Natural Resources Canada | Ressources naturelles Canada

GEOMATICS CANADA
OPEN FILE 23

Influence of Sample Distribution and Prior Probability Adjustment on Land Cover Classification

D. Pouliot, R. Latifovic, W. Parkinson

2016

Canada

GEOMATICS CANADA
OPEN FILE 23

Influence of Sample Distribution and Prior Probability Adjustment on Land Cover Classification

D. Pouliot, R. Latifovic, W. Parkinson

Natural Resources Canada, Earth Sciences Sector, Canada Centre for Mapping and Earth Observation, Canada Centre for Remote Sensing

2016

doi:10.4095/297517

This publication is available for free download through GEOSCAN (http://geoscan.nrcan.gc.ca/).

Recommended citation
Pouliot, D., Latifovic, R., and Parkinson, W., 2016. Influence of sample distribution and prior probability adjustment on land cover classification; Geomatics Canada, Open File 23, 13 p. doi:10.4095/297517

Publications in this series have not been edited; they are released as submitted by the author.

ABSTRACT

Machine learning algorithms are widely used for remote sensing land surface characterization. Successful implementation requires a representative training sample for the domain it will applied in (i.e. area of interest or validation domain). However, accessibility and cost strongly limit the acquisition of suitable training samples for large regional applications. Further, it is often desirable to use previously developed datasets where significant resources have been invested, such as data developed from extensive field survey or high resolution remotely sensed imagery. These data often only partially represent the domain of interest and can lead to various forms of sample bias (land cover distribution or class properties). Classifier spatial extension is an extreme case, where a sample is trained from one region (i.e. sample domain) and applied in another (i.e. application domain). This approach is desirable from a cost perspective, but achieving acceptable accuracy is often difficult. In this research we investigate two approaches to account for possible differences between the sample and application domain land cover distributions. The first is an iterative resampling approach to predict the application distribution and adjust the sample distribution to match. The second is the use of prior probabilities to adjust class memberships. Results reinforce the importance of the land cover distribution on accuracy for algorithms that are designed to minimize the classification error with training data. Of the adjustment methods tested resampling was superior if the application domain distribution was well known. However, if it is not then the use of prior probabilities performed similarly overall. A generic model was developed to predict if resampling or prior adjustment should be applied to enhance accuracy.

Table of Contents

1. INTRODUCTION

Machine learning algorithms are widely used for remote sensing land surface characterization. Many machine learning algorithms are sensitive to the training sample used. Sampling can bias the classification towards the most frequent class or cause it to ignore or be insensitive to rarer classes (Chen et al., 2004). Successful implementation requires a representative training sample (Abu-Mostafa et al., 2012) for the domain it will be applied in (i.e. area of interest). However, accessibility and cost strongly limit the acquisition of suitable training samples for large regional applications. Further, it is often desirable to use previously developed datasets where significant resources have been invested, such as data developed from extensive field survey or high resolution remotely sensed imagery. These data often only partially represent the domain of interest and can lead to various forms of sample bias regarding the class properties or the sampled land cover distribution. Classifier spatial extension is an extreme case, where a sample is trained from one region (i.e. sample domain) and applied in another (i.e. application/validation domain). This approach is desirable from a cost perspective, but achieving acceptable accuracy is often difficult. Performance of temporal classifier extension for simple forest/non-forest classification has been found sufficient in Pax-Lenny et al., (2001), Woodcock et al., (2001) and for change detection in Pouliot et al., (2009). Pouliot et al. (2013) et al. show that more complex land cover models can be extended in time with additional measures designed to reduce error and ensure temporal consistency. For the spatial extension case with several land cover classes performance has generally been considered poor (Minter, 1978; Fernandes et al., 2004; Olthof et al., 2005).

