To appear in *Remote Sensing Letters*
Vol. 00, No. 00, Month 20XX, 1–10

RESEARCH ARTICLE

The impact of training class proportions on binary cropland classification

François Waldner[a,*], Damien C. Jacques[a] and Fabian Löw[b]

[a] *Université catholique de Louvain, Earth and Life Institute - Environmental Sciences, Croix du Sud, Louvain-la-Neuve, Belgium*; [b] *MapTailor Geospatial Consulting GbR, Bonn, Germany*

(*Received 00 Month 20XX; accepted 00 Month 20XX*)

The collection of ground truth data required for training supervised classifiers is usually carried out as to maximize the number of samples under time, budget and accessibility constraints. Yet, the performance of machine learning classifiers is, among other factors, sensitive to the class proportions in the training set. In this letter, the joint effect of the number of calibration samples and the class proportion on the accuracy was systematically quantified using two state-of-the-art machine learning classifiers (random forests and support vector machines) in the context of binary cropland classification. The analysis focused on two contrasted agricultural landscapes. Results showed that the classifiers were more sensitive to class proportions than to sample size, though sample size had to reach 2,000 pixels before its effect leveled off. Optimal accuracies were obtained when the training class proportions were close to those actually observed on the ground. Then, synthetic minority over-sampling technique (SMOTE) was implemented to artificially regenerate the native class proportions in the training set. This resampling method led to an increase of the accuracy of up to 30%. These results have direct implications for (i) informing data collection strategies and (ii) optimizing classification accuracy. Though derived for cropland mapping, the recommendations are generic to the problem of binary classification.

Keywords: binary classification, accuracy, class proportions, cropland mapping, synthetic minority over-sampling technique

1. Introduction

Supervised image classification is a widely used technique for the extraction of land cover information from remotely sensed data. Usability and reliability of a map generated by a supervised classification depends on its accuracy. Consequently, much research has been devoted to solve the issues that prevent the increase in classification accuracy in order to yield optimal or near-optimal outputs. Yet, achieving optimal classifications remains challenging for several reasons. First, different classifiers trained with the same calibration data often yield dissimilar outputs as a result of how they use the training data and how they partition the feature space (Huang, Davis, and Townshend 2002; Foody and Mathur 2004a). However, to a large extent, accuracy also depends on the intrinsic characteristics of the calibration data sets. How a classifier is trained can have a larger impact on accuracy than the

*Corresponding author. Email: francois.waldner@uclouvain.be

classification technique itself. This situation has prompted research on the design of the training stage such as sampling design (Chen and Stow 2002), size of the training set (Foody and Arora 1997; Foody and Mathur 2004a), composition of the training set (Millard and Richardson 2015), spacing of training samples (Atkinson 1991). The size of the calibration set, *i.e.*, the number of samples available for training, has frequently attracted attention because of the costs involved in data collection. In general, studies have shown that a large number of training samples is beneficial (Pal and Mather 2004; Foody and Mathur 2004b) because it provides a more complete representation of the class populations. However, small training samples are attractive for practical reasons (Li et al. 2014). Numerous recommendations were made on the optimal size of training sets (see Mather and Koch (2011); Van Niel, McVicar, and Datt (2005) for instance). Given the costs of collecting in situ calibration data, budgeted sampling approaches are often preferred. For cropland mapping particularly, the crop-related samples are typically collected during field surveys while non-crop samples are digitized on screen based on very high resolution imagery. Further efforts to reduce the burden of ground truth data collection investigated the use of existing land cover maps as a source of calibration data (Waldner, Canto, and Defourny 2015). Conventional data collection schemes from either sources have generally dedicated more attention to the sample size than to the respective class proportions in the sample. Even with the large adoption of machine learning algorithms, which are usually more adequate to handle high-dimensional and multi-source data sets, the question of the appropriate distribution to optimize a learning algorithm remains compelling. This issue has been intensively investigated for imbalance learning, a problem that arises when one class has much more samples than the others. Imbalance heavily compromises the process of learning, because machine learning models tend to focus on the prevalent class and to ignore rare classes (Japkowicz and Stephen 2002; Visa and Ralescu 2005). In addition, all classifiers generally present some performance loss when the data is unbalanced, albeit this behavior might different among classifier algorithms (Prati, Batista, and Silva 2015).

