Geoderma 214–215 (2014) 141–154

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Predictive soil parent material mapping at a regional-scale: A Random Forest approach

Brandon Heung [a], Chuck E. Bulmer [b], Margaret G. Schmidt [a,*]

[a] Soil Science Lab, Department of Geography, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada
[b] British Columbia Ministry of Forests Lands and Natural Resources Operations, Forest Sciences Section, Vernon, BC, V1B 2C7, Canada

ARTICLE INFO

Article history:
Received 1 June 2013
Received in revised form 15 September 2013
Accepted 16 September 2013
Available online 11 October 2013

Keywords:
Soil parent material
Digital soil mapping
Knowledge discovery
Random forest
Environmental correlation
Machine learning

ABSTRACT

In this study, we evaluate the application of a Random Forest (RF) classifier as a tool for understanding and predicting the complex hierarchical relationships between soil parent material and topography using a digital elevation model (DEM) and conventional soil survey maps. Single-component soil polygons from conventional soil survey maps of the Langley–Vancouver Map Area, British Columbia (Canada), were used to generate randomized training points for 9 parent material classes. Each point was intersected with values from 27 topographic indices derived from a 100 m DEM. RF's m_{try} parameter was optimized using multiple replicates of 5-fold cross validation and parent material predictions were made for the region. Predictive parent material maps were validated through comparisons with legacy soil survey maps and 307 field points. Results show that predictions made by a non-optimized RF resulted in a kappa index of 89.6% when validated with legacy soil survey data from single-component polygons and a kappa index of 79.5% when validated with field data. Variable reduction and m_{try} optimization resulted in minimal improvements in RF predictions. Our results demonstrate the effectiveness of RF as a machine learning and data mining approach; however, the need for reliable training data was highlighted by less reliable results for polygon disaggregation in portions of the map where fewer training data points could be established.

1. Introduction

Soil parent material is the initial state of the soil system and the material from which soils are derived (Jenny, 1941). Soil type, soil development and the physical and chemical properties of soils are influenced by parent material. Information on parent material and its texture is recognized as a useful factor in soil erosion (Heung et al., 2013; le Roux et al., 2007; Weaver, 1991) and would also be beneficial to the evaluation of forest and agriculture productivity potential, hydrologic characteristics of watersheds, suitability of materials for construction and the assessment of terrain stability. Furthermore, information on soil parent material may also be used for predictive ecosystem mapping (MacMillan et al., 2007) and digital soil mapping studies (McBratney et al., 2003).

Soil parent material is the result of geomorphic processes interacting with bedrock over long periods of time. In British Columbia, glaciation during the Pleistocene Epoch was a dominant process in the evolution of the modern landscape, where the majority of parent materials in the region now consist of unconsolidated sediments deposited on the land surface by ice, gravity, water and wind (Howes and Kenk, 1997; Luttmerding, 1981). The geomorphic processes of erosion and deposition that were active at a particular location during glacial, post glacial, and modern times have also created a mosaic of distinct landforms across the region where a close association exists between the topographic landscape form and the characteristics of the unconsolidated parent material. Parent materials are classified in this area, and throughout Canada, based on their mode of formation and transport (Howes and Kenk, 1997).

The majority of digital soil mapping studies reviewed by McBratney et al. (2003) used bedrock geology as a surrogate predictor for parent material — an approach that may be adequate for environments where the soils are predominantly derived from residual materials. But for environments influenced by glaciation and where geomorphic transport processes have significantly influenced the nature and distribution of parent material, bedrock geology, when used alone, likely provides an incomplete depiction of the influence of parent materials on soil properties. Consequently, transported parent materials may not be represented and the resulting maps would potentially become biased in favor of residual materials (Lawley and Smith, 2008). For these reasons, improving the quality and accuracy of digital soil maps in glaciated areas require more detailed parent material maps that have been derived with the consideration of transport processes.

Conventional soil maps and other resource inventories are commonly developed by delineating map units based on climate, ecological features, topography, parent material, bedrock geology, soil, and vegetation (Resource Inventory Committee, 1998). The importance of parent material as a soil-environmental variable is illustrated by the use of this variable as both a fundamental and distinguishing characteristic between soil types at all mapping scales. There is an especially strong relationship between map unit boundaries and topography since the

* Corresponding author. Tel.: +1 778 782 3323; fax: +1 778 782 5841.
E-mail addresses: bha4@sfu.ca (B. Heung), Chuck.Bulmer@gov.bc.ca (C.E. Bulmer), Margaret_Schmidt@sfu.ca (M.G. Schmidt).

0016-7061/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.geoderma.2013.09.016

topography reflects the dominant geomorphic process and parent material characteristics (Hole and Campbell, 1985) and also because topography has a significant influence on vegetation and other ecological attributes that are often of interest to map makers. In addition, soil surveys are commonly based on aerial photo interpretation and the boundaries of the mapping units are determined based from the external expression of soil-environmental variables on the landscape (Beckett, 1971; Webster and Wong, 1969). Therefore, the derived map units on a conventional soil map tend to contain soil types with a defined set of parent material attributes while also maintaining a close association with topographic features in the landscape.

In British Columbia, the only comprehensive soil parent material map is the national level Soil Landscapes of Canada geographic dataset (SLC; Schut et al., 2011). The SLC database consists of 12,728 multi-component map units, with multiple taxonomic soil classes, that are generalized from detailed soil surveys and are mapped at a 1:1,000,000 scale (Geng et al., 2010). Despite having a consistent map database and comprehensive geographic coverage, the use of such highly aggregated polygon data may not be appropriate for mapping the spatial patterns of parent materials at regional or local scales. At regional-scales, existing soil surveys are available for many, but not all, parts of British Columbia and may be used to obtain information on parent materials. Examples include the 1:25,000 and 1:50,000 scale maps for the Langley–Vancouver Area (Luttmerding, 1981); 1:126,720 scale maps for the Tulameen Area (Lord and Green, 1974); and 1:20,000 scale maps for the Okanagan and Similkameen Valleys (Wittneben, 1986) with areal extents of approximately 5472 km^2, 4008 km^2, and 3895 km^2, respectively. In addition, other sources of parent material information may be taken from surficial and bedrock geology maps such as those for the New Westminster Area (Armstrong, 1957) and the Vancouver Area (Armstrong, 1956); however, such examples were mapped at smaller spatial extents in comparison to soil surveys.

Extracting the knowledge from existing soil maps is complicated because such maps typically include a large number of multi-component map units, and therefore lack a spatially explicit representation of the soil's class and attributes (Hole and Campbell, 1985; Webster and Beckett, 1968; Zhu and Band, 1994). Despite these spatial challenges, soil maps still have the potential to provide useful information about soil–landscape relationships (Bui, 2004; McBratney et al., 2003); for instance, Bui and Moran (2001) have previously used the map units from soil surveys to train the C5.0 decision tree algorithm and to validate the algorithm's outputs — an approach that was further extended in subsequent studies that mapped the soils of the Murray–Darling Basin, Australia (Bui and Moran, 2003; Moran and Bui, 2002).

Decision trees are data mining, machine learning, and rule-induction algorithms that classify data by inferring the relationships between a dependent variable and a set of predictors (Bui and Moran, 2001). They consist of nodes and leaves where each node represents an *if–then* statement and the leaves are terminal nodes where a decision is made with respect to the class variable (Breiman et al., 1984). The aim of a tree-based model is to examine all predictors in order to identify optimal node splitting rules where the within-node homogeneity is maximized. However, the manner in which the splits are made is dependent on the tree-splitting algorithm that is used.

The decision tree modeling approach has many advantages. Firstly, it is a particularly useful modeling approach for handling non-parametric data where the predictors are not characterized as having a specific distribution (Breiman et al., 1984). Secondly, decision trees are not sensitive to missing data, to the inclusion of irrelevant predictors, or to the presence of outliers. Furthermore, decision trees operate effectively using numerical, ordinal, binary, and categorical datasets. Finally, decision trees are well suited for identifying complex hierarchical relationships between predictors and response variables, as well as the relationships between predictors (Díaz-Uriarte and Alvarez de Andrés, 2006; Hastie et al., 2009). Decision trees have been used extensively to map soil classes (Bui and Moran, 2001, 2003; Bui et al., 1999; Grinand et al., 2008; Moran and Bui, 2002; Scull et al., 2005); soil properties such as pH (Henderson et al., 2005), organic C, % clay, and total N and P (Bui et al., 2006, 2009), or natural drainage (Lemercier et al., 2012). In addition, decision trees have also been used for the purposes of predictive soil parent material mapping (ie. Bui and Moran, 2001; Lacoste et al., 2011; and Lemercier et al., 2012); however, further evaluation of these methods would be valuable and incorporating detailed predictions of the distribution of parent materials, based on topographic characteristics, would likely help such efforts.

The Random Forest (RF) classifier is conceptually similar to a decision tree; except, an ensemble of decision trees are combined in order to improve the classification accuracy (Breiman, 2001; Cutler et al., 2007). For each decision tree in the RF, a random selection of predictors and training points are used to identify splits when building the tree. RF, a hierarchical non-parametric modeling approach, shares similar model advantages to decision trees (ie. insensitive to missing data, to the inclusion of irrelevant predictors and outliers, and is flexible with various types of datasets); however, RF provides a stronger prediction as it is less susceptible to over-fitting and it provides a better error measurement in comparison to decision trees (Breiman, 2001). Furthermore, RF has the advantage of incorporating 'randomness' into its predictions through reiterative bootstrap sampling and randomized variable selection when generating each decision tree. Additional characteristics of RF include its ability to provide variable importance measures and its ability to provide good predictions when noisy training data is used (Hua et al., 2005).

RF has widely been used in the field of bioinformatics (ie. Díaz-Uriarte and Alvarez de Andrés, 2006; Qi, 2012; Statnikov et al., 2008; Svetnik et al., 2003, 2004). In ecology, examples of studies that have used RF include the mapping of tree species distribution (ie. Prasad et al., 2006); land cover classification (ie. Gislason et al., 2006); ecological classification (ie. Cutler et al., 2007); the mapping of soil organic matter (Grimm et al., 2008; Wiesmeier et al., 2011); and soil texture (Ließ et al., 2012). With the exception of Häring et al. (2012) in using RF to disaggregate multi-component soil polygons, RF has not been used extensively for mapping categorical soil properties such as soil taxonomic units or parent materials.

The objectives of this study were to first evaluate the methods for extracting training data from soil survey data and the optimization of RF parameters; then, to test the reliability of using the RF classifier within single-component polygons in learning the relationship between parent material and topography; and finally, to evaluate RF as a potential method for disaggregating multi-component parent material polygons. The approach was based on the assumption that changes in parent material were closely associated with changes in topography; and hence, all environmental covariates were derived from a digital elevation model (DEM) at a 100 m spatial resolution. The proposed approach, may be extended to other resource inventory mapping studies such as ecosystem mapping (Resource Inventory Committee, 1998) and forest inventory mapping (Natural Resources Canada, 2004) where conventional mapping also uses a combination of single and multi-component map units.

