Random Forest Classification for Surficial Material Mapping in Northern Canada

by

William Parkinson

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Science

in

Geography

Carleton University

Ottawa, Ontario

Your file Votre référence
ISBN: 978-0-494-94316-8

Our file Notre référence
ISBN: 978-0-494-94316-8

Abstract

There is a need at the Geological Survey of Canada to apply improved accuracy assessments of satellite image classification and to support remote predictive mapping techniques for geological map production and field operations. Most existing image classification algorithms, however, lack any robust capabilities for assessing classification accuracy and its variability throughout the landscape. In this study, a random forest classification workflow is introduced to improve understanding of overall image classification accuracy and to better describe its spatial variability across a heterogeneous landscape in Northern Canada.

Random Forest model is a stochastic implementation of classification and regression trees, which is computationally efficient, effectively handles outlier bias can be used on non-parametric data sources. A variable selection methodology and stochastic accuracy assessment for Random Forest is introduced. Random forest provides an enhanced classification compared to the standard maximum likelihood algorithms improving predictive capacity of satellite imagery for surficial material mapping.

Acknowledgements

I would like to thank my supervisors Murray Richardson and Hazen Russell, who saw the potential in my research interests. Both permitted autonomy for this research while providing significant direction and insight. Their support and patience was an invaluable asset.

I would also like to express thanks to my partner, Sarah. This thesis is a testament to her patience, cooperation, and support. Her ability to challenge my writing improved my ability to communicate my ideas for this thesis. Though I cannot replace the nights that included the phrase "I have to finish my thesis," we can forge new ones. I hope to support her correspondingly as she builds her career.

Finally, I am forever grateful to family for their love and support during my academic career. You have taught me how to overcome adversity with fortitude, and I would not be here today without your encouragement and advice. Thank you for all that you have provided for me, as it contributed much to my success and this thesis.

This research was supported in part by the Research Affiliate Program (RAP), Government of Canada via Geomapping for Energy and Minerals (GEM) project, and the Geoscience Knowledge Management (GKM) at the Geological Survey of Canada, Natural Resources Canada.

Table of Contents

List of Tables

List of Illustrations

Chapter 1: Introduction

1.1 Glacial Geology

Glacial geology researchers examine past glaciation through analysis of landform-sediment relationships on the modern landscape. A sub-field of geomorphology, glacial geology is the study of Earth's land surface processes through observation of morphology (shape) and surface materials (Smith & Pain, 2009). Geomorphological observations allow researchers to infer formative depositional and erosional processes from past land surface changes (Smith & Pain, 2009). Canadian surficial geology research has a rich history of applied field observation – dating back as early as 1789 (Stewart, 1916; Parry, 1967). However, field observation accounts for only a small portion of glacial geology studies, as the scales for both geologic processes and land areas are quite extensive, requiring airborne and space borne remote sensing analysis (e.g. aerial photography).

1.1.1 Mapping of Surficial Geology

As information is extrapolated from field observations, a landscape perspective become available. Context (in this case a map) mixed with detailed field observations and samples provide researchers the tools to infer processes that occurred on the landscape millennia ago. The two major components in mapping glacial geology are materials and landforms. Surface materials (broadly termed surficial geology) provide a critical source of information that broaden the context of observations on the

landscape. Landform mapping investigates the morphologic features (e.g. drumlins and eskers). Together, these maps provide the framework for operational applications such as mineral exploration and infrastructure development. The difficulty in mapping lies in understanding how to use the *in-situ* observations to extrapolate to the entire region of study.

The creation of maps is complex and Wilson (1939) recognized there was a need for more landscape information than what was collected at the time. Eventually, a new data source became available – aerial photographs taken by the Royal Canadian Air Force (Wilson, 1939). Wilson (1939) used 4000 oblique aerial photographs with field notes from other geologists to map an 18 000-square mile block east of Great Slave Lake. This study is one of the first instances that used aerial photographs. Further, it sparked the idea that a better understanding of the landscape could come from a broad examination using remote sensing technology. Studies soon began to incorporate aerial photographs into geomorphology research, making remotely sensed imagery a critical source of terrain information (Wilson, 1939). During the past 50 years, space borne imagery has also become a common data source.

1.1.2 Surficial Geology Mapping and Remote Sensing

Remotely-sensed imagery is widely used in a variety of disciplines including landcover mapping (Foody, 2002; Lowry et al., 2005), soils (Irvin, Ventura et al., 1997; Odgers et al., 2011; Pennock et al., 1987), agriculture (Samaniego & Schulz, 2009), and forestry (Healey et al., 2005). Geological studies have applied airborne (Douglas & Douglas, 1949; Stokes et al., 2006), satellite (Ford, 1984; Shaw et al., 2010; Sugden,

1978), bathymetric (Andreassen et al., 2008; Domack et al., 2005) and other remotely sensed data (Heroy & Anderson, 2005; Rafaelsen et al., 2002) to the interpretation of glacial landforms and landscapes.

Using remotely sensed information in aerial photographs and multispectral data, researchers examine visual elements to classify landscape features. Rabben (1960) qualified six elements; size, shape, shadow, tone and colour, texture, and pattern. Estes, et al. (1983) added three additional elements: height, site, and association. Finally, Teng (1997) added a tenth element, time. Teng (1997) groups these elements into four meaningful categories: tone and colour, geometry of objects (size, shape, height, and shadow), spatial arrangement of tonal boundaries (texture and pattern), and context of objects and phenomena (site, association, and time). Using remotely sensed information required a new way of thinking about the landscape and therefore glacial geologists needed to develop a new skill– image interpretation.

1.1.3 Cognitive Analysis of Remotely Sensed Data

To successfully map glacial landscapes from remotely sensed information, interpreters have employed high-level semantic definitions of landforms derived from field observations, which provides a working framework for landform – sediment (material) mapping. The spatial interpretation and integration of the landform - sediment relationship provides the basis for interpretation of glacial processes. The landforms and sediments are interpreted using the integration of the nine elements (excluding time) listed above. Some of the elements provide direct information on the landform, such as relative relief, shape, and spatial association. Other elements (such as

tone and colour) may provide direct information (e.g. high brightness and white colour indicating sand and gravel), but usually serve as vegetation and moisture proxies representing particular substrates (for example, a moisture regime can indicate sediment porosity or relative position on a slope). However, landform - sediment relationships are rarely defined by visual elements (tone and colour) alone. Instead, interpreters identify landforms material associations based on *a-priori* knowledge and/or assumptions on landform genesis (examples: drumlin-till; esker ridge-sand and gravel).

Researchers have applied a number of approaches to using remote sensing data in glacial geology studies. Early research attempted to map different glacial features using thermal imagery by exploiting the fact that water content is a controlling factor on soil temperature (Schneider et al., 1979; Western, et al., 1999). Sugden (1978) mapped glacial erosion from the Laurentide Ice Sheet by exploring the variations of landscape erosion (such as areal scouring and linear erosion) on Baffin Island using Landsat-1 and field data. Ford (1984) analyzed Seasat Synthetic-Aperture Radar for texture and applied conceptual models of drumlins from Flint (1971) and Embelton and King (1975) to extract and quantify the length width ratios of drumlins. Hyperspectral imagery has been used to successfully identify glacial landforms defined by surface materials (Waldhoff, et al., 2008). In most of these cases, researchers have used visual elements (sometimes indirectly) or landform - sediment relationships to classify maps. However, visual elements do not directly dictate the interpretation of material types, as researchers may interpret similar elements differently depending on the landform-

sediment conceptual model employed.

These complexities make image interpretation time consuming and subjective. Therefore, computational approaches are favoured if they minimize subjective human components, quickly provide generalized information for researchers, or are based on some set of clearly defined rules. A number of computational approaches currently exist to augment human interpretation of land surfaces such as 3D visualization tools and multi-spectral satellite information. However, extracting useful information on surficial materials in glaciated landscapes remains a significant challenge in remote sensing applications (Smith & Pain, 2009).

1.2 Automated Approaches

Subtle changes in relief, moisture, vegetation, that control colour, tone, texture, etc. in landscape imagery can cause significant perceptual differences in interpretations. Consequently, the development of landscape – material approaches that integrate multiple components within the researcher's cognitive interpretative process is ideal (e.g. expert systems). Most remote predictive mapping approaches have relied on single or multiple datasets classified without a fully developed expert systems approach (Grunsky et al. 2009). In these approaches the ability to distinguish materials based on only colour, tone and texture with similar confidence as cognitive interpretations is beyond the machine based classification of any single spectral dataset (e.g. Landsat). Furthermore, only abstract conceptual models are available to integrate image characteristics for landform or glacial material identification (Eisank et al., 2010;

Parkinson et al., 2010). This limits the ability build the appropriate expert system or to select appropriate data sources for analysis. Because the cognitive models are multifaceted (relief, tonal, spatial), researchers select statistical models that use multiple data sources, as well as replicate human decision-making elements of manual interpretation.

1.2.1 Machine-based Classification

Researchers using machine-based predictive mapping approaches attempt to classify imagery using computer algorithms and workflows that minimize the need for human intervention in addition to providing usable information on the landscape. Two common examples of machine-based approaches include i) unsupervised classification and ii) supervised classification. Unsupervised classification commonly groups similar signatures and adjacent pixels into a predefined number of classes using clustering algorithms (e.g. k-means clustering). Once all the pixels are grouped, an interpreter assigns each group to a given/specific class. In contrast, the supervised classification relies on the identification and classification of imagery using training sites, identified by an interpreter through field data and expert knowledge. Supervised classification requires researchers to identify classes on the image (known as training areas), usually through manual interpretation, or field reconnaissance and delineation by regions or polygons (Bhatta, 2011). Supervised approaches are more common to glacial geology (Grunsky et al., 2009, Harris et al. 2012, Brown et al. 2007). In the field of glacial geology, machine-based approaches to image classification have received less attention and consequently less application relative to other fields. This delay is prevalent when

compared to the more sophisticated and established methods used in other disciplines such as forestry, soils, and landcover (Miao and Heaton, 2010; Odgers, Nathan, and Minasny, 2011). By applying algorithms used in other disciplines, this research attempts to advance predictive mapping of surficial geology approaches to match current capacity demonstrated within other disciplines.

1.2.2 Remote Sensing of Glacial Geology

Classifications have attempted to use Digital Elevation Models (DEMs) in order to incorporate topographic elements of the glacial landscape (Brown et al., 1998). In some cases, DEMs (or derivatives such as deviation from mean elevation) has proven useful (Harris et al., 2012). Other studies have used landscape derivatives, calculated from DEMs, to create geometric signatures (Brown et al., 1998; Graff & Usery, 1993; Pike, 1988). Researchers used geometric signatures to characterize continuous topography of landform composites, such as hills or plains, rather than individual landforms (Brown et al., 1998; Graff & Usery, 1993; Pike, 1988). The dependence on primitive geometric signatures presents a problem of oversimplifying the landform. As a semantic construct, geometric signatures fail to consider the importance of scale and context that glacial geologists use in image interpretation. While one data source is not robust enough to classify the landscape, using all possible elements may not be ideal either. This is because data redundancy or un-necessary data sets can lead to statistical over fitting and more assumptions that must be managed. Therefore, researchers must evaluate all data sources for their importance within the statistical model. Users must select data sources for their applicability to image-interpretative elements (usually by proxy).

Expanding on expert knowledge is crucial in order to understand how to incorporate data sources into analysis of glacial materials.

The studies mentioned above have demonstrated limited success in the application of machine-based classification due to their general reliance on single datasets (e.g. DEMs, Landsat, Radar, etc.), oversimplification of signature significance, and lack of semantic models to draw connections from classes to data sources. Understanding and incorporating interpretive processes used in glacial materials map production is necessary – even in supervised machine based approaches. Some studies have attempted to quantify these relationships within such workflows (Eisank et al., 2010; Parkinson, 2010). However, literature suggests progress is only starting to begin on selecting appropriate variables and algorithms to match cognitive models (Sinha and Mark, 2010). Furthermore, as algorithms are selected, appropriate variables must also be selected to match the visual proxies used for analysis (for example, elevation can be described using DEMs and associated derivatives or tonal reflectance can be described using Landsat bands). Unfortunately, there is a lack of adequate conceptual models suitable for integrating proxy variables into classifications, and researchers usually select variables for their ability to perform in accuracy assessments, rather than their significance to a conceptual model of landform-sediment relationships (Grunsky et al., 2009).

1.2.3 Classification Algorithms

The majority of studies classifying glaciated landscape material types have employed the Maximum Likelihood Classification (MLC) method (Grunsky et al., 2009,

Harris et al. 2012, Brown et al. 2007). However, these existing studies lack the ability to properly model the surficial material classes, manage inadequate or noisy training data, and/or handle poor spectral separability among classes. Noisy data includes anything that reduces the signal within the classified imagery. As well, outliers often occur when training captures incorrect pixels and this can directly affect classifiers that are programmed to fit all the sample data. Further, algorithms that use MLC, such as Harris' (2012) Robust Classification Method (RCM), are constrained by the underlying assumptions – most notably the assumption of parametric distributions.

Recently, interest in classifiers that do not make assumptions on the underlying distribution within supervised classifications has grown (Khalyani et al., 2012). Furthermore, the need to address common issues in remote sensing such as noisy data and training outliers is mounting (Mountrakis et al., 2011). A number of non-parametric classifiers are available from research in machine-based methods such as Support Vector Machine (SVM), Classification and Regression Trees (CART), and Random Forests (RF). SVMs are iterative learning classifiers that maximize the gap between classes as much as possible by identifying a linear boundary that minimizes misclassification (Mountrakis et al., 2011). While SVM is potentially a robust non-parametric classifier, it is not optimized to manage inherent issues of noisy data and outlier effects (Mountrakis et al., 2011). An alternative classification scheme is CART, which makes decisions at each node of a decision tree using input predictors (Liaw and Wiener, 2002). However, CART is inherently unstable to noise and outliers (Breiman, 2001). Instability causes greater variation in classifications and can lead to reducing overall accuracy or confidence in

variable significance. RF is a stochastic implementation of CART to better manage instabilities by the application of hundreds to thousands of trees that use randomly selected predictor variables at each node in a voting system (Breiman, 2001). Implementing stochastic measures have resulted in a classification that often outperforms many other classifiers including SVM (Liaw and Wiener, 2002). Additionally, RF only requires two parameters, the variables to make the prediction and the number of trees to use, and the model is usually not overly sensitive to either (Liaw or Wiener, 2002). Instability refers to the extent to which models will change when small changes are made to the input data. For example, sampling a validation subset from the training population can cause vastly different classifications using CART leading researchers to believe this model without stochastic variation is unstable (Breiman, 2001). Regardless of which supervised classification method is employed, researchers must assess performance (often expressed as accuracy) in order to gauge the strength of the statistical model.

1.2.4 Accuracy

In supervised classifications, an accuracy measure expresses the level of confidence in the final product (Brown et al., 2007; Grunsky et al., 2009; Harris et al., 2012). Confusion matrices (a common approach to describing accuracy) illustrate accuracies by tabulating validation sample data against the predicted samples for the study area. However, the sample data (derived by sampling the training data) can greatly influence the result and mislead the confusion matrix's user when underlying assumptions (such as pure pixels and discrete classes) are not met (Foody, 2002). Pure

pixels are a function of scale and indicate that the information underlying an X by Y meter pixel is homogenous, which is rarely the case. Discrete classes refer to the assumption that all conceptual classes have zero spectral overlap from one class to the other. Furthermore, sample data can be difficult to collect because of fluctuating spectral responses from seasonality or atmospheric changes. As well, if the sample data contains interpretive classes (where selection reflects assumed genetic process rather than only spectral characteristics) the data may not as always be spectrally unique (Harris et al., 2012). Harris et al. (2012) addressed this problem by implementing the RCM. RCM iteratively samples data randomly into training and validation sets to understand the impact of individual training areas and to mitigate the known errors to occur during sub setting procedures for cross-validation. This approach proved useful, as it permitted the assessment of training areas – as well as the spatial extents of accuracy – by providing a measurement of classification variability (Harris et al., 2012). Nevertheless, RCM is constrained by the underlying assumptions of MLC. Therefore, the operational feasibility of RCM or MLC across vast northern landscapes (a common problem for surficial material mapping) is problematic given model assumptions and the difficulty of creating discrete classes. An appropriate algorithm must manage the difficult conditions presented by this type of large-scale analysis. Random Forest is one such algorithm because of the non-parametric assumptions and its stochastic nature.

1.2.5 Advantages of Random Forest for Surficial Material Mapping

Random Forest models are a class of statistical classifiers that utilize multiple stochastic iterations of the CART algorithm to develop a classification (Breiman, 2001).

RF models have increasingly been used in disciplines such as machine learning (Breiman, 2001), bioinformatics (Strobl et al., 2007; Calle et al., 2011), environmental science (Kuhnert et al., 2010), psychology (Strobl et al., 2009), and has recently become an emerging tool in remote sensing (Torbick et al., 2012; Rocchini et al., 2012, Khalyani et al. 2012, Pringle et al., 2012; Dorigo, 2012). The RF approach has great potential to improve upon existing classifications because of the ability to manage (a) noisy data (Gislason et al., 2006); (b) outliers (Kuhnert et al., 2010); and (c) non-parametric classes (Breiman, 2001). The non-parametric statistical distribution of material classes and complex training make RF an ideal candidate for use in machine-based predictive mapping of surficial materials. Furthermore, a major advantage of RF classifier in all disciplines has been variable importance plots, which measure the significance of predictor variables within the statistical model. Researchers use these plots to assess the predictive capacity and to help reduce the number of input variables required (Khalyani et al., 2012).

In glacial geology mapping, there are a number of advantages to using the RF algorithm compared to other approaches. The variable importance plots can assist researchers in selecting variables, which can assist future research in building stronger conceptual models to link multispectral data to materials on the landscape. The voting system, in conjunction with overall accuracy measures, can be used to assess the predictive capacity and overall performance of the RF model. The non-parametric nature of RF allows researchers to build classes that do not have to be parametric. Finally, the ability to manage outliers and noisy data allow researchers to build useful

statistical models in limited data environments (such as Canada's North).

1.3 Objectives

This thesis attempts to harness the predictive power of RF modeling while managing constraints unique to remote sensing, particular in relation to glacial mapping. Specifically, it aims to address some of the major challenges in classifying glacial landscapes with remote sensing using a new approach based on RF modeling. Two study areas are used to evaluate a variety of circumstances common to surficial material mapping, such as classifying data within a single Landsat tile, and across multiple tiles with poor spectral balance. This thesis addresses the spectral balancing problem by using data sources considered relatively un-usable by traditional classifiers. The modified RF will allow users to evaluate the significance of the training data and data sources on the resulting classification.

The overarching objectives of this thesis are as follows:

1) Thoroughly review issues in mapping glacial landscapes and identify emerging techniques and literature to address them.

2) Use RF to develop and implement a workflow for classifying glacial landscapes and to develop a suite of diagnostics for assessing performance and variable importance of classification.

3) To apply the RF workflow in two northern Canadian landscapes and compare results with traditional classification approaches in order to evaluate operational feasibility across the northern landscape.

1.4 Structure of Thesis

This thesis contains five chapters. Chapter 1 introduces the general context and broad justification of utilizing the RF model. Chapter 2 overviews the existing literature in remote sensing and how it pertains to machine-based geological predictive mapping. Chapter 3 details a new remote sensing workflow using RF and applies it in a case study east of Great Slave Lake. Chapter 4 will highlight the robustness of the RF model and explore the potential of the algorithm to manage spectral balancing while comparing it to the widely used MLC classification approach. Chapter 5 outlines the key outcomes of this thesis.

1.5 Study Areas

Common glacial geology issues include vast study areas with poor spectral balancing between image mosaics and complex classes that do not match parametric assumptions. Two study areas have been selected to develop and test the RF algorithm under common circumstances within glacial geology mapping (Figure 1.1). This will allow for the assessment of its operational feasibility for use across the northern Canadian landscape. The first study area (Figure 1.2) is a small portion of a single Landsat scene used to introduce the modified RF workflow developed in this thesis. The second study area (Figure 1.3) is much larger, farther north, and is used to compare RF to MLC and to address issues that occur during machine-based predictive mapping (such as bimodal classes and spectral balancing).

Figure 1.1: Location of both study areas for this research thesis. Labelled 1 and 2 for their respective location in the thesis Chapter 3 and Chapter 4.

1.5.1 Chapter 3 Study Area

The first study area is north of Great Slave Lake in Northwest Territories, on the northern half of National Topographic System (NTS) sheet 75M, Mackay Lake (Figure 1.2). This study site uses one Landsat scene in order to avoid phenological and mosaicking issues, thereby addressing the major concerns associated with cross-validation and RF. The modified workflow is presented with a specific focus on how to

best utilize RF for variable selection, classification, and accuracy assessments. The modified workflow manages cross-validation issues and provides a robust digital product.

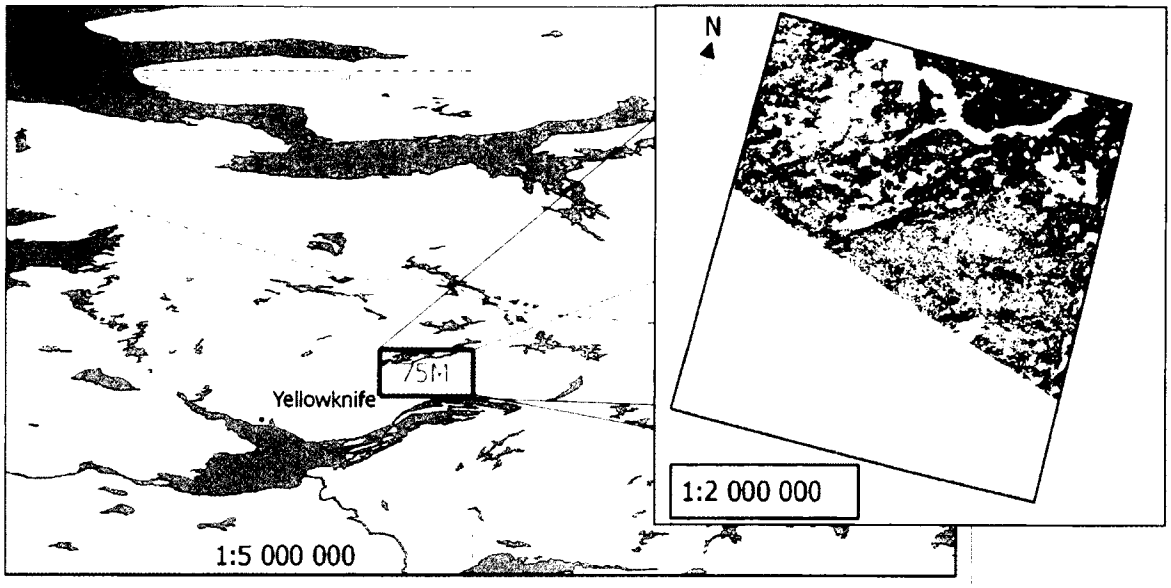


Figure 1.2: Study Area One, location within NTS sheet 75M.

1.5.2 Chapter 4 Study Area

The second study site is located on the east side of Victoria Island in Northwest Territories, and spans four 1:250,000 NTS map sheets (Figure 1.3). This study area is used to address the spectral mosaic issues, and to directly compare a RF classification to MLC. This area has a partially balanced mosaic, which Canadian Center for Remote Sensing (CCRS) produces by matching histograms to MODIS for the same area. However, the correction is not perfect and the area has three unique domains of spectral continuity. Consequently, past mapping activities have classified this region using three different sets of training and validation.

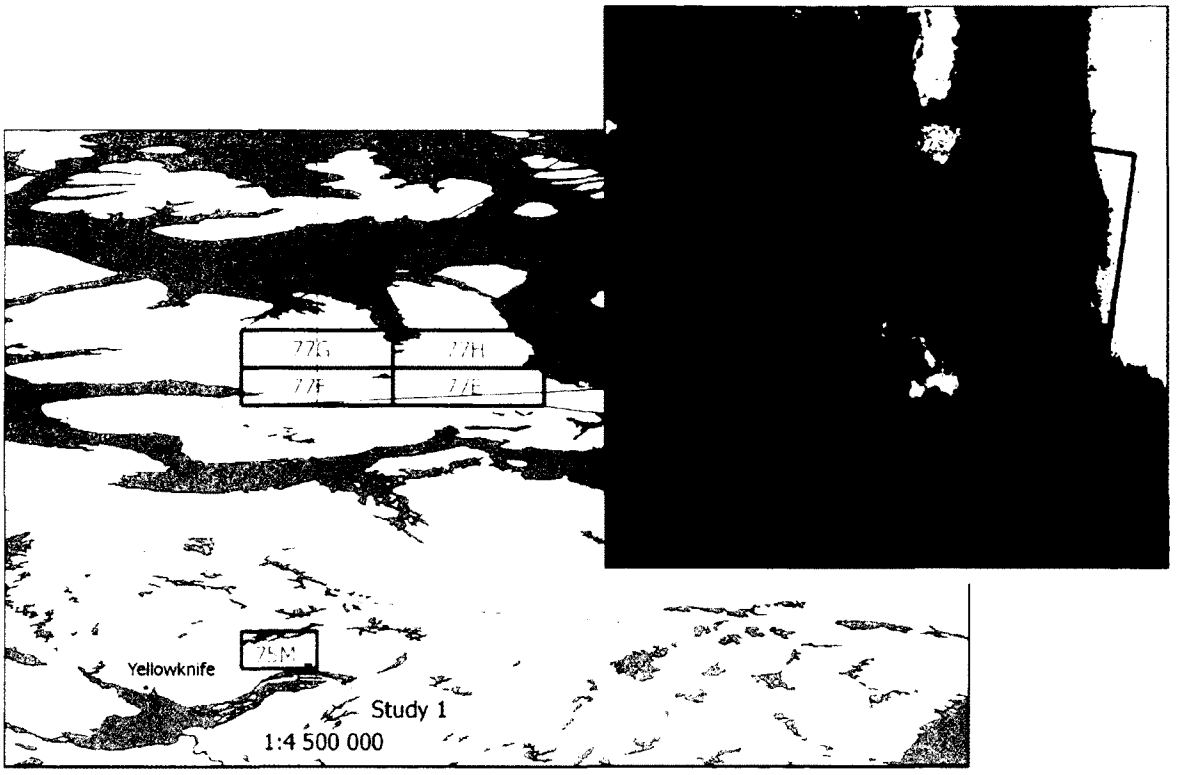


Figure 1.3: Study Area Two, East side of Victoria Island, Northwest Territories

Chapter 2: Literature Review of Remote Sensing of Glacial Geology

The field of remote sensing has developed considerably over the past few decades. However, machine-based predictive approaches for glacial geology remain limited. This chapter explores issues associated with remote sensing, with a focus on their impact to glacial geology mapping. First, the complexity associated with identifying training data in a glacial geology context is discussed. Next, the separability of glacial material classes is explored within a spectral context. Issues due to the presence of bimodal distributions, spectral balancing, and data set mosaics are also discussed. Then, an exploration of accuracy approaches is conducted to demonstrate the inherent uncertainty in accuracy measurements, and an introduction of how to manage these uncertainties (or measurements) is provided. This investigation outlines both the reasoning and details of implementing RF for surficial material mapping.

2.1 Training Datasets for Supervised Classification

Expert geological input is essential in the selection of training areas during supervised classification. Not only does training data provide the basic information for the algorithm, but it also has a larger impact on classification accuracy than the choice of algorithm itself (Campbell, 2003). Furthermore, the training samples must be representative of the different landcover types within the image (Chen, 2002; Eastman, 2002; Muchoney, 2002). Representativeness is more critical in parametric classifiers such as Maximum Likelihood Classifier (MLC), and may not be as significant in some non-parametric classifiers – for example, with Support Vector Machine (SVM), edge of

class distributions (spectral edge) are used (Foody & Mathur, 2004). Spectral edge roughly defines the region for a class in N-dimensional space in which a straight line is drawn to identify pixels in one class or another. Identifying representative samples of a population becomes a complex problem for surficial materials because they are complex to interpret through remotely sensed imagery. Normally, the analyst develops training datasets through manual interpretation. This often leads to a dataset that is biased towards the purest spectral signature and most easily identifiable land cover class. Trained classes usually consist of the purest examples corresponding to a specific range of the class' spectral signals. However, ranges of values for the classes are rarely captured and consequently are not representative of the spectral range of the imagery and thus making it unsuitable for MLC (Chen, 2002).

2.1.1 Conceptual Classes (Genesis) and Class

The genetic complexity of glacial materials poses a significant challenge to machine based predictive mapping. Although some material types – such as thick till and thin till – may be conceptually unique, their spectral signatures are very similar. This problem is manageable by incorporating additional data sources, for example using RADAR imagery for texture distinction between bedrock and boulders (Grunsky et al., 2009). Unfortunately, direct links from conceptual classes to data sources are not always available or clear. Often studies will use all available data for analysis in an attempt to improve overall accuracy, regardless of the potential application to the conceptual classes in question (LaRocque et al., 2011). In order to make the links to the data sources available, researchers must understand the internal complexities of classes and

their impact on spectral data distributions.

2.1.2 Class Spectral Complexity

Spectral signatures of classes are complex in glacial geology because of variations in moisture and vegetation. Unlike land cover (where classifications are concerned with surface phenology rather than underlying substrate), surficial materials may constrain surface vegetation but not dictate variation across the landscape. Regions of the landscape can have similar materials types, yet have entirely different moisture regimes and vegetation structures. Thick till is commonly associated with a number of topological shapes (drumlins, flutes, hummocky terrain, etc.), creating a diverse phenological and moisture regime under the similar material. Consequently, the spectral signature changes depending on the time of year, topography of the landscape, moisture of the soil, and dominant land cover type for the area. Signatures are subject to changes across regions, so no single methodological approach is appropriate and no clear signature is definable for any material type.

2.1.3 Spectral Image Mosaics

The majority of glacial mapping studies are conducted over a large study area. As such, it is often required that multiple images be mosaicked together. Landsat scenes cover an area of 33 000 km^2. For study areas such as 1:250 000 National Topographic System (NTS) map sheets (~10 000 km^2), it is still necessary to tile images either along a path or from adjacent paths because Landsat satellites have an oblique path relative to NTS tiles. Therefore, even though areas based on the NTS map sheets are smaller than

33 000 km^2, they may require multiple image tiles to cover the complete area. Unfortunately, this usually leads to acquisition of imagery on different dates and/or with different atmospheric and weather conditions. The challenge of merging the spectral signal of two adjacent images varies in severity, often depending on the magnitude of the temporal offset and associated phenological or moisture regime differences. There is greater latitudinal and longitudinal variation as the study area increases, which causes a greater challenge in northern areas where the vegetation season is short. The production of spectrally balanced images is complex, particularly when the statistical uniqueness of an image is required. Proper image balancing is accomplished by using a number of different approaches including overlapping regions, auxiliary data sources and histogram matching (Yong et al., 2001). Unfortunately, most mosaics require a similar season of acquisition for a successful merging of images.

One approach to minimize the scene balancing issues is to classify individual tiles separately in order to produce stronger predictive models and then merge after classification (Beaubien et al., 1999). A major disadvantage of this approach is the increased workload to generate multiple training datasets. There is also the potential for inconsistent training across regions due to the presentation of different spectral signals for the same classes. Additionally, this approach will often leave a visible seam on an image. Studies can rarely address the range of mosaicking issues; however, understanding how they happen can provide insight into solving classification problems and errors.

2.1.4 Data Mosaics

In addition to spectral images, other data sources must also be merged in order to achieve synoptic coverage of vast northern regions. Some studies have used ancillary data sets such as the Canadian Digital Elevation Database (CDED) that is merged together from a series of smaller tiles in order to maintain 1:50 000 precision. For example, if a study site contains a single NTS sheet (1:250 000 map product area), it uses 16 CDED tiles (each with a 1:50 000 map scale) for training and processing. This means 32 tiles (16 at 1:50 000 scale, each containing an east and west portion) must be downloaded and merged together for a single study area. Unfortunately, artefacts appear in the data mosaics. Visible artefacts can result in large-area seams, representing the former individual tiles. Consequently, derivatives produced by processing merged CDED data will reflect inconsistencies in the data. As studies increase in extent, issues associated with the use of multisource data inputs are exacerbated. As classes become more genetically complex in definition, ancillary data sources such as digital topography become more important for experts to understand the landscape.

2.2 Cross-Validation and Accuracy

Cross-validation is an integral part of classification. It involves randomly sampling training data in order to independently validate the model. Cross-validation creates a confusion matrix, which is the core of accuracy assessment (Foody, 2002). There are a number of points to consider when assessing cross-validation of geographic data: i) random subsampling, ii) spatial autocorrelation, and iii) validation sample proportions.

Random subsampling of training data is required in order to create an independent data set for validation, but this randomness (if pixels are sampled from polygons) can cause spatial autocorrelation in geographic data sets (Legendre, 1993). Spatial autocorrelation refers to individual sample bias because of their similarity to an adjacent sample. This occurs because training by regions (polygons) tends to have reduced internal variance, making any individual polygon biased and less representative of the population (Campbell, 1981; Labovitz & Masuoka, 1984). Optimal solutions include more systematic approaches to training, such as selecting signatures every n^{th} pixel or randomly selecting pixels across the scene to remove spatial correlations (Chen, 2002). The use of polygons is still acceptable due to its compatibility with field observations and human interpreters. Further, subsampling validation data by polygons rather than pixels helps mitigate the issue so training and validation data will only be internally correlated (Grunsky et al. 2009; Harris et al. 2012). However, sampling can provide misleading accuracy assessments, poor statistical models, and single interpretations of the landscape (Foody, 2002). Therefore, there is a need to generate multiple accuracy evaluations (Foody, 2002).

2.2.1 Stochastic Classification Approaches

Optimizing the use of training data and improving error estimations has led to the development of a variety of stochastic application approaches, thereby allowing for the implementation of various other classifiers. Studies have introduced stochastic methods as a solution to instability in the model caused by sampling the data for training and validation (Strobl et al., 2009). For example, to better optimize the use of

training data, multiple iterations provide the estimation of probability and variability caused by sensitivity of a classification algorithm to training data. Ensemble methods (called bagging) combine multiple classification iterations (with separate training and validation set) of a statistical model to approximate relationships and thus classifications more robustly (Strobl et al., 2009). In bagging, the combined vote from all the iterations is used to decide how to classify an individual pixel. In all training and validation sampling discussed in this thesis, data samples are drawn one of two ways, a bootstrapping (same size drawn, with replacement), or subsample (smaller size drawn, without replacement). This section reviews three stochastic classification methods: Bagging, Robust Classification Method (RCM), and Random Forest (RF). All of which use random sub-sampling of data inputs to improve classification accuracies and to produce better estimates of classification accuracy. These methods attempt to overcome some of the pitfalls inherent in other classifications in order to provide more robust analysis methodologies.

2.2.1.1 Bagging

Breiman (1996) introduced bagging predictors as a method for generating multiple iterations of classification algorithms. He used the iterations to aggregate results into a final predictor using a voting system. The underlying principal in this approach was the recognition that perturbing sample datasets (by sub-setting for validation) can cause significant changes in any single classification. Consequently, the use of a single iteration is less informative (Breiman, 1996). Bagging often uses bootstrapping where replacement is used to allow for training samples to be used more than one time,

creating artificial weights of training data. This results in an increased variance of individual classification cases, making any one iteration less useful than a single iteration without bagging (Bühlmann & Yu, 2002). Bagging works well in unstable predictor models such as Classification and Regression Trees (CART) and linear regression. It has made unstable methodologies a viable option for utilization and has led to the creation of RF, which at its core is bagging of CART decision trees. Thus, this concept applied to glacial geological mapping has great potential given the quality of training data used and the inherent sensitivities to subsampling datasets.