In this letter, the joint effect of the number of samples and the class proportions on accuracy was investigated in the context of binary cropland classification. To that aim, two specific research questions were addressed:

(1) What is the magnitude of the gain/loss in classification accuracy due to a change in class proportions and number of samples?
(2) Can resampling strategies and *a priori* information on class prevalence be combined to optimize the performance of a classifier?

The working hypotheses were that class proportions can significantly affect the accuracy of a classifier and that *a priori* information of the actual class distribution could be used to artificially modify them to optimize the classification accuracy. Throughout this article, optimal is intended in the sense of achieving the maximum possible accuracy with a given calibration data set. These two questions were systematically investigated over two contrasted agricultural landscapes and for two state-of-the-art machine learning classifiers, namely random forest (RF) and supports vector machines (SVM).

2. Data and study sites

This study focuses on two sites located in Belgium (top left: 51.00° N, 4.50° E and bottom right: 49.60° N, 5.80° E) and South Africa (top left: 26.85° S, 24.55° E and bottom right: 30.74° S, 29.77° E). They each cover an area of 60 km x 60 km and are dominated by cropland (see Bontemps et al. (2015) for an assessment of those data sets). For both sites, a 20-m multi-sensor time series consisting of SPOT-4 (Take 5) data and Landsat-8 data spanning from February to December 2013 were at hand. For Landsat, only reflectances in the green, red, near infra-red and short wave infrared wavelengths were considered. The SPOT-4 and Landsat-8 images were processed with the Multi-sensor Atmospheric Correction and Cloud Screening processor (Hagolle et al. 2015). The time series were gap filled using linear interpolation and three spectral-temporal features were extracted. These spectral-temporal features correspond to reflectance composites at the minimum and the maximum of normalized difference vegetation index as well as the mean reflectance over the season. Such features were shown to provide high discrimination power between cropland and non-cropland areas (Waldner, Canto, and Defourny 2015; Matton et al. 2015).

Ground truth observations from the corresponding growing season supplemented the satellite image time series. In Belgium, field polygons were sourced from the Land Parcel Identification System and non-cropland polygons were digitized based on very high resolution imagery. In South Africa, both cropland and non-cropland objects were both digitized based on very high resolution data. These reference data sets were evenly split at the polygon level into two independent sets to be used for calibration and validation, respectively. The calibration data sets contained between 81,000 and 85,000 pixels for Belgium and South Africa, in proportions corresponding to the actual cropland proportions (Table 1). Native cropland proportions were derived from vectorized parcel data. Hereafter, the term “prevalence” refers to cropland proportion as the native cropland proportion was dominant.

Table 1.: Characteristics of the calibration and validation data sets for the two study sites.

Site	Prevalence (%)	Size of training set (pixels)	Size of validation set (pixels)
Belgium	62	85,342	88,788
South Africa	75	81,250	99,213

3. Methodology

3.1. *Classification methods*

SVMs are based on a binary classifier concept and 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 those training samples that lie at the edge of the class distributions (Vapnik 2000). Training of SVMs includes choosing the kernel parameter γ and the regularization parameter C. C was set to ten and the widths of the kernels were defined using heuristics. The RF classifier is an ensemble of decision trees which are trained based on random bootstrapped samples of the training data (Breiman 2001). Two meta parameters must be defined: the number of trees and the number of features used to split the nodes in the trees. The number of trees was set to 500 and the number of features was set to the square root of the total number of input features (Gislason, Benediktsson, and Sveinsson 2006).

3.2. *Monte Carlo Assessment Procedure*

The RF and SVM models were calibrated with different class prevalences ranging from 5 to 95% of the full training set and different sample sizes. Ten sample sizes were tested: 200, 375, 500, 750, 1,000, 1,500, 2,000, 3,000, 4,000, 5,000; similarly twelve cropland proportions were considered: 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95% plus the native prevalence. A Monte Carlo procedure was implemented to derive robust estimates of the model performances. 25 iterations were executed per combination of prevalence and sample size. The model performance measures were averaged over the 25 model runs. In total, the Monte Carlo procedure consisted of 2,700 iterations per site. Within an iteration, the calibration set was randomly sub-sampled to generate a new calibration subset in agreement with the sample size and prevalence values for that iteration. Then, the SVM and RF models were trained with the training subset. Finally, the output of each model was run against the same validation set to evaluate the model performance. Three accuracy measures were evaluated: the overall accuracy as well as the F-score for the cropland and the non-cropland classes. A locally weighted regression interpolation (Cleveland and Devlin 1988) was implemented to interpolate the accuracy measures.

