remote sensing

MDPI

Article

Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification

Christopher A. Ramezan *, **Timothy A. Warner** and **Aaron E. Maxwell**

Department of Geology and Geography, West Virginia University, Morgantown, WV 26506, USA; tim.warner@mail.wvu.edu (T.A.W.); aaron.maxwell@mail.wvu.edu (A.E.M.)

* Correspondence: cramezan@mail.wvu.edu; Tel.: +01-304-293-4917

Received: 8 December 2018; Accepted: 16 January 2019; Published: 18 January 2019

Abstract: High spatial resolution (1–5 m) remotely sensed datasets are increasingly being used to map land covers over large geographic areas using supervised machine learning algorithms. Although many studies have compared machine learning classification methods, sample selection methods for acquiring training and validation data for machine learning, and cross-validation techniques for tuning classifier parameters are rarely investigated, particularly on large, high spatial resolution datasets. This work, therefore, examines four sample selection methods—simple random, proportional stratified random, disproportional stratified random, and deliberative sampling—as well as three cross-validation tuning approaches—k-fold, leave-one-out, and Monte Carlo methods. In addition, the effect on the accuracy of localizing sample selections to a small geographic subset of the entire area, an approach that is sometimes used to reduce costs associated with training data collection, is investigated. These methods are investigated in the context of support vector machines (SVM) classification and geographic object-based image analysis (GEOBIA), using high spatial resolution National Agricultural Imagery Program (NAIP) orthoimagery and LIDAR-derived rasters, covering a 2,609 km^2 regional-scale area in northeastern West Virginia, USA. Stratified-statistical-based sampling methods were found to generate the highest classification accuracy. Using a small number of training samples collected from only a subset of the study area provided a similar level of overall accuracy to a sample of equivalent size collected in a dispersed manner across the entire regional-scale dataset. There were minimal differences in accuracy for the different cross-validation tuning methods. The processing time for Monte Carlo and leave-one-out cross-validation were high, especially with large training sets. For this reason, k-fold cross-validation appears to be a good choice. Classifications trained with samples collected deliberately (i.e., not randomly) were less accurate than classifiers trained from statistical-based samples. This may be due to the high positive spatial autocorrelation in the deliberative training set. Thus, if possible, samples for training should be selected randomly; deliberative samples should be avoided.

Keywords: training sample selection; cross-validation; high resolution imagery; NAIP; Lidar; regional-scale

1. Introduction

With the increasing availability of high spatial resolution (HR) remotely sensed datasets (1–5 m pixels), the routine production of regional-scale HR land-cover maps has become a possibility. However, due to the large area associated with regional-scale HR remote sensing projects, the sample selection for training and assessment can be burdensome. Sampling strategies that are commonly used in remote sensing analyses involving smaller datasets may be unsuitable or impractical for regional-scale HR analyses. This is particularly true if the sampling protocol requires field observations. While much previous remote sensing research has been conducted on supervised classification sample selection methods for training [1–3] and accuracy assessments [4–7], most of these studies examine sampling methods using study sites of limited geographic extent. The limited area of these study sites is typical of classification experiments in general; Ma et al. [8] meta-reviewed over 170 supervised object-based remote sensing analyses and found that an overwhelming majority of geographic object-based image analyses (GEOBIA) studies were conducted on areas smaller than 300 ha.

This work, therefore, investigates a variety of sample selection method techniques for regional-scale land-cover classifications with large, HR remotely sensed datasets. Additionally, as the number of samples is limited in many regional studies, cross-validation for regional HR classification is also explored. Cross-validation is an approach for exploiting training and accuracy assessment samples multiple times and thus potentially improving the reliability of the results. Finally, as the acquisition of widely dispersed samples across a large region may be expensive, a sampling strategy which confines the sample selection to a small geographic subset area is also investigated. This study is conducted in the context of GEOBIA, an approach that has become increasingly popular for analyzing high-resolution remotely sensed data [8,9].

1.1. Background on Sample Selection in Remote Sensing

Samples in remote sensing analyses are typically collected for two purposes: training data for developing classification models and assessment or test data for evaluating the accuracy of the map product. Supervised classifiers, such as machine learning algorithms, use pre-labeled samples to train the classifier, which is then used to assign class labels to the remaining population. As the collection of training data inherently requires sampling, the strategies used for the sample selection must be carefully considered in the context of the characteristics of the dataset, classifier, and study objectives [10]. Although sample selection strategy is widely discussed in the remote sensing literature, there are a variety of opinions on almost every aspect of the sampling process. Nevertheless, there is a consensus that the size [5,7,11,12] and quality [12] of the training sample dataset, as well as the sample selection method used [5], can affect classification and accuracy assessments.

A variety of statistical (e.g., simple random and stratified random) and non-statistical (e.g., deliberative) sample selection methods have been used to collect training and testing samples for remote sensing analyses. Mu et al. [13] separated statistical-based sampling into two categories: spatial and aspatial approaches. Spatial sampling considers the spatial autocorrelation inherent in geographic data, while aspatial methods, which ignore potential spatial autocorrelation, include approaches such as simple random and stratified sampling. Although problems associated with aspatial sampling methods in remote sensing have been noted [14–16], spatial sampling methods can be complex and typically require a priori information about the population, which may be difficult or impractical to collect. While spatial sampling methods have been used in remote sensing analyses, currently they are far less common than aspatial methods and consequently not pursued in this study.

Simple random sampling involves the purely random selection of samples and thus gives a direct estimate of the population parameters. Although random samples for image classification on average will sample each class (or stratum) in proportion to the area of that class in the map, any single random sample will generally not do so. This can exacerbate the difficulty of dealing with rare classes. Some classifiers, including support vector machines (SVM), are sensitive to imbalanced training data sets, in which some classes are represented by a much smaller number of samples than other classes [17].

Stratified random sampling addresses this problem by forcing the number of samples in each stratum to be proportional to the area of the class. A variant on this method is equalized stratified random sampling, where the number of samples in each stratum is the same, irrespective of their area on the map. Equalized stratified random sampling may not be possible if some classes are so rare that the population of that class is smaller than the desired sample size. In such circumstances, a disproportional stratified random sample may be collected, an approach in which the sizes of the strata are specified by the user and are set to intermediate values between the proportions of the areas of the classes and a simple equalized approach.

One disadvantage of all stratified approaches is that a pre-classification is needed to identify the strata. If the samples are only to be used for an accuracy assessment (and not training), then it is possible to use the classification itself to generate the strata. However, Stehman [18] points out that if multiple classifications are to be compared and the strata are developed from just one of those classifications, the resulting accuracy statistics for the remaining maps need to be modified to account for the differences between the map used to develop the stratification and the map under consideration. Furthermore, unlike random and stratified random sampling, equalized and disproportional stratified random samplings produce samples that are not a direct estimate of the population. Thus, for samples generated with these methods, the accuracy estimates need adjustment to account for the class prior probabilities [19].

Deliberative sampling, in which samples are selected based on a non-random method, are also common in remote sensing. Deliberative sampling is necessary if access limitations or other issues constrain the sampling. One could hypothesize that deliberative sampling might, under certain circumstances, be more effective for classifier training than random sampling. Deliberative sampling allows for the incorporation of expert knowledge into the sample selection process. For example, samples can be selected to ensure that the variability of each class is well represented. Furthermore, in SVM, only training samples that define the hyperplane separating the classes are used by the classifier. Thus, for SVM, deliberative samples selected to represent potentially spectrally confused areas may be more useful than samples representing the typical class values [20].

If in situ observations are required for sample characterization and the cost of traveling between sites is high, the spatial distribution of samples becomes a central focus. This is particularly a concern for regional-scale HR datasets. While certain innovative methods such as active learning have been proposed to reduce sampling costs [21,22], these methods are complex to implement and are beyond the scope of this study. An alternative is localizing sample selection to a single subset area of the region of interest. Localizing sample selection to a small geographic subset area can be advantageous in large regional-scale analyses for reducing sampling costs, especially if field observations are required.

1.2. Background on Cross-Validation Tuning

A central tenet of accuracy assessment is that the samples used for training should not also be used for evaluation. A similar concern applies to the methods for selecting the user-specified parameters required by most machine learning methods, for example, the number of trees for random forests, sigma and C values for radial basis function kernel support vector machines, and the *k*-distance for the *k*-nearest neighbors. The value of these parameters can affect the accuracy of the classification, and thus, optimizing the chosen values (sometimes called tuning) is usually required [23–26]. Tuning is generally empirical, with various values for the parameters systematically evaluated, and the combination of values that generate the highest overall accuracy or kappa coefficient is assumed to be optimal [17,25]. Excluding training samples from the samples used for the evaluation of the candidate parameter values reduces the likelihood of overtraining and thus improves the generalization of the classifier.

If the overall number of samples is small, a fixed partition of training samples into separate training and tuning samples will further exacerbate the limitations of the small sample size, since each sample is used once and for one purpose only (e.g., training). Cross-validation is an alternative approach to a fixed partition. In cross-validation, multiple partitions are generated, potentially allowing each sample to be used multiple times for multiple purposes, with the overall aim of improving the statistical reliability of the results. Examples of cross-validation methods include k-fold, leave-one-out, and Monte Carlo. Classification parameter tuning via cross-validation has been demonstrated to improve classification accuracy in remote sensing analyses [27]. However, as with any sampling technique, it is important that the overall sample set be representative of the entire data set, otherwise the generalizability of the supervised classifier is unknown [25].

