Mapping rare and infrequent crops from space

François Waldner[1], Yang Chen[2], Roger Lawes[2], Zvi Hochman[1]

[1] CSIRO Agriculture and Food, 306 Carmody Rd, St Lucia QLD 4067, Australia; franz.waldner@csiro.au
[2] CSIRO Agriculture and Food, Underwood Ave, Floreat WA 6014, Australia; y.chen@csiro.au

Abstract
Most cropping systems around the world are organised around few dominant crops and a larger number of less frequent crops. Data about the location of infrequent crops derived from satellite data are generally inaccurate, largely owing to the class imbalance problem. Class imbalance occurs when only few instances of some classes are available for classifier training and leads to large classification errors of the infrequent classes. Here, we assessed the magnitude of the class imbalance problem in crop classification and evaluated data-level treatments to combat it by creating synthetic minority instances. We generated 18 unbalanced data sets from Sentinel-2 time series and crop type observations in Victoria, Australia. These data sets covered a wide range of complexity, number of classes, number of samples per class and spectral separability. Classification accuracy was assessed with two metrics: the Overall Accuracy (OA), which gives more weight to majority classes, and the G-Mean accuracy (GM), which is more sensitive to minority classes. We found that data-level treatments boosted GM by 0.1-0.35 and that the price for increasing the accuracy of minority classes is a drop in OA. While oversampling methods have clear potential to improve the classification of minority crop types, more control over the loss of overall accuracy needs to be gained before transitioning these methods to operations.

Key Words
Crop mapping; classification; satellite imagery; minority crops

Introduction
Compared with traditional ground-based methods, satellite remote sensing provides synoptic and repetitive observations that are well suited to monitor large agricultural areas. Over the last decades, it has thus become a prime source of data to map, monitor, and forecast crop production at a range of scales.

Classification of satellite image time series is one of the main objectives of monitoring agriculture from space as it is required to estimate crop area and yield. Improving the accuracy of classification methods to meet users' requirements has been the prime focus of decades of research. Nonetheless, evidence from the remote sensing literature shows that one can confidently identify the main crops in a region of interest with rates of success depending on data availability, and quality, landscape complexity and image spatial resolution. However rare and infrequent crop classes seem to be consistently mapped with poor accuracy. Misclassification of minority classes is often referred to as the class imbalance problem which occurs when there are many more observations of some classes than others. Under class imbalance, most conventional machine learning classifiers tend to be biased towards the prevalent classes and ignore the infrequent ones because they were primarily designed for problems with balanced class distributions. In this study, we designed a simple experimental design to quantify the impact of class imbalance and to evaluate and compare two oversampling methods to combat it.

Methods
Study area and data
The study area covers much of the Wimmera, Mallee, and Central regions of Victoria, Australia. Crop identification was carried out during two roadside surveys in September and October 2017. Field boundaries associated with each point-based label were subsequently drawn using within season Sentinel-2 and Google Earth imagery. As a result, >2.9 million pixels from 459 fields covering nine crop classes were at hand (Table 1). Four Sentinel-2 tiles covered the area of interest. Sentinel-2A and -2B imagery acquired from the 01/03/2017 to 06/12/2017 with less than 80% cloud cover were acquired. On average, 27 images were available per tile. They were corrected to surface reflectance using the multi-sensor atmospheric correction and cloud screening spectral-temporal processor.

We extracted the Green Chlorophyll Vegetation Index (GCVI; Gitelson et al., 2003) then applied harmonic regression to the GCVI time series to deal with the acquisition heterogeneity (different acquisition dates and orbits) as well as differential cloud contamination across images and tiles. Only the first five harmonics and the additive term were kept for further analysis, that is, the number of features was eleven.

Table 1. Ground truth data set for the 2017 winter season in Victoria.

Crop	# of fields	# of pixels
Barley (Ba)	85	581846
Canola (Ca)	87	542770
Chickpea (Ch)	14	100013
Faba bean (Fa)	14	67822
Field pea (Fi)	21	67822
Lentil (Le)	17	536464
Oat (Oa)	67	52087
Vetch (Ve)	13	137779
Wheat (Wh)	137	791278

We generated 18 data sets from the field data so as to cover a variety of classes, class distribution, and data complexity (Table 1). Each landscape was characterised by a crop type sequence summarising its constituting classes and associated data distribution. We then assigned a number of instances per class which we then drew from the full data set. There were four possible quantities of instance per class: low ($_l$; 250), intermediate ($_i$; 1250), high ($_h$; 2500), and very high ($_v$; 5000).