2.2.1.2 Robust Classification Method

Training issues during MLC provided the justification for developing the RCM for supervised classification of glacial geology. Harris et al. (2012) presented the RCM approach to manage training area variability due to inconsistencies in the training process. This methodology uses a handful of iterations (20-50) with random training and validation subsamples, producing a classification and confusion matrix per iteration (Harris et al., 2012). In effect, this approach is bagging of MLC classification. The variability from iteration to iteration presented components of overall accuracy and spatial measurements of certainty with the deviation of a pixel across 20 iterations. For the first time in surficial material mapping, studies were using spatial accuracy measures rather than only a confusion matrix, which aided in understanding the effects of training, validation, and subsampling (Harris et al., 2012). However, bagging methods provides little improvement when applied to highly stable models such as k-nearest neighbour methods and MLC (Breiman, 1996). Therefore, this concept must extend to a

more complex classifier that can make better use of the stochastic sampling approach.

2.2.1.3 The Random Forest Classification

The RF classification algorithm extends the use of bagging and CART by applying bootstrapped bagging. It uses omitted data for validation (data not sampled after bootstraps), and randomizes variable selection (for example Landsat bands) to further diversify classification trees (Breiman, 2001; Strobl, 2009). Breiman (2001) introduced the RF classifier as a low computational overhead model that manages noisy data. RF classification approaches have recently shown potential for remote sensing (Gislason et al., 2006; Miao & Heaton, 2010; Pal, 2005). Advantages of the RF classifiers are their ability to manage noisy data (Gislason et al., 2006), outliers (Kuhnert et al., 2010), and overtraining (Gislason et al., 2006). For data with any of these issues, the RF classifiers are more efficient than bagging (Breiman, 2001). These factors and the ability to manage non-parametric data distributions generally allow RF to outperform MLC when dealing with complex classes. Further, the RF classifiers produce variable importance measures using Gini Index or Mean Decrease accuracy that assess the predictive capacity of any data set against the training data. These variable importance measures have shown to be the most beneficial aspect of RF (Boulesteix et al., 2011; Breiman, 2001; Gislason et al., 2006; Kuhnert et al., 2010). Internal measures of RF provide data to assess model accuracy but do not take into account the spatial issues of auto-correlation. This becomes apparent when considering RF was not originally designed for remotely sensed information (Brieman, 2001). The bootstrapping that occurs to measure Out of Bag (OOB) samples selects all points randomly (in this case points are

pixels) and does not identify training regions for validation. This fact in itself makes cross-validation before RF necessary.

If cross-validation is used in order to properly assess the accuracy, it will directly influence the reliability of variable importance and the accuracy of the final product (Calle & Urrea, 2011; Strobl et al., 2007). This brings the model back to its logical inception where stochastic training and validation necessary in order to manage instability in CART caused by sub setting for training and validation. Therefore, any implementation of RF must manage this instability directly in order to use the variable importance measures and understand accuracy.

2.3 Conclusion

Issues such as obtaining unbiased yet representative training data, defining genetic (conceptual) classes, and dealing with issues pertaining to data mosaics, has led to the utilization of RF. Furthermore, the classifications of multisource data using parametric statistical techniques (such as MLC) are not appropriate because multisource data sets across large study areas often results in bimodal and skewed distributions (Gislason et al., 2006). Large study area constraints (such as bimodal distributions that result from data patching and complex classes) complicate classifications further. Class related constraints (such as genetic definitions that do not manifest spectrally and non-representative samples) exacerbate these issues. Iterative approaches have addressed cross-validation issues, particularly unstable classifiers that may affect the final product. RF must be managed to account for spatial data and possible instabilities that occur

during cross-validation. The limiting factors in surficial material mapping have made a case for the RF classifier to be an ideal candidate for a new classification approach.

Chapter 3: Random Forest Modeling for Surficial Material Mapping

3.1 Introduction

Surficial material mapping in Northern Canada provides critical information for infrastructure development, mineral exploration, and reconstruction of past glaciations. Due to low population, high operating costs, and limited resources, the need to provide accurate information through remotely sensed data is paramount. However, selection of algorithms and data sources (such as Landsat) to produce a data product is a complex procedure. Data products (a map) usually include locations of bedrock outcrops, unconsolidated sediments, and components of landform identification (Fulton, 1995; Grunsky et al., 2009). Landform-sediment relationships infer information on surrounding sediments and paleoglacial dynamics (Clark et al., 2000; Heroy & Anderson, 2005). Important sources for mapping surficial materials in Northern Canada include Landsat Enhanced Thematic Mapping Plus (ETM+), Canadian Digital Elevation Datasets (CDED), SPOT 4 and 5, and aerial photographs. However, surficial materials are complex and based on multifaceted landform-sediment relationships. Surficial materials training data is created using image interpretation and field observation, often this data is poor. Spectral signatures may be homogenous between some unique materials or completely heterogeneous over the same material. Consequently, materials can be difficult to classify and can be comprised of non-parametric spectral distributions on the landscape. This causes training data to fail parametric or representative sampling assumptions. This problem is intensified when image sources include data mosaics comprised of scenes

taken from multiple dates, thus spanning a wide variety of phonological and/or meteorological conditions. These issues complicate machine-based predictive approaches to surface materials across vast northern landscapes.

Supervised classification approaches of surficial material mapping require training data to generate the classification. In many cases, researchers generate training data through air-photo or satellite interpretation, or subset it from past mapping activities (Brown et al., 2007; Grunsky et al., 2009). Occasionally, studies create training areas using one dataset, such as air photos, and analyze using another dataset, such as satellite data (Landsat; Brown et al., 2007). However, it causes a disconnection of the data from interpretation to classification. Furthermore, if studies subsample training data from past mapping activities, the results are subject to previous interpretations, cartographic choices, and other conceptual decisions made during the original mapping process. For example, often, users simplify classes through merging as a way to compensate for poor spectral class separability (Grunsky et al., 2009). Algorithms fail to effectively exploit training data in three scenarios common to glacial geology: i) when material types are similar but their genesis is different; ii) when classes represent characteristics that are transitional between two distinct classes, such as thin till to thick till (Kerr et al., 1995; Grunsky et al., 2009; Mei & Paulen, 2009); and iii) when classes have bimodal distributions. Increasing or changing data sources can compensate for problems in achieving class separability (Grunsky et al., 2009). Selecting and understanding data, a source's contribution to the physical understanding or statistical uniqueness of classes, is not clear but vital to surficial material mapping.

Data sources are selected generally based on coverage, cost, and usability within a given analysis procedure. Landsat ETM+ is the most commonly used multispectral dataset for northern surficial material mapping due to its availability, cost, spectral resolution, and coverage across remote areas. However, its spectral bandwidths are often insufficient for separating various surficial material types with different morphology or genesis, such as bedrock and boulders (Grunsky et al., 2009). Consequently, researchers have turned to other data sources such as Radarsat and CDED (Grunsky et al., 2009; Mei & Paulen, 2009). These datasets have occasionally improved classification results; however, studies often fail to connect the physical understanding of the landscape to terrain signatures. Additional data sources do not always solve classification problems, and researchers must develop or adopt new methodologies to help identify which data sources and variables are applicable to specific study areas and classifications. In particular, to achieve maximum coverage of the landscape, surficial material mapping efforts focussed on Canada's vast northern regions should use only the minimally adequate set of input data sources. From an operational point of view, adding additional imagery types on a case-by-case basis to improve local to regional classification accuracies does not effectively advance the much-needed sub-continental scale mapping effort now underway at the GSC.

Surficial material mapping efforts have mostly focussed on traditional image classification approaches, such as Maximum Likelihood Classification (MLC) (Brown et al., 2007; Grunsky et al., 2009). An extension of the MLC is Robust Classification Method (RCM), which uses iterative subsamples of training data to assess the impact of sample

selection and stability of the model. This approach has been successful in optimizing the MLC approach. However, due to the underlying assumptions of MLC, the advantages of RCM are still somewhat limited. Specifically, MLC assumes the sample data is parametric (i.e. normally distributed) and is representative of the population for the study area. Both of these statistical traits will nearly always be violated when mapping surficial materials and integrating multiple datasets. However, these procedures are surprisingly robust, even with non-normally classified data.

Recently, researchers have explored the potential of Random Forest (RF) for satellite image classification as a solution to many of the assumptions violated when using MLC (Khalyani et al. 2012). RF is ideal for the analysis of non-parametric datasets. Recently, RF classification approaches have shown potential for remote sensing applications (Torbick et al., 2012; Rocchini et al., 2012, Khalyani et al. 2012, Pringle et al., 2012; Dorigo, 2012; Gislason et al., 2006; Miao & Heaton, 2010; Pal, 2005). RF classifiers perform well against noise (Gislason et al., 2006; Na et al., 2009), outliers (Kuhnert, Henderson, Bartley, & Herr, 2010), overtraining (Gislason et al., 2006), and process data more efficiently than bagging (Breiman, 2001). Another major strength of RF is the potential to quantitatively assess variable importance with respect to classification success (Boulesteix et al., 2011; Breiman, 2001; Gislason et al., 2006; Kuhnert et al., 2010). This characteristic may help identify which imagery types and derivatives offer the most predictive capacity for synoptic mapping of surficial materials in Canada. However, variable importance measures are unstable during cross-validation (Calle & Urrea, 2011).

RF uses bootstrapped training samples in each of many model iterations and uses the excluded data for validation, called Out-of-Bag (OOB) error. Some researchers in remote sensing and ecology have used OOB error for accuracy assessment, removing the need of standard cross-validation (Khalyani et al., 2012; Cutler et al. 2007). If researchers do not use cross-validation, the RF model can use all of that data and instability will not occur. However, OOB error in RF is different from cross-validation approaches in remote sensing. In remote sensing spatial autocorrelation occurs during subsampling and bootstrapping because training by regions (polygons) leads to a reduced internal variance, making any individual polygon biased and less representative of the population (Campbell, 1981; Labovitz & Masuoka, 1984). RF does not bootstrap using polygons (as required for unbiased samples of spatial data) making OOB errors overly optimistic. Random samples of training data can also remove spatial autocorrelation (Khalyani et al., 2012). This means any training system that uses regions must account for both RF stability during cross-validation and not use OOB error to avoid overly optimistic estimations.

When training with polygons, cross-validated samples are required in order to measure the accuracy of the RF. RF instability occurs in variable importance during data set disturbance, such as cross-validation (Calle & Urrea, 2011). There has not been a great deal of research showing the impact of cross-validation on the RF model; however, it is an issue in other bagging approaches such as RCM, and even with robust classifiers such as MLC (Harris et al., 2012). Furthermore, additional information is available when using the RF classifier for classifications. During classification, RF uses a voting system

that provides additional information regarding the strength of classification. Therefore, additional approaches are required to manage instability and better gauge the strength of the classification using all of the measures provided.

3.2 Objectives

This study used a dataset from Northern Canada to assess the suitability of the RF method for machine-based surficial geological mapping. The objectives of this study was as follows:

1) Produce a new multi-stage RF classification (MSRFC) workflow that demonstrates the ability to manage instability in variable importance and accuracy measures; and

2) Create additional diagnostic products within MSRFC that help assess the strength and capacity of the RF model during classification.

This study produces a final product unaffected by subsampling instabilities, and provides additional data products in conjunction with a standard classification to help the analyst understand strengths and deficiencies of the mapping products being produced.

3.3 Methodology

This chapter introduces and tests the MSRFC to maximize RF output while training with polygons (Figure 3). The methodology can be separated into five sections as follows: i) RF algorithm ii) data preparation, iii) OOB Error tree selection, iv) variable selection and accuracy estimation, and v) classification. Each of these steps in the MSRFC is described, beginning with an overview of the mechanics of the RF model. Data

processing steps include an overview of the creation of training data and the variables selected for use in the model. The OOB error and tree selection step is used to select the number of stochastic iterations (trees) for RF. Variable selection and accuracy assessment uses stochastic iterations of training and validation for classification. The classification iterations use a variety of variable selection parameters to identify the most suitable data for classification. The MSRFC uses classification iterations to assess the potential of individual variables and provides an estimate of overall accuracy. The map production step includes classification as well as the production of the multiple products describing model uncertainty. Together, these steps optimize RF methods for application to remote sensing and provide the framework for future use in spatial analysis.

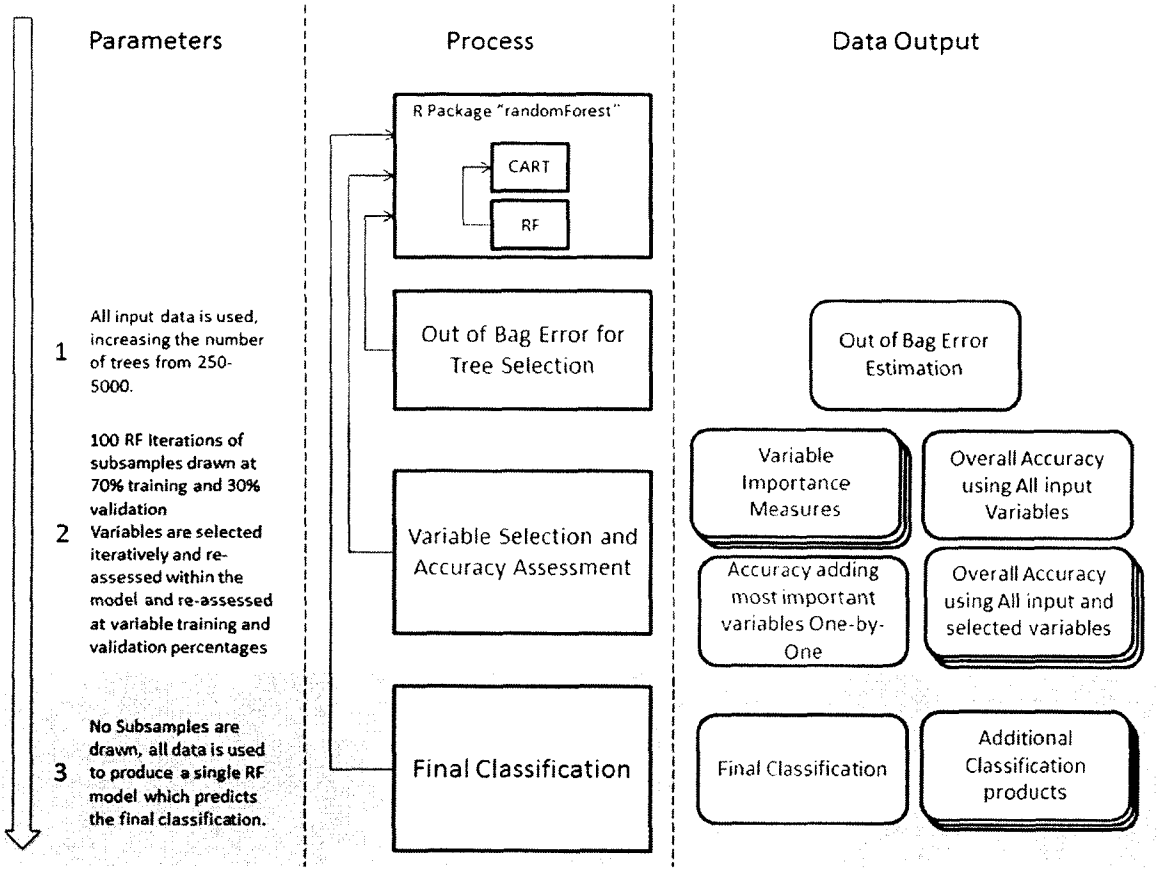


Figure 3.1: The Multi-Stage Random Forest Classification (MSRFC) workflow.

3.3.1 The Random Forest Model

The RF classifier is a stochastic implementation of the Classification and Regression Tree (CART) classifier. Each decision tree uses a random subset of input data (Breiman, 2001). Breiman (2001) introduced the RF classifier as a low computational overhead model for noisy data. RF generates training data by bootstrapping samples at each iteration (Pal, 2005). RF uses approximately 1/3 of the training data for internal cross-validation (OOB error), left over after bootstrapping (Breiman, 2001). The OOB error data is not user specified and is completely stochastic, based on un-sampled data

during bootstrapping. At each node of the tree, bootstrapped training samples are generated (Yang, 2010) and $\sqrt[2]{p}$ samples from input layers (e.g. Landsat bands) selected, where p is the number of input variables (Liaw & Wiener, 2002). The algorithm completes decision tree splits for nominal response variables (as in remote sensing classification) using the Gini index as follows:

$$Gini\ Index\ (D) = 1 - \sum_{i=1}^{m} \left(\frac{|C_{i,D}|}{|C_D|} \right)^2$$

p_i is the probability that a class C_i belongs to the dataset C_D for all of the difference classes *m*(Breiman et al., 1984). At each potential split, the algorithm calculates gini index:

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$

Where $Gini_A$ is one potential split into two different data sets D_1 and D_2. The Gini Index is a measurement of heterogeneity within classes. RF assigns individual probabilities depending on the split calculated. The algorithm continues across each node as it calculates all the possible splits. RF subtracts both Gini indices from either side of the split and from the parent node to identify the change in Gini.

RF grows the trees fully in the classifier, fitting all training data. Unlike CART, trees in a RF model are not pruned. Pruning is a process in CART where decisions at the bottom of the tree are excluded to generalize the classification and avoid over fitting of the data. Pruning can cause a loss of predictor information reducing accuracy (Pal,

2005). RF calculates the Mean Decrease Accuracy (MDA) and Mean Decrease Gini (MDG) with the trees, which measure variable importance. MDA measures the effect of perturbing a predictor variable on individual trees, estimating how important any one-predictor variable is per iteration. The MDG measures the Gini index decrease by variable across the RF model. Both MDG and MDA estimate the importance of any one-predictor variable and can aid in variable selection (Breiman L., 2001).

3.3.2 Data Processing

The study site is the northern 3931km^2 part of the Canadian National Topographic System (NTS) sheet 75M. It is located north of Great Slave Lake in the Northwest Territories, Canada (Figure 1.2). The paleoglacial landscape is north of the tree line located within the Coppermine River Upland ecoregion. The predominant land cover consists of dwarf birch, willow ericaceous shrubs, cotton grass, lichen, and moss (Environment Canada, 2012). Secondary less common growth can include stunted black spruce, tamarack, and white spruce (Environment Canada, 2012). The study area is underlain by the Slave structural province (Padgham, 1991; Padgham & Fyson, 1992) and is dominated by bedrock (Henderson, 1944). Current surficial geological mapping is limited to 1:250,000 manuscript maps (Aylsworth, unpublished) and 1:500,000 scale mapping (Aylsworth and Shilts, 1989). The, surficial geology is mapped predominantly as bedrock with areas of thin till. Local areas of thicker till are commonly associated with streamlined landforms such as drumlins. Additional local areas of sand and gravel occur in glaciofluvial corridors, characterized by eskers.

The study uses two datasets: Landsat ETM+ and CDED (Figure 2). The principal

dataset is the multispectral Landsat ETM+ image acquired on September 2, 2000 and published on GeoGratis (http://geogratis.cgdi.gc.ca/). The study area is contained within NTS 75M and therefore uses only the Landsat coverage that is contained within 75M (Figure 3.2).

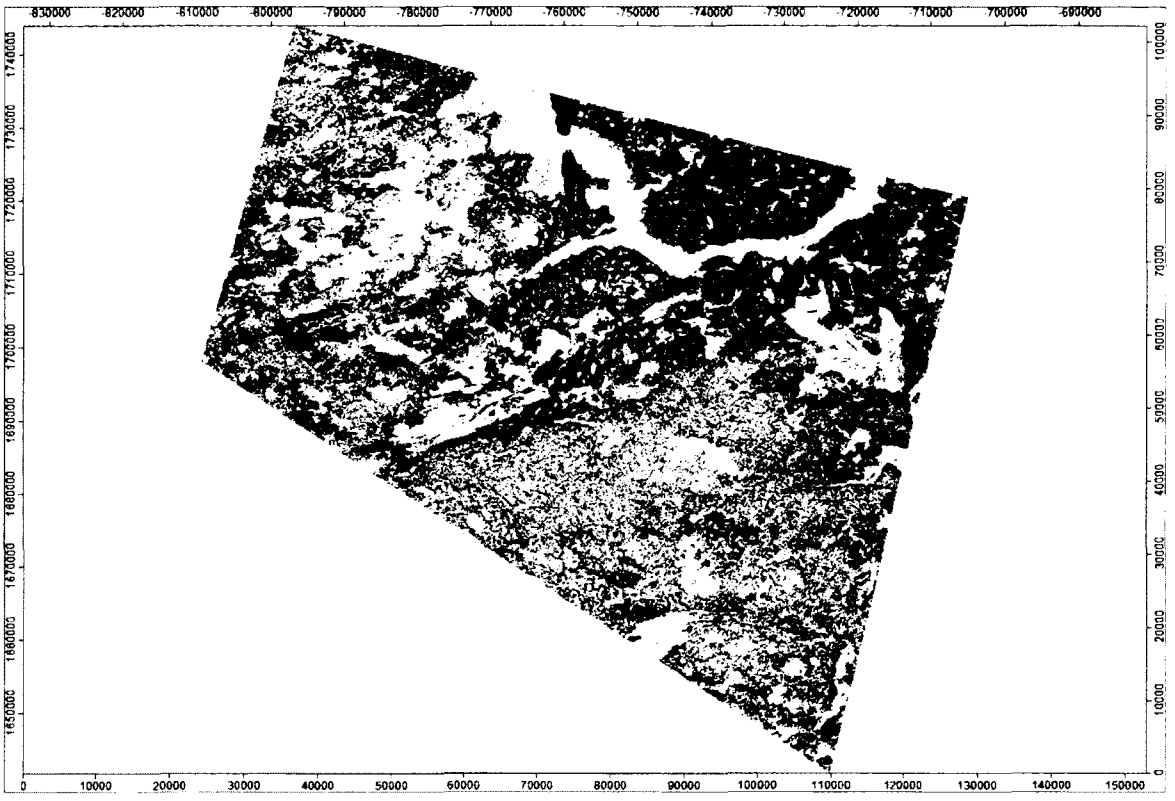


Figure 3.2: Red, Green, Blue Composite of Landsat ETM+ for the 75M study area, water is masked out using National Topograhic Database.

Since there is little physical basis in which to choose variables, an overwhelming amount of predictor data is used in order to demonstrate a data exploration exercise. The water bodies were masked out of this study area using the National Topographic database for water from GeoGratis (http://geogratis.cgdi.gc.ca/). The RF model results in variable importance assessment of a total of 78 input layers (Table 3.1). CDED data

was processed to generate 30 terrain derivative variables. Landsat ETM+ data was processed to produce nine band-ratio layers, 26 textual derivatives, and a tasselled cap transformation producing three layers. The CDED data, derived from contour data, used to generate terrain derivatives is stored in integer format and is not hydrologically conditioned this limits its ability to generate certain landscape and hydrological derivatives. There were 16 1:50 000 scale tiles with a cell size of 18m constructed a single CDED coverage for the study area. The CDED data was processed using the SAGA GIS (http://www.saga-gis.org/) software package. The terrain derivatives were generated by relating each grid cell to its neighborhood as has been useful for terrain analysis in past studies (Wilson and Gallant, 2000). Since surficial materials have a strong morphological component these derivatives were included with the hope they may provide terrain context to analysis. Each of these calculations relates the mean, difference from mean, standard deviation, range, minimum, maximum, deviation, and percentile using a user specified window size. This study uses different sizes of windows to calculate textures and DEM calculation (as described in the description of Table 3.1) to capture potentially scale dependent features in the landscape. The software package ENVI was used to produce the band ratios, tasseled cap transformation, and textural measures. The tasselled cap transformation produces three components representing brightness, greenness, and wetness. Tasseled cap transformation is a PCA on Landsat ETM+ bands that pre-defined weights in order to extract the spectral indicators of brightness, greenness and wetness (Kauth and Thomas, 1976). Textual measures were calculated using both occurrence and co-occurrence matrices at varying windows sizes

on the panchromatic band. The texture measures were scaled to Landsat ETM+ multi spectral resolution.

Table 3.1: Input Variables Used During Classification

	Name in Figures	Description
1	BAND1	Landsat ETM+ Band 1 (Blue)
2	BAND2	Landsat ETM+ Band 2 (Green)
3	BAND3	Landsat ETM+ Band 3 (Red)
4	BAND4	Landsat ETM+ Band 4 (NIR)
5	BAND5	Landsat ETM+ Band 5 (SWIR1)
6	BAND7	Landsat ETM+ Band 7 (SWIR2)
7	COCONT3	Landsat panchromatic co-occurrence texture Contrast 3x3 window
8	COCONT7	Landsat panchromatic co-occurrence texture Contrast 7x7 window
9	COCORR3	Landsat panchromatic co-occurrence texture Correlation 3x3 window
10	COCORR7	Landsat panchromatic co-occurrence texture Correlation 7x7 window
11	CODISS3	Landsat panchromatic co-occurrence texture Dissimilarity 3x3 window
12	CODISS7	Landsat panchromatic co-occurrence texture Dissimilarity 7x7 window
13	COENTR3	Landsat panchromatic co-occurrence texture Entropy 3x3 window
14	COENTR7	Landsat panchromatic co-occurrence texture Entropy 7x7 window
15	COHOMO3	Landsat panchromatic co-occurrence texture Homogeneity 3x3 window
16	COHOMO7	Landsat panchromatic co-occurrence texture Homogeneity 7x7 window
17	COMEAN3	Landsat panchromatic co-occurrence texture Mean 3x3 window
18	COMEAN7	Landsat panchromatic co-occurrence texture Mean 7x7 window
119	COSECM3	Landsat panchromatic co-occurrence texture Second Moment 3x3 window
20	COSECM7	Landsat panchromatic co-occurrence texture Second Moment 7x7 window
21	COVARI3	Landsat panchromatic co-occurrence texture Variance 3x3 window
22	COVARI7	Landsat panchromatic co-occurrence texture Variance 7x7 window
23	CURV	CDED DEM Curvature
24	DEV15	CDED DEM Deviation from Mean Elevation 15x15 window
25	DEV30	CDED DEM Deviation from Mean Elevation 30x30 window
26	DEV60	CDED DEM Deviation from Mean Elevation 60x60 window
27	DEV7	CDED DEM Deviation from Mean Elevation 7x7 window
28	DIFF15	CDED DEM Difference from Mean Elevation 15x15 window
29	DIFF30	CDED DEM Difference from Mean Elevation 30x30 window
30	DIFF60	CDED DEM Difference from Mean Elevation 60x60 window
31	DIFF7	CDED DEM Difference from Mean Elevation 7x7 window
32	LNDDG1	CDED DEM Downslope Distance Gradient
33	LNWET1	CDED DEM SAGA Wetness Index
34	MAX15	CDED DEM Max Elevation 15x15 window
35	MAX30	CDED DEM Max Elevation 30x30 window
36	MAX60	CDED DEM Max Elevation 60x60 window
37	MAX7	CDED DEM Max Elevation 7x7 window
38	MEAN15	CDED DEM Mean Elevation 15x15 window
39	MEAN30	CDED DEM Mean Elevation 30x30 window
40	MEAN60	CDED DEM Mean Elevation 60x60 window
41	MEAN7	CDED DEM Mean Elevation 7x7 window
42	OCENTRO3	Landsat panchromatic Occurrence texture Entropy 3x3 window
43	OCENTRO7	Landsat panchromatic Occurrence texture Entropy 7x7 window
44	OCMEAN3	Landsat panchromatic Occurrence texture Mean 3x3 window
45	OCMEAN7	Landsat panchromatic Occurrence texture Mean 7x7 window
46	OCRANG3	Landsat panchromatic Occurrence texture Range 3x3 window

47	OCRANG7	Landsat panchromatic Occurrence texture Range 7x7 window
48	OCSKEW3	Landsat panchromatic Occurrence texture Skewness 3x3 window
49	OCSKEW7	Landsat panchromatic Occurrence texture Skewness 7x7 window
50	OCVARIA3	Landsat panchromatic Occurrence texture Variance 3x3 window
51	OCVARIA7	Landsat panchromatic Occurrence texture Variance 7x7 window
52	PERC15	CDED DEM Percentile Elevation 15x15 window
53	PERC30	CDED DEM Percentile Elevation 30x30 window
54	PERC60	CDED DEM Percentile Elevation 60x60 window
55	PERC7	CDED DEM Percentile Elevation 7x7 window
56	PLCURV	CDED DEM Plan Curvature
57	PRCURV	CDED DEM Profile Curvature
58	R2O3	Landsat ETM+ Band 2/3
59	R3O2	Landsat ETM+ Band 3/2
60	R3O4	Landsat ETM+ Band 3/4
61	R3O5	Landsat ETM+ Band 3/5
62	R4O3	Landsat ETM+ Band 4/3
63	R4O5	Landsat ETM+ Band 4/5
64	R5O4	Landsat ETM+ Band 5/4
65	R5O6	Landsat ETM+ Band 5/6
66	R6O2	Landsat ETM+ Band 6/2
67	RANGE15	CDED DEM Elevation Range 15x15 window
68	RANGE30	CDED DEM Elevation Range 30x30 window
69	RANGE60	CDED DEM Elevation Range 60x60 window
70	RANGE7	CDED DEM Elevation Range 7x7 window
71	SLOPE	CDED DEM Slope
72	STD15	CDED DEM Standard Deviation of Elevation 15x15 window
73	STD30	CDED DEM Standard Deviation of Elevation 30x30 window
74	STD60	CDED DEM Standard Deviation of Elevation 60x60 window
75	STD7	CDED DEM Standard Deviation of Elevation 7x7 window
76	T1	Landsat ETM+ Tasseled Cap Transformation Component 1 (Brightness)
77	T2	Landsat ETM+ Tasseled Cap Transformation Component 1 (Greenness)
78	T3	Landsat ETM+ Tasseled Cap Transformation Component 1 (Wetness)

Lesemann (2010) from the Geologic Survey of Canada (GSC) identified six generic classes to map the glacial surficial materials of the area: bedrock, sediment blanket (thick), sediment veneer (thin), sand and gravel, and organics. Surficial geologists defined the classes by following past surficial geological maps within the study area and the surrounding northern and eastern zones (Kerr et al., unpublished). A surficial geologist generated training areas through interpretation of a three channel true colour Landsat image of the area. The interpreter created 7 to 25 polygon training areas, characterizing the multispectral properties of the respective units for each of the six

classes (Table 3.2). Polygon areas ranged in size from 959 400m^2 to 10 000m^2 (1066 to 12 pixels) and were subsequently converted to a point dataset representing the individual pixels. The percentage of the training areas varied, depending on the different material types. Thick till had the most training areas at 44% and organics at 1% (Figure 3.3).

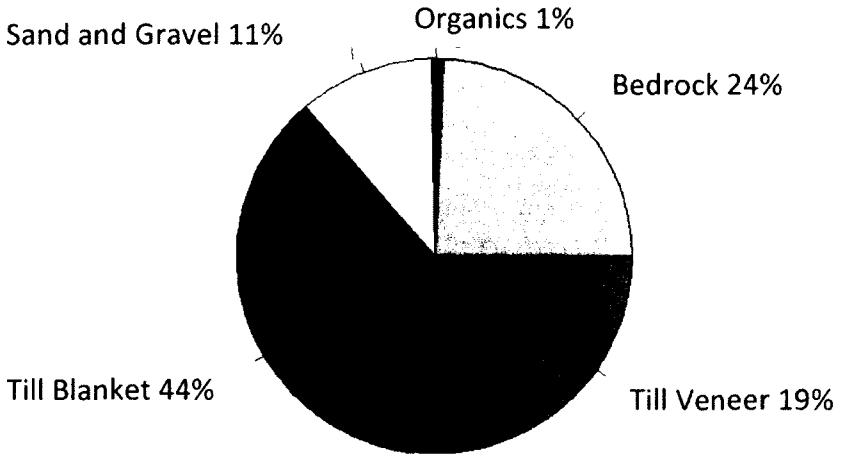


Figure 3.3: Pie Chart Showing Training Area Percentages

Table 3.2: The Number of polygons and pixels used for training.

	Class	Polygons	Pixels	Percent of Total Training
1	Bedrock	21	3452	24.4
2	Till blanket (thick)	23	6247	44.2
3	Till veneer (thin)	14	2755	19.5
4	Organics	7	142	1.07
5	Sand and Gravel	25	1552	11.0

3.3.3 Out of Bag Error and Tree Selection

The number of selected trees must be large enough for the *Strong Law of Large Numbers* to be applicable (Breiman, 2001). The *Strong Law of Large Numbers* applies when the averages of a sequence (number of trees) tends towards a constant value (Feller, 1968). When researchers select enough trees, the law applies, and statistical overfitting and classification bias is minimized (Breiman, 2001). Overfitting is not an issue because the model is stochastic by design and does not fit variables to produce optimum classification; instead, RF relies on a voting system. Furthermore, since the error rate decreases as the number of trees increases, OOB will overestimate any errors when completing cross-validation until the law applies (Breiman, 2001). Error is biased in spatial data using training samples pixels in close proximity such as in sample polygons (due to spatial autocorrelation) but is still used to assess the adequate number of trees for the model (Breiman, 2001). However, the specifics on tree selection are unclear. Some studies have used as little as 100 trees (Breiman, 2001; Pal, 2005), while others have used as many as 1000 (Sesnie et al., 2008). There are limitations on tree selection due to hardware configurations, such as available memory, and test set convergence since overfitting is not an issue (if you select enough trees). Studies indicate test set convergence occurs when the OOB error stabilizes with the addition of trees (Breiman, 2001). The number of trees used in RF was increased from 250 to 5000 by intervals of 250. Each iteration resulted in an OOB error estimate to test for model stability and if the *Strong Law of Large Numbers* would apply (Figure 3.4).

3.3.4 Variable Selection and Accuracy Assessment

Several recent studies from other non-remote sensing disciplines such as bioinformatics and machine learning have demonstrated the intricacies of variable importance stability and the bias during cross-validation (Calle & Urrea, 2011; Strobl et al., 2007). Their results imply that alternative sampling (such as sampling without replacement) is required in order to provide an unbiased variable selection approach (Strobl et al., 2007). Other researchers have maintained that the MDG can be used instead of MDA for variable selection (Calle & Urrea, 2011). This study manages RF instability by using iterative training and validation subsamples. By using iterative subsamples in RF cross-validations, a single permutation will not greatly impact the overall accuracy or the variable importance. In turn, this will remove the impact of outliers and bias in the training data. Subsampling occurs using training regions (polygons). Consequently, studies sample by polygons selecting regions for training and validation rather than individual pixels to manage spatial autocorrelation (Legendre, 1993). Subsamples of interpreted polygons were randomly split into training and validation across 100 iterations. The stochastic implementation helps mitigate bias in the cross-validation, allows for all of the training data to be assessed for importance, and provides an assessment of the stability of the model.

Variable selection is essential for selecting the approprate classification layers from the 78 data layers. Reducing the number of variables to predict the model can improve model predictions and overall accuracy, provide better interpretation of variable meaning, and/or help researchers understand layers because the model is

simpler (Andersen & Bro, 2010). Andersen and Bro (2010) indicate that the optimal method for selection is to try all possible combinations of variables and select the best results. However, testing all combinations becomes prohibitive, particularly as variables increase in number (>50) (Andersen & Bro, 2010). Consequently, this study initially uses all of the variables and then uses the RF variable importance plot for variable selection. The MDG measurements facilitate data layer selection because MDA should not be used because of stability issues during cross-validation (Calle & Urrea, 2011). The best predictive variables are automatically selected by averaging MDG values of the 100 iterations and selecting the best.

