Assessing Classification Performance for Sampled Remote Sensing Data

Tshepiso Rangongo[1], Inger Fabris-Rotelli[1], Renate Thiede[1]

[1]Department of Statistics, University of Pretoria, South Africa
u17052395@tuks.co.za, inger.fabris-rotelli@up.ac.za, renate.thiede@up.ac.za

Keywords: Sampling, Metadata, Crop Classification

Abstract

Big data poses challenges for storage, management, processing, analysis and visualisation. One technique of handling big data is the use of a representative sample of the data. This paper 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 in the sample. 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 samples and accuracy is assessed. Precision, recall and F_1-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. In addition, the classifier performed better on the proposed sample than it did on a random sample as the proposed sample due to the more informative data. This research develops an effective way of sampling big data for crop classification.

1. Introduction

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) increases the volume, velocity, and variety of geospatial data. Information from remote sensors is used for various purposes including biodiversity monitoring (Lausch et al., 2024), weather and catastrophe forecasting (Maqsood et al., 2024), as well as crop classification (Barriere et al., 2024).

The analysis of big geospatial data is difficult due to the complexity of this data (Gomes et al., 2020). 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 (Li et al., 2016). Metadata summarises big data, alleviating memory requirement in cases where metadata can be used instead of reading all the big geospatial data. One such case is sampling, such as sampling from the metadata instead of reading in all the geospatial data. This is important to consider as spatial data very quickly becomes unmanageable in size, requiring access to cloud computing for analysis and also presenting complexity in the initial storage and access of the data content.

Obtaining a representative sample of the data from a large metadata source of crop classification data is a challenge. Stratified sampling is appropriate for crop classification because this sampling technique requires that each unit must belong to only one stratum, and in crop classification, one crop can only belong to one crop type. Applications of stratified sampling in remote sensing include the quantification of spatial variability amongst peach orchids. This was in turn used to classify trees into homogenous groups (sampling strata) with the aim of decreasing sampling size in (Miranda et al., 2016). Other applications in the estimation of crop area using sampling in remote sensing can be seen in (Jiao et al., 2006, Zhu and Zhang, 2013, Schulthess et al., 2023a, Li et al., 2023).

This paper makes use of a recently proposed algorithm (Rangongo et al., 2022) 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 data frame. Then the metadata itself is used to obtain a desired sample using the algorithm. The aim of the algorithm is to achieve similar proportionality of crop types between the sample and the population as well as minimise the number of images sampled while maximising the information obtained in the images. Various resulting sample sizes are used for land cover classification, with a random forest. The performance is assessed relative to the sample size.

Section 2 provides the metadata construction. Section 3 covers the algorithm as well as its implementation. Section 4 covers classification and implementation. Section 5 discusses the results, while Section 6 concludes and proposes future research.

2. Metadata

2.1 Data Summary

The crop dataset used 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[1]. The dataset 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 until the 27th of November (48 dates). Each image has 12 bands of one area of land with

[1] Crop Type Classification Dataset for Western Cape, South Africa. Available online: `https://doi.org/10.34911/rdnt.j0co8q` (accessed on 12 March 2022)

tile ID 1114 taken by the Sentinel-2 satellite. The images were resampled by the data providers so that all the images have the same resolution of 60m.

Each image in the dataset is an area of land made up of crop fields. Each field contains only one of nine crop types, namely fallow, canola, wheat, wine grapes, weeds, small grain gazing, lucerne/medics, planted pastures (perennial), and rooibos. Each area of land (2650 locations) was captured every five days through 12 bands of the electromagnetic spectrum and the 13th image showing cloud coverage (only captured on 13 of the 48 dates) so that the whole data is made up of 1 653 000 images. The area of interest is 23 850km^2 of land of which 9 063km^2 (roughly 38%) has been labelled and it 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 are available to determine the proportion of the crop types in the population.