The spatial extension problem can be divided into two categories of 1) within landscape classifier extension and 2) between landscape extension. Within landscape extension is the simpler situation where the class properties such as the general soil and vegetation spectral characteristics do not change significantly within the area. Thus, the class properties derived from one part of the study area are expected to be representative of other parts. If an extension problem can be considered the within landscape case then the key factors controlling success involves data error and sample-classifier dependence. Data error can cause deviation in the class properties that are not related to actual land surface change. Methods for calibration, atmospheric correction, and adjustments for other factors such as phenology due to limited sensor revisit seek to address these problems. The interdependence of the sample and classifier can lead to error if the classifier is sensitive to the sample distribution and the distribution is biased. Parametric classifiers such as maximum likelihood that are based on summary statistics of the sample data would be less sensitive to the sample distribution as long as a sufficient number of samples of each class are acquired. However, some machine learning algorithms can be very sensitive to the sample land cover distribution if they specifically seek to minimize class training error (Weiss et al., 2007). To illustrate this point consider a simple two class forest/non-forest land cover application where the sample distribution was 70% forest and 30% non-forest. If the selected data features have absolutely no discriminatory capacity to separate these classes and the algorithm seeks only to minimize class training error, it will essentially force all predictions to be forest and achieve 70% accuracy. Now consider the extension case, where this model has been extended to another region with a land cover distribution of 20% forest and 80% non-forest. It will make substantial error predicting only 30% non-forest of the actual 80%. This is of course unrealistic as the selected data features usually have some capacity to discriminate classes. However, for algorithms that specifically seek to minimize the class training error, the greater the class confusion the more likely the predictions will depend on the input land cover sample distribution. Results of this study reaffirm this deduction.

The sample land cover distribution is not only a concern for the classifier extension case. Another way to consider the sample distribution problem regards the deviation of the sample relative to a local application domain (e.g. a smaller local subset of the study area). If a sample is generated to be representative of the entire domain (i.e. study area), there is likely strong local deviation from that sample with a specific smaller sub-area to be classified. For example, a large study mostly forest covered sampled at random will have a sample distribution with a high percentage of forest. However, there may be local areas within the study area that are non-forest. Accounting for this deviation can enhance the overall classification accuracy of the study area if a sample sensitive classifier is used.

In this research we first characterize the sample distribution problem for the study based on the Random Forest algorithm (Breiman, 2001) and then investigate two approaches to account for possible differences between the sample and application domain land cover distributions. The first is an iterative resampling approach to predict the application distribution and adjust the sample distribution to match. The second is the use of prior probabilities to adjust class memberships. As a final step we seek to develop a generic model to predict the potential improvement with either adjustment method as means to determine if it should be applied.

2. METHODS AND RESULTS

2.1 Study Area and Reference Data

The study area was North Eastern Alberta in the Athabasca Oil Sands Region (AOS, Figure 1A). Four Geoeye scenes were collected and classified to 7 land cover types using a combined pixel and object based approach. The classified Geoeye scenes were upscaled to 30 m using the dominant land cover fraction within a 30 m pixel footprint. An example of the land cover for the northwest Geoeye scene is provided in Figure 1B along with the class legend. Clouds and shadows were manually delineated and removed. Independent accuracy assessment of the Geoeye scenes showed the accuracy to be greater than 85% for all scenes at the 2 m resolution (Pouliot et al., 2015).