2. Methods

The workflow for this study is based on the integration of a DEM and conventional soil survey maps for the development of training data; RF for modeling the hierarchical relationships between parent material and topography; and the use of point data and a conventional soil survey map for assessing model outputs (Fig. 1). In order to select suitable training areas, the map units from a conventional soil survey map were first separated into two categories: map units with a single

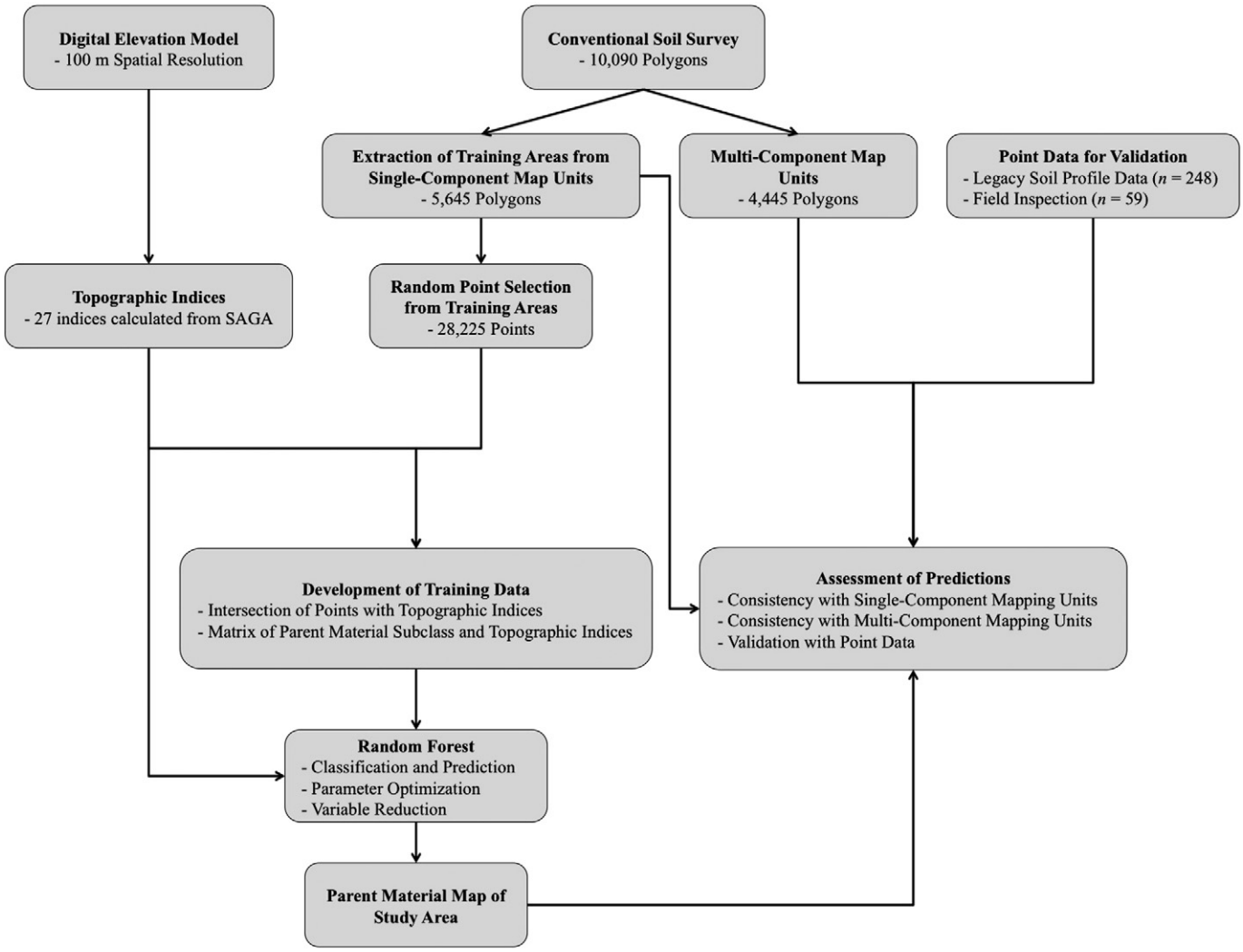

Fig. 1. Workflow diagram of predictive soil parent material mapping using a digital elevation model, conventional soil survey data, and random forest algorithm. Digital elevation model is used to generate topographic indices; the conventional soil survey is used to train the random forest model and to assess model outputs.

parent material (single-component) used as training areas and map units with multiple parent materials (multi-component). To produce a topography–parent material matrix for submission into the RF classifier, random points were generated within training areas and intersected with a suite of topographic indices derived from a DEM of the study area (see Section 2.3). Using the inputted matrix, the RF parameters were optimized and a variable reduction procedure was tested (see Section 2.4). The output of the RF classifier was a parent material map of the study area, which was then assessed using the original soil survey map and also external point data (see Section 2.5). In addition, the ability of RF to disaggregate polygons with multiple parent material components was assessed using the multi-component map units that are not used in the development of the training dataset (see Section 2.5.2).

2.1. Study area

The 5472 km^2 study area ranges from 49°00′ N to 49°56′ N latitude and 121°16′ W to 123°11′ W longitude with an elevational range of 0–2555 m above mean sea level and located in the Coastal Western Hemlock biogeoclimatic zone (Fig. 2) (Pojar et al., 1991). The zone receives mean annual precipitation of 2228 mm where snowfall constitutes less than 15% of the precipitation. The study area consists of the Lower Fraser Valley, which has predominantly an agricultural and urban land coverage, and includes portions of the predominantly forested Coastal Mountain Range located along the northern region of the area.

The pre-existing soil survey identifies 139 distinct soil series with 9 mineral parent material classes (Luttmerding, 1981). Organic parent materials are found in depressions and cover 6% of the landscape; however, they were not predicted for this study. Although the distribution of organic parent materials is affected by topography, these parent materials are also strongly dependent on climatic as well as vegetative factors, which were not included in this study. In the soil survey, the parent material classes were subdivided into 20 subclasses (Table 1). In this glaciated landscape there are very few residual materials except at high elevation, and parent materials are almost exclusively derived from the depositional and erosive processes of glaciation, gravity, wind and water. At low elevations, fluvial material is the dominant parent material; however, both marine and glaciomarine materials are also common. At higher elevations, morainal materials are the dominant parent material.

Most of the Lower Fraser Valley is underlain by sedimentary rocks from the Cretaceous period (and younger) with approximately 30 m to 150 m of unconsolidated deposits overlying the bedrock (Armstrong, 1957; Valentine et al., 1978). Due to glacial advance during the Pleistocene, ice accumulation with a thickness of 2500 m resulted in the submergence of land into the Pacific Ocean. Both glacial till and glaciofluvial materials were deposited over large areas during this time and during the subsequent ice retreat. As a result of the melting

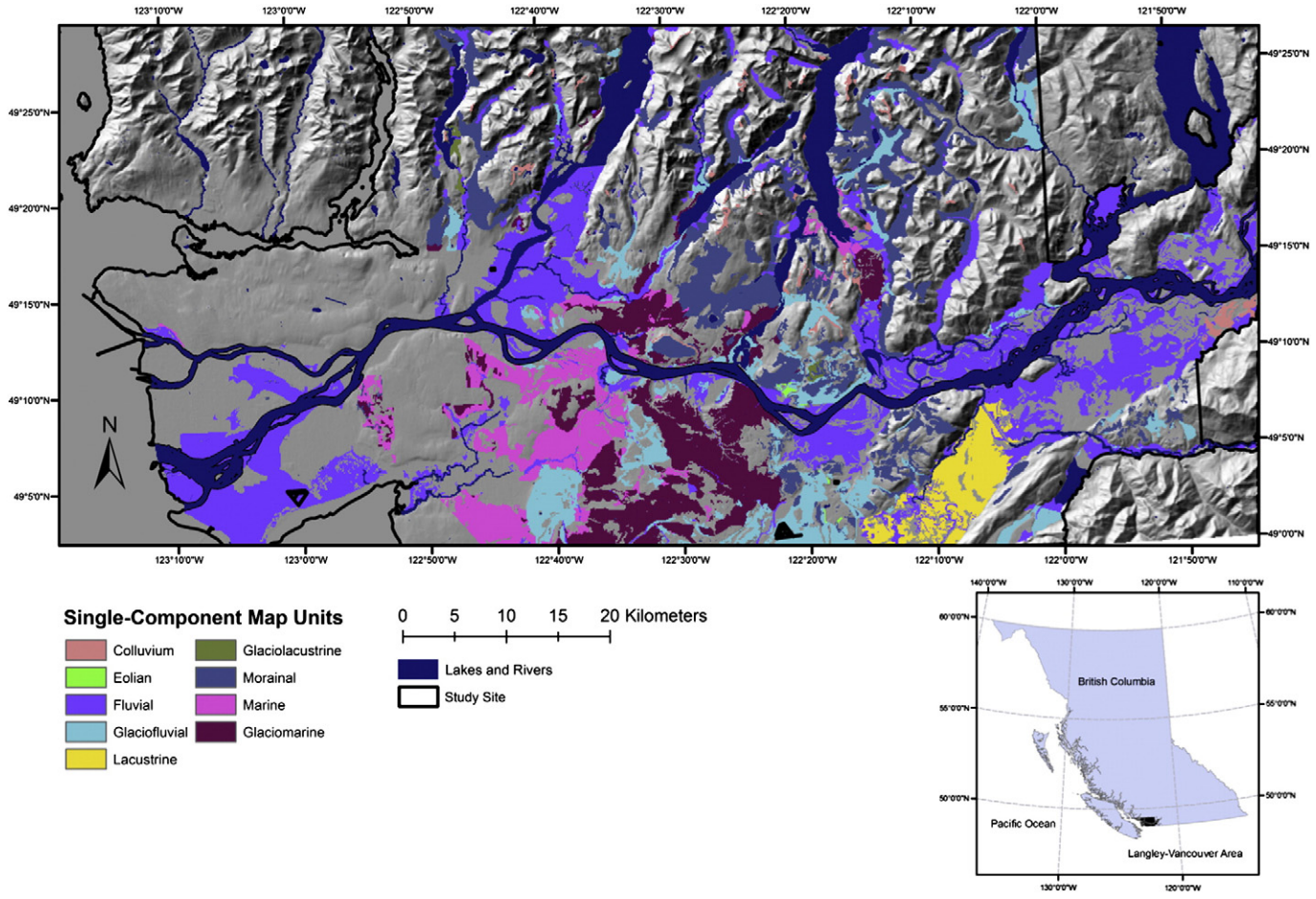

Fig. 2. Single-component parent material map units from the Langley–Vancouver Map Area (Luttmerding, 1981). Inset: study area in relation to British Columbia.

of glacial ice and isostatic rebound, marine and glaciomarine sediments from the Pacific are also common in the Lower Fraser Valley (Luttmerding, 1981). The mountainous area in the northern portion of the study area is part of the Pacific Ranges of the Coastal Mountains, where the bedrock is derived from Late Mesozoic intrusive igneous rocks (Valentine et al., 1978). On steep slopes, the dominant parent material is colluvium while depositions of glacial till are most common in areas with gentle and moderate slopes (Luttmerding, 1981). Exposed bedrock is uncommon even in the upland portions of the study area.