Table 1. Description of the 18 data sets.

id	# of classes	*Class allocation*	*# instances*	id	# of classes	*Class allocation*	*# instances*
1	3	$Wh_lCa_lLe_h$	3000	10	5	$Wh_vBa_iFi_lLe_hCa_i$	10250
2	3	$Wh_vCa_lLe_l$	5500	11	5	$Wh_hBa_iVe_lLe_lFi_h$	6750
3	3	$Ba_lLe_vFi_l$	5500	12	6	$Wh_lBa_hVe_iLe_lCh_iCa_h$	8000
4	3	$Ch_iVe_lLe_h$	4000	13	6	$Ba_iWh_lLe_hCh_iCa_lFi_h$	8000
5	3	$Wh_hCa_lBa_l$	3000	14	7	$Wh_vBa_lFi_hLe_iCh_iVe_lCa_i$	11750
6	3	$Wh_lBa_vFi_i$	6500	15	7	$Wh_lBa_iFi_hLe_iCh_iVe_hCa_l$	9250
7	4	$Ve_hCh_lLe_lCa_l$	3250	16	7	$Wh_hBa_iFi_iLe_lCh_iVe_iCa_l$	8000
8	4	$Wh_hBa_lLe_iCa_l$	4250	17	8	$Wh_iBa_hOa_hFi_hLe_iCh_lVe_hCa_l$	13000
9	4	$Wh_iBa_lFi_hLe_l$	4250	18	8	$Wh_lBa_hOa_hFi_lLe_lCh_hVe_hCa_i$	12000

Data-level treatment and classification

There are three types of resampling techniques: i) oversampling methods create a superset of the original data set either by replicating instances or by generating new instances of minority classes, ii) undersampling methods create a subset of the original data set by removing some of instances of the majority classes and iii) hybrid methods combine the two previous approaches in either order. Several studies have illustrated that oversampling approaches outperform undersampling ones, which is why we focused on assessing the former. In particular, we compared Random Oversampling (ROS) and the Synthetic Minority Oversampling Technique (SMOTE). ROS is probably the most straightforward oversampling approach (Chawla, 2009) which increases the balance between classes by random replications of the minority instances. It increases the risk of overfitting because it creates duplicates. In contrast, SMOTE instances of minority class by forming convex combinations of neighbouring instances (Chawla et al., 2002). As SMOTE creates new instances overfitting is unlikely. With oversampling methods, the number of new instances to create needs to be defined beforehand. In this paper, we used full oversampling as balancing rule, *i.e.*, perfect balance among classes was achieved.

Data sets were classified using Support Vector Machine (SVM) classifiers. SVMs rely on the notion of separating classes in a higher dimensional feature space, which is created using a kernel function. Optimal separating hyperplanes are fitted between two classes in the feature space focusing on data points lying at the edges of the class distributions (Vapnik, 2000). The kernel parameter (γ) and the regularisation parameter (C)

must be defined prior to calibration. Here, C was set to ten and γ was automatically optimised using heuristics. All features were normalised before classification.

Accuracy assessment

Map accuracy is generally assessed with the overall accuracy (OA), which expresses the proportion of correctly classified pixels in the map. By definition, OA has an emphasis on the dominant classes rather than on the rare ones. For instance, in a binary problem where the majority class represents 99% of the data, a classifier that assigns the majority class to all test case already achieves 99% of accuracy, which is misleading. As a result, additional metrics have started to come into widespread use such as the G-Mean accuracy (GM) that provides a more precise evaluation of the classifier performance for unbalanced learning (He et al., 2009). GM is mathematically defined as the geometric mean of the Producer Accuracy (PA) which indicates the probability that a value in a given class was classified correctly. As each PA value representing the classification performance of a specific class is equally accounted for, GM is indicative of the balanced performance of a learning algorithm among all classes. Relying on these two metrics provide a synoptic assessment of the accuracy of both majority and minority classes.