The k-fold cross-validation method involves randomly splitting the sample set into a series of equally sized folds (groups), where k indicates the number of partitions, or folds, the dataset is split into. For example, if a k-value of ten is used, the dataset is split into ten partitions. In this case, nine of the partitions are used for training data, while the remaining one partition is used for test data. The training is repeated ten times, each time using a different partition as the test set and the remaining nine partitions as the training data. The average of the results is then reported [28].

Leave-one-out cross-validation is similar to the k-fold cross-validation except the number of folds is set as the number of samples in the sample set. This approach can be slow with very large sample sets [29].

Monte Carlo validation works on similar principles to k-fold cross-validation except that the folds are randomly chosen with replacement, also called bootstrapping. Thus, the Monte Carlo method may result in some samples being used for both training and testing data multiple times, or some data not being used at all. Usually, Monte Carlo methods employ a large number of simulations, for example 1,000 or more, and therefore may also be slow [30].

While studies such as by Maxwell et al. [17] and Cracknell and Reading [25] demonstrated the merits of the cross-validation methods such as k-fold cross-validation for parameter tuning, very little attention has been given to examining the different cross-validation methods and their effect on parameter optimization and, by extension, machine learning classification performances.

1.3. Research Questions and Aims

This work examines sample selection and cross-validation parameter tuning on regional-scale land cover classifications using HR remotely sensed data. These issues are explored through the following interlinked research questions:

1. Which training sample selection method results in the highest classification accuracy for a supervised support vector machine (SVM) classification of a regional-scale HR remotely sensed dataset? The methods tested include both statistical (simple random, proportional stratified random, and disproportional stratified random) and non-statistical (deliberative) methods.
2. Which cross-validation method provides the highest classification accuracy? Methods tested are k-fold, leave-one-out, and Monte Carlo.
3. What is the effect on classification accuracy for the different sampling and cross-validation methods when the samples are collected from a small localized region rather than from across the entire study area?

2. Materials and Methods

2.1. Study Area and Data

The study area (Figure 1) lies within the northeastern section of West Virginia, near the borders with Maryland and Pennsylvania. The study area includes the entirety of the Preston County, as well as proportions of the neighboring Monongalia, Taylor, Barbour, and Tucker counties. This region is dominated by Appalachian mixed mesophytic forests [31] and the terrain is mountainous (548–914 m).

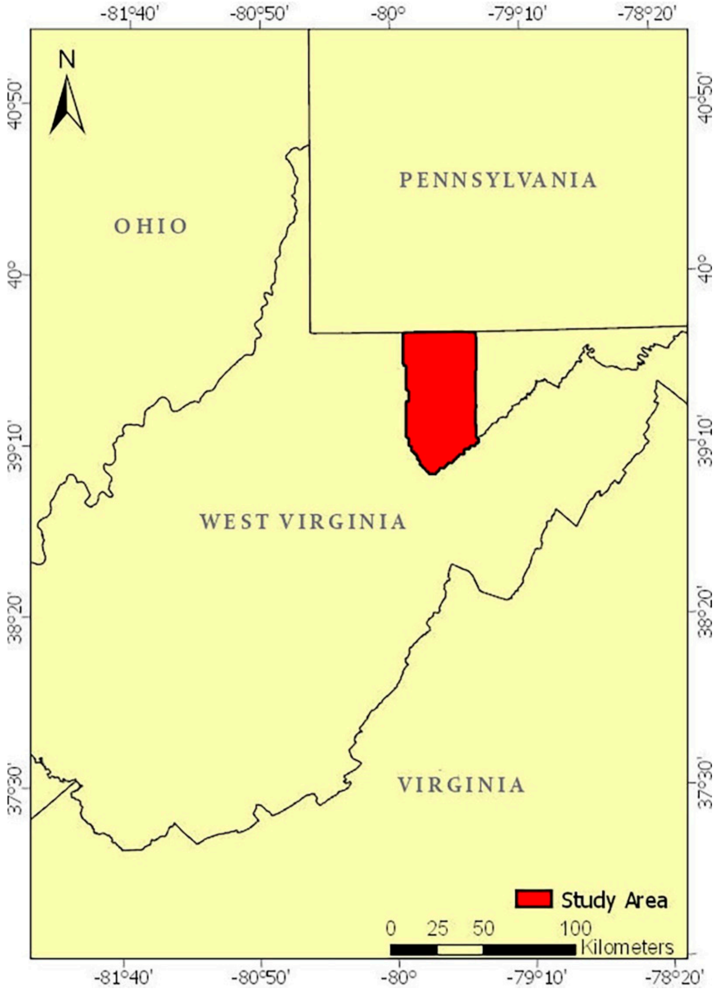


Figure 1. The regional-scale study area.

Two remotely sensed datasets were used in this analysis: optical multispectral imagery and light detection and ranging (LIDAR) point cloud data. The optical dataset comprises leaf-on National Agriculture Imagery Program (NAIP) orthoimagery collected primarily during 17–30 July 2011. A very small portion of the NAIP imagery was collected on 10 October 2011. The NAIP imagery consists of four spectral bands (red (590–675 nm), green (500–650 nm), blue (400–580 nm), and near-infrared (NIR) (675–850 nm)), with an 8-bit radiometric resolution and a spatial resolution of 1 m [32]. The data were provided as uncompressed digital orthophoto quarter quadrangles (DOQQs) in a '.tiff' format. The study area is covered by 108 individual DOQQ NAIP images, representing 260,975 ha or 4.2% of the total area of the state of West Virginia.

The LIDAR data were acquired using an Optech ALTM-3100C sensor through a series of aerial flights between 28 March and 28 April 2011. The LIDAR scanner had a 36° field of view and a frequency of 70,000 Hz. The LIDAR data were acquired at a flying height of 1,524 m above the ground with an average flight speed of 250 km/h. The flight lines of the LIDAR dataset had an average of 30% overlap. The LIDAR data include elevation, intensity, up to four returns, and a vendor-provided basic classification of the points [33]. The LIDAR data were formatted as a last version 1.2 point cloud. In total, 1164 LIDAR tiles containing a combined total of 5.6×10^9 points were used in the analysis. A preliminary investigation indicated that little change occurred during the approximately three to four-month temporal gap between the LIDAR and NAIP acquisitions.

Four land-cover classes were mapped: forest, grassland, water, and other. The forest class is primarily closed-canopy deciduous and mixed forests. The grassland class comprises areas dominated by non-woody vegetation. The water class includes both impoundments and natural waterbodies. The other class encompasses areas characterized by bare soil, exposed rock, impervious surfaces, and croplands.

2.2. Experimental Design

Sample selection includes three components: sample size, sampling region, and sampling method. The sample size specifies the number of training samples in the training set. The sampling region indicates whether samples are collected from the entire study area or only a limited sub region. The sampling method specifies the protocol for selecting samples, for example, random or deliberative.

In this study, four sampling methods are used to generate training data sets, which are then used with three cross-validation methods in SVM classifications (Figure 2). The samples are selected from the entire study area or from only a small geographic subset of the study area, and in all cases, the classifier is applied to the entire regional-scale dataset. The error for all classifications is evaluated using a large independent validation dataset acquired from the entire regional-scale dataset.

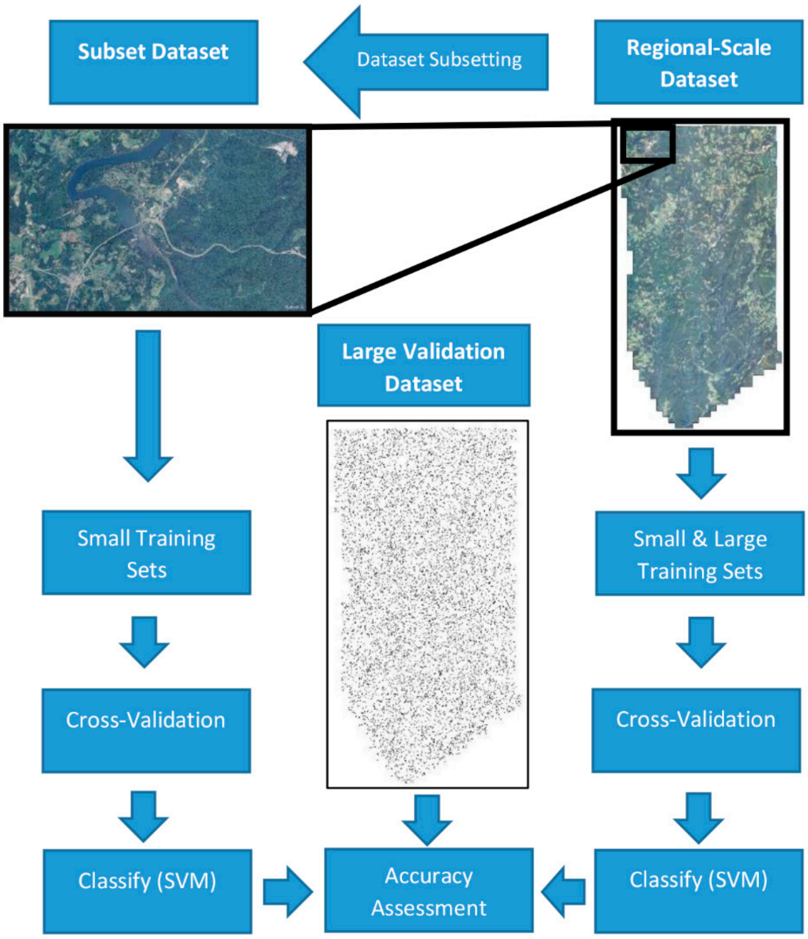


Figure 2. Overview of the experiment workflow.

