James McArdle
ESCI-442
Natasha Antonova
February 5, 2014

Unsupervised Classification of Bay to Baker Study Area For Purposes of Identifying Land Cover and Land Use

Abstract

The ISODATA unsupervised classification method was used to take the Image of the Bay to Baker Study Area and break it into 13 distinct Land/Use Land Cover types. This proccess allowed for the image to be broken into 50 distinct spectral classes based on natural groupings of spectral data. The spectral classes were then assigned to one of the 13 Land Use/Land Cover (LULC) classifications based on visual analysis of the image as well as referencing the ROI files from ground truth data. Once each of the 50 classes had been assigned, an accuracy test was performed using the ground truth data as a source for cross referencing. A total accuracy of 42.5% was achieved. All of these steps were taken in what will be called First Stage Analysis. To further better the accuracy of the classification, a Second Stage Analysis was performed. This consisted of a number of rules which were applied to the classes through the use of specific tools and models in ArcMap. The image was then once again tested for accuracy and it was found that the image classification accuracy had risen to 45.5%. Discussion of results and reasons for lack of accuracy may vary and may not be limited to a singular cause. Applications for acquisition of ground truth data may come into question as well as various methods of the classification process. The act of classification of Land Use/Land Cover through remote sensed images for satellites will have serious restrictions due to a variety of reasons, but can serve as a model and reference image.

Methods

Data Used:

For the purposes of this report multiple files of data were used. The first was an image of the Bay to Baker Study Area in Whatcom County, Washington. The image was taken in 2005 and contained 6 Thematic Mapper spectral bands (1,2,3,4,5,7). The image is a subset of the full TM scene and instead only has 2520 samples and 1510 lines. The image is projected in NAD27 UTM Zone 10. For the original purposes of the report the image was loaded with TM Bands 3,2,1, and a custom stretch was applied in order to represent visual real life color.

Figure 1.

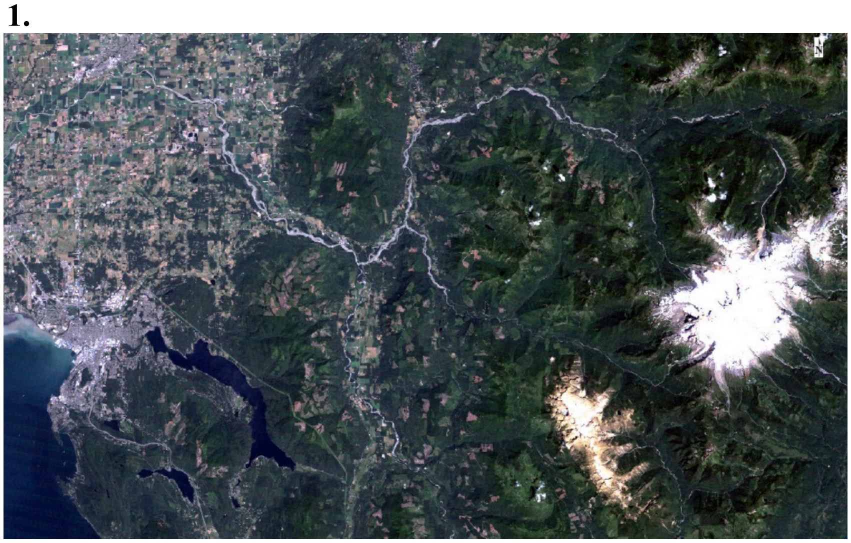

Figure 1: True-Color image of Bay to Baker Study Area

Another dataset used for the report was a set of ground truth data which had been acquired through the use of GPS units over a number of years by students in Western Washington University's ESCI-442 Class. The data was a training Region of Interest (ROI) file containing over 1000 points which had been classified as one of 18 distinct LULC types. The LULC types are as follows: Residential, Urban, Pasture, Crops, 7 Clear Cut types broken down by a range of years, Deciduous Forest, Conifer Forest, Water, Soil/Rock, Alpine, and Snow. Later a second test ROI file was used for accuracy assessment. This test file contained 846 points corresponding to one of the 13 LULC types (The previous 18 types had been reduced to 13 by merging the 7 Clear Cut values into three types: 1979-1992, 1992-2002, 2002-2005).