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 and collated with the given metadata as a pre-processing step. 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. 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. An image of another area of land thus with a different tile ID will not necessarily have the same information. 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. With the three categories brought together, metadata in the form of a database can be created. 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 since it then does not require loading and reading all the images into memory.

3. A Multivariate Stratified Sampling Algorithm

This section presents the multivariate stratified sampling algorithm in (Rangongo et al., 2022). Let N be the number of images in the population and M be the number of crop types. Let n be the sample size of images and N_i be the number of images that contain crop type i in the population. We notate A^i_{pop} and A^i_{samp} 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.

$$\mathbf{V}_{pop} = \begin{bmatrix} V^1_{pop} \\ V^2_{pop} \\ \vdots \\ V^M_{pop} \end{bmatrix}, \mathbf{V}_{samp} = \begin{bmatrix} V^1_{samp} \\ V^2_{samp} \\ \vdots \\ V^M_{samp} \end{bmatrix},$$

$$\mathbf{A}_{pop} = \begin{bmatrix} A^1_{pop} \\ A^2_{pop} \\ \vdots \\ A^M_{pop} \end{bmatrix} \text{ and } \mathbf{A}_{samp} = \begin{bmatrix} A^1_{samp} \\ A^2_{samp} \\ \vdots \\ A^M_{samp} \end{bmatrix}.$$

We propose an algorithm to obtain a sample from the population which ensures that the proportion between the population and the sample are similar while minimising 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 ($\mathbf{A}_{samp}$). 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 ϵ:

$$||\mathbf{V}_{pop} - \mathbf{V}_{samp}|| \leq \epsilon \tag{1}$$

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 algorithm starts by calculating $\mathbf{A}_{pop}$, from this the proportions of the crop types,$\mathbf{V}_{pop}$, are computed. Then a sub-dataframe that contains the only the images that have the crop type with the highest proportion in the population. The images are ordered in terms of area coverage of the crop type considered. A parameter, $cropAmax$, is imposed on crop type considered to ensure that when other crop types are considered, the area coverage of this crop type is not exceeded. When sampling, after achieving the desired $cropAmax\%$ of area coverage of the first considered crop type, these images will form a sample. From this sample, the crop type with the least representation area-wise, is considered. A new sub-dataframe is extracted that only contains images with this crop type. Note images used in the previous sample are excluded from the population. A parameter $cropBmax$ is imposed on this crop type to ensure that the desired area coverage is not exceeded. Images selected during this iteration are added to the previous sample. Another sub-dataframe is extracted that contains images with the least represented crop type in the current sample, not one of the previously considered crop types. The images selected during this iteration will be added to the sample and another crop type that is least represented will be considered until the desired area coverages of all crop types are achieved. From the final sample, $\mathbf{A}_{samp}$, the proportions in the sample $\mathbf{V}_{samp}$ are determined.

The implementation of the sampling algorithm is done in Python and the notebook containing the code for the algorithm is available on Figshare[7]. The role of the various parameters was investigated in (Rangongo et al., 2022).

[7] Sampling algorithm, Figshare, Python code, `https://doi.org/10.25403/UPresearchdata.20444061`

4. Land Cover Classification

The goal herein is to investigated the effect of the sample size on the precision of the classification of landcover. Pre-processing procedures are performed on the image data, namely feature engineering and feature selection. 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) (Zaitunah et al., 2018). The NDVI is an indicator used to assesses 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 NDVI $= \left(\frac{NIR-Red}{NIR+Red}\right)$ 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 (Gao, 1996). Hence they are used in vegetation related applications. The two NDWI used are the NDWI_green and the NDWI_blue with the following formulas. The Blue is the blue visible image band and the Green is the green visible image band.