2.3. Data Processing

A normalized-digital surface model (nDSM) and intensity rasters were generated as input variables for the classification from the LIDAR point cloud data. The LIDAR intensity raster was generated using the first returns only and the LAS Dataset to Raster function in ArcMap 10.5.1 [34]. The LIDAR intensity data has proven to be beneficial for separating land-cover surfaces such as grassland, trees, buildings, and asphalt roads. [35–37]. The LIDAR intensity data was not normalized due to the limited LIDAR metadata which prevented the normalization for distance. Previous research indicates that LIDAR intensity information is still useful for land-cover classifications without this correction [38]. The nDSM was generated by subtracting a rasterized bare earth digital elevation model (DEM) from a digital surface model (DSM) produced from the ground and first returns, respectively. The LIDAR-derived surfaces were rasterized at 1 m, matching the pixel size of the NAIP orthoimagery. nDSMs have been demonstrated to be useful for characterizing the varying heights of natural and man-made objects in GEOBIA studies [39].

The 108 NAIP orthoimages were mosaicked into a single large NAIP image mosaic using the Mosaic Pro tool within ERDAS Imagine 2016. Color-balancing was used to reduce the radiometric variations between the NAIP images, since they were acquired from different flights and times [40]. The NAIP mosaic was clipped to the extent of the LIDAR rasters. The NAIP and LIDAR rasters were then combined to form a single layer stack with six bands: four NAIP (Red, Green, Blue, and NIR) and two LIDAR (nDSM and Intensity).

2.4. Image Segmentation

The Trimble eCognition Developer 9.3 multi-resolution segmentation (MRS) algorithm was used as the segmentation method [41]. MRS is a bottom-up region-growing segmentation approach. Equal weighting was given to all six input bands for the segmentation. Preliminary segmentation trials found that a large number of artefacts were created by the image segmentation, apparently due to the "sawtooth" scanning pattern of the OPTECH ALTM 3100 sensor and the 1 m rasterization process [42]. A 5 × 5 pixel median filter was therefore applied to both the nDSM and Intensity rasters prior to segmentation to reduce the problem.

MRS has three user-set parameters: scale, shape, and compactness [43]. The scale parameter (SP) is regarded as the most important of the three parameters as it controls the size of the image objects [44–46]. The Estimation of Scale Parameter (ESP2) tool developed by Drăguţ et al. [45] was used to estimate the optimal scale parameter for the segmentation. The ESP2 tool generates image-objects using incrementally increasing SP values and calculates the local variance (LV) for each scale. The rate of change of the local variance (ROC-LV) is then plotted against the SP. In theory, peaks in the ROC-LV curve indicate segmentation levels in which segments most accurately delineate real world objects and thus optimal SPs for the segmentation [45].

Due to the high processing and memory demands of the ESP2 tool, three randomly selected subset areas were chosen to apply the ESP2 process rather than attempting to run the tool across the entire regional-scale dataset. The three subset tests indicated optimal SP values of 97, 97, and 104. The intermediate value of 100 was therefore chosen for the segmentation of the entire image. Alterations of the shape and compactness parameters from their defaults of 0.1 and 0.5 respectively did not seem to improve the quality of the segmentation, and therefore these values were left unchanged. The segmentation generated 474,614 image segments.

2.5. Dataset Subsetting

The subsetting tool in eCognition was used to extract the subset dataset from the regional dataset. The location of the subset was selected so that it included all four classes of interest. The total area of the subset dataset was approximately 4.19% of the area of the regional-scale dataset and comprised 21,777 image objects.

2.6. Segment Attributes Used for Classification

A total of 35 spectral and geometric attributes (Table 1) were generated for each image object (segment); these attributes were used as the predictor variables for the classification. Examples of the spectral attributes include the object's means and standard deviations for each band and the geometric attributes include object asymmetry, compactness, and roundness. The object's mean normalized difference vegetation index (NDVI) was also included as it is a commonly used spectral index used with NAIP data [47].

Table 1. Spectral and geometric attributes of the segments.

Attribute Type	Attributes	Number of Attributes
Spectral	Mean (Blue, Green, Intensity, NIR, Red, nDSM), Mode (Blue, Green, Intensity, NIR, Red, nDSM), Standard deviation (Blue, Green, Intensity, NIR, Red, nDSM), Skewness (Blue, Green, Intensity, NIR, Red, nDSM), Brightness	25
Geometric	Density, Roundness, Border length, Shape index, Area, Compactness, Volume, Rectangular fit, Asymmetry	9
Spectral Indices	Mean NDVI	1

2.7. Sample Data Selection

As image-objects are the base unit of analysis in this study, an object-based sampling approach was used for the collection of the samples. Two spatial scales were employed: a small subset and a regional scale, encompassing the entire study area. A large regional sample (n = 10,000) from the regional-scale dataset was collected to provide a benchmark representing an assumed maximum accuracy possible with this dataset. Since the subset area is 4.19% of the regional scale data set, the sample size for the subset area was set to n = 419 samples (4.19% of 10,000). This sample set is termed the small subset dataset. In addition, a small regional sample (n = 419) was selected from the entire regional scale data set to provide a direct comparison with the small subset sample dataset. In summary, three categories of datasets were collected at two spatial scales and two sample sizes: samples from a small limited region within the study area (small subset sample) (Figure 3) and two sets of samples collected from across the study area, one encompassing a small number of samples (small regional sample) (Figure 4) and another encompassing a large number (large regional sample) (Figure 5).

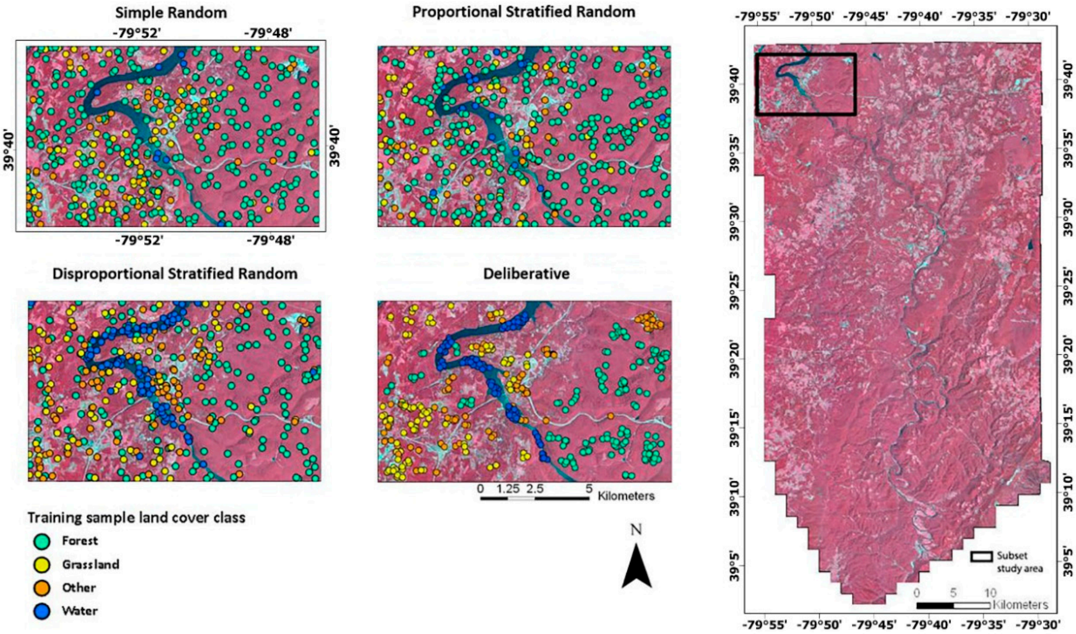


Figure 3. The subset area location and subset training samples overlaid on false color infrared composite of National Agricultural Imagery Program (NAIP) orthoimagery (Bands 4, 1, and 2 as RGB).

For each of these three categories of spatial scales and sampling sizes, four sampling methods were employed: simple random, proportional stratified random, disproportional stratified random, and deliberative. All samples were manually labeled by the analyst. In total, 53,352 samples were collected for this analysis. The number of samples for each training and validation sample sets is summarized in Table 2a–c.