2.2. Digital data

The soil map for the study area was created at a 1:25,000 scale for the Lower Fraser Valley and at a 1:50,000 scale for the Southern Sunshine Coast and Southern Coastal Mountains (Luttmerding, 1981). The soil surveys were subsequently digitized into a seamless coverage and made freely available through Agriculture and AgriFood Canada and the British Columbia Ministry of Environment (Kenney and Frank, 2010).

Data layers for topographic predictors were calculated using British Columbia's Terrain Resource Information Management (TRIM) DEM (B.C. Ministry of Sustainable Resource Management, 2002). The 25 m DEM, originally derived from a triangulated irregular network (TIN) built from TRIM mass-points and break-lines, was then aggregated to a 100 m spatial resolution. The 100 m DEM is freely available from HectaresBC.org (Hectares BC, 2012).

Three successive mean filters with window sizes of 3×3, 3×3, and 5×5 cells were applied to the DEM in order to remove anomalous pits and peaks. Similar to MacMillan et al. (2003) and Li et al. (2011), we have found through preliminary work that the successive smoothing procedure reduces local-scale noise and improves landscape-scale signals. In Grinand et al. (2008), it was also demonstrated that the application of an adaptive mean filter was able to incorporate spatial context into their outputs and improve predictions using the Multiple Additive Regression Tree algorithm (MART).

Table 1
Mineral parent material classes and subclasses from The Soils of the Langley–Vancouver Map Area (Luttmerding, 1981).

Parent material class	Code	Parent material subclass	Code
Colluvial	C	Colluvial deposits (>1 m thick)	Cb
		Shallow colluvial (<1 m thick) over bedrock	Cv
Eolian	E	Eolian	E
Fluvial	F	Fluvial deposits — deltaic (sandy)	sF-D
		Fluvial deposits—deltaic (silty or clayey)	zcF-D
		Fluvial deposits—floodplain (sandy)	sFp
		Fluvial deposits—floodplain (silty or clayey)	zcFp
		Fluvial deposits—local streams (sandy)	sF-S
		Fluvial deposits—local streams (silty or clayey)	zcF-S
		Fluvial deposits—fans	Ff
Glaciofluvial	FG	Glaciofluvial deposits	FG
		Eolian veneer over glaciofluvial deposits	E/FG
Lacustrine	L	Lacustrine deposits (sandy)	sL
		Lacustrine deposits (silty or clayey)	zcL
Glaciolacustrine	LG	Glaciolacustrine deposits	LG
Morainal	M	Morainal (glacial till) deposits	M
		Eolian veneer over morainal deposits	E/M
Marine	W	Marine deposits (clayey)	cW
		Marine deposits (lag or littoral)	W
Glaciomarine	WG	Glaciomarine deposits	WG

Table 2
Topographic derivatives derived from a 100 m spatial-resolution DEM.

Landscape representation	Terrain derivative	Code	Reference
Local landscape characteristics	Transformed aspect	ASPECT	Zevenbergen and Thorne (1987)
	Curvature	CURVE	Zevenbergen and Thorne (1987)
	Elevation	ELEV	
	Slope length factor	LS	Moore et al. (1993)
	Plan curvature	PLAN	Zevenbergen and Thorne (1987)
	Profile curvature	PROF	Zevenbergen and Thorne (1987)
	Slope	SLOPE	Zevenbergen and Thorne (1987)
	Tangential curve	TANCUR	Florinsky (1998)
	Terrain ruggedness index	TRI	Riley et al. (1999)
	Total curvature	TCURVE	Wilson and Gallant (2000)
Hydrologic characteristics	Convergence index	CONV	Koethe and Lehmeier (1996)
	Distance to nearest river	RiDIST	
	Distance to nearest stream	StDIST	
	Modified relative hydrologic slope position	mRHSP	MacMillan (2005)
	Relative hydrologic slope position	RHSP	MacMillan (2005)
	Stream power index	StPI	Moore et al. (1991)
	SAGA wetness index	SWI	Böhner et al. (2002)
	Topographic wetness index	TWI	Beven and Kirkby (1979)
Landscape context	Multiresolution ridge top flatness index	MRRTF	Gallant and Dowling (2003)
	Multiresolution valley bottom flatness index	MRVBF	Gallant and Dowling (2003)
	Midslope position	MSLOPE	SAGA Development Team (2011)
	Normalized height	NHEIGHT	SAGA Development Team (2011)
	Slope height	SLOPEH	SAGA Development Team (2011)
	Valley depth	VDEPTH	SAGA Development Team (2011)
Landscape exposure	Sky view factor	SKYVIEW	Häntzschel et al. (2005)
	Terrain view	TERVIEW	Häntzschel et al. (2005)
	Visible sky	VISSKY	SAGA Development Team (2011)

Topographic and hydrologic attributes for 27 topographic indices (Table 2) were calculated from the successively filtered DEM using the System for Automated Geoscientific Analysis (SAGA) (SAGA Development Team, 2011). The indices were selected based on their ability to represent basic landscape characteristics of the local neighborhood (ie. elevation, slope, aspect, and curvature); hydrologic characteristics at the watershed scale (ie. wetness index, convergence index, and relative hydrologic slope position); and landscape context (normalized height, slope height, sky view and terrain view). In addition, the distance to nearest stream and distance to nearest river was calculated in order to account for the presence of the Fraser River that runs through the study area.

2.3. Development of the training data

Soil survey map units include attribute data for parent material subclass where 5645 polygons contained a single parent material subclass that covered 29.8% of the study area while 3025 multi-component polygons contained either 2 or 3 subclasses that covered 55.0% of the study area (Fig. 2). The remaining 15.2% of the study area included miscellaneous land types such as anthropogenic land, bedrock, gravel pits, ice, recent alluvium, rock outcrops, talus, and tidal flats where bedrock only accounted for 0.4% of the study extent. To minimize the uncertainty in the training data, only polygons with a single-component of parent material subclass were used to develop the predictive model; however, we also recognize that these polygons may have small inclusions of other components. Overall, the dominant parent material subclasses included silty or clayey fluvial floodplain material and glaciomarine sediments for the Fraser Valley; and morainal material (glacial till) along the Coastal Mountain Range (Fig. 3). The most common parent material class is fluvial, which accounts for approximately 41% of the single-component polygon training data. In addition, many of the soils in the area have had varying amounts of eolian material added as a veneer (>1 m thick) to the surface layers. Where such additions were present only in the surface layers, or where they were considered to have a minor influence, we classified the soil parent material based on the dominant material below the eolian veneer. In other areas where eolian materials were dominant throughout the soil profile, we classified the area as having an eolian parent material.

Predictive models were developed using randomly generated training points within each single-component polygon where the points were intersected with the values for each topographic attribute and its parent material subclass. Three different methods for developing the training dataset were used with different allocations of the training points according to the following approaches: (1) equal number of points per parent material subclass, (2) equal number of points per polygon, and (3) the number of points was determined as an area-weighted proportion of the subclass' extent over the entire study area. For each sampling strategy, $n = 28225$ training points, with an average sampling density of 9.4 samples/km^2, were used as inputs for RF. The number of training points was selected based on the equal number per polygon sampling scheme where 5 points were randomly generated within each of the 5645 polygons with a single parent material subclass.

2.4. Random forest

To establish the hierarchical relationships between parent materials and topography, the *randomForest* package in the statistical software, *R*, was used (Liaw and Wiener, 2002; R Development Core Team, 2012). The RF classifier uses numerous decision trees, n_{tree}, that are grown from bootstrap samples (63%) of the entire sample population, n (Breiman, 2001). The bootstrap sampling makes RF less sensitive to over-fitting in comparison to decision trees. Initially, the RF classifier uses a bootstrapped sample to grow a single RF tree. At each binary split, the predictor that produces the best split is chosen from a random subset, m_{try}, of the entire predictor set, p, where the number of predictors tried at each split, m_{try}, is defined by the user. As a result, m_{try} is recognized as the main tuning parameter of RF and should therefore by optimized (Svetnik et al., 2003, 2004) The tree growing procedure is performed recursively until the size of the node reaches a minimum, k, which is parameterized by the user (Hastie et al., 2009). Secondly, the remaining 37% of the training dataset that was not used in the growing of a decision tree, the out-of-bag (OOB) sample, X_i, are inputted through the decision tree and a predicted class is assigned to each OOB sample, $Y_{OOB}(X_i)$.

The resulting output of the RF is a single model that is accompanied with a single aggregated error estimate — the overall OOB error rate, ER_{OOB}, using the following:

$$ER_{OOB} = n^{-1} \Sigma_{i=1}^{n} \; I[Y_{OOB}(X_i) \neq Y_i], \quad (1)$$

where the predicted class of a sample, $Y_{OOB}(X_i)$, is compared against its actual class, Y_i, using the indicator function, I (Breiman, 2001; Liaw and Wiener, 2002; Svetnik et al., 2003). In Eq. (1), I has a value of 1 when $Y_{OOB}(X_i) \neq Y_i$ — otherwise I is 0. The OOB error is similar to

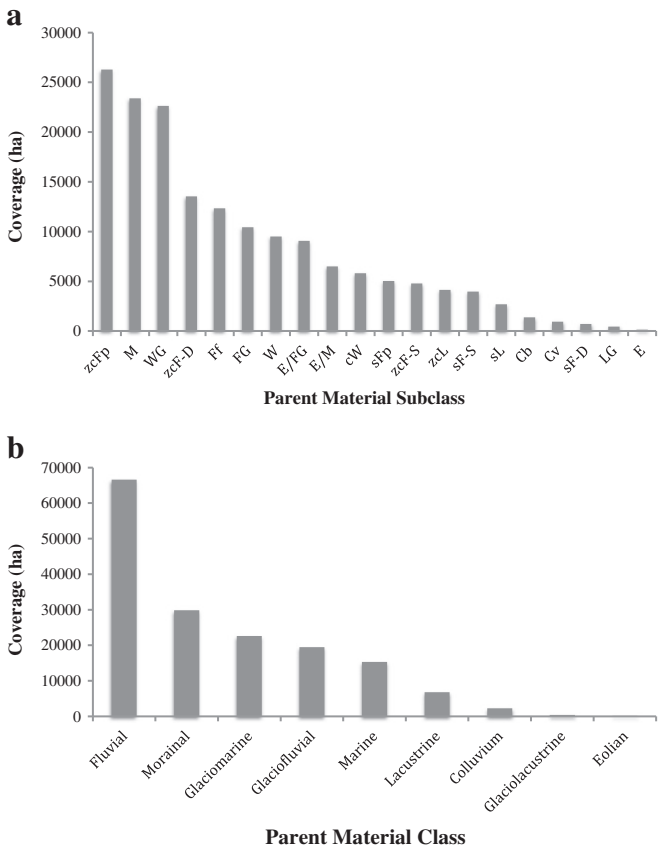

Fig. 3. Coverage of single-component parent material polygons by (a) subclass and by (b) class from Soils of the Langley–Vancouver Area (Luttmerding, 1981).

k-fold cross validation (CV) and provides comparable values (Hastie et al., 2009). As a result, RF and its OOB error rates may potentially be used when an independent validation dataset is not available.