$$\text{NDWI\_green} = \left(\frac{NIR - Green}{NIR + Green}\right) \quad (2)$$

$$\text{NDWI\_blue} = \left(\frac{NIR - Blue}{NIR + Blue}\right) \quad (3)$$

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 (Kraskov et al., 2004), minimum-redundancy-maximum-relevance (mRMR) (Berrendero et al., 2016) and the F-test (Elssied et al., 2014). The mRMR selects features that reduce their redundancy in the presence of other features while simultaneously increasing their own relevance. The F-test (correlation-based method) calculates a correlation coefficient which is then converted into a F-statistic. An F-test was performed and the statistically significant features with the highest F-statistics were chosen. The mutual information regression works to identify any sort of dependence between features and eliminates those with high dependency. The mutual information regression is determined as

$$I(X;Y) = H(X) - H(X|Y) \quad (4)$$

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.

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.

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.

The selected machine learning algorithm is random forest as not only is it easier to implement, but is widely used for crop classifications (Su and Zhang, 2021, Schulthess et al., 2023b, Tariq et al., 2023). 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. The random forest classifier was trained on the seven features. The arguments used for the classifier were default parameter values.

5. Results

5.1 10% Sample

The classifier was trained on 136 images sampled when a 10% sample was targeted. 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 F_1-score are shown in Table 1.

Crop Type	Accuracy assessments		
	Precision	Recall	F_1-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 1. Precision, recall and F_1-scores per crop type for the 10% sample.

Further, a comparison was made between samples from the proposed sampling algorithm and 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. Figure 1 shows the representations of the crop types in the two samples compared to the population.

Training the same random forest classifier on the random sample, an overall accuracy of 69.062% was achieved with an RMSE of 1.788. Figure 2 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 3 shows a graph of F_1-scores for each crop type for both the random sample and 10% proposed sample.

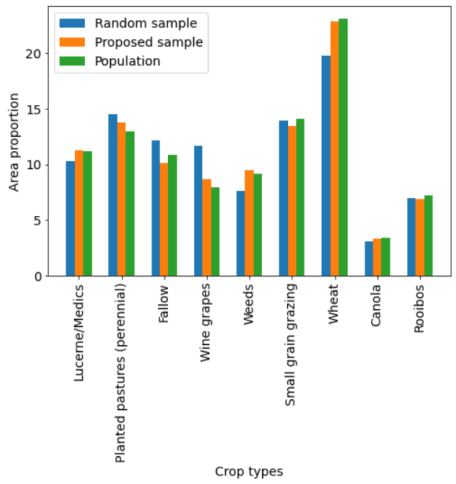

Figure 1. Area-wise proportions of the crop types in the 10% proposed sample, random sample and the population.

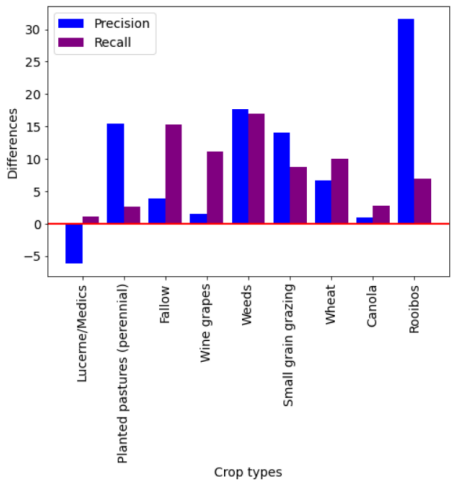

Figure 2. Difference between achieved precision and recall values in the proposed sample and the random sample.

5.2 20% Sample

The random forest classifier was also trained on a 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 accuracy assessments were measured. Table 2 contains the precision, recall and F_1-scores per crop type.

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. The random forest classifier was 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 gives an illustration of how the precision and recall values between this random sample and 20% proposed sample differ. Figure 5 gives a comparison of the F_1-scores for the two samples.

