lucerne/medics as second last as shown in Figure 4.6. Looking at Figure 4.11, the 20% proposed sample, the order is similar to what we have now as planted pastures is the third best classified followed by rooibos, but the second last is small grain grazing as opposed to fallow. When trained on the 30% proposed sample, the classifier has F1-scores that are higher than 0.51 for all crop types except for canola. Figure 5.8 gives the differences in the F1-scores of the three proposed samples. Overall, the 10% proposed sample has higher F1-scores followed by the 20% proposed sample with the ones obtained in the 30% sample being on average the lowest.

A random sample with the same number of images as in the 30% proposed sample is drawn. Figure 4.17 gives the comparison between the number of fields in these two samples. Wine grapes still has the highest number of fields in both the proposed and the random sample. However, the second most represented crop type using number of fields in this random sample is planted pastures instead of wheat. In the random sample, wheat is the crop type with the third highest number of fields. The crop type with the least number of fields is still canola followed by rooibos. The total number of fields in the random sample is 2845 whereas the ones in the proposed sample is 4596. Comparing the two samples with the same number of images we have that the sample coming from the proposed sampling algorithm gives more information than the random sample. Figure 5.9 shows the difference between the labelled and unlabelled data in the two samples. The 30% proposed sample resulted in a sample with 27.004% uninformative data while the random sample has over 61.7% uninformative data.

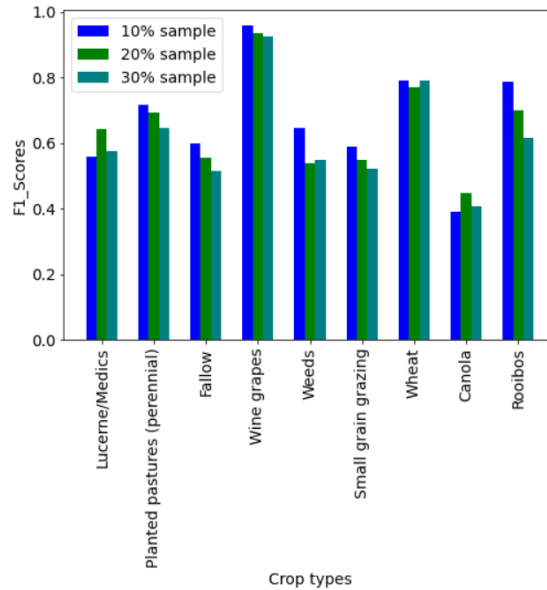


FIGURE 5.8: F1-scores of the crop types in the 10%, 20% and 30% proposed samples.

When trained on the random sample with 445 images, the classifier has an overall accuracy of 64.429% which is the lowest accuracy compared to the ones achieved on the other 5 samples. The normalised RMSE is 0.238125 which is the highest one yet. It does seem that the larger the sample size, the higher the RMSE. As the random sample sizes increases, the normalized RMSE increases and the overall accuracy decreases. Note that this is also true for the proposed samples, increasing the sample size, increased the error rate and decreased the overall accuracy. However, we do have that the classifier performed better when trained on the proposed samples than on the random samples. The accuracy has decreased by 2.219% in the proposed samples (20% and 30%) and by 2.369% in the random samples (278 and 445 images).

Looking at the per category measures in Figure 4.19, there are only a few negative differences. These are the precision values for planted pastures, fallow and canola and recall value for canola. This means the precision and recall values are higher in the 30% proposed sample than they are in the corresponding random sample. The biggest difference is the recall measure for canola of 19.758%. This is followed by the precision values of fallow of 8.655%. Figure 5.10 further illustrates