In addition, the RF algorithm also provides two measures of variable importance: mean decrease in accuracy (MDA) and mean decrease in Gini (MDG). The MDA is a permutation-based measure of variable importance based on evaluating a variable's contribution to the prediction accuracy. The MDG also measures variable importance; however, it is based on the quality of each split (node) on a variable in a decision tree. A variable that produces high homogeneity in the descendent nodes results in a high MDG (Breiman, 2001).

In this study, the parent material training points, and their associated topographic attributes were used to train the RF classifier. The resulting non-spatial RF model was then applied to all unknown points in the study area using the set of topographic indices (Table 1). The output was a map of parent material subclasses in a raster format for the entire study area.

2.4.1. Optimization of m_{try}

Based on preliminary results, we chose to use $n_{tree} = 750$ as it produced stable OOB error rates and was also small enough to maximize computational efficiency. In addition, a terminal node size of $k = 1$ was selected as an increasing k resulted in a monotonic increase in the OOB error rates.

To optimize the primary tuning parameter, m_{try} values ranging from 1 to 27 were tested and the OOB error rates from 50 replicates for each m_{try} value were assessed. In addition, m_{try} values were further assessed using error rates obtained from 20 replications of a 5-fold CV. Where $m_{try} = 1$, a random predictor variable is selected at each node; contrarily, $m_{try} = p$ has the same effect as bagging the predictors.

2.4.2. Variable reduction

Variable reduction has previously been shown to result in slight error reductions (Svetnik et al., 2003; 2004), or to have minimal effect on the RF classifier (Xiong et al., 2012) through the removal of potentially irrelevant predictor variables. In this study, variable reduction was tested in order to examine whether or not a smaller set of predictors would lead to an improvement in RF predictions based on the following algorithm adopted from Svetnik et al. (2003):

1. The RF classifier was initially applied using the entire set of predictors. Variable importance, based on the mean decrease in accuracy, was used to rank the predictor variables.
2. Using the variable rankings, the three least important predictors were removed.
3. The training data was then partitioned into 5-folds for cross-validation and the error rates for each of the 5 cross-validation partitions were aggregated into a mean error rate. 20 replicates of 5-fold CV was performed.
4. Steps 2 and 3 were repeated until 3 predictors remained.

To test the effects of variable reduction, an initial variable importance plot was generated using the default settings of RF and the area-weighted sampling approach. Variable ranking was done using the MDA as it provides a more reliable measure of variable importance in comparison to the MDG (Bureau et al., 2003).

Since the choice of m_{try} depends on the total number of predictor variables (p), m_{try} was calculated as a function of p. Here, the m_{try} functions were defined as follows: $m_{try} = p$ (bagging), $p/2$, $p/4$, and $p^{1/2}$ (default setting). A resulting parent material prediction was generated using a reduced number of predictors with an optimal m_{try} function as a basis for comparison to the map produced using the entire variable set.

2.5. Assessment of predictions

Three approaches for assessing the predictions made by RF were used. Firstly, RF predictions were compared to the single-component polygons used as training areas from the soil survey. Secondly, RF predictions were compared to the multi-component polygons from the soil survey in order to examine RF's ability to disaggregate complex mapping units. Finally, the RF predictions were validated using point data.

For assessment purposes, the raw parent material maps were reclassified by generalizing the 20 parent material subclasses to 9 classes in order to offset the limited number of field validation points for each parent material subclass. In addition, OOB error rates were recalculated in order to reflect this reclassification procedure. Furthermore, a preliminary study showed that RF performed better when the training data were derived from parent material subclasses and the results were later generalized to classes.

2.5.1. Consistency with single-component polygons

Using *Map Comparison Kit 3* (Van Vliet, 2003), *overall agreement* was calculated as the percentage of pixels that were correctly classified by RF and the *disagreement with soil survey* was calculated as the percentage of pixels that were incorrectly classified by RF. In addition, the kappa index, a measure of map agreement that considers map agreement that occurs by 'chance', was also calculated for map comparison using the following (Visser and de Nijs, 2006):

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}, \tag{2}$$

where $P(A)$ represents the actual agreement fraction and $P(E)$ represents the expected agreement fraction between the soil survey and the RF predictions. Because the 'chance' factor is taken into account, the kappa index is consistently lower than the overall agreement.

2.5.2. Disaggregation of multi-component polygons

To evaluate the effectiveness of RF in disaggregating multi-component map units, the proportion of parent material classes specified for each unit in the soil survey was compared to the proportional extent of the classes predicted by. Model residuals, $\varepsilon_{c,j}$, were calculated from the difference between RF's predicted extent, $\hat{n}_{c,j}$ [% of polygon], of a parent material class, c, for a polygon, j, and the parent material's estimated extent from the soil survey, $n_{c,j}$ [% of polygon] under the same polygon in the following:

$$\varepsilon_{c,j} = \hat{n}_{c,j} - n_{c,j} \quad (3)$$

2.5.3. Validation with point data

Legacy soil pit data from the British Columbia Soil Information System (BCSIS) (Sondheim and Suttie, 1983), which consists of $n = 248$ points, were supplemented with additional data collected from fieldwork (between April and August, 2009; $n = 59$) in order to form an external validation point dataset with $n = 307$ points. Because the BCSIS data points were primarily located in the agricultural landscapes of the Lower Fraser Valley, supplemental data points were established along the Coastal Mountain on forested landscapes. Due to the forested and mountainous terrain, the supplemental data points were located along areas with good access and in places that reflected the range of materials present. To account for uncertainty in the location of the original soil pits, two levels of validation were used. At the first level, a predicted cell was considered valid if the validation point matched the prediction at that exact location ($r = 0$). At the second level, if a validation point matched a predicted cell that was located within a radius of 1 cell ($r = 1$), or within 100 m, surrounding the validation point, the predicted cell was considered valid.

3. Results and discussion

3.1. Development of the training data

In general, the OOB error rates produced from RF's internal validation were similar to the disagreement with soil survey rates (Fig. 4a). For the equal sampling by polygon approach, however, the OOB error rates were more than 10% lower than the disagreement with soil survey for the colluvium, glaciolacustrine, and morainal parent material

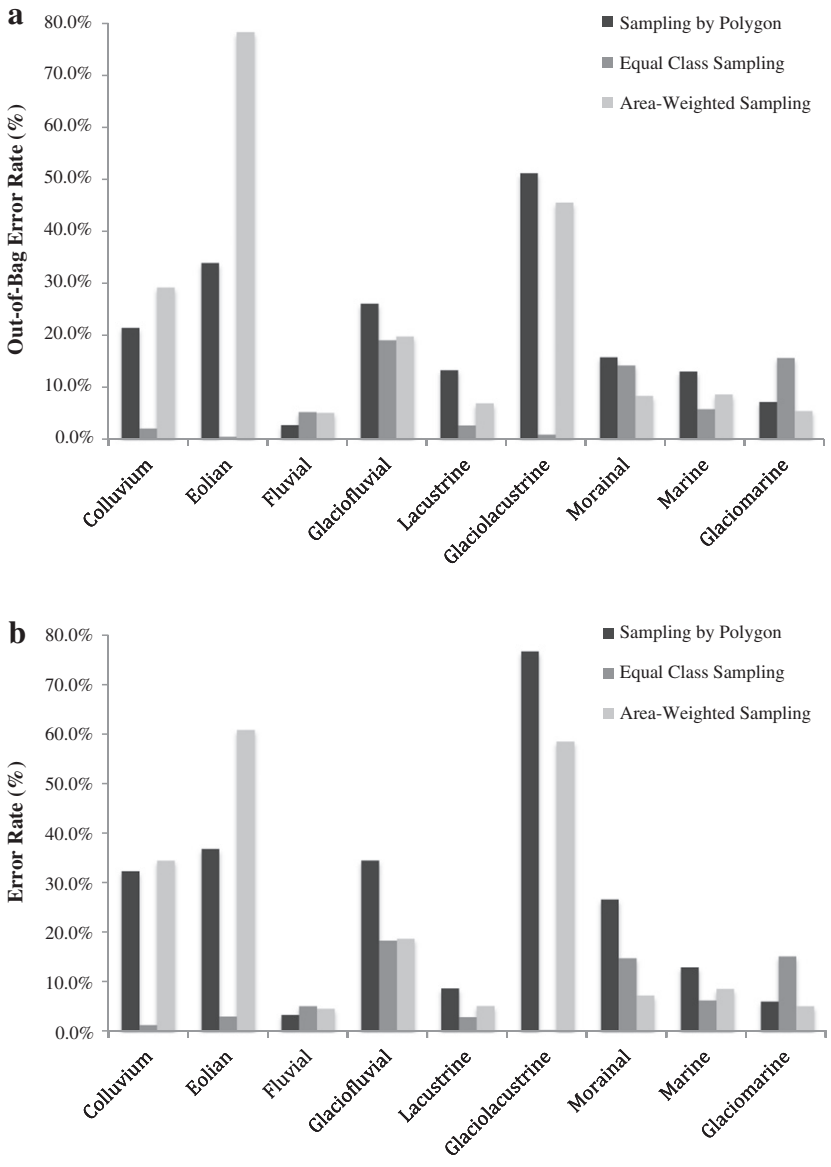

Fig. 4. (a) Out-of-bag error rates (%) and (b) error in soil survey agreement (%) by parent material class using sampling by polygon, equal-class sampling, and area-weighted sampling.

classes. These discrepancies suggest that the OOB error rates may not be the most reliable measure of class error; hence, the overall agreement with soil survey and kappa indices were used to select the sampling approach used for the parameter optimization and variable reduction analyses.

Overall agreement and kappa were highest for the area-weighted sampling approach when compared to the single-component polygons from the soil survey (Table 3). Moran and Bui (2002) noted that the area-weighted sampling approach performed better because more training data points were used to represent geographically extensive classes in order to capture a greater amount of variability that occurs under these classes. In this study, it was observed that the area-weighted sampling approach resulted in a lower error in agreement with soil survey for the most common (majority) classes, which include fluvial, morainal, and glaciomarine parent materials. Despite this, the equal-class sampling approach was superior in predicting the minority classes (ie. eolian, glaciolacustrine, colluvium, and lacustrine parent materials) (Fig. 4b). The discrepancy in performance between majority and minority classes was expected as machine-learning algorithms are recognized for their poor performance for minority classes (ie. Kubat and Matwin, 1997; Van Hulse et al., 2007).

We recognize that the potential problem that may arise with the use of the area-weighted sampling approach is that such an approach would lead to an unbalanced training dataset – a common problem for various machine-learning approaches (Galar et al., 2011; Van Hulse and Khoshgoftaar, 2009; Van Hulse et al., 2007). Despite these differences, this study was primarily aimed at producing a parent material map with the lowest overall error and hence, the area-weighted sample set was selected for all remaining analyses. A rigorous study in addressing the issue of an unbalanced dataset was beyond the scope of this study; however, such a study would be of use in cases where a study's objective is to predict the presence of rare soils or unique features in a landscape.