Through the described workflow, MSRFC selects the 15 most important variables and uses them in a second cross-validation process, this time using a fixed ratio of 70% and 30% training/validation, respectively. The 15 variables are selected using their boxplots then analyzed sequentially in order of decreasing MDG importance measures. Starting with the first one, on its own, and each time adding the next variable in the list to the group of input layers. This continues until the workflow adds all 15 layers to the analysis. Using the validation data sets, MSRFC computes change in accuracy resulting from the addition of each individual variable (relative to the previous iteration with n-1 variables). Only variables that result in a relative increase in accuracy for the final classification were. This validation process allows for a more refined assessment of predictive capability of these top 15 variables as an alternative to the prohibitively inefficient way of testing every possible permutation and combination.

The selected MDG variables were run through 50 RF iterations with different

percentages of training data starting at 25% and increasing by 5% intervals until 95%. Changing the propotions of training and validation is to demonstrate the sensitivity of training and validation datasets against the resulting output RF model. The resulting scatterplot compares overall accuracy as a function of the training and validation ratios. The resulting accuracy data show the variable suitability and strength of the model by examining the data range in the scatter plot. The slope of the scatter plot indicates how sensitive the classification is to the training and validation selection – this is potentially a function of the input variables used for classification. The researcher draws a line of best fit, and the r^2 is used to examine the model's suitability, prediction strength, and sensativity to inputs. The selected variables are suitable for classification when the line of best fit has a low r^2 value and a smaller range in overall accuracy. A low r^2 is desired because a high r^2 would indicate more sensitivity in the model to additional training data, indicated by an increase in overall accuracy as data is added. A horizontal (and thus uncorrelated line indicates a decreased sensitivity of the classifier to the training data and hence less model overfitting related to excessive numbers of predictor variables. These variable selection steps must be repeated until an adequate model is produced. This graph is used to estimate overall accuracy of the final RF model and associated classifcation

3.3.5 Classification

The final classification outputs contain more than just the model prediction maps. MSRFC outputs a set of powerful diagnostic plots that allow the analyst to assess model performance and robustness in a much more rigorous way than with other types

of RS workflows. The RF classifier uses all training data in classification. The results from variable selection and accuracy assessment are used to estimate certainty instead of cross-validation methods at final classification. MSRFC produces additional classification diagnostics including:

1. Maximum probability measuring percentage of votes used to assign a class.
2. Classifications maps that mask out pixels as unknown at different probability thresholds. E.g. a 90% Probability Classification, illustrates only pixels on a map with classes that are assigned with that probability or higher.
3. Probability difference, illustrating the difference in vote percentages between the first and second place classes.
4. Second prediction illustrating a classification using the second choice for the classifier
5. Probability sum measuring sum of vote percentage for first and second place classes.

Maximum probability is used to generate the final classification by hardening individual class percentages. Probability difference and sum illustrates the range of tree votes across classes to assess the predictive capacity of any individual pixel. The second highest vote percentage is recorded, so that the second place classification is known for future analysis. Finally, MSRFC produces individual class maps to indicate respective probabilities for each class.

3.4 Results and Discussion

The MSRFC requires specific steps that build on one another to establish a final classification. The results of OOB error are required for selection of the appropriate number of trees for further classification. Different combinations of variables are selected to measure variable importance, facilitate additional variable selection, and estimate overall accuracy. A final classification is introduced along with all of the additional of diagnostics produced during classification. This study discusses the suit of classification diagnostics produced by MSRFC with focus on how to gain much better understanding of the RF model. A thorough discussion of the use of these diagnostic models in conjunction with each other follows. Finally, there is a discussion of how to improve this research for future studies.

3.4.1 Out of Bag (OOB) Error and Tree Selection

As the number of trees increased, variability in OOB error estimates decreased. The range for all iterations was 1.2 – 1.7%, or 0.5 %. This variability is negligible and the decrease in OOB estimates as a function of increasing number of trees decreased as expected, because fewer trees reduces the strength of RF. The largest number that could be managed efficiently was 1500 trees, at this number of iterations the error was at 1.25%. beyond this number of trees the computational burden became excessive.

The low OOB error value is as expected; it does not consider spatial autocorrelation and likely exaggerates values accordingly. Cross-validation is required to assess predictive capacity of the model because RF cannot implement cross-validation in

the traditional remote sensing sense for OOB error. OOB error will presumably approach the cross validated accuracy as fewer variables are selected for classification. OOB error should approach cross validated accuracy because the model will be explaining less variance data sets that do not reflect the training data as created using cognitive processes. Therefore, as OOB error increases and begins to approach cross-validation, the reliability and stability of the model should follow. The OOB error should approach cross-validated accuracy but because OOB does not account for spatial auto-correlation, they should never meet.

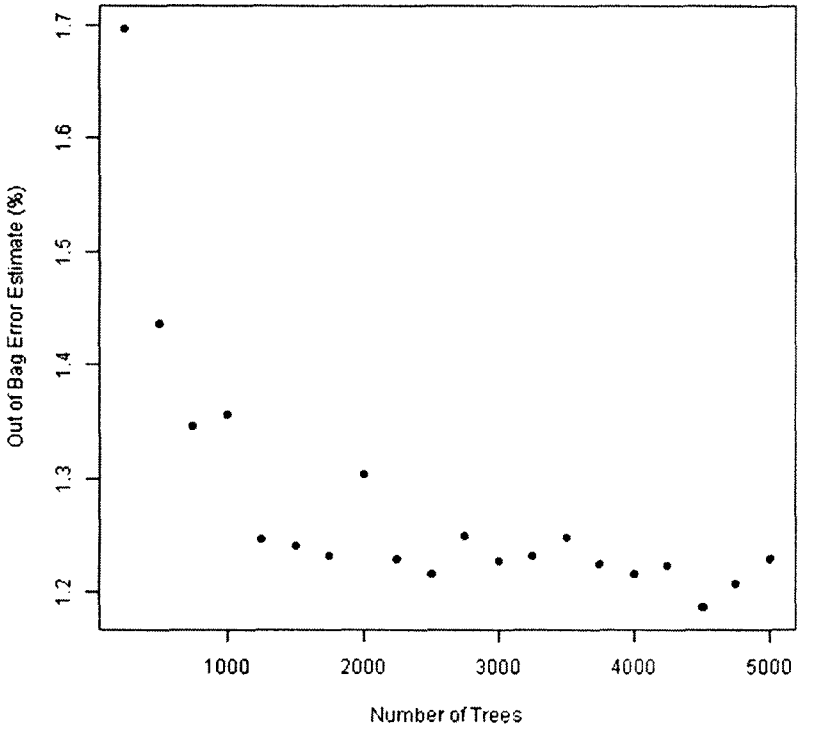


Figure 3.4: The Out of Bag (OOB) error estimates for a model vs. number of trees used.

3.4.2 Variable Selection and Accuracy Assessment

The 100 model iterations measure stability and predictive capacity of the model using all variables (Figure 3.5). The spread of data ranges from ~35% to 85%, which is indicates high variation in classification. The overall accuracy averages ~75%. Instabilities in the classification are likely to be the result of training polygons that capture sole examples of the class of interest on the landscape and unwanted relationships captured by the large amount of input layers. If cross-validation removes a polygon that captures a unique portion of a class for authentication, that manifestation on the landscape is lost. Consequently, the polygon left out for validation would likely result in a different class for that material. The model explains some variance even though the range of accuracies is high. These iterations also measure the spread of variable importance.

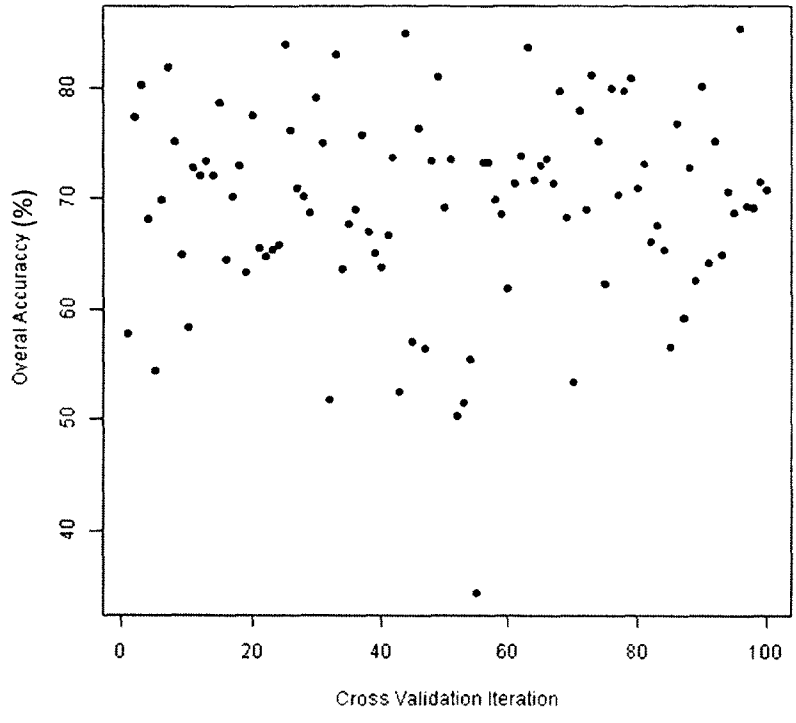


Figure 3.5: Overall accuracy using all predictor variables and 70% training and 30% Validation.

3.4.2.1 Variable Importance

The variable importance measures provide assistance in reducing data sources. The MDA and MDG analysis of the RF model provides two measures of overall variable importance (Figure 3.6 and Figure 3.7). The vertical range of the boxplots indicates spread, the black line specifies median, and the notches show 95% confidence intervals around the medians. One expects the MDA plot to have a higher range of values (particularly in correlated variables) than the MGD plot. The MDA plot also lists different variables as the most correlated variables in the model. The MDA variables are not used because of instability expected to occur during cross-validation that results in more similarly important variables, and a wider range within the boxplots. The algorithm

selects the top fifteen variables with the highest mean value from the MDG plot for further analysis.

The mean values from the MDG plot start with the two spectral ratios: 3/4 and 4/3 (Figure 3.7). These are directly correlated variables, so a similarity is expected. Next, the green component from the tasseled cap transformation has the highest mean, followed by band ratios 4/5, 5/4, 5/6. The importance of band ratios over raw bands is expected, considering acquisition late in the season, and hill shadow will influence raw spectral bands significantly. Finally, a number of DEM derivatives (mean60, max60, max30, and max15) that correspond to wider analysis windows and raw bands one and three are the next most important variables of the model. The DEM derivatives capture aspects in the training data that reflect regional relief differences.

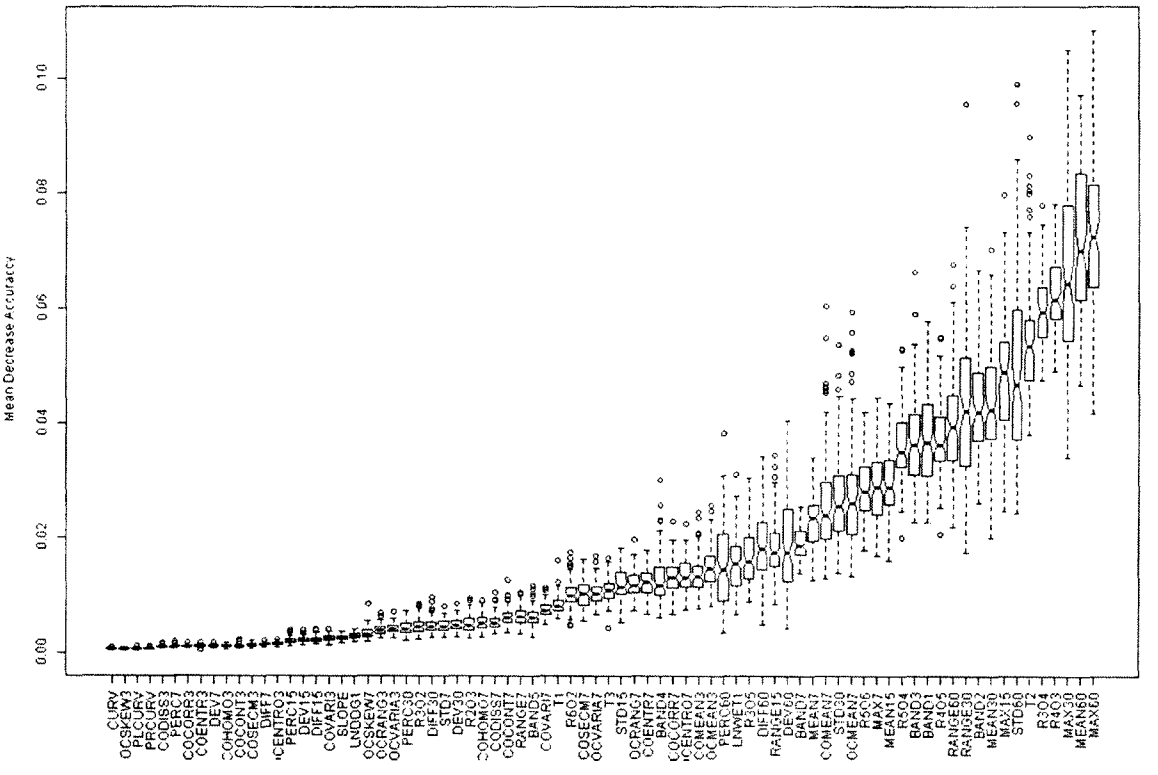


Figure 3.6: Mean decrease accuracy provided by 100 iterations of Random Forest with 70% training and 30% validation.

After the first 12 variables, there is sudden drop in the importance in the MDG plot and a decrease in the spread of the boxplots (Figure 3.7). This provides justification in selecting the first 15 variables for use in the next stage of the model. The reduced spread shows there is an overall reduction in variance, regardless of iteration which would be expected for variables of low importance. A significant reduction in importance after the first 12 variables indicates these variables are likely to provide all of the significant explanation of variance in the training data.

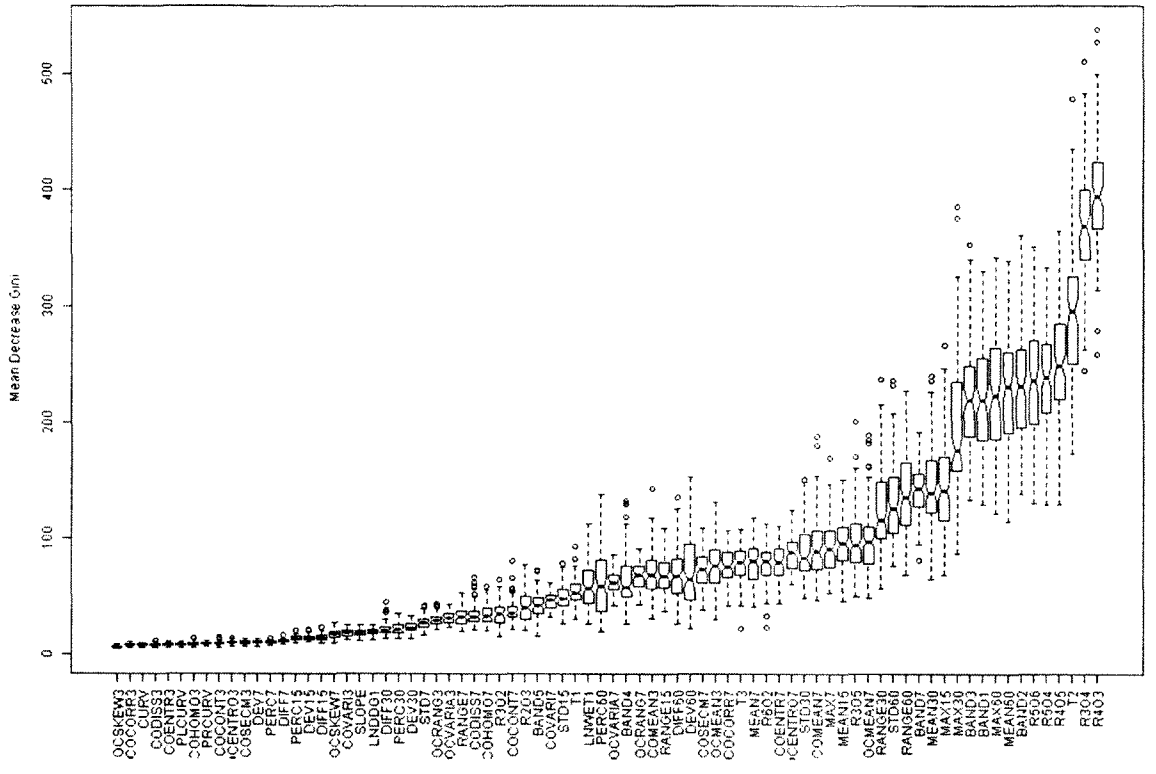


Figure 3.7: Mean Gini index provided by 100 iterations of Random Forest with 70% training and 30% validation.

3.4.2.2 Variable Selection

MSRFC adds variables one by one sequentially to the model cross validating at each step. This assesses the change in overall accuracy per variable added. MSRFC uses fifty iterations of cross-validation across the fifteen variables selected from the top MDG. Variables are added in order of importance (most to least) to assess the predictive capacity of the RF model since they would have the highest iteration impact. Iterations have fifty random cross-validations to compensate for the variability in the model. Boxplots represent the results in order of MDG importance from left to right (Figure 3.8).

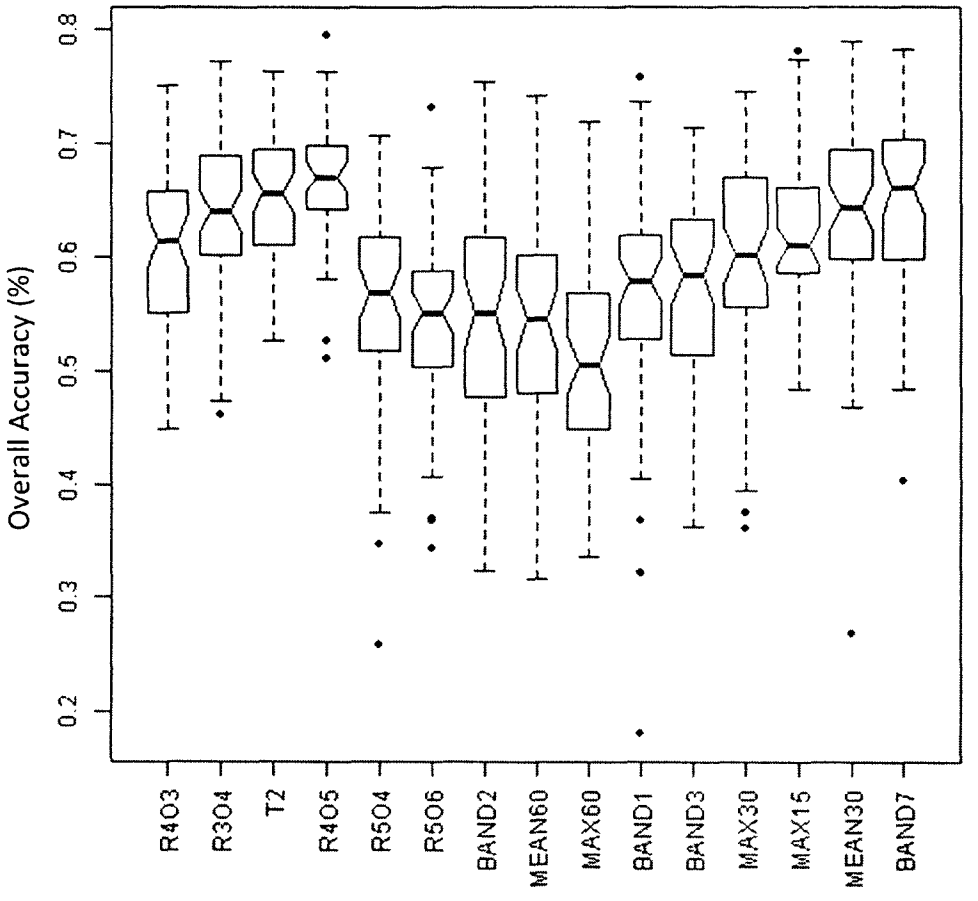


Figure 3.8: Overall accuracy from classification. MSRFC adds variables one by one to the model from left to right.

Overall accuracy tends to increase slightly with the addition of the first four variables (R4O3, R3O4, T2, R4O5; Figure 3.8). The increases are not necessarily significant, as shown by the notches on the boxplots. This is particularly clear as we examine R4O3 and the inverse R3O4 that have slight differences between the two but also contain a significant amount of overlap within the boxplots. Overlap between notches indicates that these medians are not significantly different. If R3O4 is ignored, the addition of T2 is a significant improvement, even though some correlation is

expected. As the next five variables are added to the model (R4O5, R5O6, BAND2, MEAN60, and MAX60; Figure 3.8), accuracy decreases. There is a slight accuracy increase as the final variables (BAND1, BAND3, MAX30, MAX15, and BAND7; Figure 3.8) are added. Outliers were present in almost all variables of this model (as indicated by the points), so one assumes some variability is unavoidable due to the training areas. It is possible that adding these variables to the model in a different order could influence the overall accuracy. However, RF uses $\sqrt[2]{p}$ input data layers at classification so, by design, it tests multiple combinations to vote for classification and measure variable importance. The only difference expected with a change in order for variable importance or accuracy would occur when we remove correlated variables before RF measured variable importance. Co-related variable removal could possibly increase the importance of the other variables (for example the removal of R3O4). However, even co-related variable removal is unlikely to change measurements significantly due to the random nature of selecting variables during classification. Therefore, testing multiple combinations of variables for selection is an iterative process that aims to increase the stability of the model while using the least number of variables possible. Every variable that caused an increase in mean accuracy is a possible variable for final classification.

3.4.2.3 Iterative Cross-Validation

The next step of the process examines the accuracy of the model using the variables selected. Training and validation data ranges from 25% to 95% at 5% increments with each increment using 50 iterations of stochastic subsampling. The first simulation uses 10 variables selected from Figure 3.8 as follows: R4O3, R3O4, T2, R4O5,

BAND1, BAND3, MAX30, MAX15, MEAN30, and BAND7. Each iteration is plotted and a line of line of best fit is drawn (Figure 3.9). There is a slight positive correlation and a moderately large spread of the data.

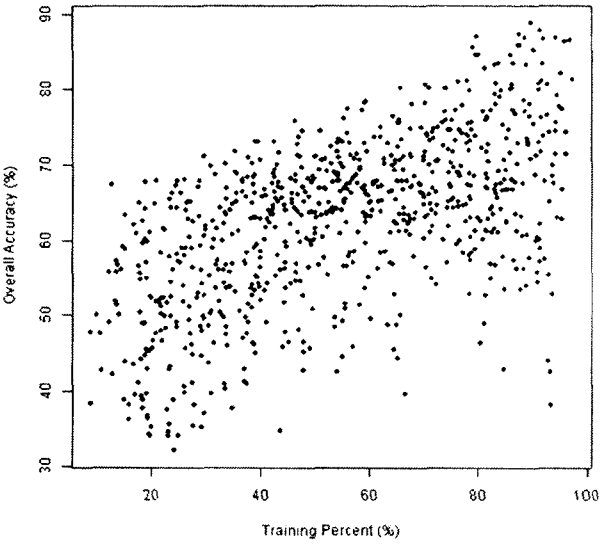


Figure 3.9: Overall accuracy of the RF models (with selected 10 variables) against the amount of training data used. Training and validation percentages use total number of pixels.

The range of accuracy values indicates high variability in the models (Figure 3.9). Usefulness of this model is doubtful, despite the significance of explaining overall variance. This is due to the range of accuracies and the line of best fit that increases as training data increases. The positive relationship indicates that the training data has not captured enough variance in the input data. If there was more training data provided, it is possible that the relationship would disappear and have reduced range of accuracy values. It is also possible that these variables may be capturing important variance, but there is not enough data to address the variability. Other variables, such as texture, are not shown as important on the landscape even though they conceptually are linked to

understanding of morphology and spectral signature. This is likely caused by the training data not properly representing the class population and possible overfitting due to a larger number of input variables. Therefore, it is prudent to explore the reduction of variables in an attempt to reduce the range of accuracy results in the process.

The next step in the workflow led to the selection of the top four variables for the final classification. Since the first two variables are the inverse of each other (R3O4 and R4O3), R4O3 was ignored (Figure 3.10).

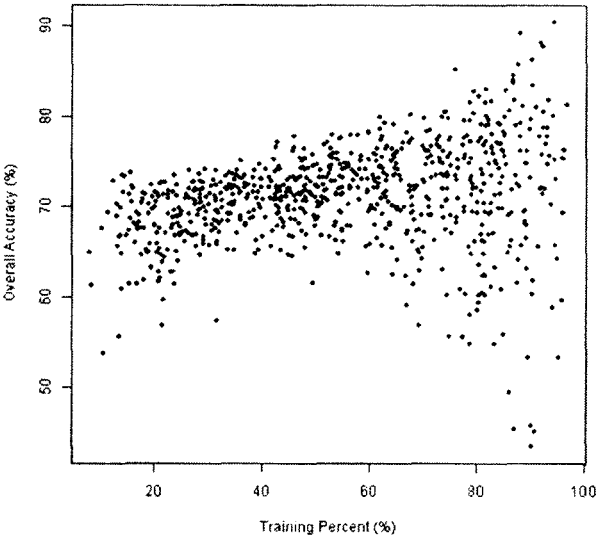


Figure 3.10: Overall Accuracy of the RF model (with selected band ratio 3/4, tasselled cap 2, and band ratio 4/5) against the amount of training data used. Training and validation percentages use total number of pixels.

The dependence of model accuracy on the proportion of training and validation data used to generate the models is reduced compared to Figure 3.9. There is less range in overall accuracy when training is below 80%. When a more parsimonious data selection is used (only three variables), overfitting is reduced and the slope of the line

decreases as well as the range in overall accuracy. These factors indicate more stability in the model, as well as explained variance in the training data aligning with captured predictor variable variance. However, variability in cross-validated accuracies increased substantially once the proportion of training data exceeds 80%. Any outlying data can offset the classification results, especially if it is inside the validation data, which only consisted of 20% of the samples. The results below that 80% threshold for training are stable. The stability of the model and reduced slope are two reasons to produce a final classified map using only the three selected variables for analysis. A more acceptable cross-validation accuracy is achieved using fewer variables. This is an ideal demonstration that simpler is better and the results demonstrate that fewer variables are leading to a more stable prediction.

3.4.3 Classification Results

Using the variable selection procedure to select the appropriate variables, a final iteration of RF is conducted using all of the training data, no training data is omitted for validation. The final classification uses the votes provided from the RF trees to produce a final prediction and uses the cross-validations from variable selection to estimate accuracy. Multiple products such as maximum probability, individual class probability, and probability sums also express the prediction quality and variability across the landscape. These metrics provide spatial assessment of the classification quality by class and as a whole. The final classification (Figure 3.11) displays the results of the MSRFC workflow with water masked out. This classification product, while useful, is only one data product provided by this workflow. Additional products produced in conjunction

with this classification provide a more comprehensive understanding of the performance of the model.

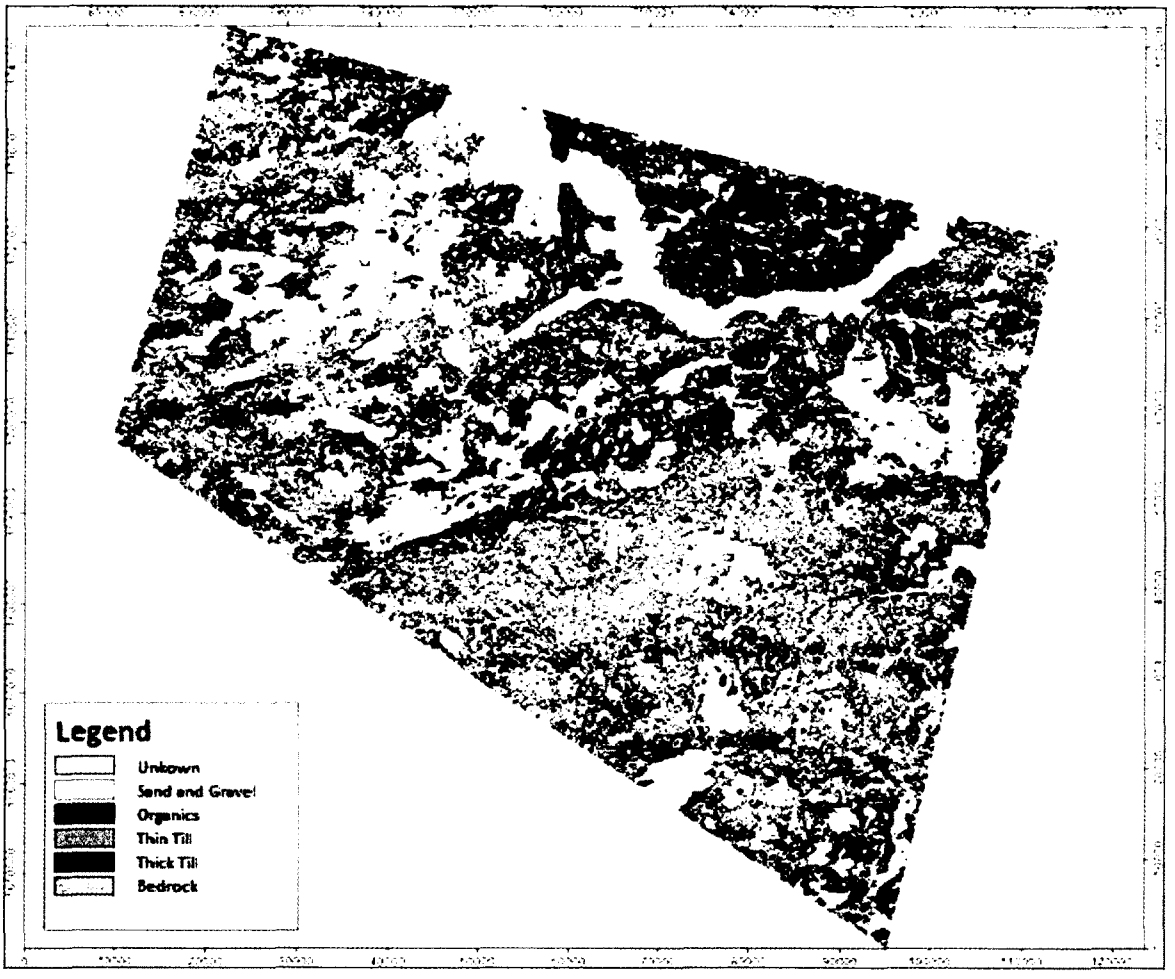


Figure 3.11: Final RF Classification Produced using maximum probability assigned from RF. Areas in white are water bodies, masked out before classification.

3.4.3.1 Prediction Probabilities

3.4.3.1.1 Maximum Probability

RF uses hard classification (classes are assigned absolute membership), which is similar to algorithms such as MLC. The maximum probability maps the percentage of votes used by RF to decide on the final classification. Figure 3.12 indicates the probability used for final classification, with red areas indicating a high probability and blue areas a low probability for classification. Some regions of high and low probability are visible in these images. The areas classified as till blanket show a consistently high probability trend (Figure 3.11 and Figure 3.12). Many areas of thick till in the western portions of the map have a much lower classification probability. Conversely, areas classified as bedrock and thin till in the centre of the study area have lower green-blue colours, corresponding to the lower classification probability. The organic, sand and gravel portions of the map tend to have a high classification probability but account for a relatively small amount of the landscape.

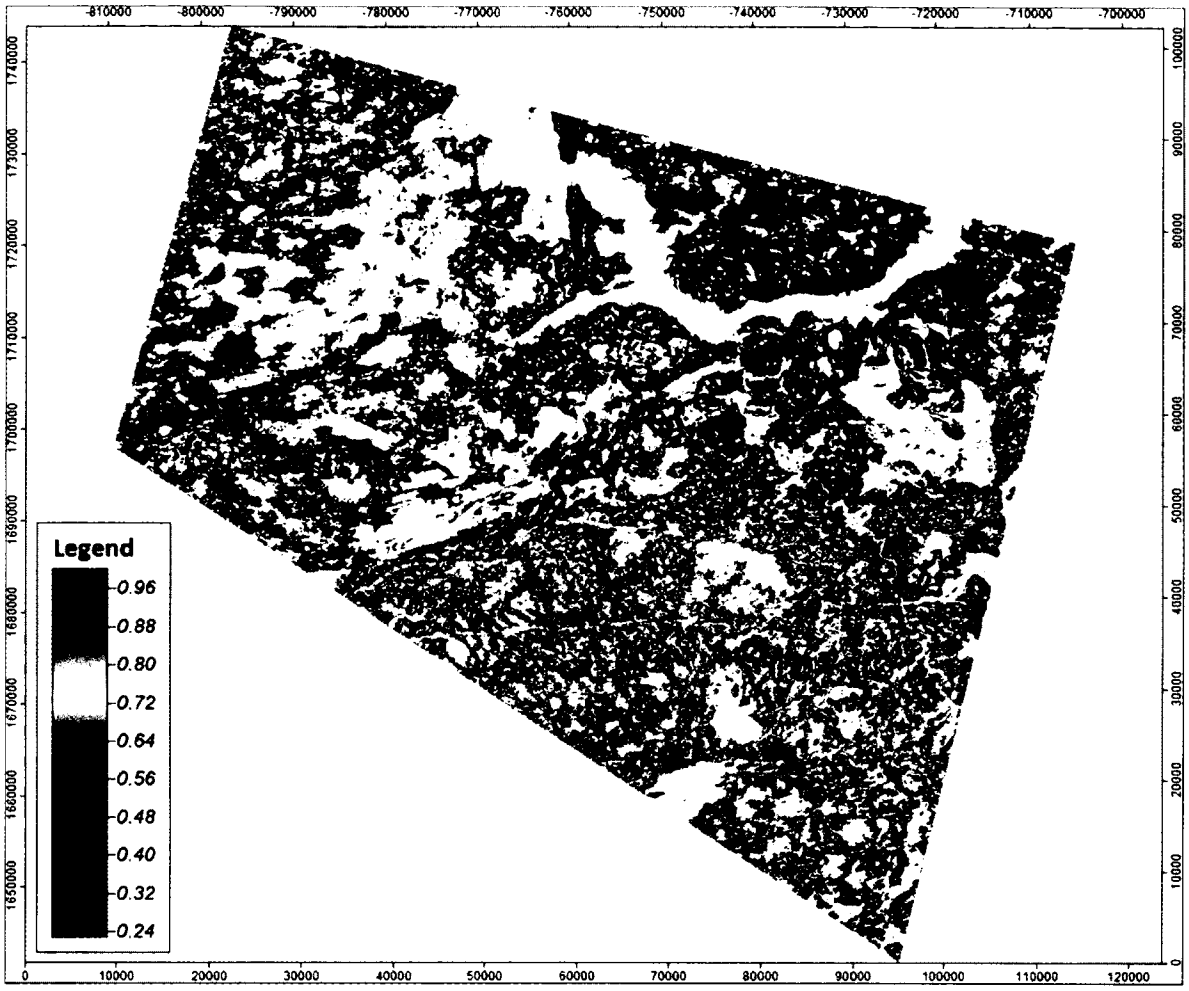


Figure 3.12: Probability used for classification

The spatial uncertainty assessment provided by the maximum probability maps is an important way to demonstrate why this remote sensing workflow is applicable for guiding additional mapping or field efforts. For example, applying the probability map to identify regions of uncertainty provides direction for further study. Applying additional training or using this map to focus additional field expeditions may both address problem areas and then be used in the model. Areas identified as having a particularly low probability for classification can have training sites added to them accordingly. This

would help the model capture unexplained variance in these regions. Summary results are difficult to gather from these images; however, probabilities appear to be autocorrelated leading to large regions of similarly high or low probabilities. While overall probability and classification capacity is gathered easily from this map, understanding the class-to-class confusions is more difficult and cannot be gleaned from the maximum probability map alone.