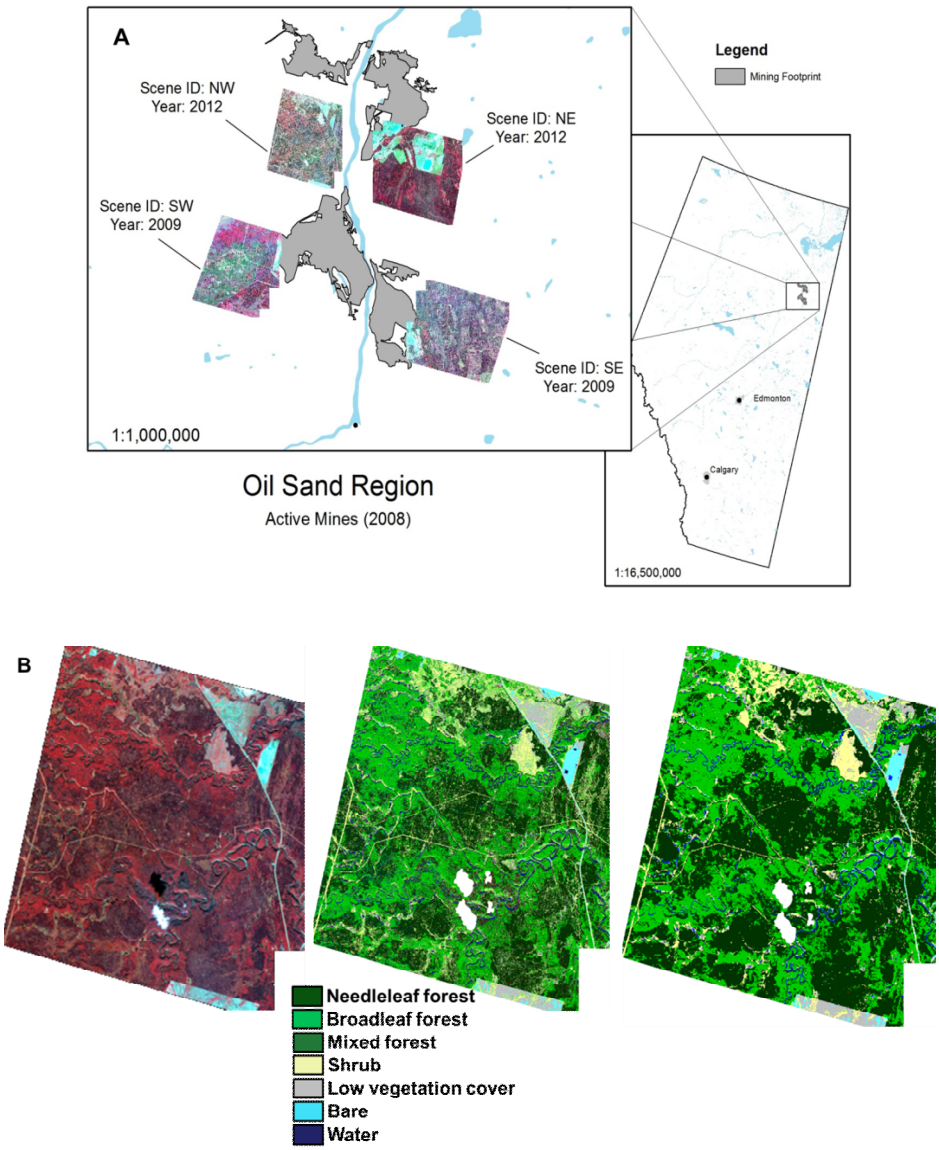

Figure 1: A) Study area and location of Geoeye scenes used to develop reference land cover. B) An example of the Geoeye land cover for the northwest scene, left is the Geoeye image at ~2 m spatial resolution, middle is the Geoeye land cover classification, and right is the upscaled land cover to 30 m spatial resolution.

2.2 Landsat Data

Landsat observations from seven scenes covering the AOS were selected for the analysis (Table 1). All spectral bands were used as well as the Normalized Difference Vegetation Index (NDVI), the Normalized Difference Moisture Index (NDMI), and a proxy for atmospheric transparency calculated as the difference between the red and blue bands, divided by the shortwave band at 1500 nm. Reprojection, scaling to top of atmosphere, and cloud and cloud shadow screening was implemented using software developed at the Canada Centre for Remote Sensing (Latifovic et al.,

2015). Processed scenes were further checked and haze or missed clouds were manually removed. Scenes were all normalized to the 2009-07-28 scene using robust regression to reduce radiometric inconsistencies (Olthof et al, 2005. Coincident Landsat and upscaled Geoeye land cover were extracted to a text file for all clear sky observations for the analysis.

Table 1: List of Landsat scenes used in the analysis.

Path	Row	Date	Sensor
41	20	2009-07-21	TM
42	20	2009-07-28	TM
42	20	2009-08-29	TM
43	20	2009-08-28	ETM+
42	20	2011-07-02	TM
42	20	2011-09-04	TM
42	20	2012-07-28	ETM+

2.3 Analysis Overview

With the four Geoeye scenes different approaches to classifier training and validation could be tested. Here a within landscape classifier extension approach was evaluated, where a classifier was trained from three of the Geoeye scenes and applied on the fourth scene. It is referred to as within landscape extension because the general soil and vegetation spectral characteristics were considered to not dramatically change within the study region.