For each landscape and oversampling method, OA and GM were estimated by means of 10-fold cross validation. The folds were defined randomly so as to ensure that all pixels from the same field belonged to the same fold, thereby guaranteeing true independence of the validation set.

Results

In the unbalanced case, OA values varied from 0.52 to 0.97 (Figure 1). The highest accuracy was achieved for data set # 2 (OA = 0.95; $Wh_vCa_lLe_l$) whereas the lowest accuracy was observed for data set # 18 (OA = 0.52; $Wh_iBa_hOa_hFi_lLe_lCh_hVe_hCa_i$). There was a strong negative correlation between OA and the number of classes (Pearson's R = -0.89). Due to class imbalance, GM values were markedly lower than OA values. The first two data sets ($Wh_lCa_lLe_h$, $Wh_vCa_lLe_l$) and the seventh ($Wh_lBa_iFi_hLe_iCh_iVe_hCa_l$) achieved accuracy that were remarkably higher than the remaining cases which oscillate around 0.45.

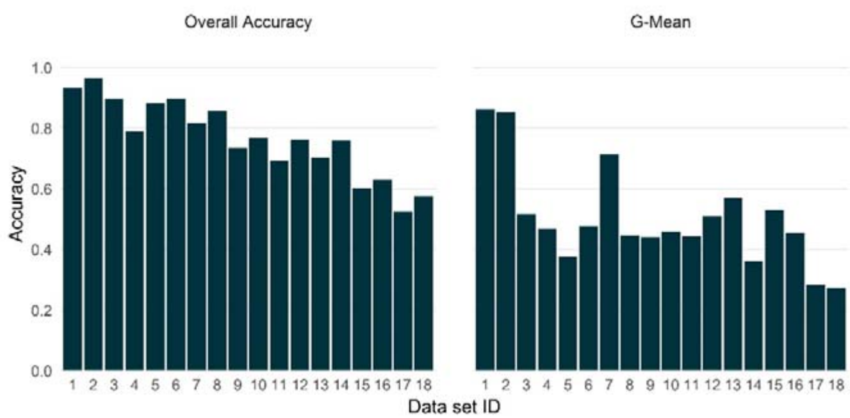

Figure 1. Accuracy measures for all data. The reader is referred to Table 2 for details about the IDs.

We then applied ROS and SMOTE to the 18 data sets and followed a fully balanced approach to restore the balance among classes prior to classification. Overall, the response to data-level treatments was highly dependent upon data sets (Figure 2). The maximum increase in GM reached 0.23 ($Wh_lBa_vFi_i$) and 0.34 ($Wh_hCa_lBa_l$) while the mean gain was 0.04 and 0.08 for RF and SVM, respectively. The average loss in overall accuracy was equal to 0.03 and the maximum losses reached 0.17 ($Wh_lBa_vFi_i$). In some cases, the loss in overall accuracy was larger than the gain in G-Mean accuracy. Overall, the correlation between oversampling methods was 0.98.

Conclusion

The intent of this study was to demonstrate how oversampling can help combat the class imbalance problem that occurs when classifying crop types using satellite imagery and machine learning. Leveraging dense time series of Sentinel-2 images and ground truth data in Victoria, we generated a set of landscapes spanning a wide range of classes, class imbalance and complexity. We then classified these data sets using support vector machines. These classifications reached overall accuracies of up to 0.92 and G-Mean accuracies of