Summary classification results are required in order to understand overall performance of the model. Land area summary pie chart (Figure 3.13) is not unique on its own but useful when used in conjunction with class probability, which summarizes the classification probability by class in boxplots (Figure 3.14). Certain classes have high classification probabilities whereas others perform poorly. Causes of these differences may be systematic error (bias) in the model. The RF model should be explored for class related bias. It is possible that the RF model preferentially chooses some classes based on these input data or that these classes (because of the large quantity of inputs) capture the most diverse spectral signatures. Therefore, in order to properly explore this issue a comprehensive experiment would have to address both spectral diversity of classes as well as training quantity of classes.

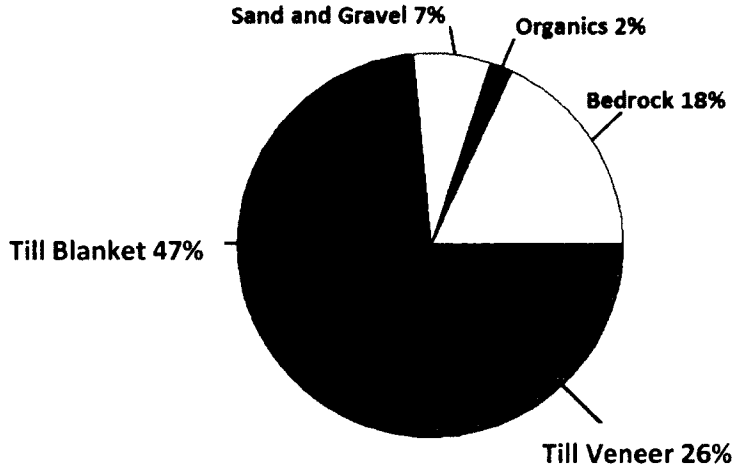


Figure 3.13: Pie chart indicating which percent each material type makes up on the final classification

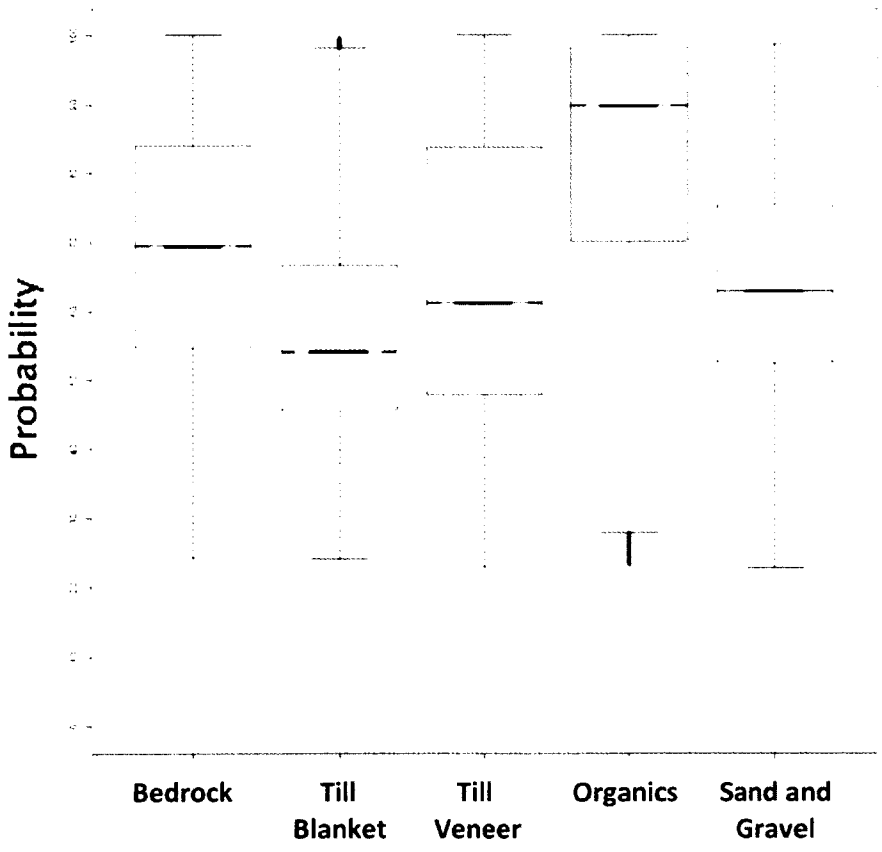


Figure 3.14: Boxplots of the probability used to make the final classification.

Thick till has the lowest probability (~55 %) but it accounts for the highest land area (~45%) in the study area. Furthermore, maximum probability and the classification indicate thick till is in high probability areas. This does not correspond to boxplots of probability that indicate thick till as one of the lowest probability classes. This association is unexpected because areas on the map that contain high probability classification also correspond to areas of thick till – this is apparent close to the water bodies in the centre and northern parts of the map (Figure 3.11 and Figure 3.12). This unexpected association suggests possible bias in the model and its classification of uncertain pixels. The size of training data used for classification could be the cause of bias in the RF model and may translate to the final classification.

Thin till is the second most common map unit at ~25% of the map area and has a ~62 % classification probability. The second most prevalent class corresponds to the second lowest probability during classification as well. Meaning most of the areas classified as thin till were possibly confused with other classes. Conversely, lower probability of this class is expected, given its transitional nature between thick till and bedrock.

Sand and gravel, bedrock, and organics account for the last three classes of the map. Sand and gravel has a ~65% probability of classification and accounts for ~7% of the map area. Bedrock has approximately a ~70 % probability of classification and accounts for ~18% of the final map area. Finally, organics has ~90 % probability of classification but only covers ~2% of the map area. The RF classification could be improved by balancing the proportions of areas and distribution of training data across

the study area. Results suggest that training data and expected outcomes should have balanced proportions in order to weight them properly within the model.

3.4.3.1.2 Classification at Probability Thresholds

MSRFC describes prediction probability two ways, using a cumulative graph of land area coverage vs. percent certainty (Figure 3.15) and mapping classes using the probabilities assigned by pixel (Figure 3.17, Figure 3.19, and Figure 3.21). Figure 3.15 illustrates the summary of mapping confidence as probability of classification increase. Using this graph one could approach classifications iteratively by addressing low probability areas for adding additional training data and assessing the result by comparing this figure from each iteration. Notice the linear negative trend showing a steady decline as we increase our probability percentages.

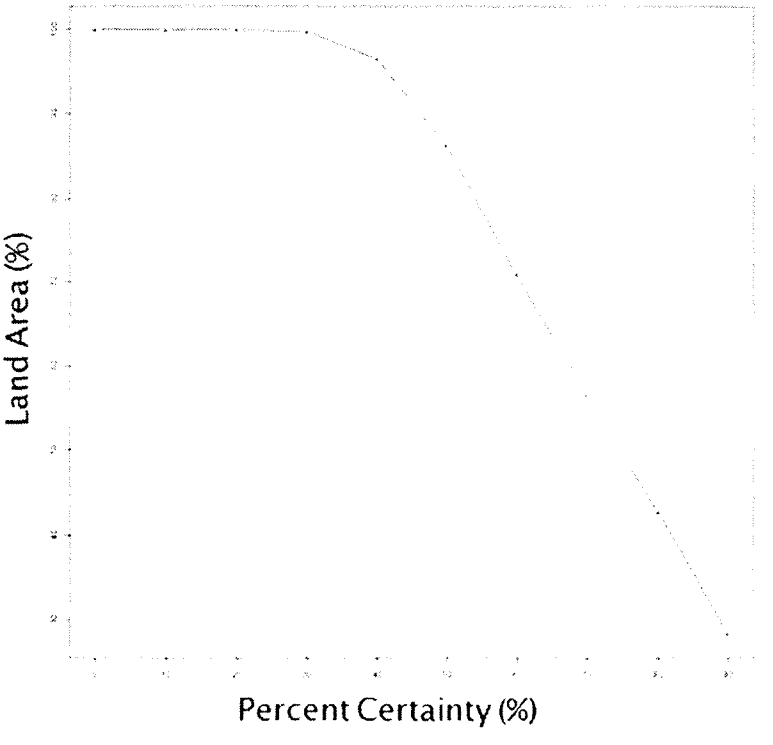


Figure 3.15: Percent certainty vs. land area percent covered.

Using maximum probability for individual pixels, MSRFC produces different maps at increasing levels of certainty (Figure 3.16 to Figure 3.21). Three illustrative images with classification probabilities of 30%, 60%, and 90%, respectively, are presented. The respective pie charts indicate the percentage of each class with pixels outside of the probability threshold excluded. These maps illustrate spatial extent of probability as the threshold is increased, examining the effect on overall classification results.

At 30 % classification probability (Figure 3.16 and Figure 3.17), the majority of the map is classified. The 30 % threshold is useful for identifying areas of highest confusion between pixels because unclassified pixels are <2% as indicated by Figure 3.15. Unclassified pixels at 30% have limited explained variance from the predictor variables or the training data and should be examined for causes.

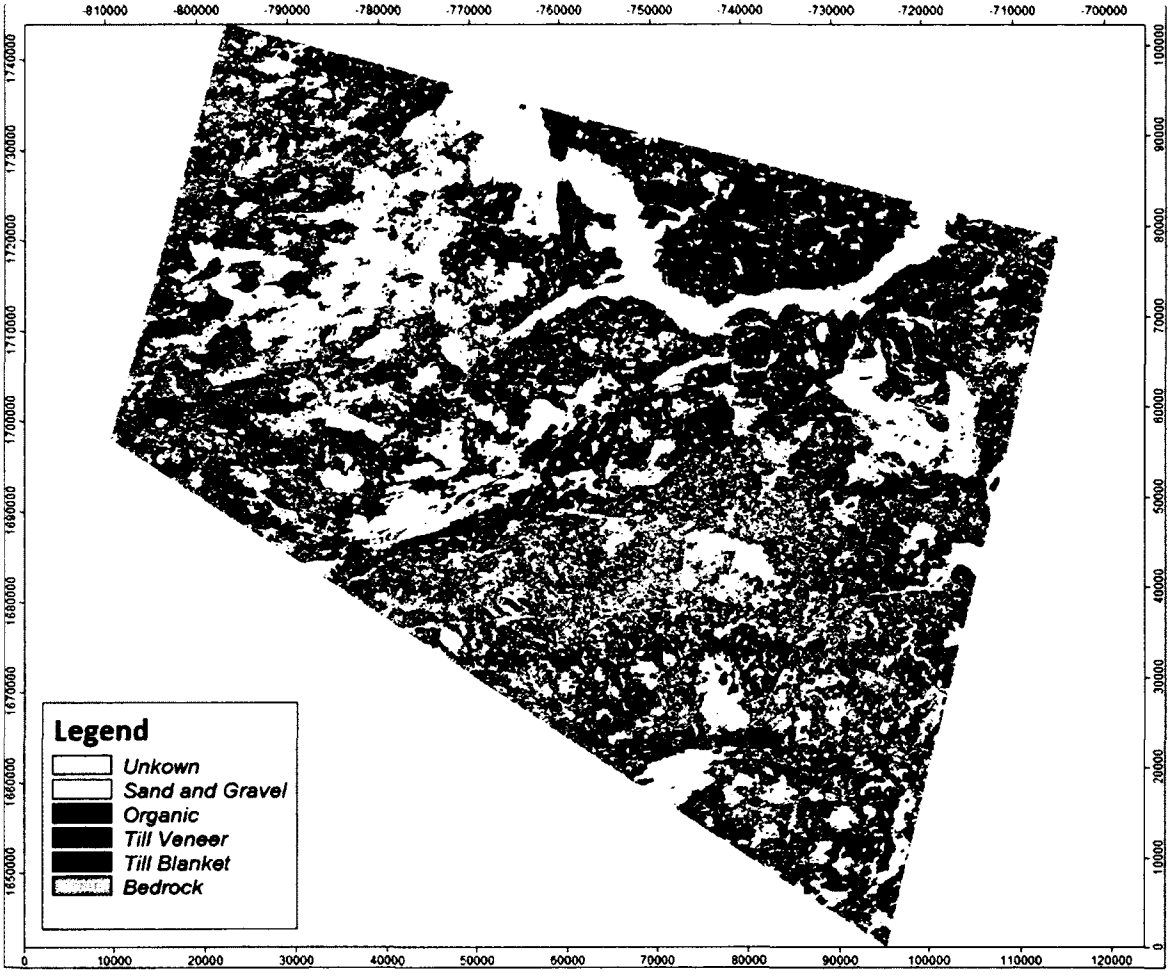


Figure 3.16: Classification at minimum 30% probabilitywith map coverage >98% of the study area.

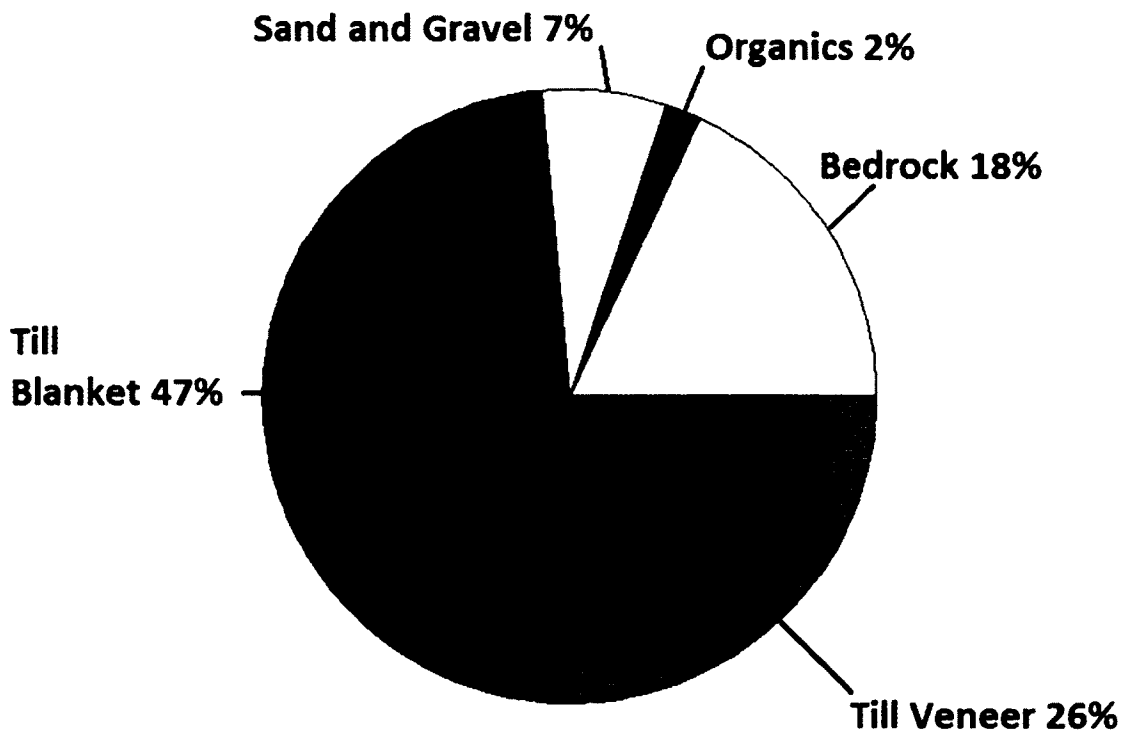


Figure 3.17: Pie chart of classification at minimum 30% probability.

The map for 60 % probability (Figure 3.18 and Figure 3.19) illustrates larger regions of uncertainty (unmapped, grey area) within the image. Uncertainty is most apparent in transition zones of bedrock and thick till in the southeast portion of the map. The map excludes lower probability till classes and regions that surround bedrock. The thick till class is largely present, resulting in a higher overall percentage of the image. The bedrock classes remain stable in spatial and overall percentage. Therefore, the most apparent problems are in the transition regions between bedrock to thick till. This transition zone causes significant confusion since it appears similar to thick till and bedrock, depending on genesis.

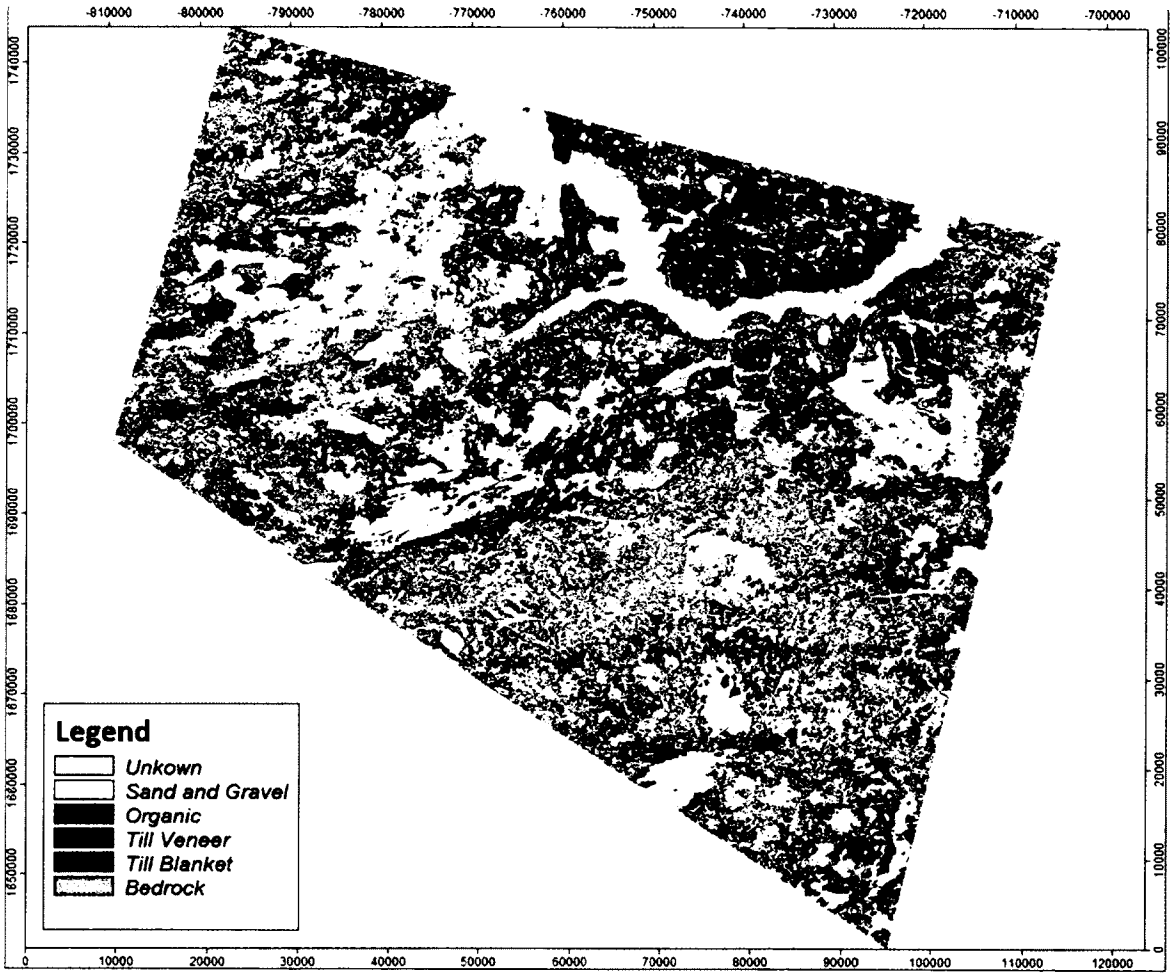


Figure 3.18: Classification at minimum 60% probability with map coverage at ~70 % of the study area.

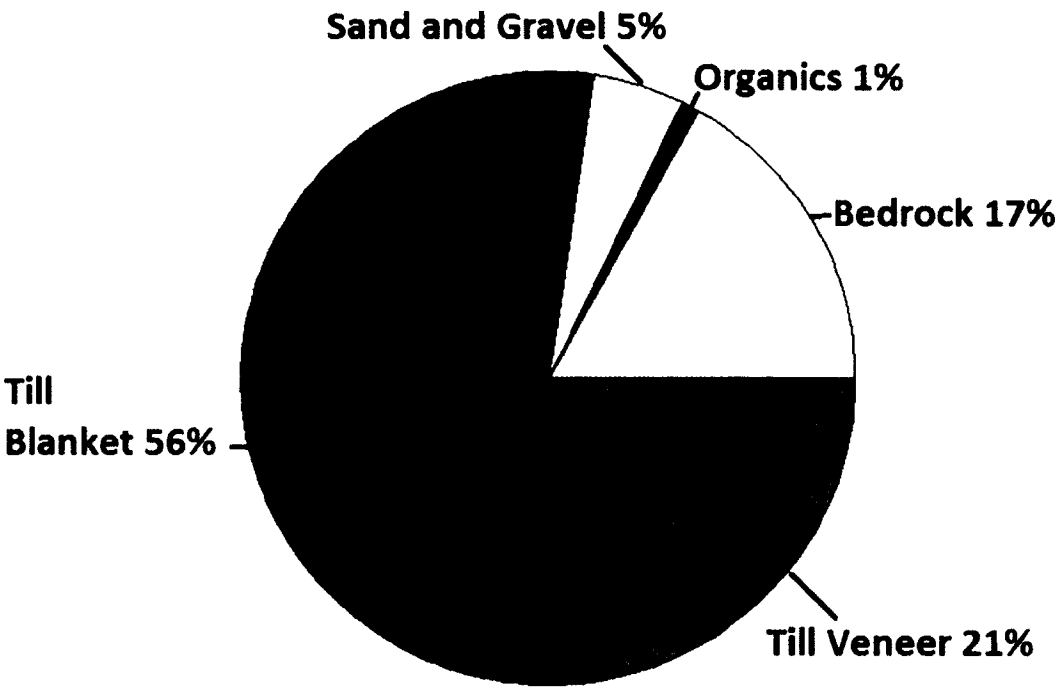


Figure 3.19: Pie chart of classification at minimum 60% probability.

The 90 % probability map (Figure 3.20 and Figure 3.21) has only ~25% map coverage and identifies unknown regions within the study area. The map illustrates significant portions of sand and gravel, bedrock, and most thin till regions as unknown. Thick till is the only persistent class on the map. The pie chart rounds the organic class to 0% in the pie chart. This is intriguing, since it is the class with the highest probability in the boxplots. Since the thick till class accounts for a high percentage of the training data, it is possible that RF biases areas of low certainty toward this class. This means that while some areas of thick till are evident (to the north and northeast of the map), RF classifies the majority of the thick till class with a low probability. This anomaly is a

strong indication that proportions of training areas affect outcomes, and that summary probability plots are not enough to understand the spatial accuracy of classes.

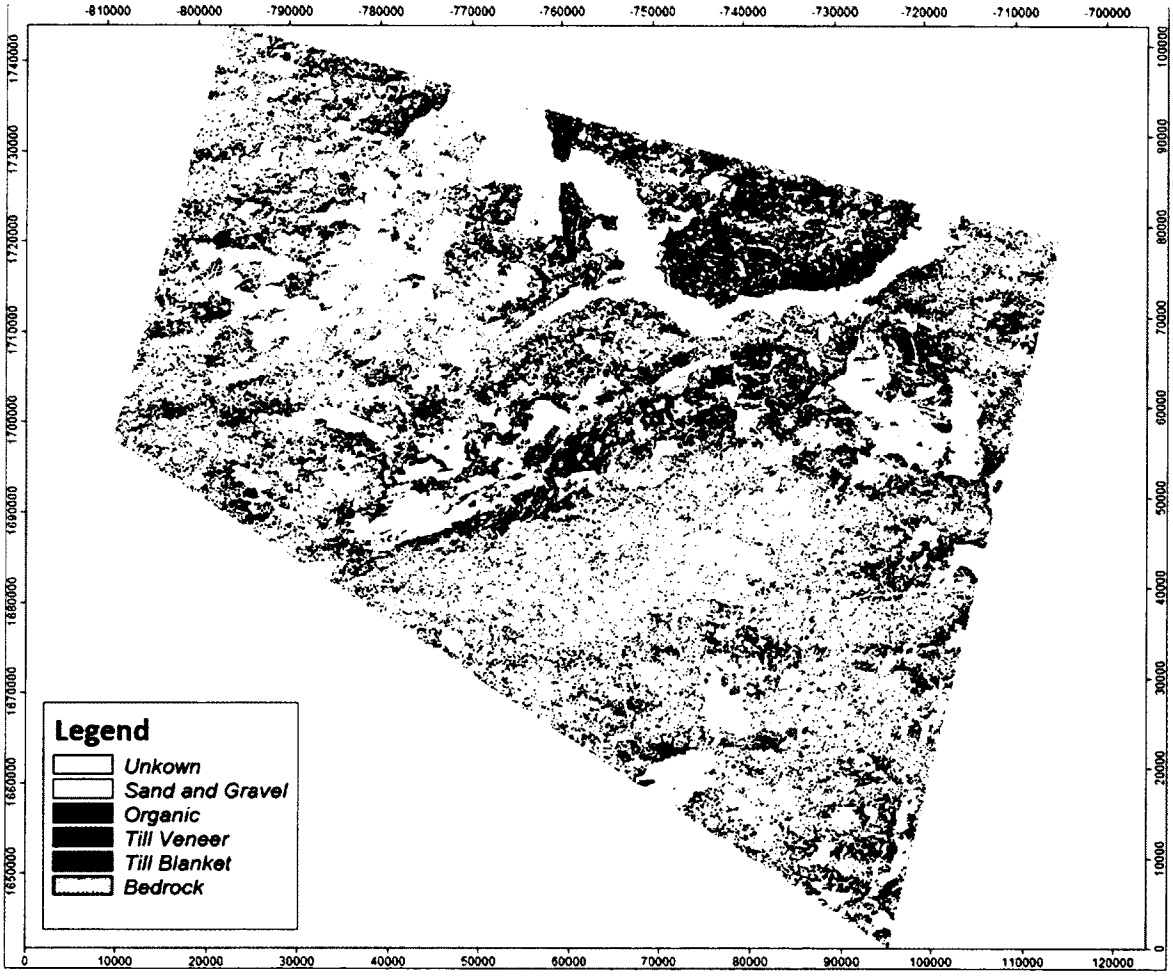


Figure 3.20: Classification at minimum 90% probability with map coverage at ~25 % of the study area.

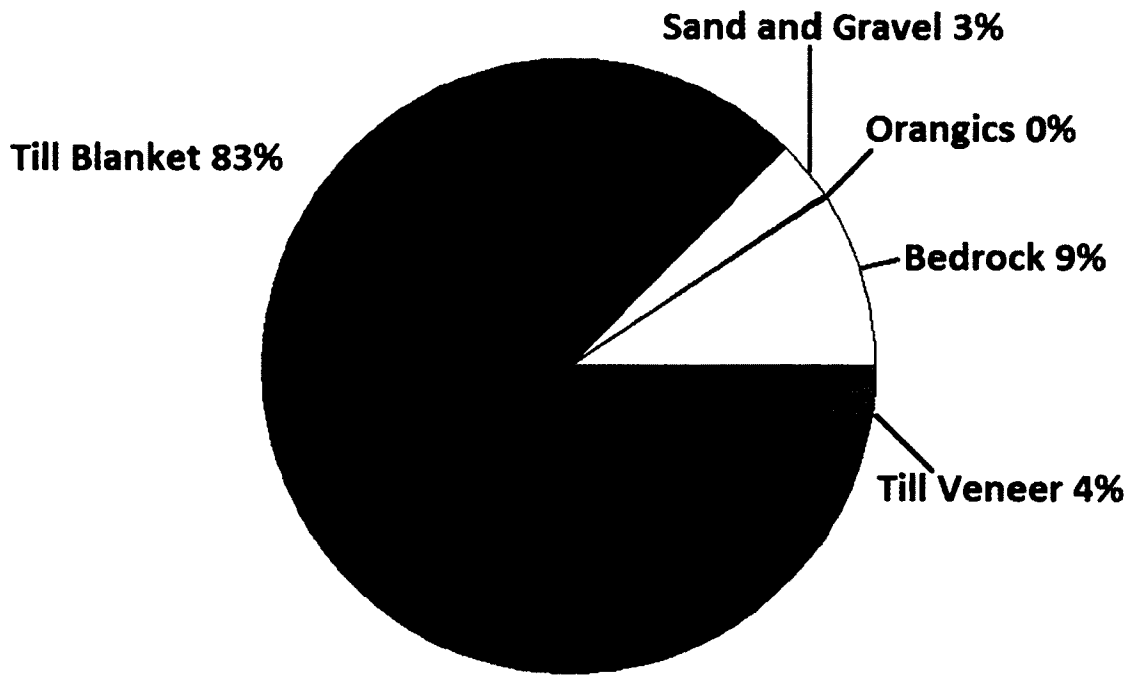


Figure 3.21: Pie chart of classification at minimum 90% probability.

3.4.3.1.3 Second Prediction

Further metrics are required in order to analyze the map for stability of classification and alternative predictions. MSRFC stores RF votes for every class at every pixel. Using the data, MSRFC gathers the second place probabilities to produce a map that represents all second place classes (Figure 3.23). Additionally, MSRFC produces additional maps representing probability difference and probability sum (Figure 3.24

and Figure 3.25). All additional maps illustrate confidence in the prediction, possible prediction alternatives, and level of confusion in pixels.

Mapping of the second prediction does not directly add value to this final classification (Figure 3.23). However, given some additional knowledge or expert analysis, one could reclassify low probability thin till pixels to a second prediction (and vice versa) for high probability organics that were not selected as the highest probability. The map exemplifies the uniform distribution of thin till as a common alternative for classification in the study area, particularly for areas of thick till and bedrock. Organic areas are also widespread in water-rich regions of thick till. Given more knowledge of the study area, one could decide on a final classification to add value to the final product. For example, one could use these metrics to identify problematic classes and regions to address creation of additional training areas or one could replace low confidence thick till classes with the organic class based on a probability threshold. These approaches would be based on the decisions of the analyst. These added metrics provide key insight into analysis, provide more information for expert analysis, and help with post classification assessment. This could foster discussion on improving and connecting physical understanding to decisions made in the model.

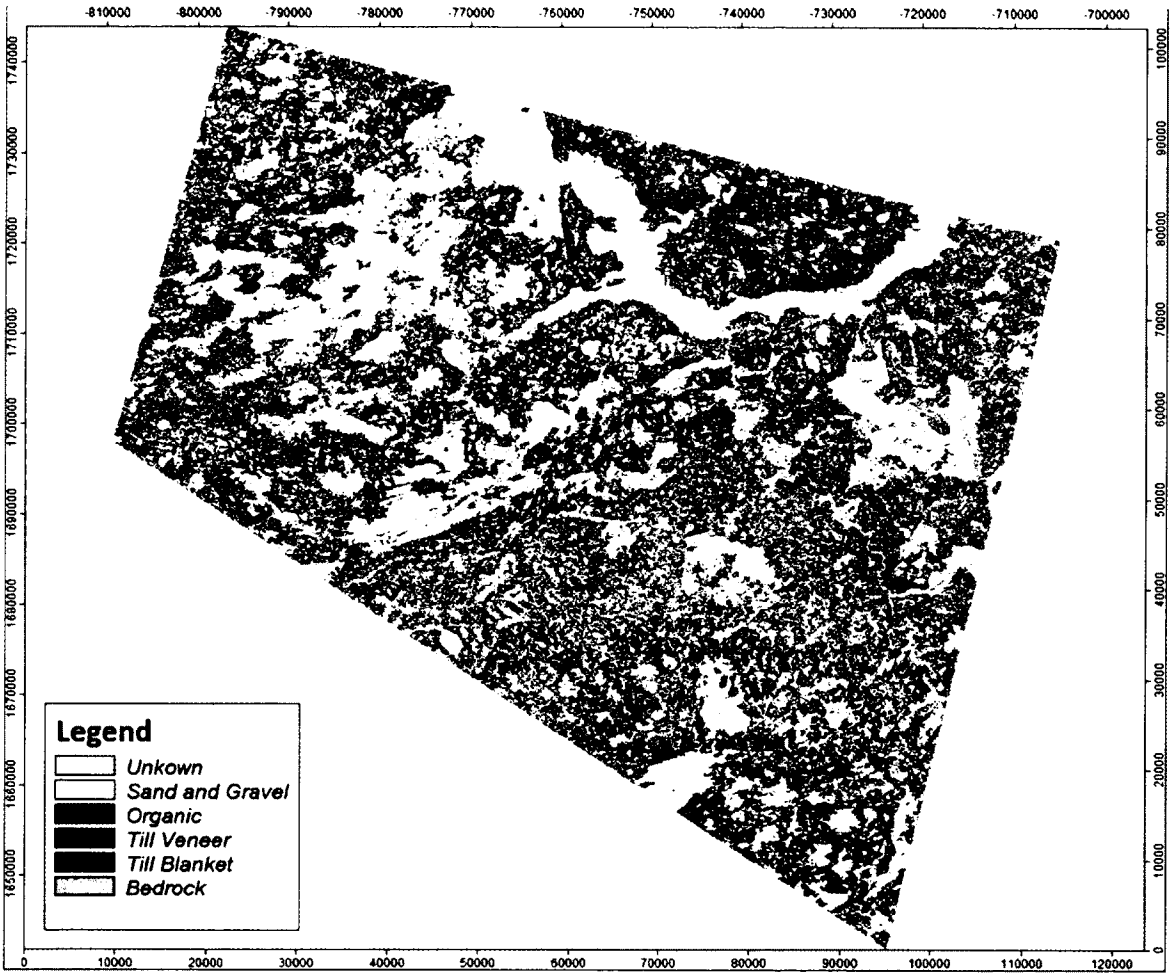


Figure 3.22: Second prediction map is generated by applying the second most probable class for classification. Each pixel represents the RF's alternative selection.

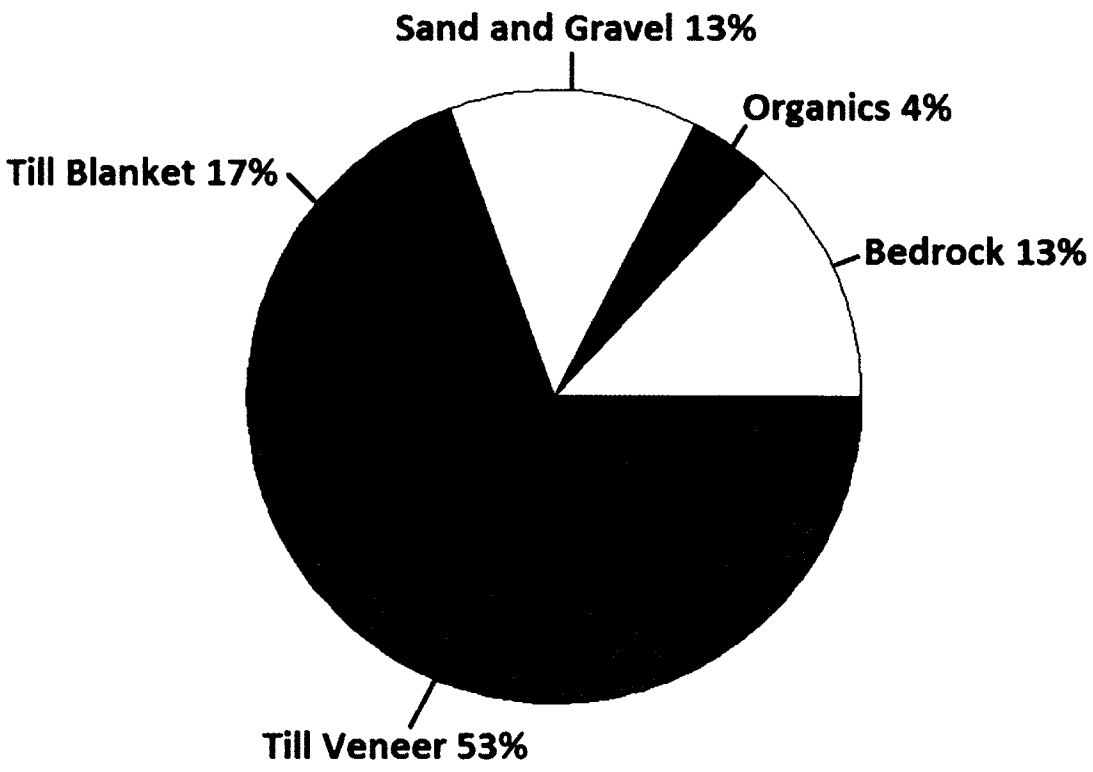


Figure 3.23: Second prediction pie chart

3.4.3.1.4 Probability Difference and Sum

The difference between the first two prediction class votes illustrates the difference in probabilities between the first and second place classes (Figure 3.24). Red pixels represent a high probability difference between first and second place classes and blue pixels represent a low difference between the first and second place probabilities. In circumstances where the range is much lower, it indicates high confusion between the first and second classes. The northern section of the map shows thick till and continues to classify areas with high differences. The sums of the first two prediction

classes illustrate the total amount of probability represented by the first and second place classes (Figure 3.25). Figure 3.25 displays the sum of the first two classes. Areas that show a low difference between the first and second class as well as a low probability sum could indicate the model is confused between all of the classes. Conversely, when there is a low probability difference and a high probability sum, most of the predictions are one of the two classes.