A set of 65 land cover distributions were randomly sampled from the 2010 Alberta Biodiversity Monitoring Institute (ABMI) land cover map based on the approximate footprint of a high resolution image of 18 by 35 km. This concept is depicted in Figure 2A where the grid represents possible selections of high resolution imagery to be used for training. The implementation, however, used a random selection of scene centres instead of the grid shown. In the figure two grid squares are marked as 1 and 2 and highlight the differences in land cover distribution that can be found with the study area, where 1 is broadleaf dominated and 2 is needleleaf dominated. The set of 65 land cover distributions were used to select a training sample from three of the Geoeye scenes and used to predict the land cover of the fourth. All combinations of selecting three scenes and evaluating on the fourth were tested to provide 65×4= 260 cases for analysis.

To account for possible effects of the land cover sample distribution on model prediction two approaches were tested. The first is referred to as sample adjustment and it used an initial sample to predict the land cover for the application domain then resampled from the training data pool (3 Geoeye scenes) to match this predicted distribution in a second iteration . The second approach, used the initial land cover prediction to generate prior probabilities which were used to adjust the class memberships and the maximum membership taken as the final class. Both were tested for the classifier extension problem as detailed in Figure 2B.

To summarize, the sample distribution refers to the land cover distribution of the training sample, which is used to create a land cover classification model. This model is applied to the application domain or in this case the validation Geoeye scene. To adjust this sample to account for differences between the sample and validation land cover distributions two approaches were used: 1) resampling to match an estimate of the validation land cover distribution, or 2) applying prior probability land cover estimates.

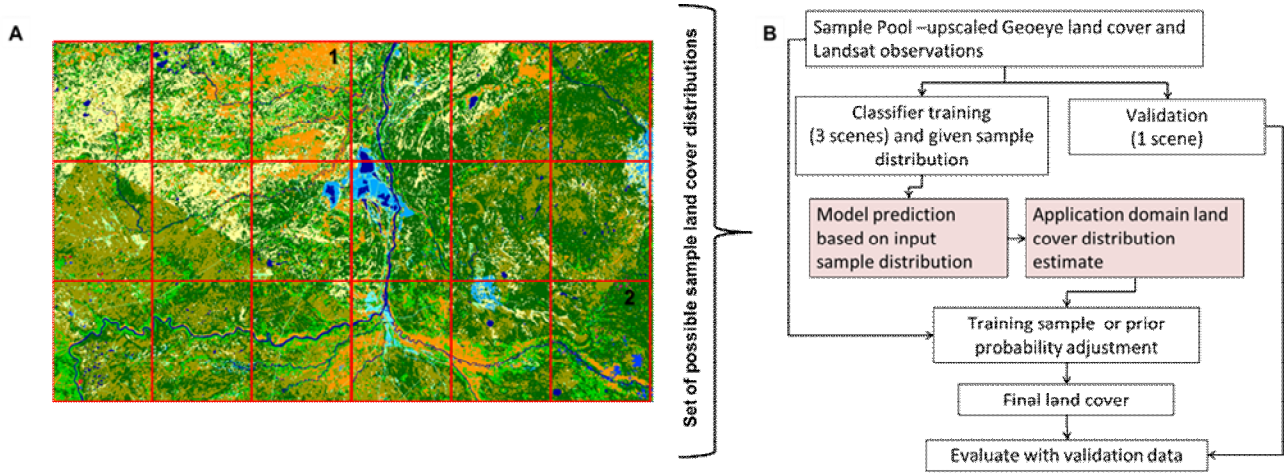

Figure 2: A) Potential land cover distributions possible within the study area based on the ABMI land cover 2010 and approximate footprint of a Geoeye scene . B) shows a flowchart of the analysis undertaken in this research to examine the effect of sample distribution on classification accuracy.

2.4 Evaluating the Effect of Sample Distribution on Land Cover Accuracy

The overall accuracy results from the sampling runs were plotted by the absolute difference between the sample land cover and application domain/validation land cover distributions in Figure 3. It reveals that the greater the difference between land cover distributions lower accuracy was achieved. The rate was specific to a given land cover classification problem (i.e. validation for NE, NW, SE, or SW). This is due to differences in the class confusion resulting from the training data and how that data is representative of the application domain. Thus, higher classification accuracy equated to lower confusion and a reduced sensitivity to the sample-application domain/validation distribution difference. For the NE case the results were not strongly dependent on the sample distribution with only a small change in accuracy for a large difference between the sample and application domain/validation land cover distributions. The SW case was more sensitive and was reduced by as much as 30%.