5.3 30% Sample

When trained on the 30% proposed sample (445 images), the overall accuracy of the random forest classification algorithm was 74.260% with a RMSE of 1.695. To illustrate how accurate

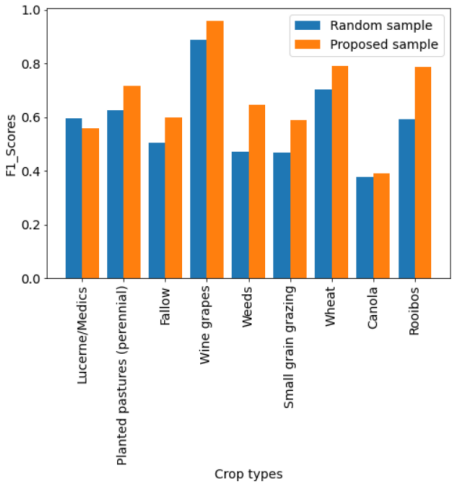

Figure 3. F_1-scores between the random sample and the proposed sample.

Crop Type	Achieved area coverage		
	Precision	Recall	F_1-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 2. Precision, recall and F_1-scores per crop type for the 20% sample.

the classifier is per category, precision, recall as well as F_1-scores are computed for each crop type in Table 3.

Again a random sample with the same number of images as the 30% proposed sample was drawn. The overall accuracy of the classifier when trained on this random sample is 64.429% with a root mean square error of 1.905. Figure 6 shows the precision and recall values between this random sample and the 30% proposed sample. Figure 7 gives a comparison of the F_1-scores for the two samples.

6. Discussion

A random forest classifier was trained on three samples obtained using the proposed sampling algorithm. Random samples with the same number of images were drawn and also trained for comparison. The classifier was assessed using overall accuracy, normalised RMSE, precision, recall, and F_1-scores. The improved sampling 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.

Feature engineering and feature selection, as pre-processing techniques, were 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,

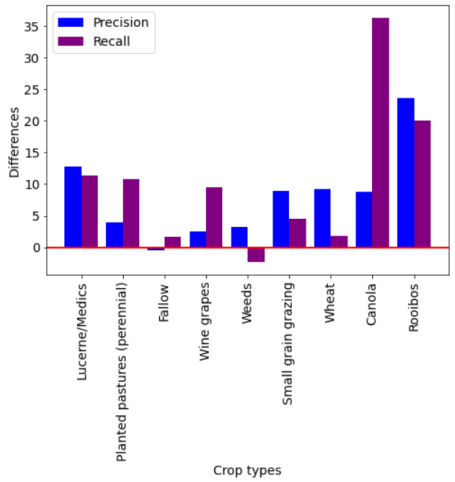

Figure 4. Difference between achieved precision and recall values in the proposed 20% sample and the random sample.

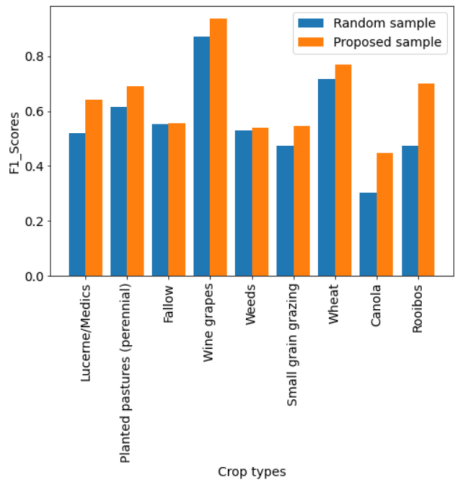

Figure 5. F_1-scores between the proposed 20% sample and random sample.

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 10% sample, using the new sampling algorithm, contained 136 images. Note that the 10% sample has the lowest Euclidean norm not just compared to other 10% samples, but compared to all the other samples considered from 10% to 90%. Thus the area-wise proportions of the crop types are closest to those in the population.

