Using the Random Forest Algorithm to Map Land Cover in the Coastal Watershed of New Hampshire from Landsat 8 Satellite Imagery

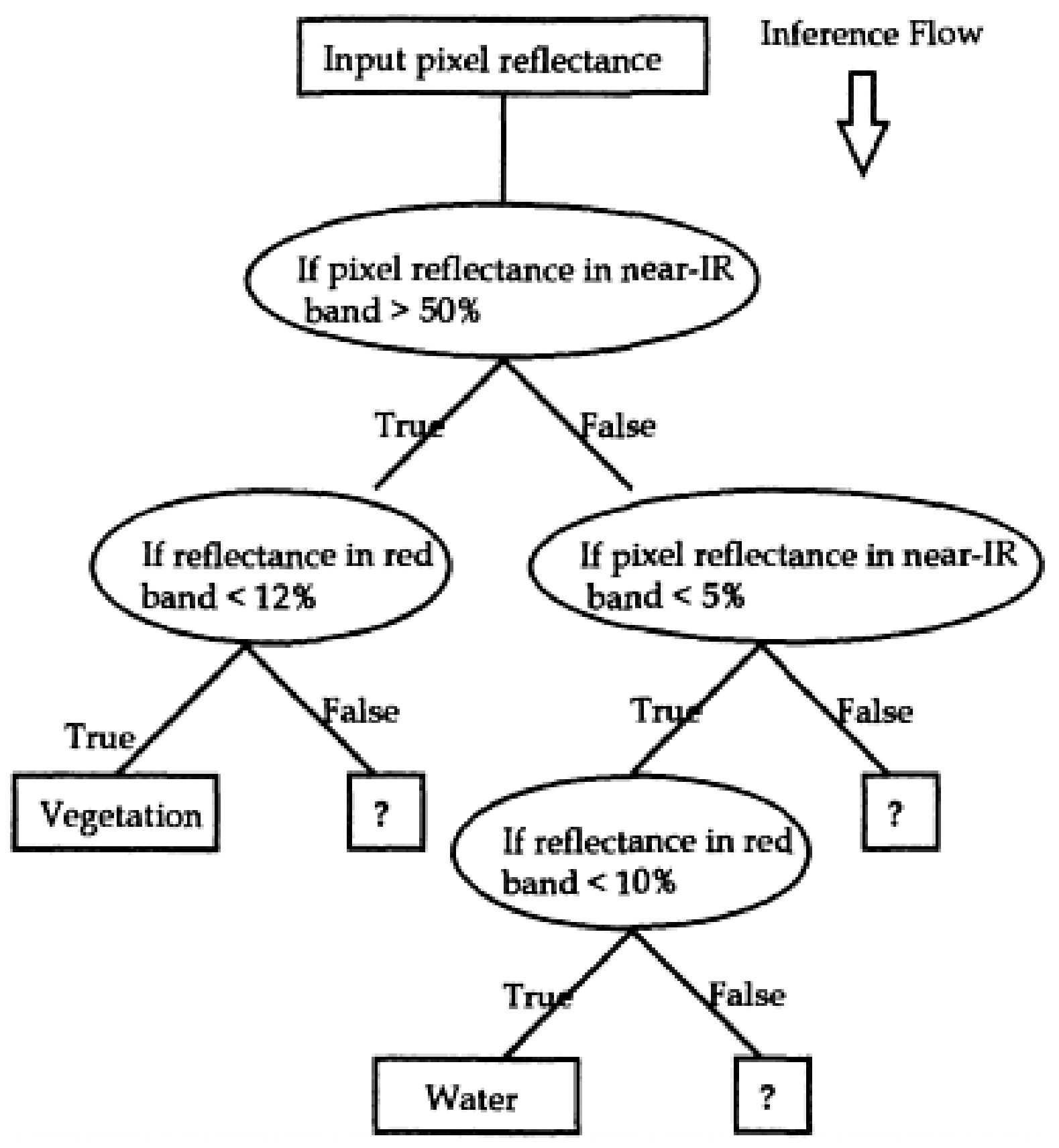
Lindsay Ledoux and Russell G. Congalton
Department of Natural Resources and the Environment
University of New Hampshire

Background

The data obtained from remotely sensed imagery is most useful when converted into meaningful information (e.g., thematic maps). These are maps that convey information related to a specific subject, or serve a specific purpose.