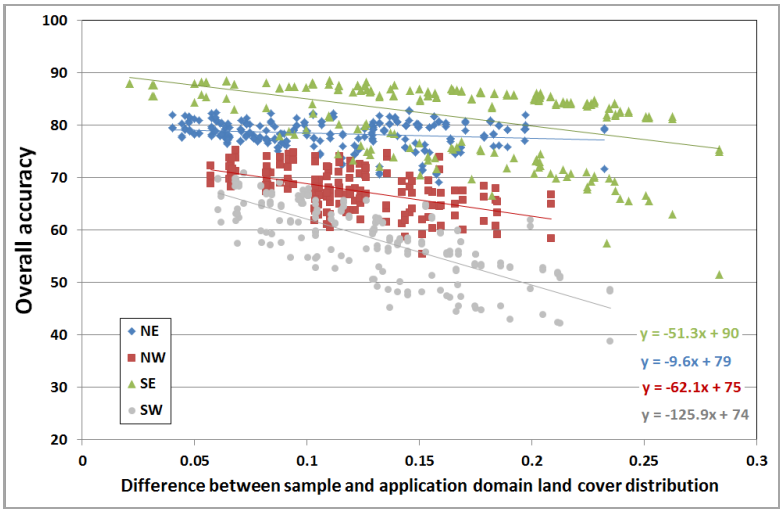

Figure 3: Relation between overall accuracy and absolute difference in sample and application (validation) land cover distributions.

2.5 Sample/Prior Probability Adjustment to Account for Sample Distribution Effects

To quantify the effect of the improvement with either sample or prior probability adjustment the overall accuracy result of either adjustment was subtracted from the pre-adjustment results. This was done for the case where the validation land cover distribution was known (directly from the Geoeye classification) or estimated based on the initial sample and classification. Results show that if the land cover distribution of the application domain is known then the sample adjustment method was found to perform better than the use of prior probabilities (Figure 4A). The x-axis in the figure is grouped by the validation dataset (NE, NW, SE, or SW). If the application distribution is not known and has to be estimated based on the initial sample, the performance between sample and prior probability adjustment are overall similar (Figure 4B). In both cases, the potential for a small reduction in accuracy was observed. This finding warrants caution in the application of either adjustment approach if the application land cover distribution is not well specified.