3.3. *Resampling with synthetic minority over-sampling*

In the context of supervised classification, imbalanced calibration samples are often handled by over- and under-sampling to achieve a more balanced calibration data set. In this study, a synthetic minority over-sampling technique (SMOTE; Chawla et al. (2002)) was applied to generate synthetic calibration samples for the minority class so that the training class proportion respect their native occurrence. SMOTE's core idea is to artificially generate new samples of the minority class using bootstrapping and k-nearest neighbors. As a hybrid method, SMOTE features both oversampling of the minority class and undersampling of the majority class. To create a synthetic sample, one randomly chosen minority class sample as well as one of its randomly chosen next neighbors were interpolated, so that finally a new artificial sample of the minority class is created. In addition, the dominant class samples were randomly undersampled (only 90% of the samples belonging to the dominant class were kept). The performance of resampling the calibration data set to the native cropland proportion with SMOTE were evaluated following the Monte Carlo assessment procedure detailed previously. Paired t-tests were performed to assess if the differences in accuracy were significant at the 5% level.

4. Results

4.1. *The respective impact of the training set size and the class prevalence*

The performance of the RF and SVM classifiers was systematically tracked for different sample sizes and prevalences (Figure 1; left-hand side column). The y-axis is the proportion of cropland samples in the training data; the x-axis gives the accuracy of the model on the test set, averaged over 25 random draws of the training set. The black horizontal line indicates the native class prevalence.

The classification accuracies exhibited different patterns depending on the study site but similar trends were obtained for both algorithms. Overall, SVMs yielded consistently slightly higher accuracies than RFs. In Belgium (Figure 1a), the overall accuracy and the F-scores were generally high (>0.9) with exceptions for large (>0.75) and low (<0.25) cropland prevalences. The highest accuracies were reached when the native cropland prevalence was used. In South Africa (Figure 1c), the overall accuracy increased with the cropland prevalence from 0.55 to 0.65. Highest F-scores for the non-cropland class were clearly obtained for near-native class prevalences. The results demonstrate that the overall accuracy was mostly driven by the accuracy of the dominant class (the cropland class). Optimizing the accuracy of the classification depends on which measure one wants to optimize because different accuracy measures do not necessarily behave similarly (see the F-scores in South Africa for instance). In the present case, the goal was to identify the rare class, *i.e.*, non-cropland. The lowest accuracies were obtained for a combination of extreme prevalences and very low sample sizes. Highest accuracies were found when maintaining approximately the native class prevalences. Equal class prevalences always showed sub-optimal accuracies compared to the native prevalences case. Results also showed a stronger sensitivity to class prevalence than to the size of the training data set. This effect is further illustrated in Figure 2 (red line) which charts the evolution of accuracy observed at a training set prevalence of 25% for each sample size. It highlights that once a certain sample size is reached (2,000-2,500) the impact of adding more the calibration samples levels off.

4.2. *The impact of artificially re-generating the native prevalence*

To maximize the classification accuracy, two main observations can be derived from Figures 1a and 1c. First, balancing class prevalence before training (50/50) does not systematically improve classification accuracy. It will do so only if it brings the sample class prevalence closer to the native class prevalence. Second, one can achieve a higher accuracy with fewer samples if their prevalence is closer to the native prevalence. Figures 1b and 1d illustrate the effect of artificially recreating the native prevalence within the calibration data set before training the classifier. The region outside the black lines and the grey lines correspond to statistically significant improvements as measured by the paired t-tests at the 0.95 confidence level.