3.2. Optimization of m_{try}

The CV error rates reached a minimum when m_{try} ranged from 15 to 21; although, the increase of m_{try} from 11 to 15 only amounted to a minor decrease in CV error rate of 0.1% (Fig. 5). In the optimization of RF's main tuning parameter, m_{try}, it was determined that the OOB error rate was a fairly adequate measure of the model error when compared to the 5-fold CV error rates. Generally, the OOB error rates were consistently lower than the CV error rates by a margin of roughly 1%. The lower OOB error rates were expected because fewer training points were used to build a RF using the partitioned 5-fold CV training dataset, which was further partitioned into a 63% bootstrap sample. Based on 20 replicates of 5-fold CV, we have chosen to use $m_{try} = 11$; in addition, the smaller value of m_{try} was used in order to retain more of the 'randomness' in RF's randomized variable selection process.

3.3. Variable reduction

Based on the MDA values from the variable importance plot generated by RF, it was observed that the most important variables were aspect, distance to nearest stream, convergence index, and distance to nearest river whereas slope-length, slope, plan curvature, and profile curvature were the least important (Fig. 6). A detailed further examination between environmental covariates and parent material classes was beyond the scope of this study since the topographic indices were all derived from the same DEM; and hence,

Table 3
Overall agreement within single-component soil survey polygons based on sampling by equal number per polygon, equal-class, and area-weighted.

Sampling method	RF internal validation	Soil survey	
	Out-of-bag error	Overall agreement	Kappa
	(%)	(%)	(%)
By polygon	7.8	86.6	82.9
Equal-class	7.0	90.3	87.1
Area-weighted	8.3	92.2	89.6

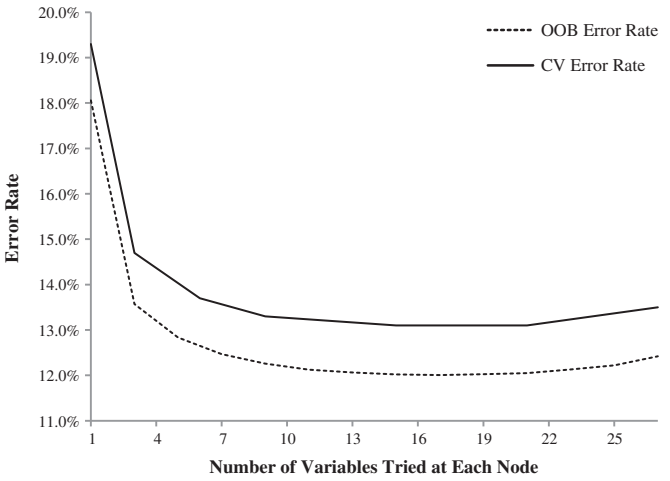

Fig. 5. Non-aggregated overall out-of-bag error rates and mean 5-fold CV error rates vs. number of predictors tried at each node (m_{try}).

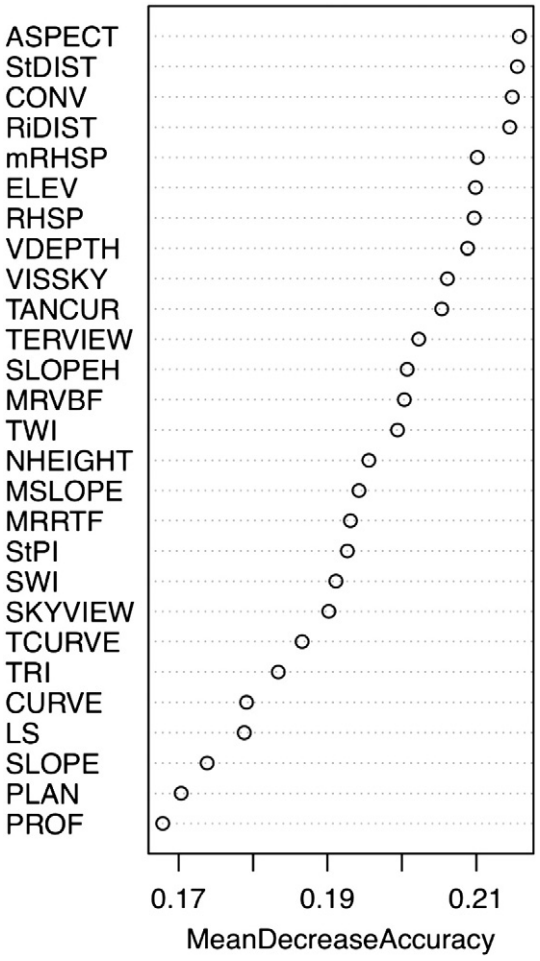

Fig. 6. Variable importance plots based on mean decrease in accuracy using area-weighted sampling. See Table 2 for the description of predictor variables.

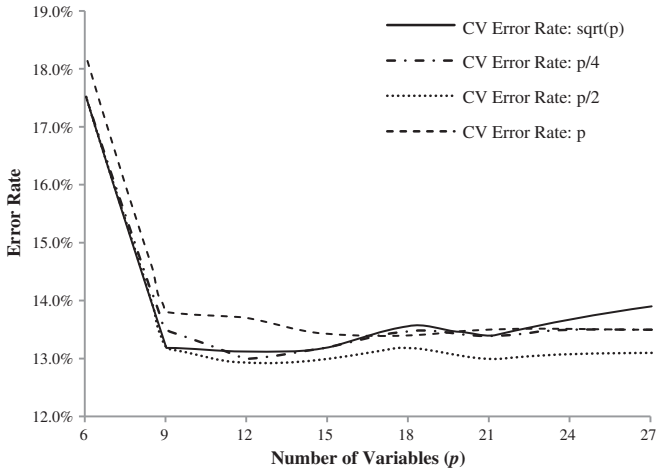

Fig. 7. Non-aggregated mean CV test error rates with 3 predictors removed at each step using various m_{try} functions: $m_{try} = \text{sqrt}(p)$; $p/4$; $p/2$; and p.

an inherently high level of covariance between the indices would make such a detailed analysis to be highly complex.

From this study it was determined that the CV error rates produced when $m_{try} = p$, $p/4$, $p/2$, and $p^{1/2}$ remained fairly consistent until the number of predictors were reduced to $p = 9$ (Fig. 7). As the number of variables reduced to $p = 3$, the CV error rates increased to 64% for each m_{try} function (not shown in Fig. 7). Overall, $m_{try} = p/2$ resulted in a slightly better overall performance; however, the difference in CV error rates in comparison to other m_{try} functions were less than 0.5%. Hence, it was found that variable reduction did not necessarily result in an improvement in RF performance with respect to the CV error rates. Furthermore, the minimal degradation in RF predictions when the predictors were reduced to $p = 9$ indicates that, for our study area, RF is insensitive to the presence of irrelevant predictors. These findings corroborate the results in Svetnik et al. (2003) and Xiong et al. (2012) where variable reduction algorithms were also tested. Although this study only examined a single approach for variable reduction, further studies may explore alternative dimension reduction approaches, such as the use of principal components as predictors for RF.

3.4. Assessment of predictions

3.4.1. Consistency with single-component polygons

There was a high overall agreement between the RF predictions and single-component polygons from soil survey data (Table 4). Based on the overall agreement with soil survey data, m_{try} optimization resulted in a minimal effect when the entire variable set was used while variable reduction increased the overall agreement by 0.9% when compared to the optimized RF using 27 predictors. Kappa indices indicated a high overall agreement between predicted parent material maps and the soil survey data. The optimization of m_{try} and variable reduction, however, increased kappa minimally.

By examining the various error rates for each parent material class (Fig. 8), it was observed that m_{try} optimization had a minimal effect on improving the agreement with soil survey data in the cases of fluvial, lacustrine, morainal, and marine parent materials. Improvements in agreement for these classes were less than 0.5%. Improvements in agreement with soil survey data occurred primarily for the minority classes such as eolian, colluvium, and glaciolacustrine, which had an increase in agreement of 3.0%, 4.3%, and 12.5%, respectively.

By way of a visual comparison between the single-component parent material polygons and the continuous surface generated using RF (Fig. 9), it was observed that RF was able to produce results that had patterns and boundaries that were qualitatively similar to the single-component polygons. Fig. 9A shows a close-up of an area with low relief terrain, adjacent to the Fraser River, which is typical of the southern region of the study area and where fluvial, glaciofluvial, marine, and glaciomarine parent materials are most common. Based on the visual comparison and a low disagreement with soil survey for the listed parent materials, we are confident that the RF results were fairly consistent with the single-component polygons for low relief terrain. This was to be expected since the majority of the training points were located in low relief terrains. In comparison, Fig. 9B shows a close-up of an area with high relief terrain, which is typical of the northern region where the dominant parent material is moraine; however, the presence of colluvial, fluvial, and glaciofluvial deposits are also common. RF was able to produce parent material boundaries that were similar to single-component polygon boundaries; however, we have also noted that RF over-predicted the presence of morainal deposits and under-predicted the presence of colluvial deposits since the steep slopes were classified as morainal when colluvial materials were expected.

3.4.2. Disaggregation of multi-component polygons

Histograms of the distribution of model residuals, $\varepsilon_{c,j}$, were produced based on polygons where a parent material class, c, was a component of multi-component polygon, j (Eq. (3)). Examples for glaciomarine, fluvial, colluvial, and morainal parent materials using an optimized RF and $p = 27$ predictors are presented in Fig. 10. In Fig. 10, $\varepsilon_{c,j} = 0$ represents cases of multi-component polygons where the proportional extent of a parent material class estimated by the soil survey, $n_{c,j}$, matched the extent of the same parent material class predicted by RF, $\hat{n}_{c,j}$. Where $\varepsilon_{c,j} > 0$, the parent material class was over-predicted; conversely, where $\varepsilon_{c,j} < 0$, the parent material class was under-predicted for polygon, j.

Overall, RF was effective in the disaggregation of both glaciomarine and fluvial materials when they were a component of a multi-component polygon as $\varepsilon_{c,j}$ was most frequently 0. In comparison, colluvial materials were poorly disaggregated from multi-component polygons and were largely under-predicted by RF while morainal materials were over-predicted — findings that were similar to the analysis of the single-component polygons.

Table 4
Classification accuracy measurements using non-optimized RF, optimized RF with no variable reduction, and optimized RF with variable reduction.