For the second stage of analysis, performed in Arc Map, the produced Classification image (derived from the original Bay to Baker Image) was used in conjunction with 3 other files. The first was a road buffer raster file containing 2 classifications, one of 100 meter distance from the road and the other classified by a distance larger 100 meters from the road. The second image was a raster file of the slope of region. This file had been previously derived from a Digital Elevation Model, which also happened to be the third file used in the analysis.

Analysis:

The first analysis done was to group the spectral values of each pixel in the Bay to Baker image into 50 distinct spectral classes. This was done by using an ISODATA unsupervised classification method. A tool in the program ENVI performed cluster analysis by grouping like image values until there were 50 distinct spectral classes. At this point each of the spectral classes was assigned to one of the 13 previously mentioned LULC types. This was done manually for each spectral class. Two methods were used. The first was the visual analysis performed by overlaying an individual spectral class over the true-color image to see if an obvious connection could be made. The second method used for assigning the spectral classes was the use of a table of each GPS point, the spectral class it fell in, and the LULC type assigned to it. This data can be seen in Table 1.

Table 1. Region of Interest Table

Class	Snow	Alpine	Soil	Water	Conifer	Deciduous	Clear02_05	Clear92_02	Clea73_92	Crops	Pasture	Urban	Residenti al
1			1	28		1							
2				1									
3				1	1						1		
4												1	
5					1				1		1		
6					5								
7				1	6								
8			1		10				1				
9			1	1	12	1			1				1
10			1		9	1			9				
11					4	1		1	17		1		
12			1		9	4		1	1			1	
13					11	5		2	4				
14					6	2		1	5		1		
15					6	4		2	6			1	
16				1	3	3		5	12	2	2		
17					1	7		5	4	3	2		
18					1	5		8	9				
19			1	1	1	2				1	1	1	1
20					4		1				1		1
21					1								
22					1			8		1	5		
23			1	1		1		8	3	2	5		
24		1						5		12	7		
25			1				5			1	2		
26			1									26	7
27													
28	6												
29				2								2	1
30			5		1					2		4	1
31						1		1	1		1	1	3
32				1	2			1			4		7
33			1			1		3	2	2	2		
34		1						1			4		
35					1			3	1	2	2	1	
36						1	2	3		7	7		
37						1		5		5	2		
38						1		1			10		
39	5												
40												3	
41													
42	1		5								1	2	
43					1						1	1	
44			2			2		1			3	4	16
45					2	1				1	3	1	13
46		4						1		3	1		
47			1			1	6			4	3	1	
48										1	3	1	
49					1	2				3	9		2
50			2				1			2	1		
Class	Snow	Alpine	Soil	Water	Conifer	Deciduous	Clear02_05	Clear92_02	Clea73_92	Crops	Pasture	Urban	Residenti al

Table 1: All 50 spectral classes and the number of DN points (from Ground Truth train data) in each information class

The table was used as a reference in collaboration with visual analysis to determine which spectral classes should be assigned into each LULC type. The assigned spectral classes can be seen in Table 2. After each spectral class had been assigned a cover type and the Combined Class Image had been produced, an accuracy assessment test was performed. To achieve this, the 13 assigned information classes were tested against the ground truth test data (ROI file) and a confusion matrix was generated. The results of the confusion matrix were then placed into a sample error matrix. The accuracy for the classification was then performed by calculating the Kappa Statistic in an error matrix in order to produce values for both the producer's accuracy and the user's accuracy, as well as an overall accuracy.

Table 2. Assigned Spectral Classes

Spectral Class	Primary LULC Classification	Secondary LULC Classification
1	Water	Shadow
2	Water	
3	Conifer	Pasture
4	Conifer	
5	Conifer	Clear73_92
6	Conifer	
7	Conifer	
8	Conifer	
9	Conifer	
10	Conifer	Clear73_92
11	Clear73_92	
12	Conifer	
13	Conifer	Deciduous
14	Clear73_92	Conifer
15	Clear73_92	Conifer
16	Clear73_92	Conifer
17	Deciduous	Clear92_02
18	Clear73_92	Clear92_02
19	Deciduous	
20	Conifer	
21	Conifer	
22	Clear92_02	
23	Clear92_02	
24	Crops	
25	Clear02_05	
26	Alpine	Urban
27	Snow	
28	Snow	
29	Water	
30	Soil	
31	Residential	
32	Pasture	
33	Clear92_02	
34	Pasture	
35	Clear92_02	
36	Crops	Pasture
37	Clear92_02	Crops
38	Pasture	
39	Snow	
40	Urban	
41	Snow	Alpine
42	Soil	
43	Pasture	Urban
44	Residential	
45	Residential	
46	Crops	
47	Clear02_05	
48	Pasture	
49	Pasture	
50	Soil	Crops