Overall, SMOTEing the calibration data sets had a leveling-up effect across all accuracy measures for both sites and algorithms. Besides, it narrowed the spread of accuracies while at the same time increasing it overall. Still, the lowest accuracies were obtained for extreme prevalences and low sample sizes. It should be mentioned that SMOTE artificially modifies the size of the sample to reach the correct

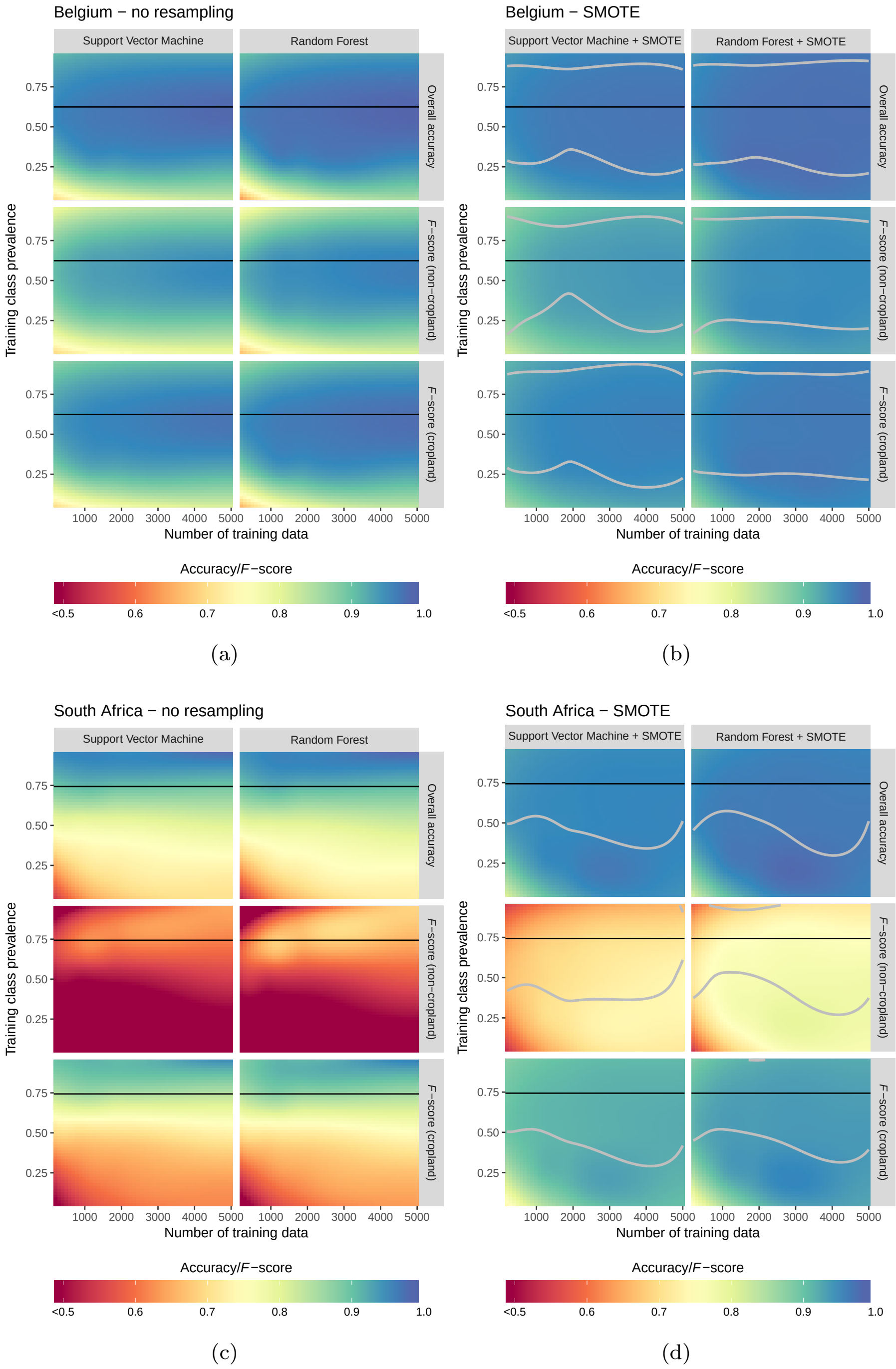
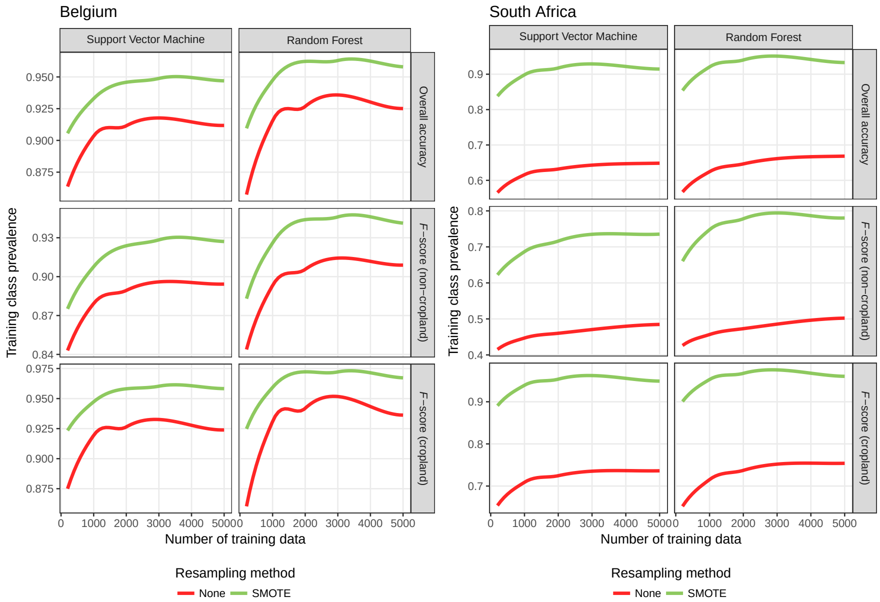