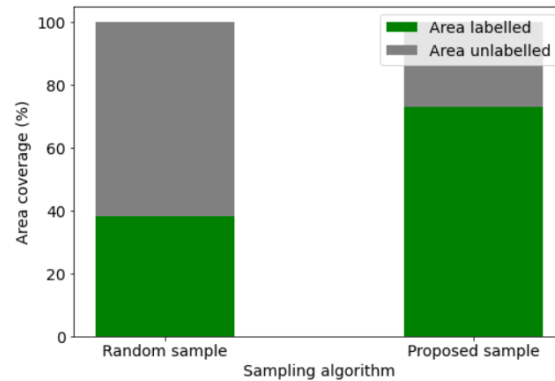


FIGURE 5.9: Labelled and unlabelled data in the 30% proposed and random sample with same number of images.

how the precision and recall values differ for the different random samples. A positive difference means a precision or recall value achieved in the random sample with 445 images is higher than in the random sample with 278 images.

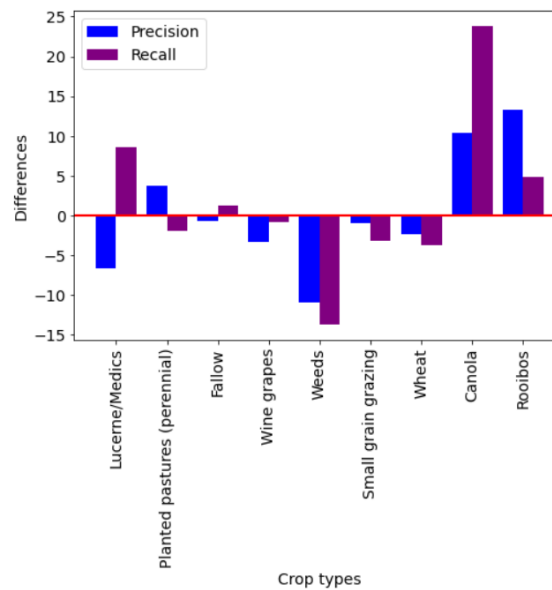


FIGURE 5.10: Differences in the precision and recall values for the random samples.

From Figure 5.10, the random sample with 278 images has more higher precision and recall values than the other random sample with 445 images. The biggest difference is the recall value for crop type canola with a decline of 23.864% followed by a decline in precision value for rooibos of 13.259%. One can conclude based on

the precision, recall, overall accuracy as well as normalised RMSE that the model performs better when trained on the random sample with 278 images as opposed to the one with 445 images. So far, the classifier performs best when trained on the 10% proposed sample. The list below shows the order in which the classifier performs, from best to least, when trained on the different samples.

1. 10% proposed sample (136 images).
2. 20% proposed sample (278 images).
3. 30% proposed sample (445 images).
4. Random sample with 136 images.
5. Random sample with 278 images.
6. Random sample with 445 images.

Figure 5.11 is a plot of accuracy values achieved by the classifier when trained on the different random and proposed samples.

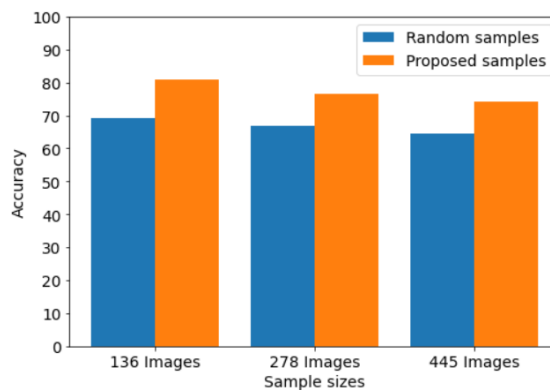


FIGURE 5.11: Accuracy values of the classifier when trained on the different random and proposed samples.

## Summary

Before the classification, feature engineering and selection processes are performed on the images. With image bands being the features, three additional features are created, namely the vegetation index (NDVI) and two water indices (NDWI\_blue

and NDWI\_green). Then three feature selection process, namely mutual information regression, mRMR and the F-test are performed. These revealed that the additional indices, the B04, B06, B07 and B8A are the most informative features. A random forest classifier with default arguments is trained on these samples as well as random samples with the same number of images as in the proposed samples for comparison. The 10%, 20% and 30% samples with the lowest Euclidean norms are used for classification. These are the samples with 136, 278 and 445 images respectively.