Number of predictors	RF internal validation	Soil survey		External validation ($r = 0$ cell)		External validation ($r = 1$ cell)	
	Out-of-bag error	Overall agreement	Kappa	Overall accuracy	Kappa	Overall accuracy	Kappa
	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Not optimized							
27	8.3	92.2	89.6	77.5	69.1	85.0	79.5
Optimized							
12	7.3	93.0	90.7	77.9	69.7	85.7	80.3
27	7.7	92.8	90.4	77.5	69.2	85.7	80.4

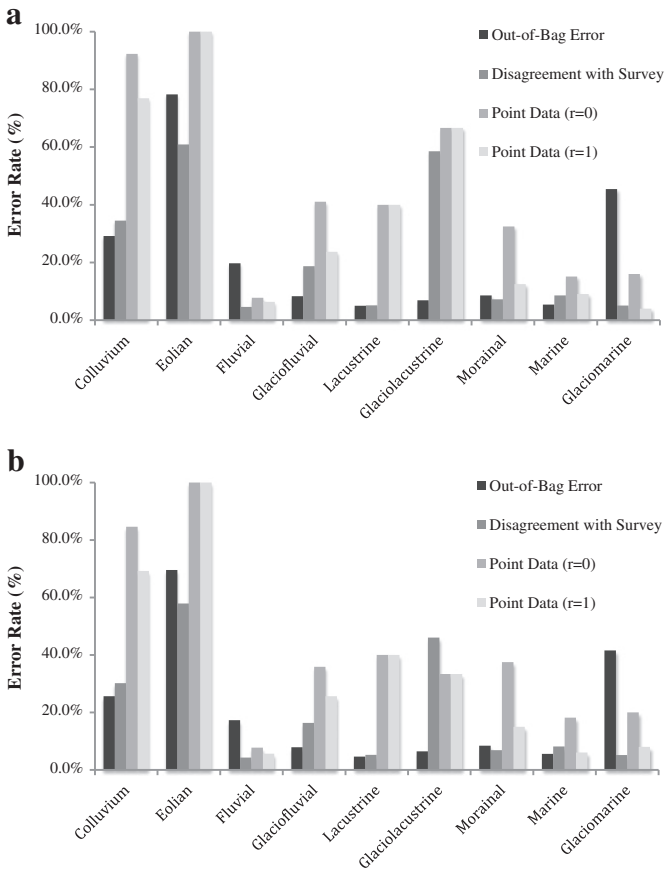

Fig. 8. Out-of-bag error rates (%), error in agreement with soil survey rates (%), and error rates using validation points (%) by parent material class using (a) non-optimized random forest; (b) optimized random forest; and (c) optimized random forest with variable reduction.

These results confirm the initial visual assessment from Section 3.4 and suggest that the topographic distinctions between morainal and colluvial deposits may be difficult for RF to detect. The distinctions between these parent materials were not entirely clear in the soil survey map (Luttmerding, 1981) for several reasons. Firstly, colluvial and morainal components were frequently coupled together in 481 multi-component polygons; and therefore not included in the training dataset. Secondly, the soil survey was carried out at a 1:50,000 scale for the Coastal Mountains whereas the Lower Fraser Valley was carried out at the 1:25,000 scale. Consequently, the smaller scale mapping of the Coastal Mountains inherently resulted in greater generalization of the map units where localized colluvial deposits would not have been mapped as a single-component map unit, but rather, a multi-component unit. The generalization of morainal and colluvial deposits into multi-component polygons would also explain the small number of single-component colluvium polygons that were available to train the RF and hence, the poor disaggregation of morainal-colluvial complexes.

The findings of using RF for polygon disaggregation are further summarized in Table 5. Most parent materials were not predicted by RF when they were not a component of a soil survey polygon with the exceptions of morainal and fluvial materials. Both the optimization of m_{try} and variable reduction had little influence on the disaggregation of multi-component polygons. Colluvial, eolian, fluvial, glaciofluvial, marine, and glaciomarine materials were under-predicted when they were included as components of a polygon; whereas, lacustrine and morainal materials were over-predicted. There were 815 instances where morainal materials were predicted in polygons where they were not identified in the soil survey; consequently, this inherently would have contributed to the under-prediction of the parent materials that were mapped as a component of a polygon. In contrast to the polygon disaggregation study in Häring et al. (2012), this study did not constrain the number of different parent material classes to the ones identified by the multi-component polygons in order to account for the inclusion of parent materials that were not recognized. Hence, the polygon components identified by the soil survey would have

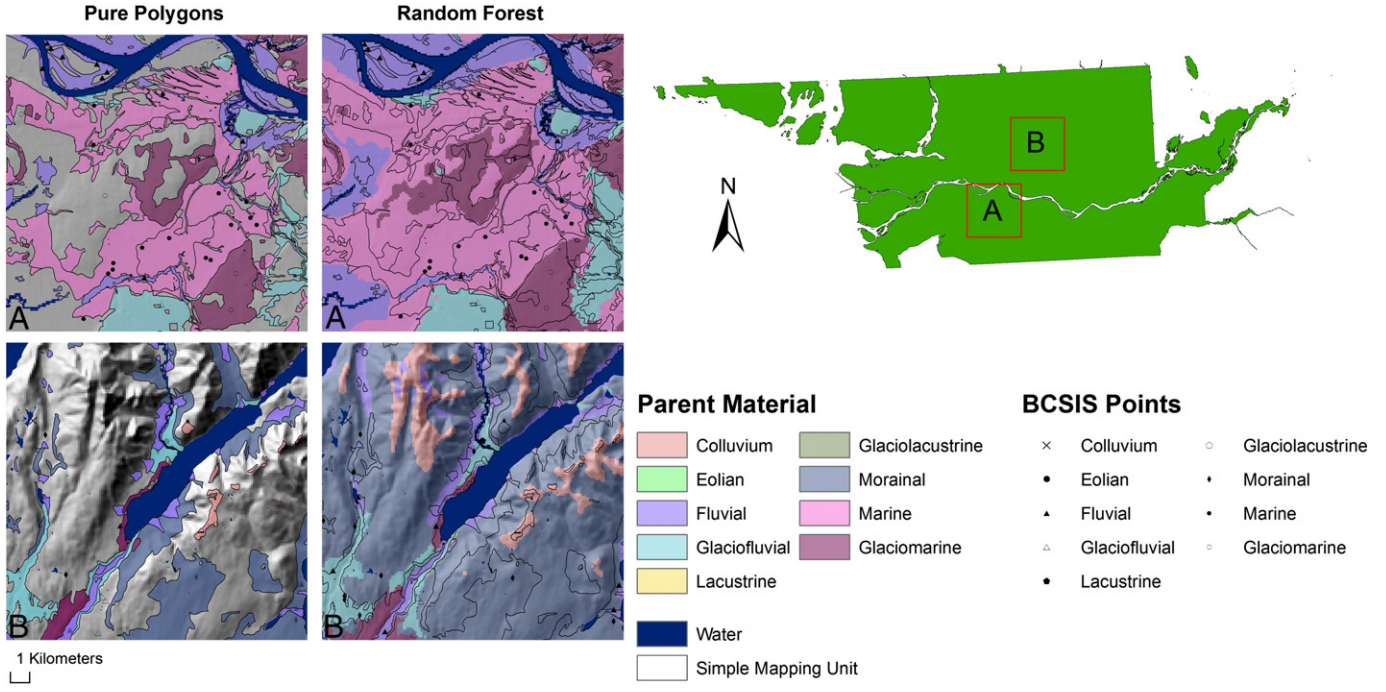

Fig. 9. Close-up map of single-component parent material polygons, RF results, and sample points overlaid on a hill-shade for (A) a low relief terrain and (B) a high relief terrain. The map uses the optimized parameter settings with $p = 27$ predictor variables.

been under-predicted due to the presence of small inclusions of other parent materials in the polygons.

3.4.3. Validation with point data

When the predicted parent material maps were compared to the validation points (Fig. 11), the overall accuracy and kappa indices were lower than the agreement with soil survey data (Table 4). It was observed that between 77% and 78% of the validation points matched the predicted parent material map exactly. Comparing the overall accuracy to the kappa index, we observed a difference of 8%. Differences between the overall accuracy and kappa index suggest that there was a low to moderate probability that cells were correctly classified by chance. When examining the cells that were within a 1-cell (100 m) radius of a validation point, it was observed that the overall accuracy and kappa index increased by 8% and 10%, respectively on average. This suggests that the RF produced a fairly accurate map within 100 m with an average overall accuracy of 85% and an average kappa index of 80%.

Based on the validation using point data, m_{try} optimization and variable reduction resulted in little improvement in predictions for each parent material class (Fig. 8). A comparison between the results produced with optimized RF and the results with variable reduction were similar with minimal differences (<1 %) in agreement with soil survey were minimal as well as with the validation points. For glaciolacustrine and colluvium classes, m_{try} optimization resulted in a 33% and 7.7% increase for those respective classes; however, the seemingly large increase is the result of a small sample size for those classes.

Map comparison using the single-component polygons from the soil survey data resulted in higher prediction accuracy compared to the prediction accuracy using the point data. These differences in accuracy are likely caused in part by the initial use of the soil survey data to stratify the training data for the RF model. Secondly, the soil survey polygons represent an aggregation of the soil-environmental conditions for each map unit whereas the point data may not necessarily be representative of the average environmental conditions from which map units are derived and from which the RF model is based on.

4. Conclusions

The objective of this study was to first evaluate methods for the extraction of training data from legacy soil data and the optimization of RF parameters. It was determined that the imbalanced area-weighted

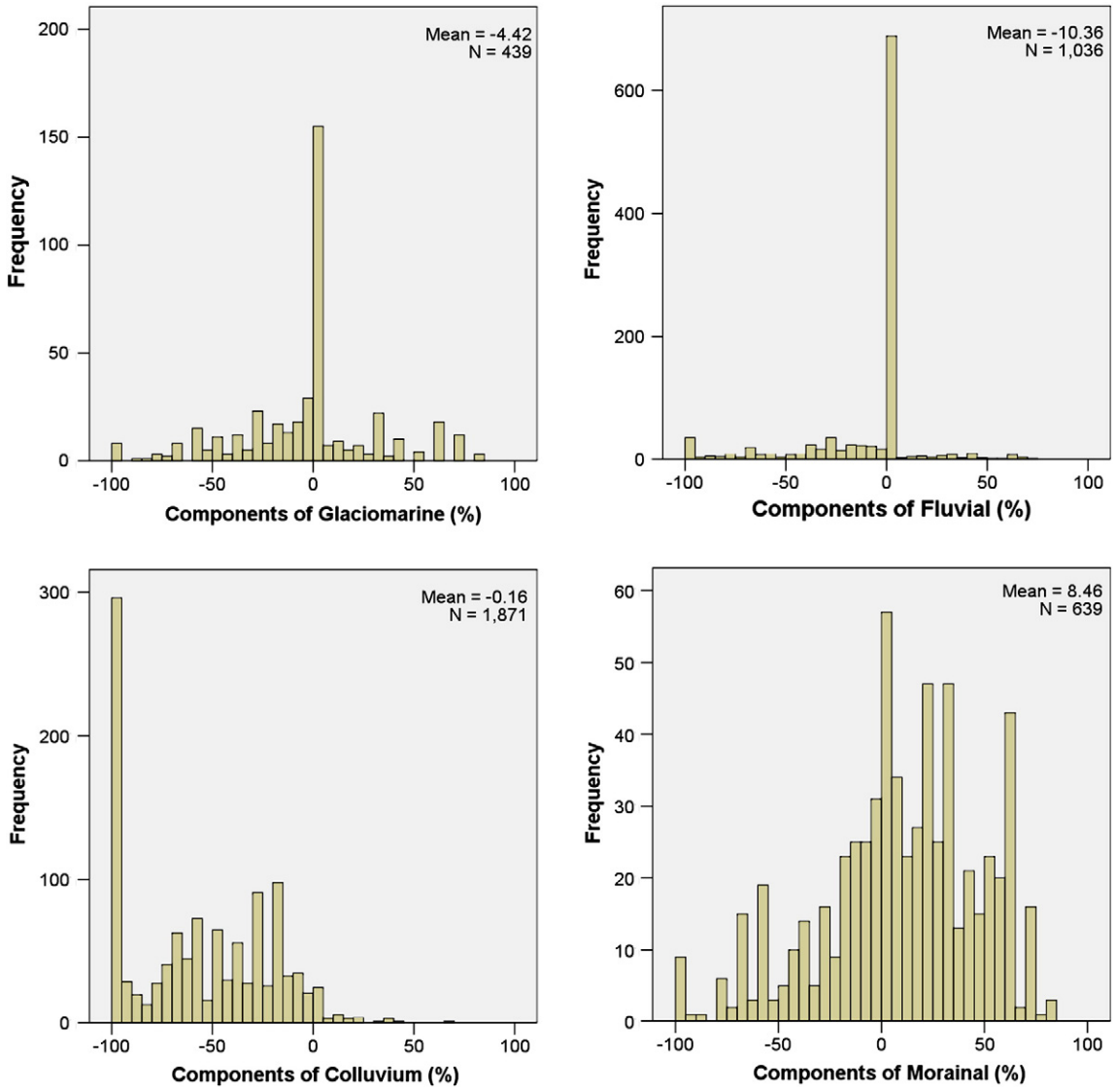

Fig. 10. Histograms of model residuals, $\varepsilon_{c,j}$, calculated as the difference between the predicted RF extent and the soil survey extent for each multi-component polygon, j. Histograms only consider polygons where parent material class, c, is a component of polygon, j. Histograms are based on an optimized m_{try} with $p = 27$ predictor variables.