Figure 1.: Evolution of the accuracy measures as a results of varying sample sizes and the class prevalences. The top row refers to the Belgian site and the bottom row to the South African site. Figures on the right-hand side integrate a SMOTE resampling whereas figures on the left-hand side do not.

prevalence. The assessment of the impact of size and proportion might therefore be blurred and more difficult to isolate. The magnitude of the effect of SMOTE was site-specific (Figures 1a and 1c). For Belgium, where the average accuracy was already high, a limited increase was observed (1-5%). For prevalences close to the native



(a) (b)

Figure 2.: Evolution of the classification accuracy for a cropland class prevalence of 25% and for different calibration set sizes. Accuracy reaches a plateau at around 2,000-3,000 samples while the use of SMOTE induces a systematic shift towards higher accuracies. The magnitude of the shift is 3-5% in Belgium and 20-30% in South Africa.

one, a rather small decrease in accuracy was measures (3% maximum) and it was not statistically significant. However, for a low cropland prevalence, the accuracy increased up to 15% after applying SMOTE which was a statistically significant. In South Africa, the benefit of SMOTE was even more striking, with significant improvements in accuracy of up to 30%. The effect of SMOTE is also highlighted in Figure 2 for a prevalence of 25%, *i.e.*, far from the native prevalence of both sites. At this prevalence, differences between the SMOTE (green lines) and the non-SMOTE (red lines) approaches are statistically significant for both sites. Both approaches have a similar shape with respect to the site and the accuracy measure but SMOTE consistently shifted the performance of the classifiers towards higher accuracies. This increase was of about 3-5% for Belgium and 20-30% for South Africa. It highlights the interest of SMOTE for very unbalanced calibration data set.

5. Discussion

Results confirmed that training data have a more pronounced impact on accuracy than the choice of the classifier algorithm (Foody and Arora 1997) and that this magnitude and patterns of the response are agrosystem-specific (Waldner et al. 2016). Furthermore, the results are also in line with previous research recommending that the class proportions of the training data should be representative of their actual proportions in the landscape (Millard and Richardson 2015; Zhu et al. 2016). However analyzing the effect of class proportions for multi-class classifications remained generally bounded to two cases (native and equalized proportions) due to the difficulty to draw some general conclusions for other class distribution. The binary classification approach allowed quantifying the evolution of accuracy due to changes in the training class prevalence in a systematic fashion. Algorithms were found more sensitive to the class prevalence than to the sample size, especially when a minimum number of samples is provided (∼2,000). *A priori* knowledge on the native class proportions could therefore be used to inform both data collection and resampling strategies which in turn will optimize the classification accuracy. Conventional field data collection for crop mapping implies that entire fields are sampled so that the proportion of field belonging to each class is representative of the actual prevalence but not necessary the number of pixels. Caution should therefore be exercised before training a classifier such as RF or SVM on the whole sample set regardless of the pixel class proportions.