The classifier performed best when trained on the 10% proposed sample as it gave the highest accuracy, lowest normalised RMSE and overall higher F1-scores. This is then followed by the 20% proposed sample then by the 30% proposed sample. The classifier performs better when trained on the proposed samples than it does on the random samples with the same number of images. Also, increasing the sample sizes decreases the overall accuracy and increases the training error, this is true for both the proposed samples and the random samples. The main reason for this is the amount of informative data in the samples. The proposed samples have higher informative data than the random samples. And in both samples, the more data sampled, the more uninformative the data is.



## Chapter 6

# Conclusion

This mini-dissertation's objectives listed in Chapter 1 have been achieved as follows: First, metadata is obtained and constructed in the form of a dataframe that contains descriptive information of the images to be used for sampling. Next, a multivariate stratified sampling strategy is developed that aims to minimise the number of images sampled, keep the area-wise proportions in the sample and the population similar and maximise the information obtained from the images sampled. The proposed algorithm has two parameters, namely *cropAmax* and *cropBmax*. *cropAmax* is imposed on the first considered crop type (the highest contained crop type in the population) and *cropBmax* is imposed on the second considered crop type (least contained crop type in the population). The Euclidean norm is used to measure the closeness of the area-wise proportions in the samples to the ones in the population. The 10%, 20% and 30% samples with the lowest Euclidean norms are used for classification. These are the samples with 136, 278 and 445 images respectively.

A random forest classifier with default arguments is trained on these samples as well as random samples with the same number of images as in the proposed samples for comparison. The classifier performed best when trained on the 10% proposed sample as it gave the highest accuracy, lowest normalised RMSE and overall higher F1-scores. This is then followed by the 20% proposed sample then by the 30% proposed sample. The classifier performs better when trained on the proposed samples than it does on the random samples with the same number of images. Also, increasing the sample sizes decreases the overall accuracy and increases the training error, this is true for both the proposed samples and the random samples. The main reason for this is the amount of informative data in the samples. The proposed samples have higher informative data than the random samples. And in

both samples, the more data sampled, the more the uninformative data.

The importance of this research is to alleviate the memory requirement problem that often occurs when handling big geospatial data. The use of metadata and sampling from the metadata avoids having to read in all images whenever one wants to use them. The proposed sampling algorithm returned the lowest Euclidean norm for a 10% sample, which means the sample with least number of images has proportions closer to the population compared to the other samples (20% and above). Feature selection and engineering reduced the number of features from 12 bands to 7 informative bands, meaning less data to train on. In classification, the classifier performed best when trained on the proposed sample with the least number of images since it has the most informative images. Not only do all these alleviate the memory requirement problem, but also reduces time-consumption and the need for complex solutions. In summary:

- In Chapter 2, metadata of the large database of images is constructed.
- In Chapter 3, a stratified algorithm is developed that aims to keep proportions the same while minimising number of images sampled.
- The proposed sampling algorithm proved to be efficient and gave samples that are representative of the population.
- In Chapter 4, the various samples obtained from the proposed algorithm gave high accuracy values and low error values when a random forest classifier is trained on them.
- The classifier performed better when trained on the proposed samples than it did on the random samples with equivalent number of images.

One limitation includes lack of understanding in the changes of the recall and precision values for crop type canola in the different samples. A further investigation may address this issue. Even though the 30% proposed sample has 445 images, which is almost 17% of all images, training on these images is computationally heavy, hence higher samples were not considered. Proposals for future research include the usage of other land cover datasets to thoroughly assess the effectiveness of the proposed sampling algorithm.

Considering all the information the accuracy measures provided, the usage of metadata as well as the proposed sampling algorithm is beneficial for land cover detection purposes. This will help with the extraction of information, choosing a sample that best represents the population with least number of images but lot of information.