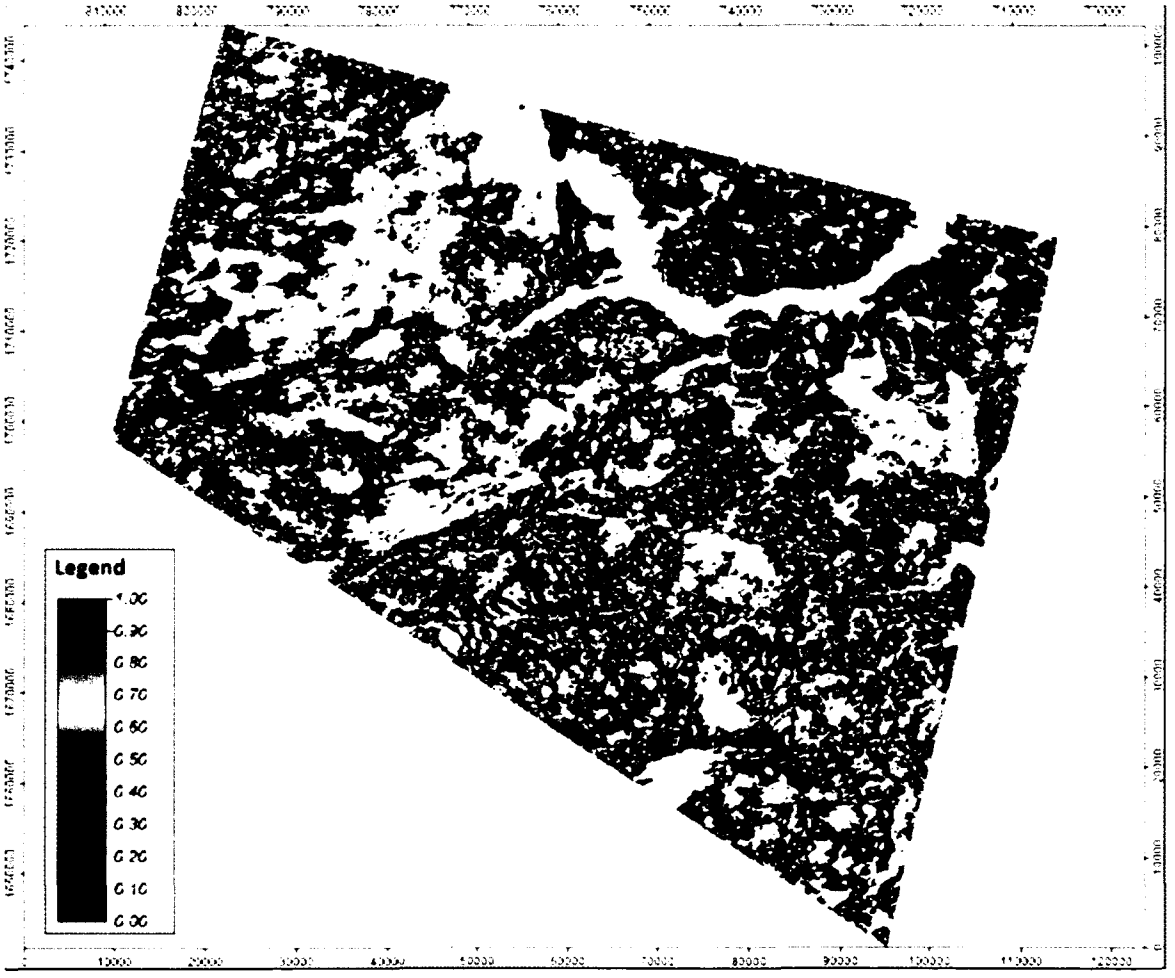


Figure 3.24: Probability difference between first and second place class.

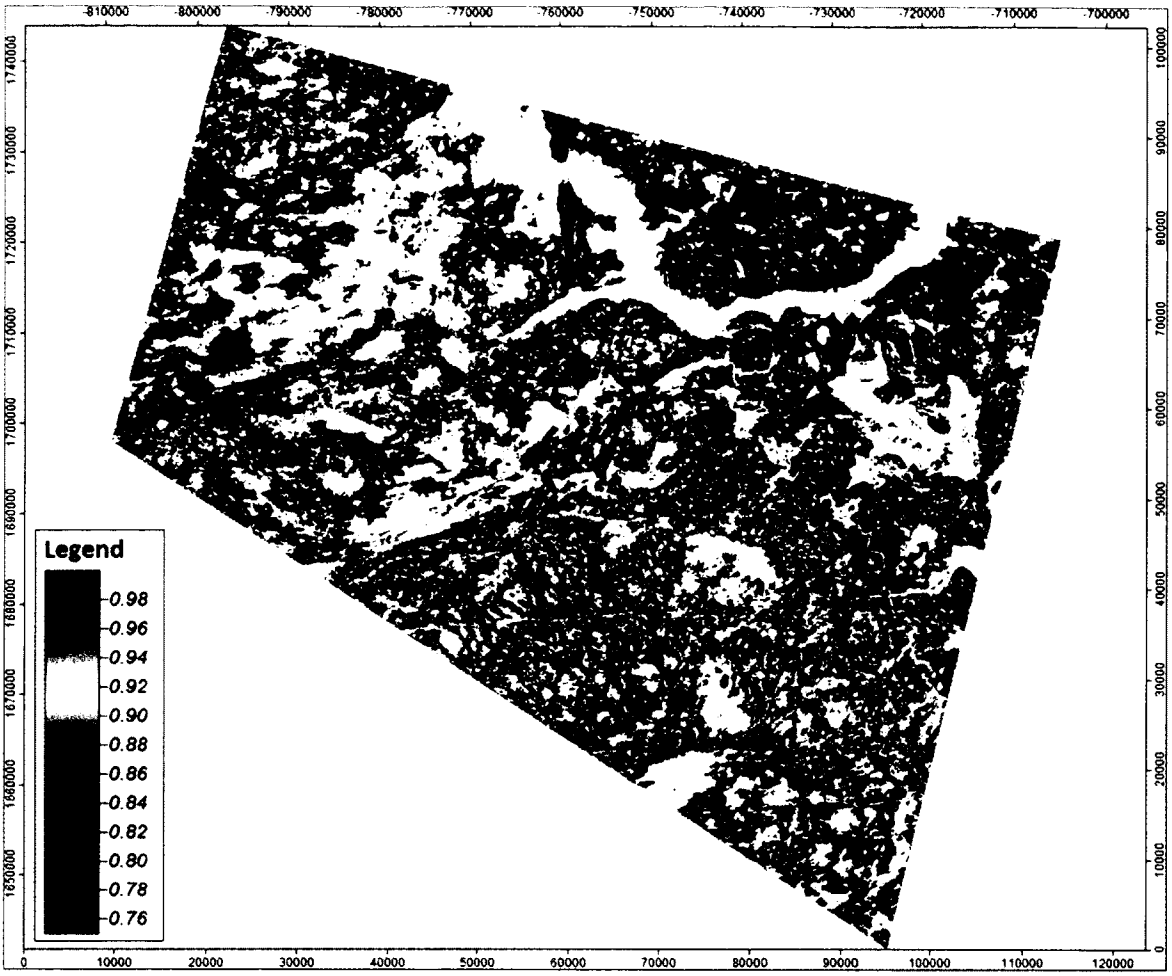


Figure 3.25: Probability sum of first and second class.

Areas in the center of the map associated with bedrock and thin till classes tend to have lower certainty. Persistent thick till areas indicate higher certainty regions, as outlined in the previous section. These maps help to clarify these areas of certainty. The maps demonstrate areas of confusion, which RF classifies as thick till. This is likely a result of the model, considering the high probability classification maps contain mostly thick till (even though it accounts for the lowest probability used in classification). MSRFC now demonstrates variability and confusion across the study region and perhaps can even aid in facilitating further research in ambiguous areas.

3.4.4 Future Direction

The use of surficial material maps requires expert knowledge and incorporation of diverse data sources. As data products become available, methodologies must exist in order to properly incorporate and assess their reliability and usefulness. The MSRFC is particularly helpful as a way to evaluate new data products for surficial material mapping. That is, analysts can use MSRFC to help assess data sources as they become available. Furthermore, as new training data or field information is gathered, the parametric assumptions are not required, which makes it easier to update classification products.

Parametric methodologies split non-parametric data into different spectral classes, even if they are conceptually part of the same class (Harris et al., 2012). MSRFC does not require spectral splits of the data in order to maximize separability, since multi-modal distributions will not be an issue in classification. Spectral separability between classes is a necessity; however, it should not require normal distributions of the same classes when other classifiers are available.

Time is required to compensate for spectral balancing issues between scenes taken at different phenological or meteorological intervals. Classification approaches must manage seasonality issues by pre-processing the datasets, or by using separate classification algorithms for each area. Future studies should explore this interest across a study area with misbalanced scenes. Ideally, such exploration will examine a study area with more than one misbalance seam.

The next steps of study should encompass handling large study area problems,

such as scene misbalancing and multi-modal classes. Additionally, research should be conducted to assess balancing issues present at multiple scales. Finally, multi-model classes must also be present in order statistically capture subtle differences between classes in the landscape to produce a classification. How RF handles these issues will further demonstrate the potential of the RF model in remote sensing classification projects.

3.5 Conclusion

The RF model is a powerful algorithm that classifies noisy data for training areas that violate parametric assumptions associated with common Gaussian classification models (e.g. MLC). This study also demonstrated additional diagnostic information resulting from the MSRFC. Additional stochastic steps within the workflow can manage many of the issues inherent in remote sensing application of RF. MSRFC created additional pre-classification and post-classification products to provide insight into the spatial dynamics of classifiers on the landscape. Finally, the MSRFC post classification products can be used to integrate expert knowledge to further refine input data sources and training data. These factors make the RF model the best-suited classifier to meet the objectives of remote predictive mapping of surficial materials in northern Canada. All of the value added products provide analysis with a robust set of additional diagnostics and post classification tools.

3.5.1 Pre-Classification

The introduced workflow manages model stability by using additional stochastic training and validation steps. This gauges the quality of training data. The workflow properly identifies instabilities, allowing for a thorough assessment of the model. Therefore, researcher can use expert judgement to compensate for errors. High variations in importance are indicative of variables that are significant, depending on the circumstance. Conversely, low variability tends toward a consistently high or low importance, depending on its overall mean. These steps are crucial to understanding variable importance and the impact of individual variables on the model.

Iterative addition of variables during classification provides a cross-validated understanding of importance. Cross-validation steps quantify useful variables for model prediction within these objectives, rather than simply explaining variance that may be inapplicable. DEM derivatives are significant to the model (shown by Figure 3.7); however, cross-validation indicates they are not useful for improving model accuracy (Figure 3.8). Therefore, these factors have a direct influence on understanding physical differences of classes.

3.5.2 Expert Incorporation

Glacial geology experts can investigate the implications of different variables provided for future study. For example, variance from DEM derivative classes indicates a relationship between the variables and the training sites. The classification scheme is likely capturing a variance that interpreters have not accounted for in the training data.

Different Landsat bands (particularly band ratios) were moderately successful compared to others variables in this study. Therefore, the ability to identify mositure and vegetation is significant to understanding the surficial materials. Furthermore, the model quantifies regions of low probability, which researchers may use to identify classes and areas for future study.

3.5.3 Classifications

The final classification product is a part of a data library that contains many classifications and probability products. These additional products represent one component of the significant advances using this methodology. Class boxplots give the user understanding of the level of certainty the model has per class. Users gain understanding of the generated results when using boxplots in conjunction with probability classifications. Using the plots produced by the introduced workflow in conjunction with each other, this study could identify the anomalous result that thick till has the highest probability displayed using 90% thresholds and lowest probability when summarized using boxplots. Furthermore, researchers can make future comparisons by assessing the probability versus land area covered plots. Stochastic cross-validation gives insightful measures into overall accuracy and spread of the model. These measures allowed the user to use all training data to produce the final classification. However, the significance of training proportions became apparent as RF preferentially classified uncertain pixels as classes with more training data.

Chapter 4: Large Area Random Forest Comparisons to Maximum Likelihood Classification

4.1 Introduction

During the past decade, there has been an increase in studies engaged in large-area mapping applications focussed on northern Canada and the Arctic (Scambos et al., 2007). Therefore, there has been an increasing need to apply machine-based predictive mapping approaches to classify imagery through workflows that minimize the need for human intervention while providing usable information on the landscape. However, machine-based predictive approaches to surficial material mapping continue to face four main complications that must be managed: they must have a sufficiently large amount of training samples, genesis-related class descriptions, assumptions of normality, and a-priori knowledge of data distributions (Mountrakis et al., 2011). Complex classification algorithms such as Support Vector Machines (SVM) and neural networks aid in managing these issues, but many algorithms require excessive computational power to produce a final classification product (Breiman, 2001). Breiman (2001) used Random Forest (RF) as an alternative to other high-level processes due to its ability to match or outperform other methods in performance and accuracy. Further, chapter 3 demonstrated that RF can be used to generate additional products using the multi-stage random forest classification (MSRFC).

Maximum Likelihood Classification (MLC) is one of the most widely used classification algorithms (Conese, 1992; Grunsky et al., 2009; Kleinbaum & Klein, 2010).

Like all Gaussian classifiers, MLC is constrained by two key factors: first, training samples must fit a normal distribution (Gao, 2009); second, there is an incompatibility with combining auxiliary non-remote sensing data formats such as thematic map data (Gao, 2009). Other relevant factors that can affect classification include representativeness of training data with respect to the population, data scales, and a-priori knowledge of class proportions. Amplification of these issues occurs when seasonality and spectral balancing influence the study area.

Prior knowledge of study areas is difficult to obtain in predictive mapping exercises. In northern regions, for example, machine based predictive mapping is prioritized for areas where there is limited field information (Harris et al., 2012). It is usually only after the predictive mapping is complete that planned field excursions take place. These factors limit the possibility of using *a-priori* probabilities. Therefore, when interpreters produce training data through manual delineation, they tend to select the easily identifiable material examples because of the complexities involved in interpreting medium resolution multispectral imagery, which is often the only type of data available or feasible to use. This bias in training data selection can result in increased errors and misclassification because the purpose of obtaining training data is to obtain a statistical representation of the population.

Predictive maps created by MLC applications have resulted in confusion between classes. Grunsky et al. (2009) experienced this issue, finding significant confusion between thick and thin till, bedrock and boulder, and sand and gravel. Confusion between classes is difficult to resolve, since conceptual genesis does not imply

uniqueness in spectral reflectance or physical material types. Heterogeneity within a class, influenced by lithology, moisture regime, vegetation complexity, and season of acquisition, can further exaggerate these issues. In order to advance surficial material mapping activities into usable products, researchers must address class definition in relation to surface reflectance.

Scene balancing is a critical aspect of image mosaicking. The user usually stretches or scales one scene to match the other. This process continues as more scenes are added to a mosaic (Gao, 2009). However, spectral consistency can degrade when there are non-linear distortions or statistical anomalies (such as cloud and haze) in the image (Beaubien et al., 1999; Gao, 2009). Consequently, when users combine scenes from different seasons, all distortions are rarely accounted for (Olthof et al., 2005). Scene balancing issues are prevalent as study area boundaries often obliquely intersect the satellite acquisition paths.

4.2 Objectives

The traditional ways of handling bimodal classes and image mosaicking in northern surficial material mapping have many limitations. For example, multiple spectral responses for individual classes are common in surficial material mapping. Characteristics including, but not limited to, unique lithological structures, changing moisture regimes, phenological gradients, and topographic position violate normality, causing classes of the same material type to have multiple spectral signatures. Researchers must capture these complex classes appropriately in order to capture the

population.

Researchers assume the interpretation accounts for every possible manifestation of a material type in that landscape (representative sampling of population). Representative sampling, however, is rarely achieved because there is often a bias in the selection of training areas. When scenes are not balanced, artificial bimodal classes are more prevalent resulting in one material type having different spectral signatures across scene boundaries. In order to compensate for this, training areas have been interpreted using different datasets for each scene. Using different data sets is time consuming and operationally complex when study areas and classes become multifaceted. Therefore, the objectives of this section are as follows:

1) Apply the MSRFC workflow to a large area on Victoria Island in the Canadian North and compare results with traditional classification approaches.
2) Introduce an alternative approach to handling classification issues in northern surficial material mapping by incorporating a nominal data (category based) source into the RF classification procedure.
3) Compare an RF classification to both itself (by using and then omitting a nominal index data source) and the standard MLC.

4.3 Methodology

This study compares the MSRFC to the standard MLC classifier. The addition of nominal data to the predictor classes for MSRFC is explored as an attempt to manage spectral issues using a nominal data source (Figure 4.1). This study does not discuss in

detail all the advantages and additional metrics provided by MSRFC (see Chapter 3). It only uses a single iteration of MLC for classification and compares that to the two MSRFC algorithms. Both MSRFC and MLC used the same training data without sub setting for final classification. Therefore, significant bias can occur in the quantitative comparisons of the two models but the final prediction can have direct comparisons from the same data input. This study focuses on qualitative aspects using post classification. Knowledge-driven assessments take precedence over cross-validated quantitative assessments. These comparisons are subjective; however, there is significant interest for broad, qualitative geologic implications of violating classification algorithm assumptions, rather than directly comparing accuracies. Furthermore, since the MLC algorithm assumptions, such as Gaussian distribution, are violated, a comparison is biased and predestined to fail when compared to RF. The MLC classification contains extensive errors associated with broken assumptions when using the algorithm. This study aims to assess whether synoptic mapping across multiple scene boundaries can be executed more effectively by using the RF workflow presented in Chapter 3.

4.3.1 Study Site

The study site is located on the eastern portion of Victoria Island in Northwest Territories, Canada (Figure 1.3). It is north of the tree line and is part of the polar desert. The land cover is predominantly lichen, small shrubs, and exposed surface materials. The remote location and presence of a polar desert cause vast regions of the landscape to be relatively unmodified since the last glaciation. The study area contains the low

arctic ecosystem with dwarf shrubs and herbaceous legumes (Sharpe, 1991). In many cases, the plant distribution is closely linked to soil types, drainage, and moisture regime (Sharpe, 1991). In other regions, significant wind erosion and aeolian deposits have created recent complexities (Sharpe, 1991). Large regions of streamlined forms indicate a dominance of thick sediment (Sharpe, 1991).. Lighter toned ridges on some streamline forms indicate dryer regions on the peaks of these forms (Sharpe, 1991).. Heavy winds have eroded finer particles from the surface, exposing denser sediment, boulders, and bedrock (Sharpe, 1991). There is extensive permafrost in the area with regions of ground ice and local thermal karst (Sharpe, 1991). These complex ground conditions make this study area particularly complex for classification.

4.3.2 Data Processing

This study used Landsat data from GeoGratis (http://geogratis.cgdi.gc.ca/). The Canadian Centre for Remote Sensing compiled, mosaicked, and balanced the study scenes for this image by using linear regression of the over lapping regions between scenes (Harris, J. Personal Communication, 2011). Compiling and mosaicking the scenes consisted of matching peak season scenes (when possible) to each other in the region. Cloud cover and the limited data availability made this balancing operation imperfect. Due to imperfect balancing, the resulting image contains three distinct spectral regions (Figure 4.1). An interpreter generated training areas for the study region through operational mapping within the Geological Survey of Canada. This mapping procedure involved splitting the scenes and classes to handle classification of bi-modal spectral distributions during MLC classification (Harris, J. Personal Communication, 2011). The

GSC exercise was successful because parametric assumptions within the training data were explicitly assessed and classes were split when their makeup would not form a normal distribution. Spatial regions were identified for classification when spectral balancing issues could not be resolved. The resulting classification used three different regions with separate training and validation for each. This study merged the resulting training data combining similar classes and material types instead of separating each of them by scene. The disadvantage of the GSC approach, however, is the substantial increase in time and effort required to undertake unique classifications for each area with the mosaic. This issue is a major impediment to operational mapping of vast northern regions using mosaics of hundreds of individual satellite images.

Figure 4.1: Study Site broken down by scene balancing regions.

Table 4.1: Training Data Used

CLASSNAME	INDEX	Merged Class	Polygons	Pixels	Percentage
Sand and Gravel	1	Sand_Gravel	26	11902	14.08
Sand and Gravel Eolian	2	Sand_Gravel			
Sand and Gravel	2	Sand_Gravel			
Sand and Gravel	3	Sand_Gravel			
Sand and Gravel Eolian	3	Sand_Gravel			
Sand and Gravel Esker	1	Sand_Gravel_Esker	14	8804	10.42
Very Thin Sediment	2	Very Thin Sediment	4	526	0.62
Thin Sediment	1	Thin_Sed	25	9476	11.21
Thin Sediment	3	Thin_Sed			
Thick Sediment	1	Thick_sed	85	22724	26.89
Thick Sediment	2	Thick_Sed			
Thick Sediment	3	Thick_Sed			
Very Thick Sediment	3	Very Thick Sediment	58	14653	17.34
Very Thick Sediment	1	Very Thick Sediment			
Very Thick Sediment	2	Very Thick Sediment			
Bedrock	1	Bedrock	60	16407	19.41
Bedrock 1	2	Bedrock			
Bedrock 2	2	Bedrock			
Bedrock 2	3	Bedrock			

With the exception of very thin sediment, all of the training classes have proportions ranging from 10 to 27% of the total training data population (Table 4.1). The training data is distributed evenly across the study area and within each of the three spectral regions (Figure 4.2).

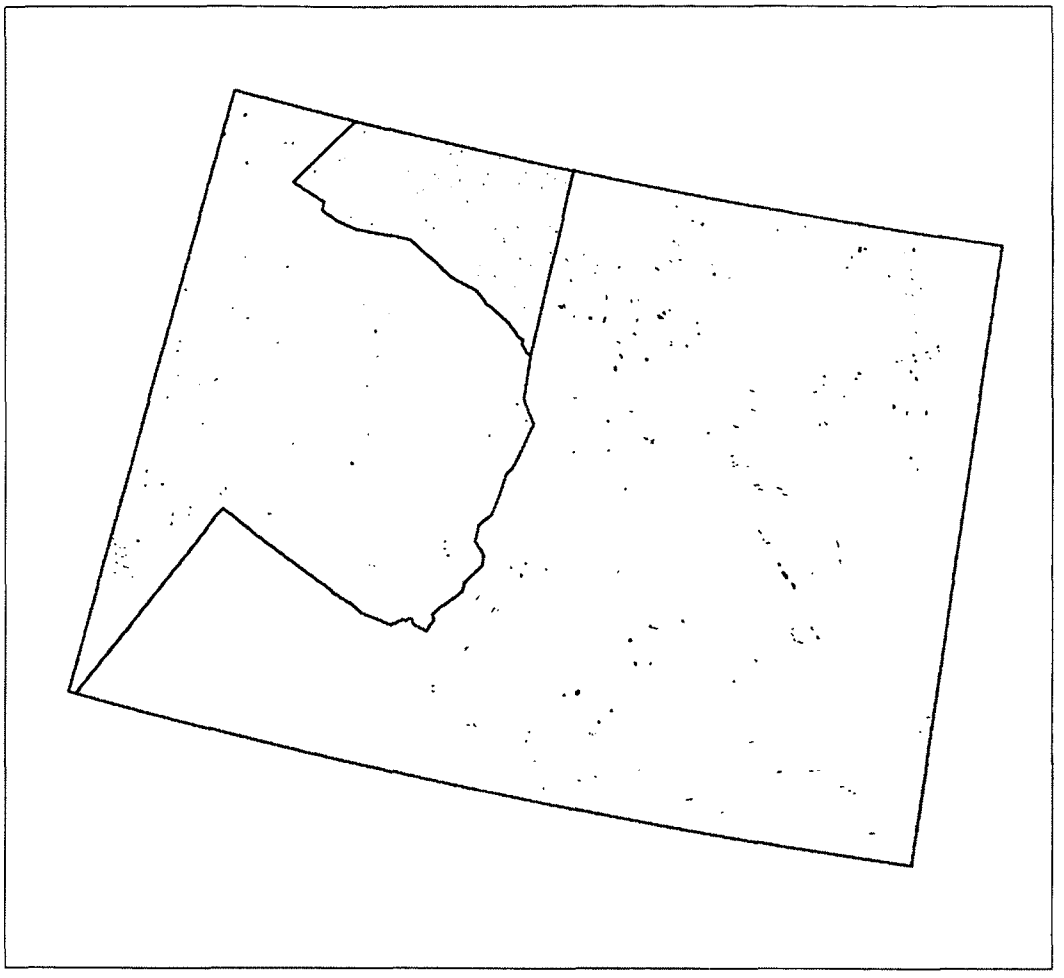


Figure 4.2: Distribution of training data across the study site.

The training data contains multiple different distributions, which are displayed in Appendix A by band. The Bedrock class is bimodal in all Bands. The Sand and Gravel Class is bimodal in all bands with the exception of the SWIR1 band. The thin sediment and very thin sediment have non-parametric skewed distributions across all the Landsat bands. The Sand and Gravel Esker and Thick sediment, and Very thick sediment classes have relatively parametric distributions across all bands. Although not all training sites are non-parametric, the complex classes justify the uses of non-parametric approaches

to manage the distributions. With the completion of training data processing, classification could begin using the mosaic for this study.

4.3.3 Classification

RF comparisons are facilitated by running two iterations one with and one without the Landsat ETM+ scene location index (Figure 4.1). The scene index essentially consists of an image that only contains three values. These values describe the three spectral domains for classification: (1) central, (2) western, and (3) northern (Figure 4.2). The MLC does not use the scene index because it cannot use nominal data sources as input training data. The purpose of comparing these three approaches is to provide insight regarding the impact of different data sources on RF.

4.3.3.1 Maximum Likelihood

MLC relies on second-order statistics of the Gaussian probability function model to discriminate between classes (Gao, 2009). These functions can be weighted using *a-priori* probabilities related to the expected relative proportions of classes (Gao, 2009). Without *a-priori* knowledge, classes are equally weighted (the case in this study) or generated using alternative classification approaches (Gao, 2009). These parameters make it unreasonable to classify study areas when limited knowledge of the region is available. Further, even if *a-priori* probabilities can be applied, quantifying uncertainty remains a problem.

While it has been difficult for researchers to apply classification algorithms to deal with biases and normality issues, classification accuracy improvements have

advanced quickly (Conese, 1992; Maselli et al., 1992; Maselli et al., 1990). There have been many steps taken in understanding and quantifying uncertainty and improving accuracy; for example, error matrices (Conese, 1992), probabilities from external knowledge, non-parametric exploration (Maselli et al., 1992; Maselli et al., 1990), conditional classifications (Kleinbaum & Klein, 2010), and bagging (Harris et al., 2012b). For a more complete review of MLC approaches, see the work of Kleinbaum and Klein (2010). For assessing accuracy of the MLC results in this study, a confusion matrix is used.

A confusion matrix (also called an error matrix) makes a comparison on a category-by-category basis or the relationship between reference data and the corresponding classification result (Lillesand and Kiefer, 1999). Reference data, often identified via field or high resolution image information, thus provides the basis for assessing accuracy. However, classification algorithms and associated accuracy assessments are limited by the quality of data provided (field data accuracy and precision of coordinates) and the often-limited spatial extent of data used to validate. Therefore, unless additional steps or fuzzy classifications are used, accuracy assessments are biased towards known regions or interpreted best case examples of classes created during the training step of classification.

4.3.3.2 Random Forest

The workflow in this chapter uses the MSRFC algorithm introduced in the previous section, which manages classification instability and bias through additional stochastic training and cross-validation. The internal matrix measurements, out of bag

(OOB) error, are assumed to be biased due to spatial autocorrelation because sampling for RF is random on individual pixel values rather than regions. Overall accuracy of RF is measured before variable selection using 100 iterations of training and validation using 60/40 subsets. Therefore, alternative accuracy measures are required. Since multiple stochastic iterations of training and validation are used, overall trends of accuracy across hundreds of different RF iterations are used to estimate accuracy. In order to gauge robustness model comparisons are made amongst the two RF models and the MLC.

4.3.4 Cross-Validation and Accuracy Assessments

Comparisons between methodological approaches are completed on the model itself (usually using cross-validation or a variant of accuracy) and between the models. This research quantifies classification accuracy of the RF model using a series of classifications and recording the resulting cross-validation results, using the same procedure described in Chapter 3. Essentially, the workflow uses a series of stochastic subsamples to cross validate the model and assess prediction sensitivity to number of trees selected and cross-validation subsets used. Next, a classification is produced which uses all of the resulting data from the analysis. The MLC classifier, on the other hand, does not use the same rigorous stochastic training and validation that the RF model uses (such as RCM) because its accuracy tends to be less sensitive to training and validation subsets and does not improve when a stochastic approach is implemented (Harris et al., 2012). Therefore, the classification accuracy of the MLC model is quantified using a simple confusion matrix. This confusion matrix is generated from a single iteration,

based on a random subset of 60% training and 40% validation. All final classifications apply all of the data to the model for a direct, unbiased comparison of the two approaches. The workflow is used to estimate final accuracy but does not influence the actual classification because no sub-sampling occurs in the final classification step.

4.3.5 Product Comparisons

The differences between final classification products are essential to understanding the strengths and weaknesses of these different approaches. This study uses multiple post-classification assessments to assess qualitatively the differences between approaches. First, overall accuracy is compared using estimates from the modified RF workflow and a single iteration from MLC. Since assumptions of MLC are violated in its application to this study (demonstrated with class distributions in Appendix A), the value of statistical comparisons of accuracy is questionable. This study deliberately violates MLC assumptions in order to demonstrate a common practice within surficial geological mapping in order to discuss and compare the results of such a methodological approach. Next, cross tabulation is used to compare the different classifications pixel by pixel. Finally, regional differences between classifications are explored by manually identifying regions of confused classification between products. This assessment examines the differences between the two classification approaches and points out scenarios for which each classifier seems to be performing better.

4.3.5.1 Estimations of Accuracy

The MLC uses the validation data sets to produce the confusion matrix. The matrix assesses accuracy and is generally a common accuracy standard at the core of many assessments (Foody, 2002). The error matrix is a square array of numbers in rows and columns. The columns represent the reference data and the rows represent the classification result (Congalton & Mead, 1986). Many studies use the error matrix as a standard for classification error assessment (Conese, 1992; Foody & Cox, 1994). However, the error matrix has problems, most of which stem from assumptions regarding the data that result in an exaggeration of accuracy (Foody, 2002). For example, rarely is all of the validation data appropriate to use due to mixed pixels, or because training data is misregistered to the remotely sensed datasets (Foody, 2002). This study is not primarily concerned with the overall accuracy statistic and therefore does not directly compensate for MLC exaggeration. However, the study will utilise both visual and statistical analysis in order to understand the differences between the two classification approaches.

At each iteration of the modified RF classifier, overall accuracy is assessed by using the confusion matrix and by ignoring the internal matrix measurements. Estimates of overall accuracy based on the final trend of the accuracy measures from the stochastic iterations of RF are used. The RF workflow also produces boxplots that represent the overall number of votes for each class. These boxplots summarize the probability used to classify each class within the study area. The probability is determined using the number of votes from the trees that voted for a specific class

within a pixel. These probabilities serve as relative metrics of prediction confidence by class. However, this method has a tendency to oversimplify analysis by allocating uncertain pixels to the same class (see Chapter 3).

4.3.5.2 Cross Tabulation

Direct comparisons are useful in assessing the overall conformity of one map to another. Cross tabulation is the core tool for direct map comparisons using similar classes (Pontius & Cheuk, 2006). Cross tabulation summarizes the conformity of each map to the other pixel by pixel, providing an overall percentage of the correspondence of one map to the other. This measure, common to land cover change, should have a resulting conformity similar to the product of multiplying the overall accuracies of both maps (Lambin and Strahlers, 1994; Serra, Pons, and Sauri 2003; Singh, 1989). Its similarity to the confusion matrix makes it particularly desirable for easy interpretation, since it compares maps directly on a per pixel basis across the entire scene. This study produces a comparison of the entire scene between the different classification models using a cross tabulation matrix. This allows for overall class and conformity between maps to be estimated. Finally, this study produces two maps to compare random forest to itself and MLC to assess the classification conformity across the study area.

4.3.5.3 Regional Comparison

Regional comparisons are essential in order to test the accuracy of classifications. They can confirm how various methods perform outside of measured training areas and validation boundaries. We can verify any conceptual understanding of

the area, validate known results, highlight areas of confusion, establish differences between classifiers, and focus on any specific issues or problems in classification. It is for these reasons that this study uses regional comparisons. This thesis includes comparisons, discussions, and exploration of issues in the final classification for the regional comparisons within a workgroup of surficial geologists. Regional differences among maps are explored by focusing on specific study sites. The underlying reasons for observed errors are reported, with a specific focus on how to reduce them in future classifications. Specifically, this comparison will focus on known lithologic and vegetative regions identified during training but not selected for quantitative analysis. The evaluation assesses and discusses areas of confusion to determine their impact on classification and the final classification map.

4.3.5.4 Scene Boundaries

Finally, the product comparison focuses on scene balancing issues and the resulting boundaries. In Chapter 3 it was hypothesised that RF can handle scene-balancing issues that occur when study areas do not coincide with scene acquisition. This hypothesis is in the present chapter using by incorporating the scene number in the RF model and then excluding it. This study tests the hypothesis by directly assessing these two RF maps for their application across scene boundaries. The effect of spectral balancing on individual classes and the whole scene was explored in order to gauge the effect of the boundaries on the different models.

4.4 Results and Discussion

The methodology described above resulted in a number of products as well as many statistical and visual accuracy comparisons. This includes overall accuracy comparisons and cross tabulations. The overall accuracies are not directly comparable to each other because RF cross-validation accuracy is stochastic and MLC is not. The second comparison (cross tabulation) is numerical and more comparable than overall accuracy because both models use all of the data to produce the final classification. The third regional comparison focuses on specific points of interest on the map (selected for poor classification) to compare how the different algorithms performed. Scene balancing issues focus on the effect the spectral balancing had on the classification boundary seams. The final classification and comparisons used in this methodology provides a broad synoptic description of the landscape, which aids in analyzing the performance of each model. These comparisons help us to understand whether RF or MLC is more appropriate for classification in this study area.