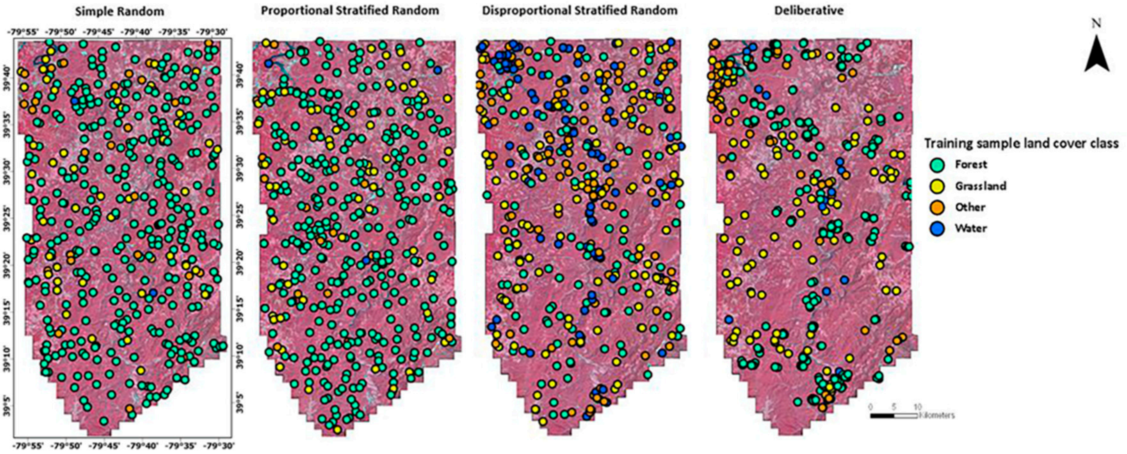


Figure 4. The small regional training samples displayed over false color infrared composite of NAIP orthoimagery (Bands 4, 1, and 2 as RGB).

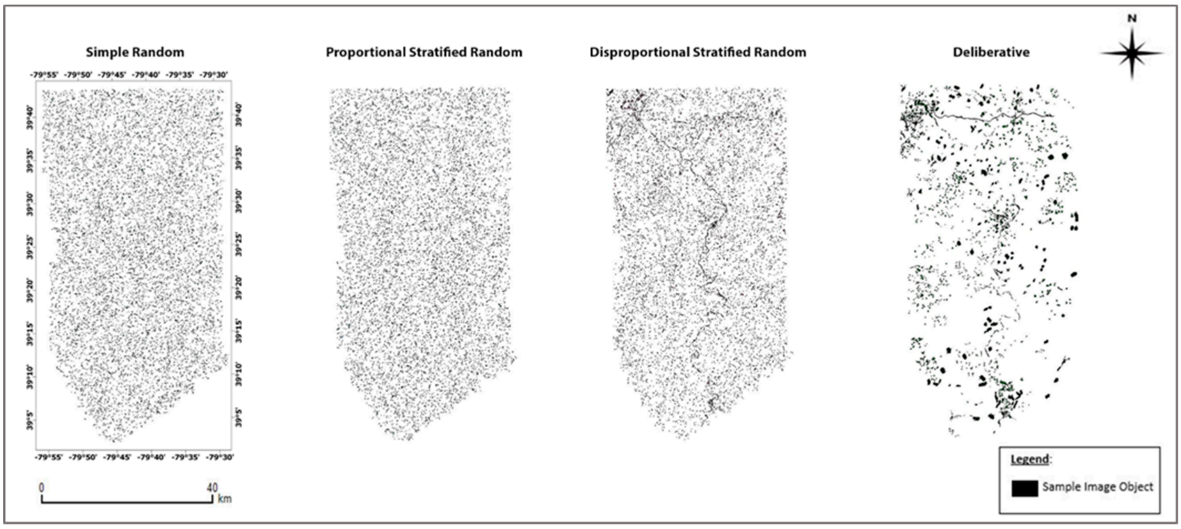


Figure 5. Large regional-scale training sample datasets.

Table 2a. Small subset sample sets.

Sample Selection Method	**Number of Samples per Class**				
	Forest	**Grass**	**Other**	**Water**	**Total # of Samples**
Small Subset Simple Random	290	67	53	9	419
Small Subset Proportional Stratified Random	305	59	35	20	419
Small Subset Disproportional Stratified Random	209	84	84	42	419
Small Subset Deliberative	139	100	100	80	419

Table 2b. Small regional sample sets.

Sample Selection Method	Number of Samples per Class				
	Forest	Grass	Other	Water	Total # of Samples
Small Regional Simple Random	341	50	26	2	419
Small Regional Proportional Stratified Random	333	65	18	3	419
Small Regional Disproportional Stratified Random	209	84	84	42	419
Small Regional Deliberative	254	80	69	16	419

Table 2c. Small regional sample sets.

Sample Name	Number of Samples per Class				
	Forest	Grass	Other	Water	Total # of Samples
Large Regional Simple Random	8183	1178	600	39	10,000
Large Regional Proportional Stratified Random	7984	1553	408	55	10,000
Large Regional Disproportional Stratified Random	5000	2000	2000	1000	10,000
Large Regional Deliberative	6087	1897	1651	365	10,000

2.7.1. Simple Random Sampling

The eCognition version 9.3 client does not offer a tool for selecting random samples, and therefore the select random polygon tool in QGIS was used.

2.7.2. Proportional Stratified Random Sampling

Because a stratified approach requires a priori strata, a rule-based classification developed through the expert system was applied to both the regional-scale dataset and the subset dataset (Figure 6) to estimate the strata sizes for the subset and regional datasets.

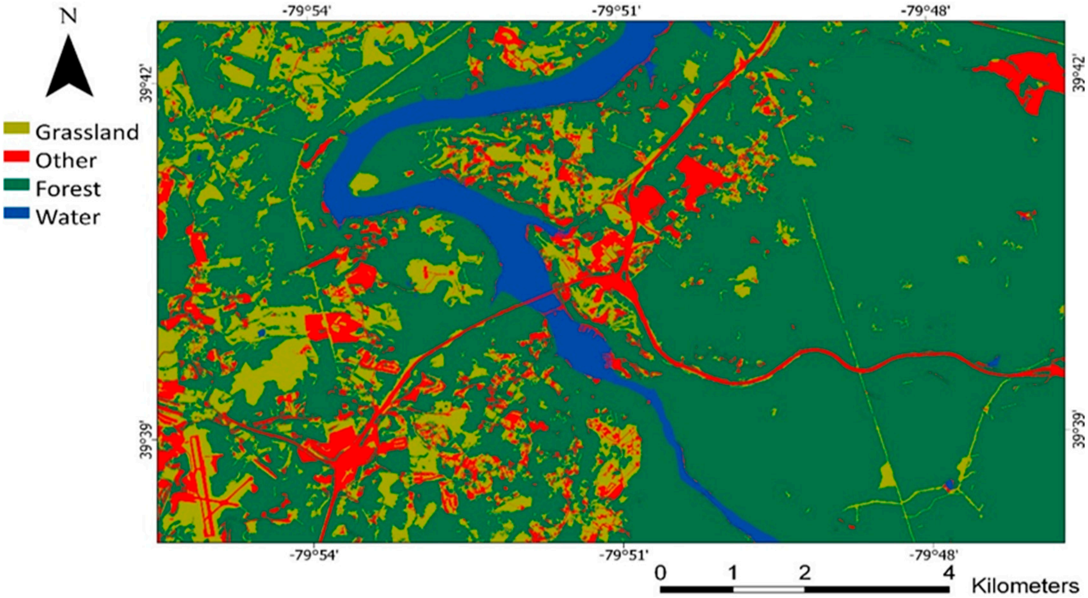


Figure 6. Rule-based classification of subset area.

The ruleset contained 16 individual rules. The accuracy of the rule-based classifications was evaluated using the samples from the large regional-scale validation dataset and had an overall accuracy of 98.1%. The strata size for both the subset and regional-scale datasets were determined by the total area occupied by each class. Table 3 summarizes the proportions of the strata for both datasets. Simple random sampling was used within each stratum to obtain samples for both the subset and regional-scale datasets.

Table 3. Class strata sizes for subset and regional datasets.

Class	**Proportion of Total Area Occupied**	
	Subset Dataset	**Regional Dataset**
Forest	72.73%	79.84%
Grassland	14.10%	15.53%
Other	8.34%	4.08%
Water	4.84%	0.55%

2.7.3. Disproportional Stratified Random Sampling

It was not possible to test an equalized stratified sampling approach because the water class is too rare to provide sufficient samples for a 25% proportion. Consequently, a disproportional stratified approach was chosen. For this sample, the class proportions were defined as 50% forest, 20% grassland, 20% other, and 10% water. These proportions were selected as intermediate values between the random and equalized stratified proportions to ensure a larger representation of the less common classes than in the random dataset. The same values were used for the small subset and small and large regional sample sets.

2.7.4. Deliberative Sampling

The deliberative sample was produced via on-screen digitizing by the analyst using the sample selection tool in eCognition Developer. No attempt was made to avoid spatial autocorrelation in the samples selected, for example, by avoiding samples that were spatially adjacent, because manual selection of samples is generally characterized by autocorrelation [48].

2.8. Cross-Validation Strategies

The cross-validation tuning methods were conducted using the trainControl function within the caret package [49] in R Studio 1.1.383. A separate classification was conducted for each cross-validation tuning method used and each sample set. The three cross-validation strategies tested were k-fold, leave-one-out, and Monte Carlo.

2.9. Supervised Classification

A radial basis function kernel (RBF) Support Vector Machines (SVM) was chosen as the supervised machine learning classifier for several reasons:

1. SVM is a commonly used supervised classifier in remote sensing analyses [17].
2. SVM is a non-parametric classifier, meaning it makes no assumption regarding the underlying data distribution. This may be advantageous for a small sample set [50].
3. SVMs are able to perform well with relatively small training datasets when compared to other commonly used classifiers.
4. SVMs are attractive for their ability to find a balance between accuracy and generalization [51].