Traditionally, thematic map generation through classification has been grouped into three categories: unsupervised classification, supervised classification, and advanced classification algorithms such as decision-tree classifiers.

Computer technology has improved, and so have the ways to classify remotely sensed imagery. One of these innovative approaches is known as Random Forest (Breiman, 2001). The Random Forest (RF) classifier is a machine-learning algorithm in which multiple classification trees are networked. The foundation of the classifier is to build a set of randomly generated decision trees that are independent of each other. Each tree is constructed using a subset that is randomly generated from the original training data with sample replacement (Rodriguez-Galiano *et al.*, 2012). The results of the different trees are then manipulated by majority vote, and this is used to produce the final classification outcome.


Input pixel reflectance
Inference Flow
If pixel reflectance in near-IR band > 50%
True
False
If reflectance in red band < 12%
If pixel reflectance in near-IR band < 5%
True
False
Vegetation
?
True
False
If reflectance in red band < 10%
?
True
False
Water
?

Example of a Random Forest tree.
http://flylib.com/books/en/3.365.1.316/1/

Advantages of the Random Forest classifier include:

- An ability to withstand outliers and overfitting (Breiman, 2001)
- Efficiency in calculations
- Smooth handling of various visual features (e.g. color, shape, texture, etc.)
- Is non-parametric, so it makes no assumptions about the data distribution

Using this novel approach to classification, as well as using object-based image analysis, and data derived from the newest Landsat 8 satellite sensor, three thematic maps were generated, and their accuracies evaluated.

Objectives

- To create the most accurate land cover thematic map of the Great Bay Watershed possible using the latest improvements in image analysis techniques (including Random Forest) and Landsat 8 imagery
- Create a map utilizing all of the bands of Landsat 8
- Create a map utilizing only the equivalent Landsat 7 bands on the Landsat 8 sensor
- Assess the accuracy and compare the results of the maps produced in this study