# Bibliography

- [1] Umut A Acar and Yan Chen. “Streaming big data with self-adjusting computation”. In: *Proceedings of the 2013 workshop on Data driven functional programming*. 2013, pp. 15–18.
- [2] Radiant Earth Foundation Western Cape Department of Agriculture. *Crop Type Classification Dataset for Western Cape, South Africa*. <https://doi.org/10.34911/rdnt.j0co8q>. May 2021.
- [3] Pradeep Anand. “Big Data is a big deal”. In: *Journal of Petroleum Technology* 65.04 (2013), pp. 18–21.
- [4] T.E. Avery and G.L. Berlin. *Fundamentals of Remote Sensing and Airphoto Interpretation*. Macmillan, 1992.
- [5] G Bryan Bailey, Donald T Lauer, and David M Carnegie. “International collaboration: the cornerstone of satellite land remote sensing in the 21st century”. In: *Space Policy* 17.3 (2001), pp. 161–169.
- [6] P. Dimitropoulos K. Barmpoutis P. Papaioannou and N Grammalidis. “A Review on Early Forest Fire Detection Systems Using Optical Remote Sensing”. In: *Sensors* 20.22 (2020), p. 6442.
- [7] Vic Barnett. *Sample survey principles and methods*. Arnold, 2002.
- [8] José R Berrendero, Antonio Cuevas, and José L Torrecilla. “The mRMR variable selection method: a comparative study for functional data”. In: *Journal of Statistical Computation and Simulation* 86.5 (2016), pp. 891–907.
- [9] Dempsey Caitlin. *Types of GIS Data Explored: Vector and Raster*. <https://www.gislounge.com/geodatabases-explored-vector-and-raster-data/>. May 2021.
- [10] Xue-Qi Cheng et al. “Survey on big data system and analytic technology”. In: *Journal of software* 25.9 (2014), pp. 1889–1908.

- [11] George Christakos. *Random field models in earth sciences*. Courier Corporation, 2012.
- [12] William G Cochran. *Sampling techniques*. John Wiley & Sons, 1977.
- [13] Nadir Omer Fadl Elssied, Othman Ibrahim, and Ahmed Hamza Osman. "A novel feature selection based on one-way anova f-test for e-mail spam classification". In: *Research Journal of Applied Sciences, Engineering and Technology* 7.3 (2014), pp. 625–638.
- [14] M Fahimi. "Cluster sampling". In: *Encyclopedia of Survey Research Methods* (2008), p. 99.
- [15] H. Dierssen R. Pettersen M. Van Ardelan G. Johnsen Z. Volent and F. Soriede. "Underwater hyperspectral imagery to create biogeochemical maps of seafloor properties". In: *Subsea Optics and Imaging*. Elsevier, 2013, 508–540e.
- [16] Bo-Cai Gao. "NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space". In: *Remote sensing of environment* 58.3 (1996), pp. 257–266.
- [17] Rober Haining. "Estimating spatial means with an application to remotely sensed data". In: *Communications in Statistics-Theory and Methods* 17.2 (1988), pp. 573–597.
- [18] Robert P Haining and Robert Haining. *Spatial data analysis: theory and practice*. Cambridge university press, 2003.
- [19] Christopher C Heyde and Eugene Seneta. *Statisticians of the Centuries*. Springer, 2001.
- [20] R Hoffer and M Roger. "Natural resource mapping in mountainous terrain by computer analysis of ERTS-1 satellite data". In: *Purdue Univ, Indiana* (1975).
- [21] Roger M Hoffer and Michael D Fleming. "Mapping vegetative cover by computer-aided analysis of satellite data". In: *General Technical Report RM*. (1978), p. 227.
- [22] A. Gruber G. E. Hunt House F. B. and A. T. Mecherikunnel. "History of satellite missions and measurements of the Earth radiation budget (1957–1984)". In: *Reviews of Geophysics* 24.2 (1986), pp. 357–377.