A total of 36 individual classifications were conducted, each using a different combination of sample and validation methods: 3 categories of approaches at different spatial scales and sample sizes (small subset, small regional, and large regional) × 4 sample selection methods (simple random, proportional stratified random, disproportional stratified random, and deliberative) × 3 cross-validation tuning methods (k-fold, leave-one-out, and Monte Carlo) = 36 classifications.

Table 4 details all subset-trained and regional-trained classifications. The SVM classifications were conducted within R Studio client version 1.1.383 using the e1071 [52] and caret packages [49] on a Dell Optiplex 980 workstation with an Intel i7 2.80 GHz processor with 16.0 GB of memory running Windows 8.1 Enterprise. The processing time for all classifications were recorded using the microbenchmark package [53]. Processing runtime values should be interpreted as indications of relative speed and not as absolute values as they are highly dependent on the system architecture, CPU allocation, memory availability, and background system processes, among other factors.

Table 4. Classifications and associated abbreviations based on the sample selection method, training sample size, region of area collected, and cross-validation method.

		Cross-Validation Method								
		k-Fold (KF)	**Monte Carlo (MC)**	**Leave-One-Out (LOO)**	**k-Fold (KF)**	**Monte Carlo (MC)**	**Leave-One-Out (LOO)**	**k-Fold (KF)**	**Monte Carlo (MC)**	**Leave-One-Out (LOO)**
		Small Subset-Trained Classification			**Small Regional-Scale Trained Classification**			**Large Regional-Scale Trained Classification**		
Sample Selection Method	**Simple Random (SR)**	Small-Subset-(SR-KF)	Small-Subset-(SR-MC)	Small-Subset-(SR-LOO)	Small-Regional-(SR-KF)	Small-Regional-(SR-MC)	Small-Regional-(SR-LOO)	Large-Regional-(SR-KF)	Large-Regional-(SR-MC)	Large-Regional-(SR-LOO)
	Proportional Stratified Random (PSTR)	Small-Subset-(PSTR-KF)	Small-Subset-(PSTR-MC)	Small-Subset-(PSTR-LOO)	Small-Regional-(PSTR-KF)	Small-Regional-(PSTR-MC)	Small-Regional-(PSTR-LOO)	Large-Regional-(PSTR-KF)	Large-Regional-(PSTR-MC)	Large-Regional-(PSTR-LOO)
	Disproportional Stratified Random (DSTR)	Small-Subset-(DSTR-KF)	Small-Subset-(DSTR-MC)	Small-Subset-(DSTR-LOO)	Small-Regional-(DSTR-KF)	Small-Regional-(DSTR-MC)	Small-Regional-(DSTR-LOO)	Large-Regional-(DSTR-KF)	Large-Regional-(DSTR-MC)	Large-Regional-(DSTR-LOO)
	Deliberative (DL)	Small-Subset-(DL-KF)	Small-Subset-(DL-MC)	Small-Subset-(DL-LOO)	Small-Regional-(DL-KF)	Small-Regional-(DL-MC)	Small-Regional-(DL-LOO)	Large-Regional-(DL-KF)	Large-Regional-(DL-MC)	Large-Regional-(DL-LOO)

2.10. Error Assessment

Each of the trained classifications was tested against a large, randomly sampled validation dataset (n = 10,000) collected from the regional-scale dataset. Results for each classification were reported in a confusion matrix programmed via the caret package in the R statistical client. User's and producer's accuracies were calculated as well as overall accuracy and the kappa coefficient. Additionally, McNemar's test [54] was used to evaluate the statistical significance of differences observed between the k-fold tuned classifications. McNemar's test is a non-parametric evaluation of the statistical differences between two classifications with related samples [55]. A *p*-value smaller than 0.05 indicates a one-sided 95% confidence that the differences in accuracy between the classifications are statistically significant.

3. Results and Discussion

3.1. Performance of Sample Selection Methods

Figure 7 summarizes the overall accuracies of the classification of the entire regional dataset using the various training samples and k-fold (k = 10) cross-validation. Within each spatial scale and sample size (i.e., subset, small regional, and large regional), the disproportional stratified random (DSTR) samples consistently resulted in the highest overall accuracy although it is notable that variations between the performance of the classifications trained using the different statistical-based sampling methods were small, less than 2%.

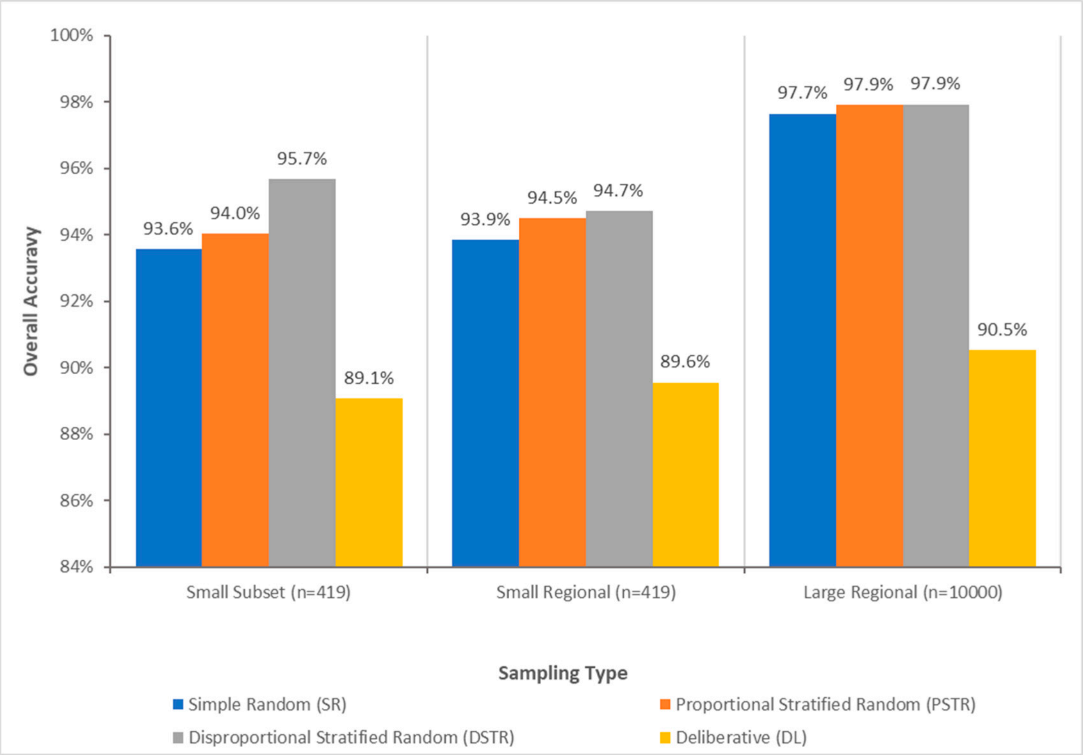


Figure 7. Overall accuracies of the regional classifications using small subset, small regional, and large regional training datasets and k-fold (k = 10) cross-validation tuning.

Despite the small differences between some of the classification accuracies, the McNemar's test results, shown in Table 5, indicate that most of the differences were statistically significant. The only exceptions were the differences in the classification accuracies for the large regional-trained SR, PSTR, and DSTR, which indicates that when the sample size is very large (n = 10,000 in this case), differences between sampling methods is less important.

Classifications trained with the SR sample resulted in a slightly lower accuracy than those trained with the PSTR and DSTR samples. This suggests that sample stratification is advantageous for SVM classifiers, as stratification ensures that rare classes are sampled at a rate that is either equal to or greater than their proportion in the dataset, depending on whether proportional or disproportional stratified approaches are selected.

The DSTR sampling method was designed to provide a much larger number of samples from minority classes, such as the water and other classes, than the simple random or proportional stratified random sampling methods (Tables 2a–2c). Using SR sampling, the number of samples from the minority classes may vary greatly, depending on random chance, especially when the total number of samples is small (e.g., 419 in this case). This can be seen in the small-regional SR and PSTR sample sets, where only 2 and 3 samples, respectively, were collected for water (Table 2b). The number of samples selected for the rare classes is important; Waske et al. [56] found that SVM was negatively affected by dataset imbalance. Stehman [57] also found that sample stratification resulted in improved classification accuracy due to an increased sample selection from the minority classes. The results of our study emphasize the value of disproportional stratified sample selection to reduce class imbalance and ensure minority class representation in the training set.

Table 5. McNemar's test *p*-values for small subset, small regional, and large regional-trained classifications using k-fold (k = 10) cross-validation tuning. (* Indicates the differences between classifications that are statistically significant, $p < 0.05$.)