Study Area

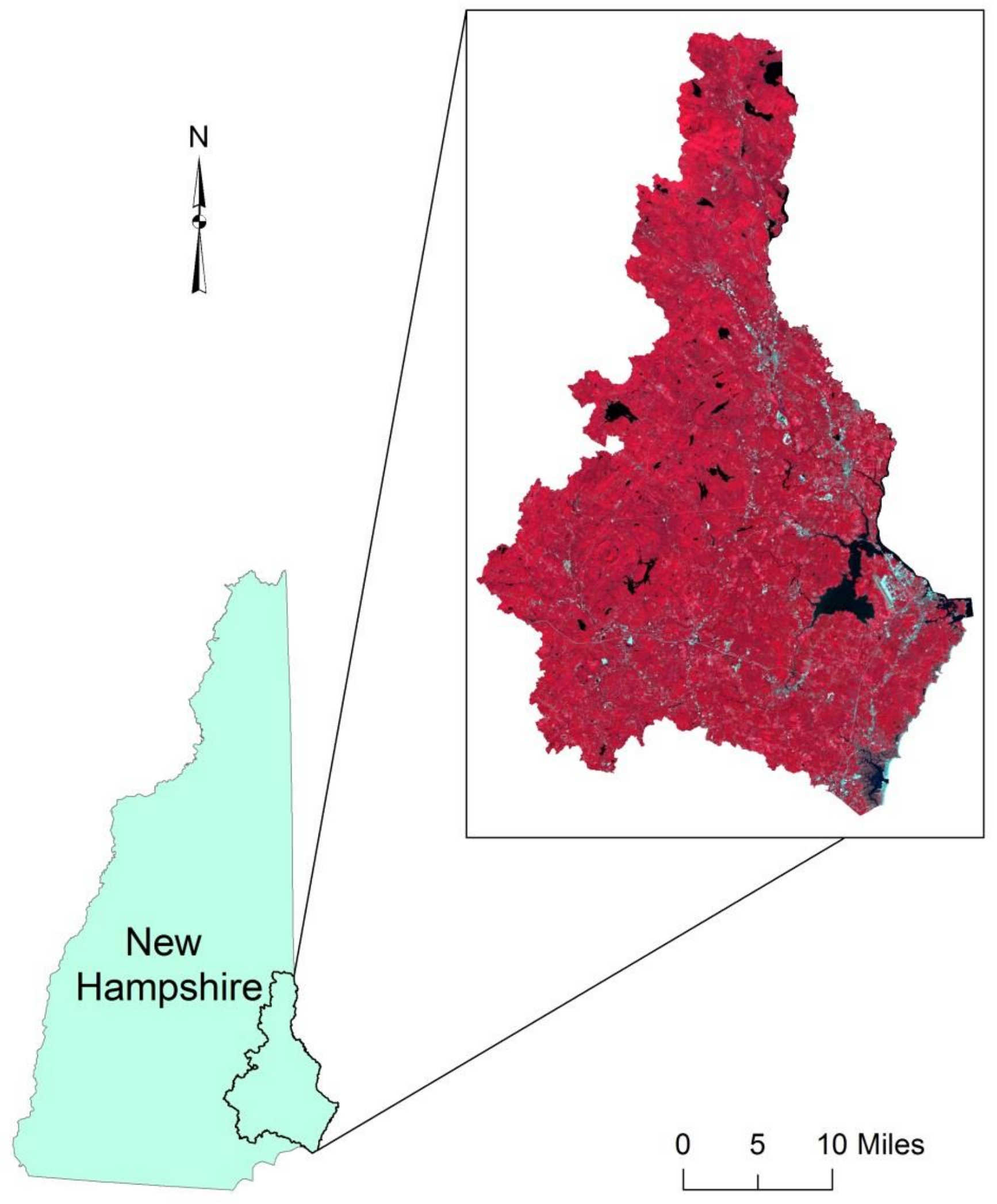


Figure 1. Study area of coastal Watershed of Southeastern New Hampshire relative to state. Inset is a Landsat 8 image of the study area in Color Infrared composite.