- [23] XF Jiao, BJ Yang, ZY Pei, et al. "Paddy rice area estimation using a stratified sampling method with remote sensing in China". In: *Transactions of the CSAE* 22.5 (2006), pp. 105–110.
- [24] Xiaolong Jin et al. "Significance and challenges of big data research". In: *Big data research* 2.2 (2015), pp. 59–64.
- [25] Navid Kardani et al. "Improved prediction of slope stability using a hybrid stacking ensemble method based on finite element analysis and field data". In: *Journal of Rock Mechanics and Geotechnical Engineering* 13.1 (2021), pp. 188–201.
- [26] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. "Estimating mutual information". In: *Physical review E* 69.6 (2004), p. 066138.
- [27] V Kurama. "A Guide to AdaBoost: Boosting To Save The Day". In: *Paperspace Blog* 23 (2020).
- [28] Doug Laney et al. "3D data management: Controlling data volume, velocity and variety". In: *META group research note* 6.70 (2001), p. 1.
- [29] L Lanza and M Conti. "Cloud tracking using satellite data for predicting the probability of heavy rainfall events in the Mediterranean area". In: *Surveys in Geophysics* 16.2 (1995), pp. 163–181.
- [30] Teerawong Laosuwan, Torsak Gomasathit, and Tanutdech Rotjanakusol. "Application of remote sensing for temperature monitoring: The technique for land surface temperature analysis". In: *Journal of Ecological Engineering* 18.3 (2017).
- [31] Kisung Lee et al. "Efficient spatial query processing for big data". In: *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 2014, pp. 469–472.
- [32] Tae-Hwy Lee, Aman Ullah, and Ran Wang. "Bootstrap aggregating and random forest". In: *Macroeconomic forecasting in the era of big data*. Springer, 2020, pp. 389–429.
- [33] Guo-jie Li and Xue-qi Cheng. "Research status and scientific thinking of big data". In: *Bulletin of Chinese Academy of Sciences* 27.6 (2012), pp. 647–657.

- [34] Songnian Li et al. "Geospatial big data handling theory and methods: A review and research challenges". In: *ISPRS journal of Photogrammetry and Remote Sensing* 115 (2016), pp. 119–133.
- [35] Thomas Lillesand, Ralph W Kiefer, and Jonathan Chipman. *Remote Sensing and Image Interpretation*. John Wiley & Sons, 2015.
- [36] Sharon L Lohr. *Sampling: design and analysis*. Chapman and Hall/CRC, 2021.
- [37] Seema Maitrey and CK Jha. "Handling big data efficiently by using map reduce technique". In: *2015 IEEE International Conference on Computational Intelligence & Communication Technology*. IEEE. 2015, pp. 703–708.
- [38] Bertil Matérn. *Spatial variation*. Vol. 36. Springer Science & Business Media, 2013.
- [39] G Matheron. "The theory of regionalised variables and its applications". In: *Les Cahiers du Centre de Morphologie Mathématique* 5 (1971), p. 212.
- [40] Georges Matheron. "Principles of geostatistics". In: *Economic geology* 58.8 (1963), pp. 1246–1266.
- [41] Roy A Mead and Merle P Meyer. "Landsat digital data application to forest vegetation and land use classification in Minnesota". In: *LARS Symposia*. 1977, p. 220.
- [42] A Milne. "The centric systematic area-sample treated as a random sample". In: *Biometrics* 15.2 (1959), pp. 270–297.
- [43] Sava Mintchev. "User-Defined Rules Made Simple with Functional Programming". In: *International Conference on Business Information Systems*. Springer. 2014, pp. 229–240.
- [44] Carlos Miranda et al. "Sampling stratification using aerial imagery to estimate fruit load in peach tree orchards". In: *Agriculture* 8.6 (2018), p. 78.
- [45] Emad A Mohammed, Behrouz H Far, and Christopher Naugler. "Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends". In: *BioData mining* 7.1 (2014), pp. 1–23.
- [46] David Opitz and Richard Maclin. "Popular ensemble methods: An empirical study". In: *Journal of artificial intelligence research* 11 (1999), pp. 169–198.