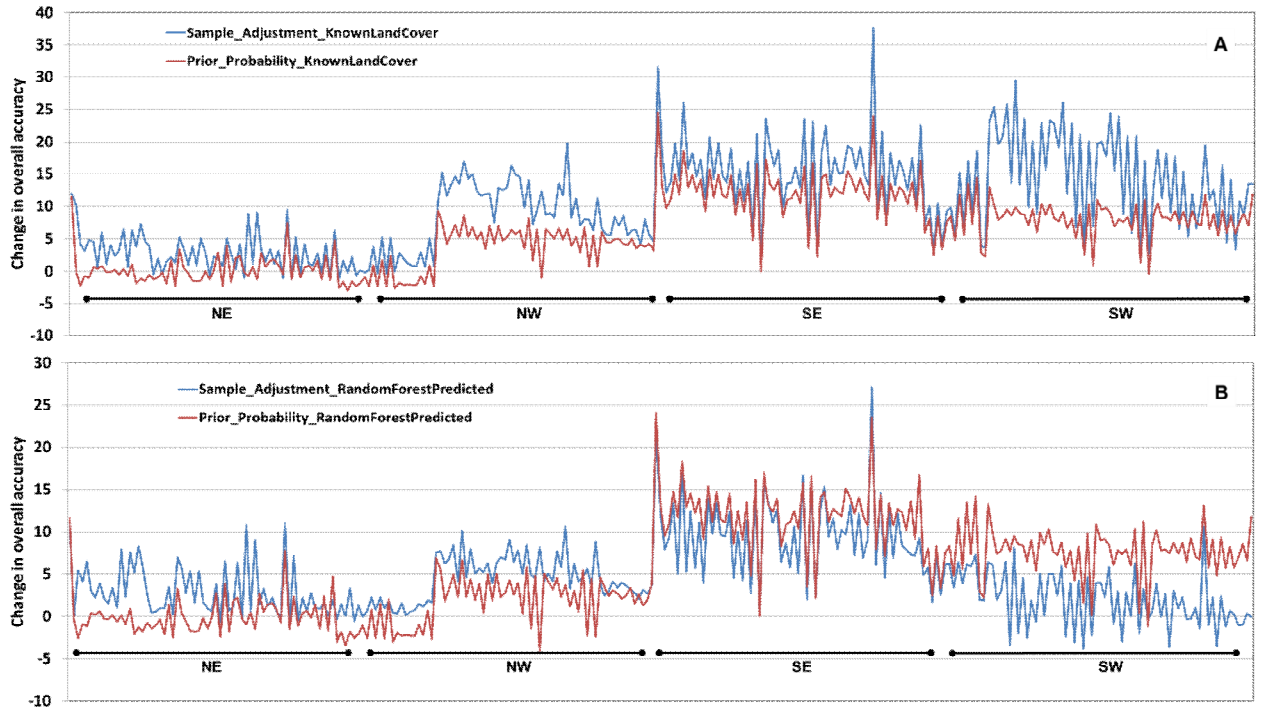

Figure 4: A) Improvement in overall accuracy for sample adjustment and prior probabilities with a known land cover distribution. B) Improvement in overall accuracy for sample adjustment and prior probabilities where the land cover distribution was initially estimated based on the input sample.

2.6 Modeling Improvement

To determine if either adjustment should be applied a generic model was developed to estimate the improvement in land cover calculated as:

$$I = C \sum_{i=1}^{N} S_i Fp_i |Fs_i - Fp_i|$$

Where i = land cover class from 1 to N classes. Fs_i is the sample frequency for class i scaled between 0 and 1. Fp_i is the predicted class frequency for the application domain (the region for which the land cover is being predicted) scaled between 0 and 1. S_i is a measure of class sensitivity to sampling. It was estimated by perturbing the input sample by ±20% for each class in 10 iterations. The standard deviation of the predicted class frequency was calculated using the results from the 10 iterations. C is a scaling coefficient to adjust the result to represent percent improvement (I). This index provides a generalized means to estimate if sample/prior adjustment should be applied. Figure 5 shows the relation between the estimated improvement I and the actual improvement. From Figure 5 , if an improvement of 4% or greater is estimated then it is likely safe to apply sample/prior probability adjustment.

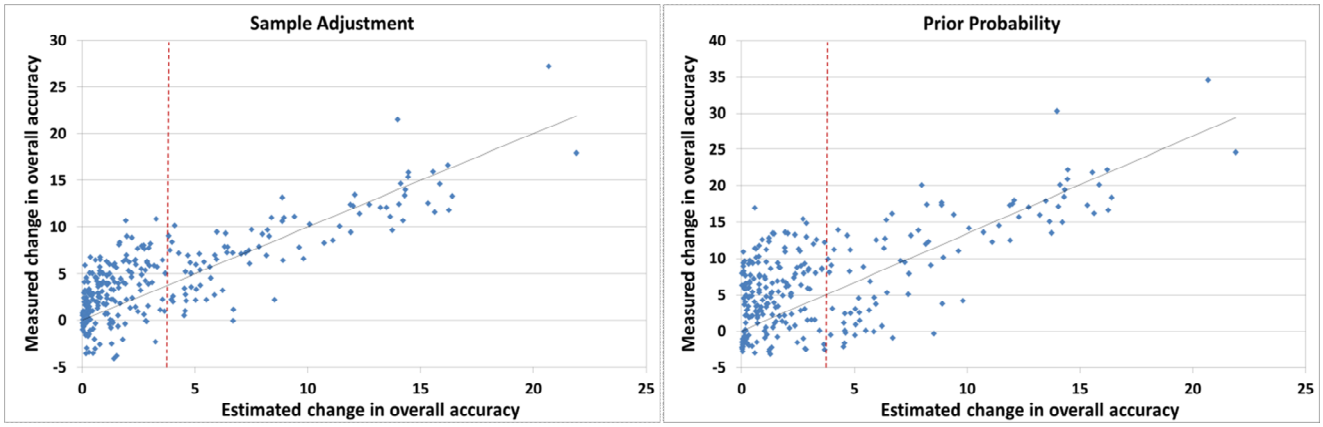

Figure 5: Relation between the estimated improvement I and the actual improvement for sample and prior probabilty adjustment approaches.