Considering the random 10% sample, most of the images contain unlabelled data (roughly 62% of database of images are not labelled). Ideally, a sample should contain sample images with the most labelled data, i.e. informative images. Comparing the two samples with the same number of images, we see that the sample coming from the proposed sampling algorithm contains around twice more information than from the random sample. Figure 8 shows the difference between the labelled and

Crop Type	Achieved area coverage		
	Precision	Recall	F_1-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 3. Precision, recall and F_1-scores per crop type for the 30% sample

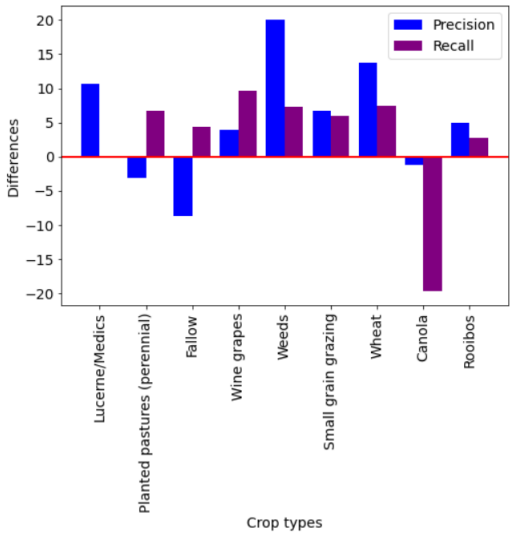

Figure 6. Difference between achieved precision and recall values in the proposed 30% sample and the random sample.

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.

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. However, 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% sample from the proposed algorithm. Hence the noise in the 20% sample is higher than the one in the 10% proposed sample.

A random sample of 278 images is drawn to be compared to the 20% proposed sample. More fields are in the proposed sample than in the random sample for each crop type. The number of fields in the proposed samples (10% and 20%) are generally higher than the number of fields in their corresponding random

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-2024-279-2024 | © Author(s) 2024. CC BY 4.0 License.

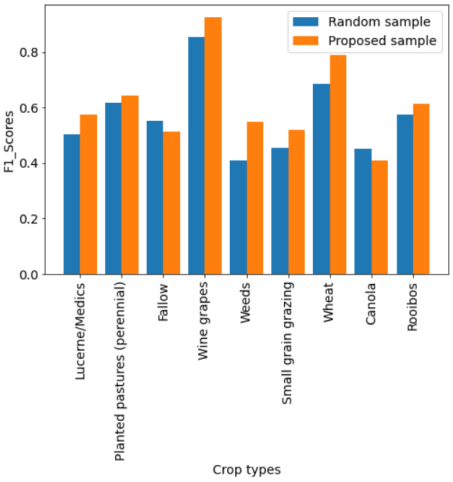

Figure 7. F_1-scores between the proposed 30% sample and random sample.

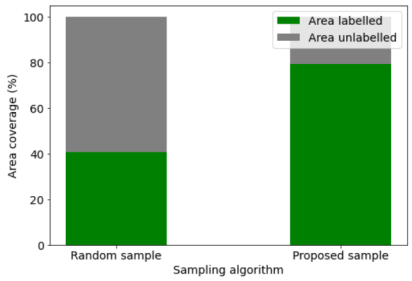

Figure 8. Labelled and unlabelled data between the 10% proposed sample and random sample.

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 9 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.

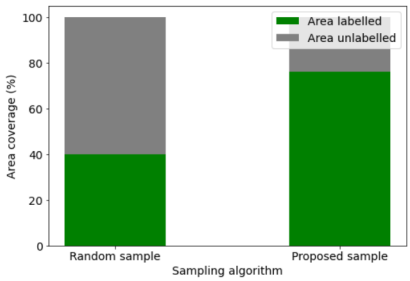

Figure 9. Labelled and unlabelled data in the 20% proposed and random sample with same number of images.

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.

In the 30% proposed sample the 445 images represent 16.79% of the total 2650 images. This sample had a Euclidean norm of 7.47 which is higher than the norm for the 20% proposed sample of 4.82 and that of 10% proposed sample of 2.79. This mean 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. 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 values 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 F_1-scores are computed for each crop type in the 30% proposed sample.