4.4.1 Accuracy

The accuracy for the MLC classification (Table 4.2) indicates an overall accuracy of ~48.6%. This represents a poor classification (less than 50% accuracy). Bedrock was 37%, thin sediment was 27%, and very thin sediment is 6% - all of which are far below the overall accuracy. Low percentages are likely a result of the complexities of the classes (such as light and dark lithology) resulting in non-parametric spectral distributions. The thick sediment and sand and gravel classes score much higher at

58.4% and 60.5% respectively. Each class shows some confusion between similar material classes (for example, sand and gravel eskers is often confused with sand and gravel). This confusion is expected and is an example of how similar material types may belong to separate genesis.

Table 4.2: Confusion Matrix for MLC using 60% training 40% validation.

	Bedrock	Sand Gravel	Sand Gravel Esker	Thick Sediment	Thin Sediment	Very Thick Sediment	Very thin Sediment	Omission Total	Users Accuracy
Bedrock	675	1317	226	1038	1588	1919	145	6908	0.09
Sand Gravel	710	4448	53	153	60	0	24	5448	0.81
Sand Gravel Esker	191	1491	373	503	762	128	33	3481	0.10
Thick Sediment	166	30	84	3971	2166	1372	274	8063	0.49
Thin Sediment	32	0	0	797	1733	24	1125	3711	0.47
Very Thick Sediment	0	0	0	337	0	4651	0	4988	0.93
Very thin Sediment	35	58	1	0	29	0	105	228	0.46
Commission Total	1809	7344	737	6799	6338	8094	1706	32827	NA
Producers Accuracy	0.37	0.60	0.50	0.58	0.27	0.57	0.06	NA	NA

The OOB error and overall accuracy of the RF model using the scene location as an input variable are depicted in Figure 4.3 and Figure 4.4, respectively. The OOB error averages (Figure 4.6) at ~13.9% indicating 86.1% certainty of the model (Figure 4.3). The average overall accuracy, produced by stochastic cross-validation assessments, is ~78% (Figure 4.4). This represents a difference of ~8.1% between OOB and overall accuracy. The higher OOB error relative to overall accuracy indicates that some overestimation of accuracy is taking place from internal measurements (demonstrated in Chapter 3). This anomaly is likely due to spatial autocorrelation within the OOB error measures because pixel based validation tends to overestimate classification accuracy (Muchoney, 2002). Spatial autocorrelation is managed in the training and validation data sets by sampling

polygons used for training rather than points. These likely accounts for the ~8.1% difference between the OOB error and overall accuracy. Overall, the RF model performs well as a classifier, given the high accuracy values for both the cross-validated and OOB error estimates.

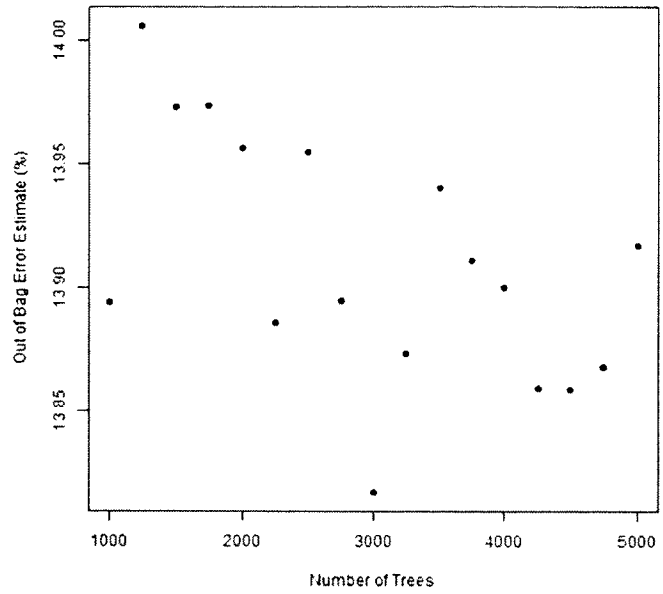


Figure 4.3: Out of bag error using Landsat ETM+ Mosaic bands.

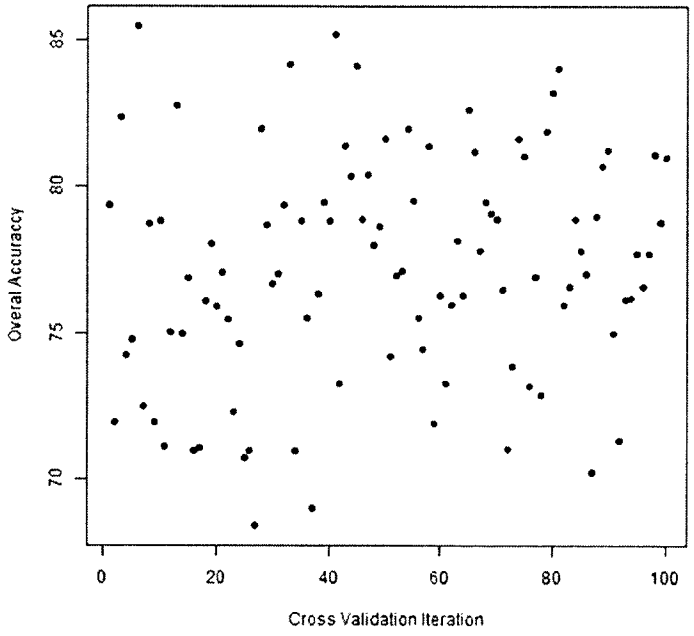


Figure 4.4: Overall accuracy results from 100 confusion matrices and Landsat ETM+ bands.

Mean decrease Gini boxplots (Figure 4.5) assess overall variable importance in the model. It does this by displaying the variable of importance, for example GREEN, measured across 100 iterations. While this study uses an additional index variable to separate the different Landsat scenes, it is the least important component in the analysis. Conversely, the SWIR2, blue, and NIR bands are the most important for classification. The raw green band is the least important while SWIR1 and red bands have similar importance, with little difference between the two. We can assume that all Landsat bands are useful for the classification and the index band may not be as significant.

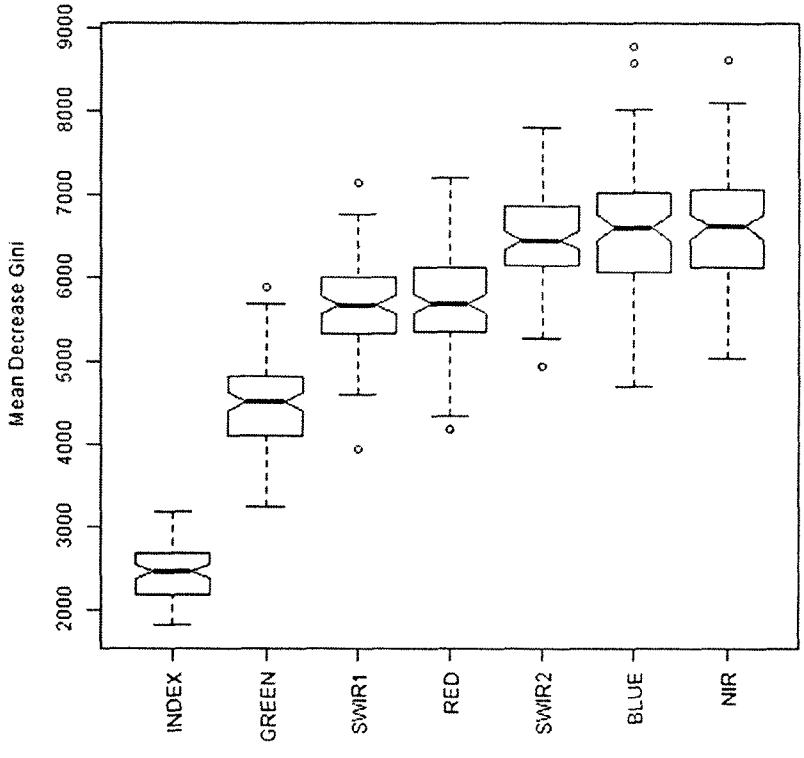


Figure 4.5: Mean decrease Gini from random forest

The Gini index provides a framework for adding variables to the model. Starting with the highest value (NIR), the workflow runs the variables through the RF model 50 times using 70% training and 30% validation (Figure 4.6). It analyzes the first variable and adds an additional variable to the collection – one at a time –until all have been combined for analysis (NIR; NIR and BLUE; NIR, BLUE and SWIR; etc). This step is not required for variable selection, but it provides insight into the amount of prediction variation expected from the RF model. The most important band, NIR, resulted in 25% accuracy when it was the only one used for classification. When the workflow added the BLUE band, the classification accuracy jumped to 45%, which indicates there is a direct impact on the model, despite insignificant differences in Gini measurements. This supports further exploration for variable selection and suggests that there is potentially bias within variable importance measures reported in the remote sensing literature (Strobl et al., 2007). This bias is usually associated with comparing nominal data to interval or continuous data (Strobl et al., 2007). This case is different for the BLUE band because the scale of data (integer 1-255 for all bands) should not directly cause biased importance results. However, when ratio variables are used in conjunction with nominal data (in our case the scene Index), importance measures are misleading and unreliable (Strobl et al., 2007).

Accuracy increased with the addition of each additional variable except for the index variable, which did not lead to improved accuracy. The low increase from the index variable is an indication that it does appear to benefit spectral balancing management for the study area. Since this index value shows little importance in Gini

and little increase with overall accuracy it is not expected to provide significant increase in mapping predictions for the study area.

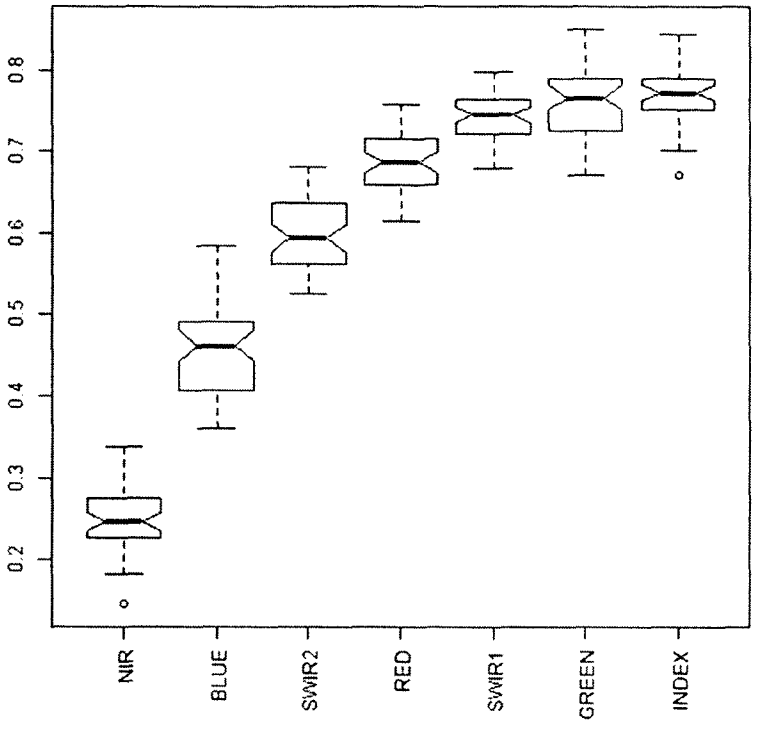


Figure 4.6: Overall accuracy adding each variable by order of GINI importance across 50 iterations.

The scatter plot of accuracy as a function of proportion of the training data withheld for cross-validation (using all the variables as model inputs) was used to estimate the overall classification accuracy (Figure 4.7). The overall accuracy trend increases with the addition of more training data. This suggests that the training data has little redundancy because more data is explaining more variance in the model and thus providing stronger results. This means an individual polygon likely represents a unique aspect of the landscape, which is supported by the outlier accuracy measures when training goes above 80%. It is not advisable to exclude any variables as a validation

sample for classification because they may be crucial data for the model. Final classification must use all training data for classification using Figure 4.7 to estimate overall accuracy at ~78%. This accuracy is significantly higher than the MLC accuracy of 48.6%. At its lowest values, the MSFRC model still improves upon the MLC accuracy measure. It is probable that a similar cross-validation approach to MLC would produce a similarly shaped plot. However, across 100s of iterations for RF, no value was below 50% (most were above 60%). Therefore, MLC's single iteration of accuracy represents a considerably lower accuracy when compared to the RF model.

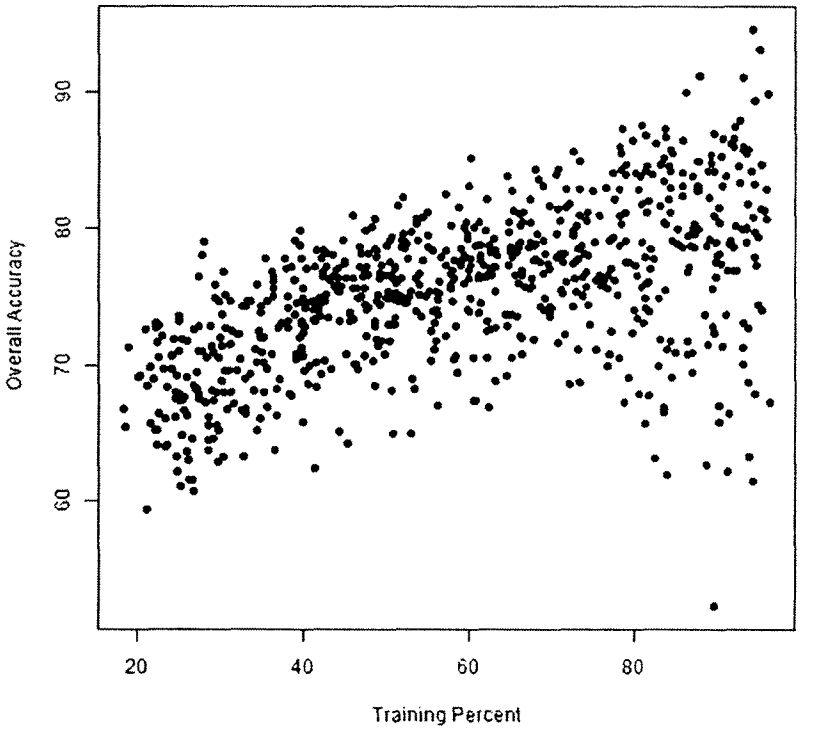


Figure 4.7: Overall accuracy using all variables and increasing training percentages.

4.4.2 Cross Tabulation

The workflow resulted in two cross tabulations to aid in the comparisons of different classification models. First, a cross tabulation was conducted between the MLC and the RF with an index (Table 4.3). Second a cross tabulation was produced which compared both RF models with and without index (Table 4.4). RF and MLC have a ~65% conformity to each other, which is higher than anticipated when cross tabulating these results. An estimate more similar to the product of these two accuracies, in this case ~37% was expected (Singh, 1989). This discrepancy indicates that MLC and/or RF are more similar than their accuracies imply. It is also likely related to specific classes which take up large areas of the map that where classified similarly. There was a high level of agreement (greater than 60%) for thick sediment, very thick sediment, and sand and gravel classes. The other classes (sand and gravel esker, very thin sediment, bedrock, and thin sediment) were classified differently between the two models – bedrock was at ~2% for MLC to RF and 0.5% for RF to MLC. This indicates little to no conformity for the bedrock class, where MLC classified bedrock, RF did not and vice versa. Low correspondence in the cross tabulation is likely due to spectral complexities within the bedrock class. As indicated earlier, the bedrock class has multiple lithological characteristics across multiple scene boundaries. This causes confusion with other material types.

Table 4.3: Cross Tabulation between Random Forest with Index and Maximum Likelihood classification.

Maximum Likelihood Classification	Random Forest With Index							
	Thick Sediment	Very Thick Sediment	Sand and Gravel	Sand and Gravel Esker	Very Thin Sediment	Bedrock	Thin Sediment	Percent Agreement
Thick Sediment	20 363 697	1 238 948	190 045	1 311 487	152	931 628	755 751	0.82
Very Thick Sediment	5 404 490	12 291 701	0	680 130	0	1 299 008	29 871	0.62
Sand and Gravel	1 429	0	654 993	183 173	726	81 870	802	0.71
Sand and Gravel Esker	6 691	232	94 585	97 238	0	23 378	8 330	0.42
Very Thin Sediment	546 728	415	257 685	89 123	162 463	185 115	387 413	0.1
Bedrock	176	719 465	20 135	235 301	0	15 713	860	0.02
Thin Sediment	3 000 757	5 138	51 356	397 685	94 260	494 446	1 022 940	0.2
Percent Agreement	0.69	0.86	0.52	0.03	0.63	0.005	0.46	0.65

Table 4.4: Cross Tabulation between Random Forest with Index and Random Forest without Index

Random Forest No Index	Random Forest With Index							
	Thick Sediment	Very Thick Sediment	Sand and Gravel	Sand and Gravel Esker	Very Thin Sediment	Bedrock	Thin Sediment	Percent Agreement
Thick Sediment	27 353 311	387 823	158 375	297 781	45 342	392 869	229 669	0.95
Very Thick Sediment	403005	13 576 028	3 631	202 864	129	95 734	2 603	0.95
Sand and Gravel	46047	570	926563	17 269	2 956	10 406	6 259	0.92
Sand and Gravel Esker	737477	217 450	108801	2 386 266	6 027	70 680	147 117	0.65
Very Thin Sediment	118347	2	2531	4 512	195 613	1 167	12 873	0.58
Bedrock	292023	49 056	65248	60 873	4 619	2 449 559	21 454	0.83
Thin Sediment	373758	24 970	3650	24 572	2 915	10 743	1 785 992	0.8
Percent Agreement	0.93	0.95	0.73	0.8	0.76	0.81	0.81	0.91

In the northeast portion of the map, areas of sediment have bedrock characteristics. As bedrock outcrops erode, they distribute along the surface of sediment, making the two classes spectrally similar. The sand and gravel esker class has a value of 42% for the MLC classification and 3% for the Random Forest classification. These results show that 42% of the eskers classified by MLC are also classified as eskers using the RF model. Conversely, only 3% of RF-classified eskers corresponded with the same classification in MLC. Therefore, RF is classifying more of the landscape as eskers than MLC, which indicates either an RF esker exaggeration or an underestimation of eskers by MLC.

The two RF models – one which had an index, and one which did not – created a map where ~91% of the pixels were the same (Table 4.4). The 9% difference between the two may be in small part due to stochastic variation and also the addition of the index variable. The lowest conformity class is the very thin sediment class, where 58% from the index-free RF matches the indexed RF. The RF model classifies more sand and gravel eskers before the addition of the index class. The absolute numbers for classification show that the RF without an index classifies 6.8% of the map as sand and gravel eskers. Similarly, the RF with an index classifies 5.6% as sand and gravel eskers. Due to its small area on the landscape, this class does not significantly affect overall accuracy. Similarities between sand and gravel eskers and sand and gravel classes are likely difficult to distinguish using just spectral attributes and therefore are more likely to be confused. The rest of the classes have a relatively high conformity, so it is assumed that the overall differences have not changed significantly.

Additional products are used to compare the difference between the RF classifications. The land area coverage for these maps (Figure 4.8 and Figure 3.10) show the amount of land area covered at given probability intervals. A slight increase in land area coverage occurs with the addition of the index value (4.4% higher at 90% certainty). These figures indicate a slight increase in model certainty with the addition of the index variable.

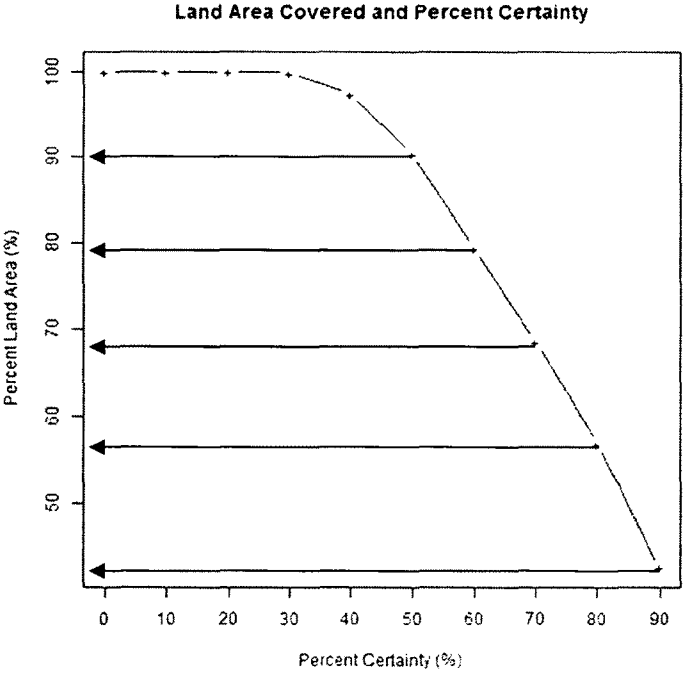


Figure 4.8: Overall probability vs. land area percent coverage for RF without index used. Horizontal lines are guides to show %land area at 50%, 60%, 70%, 80% and 90% probabilities, which can be used for comparison with alternative models/classifications (e.g. Figure 4.9).

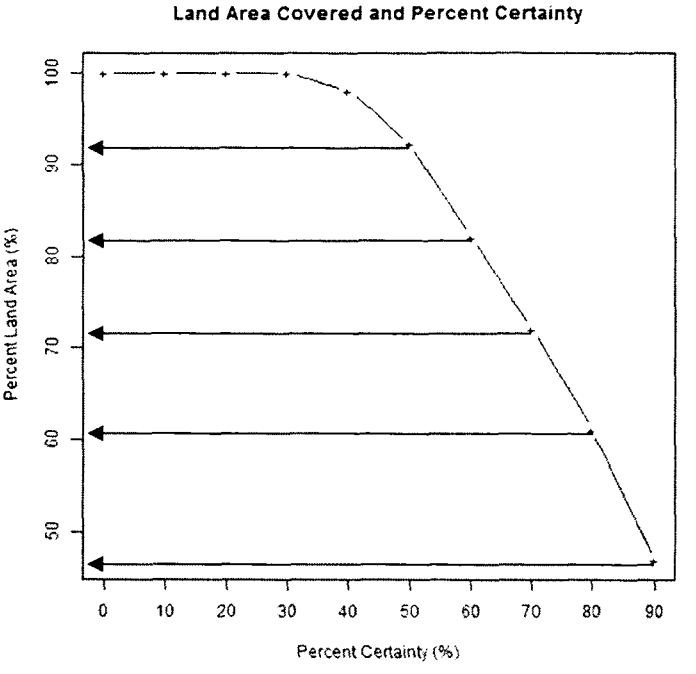


Figure 4.9: Overall probability vs. land area percent coverage for RF with index used. Horizontal lines are guides to show %land area at 50%, 60%, 70%, 80% and 90% probabilities, which can be used for comparison with alternative models/classifications (e.g. Figure 4.8)

In both comparisons (MLC vs. RF and RF vs. RF) the true impact on the classification of surficial materials is unclear. Further, the relatively high cross tabulation for MLC vs. RF of ~65% indicates the accuracy of MLC is likely more similar than overall accuracies indicate because 0.65 should be close to the product of both accuracies (Lambin and Strahlers, 1994; Serra, Pons, and Sauri 2003; Singh, 1989). If this theory holds then accuracies of 80% each should result in a cross tabulation of ~65%. Conversely, the high cross tabulation of ~91% between the two RF models indicates that results are so similar that statistical significance is negligible. In both cases, a qualitative approach is used in order to understand how these models and their differences behave for surficial materials.

4.4.3 Regional Comparisons

In order to understand the impact that the accuracies above have on the quality of classification, qualitative assessment is necessary. Most close up comparisons visually compare different classifications in specific areas of interest. In all cases where a classification is presented, the same legend colour scheme is used to reduce confusion (Figure 4.10). These study areas identified represent regions of known difficulties in past classifications, or areas that are identified during training but and not used for classification. These areas are identified subjectively but do represent classification scenarios that are expected to be more complex and have the least amount of accuracy.

Figure 4.10: Classification Legend

The northwestern portion of the study area is composed of sediment deposits derived from the adjacent mafic bedrock in the area (Figure 4.11). Spectrally dark, mafic lithological material causes the overriding surface to be much darker. Consequently, a sediment blanket cannot be easily distinguished from mafic bedrock. The MLC captures little bedrock for this region, which is expected considering past field excursions and air

photo interpretations implied significant sediment deposits. In regions where classifiers mapped the bulk of the region as bedrock, both of the RF classifications showed similar results. The mafic sediment deposits cause an exaggeration of bedrock, but some of the bedrock classifications may correspond to real bedrock occurrence. However, all three classifications selected the outcrops as sand and gravel, appearing as concentric bands. The algorithms likely selected the outcrops as sand and gravel due to higher reflectance caused by sun angle and lack of lichen cover. Due to the overall under-classification of bedrock across the entire study area by MLC, the MSRFC classifiers performed better in this specific region.

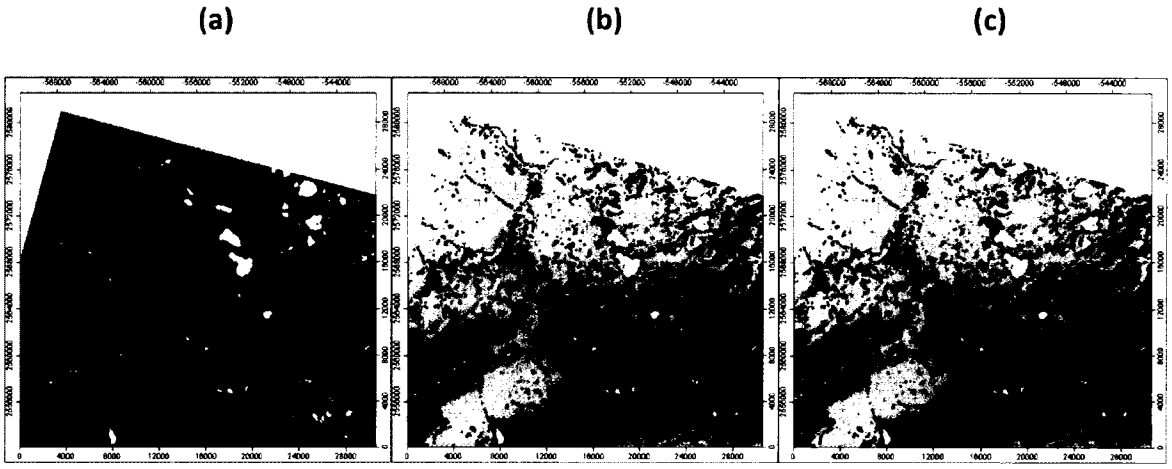


Figure 4.11: North Western Portion of the study area for (a) MLC, (b) RF without Index, and (c) RF with index. See Figure 4.10 for legend.

Figure 4.12 displays a selected study area with Sand and gravel deposits. These Sand and gravel deposits are located in the center of the study area, identified by their bright spectral signature in Landsat bands 1 through 4. The signature has a darker tone along edge of the deposit, associated with moisture differences. All three classifications

predicted the sand and gravel esker class incorrectly. The deposit's higher moisture regions, resulting in higher vegetation and lower brightness in all the spectral bands, cause some confusion for the area. The MLC classifier also selects some areas as very thin sediment when they are in fact sand and gravel esker. Both RF classifiers exaggerate sand and gravel eskers; however, MLC's classification of thin sediment in these areas is less appropriate than sand and gravel esker. Areas on the edge of aeolian deposits represent a spectral signature not sampled by interpretation during the training and validation stage.

(a) (b) (c)

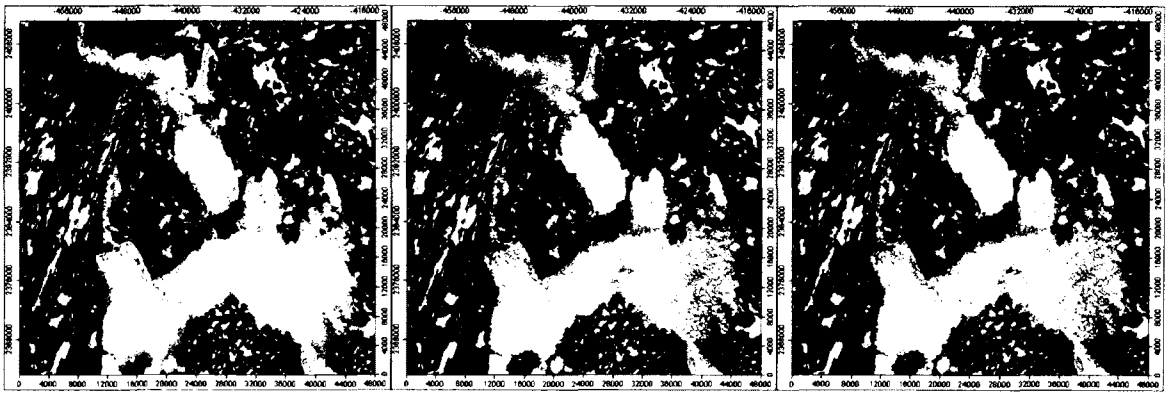


Figure 4.12: Aeolian Deposits in the center of the study area for (a) MLC, (b) RF without Index, and (c) RF with index. See Figure 4.10 for legend.

A single scene that is misbalanced relative to the southeast portion of the mosaic, located at the northern portion of the study area is depicted in Figure 4.13. Just east of the portion of mafic bedrock, there is a significant amount of variability between classifications. The MLC selects most of this area as very thin sediment and thin

sediment. Thick sediment also corresponds with the moisture regime. The RF with no index over-classifies the sand and gravel esker components of this region. However, it does show some of the bedrock outcrops, which, according to interpretation, was expected in this area. Much of the area that MLC mapped as bedrock is likely to be sand and gravel or thin sediment, as it tends to correspond with moisture flow and sand exposures. The indexed RF classifies large areas as thin sediment and thick sediment. Further, RF with index also appropriately classifies bedrock outcrops and sand and gravel. Although scene balancing issues are not entirely resolved (as seen in the southern part of the figures inside the ellipse), the RF with index produced the strongest classification. This is apparent with the absence of the sand and gravel esker class. The reduction of this class in this area is an indication that the most appropriate classification uses the index for this region.

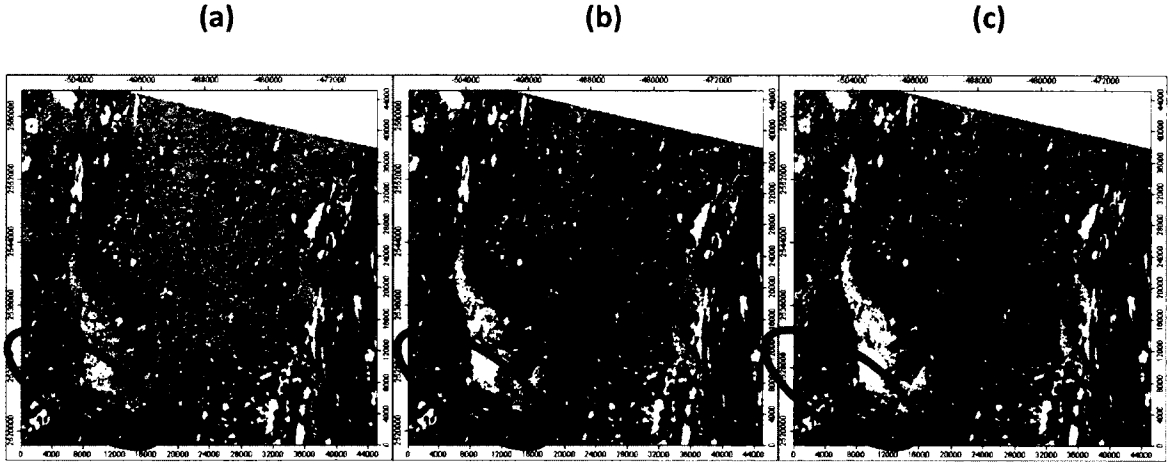


Figure 4.13: Northern centre portion of the study area for (a) MLC, (b) RF without Index, and (c) RF with index. The ellipses indicate the region where spectral balancing issues lead ot sharp discontinuities in classification results, increasing misclassification results. See Figure 4.10 for legend.

The southwestern portion of the study area contains a higher level of moisture and well-established drainage networks. This area (Figure 4.14) is predominantly thick sediment and has a relatively consistent classification varying between thin sediment and very thick sediment. The centre of the image shows a meandering river network, ending with possible outwash at the southern terminus in a fan shaped deposit. The MLC classifier shows the regions adjacent to the study area as thin sediment, while both RF classifiers classify the regions as bedrock, thin sediment, and sand and gravel. Sand and gravel and possibly bedrock are likely the most appropriate classes for that area given that these are surrounding streams. In the RF, scene-balancing issues occur at the boundary along the river between bedrock and sand and gravel. Therefore, the index has a limited effect on this part of the study area – the same boundary occurs, in both RF models. Finally, the MLC classifier shows thin sediment to the south. These areas, which are likely sandy, may have some vegetation. The RF classifies these areas as sand and gravel eskers, which is still incorrect but more appropriate, given the similarity to the other sand and gravel class. The RF model outperforms the MLC significantly but the addition of the Scene Index class has little effect on the classification.

(a) (b) (c)

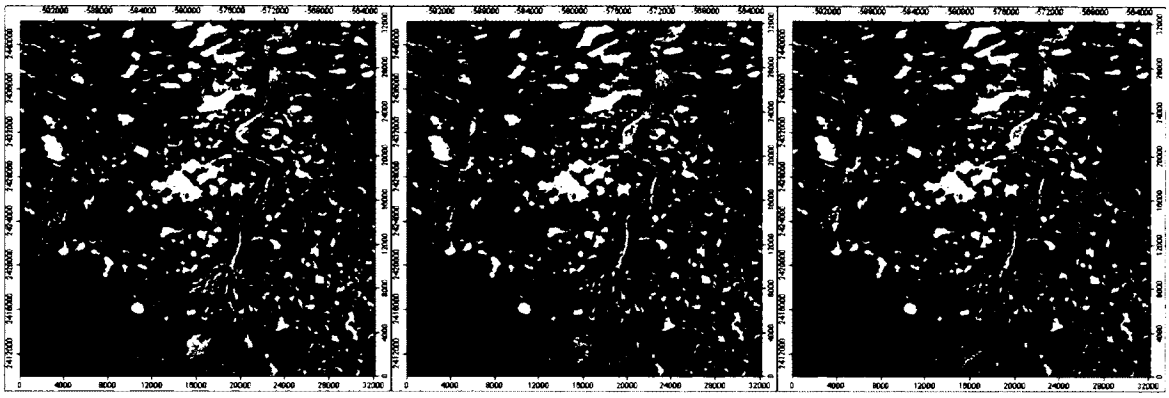


Figure 4.14: Southwestern Portion, a river delta for (a) MLC, (b) RF without Index, and (c) RF with index. See Figure 4.10 for legend.

4.4.4 Scene Boundaries

The study area is composed of multiple scenes mostly obtained from either the same acquisition date or at least the same season. However, two scenes on the western and northern portion of the map were acquired in different seasons causing seams. The resulting classification reflected this inconsistency and further showed the ability of RF to handle classification of image mosaics that are based on spectrally imbalanced images or images acquired at different times of the year. Comparing classification artefacts caused by spectral discontinuities at scene boundaries is therefore used as a way to evaluate the success of the RF model's ability to manage issues during classification.

The northern scene's mismatching (Figure 4.15) caused abrupt changes in the classifications for all three algorithms. The scene has significant amounts of haze, resulting in a brighter spectral response for all materials. In turn, the MLC designated

most of the region as thin sediment. The RF classifies more thick sediment, which corresponds to the southern materials of the boundary. The RF with index defines the area as sand and gravel eskers. The western portion's bedrock signature is more prevalent in the RF with index, while it is non-existent in the MLC classification. The western portion of the seam has a less prevalent boundary, as indicated by the smooth transition. This is likely a result of the dark mafic sediment and bedrock exposures, which outweigh the increase in brightness caused by balancing issues.