A resampling algorithm (SMOTE) that adjusts the training sample proportions was successfully implemented to optimize accuracy. Such resampling techniques could prove instrumental to meet the standards of accuracy associated with reliable area estimation (Waldner and Defourny 2017). Whilst simple over-sampling generally produced unwanted effects such as over-fitting and time overhead, under-sampling resulted in information loss as potentially valuable data points might be discarded. SMOTE addresses these limitations by changing the distribution of the data by interpolation method while increasing the number of minority class samples. It adopts artificially generated examples rather than randomly copied examples. Thus SMOTE can avoid the problem of over-fitting, but may also introduce noise. Results obtained here apply for the case of limited native class imbalance ratios (1:4 and 1:3) and might differ in the case of stronger imbalance, *e.g.*, 1:1,000. It should be noted that the original design of the SMOTE algorithm was to boost an extremely rare class and consequently modify the native class proportion. Its implementation here differs thus from the classical implementation in class-imbalanced classifiers (Lusa et al. 2010). Future research could extend this approach to a larger diversity of agrosystems, classifiers and benchmark resampling algorithms.

6. Conclusion

The main objective of this letter was to quantify the combined effect of the class proportions and size of the calibration data on classification accuracy as well as to assess the potential of a straight-forward resampling strategy (SMOTE) to compensate for potential negative effects of these two issues on the classification accuracy. Based on the results, we found that the effect of class prevalence, *i.e.*, the class proportion of the calibration samples had a much stronger impact on classification accuracy than the total number of calibration samples when using machine learning algorithms for solving a binary classification problem (here: cropland classification). The results of this research further suggest that the class proportion of a calibration data set can be advantageously adjusted using a resampling strategy like SMOTE. This offers some positive prospects in situations when a sampling scheme considering the actual class prevalence cannot be realized due to, for instance, financial or access limitations in the study area. Further, it allows to consistently increase the accuracy of a RF or SVM classifier, especially when the calibration sampling prevalence is strongly deviating from the actual one. The magnitude of the increase of accuracy when using SMOTE and imbalanced calibration samples was as high as 30% in some cases. The findings of this study allow formulating a set of general recommendations for performing an efficient calibration sampling in the case of binary cropland classification:

- Ideally, collection of calibration samples should be carried out in a statistical framework that allows a robust estimation of the real class proportions. In the case of binary cropland classification, it implies to collect samples from both cropland and non-cropland areas. In particular, balancing class proportions equally by default is contraindicated for both RFs and SVMs.
- If the sampling units are objects and the calibration is pixel-based, the training sample proportions should be adapted using a resampling algorithm such as SMOTE to ensure that the pixel proportions follows the object proportions.
- If the prevalence of each class is not statistically assessed, it could be approximated thanks to existing land cover maps or statistics. Such information on class proportions could then be used for resampling the calibration data set to the native proportions.

Acknowledgments

The authors acknowledge Terence Newby for providing data over South Africa. The SPOT4 imagery was obtained under the SPOT4/Take5 program.

References

Atkinson, Peter M. 1991. "Optimal ground-based sampling for remote sensing investigations: estimating the regional meant." *International Journal of Remote Sensing* 12 (3): 559–567.

Bontemps, Sophie, Marcela Arias, Cosmin Cara, Gérard Dedieu, Eric Guzzonato, Olivier Hagolle, Jordi Inglada, et al. 2015. "Building a data set over 12 globally distributed sites to support the development of agriculture monitoring applications with Sentinel-2." *Remote Sensing* 7 (12): 16062–16090.

Breiman, Leo. 2001. "Random forests." *Machine learning* 45 (1): 5–32.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16: 321–357.

Chen, D, and Douglas Stow. 2002. "The effect of training strategies on supervised classification at different spatial resolutions." *Photogrammetric Engineering and Remote Sensing* 68 (11): 1155–1162.

Cleveland, William S, and Susan J Devlin. 1988. "Locally weighted regression: an approach to regression analysis by local fitting." *Journal of the American statistical association* 83 (403): 596–610.

Foody, GM, and MK Arora. 1997. "An evaluation of some factors affecting the accuracy of classification by an artificial neural network." *International Journal of Remote Sensing* 18 (4): 799–810.

Foody, Giles M, and Ajay Mathur. 2004a. "A relative evaluation of multiclass image classification by support vector machines." *IEEE Transactions on geoscience and remote sensing* 42 (6): 1335–1343.

Foody, Giles M, and Ajay Mathur. 2004b. "Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification." *Remote Sensing of Environment* 93 (1): 107–117.