A random sample with the same number of images as in the 30% proposed sample is drawn. 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 10 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.

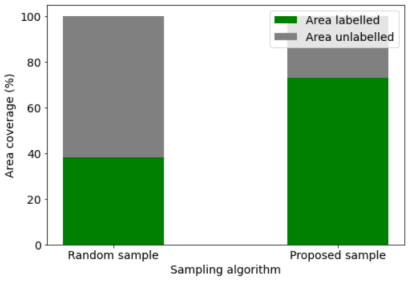

Figure 10. Labelled and unlabelled data in the 30% proposed and random sample with same number of images.

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).

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. Figure 11 is a plot of accuracy values achieved by the classifier when trained on the different random and proposed samples. The main reason for these results is the amount of informative data in the samples. The proposed samples have higher informative data than the random samples. In both samples, the more data sampled, the more uninformative the data is since the algorithm acts by determining most representative data first.

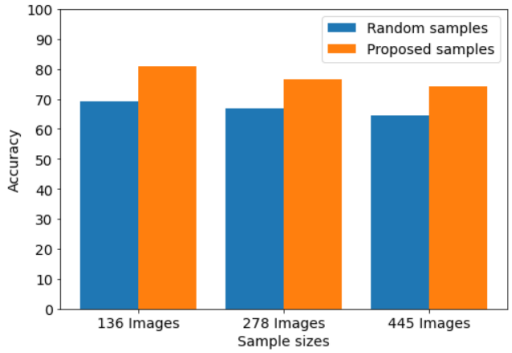

Figure 11. Accuracy values of the classifier when trained on the different random and proposed samples.

7. Conclusions

The big data used in this research is crop classification remote sensing data, that is freely available. First, metadata was obtained and constructed in the form of a dataframe that contains descriptive information of the images to be used for sampling. The construction of metadata can also be done on any land cover dataset, it is not limited to crop data only. The usage of metadata alleviated the memory requirement that an application may need as instead of reading all images in, metadata can be used, as this also saves the time required.

Next, a multivariate stratified sampling strategy is made use of that aims to minimise the number of images sampled, keep the area-wise proportions in the sample and the population similar while maximising the information obtained from the images sampled. Euclidean norms were used to measure the closeness of the proportions in the samples and the population. We had that the sample with the least number of images had proportions closest to the ones in the population. Also, from the samples achieved, you get twice as much information with half the number of images, for example, the achieved 10% sample is made up of 5% of the total images. This means the number of images sampled is minimised whilst the information obtained is maximised, making the proposed sampling strategy a practical solution.

When training the samples, the one with the least number of images has the highest accuracy measures as well as lowest training error. This sample contains the most informative data and is easier to train on. Future research should investigate the changes of the recall and precision values for specific crop types in the different samples. Other land cover data sets (used for classification purposes), not just crops, should be considered to assess the versatility of the sampling algorithm. Even though the largest considered sample has 445 images, which is almost 17% of all images, training on these images is computationally heavy, hence higher samples were not considered. In addition, the accuracy continued to decrease as the sample size increased.

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 the least number of images but a lot of information as well as the training of data for classification purposes. This paper focussed on crop classification as the application area. The methodology developed could be easily applied to any imagery with labelled areas. For example, land use data is widely available, and if combined with imagery to do classification, will encounter the same computational difficulties as this current focus.

8. Acknowledgements

This work is based on the research supported wholly in part by the National Research Foundation of South Africa (Grant Numbers: 137785) and is part of an unpublished MSc thesis (Rangongo, 2022).

References

Barriere, V., Claverie, M., Schneider, M., Lemoine, G., d’Andrimont, R., 2024. Boosting crop classification by hierarchically fusing satellite, rotational, and contextual data. *Remote Sensing of Environment*, 305, 114110.

Berrendero, J. R., Cuevas, A., Torrecilla, J. L., 2016. The mRMR variable selection method: a comparative study for functional data. *Journal of Statistical Computation and Simulation*, 86, 891–907.