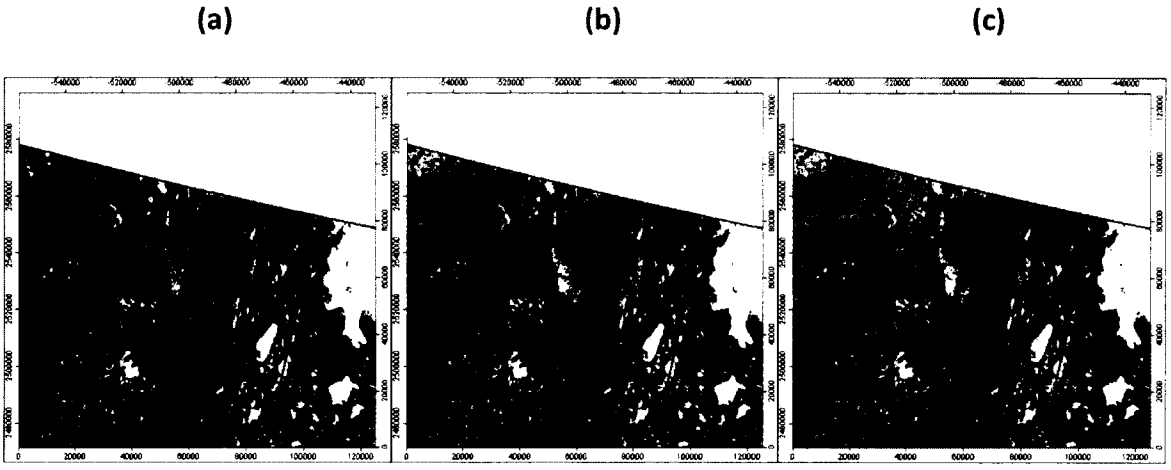


Figure 4.15: Northern scene balance issue for (a) MLC, (b) RF without Index, and (c) RF with index. See Figure 4.10 for legend.

The western scene portionis very large, showing a clear boundary in some regions and a lack of boundaries in others. Figure 4.16 illustrates a scenario where MLC performs more consistently across the boundary. Significant abruptness for RF along the river delta occurs while none exists for MLC. The thin sediment class present in the MLC is likely an inappropriate classification for the study area. Based on expert knowledge underlying the training datasets and morphology of fluvial systems, the spectral

signature should correspond to the sand and gravel class. This classification confusion is common because immobile gravel, thin sediment, and bedrock all host similar vegetation (lichen) as they can have similar lithologies and moisture content (e.g. Grunsky et al., 2009). Both RF classifications look similar and show the fluvial classification of sand and gravel, but then transitions abruptly to bedrock as it crosses the boundary. The RF selected the appropriate class on one scene, but transitions to an incorrect classification at the boundary.

(a) (b) (c)

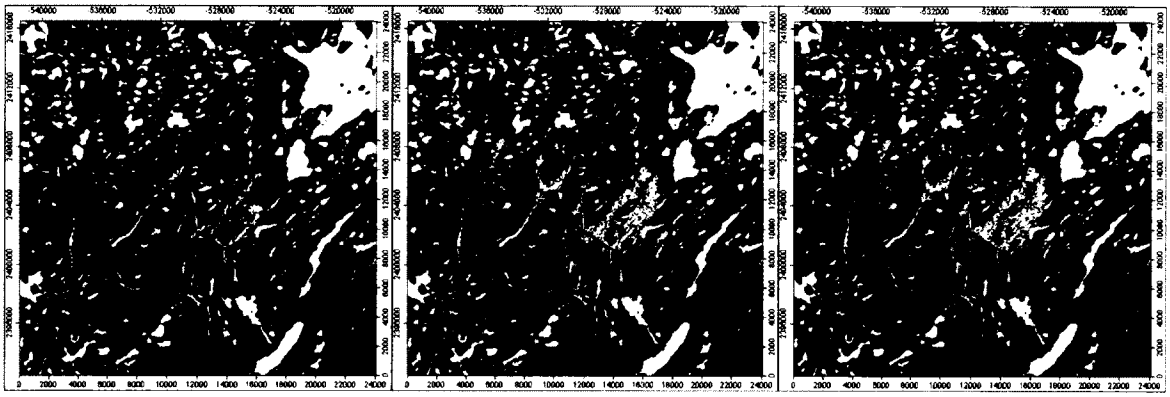


Figure 4.16: A portion of western portion of scene balancing problem for (a) MLC, (b) RF without Index, and (c) RF with index. See Figure 4.10 for legend.

Using RF probability maps we can explore sites for how certain the output model was. This aids in gauging the classification of regions by assigning the corresponding maximum probability certainty to them. Maximum probability images on the northern seam for RF without index and the RF with index illustrate the change in certainty caused by one variable addition (Figure 4.17 and Figure 4.18). As this boundary is crossed, a significant drop in the model's certainty occurs. This indicates that the classes

from model analysis may be incorrect. By comparing these two images, we can see that adding the index class increases the certainty in the thicker sediment class. The bedrock and sand and gravel exposures also have a strong probability response, likely due to the distinctiveness of their signature classes.

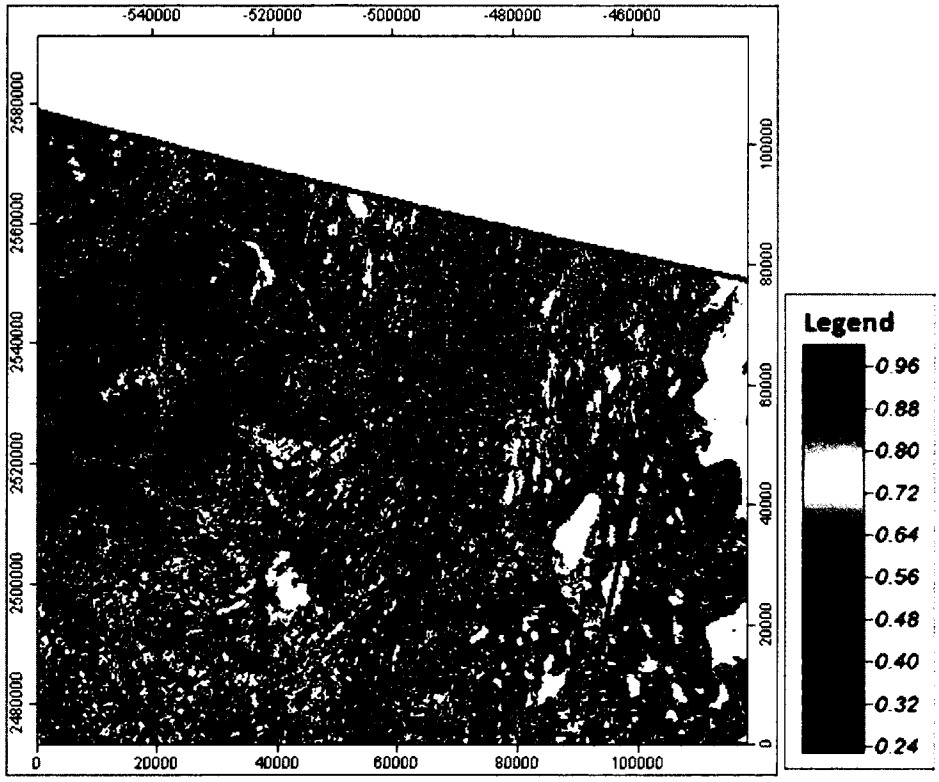


Figure 4.17: Max Probability of RF without Index at Northern Scene Balancing Location

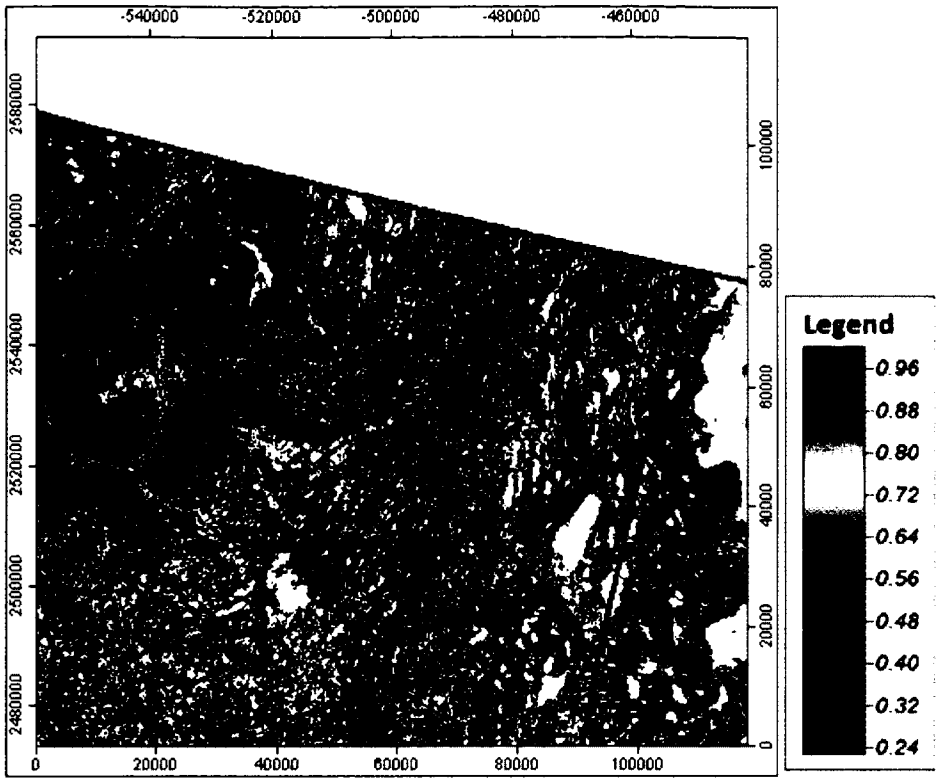


Figure 4.18: Max Probability of RF with Index for At Northern Scene Balancing Location

The western scene balancing areas are less severe than the northern. Both RF without an index and RF with an index have a slight change in max probability at the scene boundary (Figure 4.19 and Figure 4.20). By exploring these probability maps along with the regional comparisons the areas where probability drops tend to correspond with incorrect predictions for RF. Finally, areas with a higher degree of certainty for the RF model relative to the rest of the study area are tending towards appropriate classifications. It is possible that certainty transitions across boundaries are caused by the training area percentages. For example, if one class' training was exclusive to or the majority of them occurred in one spectral scene it would represent one sample of the population for that class.

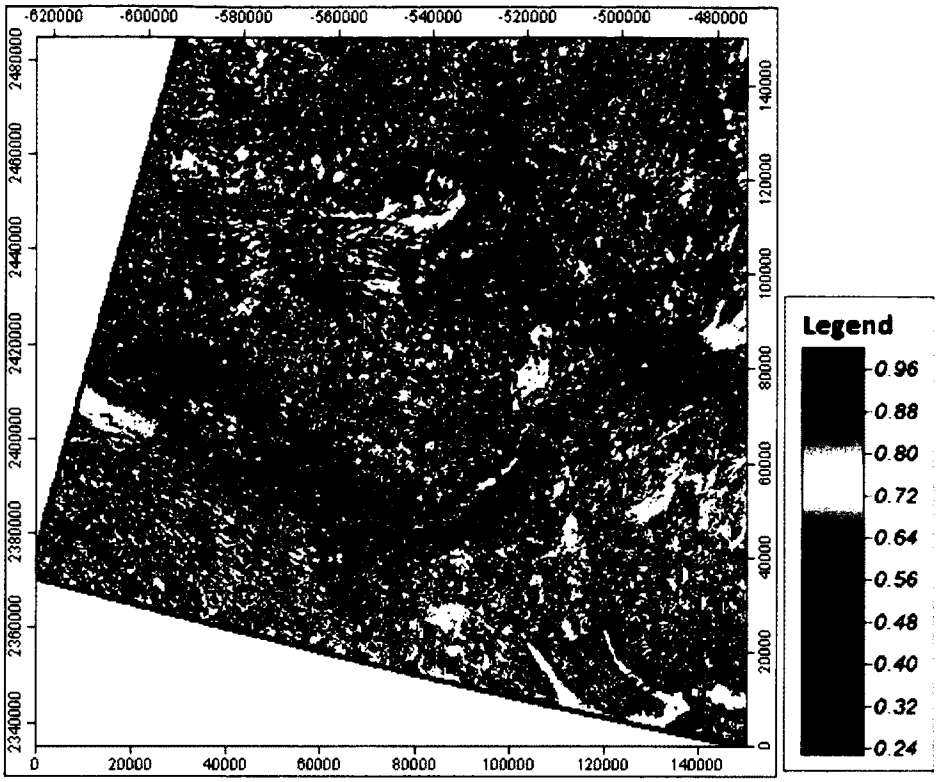


Figure 4.19: Max Probability Map of RF without Index for Southwestern Scene balance location.

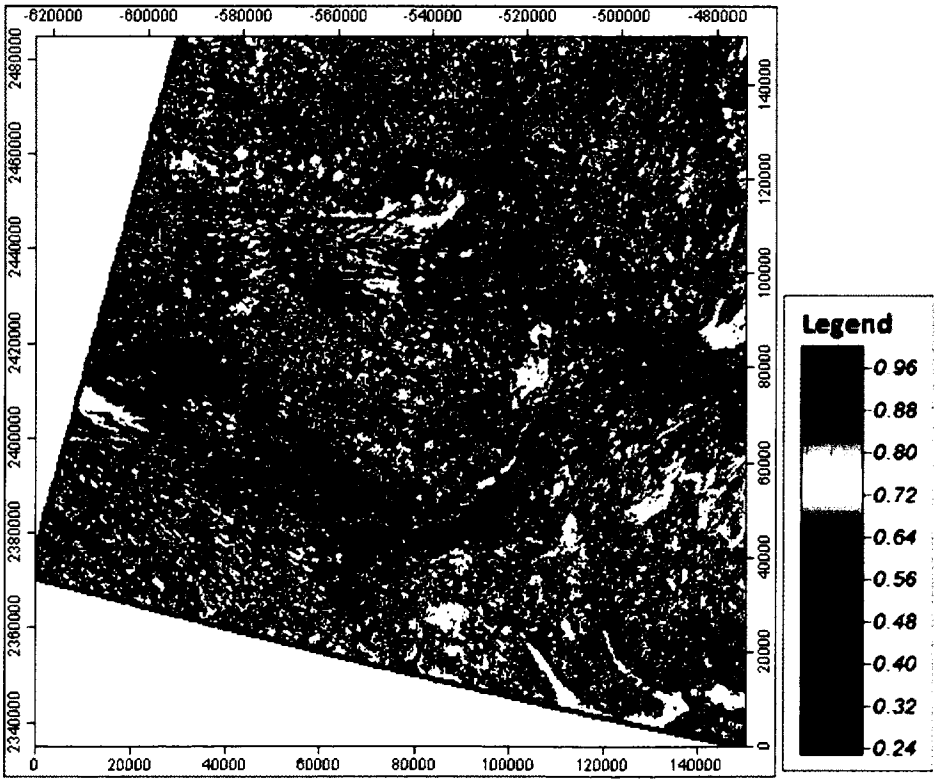


Figure 4.20: Max Probability Map of RF with Index for Southwestern Scene balance location

4.4.5 Final Classification

The three classification products all have significant errors as was demonstrated by the max probability and comparison plots. Some errors were systematic across all three classifiers (usually associated with scene balancing) and some errors were isolated to certain algorithms and surficial material classes. The MLC algorithm (Figure 4.21) had very little classified as bedrock in the northwestern and northeastern portions of the map. MLC classified the aeolian deposits as sand and gravel, showing a higher correctness in this region. Both RF models classified bedrock in the northeast and west (Figure 4.22 and Figure 4.23), but the extents of the predicted areas were exaggerated based on expert knowledge of the area. This was probably due to the mafic nature of the bedrock and the sediment in the northwest. In areas where eskers were expected based on *a-priori* understanding of this landscape, both RF models correctly classified the eskers. However, similar spectral signatures of eskers to sand and gravel may be leading all classifiers to over-predict this class.

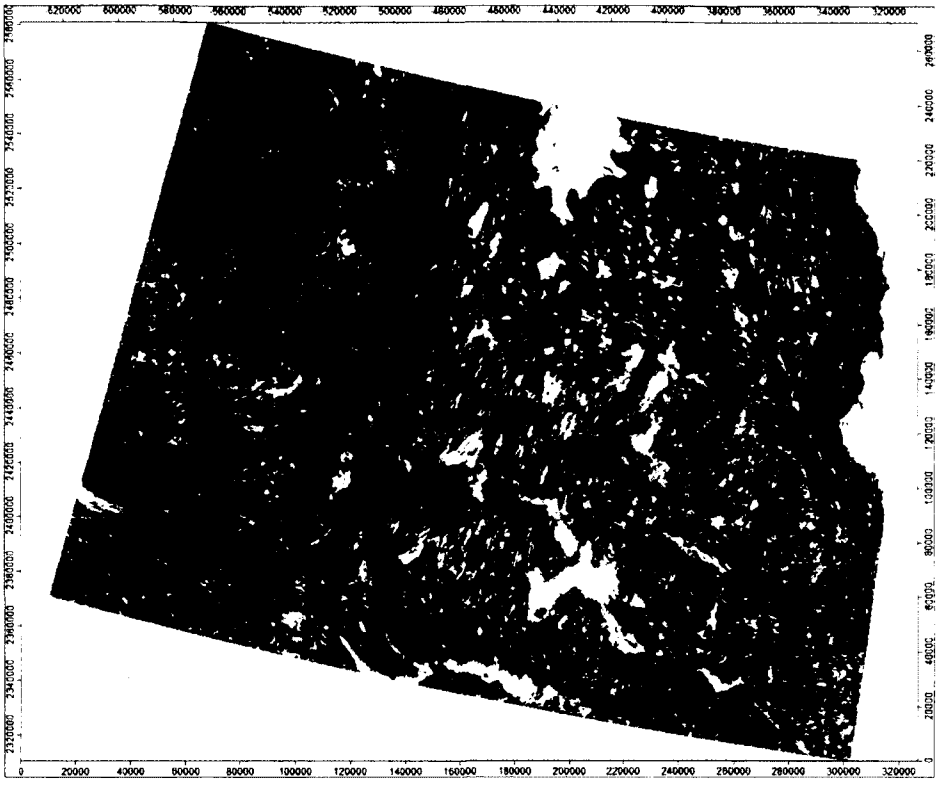


Figure 4.21: Classification Using the MLC Algorithm

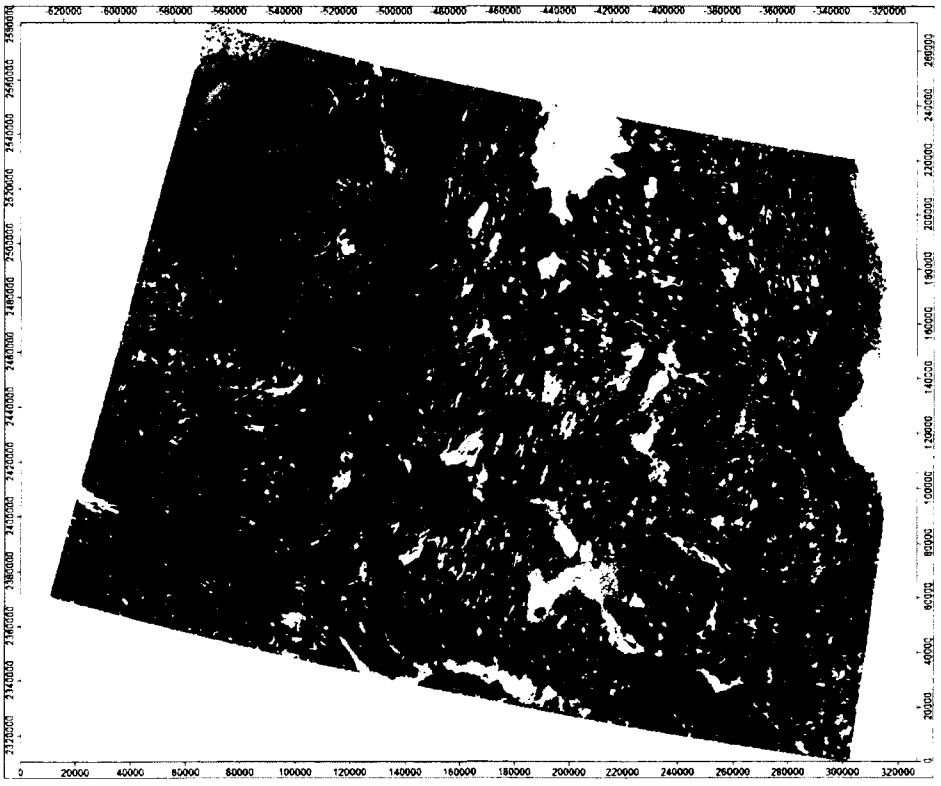


Figure 4.22: Classification using the RF algorithm without Index

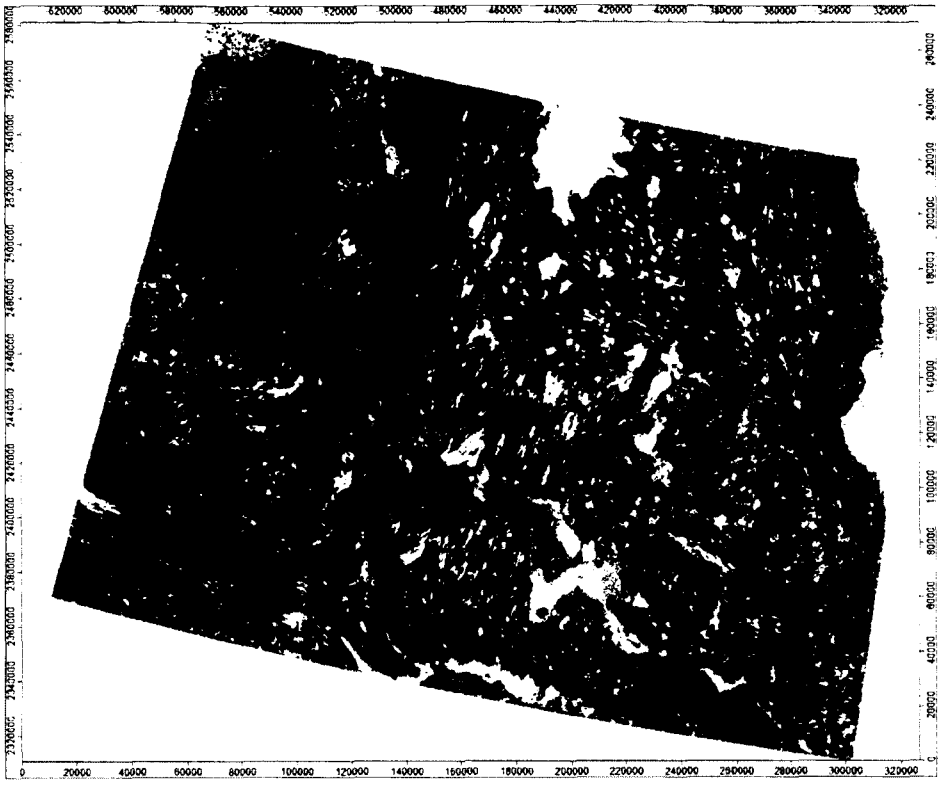


Figure 4.23: Classification using the RF algorithm with Index

The cross-validated accuracy of the RF model was ~30% higher than MLC. While some regions performed better using MLC, RF classified the bulk of the study area more accurately. RF probability maps clearly illustrated the image mosaic seams, resulting from scene balancing issues. These maps provide an important diagnostic for assessing problem areas in RF classification. This indicates that the model could not directly manage the classification issues that occurred from scene balancing. The model was, however, able to better quantify where uncertainty was the greatest. Changes in certainty across these boundaries were not systemic to all classes or the entire boundary, that some classes are more sensitive to scene balancing issues than others. While the inclusion of a nominal scene index class did not provide any notable increase

in overall accuracy, it led to a modest increase in percent landscape areas falling within each classification probability class (Figure 4.8 and Figure 4.9). These diagnostic plots provide a quick way in which analysts can evaluate mapping quality by using quantitative benchmarks to assess how prediction probabilities respond to additional variables or training information.

4.5 Conclusion

RF has proven to be a more robust classifier compared to MLC under non-parametric data conditions such as scene balancing issues typically encountered during synoptic surficial material mapping efforts across northern Canada. RF better manages non-parametric data sources and classes with higher accuracy measures. The RF model outperformed the MLC classification algorithm with a cross-validated result that is nearly 30% higher. The MLC classification produced the lowest cross-validated accuracy, with a value of 48.6%, whereas the RF produced a cross-validated accuracy of ~78%. The discrepancies between RF and MLC were demonstrated in the qualitative analysis across the study area. Often classes known to have some confusion between them were confused further when poor data was added along with the MLC classifier. Furthermore, by using MSRFC, RF was able to identify regions of high and low certainty for different classes.

There is significant room for improvement of the model, particularly in how it handles training areas and scene balancing. The RF model was not able to directly manage scene-balancing issues, nor did it completely solve the confusions associated

with classes in surficial material mapping. Nevertheless, the RF model allows researchers to improve understanding of the complexity of classification results and allows for the identification of certainty across the study area. Areas of uncertainty are of particular significance to surficial material mapping because often these classifications can serve as a tool for further field exploration. Thus, areas of particular importance can be identified by model deficiencies and possible sources. These areas could become targets for future field verification. RF, and the modified MSRFC, is better suited for classification of surficial materials than MLC.

Chapter 5: Conclusion

This thesis introduced Random Forest (RF) as an effective image classifier for managing common issues in surficial material mapping. The RF algorithm is more suitable for managing training area limitations such as unrepresentative sampling, genetically defined classes, and non-parametric distributions. RF can integrate multisource data and improve the management of spectral issues that are prevalent when creating data mosaics. The application of RF to spatial data introduces specific challenges that require additional steps in order to manage classification quality. Comparisons between RF and MLC show that RF performs better than MLC in both cross-validated accuracy and qualitative assessments. This thesis demonstrates numerous benefits of RF as an image classifier for mapping surficial materials in Canada's north.

Overall, the MSRFC has strong potential for operational implementation of machine-based predictive mapping of surficial materials. The MSRFC improves accuracy assessment through an iterative cross-validation procedure. MSRFC outlines a clear methodological approach for variable selection that is also stochastic and uses cross-validation accuracies to aid in selection. MSRFC does not rely on the internal accuracy assessment metrics and provides powerful diagnostics for classification assessment. These products allow researchers to explore classifications by identifying areas where the model does not work well or classes that cannot be determined spectrally.

RF manages common problems associated with large-scale surficial material

mapping issues. For example, bimodal distributions that commonly occur in regions when a study area crosses a lithological parent material, when different data sources are used, or when balancing of spectral datasets is poor. Past studies have explored the impact of training data on surficial material mapping using RCM. However, iterative approaches (e.g. bagging) can affect the outcome of the prediction depending on stability of the classifier used. The original RF algorithm was designed for complex statistical analysis and classification of non-spatial data. Therefore, implications of spatial data and procedures had to be explored for use in image classification.

The MSRFC workflow provides robust measures of classification accuracy, spatial uncertainty, and variable selection. The MSRFC introduces an additional set of methodological steps to use before and after the standard RF predictive model. This procedure still uses the same classification as RF; however, it provides a framework for variable selection, accuracy assessment, and model certainty. Variable selection is a key component of RF and was explored using the MSRFC methodology to provide researchers with variable importance assessment. MSRFC also managed sensitivity to sub setting for training and validation during pre-classification steps. Once key variables were identified using Gini Index, the predictive capacity of variables was investigated using an iterative cross-validation procedure that added variables by importance to assess their effect on accuracy of a final product.

New accuracy assessments were provided to estimate accuracy of classifications through stochastic iterations of cross-validation at varying percentages of training and validation. Additional post-classification model certainty metrics were delivered to

provide insight into the quality of classifications produced. Once a final classification was produced using all of the training data, the pre-classification stochastic accuracy estimations and post-classification probability maps measure accuracy and certainty for the model. Boxplots of class probability used during classification, model certainty maps, and land area versus percentage certainty charts provide an analysis framework for researchers to base their confidence in a predictive product. These steps outline the major advantages of using the new MSRFC in a predictive model, rather than just RF for classification of spatial information.

RF did not manage spectral balancing issues as desired in the objectives of this thesis. Inadequate training for each unique spectral region could be responsible. However, the variance between two spectral regions is too great without adequate pre-processing and balancing. Using additional nominal data for predictor variables had a minor impact in quantitative and qualitative comparisons. Overall, the MSRFC outperformed MLC both in quantitative and qualitative assessments. Further research is required to examine the MSRFC's limitations during classification, particularly as it pertains to training areas and variable selection. It became apparent that training area quantities directly influence the classification for uncertain pixels. For example, in Chapter 3, thick till class was the most commonly categorized class for uncertain pixels, and almost the only class categorized in pixels above a 90% probability. The model had a bias for abundant training data giving uncertain pixels classification preference, indicating some bias has occurred based on quantities of training data. This problem could be approached in two ways: 1) re-assessing how training area sizes and quantities

are defined; and 2) exploring how to improve the use of data in the RF model in order to reduce bias (e.g. sampling strategies within polygons). Both approaches require an exploration into more direct causes in order to understand and then address this issue.

The MSRFC workflow developed here can provide robust accuracy assessment and diagnostics tools compared to more conventional MLC and RCM classifiers used in surficial material mapping. Furthermore, it is non-parametric by design and therefore is useful to identify optimum predictor variables for image classification. Variable selection facilitated by MSRFC allows researchers to explore the potential of new data sources, and helps to reduce large data sets to more specific ones. Classification accuracy assessments help researchers to understand how sensitive the model is to the training and how much confidence can be placed in the final classification. The MSRFC adds more benefit to the application of the RF algorithm for classification of surficial materials. All of the benefits of RF and the MSRFC workflow make this approach highly suitable for classification of surficial materials in the Canadian North.

Bibliography

Andreassen, K., Laberg, J., & Vorren, T. (2008). Seafloor geomorphology of the SW Barents Sea and its glaci-dynamic implications. *Geomorphology*, 97(1-2), 157-177.

Andersen, C. M., & Bro, R. (2010). Variable selection in regression-a tutorial. *Journal of Chemometrics*, *24*(11-12), 728-737. doi:10.1002/cem.1360

Aylsworth J.M. and Shilts, W.W., 1989. Glacial features around the Keewatin Ice Divide, Districts of Mackenzie and Keewatin. Geol. Surv. Can., Pg., 88-24.

Beaubien, J., Cihlar, J., Simard, G., & Latifovic, R. (1999). Land cover from multiple thematic mapper scenes using a new enhancement-classification methodology Visual assessment. *Journal of Geophysical Research*, 104(D22), 27,909-27,920.

Bhatta, B. (2011). Remote Sensing and GIS (2nd ed.). New Delhi: Oxford University Press.

Boulesteix, A.-laure, Bender, A., Bermejo, J. L., & Strobl, C. (2011). Random forest Gini importance favors SNPs with large minor allele frequency. *Statistics*, (106).

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. (C. Hall Crc, Ed.)The Wadsworth statisticsprobability series (Vol. 19, p. 368).

Breiman, L. (1996). Bagging predictors. *Machine learning*, *140*, 123-140.

Breiman, L. (2001). Random forests. *Machine learning*, *9*, 29.

Brown, O., Harris, J. R., Utting, D., & Little, E. C. (2007). Remote predictive mapping of surficial materials on northern Baffin Island: developing and testing techniques

using Landsat TM and digital elevation data. *Geological Survey of Canada, Current Research* 2007-B1, 12.

Brown, D., Lusch, D., & Duda, K. (1998). Supervised classification of types of glaciated landscapes using digital elevation data. *Geomorphology*, 21(3-4), 233-250.

Bühlmann, P., & Yu, B. (2002). Analyzing bagging. *The Annals of Statistics*, *30*(4), 927-961.

Calle, M. L., & Urrea, V. (2011). Letter to the editor: Stability of Random Forest importance measures. *Briefings in bioinformatics*, *12*(1), 86-9.

Campbell, J. B. (1981). Spatial correlation effects upon accuracy of supervised classification of land cover. *Photogrammetric Engineering & Remote Sensing*, *47*(3), 355-363.

Campbell, J. B. (2003). *Introduction to remote sensing* (3rd ed.). London: Taylor and Francis.

Chen, D. (2002). The effect of training strategies on supervised classification at different spatial resolutions. *Photogrammetric Engineering and Remote Sensing*, *68*(11), 1155-1161.

Clark, C. D., Knight, J. K., & Gray, J. T. (2000). Geomorphological reconstruction of the Labrador Sector of the Laurentide Ice Sheet. *Quaternary Science Reviews*, 19, 1343-1366.

Conese, C. (1992). Use of error matrices to improve area estimates with maximum likelihood classification procedures. *Remote sensing of Environment*, 124, 113-124.

Congalton, R., & Mead, R. (1986). A Review of Three Discrete Multivariate Analysis Techniques Used in Assessing the Accuracy of Remotely Sensed Data from Error Matrices. *IEEE Transactions on Geoscience and Remote Sensing*, GE-24(1), 169-174.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Kyle, T., Gibson, J., & Lawler, J. J. (2007). Random Forests for Classification in Ecology. *Ecology*, 88(11), 2783-2792.

Douglas, G. V., & Douglas, M. C. V. (1949). Ungava (New Quebec) and Interior of Labrador, 3 vols.

Domack, E., Amblas, D., Gilbert, R., Brachfeld, S., Camerlenghi, A., Rebesco, M., Canals, M., et al. (2005). Subglacial morphology and glacial evolution of the Palmer deep outlet system, Antarctic Peninsula. *Geomorphology*, 75, 125 - 142.

Dorigo, W., Lucieer, A., Podobnikar, T., & Čarni, A. (2012). Mapping invasive Fallopia japonica by combined spectral, spatial, and temporal analysis of digital orthophotos. *International Journal of Applied Earth Observation and Geoinformation*, 19, 185-195.

Eastman, J. (2002). Bayesian Soft Classification for Sub-Pixel Analysis: A Critical Evaluation. *Photogrammetric Engineering and Remote Sensing*, *68*(I).

Eisank, C., Blaschke, T., & Götz, J. (2010). Developing A Semantic Model Of Glacial Landforms For Object- Based Terrain Classification – The Example Of Glacial Cirques. *Remote Sensing and Spatial Information Sciences*, 38(4).

Embelton, C., & King, C. A. M. (1975). *Glacial geomorphology*. Wiley. New York.

Environment Canada. (2012). Ecoregions of Canada, Coppermine River Upland.

Ecological Framework of Canada. Retrieved from http://ecozones.ca/english/region/68.html

Estes, J. E., Hajic, E. J., & Tinney, L. R. (1983). Fundamentals of Image Analysis: Anlysis of Visible and Thermal Infrared Data. In R. N. Colwell (Ed.), Manual of Remote Sensing (2nd ed., pp. 987-1124). Falls Church, Virginia: *American Society of Photogrammetry.*

Feller, W. "The Strong Law of Large Numbers." §10.7 in An Introduction to Probability Theory and Its Applications, Vol. 1, 3rd ed. New York: Wiley, pp. 243-245, 1968.

Foody, G. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, 80(1), 185-201.

Foody, G., & Cox, D. (1994). Sub-pixel land cover composition estimation using a linear mixture model and fuzzy membership functions. *International Journal of Remote Sensing*, 15(3), 619–631. Taylor and Francis.

Foody, G. & Mathur, A. (2004). Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification. *Remote Sensing of Environment*, *93*(1-2), 107-117.

Ford, J. P. (1984). Landforms from Seasat Radar Images. Quaternary Research, 22, 314-321.

Fulton, R. J. (1995). Surficial materials of Canada / Matera superfi- cies du Canada. Geological Survey of Canada, Map 1880A, map scale 1 : 5 000 000.