Subset-SR-KF	Subset-PSTR-KF	Subset-DSTR-KF	Subset-DL-KF	Small-Regional-SR-KF	Small-Regional-PSTR-KF	Small-Regional-DSTR-KF	Small-Regional-DL-KF	Large-Regional-SR-KF	Large-Regional-PSTR-KF	Large-Regional-DSTR-KF	Large-Regional-DL-KF	
	<0.001 *	<0.001 *	<0.001 *	<0.001 *	<0.001 *	<0.001 *	<0.001 *	<0.001 *	<0.001 *	<0.001 *	<0.001 *	Subset-SR-KF
		0.007 *	<0.001 *	<0.001 *	<0.001 *	0.004 *	<0.001 *	<0.001 *	<0.001 *	<0.001 *	<0.001 *	Subset-PSTR-KF
			<0.001 *	0.043 *	<0.001 *	<0.001 *	<0.001 *	<0.001 *	<0.001 *	<0.001 *	<0.001 *	Subset-DSTR-KF
				<0.001 *	<0.001 *	<0.001 *	0.009 *	<0.001 *	<0.001 *	<0.001 *	<0.001 *	Subset-DL-KF
					0.013 *	0.002 *	<0.001 *	<0.001 *	<0.001 *	<0.001 *	<0.001 *	Small-Regional-SR-KF
						0.031 *	<0.001 *	<0.001 *	<0.001 *	<0.001 *	<0.001 *	Small-Regional-PSTR-KF
							<0.001 *	<0.001 *	<0.001 *	<0.001 *	<0.001 *	Small-Regional-DSTR-KF
								<0.001 *	<0.001 *	<0.001 *	<0.001 *	Small-Regional-DL-KF
									0.108	0.113	<0.001 *	Large-Regional-SR-KF
										0.162	<0.001 *	Large-Regional-PSTR-KF
											<0.001 *	Large-Regional-DSTR-KF
												Large-Regional-DL-KF

The classifications trained from the deliberative (DL) samples consistently had lower accuracies across all sample sets (Figure 7). The low accuracy for the classifications with the DL samples indicates that samples acquired though expert selection of the training data did not adequately characterize the dataset. Notably, the DL samples have higher spatial autocorrelation than samples collected via the statistical-based methods (Figure 5). This is not surprising; as noted previously, human-based deliberative sampling has a high potential for spatial autocorrelation [48]. High spatial autocorrelation in sample sets may result in a reduction in the effective sample size [13]. A univariate Moran's I test indicates that the small subset, small regional, and large regional DL samples all show positive spatial autocorrelation, with values of 0.950, 0.692, and 0.985, respectively. While the SR, PSTR, and DSTR samples also contained positive spatial autocorrelation, ranging from 0.208 (subset-SR) to 0.661 (large-regional-DSTR), they showed less positive spatial autocorrelation than all DL sample sets. Stratified sampling, especially disproportional stratified sampling, tends to favor some autocorrelation, since samples are, by definition, not completely random.

The similar performance between the classifications trained from the small subset and small regional-scale samples and the much higher accuracy reported from the large regional-scale sample indicate that the geographic location of the sample may not be as important as the sample size in determining the accuracy of the supervised SVM classifications. This is notable as selecting a small sample from a subset area is less expensive in terms of effort than selecting a small sample from a regional-scale area, especially if field data collection is needed. It should also be mentioned that the regional-scale area in this analysis was generally homogenous, which allowed the selection of a single subset area that contained many examples of all four classes of interest. In areas or datasets that are more heterogeneous or contain extreme minority classes limited to separate geographic regions of a regional-scale dataset, multiple subset areas may be needed for subset-based sampling to be effective. The fact that the classifications trained from both small sample sets were less accurate than the large regional-trained benchmark classification emphasizes that large numbers of statistical-based samples can raise the accuracy of SVM classifications substantially.

3.2. Performance of Cross-Validation Tuning Methods

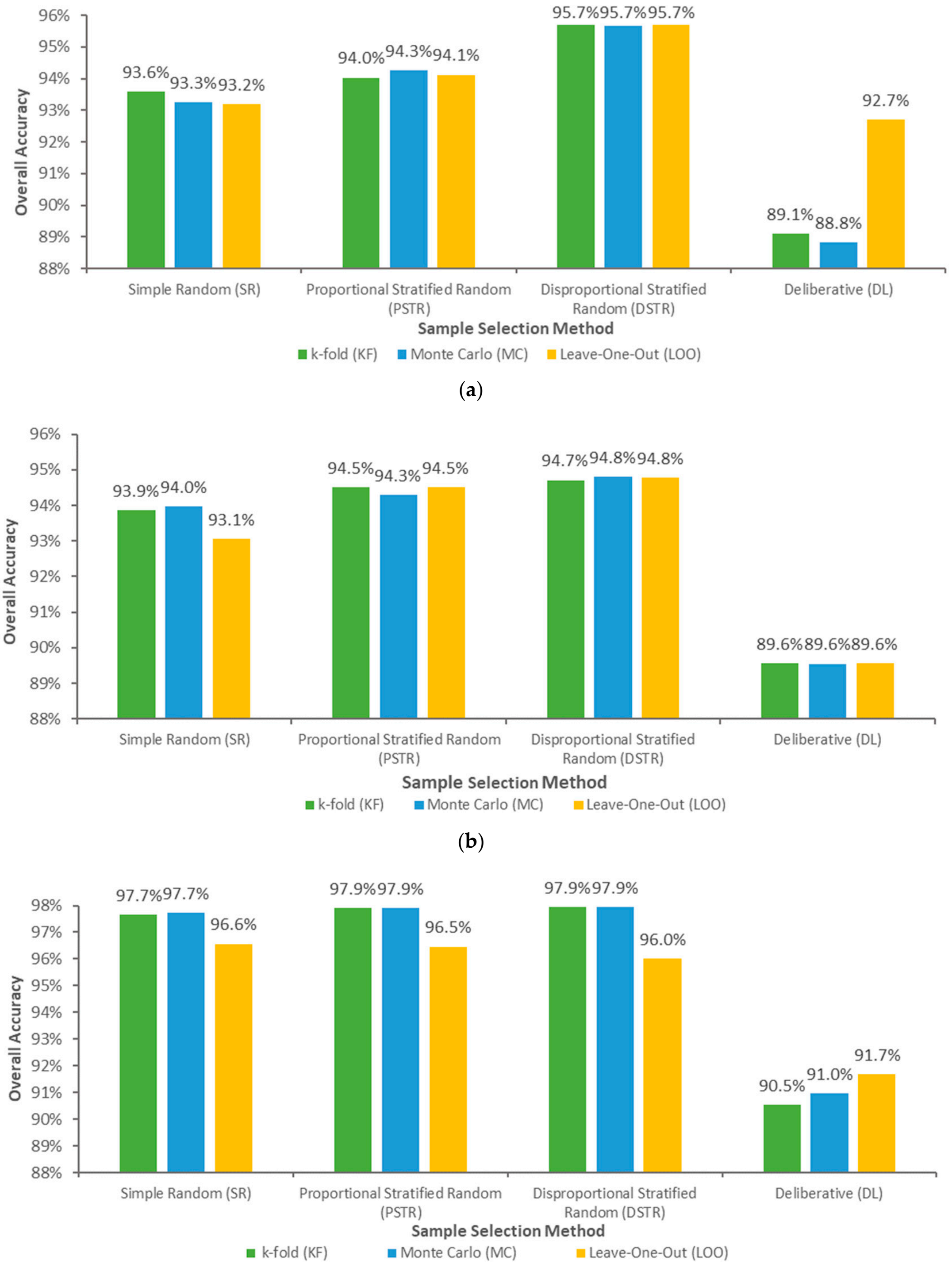
There was no consistent pattern for the overall accuracy of the classifications trained from the small subset samples and tuned using the three cross-validation methods (k-fold, Monte Carlo, and leave-one-out). As seen in Figure 8a, when the SR samples were used as training data, the k-fold (KF) method provided slightly higher accuracy than the Monte Carlo (MC) or leave-one-out (LOO) cross-validation. MC proved the best method for the PSTR samples. KF, MC, and LOO tuning resulted in equal overall accuracies for the DSTR samples. Overall, the differences in overall accuracy between the tuning methods on the small subset statistical-sample trained classifications was less than 1%.

The cross-validation tuning methods using the small regional SR, PSTR, and DSTR training data applied to the SVM classification all showed high values for overall accuracy (Figure 8b) and inconsistent results for the different tuning methods, similar to the results of the small subset training data. LOO had slightly lower performance on the SR classifications, but this decrease in performance was also less than 1%. The DSTR classifications had the highest overall accuracy, irrespective of the cross-validation tuning methods, with the MC- and LOO-tuned DSTR classifications resulting in 94.8% and the k-fold tuned-DSTR resulting in 94.7% overall accuracy.

For the large-regional statistical sample-trained classifications, the MC- and KF-tuned classifications consistently outperformed LOO. This indicates that LOO is less effective for tuning than KF and MC when dealing with large statistical-based sample sets. Both MC and KF had the same overall accuracy for the large-regional SR, PSTR, and DSTR classifications.

The deliberative-trained classifications for both the small subset and regional scales had much lower performances than the statistical based classifications. For the small-regional DL classifications, LOO matched the overall accuracy of KF and MC at 89.6% (Figure 8b). However, for both the small-subset (Figure 8a) and large-regional DL classifications (Figure 8c), LOO tuning improved overall accuracy by 3.6% and 1.2%, respectively over the KF-DL classifications.

The confusion matrices of the small subset-DL-KF (Table 6), MC (Table 7), and LOO (Table 8) show that the increase in performance of the small subset-DL-LOO classification was due to a substantial increase in the producer's accuracy of the forest class and the increased user's accuracy of the grassland class, though at the cost of decreases in the other class accuracies, most notably in the grassland producer's accuracy. As the forest and grassland classes combined make up 93.4% of the validation dataset, improving the average class accuracies of these two classes led to a marked improvement of the overall accuracy.