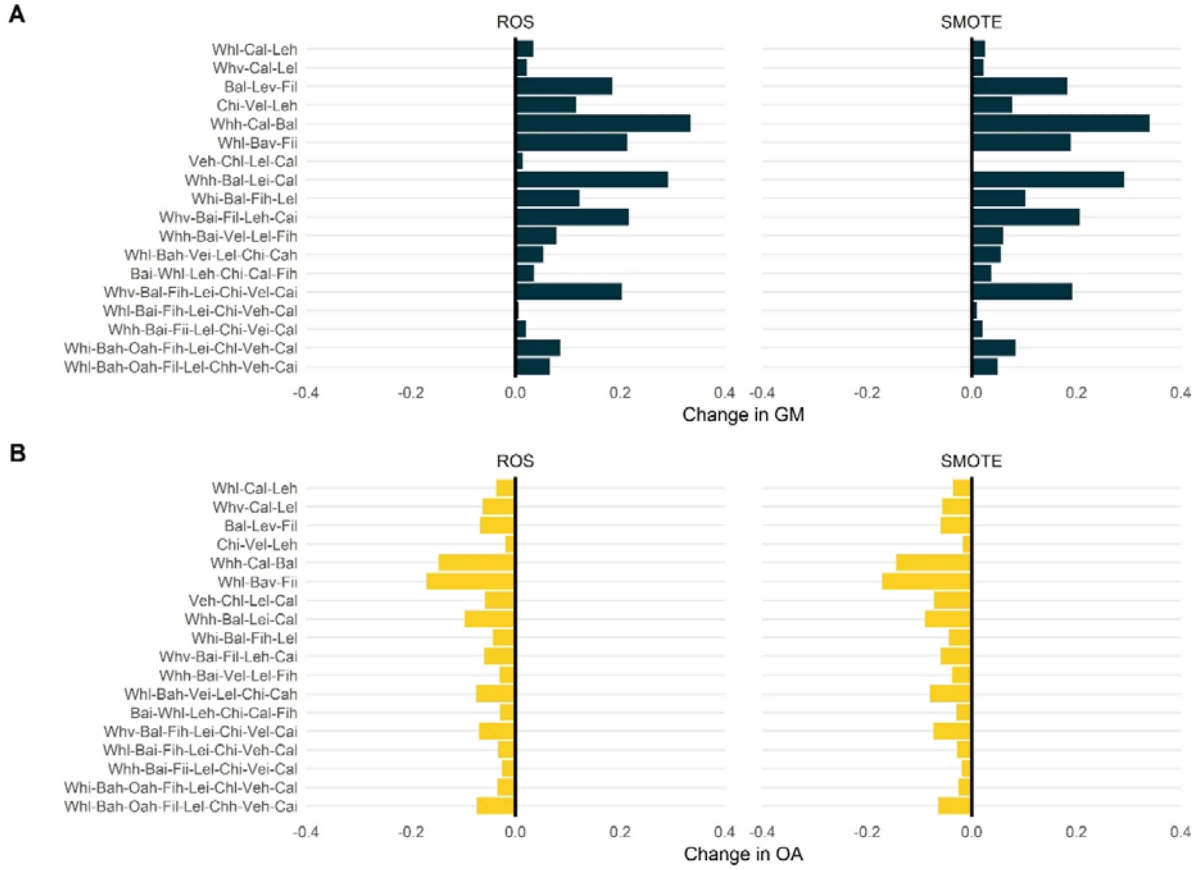

Figure 2.Changes in G-Mean and overall accuracy before and after oversampling

around 0.45. Oversampling methods helped improve the accuracy of minority classes as shown by increases in G-Mean accuracy of up to 0.35. The associated cost was a drop in overall accuracy of up to 0.17. In some cases, the drop in overall accuracy seemed to exceed the gain in G-Mean accuracy. Therefore, we conclude that while oversampling techniques are promising to combat class imbalance, future work should propose a framework to maximise the accuracy of the minority classes and help control the drop in overall accuracy.

References

Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.

Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. In Data mining and knowledge discovery handbook. Springer, Boston, MA.

Chawla, N.V., 2009. Data mining for imbalanced datasets: An overview. In Data mining and knowledge discovery handbook (pp. 875-886). Springer, Boston, MA.

Gitelson, A.A., Viña, A., Arkebauer, T.J., Rundquist, D.C., Keydan, G. and Leavitt, B., 2003. Remote estimation of leaf area index and green leaf biomass in maize canopies. Geophysical Research Letters, 30(5).

Gislason, P.O., Benediktsson, J.A. and Sveinsson, J.R., 2006. Random forests for land cover classification. Pattern Recognition Letters, 27(4), pp.294-300.

He, H. and Garcia, E.A., 2008. Learning from imbalanced data. IEEE Transactions on Knowledge & Data Engineering, (9), pp.1263-1284.

Vapnik, V., 2000. Statistics for engineering and information science. The Nature of Statistical Learning Theory.