Table 5
Descriptive statistics for model residuals, $\varepsilon_{c,j}$, calculated as the difference between the predicted RF extent and the soil survey extent of each parent material class, *c*, for each multi-component polygon, *j*.

	Parent material class	*n*	27 Variables		27 Variables + optimization		12 Variables + optimization	
			Mean	St. dev.	Mean	St. dev.	Mean	St. dev.
			(%)	(%)	(%)	(%)	(%)	(%)
Non-component of polygons[a]	Colluvium	1871	0.2	0.1	0.2	0.1	0.2	0.1
	Eolian	2969	0.0	0.0	0.0	0.0	0.0	0.0
	Fluvial	1989	7.4	0.5	7.8	0.5	7.8	0.5
	Glaciofluvial	2758	3.5	0.3	4.1	0.3	5.2	0.3
	Lacustrine	2924	0.2	0.1	0.2	0.1	0.2	0.1
	Glaciolacustrine	3015	0.0	0.0	0.0	0.0	0.0	0.0
	Morainal	2386	22.7	0.7	21.5	0.7	21.3	0.7
	Marine	2654	1.3	0.2	1.3	0.2	1.2	0.2
	Glaciomarine	2586	3.5	0.3	3.4	0.3	3.3	0.3
Component of polygons[b]	Colluvium	1154	−57.4	1.0	−56.8	1.0	−57.7	1.0
	Eolian	56	−61.4	4.1	−60.4	4.2	−61.1	4.1
	Fluvial	1036	−10.7	0.9	−10.4	0.9	−10.6	0.9
	Glaciofluvial	267	−29.7	2.4	−27.0	2.4	−25.9	2.5
	Lacustrine	101	16.3	3.5	17.6	3.5	15.9	3.8
	Glaciolacustrine	10	23.5	10.1	−65.9	7.8	−65.9	7.7
	Morainal	639	10.0	1.5	8.5	1.5	7.9	1.5
	Marine	371	−25.7	2.1	−24.4	2.1	−23.3	2.2
	Glaciomarine	439	−2.8	1.6	−4.4	1.6	−6.2	1.7

[a] Polygons where parent material class, *c*, is not a component of multi-component polygon, *j*.
[b] Polygons where parent material class, *c*, is a component of multi-component polygon, *j*.

sampling resulted in higher overall agreement with soil survey with lower error rates for majority parent material classes such as fluvial, morainal, marine and glaciomarine materials; however, the prediction of minority classes was less successful. Using a balanced dataset improved the prediction of the minority parent material classes such as eolian, glaciolacustrine, colluvium, and lacustrine materials; however, overall agreement with soil survey decreased as a consequence. This research suggests that the selection of a sampling approach for training data should reflect the objectives of the study and whether the goal is to maximize overall accuracy or to maximize the accuracy of

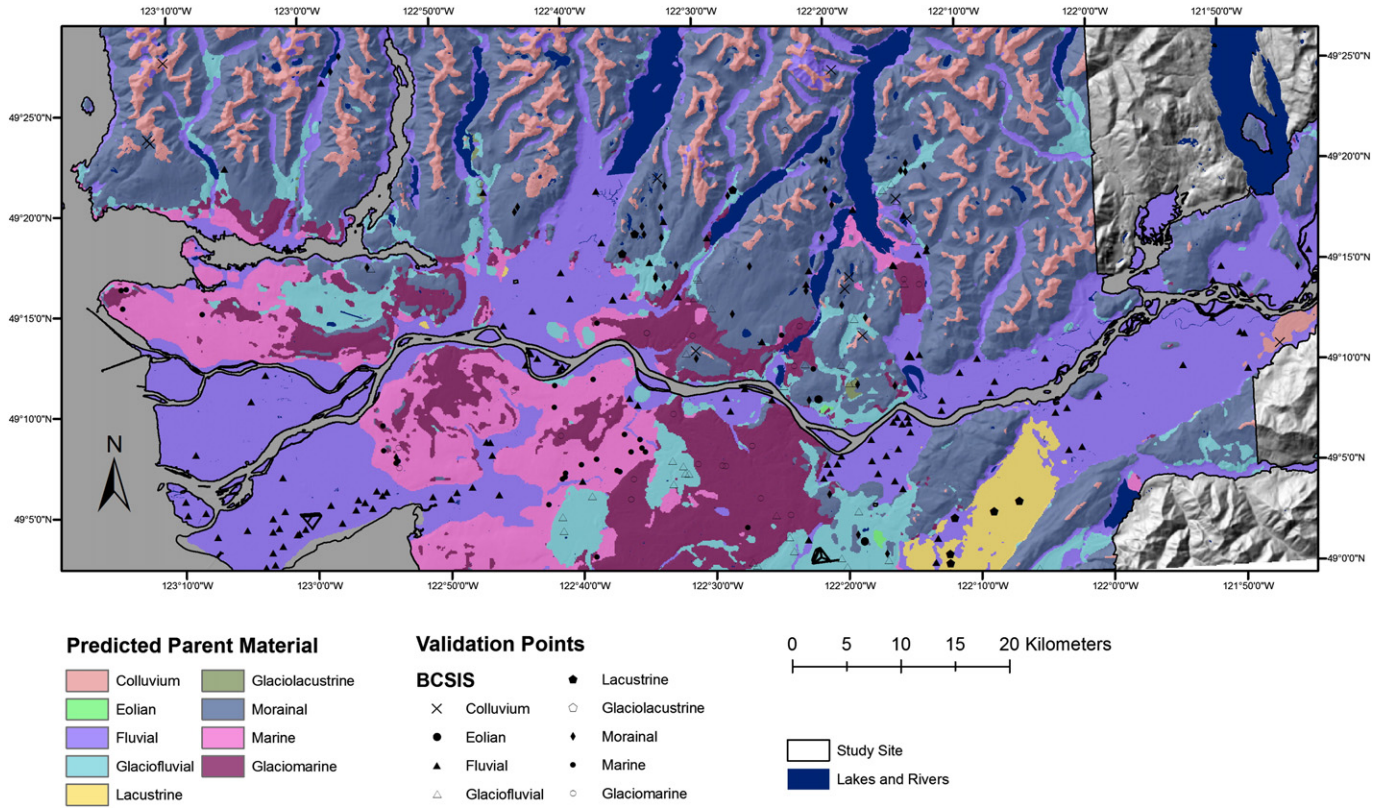

Fig. 11. Predictive parent material map using random forest at a 100 m spatial resolution with underlying hill-shade and overlying sample points for the Langley–Vancouver Map Area, British Columbia.

the minority classes. Furthermore, this research also suggests that the relationship between imbalanced multi-class training data and machine-learning approaches should be investigated further.

In terms of the optimization of RF parameters, this study has found through extensive CV testing, that both the m_{try} optimization and variable reduction had little effect in improving RF outputs. As a result, it was concluded that RF performs well with minimal user intervention through the parameterization of the model. In addition, it was found that RF was able to identify important predictors, internally, as the reduction of predictors resulted in marginal improvements in overall agreement with soil survey and overall accuracy.

The second objective of this study was to assess the reliability of RF outputs within single-component polygons. It was determined that RF produced maps that had a high overall agreement with soil surveys and that RF was effective in extracting the relationships between parent material and topography. In comparison, however, it was also concluded that RF was not as effective in the disaggregation of multi-component parent material polygons. These results may illustrate the importance of the training dataset as much as the characteristics of RF, since our training data was concentrated in areas of the map with single-component polygons. This study has found that the RF classifier is an effective machine learning and data mining approach. Our approach to developing a training dataset by extracting points from single-component polygons likely limited the performance of RF for disaggregation of multi-component polygons.

Acknowledgments

The authors are thankful for the support from the Forest Science Program of the Ministry of Forests, Lands, and Natural Resource Operations; the Future Forest Ecosystem Scientific Council of British Columbia and from a Natural Sciences and Engineering Research Council Discovery grant to M. Schmidt. The authors are also grateful for the field and lab assistance provided by Darren Murray, Sarah Robertson, Maciej Jamrozik, Jin Zhang, Lillian Fan, and in particular, Darrell Hoffman.