Table 2: Each Spectral Class and the LULC Type in which it was assigned.

After this process was complete, the Second Stage Analysis was performed in ArcMap as an attempt to improve the accuracy. A set of conditional models were run on the classified image in order to reclassify some LULC types that had been perhaps incorrectly classified. Models 1-3 were used with input from the road buffer raster file, Model 4 was used with the slope file as an input, and Models 5-10 used the DEM as an input. The following models were run:

1. Residential Class was reclassified to Rock if class was more than 100 meters from a road.
2. Urban Class was reclassified to Rock if class was more than 100 meters from a road.
3. Rock Class was reclassified to Urban if class was within 100 meters of a road.
4. Water Class was reclassified to Shadow (class was added) if the slope of the Class was not 0 degrees
5. Pasture Class was reclassified to 92-02 Clear Cut (CC) if class elevation was between 200 meters and 1500 meters and was reclassified to Alpine if class elevation was above 1500 meters.
6. Crops Class was reclassified to 92-02 CC if class elevation was between 200 meters and 1500 meters and was reclassified to Alpine if class elevation was above 1500 meters.
7. 73-92 CC Class was reclassified to Crops if class elevation was below 200 meters and was reclassified to Alpine if class elevation was above 1500 meters.
8. 92-02 CC Class was reclassified to Crops if class elevation was below 200 meters and was reclassified to Alpine if class elevation was above 1500 meters.
9. 02-05 CC Class was reclassified to Crops if class elevation was below 200 meters and was reclassified to Alpine if class elevation was above 1500 meters.
10. Alpine Class was reclassified to Crops if class elevation was below 200 meters and was reclassified to 92-05 if class elevation was above 200 meters and below 1500 meters.

Once the models were completed and the classification image had been updated, the new classified image was then run through another accuracy assessment. The same process previously described (using ground truth data, generating a confusion matrix, and calculating a Kappa Statistic in an error matrix) was then done to achieve the desired results. All methods were taken from: (Antonova and Wallin, 2014)

Results

In the first stage (analysis performed in ENVI), the classification performed produced and image with an overall accuracy of 42.5%. When examining the image, based on visual analysis alone, it is easy to see where some of the errors come into play. Specifically around Mount Baker, where there are large areas of Urban and Water, when in actuality there is no real world equivalent. Likewise much of the area surrounding downtown Bellingham has been incorrectly classified as Alpine. The results of the first stage classification can be seen in Figure 2 and Table 3.

Table 3. Stats for All Classes

LULC Classification	# of DN Points	Percent of Total Coverage	Total Area (Hectares)
Residential	131,838	3.465	8,239.8750
Urban	6,383	0.180	427.3750
Pasture	202,199	5.314	12,637.4375
Crops	113,976	2.995	7,123.5000
Clear Cut 73-92	856,831	22.517	53,551.9375
Clear Cut 92-02	284,040	7.465	17,752.5000
Clear Cur 02-05	67,488	1.774	4,218.0000
Deciduous	207,887	5.466	12,999.0625
Conifer	1,520,074	39.947	95,004.6250
Water	207,887	5.463	12,992.9375
Soil	67,117	1.764	4,194.8125
Alpine	48,249	1.268	3,015.5625
Snow	90,678	2.383	5,667.3750

Table 3: First Stage Classification Results of total points, percentage of coverage, and total area

Figure 2.