Gislason, Pall Oskar, Jon Atli Benediktsson, and Johannes R Sveinsson. 2006. "Random forests for land cover classification." *Pattern Recognition Letters* 27 (4): 294–300.

Hagolle, Olivier, Sylvia Sylvander, Mireille Huc, Martin Claverie, Dominique Clesse, Cécile Dechoz, Vincent Lonjou, and Vincent Poulain. 2015. “SPOT-4 (Take 5): Simulation of Sentinel-2 Time Series on 45 Large Sites.” *Remote Sensing* 7 (9): 12242–12264.

Huang, C, LS Davis, and JRG Townshend. 2002. “An assessment of support vector machines for land cover classification.” *International Journal of remote sensing* 23 (4): 725–749.

Japkowicz, Nathalie, and Shaju Stephen. 2002. “The class imbalance problem: A systematic study.” *Intelligent data analysis* 6 (5): 429–449.

Li, Congcong, Jie Wang, Lei Wang, Luanyun Hu, and Peng Gong. 2014. “Comparison of classification algorithms and training sample sizes in urban land classification with Landsat thematic mapper imagery.” *Remote Sensing* 6 (2): 964–983.

Lusa, Lara, et al. 2010. “Class prediction for high-dimensional class-imbalanced data.” *BMC bioinformatics* 11 (1): 523.

Mather, Paul M, and Magaly Koch. 2011. *Computer processing of remotely-sensed images: an introduction*. John Wiley & Sons.

Matton, Nicolas, Guadalupe Sepulcre Canto, François Waldner, Silvia Valero, David Morin, Jordi Inglada, Marcela Arias, Sophie Bontemps, Benjamin Koetz, and Pierre Defourny. 2015. “An Automated Method for Annual Cropland Mapping along the Season for Various Globally-Distributed Agrosystems Using High Spatial and Temporal Resolution Time Series.” *Remote Sensing* 7 (10): 13208–13232.

Millard, Koreen, and Murray Richardson. 2015. “On the importance of training data sample selection in random forest image classification: A case study in peatland ecosystem mapping.” *Remote sensing* 7 (7): 8489–8515.

Pal, Mahesh, and Paul M Mather. 2004. “Assessment of the effectiveness of support vector machines for hyperspectral data.” *Future Generation Computer Systems* 20 (7): 1215–1225.

Prati, Ronaldo C, Gustavo EAPA Batista, and Diego F Silva. 2015. “Class imbalance revisited: a new experimental setup to assess the performance of treatment methods.” *Knowledge and Information Systems* 45 (1): 247–270.

Van Niel, Thomas G, Tim R McVicar, and Bisun Datt. 2005. “On the relationship between training sample size and data dimensionality: Monte Carlo analysis of broadband multi-temporal classification.” *Remote Sensing of Environment* 98 (4): 468–480.

Vapnik, VN. 2000. “The nature of statistical learning theory. Statistics for engineering and information science..” .

Visa, Sofia, and Anca Ralescu. 2005. “Issues in mining imbalanced data sets-a review paper.” In *Proceedings of the sixteen midwest artificial intelligence and cognitive science conference,* Vol. 200567–73. sn.

Waldner, François, Guadalupe Sepulcre Canto, and Pierre Defourny. 2015. “Automated annual cropland mapping using knowledge-based temporal features.” *ISPRS Journal of Photogrammetry and Remote Sensing* 110: 1–13.

Waldner, François, Diego De Abelleyra, Santiago R Verón, Miao Zhang, Bingfang Wu, Dmitry Plotnikov, Sergey Bartalev, et al. 2016. “Towards a set of agrosystem-specific cropland mapping methods to address the global cropland diversity.” *International Journal of Remote Sensing* 37 (14): 3196–3231.

Waldner, François, and Pierre Defourny. 2017. “Where can pixel counting area estimates meet user-defined accuracy requirements?.” *International Journal of Applied Earth Observation and Geoinformation* 60: 1–10.

Zhu, Zhe, Alisa L Gallant, Curtis E Woodcock, Bruce Pengra, Pontus Olofsson, Thomas R Loveland, Suming Jin, Devendra Dahal, Limin Yang, and Roger F Auch. 2016. “Optimizing selection of training and auxiliary data for operational land cover classification for the LCMAP initiative.” *ISPRS Journal of Photogrammetry and Remote Sensing* 122: 206–221.