References

Armstrong, J.E., 1956. Surficial geology of Vancouver area, British Columbia. Paper 55–40, Geological Survey of Canada, Ottawa.
Armstrong, J.E., 1957. Surficial geology of New Westminster map-area, British Columbia. Paper 57–5 and Map 16.Department of Mines and Technical Surveys, Ottawa.
B.C. Ministry of Sustainable Resource Management, 2002. Gridded Digital Elevation Model Product Specifications, 2nd edition. Base Mapping and Geomatics Services Branch Ministry of Sustainable Resource Management, Victoria.
Beckett, P.H.T., 1971. The cost-effectiveness of soil survey. Outlook Agric. 6, 191–198.
Beven, K.J., Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology. Hydrol. Sci. Bull. 24, 43–69.
Böhner, J., Köthe, R., Conrad, O., Gross, J., Ringeler, A., Selige, T., 2002. Soil regionalization by means of terrain analysis and process parameterization. European Soil Bureau – Research Report, 7 213–222.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC Press LLC, Boca Raton, FL.
Bui, E.N., 2004. Soil survey as a knowledge system. Geoderma 120, 17–26.
Bui, E.N., Moran, C.J., 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modeling and legacy data. Geoderma 103, 79–94.
Bui, E.N., Moran, C.J., 2003. A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray–Darling basin of Australia. Geoderma 111, 21–44.
Bui, E.N., Lougheed, A., Corner, R., 1999. Extracting soil-landscape rules from previous soil surveys. Acad. J. Sci. Res. 37, 495–508.
Bui, E.N., Henderson, B.L., Viergever, K., 2006. Knowledge discovery from models of soil properties developed through data mining. Ecol. Model. 191, 431–446.
Bui, E.N., Henderson, B.L., Viergever, K., 2009. Using knowledge discovery with data mining from the Australian Soil Resource Information System database to inform soil carbon mapping in Australia. Glob. Biogeochem. Cycle 23, GB4033. http://dx.doi.org/10.1029/2009GB003506.
Bureau, A., Dupuis, J., Hayward, B., Falls, K., Van Eerdewegh, P., 2003. Mapping complex traits using Random Forest. BMC Genet. 4, S64.
Cutler, R.D., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88, 2783–2792.
Díaz-Uriarte, R., Alvarez de Andrés, S., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 3–15.
Florinsky, I.V., 1998. Accuracy of local topographic variables derived from digital elevation models. Int. J. Geogr. Inf. Sci. 12, 47–61.
Galar, M., Fernández, A., Barrenechea, E., Bustince, H., Herrera, F., 2011. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 42, 463–484.
Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resour. Res. 39, 1347–1359.
Geng, X., Fraser, W., VandenBygaart, B., Smith, C.A.S., Wadell, A., Jiao, Y., Patterson, G., 2010. Towards digital soil mapping in Canada: existing soil survey data and related expert knowledge. Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Springer, pp. 325–337.
Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random forests for land cover classification. Pattern Recognit. Remote Sens. 27, 294–300.
Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island — digital soil mapping using random forests analysis. Geoderma 146, 102–113.
Grinand, C., Arrouays, D., Laroche, B., Martin, M.P., 2008. Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. Geoderma 143, 180–190.
Häntzschel, J., Goldberg, V., Bernhofer, C., 2005. GIS-based regionalization of radiation, temperature and coupling measures in complex terrain for low mountain ranges. Meteorol. Appl. 12, 33–42.
Häring, T., Dietz, E., Osenstetter, S., Koschitzki, T., Schröder, B., 2012. Spatial disaggregation of complex soil map units: a decision-tree based approach in Bavarian forest soils. Geoderma 185–186, 37–47.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Springer, New York, NY (734 pp.).
Hectares, B.C., 2012. Hectares BC. Available at http://hectaresbc.org/app/habc/HaBC.html (verified 16 May 2012).
Henderson, B.L., Bui, E.N., Moran, C.J., Simon, D.A.P., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma 124, 383–398.
Heung, B., Bakker, L., Schmidt, M.G., Dragićević, S., 2013. Modelling the dynamics of soil redistribution induced by sheet erosion using the Universal Soil Loss Equation and cellular automata. Geoderma 202–203, 112–125.
Hole, F.D., Campbell, J.B., 1985. Soil Landscape Analysis. Rowman and Allanheld, Totowa, NJ.
Howes, D.E., Kenk, E., 1997. Terrain Classification System for British Columbia, Version 2. Resource Inventory Branch, Ministry of Environment, Lands and Parks, Victoria, BC.
Hua, J., Xiong, Z., Lowely, J., Suh, E., Dougherty, E.R., 2005. Optimal number of features as a function of sample size for various classification rules. Bioinformatics 21, 1509–1515.
Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill, NY.
Kenney, E., Frank, G., 2010. Creating a seamless soil dataset for the Okanagan Basin, British Columbia. Proceedings of the Western Regional Cooperative Soil Survey Conference, Las Vegas, NV. USDA-NRCS (ftp://ftp-fc.sc.egov.usda.gov/NSSC/NCSS/Conferences/regional/2010/west/kenney.pdf (verified 8 Nov 2012).
Koethe, R., Lehmeier, F., 1996. SARA-Systeme Zur Automatischen Relief-Analyse, Benutzerhandbuch, 2. Goettingen University.
Kubat, M., Matwin, S., 1997. Addressing the curse of imbalanced training sets: one-sided selection. Proceedings of the 14th Annual International Conference on Machine Learning, Nashville, TN, pp. 179–186.
Lacoste, M., Lemercier, B., Walter, C., 2011. Regional mapping of soil parent material by machine learning based on point data. Geomorphology 133, 90–99.
Lawley, R., Smith, B., 2008. Digital soil mapping at a national scale: a knowledge and GIS based approach to improving parent material and property information. Digital Soil Mapping with Limited Data. Springer, pp. 173–182.
le Roux, J.J., Newby, T.S., Sumner, P.D., 2007. Monitoring soil erosion in South Africa at a regional scale: review and recommendations. S. Afr. J. Sci. 103, 329–335.
Lemercier, B., Lacoste, M., Loum, M., Walter, C., 2012. Extrapolation at regional scale of local soil knowledge using boosted classification trees: a two-step approach. Geoderma 171–172, 75–84.
Li, S., MacMillan, R.A., Lobb, D.A., McConkey, B.G., Moulin, A., Fraser, W.R., 2011. Lidar DEM error analyses and topographic depression identification in a hummocky landscape in the prairie region of Canada. Geomorphology 129, 263–275.
Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2, 18–22.
Ließ, M., Glaser, B., Huwe, B., 2012. Uncertainty in the spatial prediction of soil texture: comparison of regression tree and Random Forest models. Geoderma 170, 70–79.
Lord, T.M., Green, A.J., 1974. Soils of the Tulameen Area of British Columbia. Report No. 13, British Columbia Soil Survey.Research Branch, Canada Department of Agriculture, Ottawa, ON, Canada.
Luttmerding, H.A., 1981. Soils of the Langley–Vancouver Map Area. Report No. 15, British Columbia Soil Survey.BC Ministry of Environment, Kelowna, BC, Canada.
MacMillan, R.A., 2005. A new approach to automated extraction and classification of repeating landform types. . Naples Florida Frontiers in Pedometrics, p. 54 (http://www.conference.ifas.ufl.edu/pedometrics/Abstract%20Book.pdf (verified 8 Nov 2012)).
MacMillan, R.A., Martin, T.C., Earle, T.J., McNabb, D.H., 2003. Automated analysis and classification of landforms using high-resolution digital elevation data: applications and issues. Can. J. Remote. Sens. 29, 592–606.
MacMillan, R.A., Moon, D.E., Coupé, R.A., 2007. Automated predictive ecological mapping in a forest region of B.C., 2001–2005. Geoderma 140, 353–373.
McBratney, A.B., Mendoça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.
Moore, I.D., Grayson, R.B., Ladson, A.R., 1991. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydrol. Process. 5, 3–30.

Moore, I.D., Turner, A.K., Wilson, J.P., Jenson, S.K., Band, L.E., 1993. GIS and land-surface–subsurface process modeling. Environmental Modeling with GIS. Oxford University Press, pp. 196–230.
Moran, C.J., Bui, E.N., 2002. Spatial data mining for enhanced soil map modeling. Int. J. Geogr. Inf. Syst. 16, 533–549.
Natural Resources Canada, 2004. Canada's National Forest Inventory Photo Plot Guidelines, version. 1.1. Canadian Forest Service, Pacific Forestry Centre, Victoria, BC.
Pojar, J., Klinka, K., Demarchi, D., 1991. Ecosystems of British Columbia. Br. Columbia Minist. For. Ranges 95–111.
Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9, 181–199.
Qi, Y., 2012. Random forest for bioinformatics. Ensemble Machine Learning. Springer, pp. 307–323.
R Development Core Team, 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (URL http://www.R-project.org.).
Resource Inventory Committee, 1998. Standards for Terrestrial Ecosystem Mapping in British Columbia. Ecosystem Working Group, Terrestrial Ecosystem Task Force, Resource Inventory Committee, Victoria, BC.
Riley, S.J., DeGloria, S.D., Elliot, R., 1999. A terrain ruggedness index that quantifies topographic heterogeneity. Intermountain J. Sci. 5, 23–27.
Saga Development Team, 2011. System for Automated Geoscientific Analyses (SAGA). Available at http://www.saga-gis.org/en/index.html (verified 12 August, 2012).
Schut, P., Smith, S., Fraser, W., Geng, X., Kroetsch, D., 2011. Soil Landscapes of Canada: building a national framework for environmental information. Geomatica 65, 293–309.
Scull, P., Franklin, J., Chadwick, O.A., 2005. The application of classification tree analysis to soil type prediction in a desert landscape. Ecol. Model. 181, 1–15.
Sondheim, M., Suttie, K., 1983. User Manual for the British Columbia Soil Information System, 1. BC Ministry of Forests Publication R28-82053, Victoria, BC.
Statnikov, A., Wang, L., Aliferis, C.F., 2008. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 9, 319–329.
Svetnik, V., Liaw, A., Tong, C., Culberson, C., Sheridan, R.P., Feuston, B.P., 2003. Random Forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. 43, 1947–1958.
Svetnik, V., Liaw, A., Tong, C., Wang, T., 2004. Application of Breiman's Random Forest to modeling structure–activity relationships of pharmaceutical molecules. In: Roli, F., Kittler, J., Windeatt, T. (Eds.), Multiple Classifier Systems, Fifth International Workshop, MCS 2004, Proceedings, 9–11 June 2004, Cagliari, Italy. Lecture Notes in Computer Science, vol. 3077. Springer, Berlin, pp. 334–343.
Valentine, K.W.G., Sprout, P.N., Baker, T.E., Lavkulich, L.M., 1978. The Soil Landscapes of British Columbia. BC Ministry of Environment, Victoria, BC, Canada.
Van Hulse, J., Khoshgoftaar, T., 2009. Knowledge discovery from imbalanced and noisy data. Data Knowl. Eng. 68, 1513–1542.
Van Hulse, J., Khoshgoftaar, T.M., Napolitano, A., 2007. Experimental perspectives on learning from imbalanced data. Proceedings of the 24th Annual International Conference on Machine Learning (ICML 2007), Corvalis, OR, pp. 935–942.
Van Vliet, J., 2003. Map Comparison Kit 3: User Manual. Research Institute for Knowledge Systems, Maastricht.
Visser, H., de Nijs, T., 2006. The map comparison kit. Environ. Model. Software 21, 346–358.
Weaver, A., 1991. The distribution of soil erosion as a function of slope aspect and parent material in Ciskei, Southern Africa. Geojournal 23, 29–34.
Webster, R., Beckett, P.H.T., 1968. Quality and usefulness of soil maps. Nature 219, 680–682.
Webster, R., Wong, I.F.T., 1969. A numerical procedure for testing soil boundaries interpreted from air photographs. Photogrammetria 24, 59–72.
Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil organic matter stocks using random forest modeling in a semi-arid steppe ecosystem. Plant Soil 340, 7–24.
Wilson, J.P., Gallant, J.C. (Eds.), 2000. Terrain Analysis: Principles and Applications. John Wiley & Sons, New York, NY.
Wittneben, U., 1986. Soils of the Okanagan and SimilkameenValleys. Report No. 52, British Columbia Soil Survey.Survey and Resource Mapping Branch, BC Ministry of Environment, Victoria, BC, Canada.
Xiong, X., Grunwald, S., Myers, D.B., Kim, J., Harris, W.G., Comerford, N.B., 2012. Which covariates are needed for soil carbon models in Florida. Digital Soil Assessment and Beyond. CRC Press, pp. 109–113.
Zevenbergen, L.W., Thorne, C.R., 1987. Quantitative analysis of land surface topography. Earth Surf. Proc. Land. 12, 47–56.
Zhu, A.X., Band, L.E., 1994. A knowledge-based approach to data integration for soil mapping. Can. J. Remote. Sens. 20, 408–418.