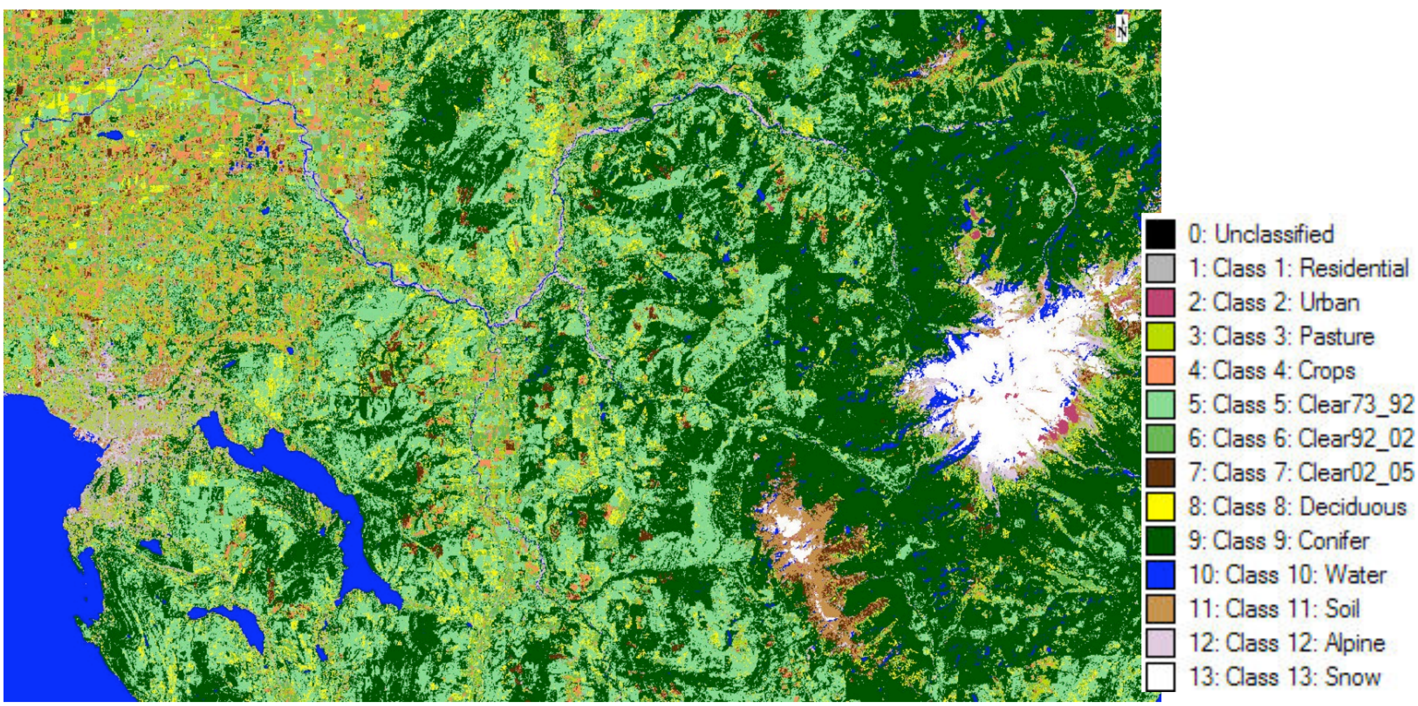

Figure 2: The First Stage Classification Resulting Image with 13 classified types of Land Use / Land Cover.

When further breaking down the accuracy, more errors become clear. The producer accuracy, or the error of exclusion, resulted in the Alpine Class (0%) and the Urban Class (2.7%) with the lowest totals and the 2002-2005 Clear Cut Class (92.6%) with the highest accuracy. The user accuracy, or error of inclusion, demonstrated that the Alpine Class was once again the least accurate (0%) while the most accurate shifted to the Snow Class (70.6%). The results of the first stage accuracy assessment can be seen in Table 4.