Methods

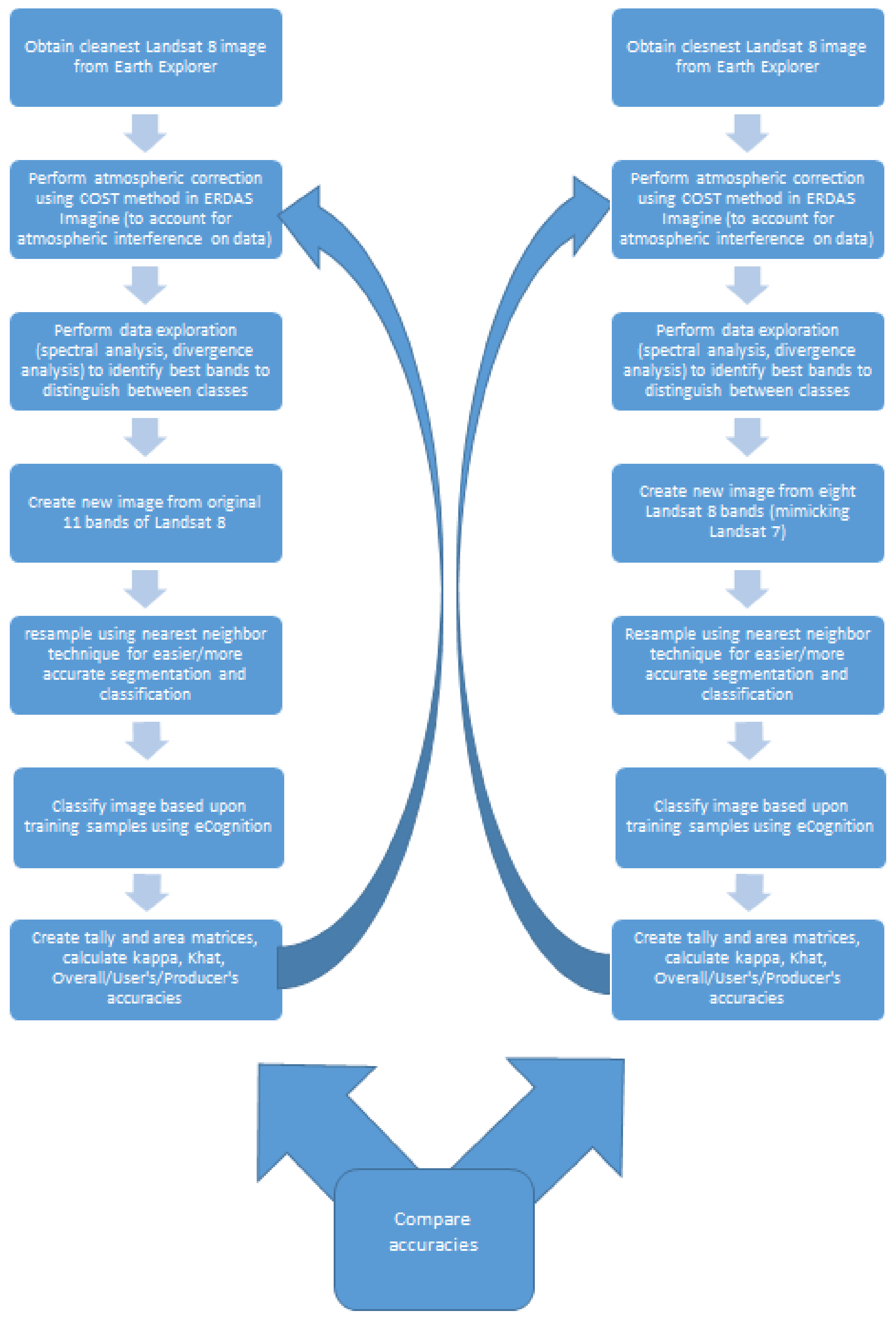

Obtain cleanest Landsat 8 image from Earth Explorer
Perform atmospheric correction using COST method in ERDAS Imagine (to account for atmospheric interference on data)
Perform data exploration (spectral analysis, divergence analysis) to identify best bands to distinguish between classes
Create new image from original 11 bands of Landsat 8
resample using nearest neighbor technique for easier/more accurate segmentation and classification
Classify image based upon training samples using eCognition
Create tally and area matrices, calculate kappa, Khat, Overall/User's/Producer's accuracies
Obtain clesnest Landsat 8 image from Earth Explorer
Perform atmospheric correction using COST method in ERDAS Imagine (to account for atmospheric interference on data)
Perform data exploration (spectral analysis, divergence analysis) to identify best bands to distinguish between classes
Create new image from eight Landsat 8 bands (mimicking Landsat 7)
Resample using nearest neighbor technique for easier/more accurate segmentation and classification
Classify image based upon training samples using eCognition
Create tally and area matrices, calculate kappa, Khat, Overall/User's/Producer's accuracies
Compare accuracies

Figure 2. Process chart of methodology used in this study, focusing on the comparison of the Landsat 8 data with the Landsat 8 mimicking Landsat 7 ETM+ data.

Jensen, J.R., 2005.