- [47] Jeffrey Pomerantz. *Metadata*. MIT Press, 2015.
- [48] Julien Radoux and Patrick Bogaert. “Good practices for object-based accuracy assessment”. In: *Remote Sensing* 9.7 (2017), p. 646.
- [49] Shashi Shekhar et al. “Benchmarking spatial big data”. In: *Specifying Big Data Benchmarks*. Springer, 2012, pp. 81–93.
- [50] Robert R Sokal. “Classification: purposes, principles, progress, prospects”. In: *Science* 185.4157 (1974), pp. 1115–1123.
- [51] Alfred Stein and Christien Ettema. “An overview of spatial sampling procedures and experimental design of spatial studies for ecosystem comparisons”. In: *Agriculture, Ecosystems & Environment* 94.1 (2003), pp. 31–47.
- [52] MD Steven and Jeremy Austin Clark. *Applications of Remote Sensing in Agriculture*. Elsevier, 2013.
- [53] Joris Toonders. “Data is the new oil of the digital economy”. In: *Wired* (2014).
- [54] Nam-Luc Tran et al. “Arom: Processing big data with data flow graphs and functional programming”. In: *4th IEEE International Conference on Cloud Computing Technology and Science Proceedings*. IEEE. 2012, pp. 875–882.
- [55] CJ Van Westen. “Remote sensing for natural disaster management”. In: *International Archives of Photogrammetry and Remote Sensing* 33.B7/4; PART 7 (2000), pp. 1609–1617.
- [56] Velocity Case Studies on Volume, Xun Zhou Variety Michael R. Evans Dev Oliver, and Shashi Shekhar. “Spatial Big Data: Case Studies on Volume, Velocity, and Variety”. In: *Big Data: Techniques and Technologies in Geoinformatics*. CRC Press, 2019, pp. 163–190.
- [57] HW Walton. *Understanding and controlling distortion in large bearing rings-some practical aspects*. Tech. rep. ASM International, Materials Park, OH (United States, 1996.
- [58] Shuliang Wang and Hanning Yuan. “Spatial data mining in the context of big data”. In: *2013 International Conference on Parallel and Distributed Systems*. IEEE. 2013, pp. 486–491.
- [59] Yuanzhuo Wang, Xiaolong Jin, and XQ Cheng. “Network big data: present and future”. In: *Chinese Journal of Computers* 36.6 (2013), pp. 1125–1138.



- [60] Andreas Weingessel, Evgenia Dimitriadou, and Kurt Hornik. *An ensemble method for clustering*. 2003.
- [61] Li Wen and Michael Hughes. "Coastal wetland mapping using ensemble learning algorithms: A comparative study of bagging, boosting and stacking techniques". In: *Remote Sensing* 12.10 (2020), p. 1683.
- [62] A Zaitunah, AG Ahmad, RA Safitri, et al. "Normalized difference vegetation index (ndvi) analysis for land cover types using landsat 8 oli in besitang watershed, Indonesia". In: *IOP Conference Series: Earth and Environmental Science*. Vol. 126. 1. IOP Publishing. 2018, p. 012112.
- [63] Gunter Zeug and Sandra Eckert. "Population growth and its expression in spatial built-up patterns: The Sana'a, Yemen case study". In: *Remote Sensing* 2.4 (2010), pp. 1014–1034.
- [64] Shuang Zhu and Jinshui Zhang. "Provincial agricultural stratification method for crop area estimation by remote sensing". In: *Transactions of the Chinese Society of Agricultural Engineering* 29.2 (2013), pp. 184–191.