Table 4. Confusion Matrix Results

Predicted	**Ground Truth (Reference Data)**														Row		User'
Class	Residenti al	Urban	Pasture	Crops	73-92 Clearcut	92-02 Clearcut	02-05 Clearcut	Deciduous	Conifer	Water	Rock	Alpine	Snow		Total		Acc ('
Residenti al	35	7	8	9	0	1	0	5	6	1	4	0	0		76		
Urban	0	2	0	0	0	0	0	0	0	0	0	0	1		3		
Pasture	19	4	25	21	0	1	0	2	3	0	3	2	0		80		
Crops	0	0	16	19	0	1	1	0	0	1	0	1	0		39		
73-92 Clearcut	2	1	4	1	55	17	0	30	22	4	3	0	0		139		
92-02 Clearcut	1	1	17	19	6	14	0	4	3	0	2	1	0		68		
02-05 Clearcut	1	1	5	4	1	1	25	0	0	0	1	0	0		39		
Deciduous	3	0	6	1	6	10	0	7	3	1	4	0	0		41		
Conifer	4	2	3	0	20	3	0	28	93	7	3	0	0		163		
Water	0	1	0	0	0	0	0	0	1	51	3	0	0		56		
Rock	1	14	10	5	0	0	1	1	0	1	10	2	2		47		
Alpine	6	37	1	0	0	0	0	2	0	1	4	0	0		51		
Snow	0	4	1	0	0	0	0	0	0	0	0	0	12		17		
Column Total	72	74	96	79	88	48	27	79	131	67	37	6	15		819	Total Sampl	
															348	Total # Corr	
Prod Acc (%)	48.6	2.7	26.0	24.1	62.5	29.2	92.6	8.9	71.0	76.1	27.0	0.0	80.0		**42.5**	**Overall Acc**	

Table 4: Error Matrix showing the Predicted Classes, Ground Truth Data and the corresponding accuracy for both producer and user for the First Stage Analysis.

After running the image through the Second Stage Analysis in order to increase accuracy, a plain visual examination of the Figure 3 demonstrates an apparent improvement in classification, specifically in the areas of higher elevation. Also, after examination of the percent totals of Table 3 and the percent totals of Table 5, it is apparent that one of the biggest changes to occur took place in the Crop Class.

Table 5. Stats for All Classes

LULC Classification	# of DN Points	Percent of Total Coverage	Total Area (Hectares)
Residential	70,141	1.843	4,383.8125
Urban	15,310	0.402	956.8750
Pasture	169,974	4.467	10,623.3750
Crops	547,462	14.387	34,216.3750
Clear Cut 73-92	638,843	16.789	39,927.6875
Clear Cut 92-02	121,106	3.183	7,569.1250
Clear Cur 02-05	34,038	0.895	2,127.3750
Deciduous	207,985	5.466	12,999.0625
Conifer	1,520,074	39.947	95,004.6250
Water	164,672	4.328	10,292.0000
Soil	120,342	3.163	7,521.3750
Alpine	61,360	1.613	3,835.0000
Snow	90,678	2.383	5,667.3750
Shadow	43,215	1.136	2,700.9375

Table 5:Second Stage Classification Results of total points, percentage of coverage, and total area

Figure 3.