Results

Table 1. Traditional tally-based error matrix of classification of image using Landsat 8 and derivative bands.
AA= Active agriculture, BO= Beech/oak, COO= Cleared/other open, D= Developed, H= Hemlock, MF= Mixed forest, OW= Open water, OH= Other hardwoods, W= Wetlands, WRP= White/red pine

	Reference Data											
Map Data	Active agriculture	Cleared/ other open	Developed	Open water	Wetlands	Mixed Forest	Beech/ oak	Other hardwoods	Hemlock	White/ red pine	Row Totals	User's Accuracy
Active agriculture	88	28	5	0	3	0	1	0	0	0	125	70.40%
Cleared/other open	9	67	6	0	7	1	0	0	1	0	91	73.63%
Developed	1	5	87	1	7	0	0	0	1	1	103	84.47%
Open water	0	0	0	48	2	0	0	0	0	0	50	96.00%
Wetlands	0	0	1	1	66	1	1	0	1	1	72	91.67%
Mixed Forest	0	0	0	0	1	16	2	2	2	1	24	66.67%
Beech/oak	4	0	0	0	5	9	27	14	1	0	60	45.00%
Other hardwoods	0	0	0	0	4	17	19	34	2	0	76	44.74%
Hemlock	0	0	0	0	1	2	0	0	4	12	19	21.05%
White/red pine	0	0	1	0	4	4	0	0	9	35	53	66.04%
Column Totals	102	100	100	50	100	50	50	50	21	50	673	
Producer's Accuracy	86.27%	67.00%	87.00%	96.00%	66.00%	32.00%	54.00%	68.00%	19.05%	70.00%		**70.13%**

Congalton, R. and K. Green, 1999.