Elssied, N. O. F., Ibrahim, O., Osman, A. H., 2014. A novel feature selection based on one-way anova f-test for e-mail spam classification. *Research Journal of Applied Sciences, Engineering and Technology*, 7(3), 625–638.

Gao, B.-C., 1996. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of the Environment*, 58, 257–266.

Gomes, V. C., Queiroz, G. R., Ferreira, K. R., 2020. An overview of platforms for big earth observation data management and analysis. *Remote Sensing*, 12(8), 1253.

Jiao, X., Yang, B., Pei, Z., 2006. Paddy rice area estimation using a stratified sampling method with remote sensing in China. *Transactions of the CSAE*, 22, 105–110.

Kraskov, A., Stögbauer, H., Grassberger, P., 2004. Estimating mutual information. *Physical Review E*, 69.

Lausch, A., Selsam, P., Pause, M., Bumberger, J., 2024. Monitoring vegetation-and geodiversity with remote sensing and traits. *Philosophical Transactions of the Royal Society A*, 382(2269), 20230058.

Li, H., Song, X.-P., Hansen, M. C., Becker-Reshef, I., Adusei, B., Pickering, J., Wang, L., Wang, L., Lin, Z., Zalles, V. et al., 2023. Development of a 10-m resolution maize and soybean map over China: Matching satellite-based crop classification with sample-based area estimation. *Remote Sensing of Environment*, 294, 113623.

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A., 2016. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115, 119–133.

Maqsood, M. H., Mumtaz, R., Khan, M. A., 2024. Deforestation detection and reforestation potential due to natural disasters—A case study of floods. *Remote Sensing Applications: Society and Environment*, 101188.

Miranda, C., Santesteban, L. G., Urrestarazu, J., Loidi, M., Royo, J. B., 2016. Sampling stratification using aerial imagery to estimate fruit load in peach tree orchards. *Agriculture*, 8.

Rangongo, T., 2022. Assessing classification performance for sampled remote sensing data. Master’s thesis, University of Pretoria, South Africa.

Rangongo, T., Fabris-Rotelli, I., Thiede, R., 2022. Multivariate big data sampling for crop area coverage. *Proceedings of the 63rd Annual Conference of the South African Statistical Association, 30 November – 2 December 2022, George, South Africa.*

Schulthess, U., Rodrigues, F., Taymans, M., Bellemans, N., Bontemps, S., Ortiz-Monasterio, I., Gérard, B., Defourny, P., 2023a. Optimal sample size and composition for crop classification with Sen2-Agri’s random forest classifier. *Remote Sensing*, 15(3), 608.

Schulthess, U., Rodrigues, F., Taymans, M., Bellemans, N., Bontemps, S., Ortiz-Monasterio, I., Gérard, B., Defourny, P., 2023b. Optimal Sample Size and Composition for Crop Classification with Sen2-Agri’s Random Forest Classifier. *Remote Sensing*, 15(3).

Su, T., Zhang, S., 2021. Object-based crop classification in Hetao plain using random forest. *Earth Sciene Informatics*, 14, 119–131.

Tariq, A., Yan, J., Gagnon, A. S., Riaz Khan, M., Mumtaz, F., 2023. Mapping of cropland, cropping patterns and crop types by combining optical remote sensing images with decision tree classifier and random forest. *Geo-Spatial Information Science*, 26(3), 302–320.

Zaitunah, A., Ahmad, A., Safitri, R., 2018. Normalized difference vegetation index (ndvi) analysis for land cover types using landsat 8 oli in besitang watershed, indonesia. *Proceedings of the IOP Conference Series: Earth and Environmental Science*.

Zhu, S., Zhang, J., 2013. Provincial agricultural stratification method for crop area estimation by remote sensing. *Transactions of the Chinese Society of Agricultural Engineering*, 29, 184–191.