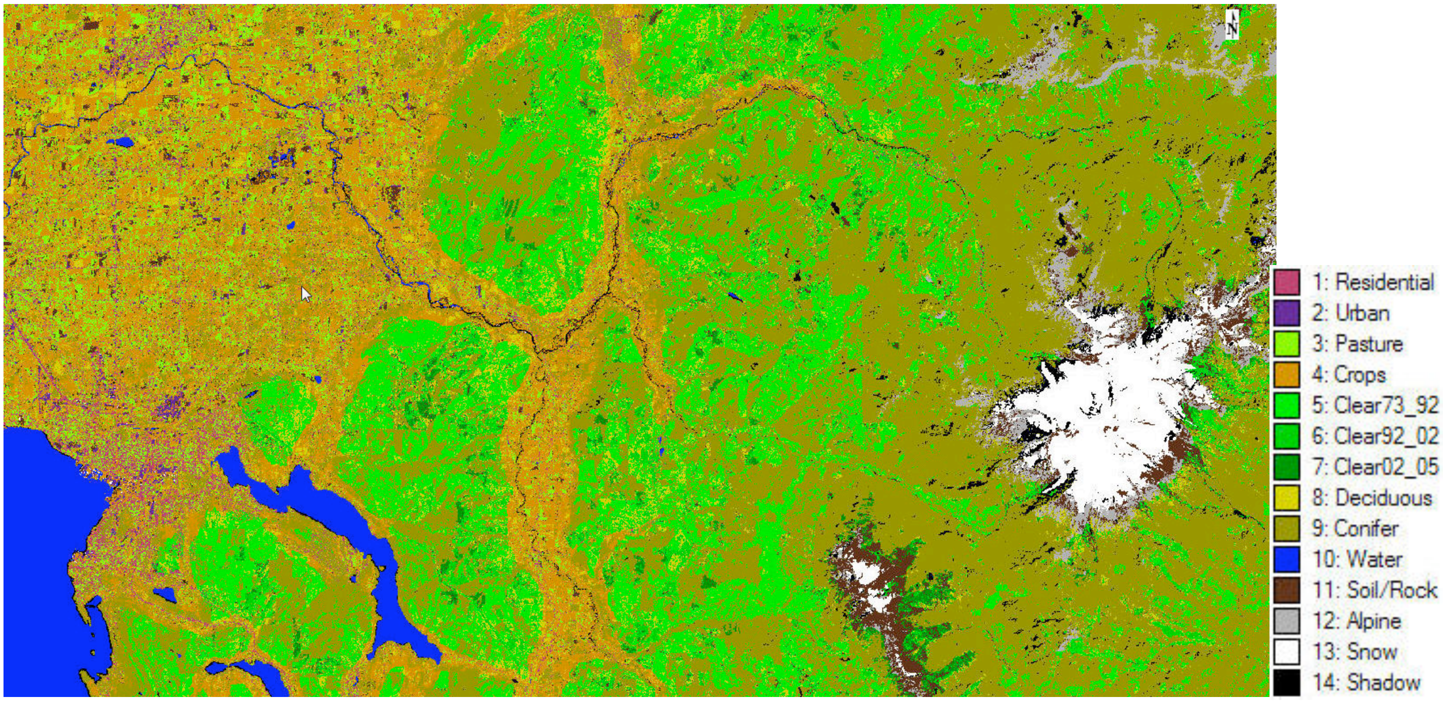

Figure 3:The Second Stage Classification Resulting Image with 13 classified types of Land Use / Land Cover.

While there are significant areas of improvement, some other problem areas become more glaring. Specifically the area around downtown Bellingham on the coast show a significant area of Crops when in reality, much of the area is either Urban or Residential. Examination of the accuracy assessment shows just that. While the overall accuracy does show an improvement of 3% up to 45.5%, a comparison of Table 4 and Table 6 shows a substantial drop in the user accuracy in the Crops Class (from 48.7% to 20.6%). This likely to be a result of running model #10; assigning all low elevation Alpine Classes as Crops, when in reality much of it should have been reclassified as Urban.

Table 6. Confusion Matrix Results

Predicted	Ground Truth (Reference Data)													Row		User's
Class	Residential	Urban	Pasture	Crops	73-92 Clearcut	92-02 Clearcut	02-05 Clearcut	Deciduous	Conifer	Water	Rock	Alpine	Snow	Total		Acc (%)
Residential	33	7	6	8	0	0	0	5	4	1	4	0	0	68		48
Urban	1	15	6	5	0	0	0	1	0	0	5	0	0	33		45
Pasture	19	4	25	21	0	0	0	2	3	0	3	1	0	78		32
Crops	10	40	42	43	5	6	0	33	18	5	7	0	0	209		20
73-92 Clearcut	0	0	0	0	50	15	0	3	7	1	3	0	0	79		63
92-02 Clearcut	0	0	1	0	6	13	1	0	0	0	0	0	0	21		61
02-05 Clearcut	0	0	0	0	1	0	25	0	0	0	0	0	0	26		96
Deciduous	3	0	6	1	6	10	0	7	3	1	4	0	0	41		17
Conifer	4	2	3	0	20	3	0	28	93	7	3	0	0	163		57
Water	0	0	0	0	0	0	0	0	0	43	1	0	0	44		97
Rock	2	1	6	1	0	1	1	0	2	1	5	2	3	25		20
Alpine	0	0	0	0	0	0	0	0	0	0	0	3	0	3		100
Snow	0	4	1	0	0	0	0	0	0	0	0	0	12	17		70
Column Total	72	73	96	79	88	48	27	79	130	59	35	6	15	807	Total Samples	
														367	Total # Correct	
Prod Acc (%)	45.8	20.5	26.0	54.4	56.8	27.1	92.6	8.9	71.5	72.9	14.3	50.0	80.0	**45.5**	**Overall Accurac**	