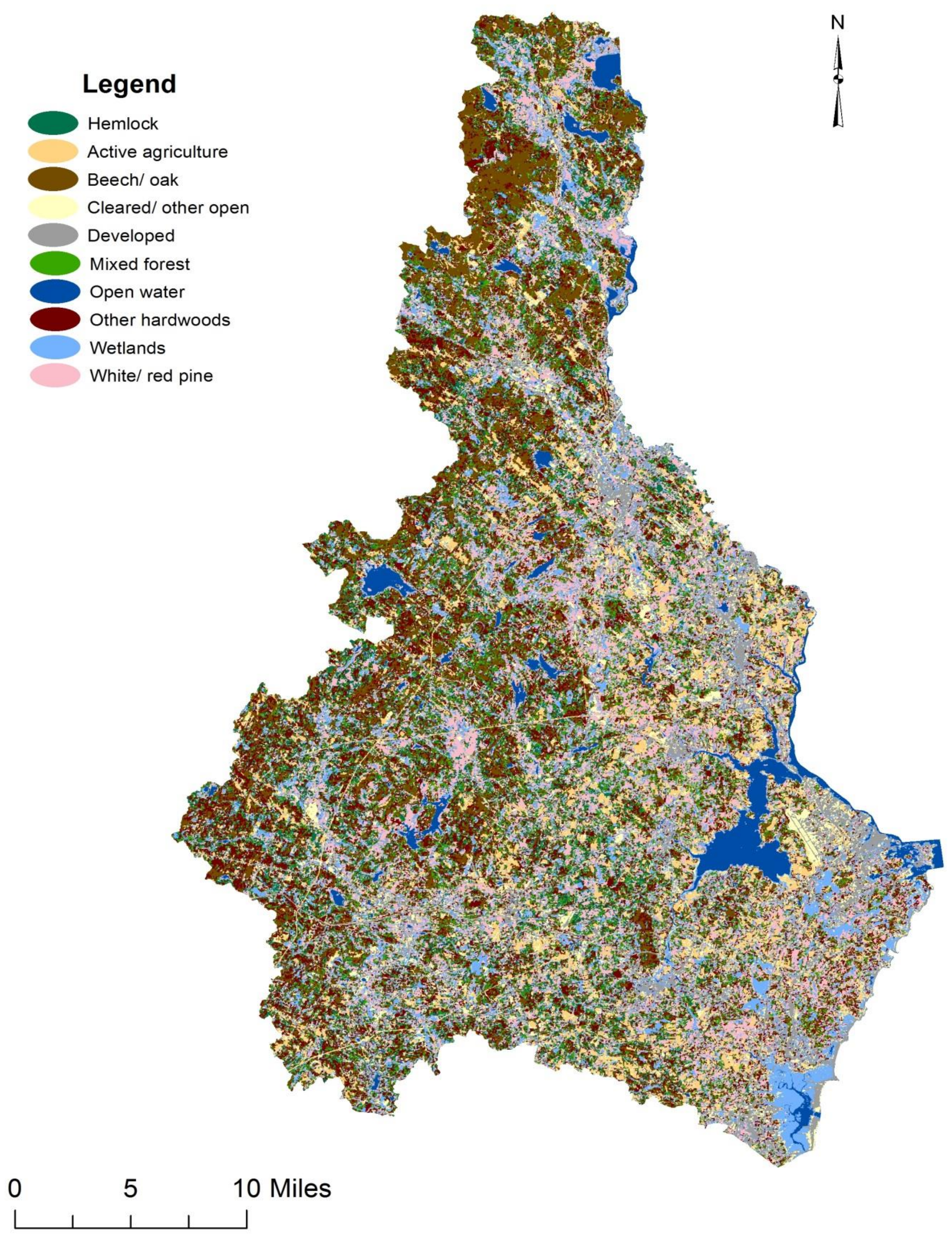


Figure 3. Final land cover thematic map generated by Random Forest and Landsat 8 and derivative bands.

Discussion

The accuracies of each class varied depending on the level of classification. The Level 1 land cover classes had the least omission and commission errors, while the more specific land cover classes (Level 3) contained the most omission and commission errors.

Statistic	Landsat 8 as Landsat 7	Landsat 8 all bands	final Landsat 8
KHAT (Kappa)	0.5159754	0.4536462	0.6618429
Z Test statistic	** 24.3069081	** 21.1279045	** 33.6054546
Z Test statistic	** 2.0643451		
Z Test statistic		** 7.1457314	
Z Test statistic	** 5.0374631 with final L8		

*** significant at a 95% confidence interval*

- A Khat value ranges from 0 to 1, 0 = random assignment of classes, 1= total agreement of classes These alone do not draw many conclusions, other than both classified images contain objects with moderate agreement when compared with the accuracy polygons

- These Z-scores indicate that the final Landsat 8 classified map performed significantly better than if the classes were assigned at random

Thematic maps were also generated with the maximum likelihood algorithm, and data using all Landsat 8 bands, and from data using Landsat 8 bands mimicking those available on Landsat 7. The overall accuracy of these maps (51.86%, and 57.36%, respectively) were well below the accuracy generated by the thematic map using the Random Forest approach and Landsat 8 and derivative bands.

Overall, the thematic maps created utilizing Landsat 8 bands produced reasonable accuracies. However, the comparison of the Landsat 8 11 banded image with the Landsat 8 bands equivalent to Landsat 7 bands image, demonstrated that the additional spectral bands of the Operational Land Imager portion of the Landsat 8 satellite did not improve classification accuracies for these map classes. The Random Forest classification method produced thematic maps with higher accuracies than those using the maximum likelihood algorithm.

Congalton, R. and K. Green, 1999.

References

Breiman, Leo, 2001. Random Forests. *Machine Learning*, 45(1): 5-32.
Congalton, R. and K. Green, 1999. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, Boca Raton, FL: Lewis Publishers, 137p.
Jensen, J.R., 2005. *Introductory Digital Image Processing: A Remote Sensing Perspective*, Upper Saddle River, NJ: Pearson Prentice Hall, 508p.
Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., and J.P. Rigol-Sanchez, 2012. An Assessment of the Effectiveness of a Random Forest Classifier for Land-Cover Classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67: 93-104.
Tso, B., and P. Mather, nd. "Classification Methods for Remotely Sensed Data". Web. <http://flylib.com/books/en/3.365.1.316/1/>.

Acknowledgments

This project was supported, in part, by Grant #G14AP000002 from the Department of Interior, US Geological Survey to AmericaView. Additional support was from the College of Life Sciences and Agriculture, University of New Hampshire.