3. CONCLUSIONS

Results of the analysis show that significant improvement can be made regarding within landscape classifier extension with sample adjustment or prior probabilities. These findings are also likely informative for the local sub-region deviation problem as well, but this approach was not tested in this research. The level of improvement can be estimated and used to determine if an adjustment should be applied. In some cases adjustment may lead to reduced accuracy when the improvement is small relative to the classifier variability and classification accuracy. Further evaluation is required to better determine the generality and robustness of the approach. Additional research is will seek to apply these concepts to additional study areas and refine the general model developed here.

ACKNOWLEDGEMENTS

This research was partly support by the Canadian Space Agency, Government Related Initiatives Program IMOU # 13MOA41001, Enabling Responsible Resources Development through Earth Observation.

REFERENCES

Abu-Mostafa, Y., Magdon-Ismail, M., and Lin, H. (2012). Learning from Data. AMLBook.com

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

Chen, C., Liaw, A., Breiman, L. (2004). Using random forest to learn imbalanced data. University of California, Berkeley.

Fernandes, R., Fraser, R., Latifovic, R., Cihlar, J., Beaubien, J., and Du, Y. (2004). Approaches to fractional land cover and continuous field mapping: A comparative assessment over the BOREAS study region. Remote Sensing of Environment, *89*(2), 234-251.

Latifovic, R., Pouliot, D., Sun, L., Schwarz, J., and Parkinson, W. (2015). Moderate Resolution Time Series Data Management and Analysis: Automated Large Area Mosaicking and Quality Control; Geomatics Canada, Open File 6, 25 p. doi:10.4095/296204.

Minter, T. C. (1978). Methods of extending crop signatures from one area to another. Proceedings, the LACIE symposium, a technical description of the large area crop inventory experiment (LACIE), October 23–26, 1978, Houston, TX.

Olhtof, I., Butson, C., and Fraser, R. (2005). Signature extension through space for northern landcover classification: a comparison of radiometric correction methods. Remote Sensing of Environment, 95:290-302.

Olthof, I., Pouliot, D., Fernandes, R., and Latifovic, R. (2005). ETM+ radiometric normalization comparison for northern mapping applications. Remote Sensing of Environment, 95:388-398.

Pax-Lenney, M., Woodcock, C.E., Macomber, S.A., Sucharita, G., and Song, C. (2001). Forest mapping with generalized classifier and Landsat TM data. Remote Sensing of Environment, 77:241-250.

Pouliot, D., R. Latifovic, I. Olthof, and R. Fraser. (2012). Supervised classification approaches for the development of land cover time series. In Remote sensing of land use and land cover. Edited by C. Giri. CRC press, pp 177-190.

Pouliot, D., R. Latifovic, N. Zabcic, L. Guindon, and I. Olthof. (2013). Development and assessment of a 250 m spatial resolution MODIS annual land cover time series (2000-2011) for the forest region of Canada derived from change-based updating. Remote Sensing of Environment, 140:731-743.

Pouliot, D., R. Latifovic, R. Fernandes, and I. Olthof. (2009). Evaluation of annual forest disturbance monitoring using decision trees and MODIS 250m data. Remote Sensing of Environment, 113:1749-1759.

Pouliot, D., Parkinson, W., and Latifovic, R. (2015). Evaluation of Landsat based fractional land cover mapping in the Alberta Oil Sands Region; Geomatics Canada, Open File 17, 16 p. doi:10.4095/296802.

Weiss, G., Mccarthy, K., and Zabar, B. (2007). Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs? International Conference on Data Mining, pp. 35-41.

Woodcock, C. E., Macomber, S. A., Pax-Lenney, M., and Cohen, W. B. (2001). Monitoring large areas for forest change using Landsat: Generalization across space, time and Landsat sensors. Remote Sensing of Environment, 78, 194– 203.