Table 6: Error Matrix showing the Predicted Classes, Ground Truth Data and the corresponding accuracy for both producer and user for the Second Stage Analysis

Discussion

The process of an unsupervised classification of a remote sensed image will always have its challenges. Beginning with the meaning of the terms unsupervised and remote sensed automatically lead to the problem of classifying a land coverage which has not been evaluated first hand. Given that after two distinct stages of analysis attempting to produce an accurate classification, the end result was still an accuracy less than 50%, the methods have much room for improvement. In examination of the Second Stage Error Matrix, it is clear that one of the largest misclassifications occurred in the labeling of 40 ground truth Urban data points as Crops in the image. In seeing that the overall accuracy increased despite the large decrease in user accuracy of Crops, it is apparent that if this one misclassification were rectified, the overall accuracy would see a large uptick. Given that the Crops user accuracy was better before the Second Stage Analysis, it is clear that the assigning of all Alpine Classes below 200 meters was a mistake. The problem, however, is that if the model is changed to reclassify all Alpine Classes below 200 meters as Urban, then this will most likely severely reduce the producer's accuracy levels. Therefore the best solution to rectify this problem would be to change a classification in the First Stage Analysis. When examining Table 1: Region of Interest, the 26th spectral class shows that it had 26 DN points that had been attributed as Urban by the train ground truth data. However, due to improper visual analysis, and persuasion at the hands of a lab tutorial, the 26th spectral class was assigned to the Alpine information class. If the same process was redone, only this time assigning spectral class 26 as Urban, the results would be more accurate in the first stage analysis. This would leave the image with the inverse problem that did occur in the process performed. Instead of having areas of downtown Bellingham classified as Alpine, much of the area surrounding Mount Baker would now be classified as Urban. This could, however, be more easily corrected during the Second Stage Analysis. Instead of model #10 reclassifying all Alpine below 200 meters as Crops (which incorrectly classified much of Urban Bellingham as Crops) the model could reclassify any Urban Class above 1500 meters as Alpine. This approach would likely produce a more accurate image classification. To further improve accuracy the same approach could be taken by examining any large discrepancy in the error matrix (such as the conflict in classification of Conifer and 73-92 CC or Deciduous and Conifer) and investigating where the error occurred.

Additionally, other reasons for lack of accuracy could be due to Ground Truth data misclassification. Potentially, some of the points classified and recorded by the GPS units may be inaccurate due to a variety of reasons. One or more of the areas sampled could have been to small an area to show up as a pixel when the spatial resolution is 30 meters by 30 meters. Also the Northwest is notoriously difficult place to get accurate GPS readings due cloud cover, steep terrain and thick canopy cover, all of which could lead to the multi-pathing of the GPS signal. Likewise the atmosphere could have affected the reflectance values for the pixel in the original Bay to Baker image.

The methods used above for the classification were used as a result of trying to classify a large swath of land. While unsupervised classification can be inaccurate at times due to lack of firsthand knowledge, it is an efficient and effective way to classify data in large portions. In this particular case, these methods were used not only because of the size of the area being classified, but also because of the remoteness of much of the land. This becomes increasingly apparent when evaluating the ground truth data, as many of the point are localized around Bellingham and the Western portion of the image.

With slight modifications and perhaps greater attention to detail both in the collection of the ground truth data and the assigning of spectral classes to information classes, the process could be much more effective and thus useful for repetition in both later dates of the same area and other areas. In addition, with further investigation of the error matrix, more models could be applied to increase accuracy. It seems obvious that a stream buffer layer could be useful. Also an aspect rasterset could prove to be useful if trying to limit reclassify a shadows area. Overall the approach proved to be a good starting point in LULC classification, and even with a less than 50% accuracy, the results could easily be used for general reference.

Literature Cited

Antonova, N. & Wallin, D. 2014 *Lab 4:Unsupervised Classification with ENVI.* ESCI 442/542: Introduction into Remote Sensing
< http://staff.wwu.edu/antonon/envr442/ENVI/442_unsup_class_ENVI.html>