(**a**)

(**b**)

(**c**)

Figure 8. (**a**) Overall accuracies of the small-subset SVM training data classifications using k-fold (k = 10), Monte Carlo, and leave-one-out cross-validation tuning. (**b**) Overall accuracies of the small-regional training data SVM classifications using k-fold (k = 10), Monte Carlo, and leave-one-out cross-validation tuning. (**c**) Overall accuracies of the classifications using large-regional training data and k-fold (k = 10), Monte Carlo, and leave-one-out cross-validation tuning.

Table 6. Confusion matrix for the classification trained with the subset-DL-KF data set.

		Reference Data (No. Objects)					
		Forest	**Grassland**	**Other**	**Water**	**Total**	**User's Accuracy**
Classified data (No. objects)	**Forest**	7238	31	4	0	7273	99.5%
	Grassland	699	1151	85	0	1935	59.5%
	Other	136	73	479	28	716	66.9%
	Water	12	1	22	41	76	53.9%
	Total	8085	1256	590	69	10,000	**Overall accuracy: 89.1%**
	Producer's accuracy	89.5%	91.6%	81.2%	59.4%		

Table 7. Confusion Matrix for the classification trained with the subset-DL-MC data set.

		Reference Data (No. Objects)					
		Forest	**Grassland**	**Other**	**Water**	**Total**	**User's Accuracy**
Classified data (No. objects)	**Forest**	7723	36	6	0	7265	99.4%
	Grassland	723	1133	80	0	1936	58.5%
	Other	129	86	487	29	731	66.6%
	Water	10	1	17	40	68	58.8%
	Total	8085	1256	590	69	10,000	**Overall accuracy: 88.8%**
	Producer's accuracy	89.3%	90.2%	82.5%	58.0%		

Table 8. Confusion Matrix for the classification trained with the subset-DL-LOO data set.

		Reference Data (No. Objects)					
		Forest	**Grassland**	**Other**	**Water**	**Total**	**User's Accuracy**
Classified data (No. objects)	**Forest**	7728	174	5	0	7907	97.7%
	Grassland	176	1034	95	1	1306	79.2%
	Other	176	48	466	26	716	65.1%
	Water	5	0	24	42	71	59.2%
	Total	8085	1256	590	69	10,000	**Overall accuracy: 92.7%**
	Producer's accuracy	95.6%	82.3%	79.0%	60.9%		

However, both the leave-one-out and Monte Carlo tuning required longer processing time than the k-fold (k = 10) cross-validation tuning (Table 9). When sample sizes become very large (n = 10,000), leave-one-out tuning may become prohibitively slow; though with advances in processor technology, this may be less of a concern for the future.

Table 9. Processing time metrics.

Classification	Processing Time (seconds)
SVM-Subset-KF	10
SVM-Large-Regional-KF	468
SVM-Subset-MC	17
SVM-Large-Regional-MC	876
SVM-Subset-LOO	313
SVM-Large-Regional-LOO	489,960

Since no cross-validation method was consistently superior for tuning SVM classifications trained from the SR, PSTR, and DSTR sample sets and for all sample sets, k-fold may be the most effective and efficient method for cross-validation parameter tuning for SVM classifiers.

4. Conclusions

This investigation explored the effects sample acquisition method, sample geographic distribution, and cross-validation tuning methods in regional-scale land-cover classifications of HR remotely sensed data. Based upon the results presented in this analysis, a random sample, possibly combined with stratification techniques to ensure adequate representation of minority classes within training sample sets, is recommended. Deliberative samples should be avoided, possibly because of the tendency of humans to collect excessively highly autocorrelated samples. Stehman, [57] recommends using an underlying systematic sampling scheme to minimize spatial autocorrelation in sample sets.

Classifications trained from the small subset-based sample sets were found to have comparable performance to classifications trained from small sample sets acquired in a dispersed manner across the entire regional-scale study site. This is an important finding because if sample selection is expensive, especially if field checking is required, a relatively small sample set collected from a subset area of the regional-scale study area can be used. However, it is important to note that since the study area for this analysis was broadly homogenous, it was possible to select a single subset area that contained adequate examples of all four classes of interest for training data collection. In more heterogeneous environments, multiple subset areas may be needed to obtain the samples. Future research is needed on large-scale sample selection strategies in highly heterogeneous environments.

The relative accuracy of classifications produced with k-fold (k = 10), leave-one-out, and Monte Carlo cross-validation tuning methods when trained with the small subset, small regional, and large regional SR, PSTR, and DSTR data sets were not consistent. As the Monte Carlo and leave-one-out cross-validation tuning methods required greater processing resources and time, the k-fold cross-validation method may be preferable, especially for large sample sets. Regarding deliberative sampling methods, in both the small subset and large regional classifications, leave-one-out cross-validation tuning was more effective in increasing classifier performance when compared to the k-fold and Monte Carlo tuning.

In summary, for large regional-scale HR classifications, deliberative sampling should be avoided not only for accuracy assessment data but also for training data collection. Random samples are preferable, and data collected randomly from a small subset region is adequate, at least in relatively homogenous areas. Disproportional stratified sampling can be used to reduce the effect of imbalanced samples. Tuning is important, though the type of method used does not seem to have a large effect. k-fold tuning is possibly a good choice because it is relatively rapid.

Author Contributions: Conceptualization, C.A.R. and T.A.W.; Methodology, C.A.R. and T.A.W.; Software, C.A.R. and A.E.M.; Validation, C.A.R.; Formal Analysis, C.A.R.; Investigation, C.A.R.; Resources, C.A.R., and A.E.M.; Data Curation, C.A.R.; Writing-Original Draft Preparation, C.A.R.; Writing-Review & Editing, C.A.R., T.A.W., and A.E.M.; Visualization, C.A.R.; Supervision, T.A.W.; Project Administration, T.A.W.; Funding Acquisition, C.A.R.

Funding: The APC was funded by [West Virginia University Open Access Author Fund (OAAF)].

Acknowledgments: The authors would like to thank the three anonymous referees whose comments greatly strengthened this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fassnacht, F.E.; Hartig, F.; Latifi, H.; Berger, C.; Hernandez, J.; Corvalan, P.; Koch, B. Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. *Remote Sens. Environ.* **2014**, *154*, 102–114. [CrossRef]

2. Guo, Y.; Ma, L.; Zhu, F.; Liu, F. Selecting Training Samples from Large-Scale Remote-Sensing Samples Using an Active Learning Algorithm. In *Computational Intelligence and Intelligent Systems*; Li, K., Li, J., Liu, Y., Castiglione, A., Eds.; Springer: Singapore, 2016; pp. 40–51.
3. Lu, D.; Weng, Q. A survey of image classification methods and techniques for improving classification performance. *Int. J. Remote Sens.* **2007**, *28*, 823–870. [CrossRef]
4. Foody, G.M. Sample size determination for image classification accuracy assessment and comparison. *Int. J. Remote Sens.* **2009**, *30*, 5273–5291. [CrossRef]
5. Jin, H.; Stehman, S.V.; Mountrakis, G. Assessing the impact of training sample selection of accuracy of an urban classification: A case study in Denver, Colorado. *Int. J. Remote Sens.* **2014**, *35*, 2067–2081. [CrossRef]
6. Radoux, J.; Bogaert, P.; Fasbender, D.; Defourny, P. Thematic accuracy assessment of geographic object-based image classification. *Int. J. Geogr. Inf. Sci.* **2011**, *25*, 895–911. [CrossRef]
7. Stehman, S.V. Impact of sample size allocation when using stratified random sampling to estimate accuracy and area of land-cover change. *Remote Sens. Lett.* **2012**, *3*, 111–120. [CrossRef]
8. Ma, L.; Li, M.; Ma, X.; Cheng, K.; Du, P.; Liu, Y. A review of supervised object-based land-cover image classification. *ISPRS J. Photogramm. Remote Sens.* **2017**, *130*, 277–293. [CrossRef]
9. Blaschke, T. Object based image analysis for remote sensing. *ISPRS J. Photogramm. Remote Sens.* **2010**, *65*, 2–16. [CrossRef]
10. Foody, G.M.; Pal, M.; Rocchini, D.; Garzon-Lopez, C.X.; Bastin, L. The Sensitivity of mapping Methods to Reference Data Quality: Training Supervised Image Classifications with Imperfect Reference Data. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 199. [CrossRef]
11. Congalton, R.G. A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data. *Remote Sens. Environ.* **1991**, *37*, 35–46. [CrossRef]
12. Foody, G.M.; Mathur, A.; Sanchez-Hernandez, C.; Boyd, D.S. Training Set Size Requirements for the Classification of a Specific Class. *Remote Sens. Environ.* **2006**, *104*, 1–14. [CrossRef]
13. Mu, X.; Hu, M.; Song, W.; Ruan, G.; Ge, Y.; Wang, J.; Huang, S.; Yan, G. Evaluation of Sampling Methods for Validation of Remotely Sensed Fractional Vegetation Cover. *Remote Sens.* **2015**, *7*, 16164–16182. [CrossRef]
14. Chen, D.M.; Stow, D. The Effect of Training Strategies on Supervised Classification at Different Spatial Resolutions. *Photogramm. Eng. Remote Sens.* **2002**, *68*, 1155–1161.
15. Chen, D.; Stow, D.A.; Gong, P. Examining the effect of spatial resolution and texture windows size on classification accuracy: An urban environment case. *Int. J. Remote Sens.* **2004**, *25*, 2177–2192. [CrossRef]
16. Congalton, R.G. A comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data. *Photogramm. Eng. Remote Sens.* **1988**, *54*, 593–600.
17. Maxwell, A.E.; Warner, T.A.; Fang, F. Implementation of machine-learning classification in remote sensing: An applied review. *Int. J. Remote Sens.* **2018**, *39*, 2784–2817. [CrossRef]
18. Stehman, S.V. Estimating area and map accuracy for stratified random sampling when the strata are different from the map classes. *Int. J. Remote Sens.* **2014**, *35*, 4923–4939. [CrossRef]
19. Stehman, S.V.; Foody, G.M. Accuracy assessment. In *The SAGE Handbook of Remote Sensing*; Warner, T.A., Nellis, M.D., Foody, G.M., Eds.; Sage Publications Ltd.: London, UK, 2009; pp. 129–145. ISBN 9781412936163.
20. Pal, M.; Foody, G.M. Evaluation of SVM, RVM and SMLR for accurate image classification with limited ground data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 1344–1355. [CrossRef]
21. Demir, B.; Minello, L.; Bruzzone, L. An Effective Strategy to Reduce the Labeling Cost in the Definition of Training Sets by Active Learning. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 79–83. [CrossRef]
22. Wuttke, S.; Middlemann, W.; Stilla, U. Concept for a compound analysis in active learning remote sensing. In Proceedings of the International Archives of the Photogrammetry, Remote Sensing, and Spatial Information Sciences, Munich, Germany, 25–27 March 2015; Volume XL-3(W2), pp. 273–279. [CrossRef]
23. Babcock, C.; Finely, A.O.; Bradford, J.B.; Kolka, R.K.; Birdsey, R.A.; Ryan, M.G. LiDAR based prediction of forest biomass using hierarchial models with spatially varying coefficients. *Remote Sens. Environ.* **2015**, *169*, 113–127. [CrossRef]
24. Brenning, A. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. In Proceedings of the 2012 IEEE International Geoscience and Remote Sensing Symposium, Munich, Germany, 22–27 July 2012; pp. 5372–5375. [CrossRef]