Gislason, P., Benediktsson, J., & Sveinsson, J. (2006). Random Forests for land cover classification. *Pattern Recognition Letters*, *27*(4), 294-300.

Gao, J. (2009). Digital Analysis of Remotely Sensed Imagery. McGraw-Hill Companies.

Graff, L. H., & Usery, E. L. (1993). Automated of Generic Classification Terrain Features in Digital Elevation Models. *Society*, 59(9).

Grunsky, E., Harris, J., & McMartin, I. (2009). Predictive Mapping of Surficial Materials, Schultz Lake Area (NTS 66A), Nunavut, Canada. *Reviews in Economic Geology*, 16, 177-198.

Harris, J. R., Grunsky, E. C., He, J., Gorodetzky, D., & Brown, N. (2012)a. A robust, cross-validation classification method (RCM) for improved mapping accuracy and confidence metrics. *Canadian Journal of Remote Sensing*, *38*(1), 69-90.

Harris, J. R., Parkinson, W., Dyke, A., Kerr, D., Russell, H., Eagles, S., Richardson, M., et al. (2012)b. Predictive surficial geological mapping of Hall Peninsula and Foxe Basin Plateau, Baffin Island using LANDSAT and DEM data.

Healey, S., Cohen, W., Zhiqiang, Y., & Krankina, O. (2005). Comparison of Tasseled Cap-based Landsat data structures for use in forest disturbance detection. *Remote Sensing of Environment*, 97(3), 301-310.

Henderson, J.F., 1944. Mackay Lake, District of Mackenzie, Northwest Territories. *Geological Survey of Canada*, Map 738A.

Heroy, D. C., & Anderson, J. B. (2005). Ice-sheet extent of the Antarctic Peninsula region during the Last Glacial Maximum (LGM)—Insights from glacial geomorphology. *Geological Society of America Bulletin*, 117(11), 1497.

Irvin, B., Ventura, S., & Slater, B. (1997). Fuzzy and isodata classification of landform elements from digital terrain data in Pleasant Valley, Wisconsin. *Geoderma*,

77(2-4), 137-154.

Kauth, R. J. and Thomas, G. S. (1976). The tasseled cap-A graphic description of the spectral-temporal development of agricultural crops as seen by Landsat, Proc. the Symposium on Machine Processing of Remotely Sensed Data, pp.4b-41 -4b-50 1976

Khalyani, A., Falkowski, M., & Mayer, A. (2012). Classification of Landsat images based on spectral and topographic variables for land-cover change detection in Zagros forests. *International Journal of Remote Sensing*, (August 2012), 37-41.

Kleinbaum, D. G., & Klein, M. (2010). Maximum Likelihood Techniques: An Overview. Logical Regression (pp. 103-127). New York, NY: Springer New York.

Kor, P. S. G., Shaw, J., & Sharpe, D. R. (1991). Erosion of bedrock by subglacial meltwater , Georgian Bay , Ontario : a regional view1. *Canadian Journal of Earth Sciences*, 28, 623-642.

Kerr, D.E., Ward, B.C., and Dredge, L.A. 1995. Surficial geology, Winter Lake, District of Mackenzie, Northwest Territories; *Geological survey of Canada*, Map 1871A, (South half), scale 1:250,000.

Kuhnert, P. M., Henderson, A.-kinsey, Bartley, R., & Herr, A. (2010). Incorporating uncertainty in gully erosion calculations using the random forests modelling approach. *Environmetrics*, (August 2008), 493-509.

Labovitz, M. L., & Masuoka, E. J. (1984). The influence of autocorrela- tion in signature extraction-An example from a geobotanical investigation of Cotter Basin. *International Journal of Remote Sensing*, 5(2), 315-332.

Lambin, E. F., & Strahlers, A. H. (1994). Change-vector analysis in multitemporal space: A tool to detect and categorize land-cover change processes using high temporal-resolution satellite data. *Remote Sensing of Environment, 48*(2), 231-244.

LaRocque, A., Leblon, B., Shelat, Y., & Harris, J. (2011). Use of multi-beam RADARSAT-2 dual-polarization C-HH and C-HV imagery for surficial geology mapping in Nunavut, Canada. *geohydro2011.ca*.

Legendre, P. (1993). Spatial Autocorrelation : Trouble or New Paradigm. *Ecology, 74*(6), 1659-1673.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News, 2(3), 18-22. Retrieved from http://cran.r-project.org/doc/Rnews/

Lillesand, T. M., & Kiefer, R. W. (1999). *Remote sensing and image interpretation* (4th ed.). New York ;Toronto: John Wiley & Sons.

Lowry, J., Ramsey, R., Boykin, K., & Bradford, D. (2005). Final report on landcover mapping methods. USGS.

Maselli, F., Conese, C., & Petkov, L. (1992). Inclusion of prior probabilities derived from a nonparametric process into the maximum-likelihood classifier. *Engineering and remote sensing*, 58(2), 201-207.

Maselli, F., Conese, C., & Zipoli, G. (1990). Use of error probabilities to improve area estimates based on maximum likelihood classifications. *Remote Sensing of Enviornment*, 160(August 1989), 155-160.

Mei, S., & Paulen, R. (2009). Using multi-beam RADARSAT-1 imagery to augment mapping surficial geology in northwest Alberta , Canada, 35(1), 1-22.

Miao, X., & Heaton, J. S. (2010). A comparison of random forest and Adaboost tree in ecosystem classification in east Mojave Desert. *Geoinformatics, 2010 18th International Conference on* (pp. 1–6).

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. ISPRS Journal of Photogrammetry and Remote Sensing, 66(3), 247-259. Elsevier B.V.

Muchoney, D. (2002). Pixel-and site-based calibration and validation methods for evaluating supervised classification of remotely sensed data. *Remote Sensing of Environment*, *81*, 290-299.

Odgers, N. P., McBratney, A. B., & Minasny, B. (2011). Bottom-up digital soil mapping. Soil layer classes. *Geoderma*, 163(1-2), 38-44.

Olthof, I., Butson, C., Fernandes, R., Fraser, R., Latifovic, R., & Orazietti, J. (2005). Landsat ETM+ mosaic of northern Canada. *Canadian Journal of Remote Sensing*, 31(5), 412-419.

Padgham, W.A., 1991. The Slave Province: an overview. in Mineral Deposits of the Slave Province, Northwest Territories, Padgham, W.A. and Atkinson, D. (eds.), *Geological Survey of Canada*, Open File 2168, p. 1 - 40.

Padgham, W.A. and Fyson, W.K., 1992. The Slave Province: a distinct Archean craton. Canadian Journal of Earth Sciences, v. 29, p. 2066-2071.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1), 217-222.

Parkinson, W., Richardson, M., & Russell, H. (2010). Defining Eskers for Classification within Object-Based Image Analysis Framework. *geohydro2011*.

Pennock, D., Zebarth, B., & Dejong, E. (1987). Landform classification and soil distribution in hummocky terrain, Saskatchewan, Canada. *Geoderma*, 40(3-4), 297-315.

Parry, J. T. (1967). Geomorphology in Canada. *The Canadian Geographer/Le Géographe canadien*, 11(4), 280-311.

Pike, R. J. (1988). The geometric signature: Quantifying landslide-terrain types from digital elevation models. *Mathematical Geology*, *20*(5), 491-511.

Pontius, R. G., & Cheuk, M. L. (2006). A generalized cross - tabulation matrix to compare soft classified maps at multiple resolutions. International Journal of Geographical Information Science, 20(1), 1-30.

Pringle, M. J., Denham, R. J., & Devadas, R. (2012). Identification of cropping activity in central and southern Queensland, Australia, with the aid of MODIS MOD13Q1 imagery. *International Journal of Applied Earth Observation and Geoinformation*, 19(February 2000), 276-285. Elsevier B.V.

Rabben, E. L. (1960). Fundamentals of Photo Interpretation. In R. N. Colwell (Ed.), Manual of Photographic Interpretation (1st ed., pp. 99-168). Washington, D.C.: *American Society of Photogrammetry*.

Rafaelsen, B., Andreassen, K., Kuilman, L. W., Lebesbye, E., Hogstad, K., & Midtbo, M. (2002). Geomorphology of buried glacigenic horizons in the Barents Sea from three-dimensional seismic data. Geological Society, London, Special Publications,

203(1), 259-276.

Rocchini, D., Foody, G. M., Nagendra, H., Ricotta, C., Anand, M., He, K. S., Amici, V., et al. (2012). Uncertainty in ecosystem mapping by remote sensing. *Computers & Geosciences*. Elsevier.

Samaniego, L., & Schulz, K. (2009). Supervised Classification of Agricultural Land Cover Using a Modified k-NN Technique (MNN) and Landsat Remote Sensing Imagery. Remote Sensing, 1(4), 875-895.

Sesnie, S. E., Gessler, P. E., Finegan, B., & Thessler, S. (2008). Integrating Landsat TM and SRTM-DEM derived variables with decision trees for habitat classification and change detection in complex neotropical environments. *Remote Sensing of Environment*, *112*(5), 2145-2159.

Serra, P., Pons, X., & Saurí, D. (2003). Post-classification change detection with data from different sensors: Some accuracy considerations. *International Journal of Remote Sensing, 24*(16), 3311-3340.

Scambos, T. a., Haran, T. M., Fahnestock, M. a., Painter, T. H., & Bohlander, J. (2007). MODIS-based Mosaic of Antarctica (MOA) data sets: Continent-wide surface morphology and snow grain size. *Remote Sensing of Environment*, 111(2-3), 242-257.

Schneider, S. R., Mcginnis, D. F., & Pritchard, J. A. (1979). Use of Satellite Infrared Data for Geomorphology Studies. Remote Sensing of Environment, 8, 313-330.

Singh, A. (1989). Review article digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing, 10*(6), 989-1003.

Sinha G, Mark D M (2010). Cognition-based extraction and modelling of topographic eminences. *Cartographica: The International Journal for Geographic Information and Geovisualization*, Vol. 45(2), pp. 105-112. DOI: 10.3138/carto.45.2.105.

Smith, M. J., & Pain, C. F. (2009). Applications of remote sensing in Geomorphology. Progress in Physical Geography, 33(4), 568-582.

Shaw, J., Sharpe, D., & Harris, J. (2010). A flowline map of glaciated Canada based on remote sensing data. *Canadian Journal of Earth Sciences*, 47(1), 89-101.

Sharpe, D. R. (1991). *Glacial Sediments and Landforms*. University of Ottawa.

Stewart, W. M. (1916). David Thompson's Surveys in The North-West. *University of Toronto Press*, 2, 289-303.

Stokes, C. R., Clark, C. D., & Winsborrow, M. C. M. (2006). Subglacial bedform evidence for a major palaeo-ice stream and its retreat phases in Amundsen Gulf, Canadian Arctic Archipelago. *Journal of Quaternary Science*, 21(4), 399-412.

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC bioinformatics*, *8*, 25.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychological methods, 14(4), 323-48. *American Psychological Association*.

Sugden, D. E. (1978). Glacial Erosion by the Laurentide Ice Sheet. *Journal of Glaciology*, 20(83), 367-391.

Teng, W. L. (1997). Fundamentals of Photographic Interpretation. In W. R. Philipson (Ed.), Manual of Photographic Interpretation (2nd ed., pp. 49-116). Bethesda, Maryland: *American Society for Photogrammetry and Remote Sensing*.

Torbick, N., Persson, A., Olefeldt, D., Frolking, S., Salas, W., Hagen, S., Crill, P., et al. (2012). High Resolution Mapping of Peatland Hydroperiod at a High-Latitude Swedish Mire. *Remote Sensing*, 4(7), 1974-1994.

Waldhoff, G., Bubenzer, O., Bolten, A., Koppe, W., & Bareth, G. (2008). Spectral Analysis Of Aster , Hyperion , And Quickbird Data For Geomorphological And Geological Research In Egypt (Dakhla Oasis , Western Desert)The International Archives of the Photogrammetry, *Remote Sensing and Spatial Information Sciences*, 37(B8), 1201-1206.

Western, A. W., Grayson, R. B., Bschl, G., Willgoose, G. R., & McMahon, T. A. (1999). Observed spatial organization of soil moisture and its relation to terrain indices. Water Resources Research, 35(3), 797.

Wilson, J. T. (1939). Eskers North-east of Great Slave Lake. Trans. *Royal Society of Canada*, 33(3), 119-130.

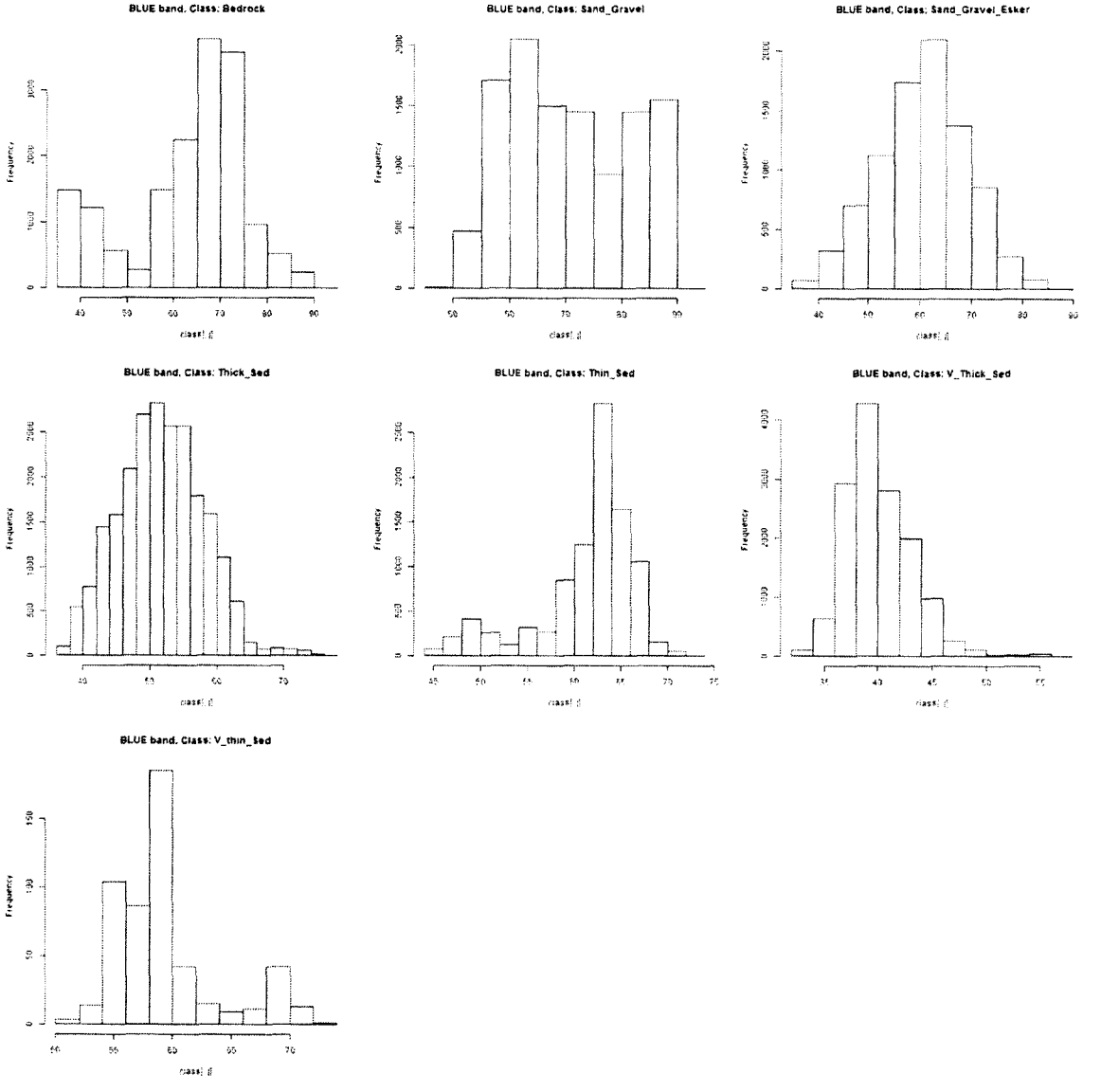
Wilson, J.P., Gallant, J.C., [Eds.] (2000): 'Terrain analysis - principles and applications', New York, John Wiley & Sons, Inc.

Yang, Z. R. (2010). Machine Learning Approaches to Bioinformatics (pp. 120-132). Singapore: *World Scientific Publishing* Co. Pte. Ltd.

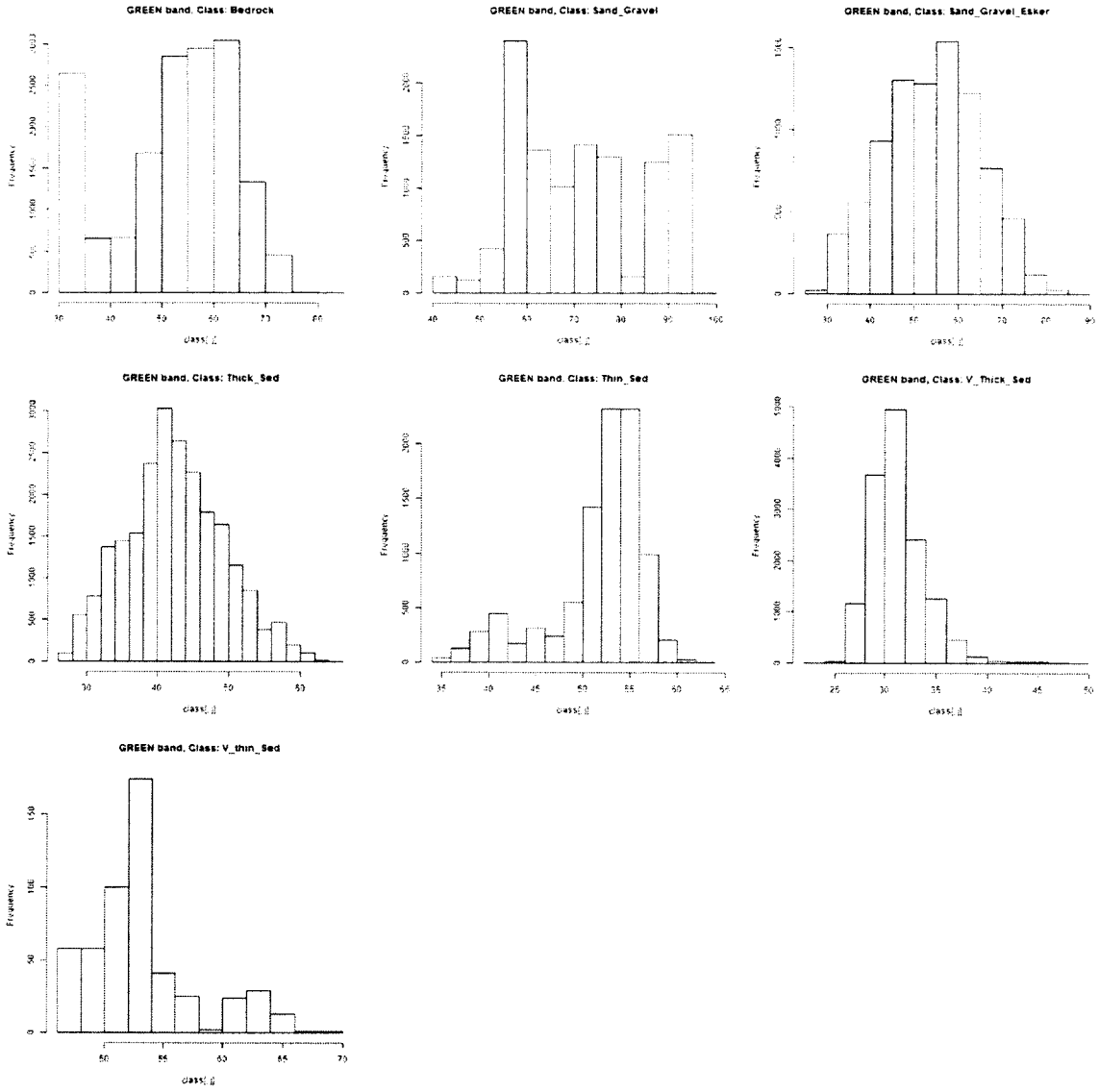
Yong, D., Cihlar, J., Beaubien, J., & Latifovic, R. (2001). Radiometric Normalization, Compositing, and Quality Control for Satellite High Resolution Image Mosaics over Large Areas. *IEEE Transactions on Geoscience and Remote Sensing*, 39(3), 623-634.

Appendix A

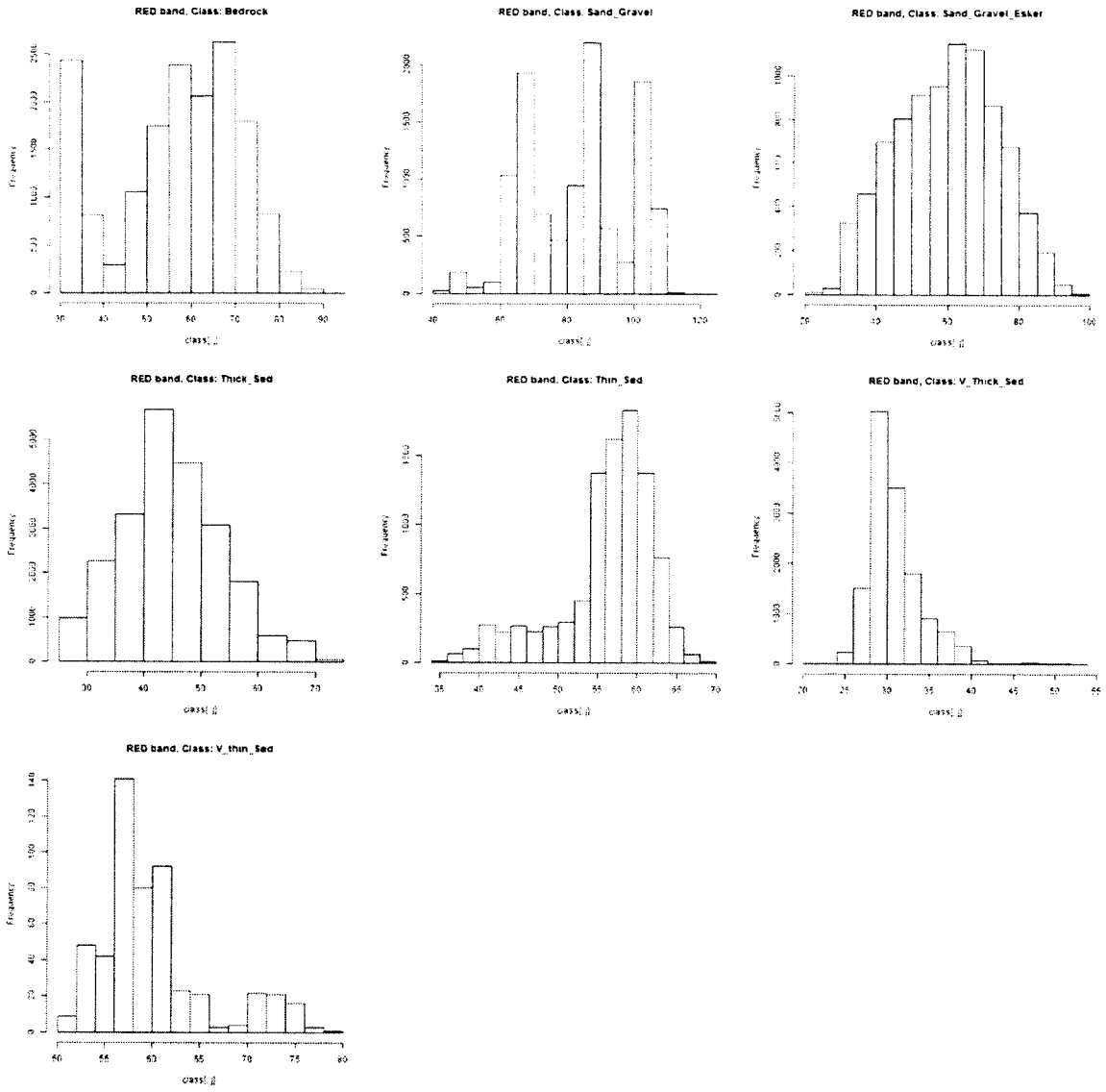
This section contains the distributions by Band for each class used in chapter 4. This is included to demonstrate that these classes are not all parametric and therefore should not be used with parametric statistics.



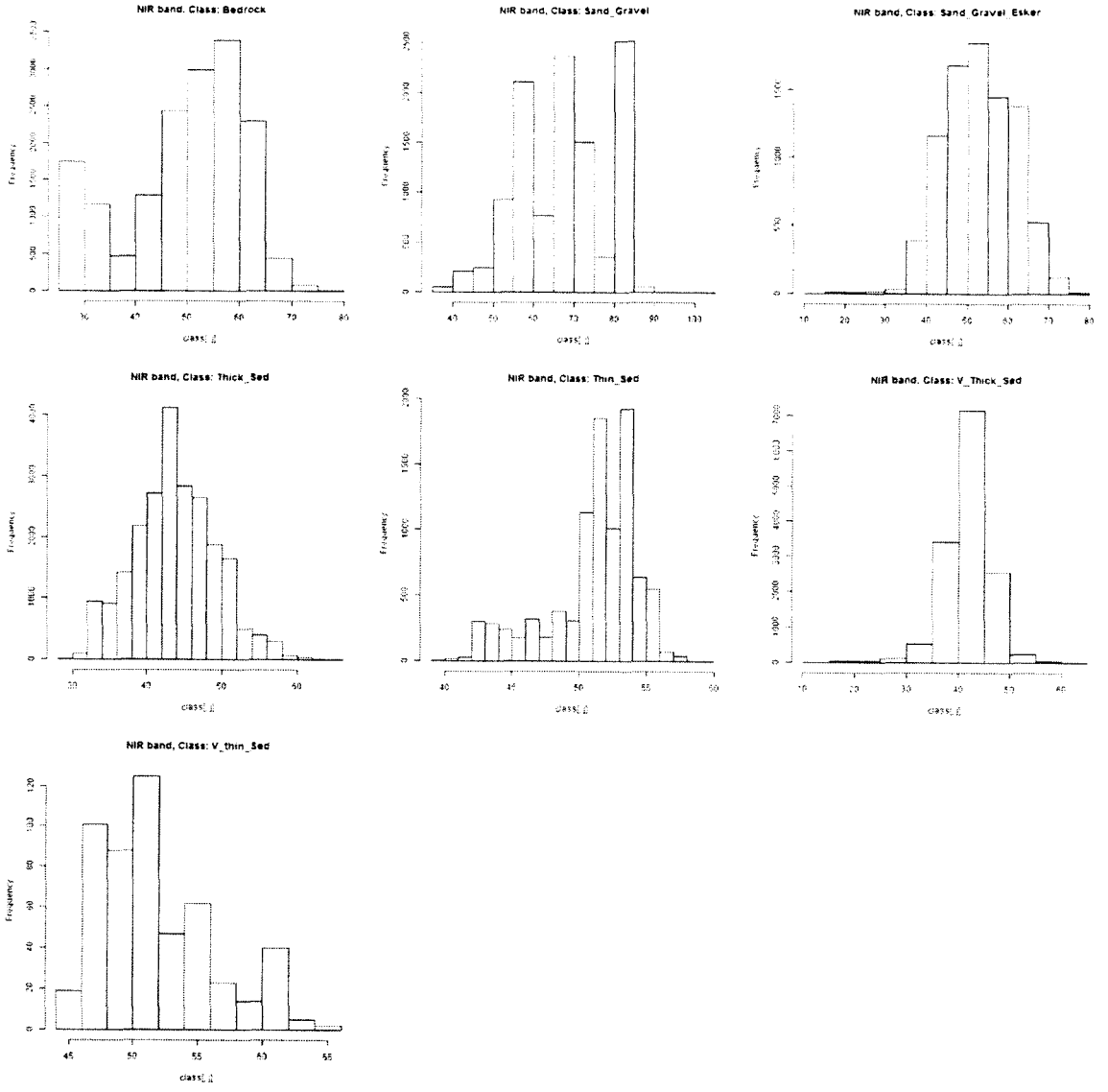
Landsat ETM+ blue band distributions



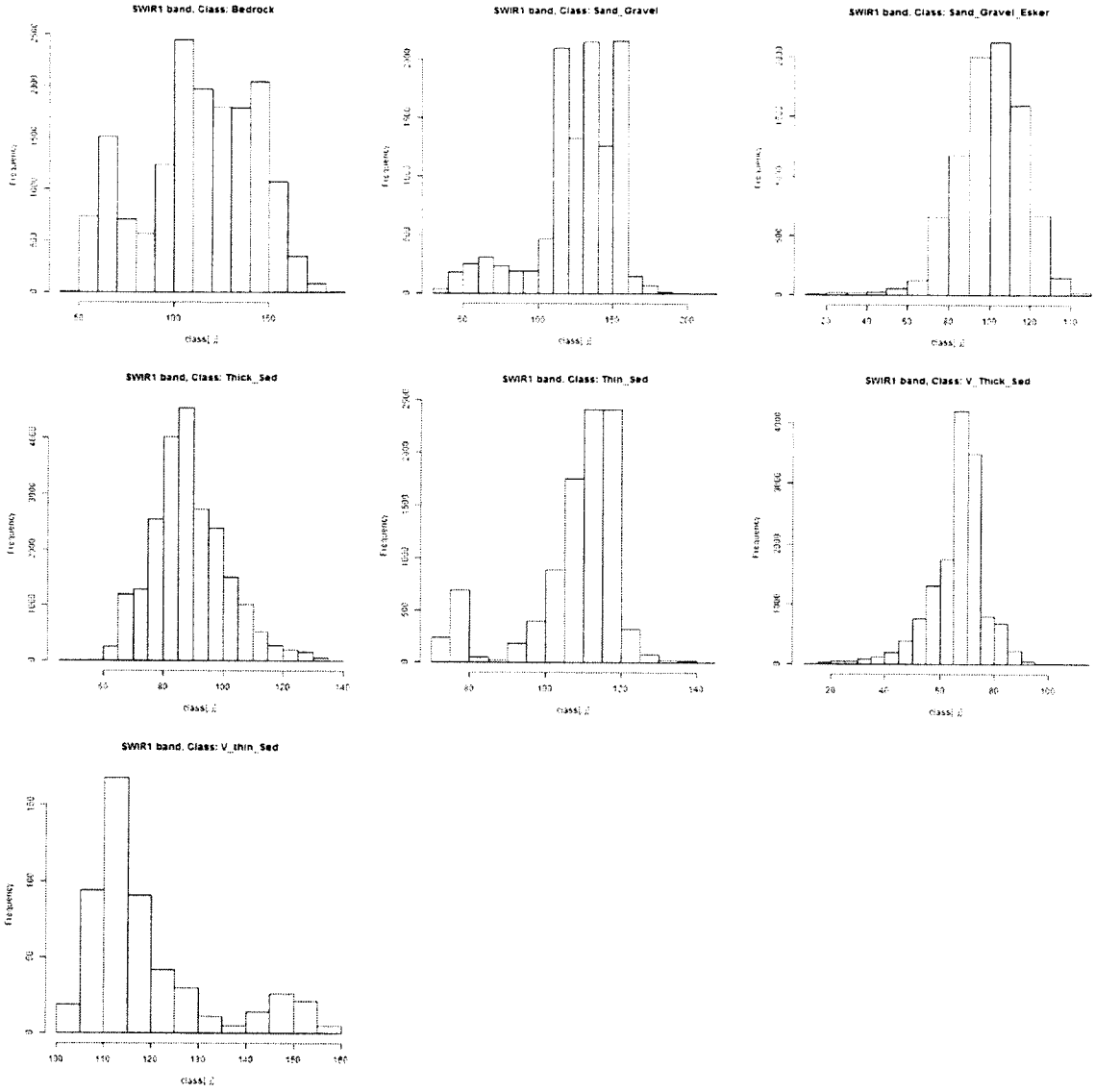
Landsat ETM+ green band distributions



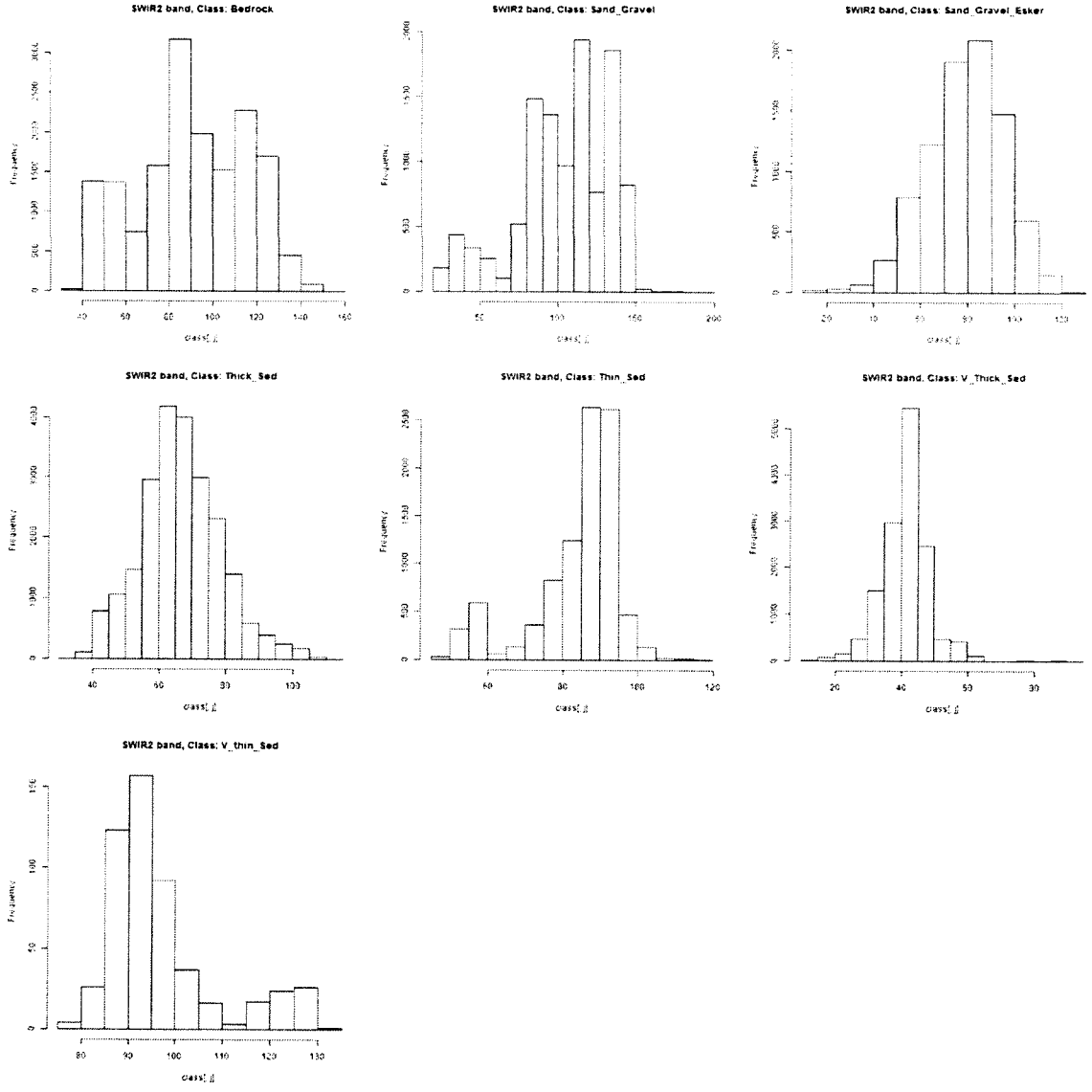
Landsat ETM+ red band distributions



Landsat ETM+ NIR band distributions



Landsat ETM+ SWIR1 band distributions



Landsat ETM+ SWIR2 band distributions