25. Cracknell, M.J.; Reading, A.M. Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Comput. Geosci.* **2014**, *63*, 22–33. [CrossRef]
26. Sharma, R.C.; Hara, K.; Hirayama, H. A Machine Learning and Cross-Validation Approach for the Discrimination of Vegetation Physiognomic Types Using Satellite Based Multispectral and Multitemporal Data. *Scientifica* **2017**, *2017*, 9806479. [CrossRef] [PubMed]
27. Duro, D.C.; Franklin, S.E.; Dubé, M.G. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote Sens. Environ.* **2012**, *118*, 259–272. [CrossRef]
28. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc.* **1974**, *36*, 111–147. [CrossRef]
29. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer: New York, NY, USA, 2009.
30. Picard, R.R.; Cook, R.D. Cross-Validation of Regression Models. *J. Am. Stat. Assoc.* **1984**, *387*, 575–583. [CrossRef]
31. Braun, E.L. *Deciduous Forests of Eastern North America*; Hafner Publishing Company: New York, NY, USA, 1950.
32. Maxwell, A.E.; Strager, M.P.; Warner, T.A.; Zegre, N.P.; Yuill, C.B. Comparison of NAIP orthophotography and RapidEye satellite imagery for mapping of mining and mine reclamation. *GISci. Remote Sens.* **2014**, *51*, 301–320. [CrossRef]
33. WVU NRAC. Aerial Lidar Acquistion Report: Preston County and North Branch (Potomac) LIDAR *.LAS 1.2 Data Comprehensive and Bare Earth. West Virginia Department of Environmental Protection. Available online: http://wvgis.wvu.edu/lidar/data/WVDEP_2011_Deliverable4/WVDEP_deliverable_4_Project_Report.pdf (accessed on 1 December 2018).
34. ESRI. *ArcGIS Desktop: Release 10.5.1*; Environmental Systems Research Institute: Redlands, CA, USA, 2017.
35. Charaniya, A.P.; Manduchi, R.; Lodha, S.K. Supervised parametric classification of aerial LIDAR data. In Proceedings of the IEEE 2004 Conferences on Computer Vision and Pattern Recognition Workshop, Washington, DC, USA, 27 June–2 July 2004.
36. Kashani, A.G.; Olsen, M.; Parrish, C.; Wilson, N. A Review of LIDAR Radiometric Processing: From Ad Hoc Intensity correction to Rigorous Radiometric Calibration. *Sensors* **2015**, *15*, 28099–28128. [CrossRef] [PubMed]
37. Song, J.H.; Han, S.H.; Yu, K.Y.; Kim, Y.I. Assessing the possibility of land-cover classification using LIDAR intensity data. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2002**, *34*, 259–262.
38. Maxwell, A.E.; Warner, T.A.; Strager, M.P.; Conley, J.F.; Sharp, A.L. Assessing machine learning algorithms and image- and LiDAR-derived variables for GEOBIA classification of mining and mine reclamation. *Int. J. Remote Sens.* **2015**, *36*, 954–978. [CrossRef]
39. Beşol, B.; Alganci, U.; Sertel, E. The use of object based classification with nDSM to increase the accuracy of building detection. In Proceedings of the 25th Signal Processing and Communications Applications Conference (SIU), Antalya, Turkey, 15–18 May 2017.
40. Lear, R.F. NAIP Quality Samples. United States Department of Agriculture Aerial Photography Field Office. Available online: https://www.fsa.usda.gov/Internet/FSA_File/naip_quality_samples_pdf.pdf (accessed on 28 December 2018).
41. Trimble. *Trimble eCognition Suite 9.3.2*; Trimble Germany GmbH: Munich, Germany, 2018.
42. Petrie, G.; Toth, C.K. Airborne and Spaceborne Laser Profilers and Scanners. In *Topographic Laser Ranging and Scanning: Principles and Processing*; Shan, J., Toth, C.K., Eds.; CRC Press: Boca Raton, FL, USA, 2008.
43. Baatz, M.; Schäpe, A. Multiresolution segmentation—An optimization approach for high quality multi-scale image segmentation. In Proceedings of the Angewandte Geographische Informations-Verarbeitung XII, Karlsruhe, Germany, 30 June 2000; pp. 12–23.
44. Belgiu, M.; Drăgut, L. Comparing supervised and unsupervised multiresolution segmentation approaches for extracting buildings from very high resolution imagery. *ISPRS J. Photogramm. Remote Sens.* **2014**, *96*, 67–75. [CrossRef]
45. Drăgut, L.; Csillik, O.; Eisank, C.; Tiede, D. Automated parameterization for multi-scale image segmentation on multiple layers. *ISPRS J. Photogramm. Remote Sens.* **2014**, *88*, 119–127. [CrossRef] [PubMed]

46. Kim, M.; Warner, T.A.; Madden, M.; Atkinson, D. Multi-scale texture segmentation and classification of salt marsh using digital aerial imagery with very high spatial resolution. *Int. J. Remote Sens.* **2011**, *32*, 2825–2850. [CrossRef]
47. Maguigan, M.; Rodgers, J.; Dash, P.; Meng, Q. Assessing Net Primary Production in Montane Wetlands from Proximal, Airborne, and Satellite Remote Sensing. *Adv. Remote Sens.* **2016**, *5*, 118–130. [CrossRef]
48. Griffith, D.A. Establishing Qualitative Geographic Sample Size in the Presence of Spatial Autocorrelation. *Ann. Assoc. Am. Geogr.* **2013**, *103*, 1107–1122. [CrossRef]
49. Kuhn, M. Caret: Classification and Regression Training. R package Version 6.0-71. 2016. Available online: https://CRAN.R-project.org/package=caret (accessed on 21 February 2018).
50. Scheuenemeyer, J.H.; Drew, L.J. *Statistics for Earth and Environmental Scientists*; John Wiley & Sons: Hoboken, NJ, USA, 2010; ISBN 9780470650707.
51. Mountrakis, G.; Im, J.; Ogole, C. Support vector machines in remote sensing: A review. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 247–259. [CrossRef]
52. Meyer, D. Support Vector Machines: The Interface to Libsvm in Package e1071. R Package Version 6.0-71. 2012. Available online: https://CRAN.R-project.org/package=e1071 (accessed on 21 February 2018).
53. Ulrich, J.M. Microbenchmark: Accurate Timing Functions. R Package Version 1.4-4. 2018. Available online: https://cran.r-project.org/web/packages/microbenchmark/microbenchmark.pdf (accessed on 21 February 2018).
54. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157. [CrossRef]
55. Foody, G.M. Thematic Map Comparison: Evaluating the Statistical Significance of Differences in Classification Accuracy. *Photogramm. Eng. Remote Sens.* **2004**, *70*, 627–633. [CrossRef]
56. Waske, B.; Benediktsson, J.A.; Sveinsson, J.R. *Classifying Remote Sensing Data with Support Vector Machines and Imbalanced Training Data*; CMS 2009, LNCS 5519; Benediktsson, J.A., Kittler, J., Roli, F., Eds.; Springer: Berlin/Heidleberg, Germany, 2009; pp. 375–384.
57. Stehman, S.V. Sampling designs for accuracy assessment of land cover. *Int. J. Remote Sens.* **2009**, *30*, 5243–5272. [CrossRef]