NASA Graduate Student Summer Program 2010

Satellite-Based Estimation of Impervious/Pervious Land Cover Percentage of the Chesapeake Bay Watershed

SARAH SMITH

Eastern Michigan University, Ypsilanti, Michigan

Department of Geography and Geology

In conjunction with:

GILMAN OUELLETTE

The Pennsylvania State University, University Park, Pennsylvania

Department of Geography

Mentor: Dr. Eric Brown de Colstoun

NASA Goddard Space Flight Center

ABSTRACT

Scientists are developing new satellite-based maps of various areas all around the world. Previously the attached price tag for this process was high, until recently when satellite imagery became freely available. Now satellite-based maps are the norm, and scientists are working hard at creating new maps for the Baltimore/Washington D.C. region. The focus of this project is on the Chesapeake Bay Watershed (CBW) region and utilizing Landsat data to determine the percentage of pervious/impervious cover of the area utilizing ENVI software. The Landsat data will be 'stacked' and cloud-free; recent images (2009 to present) were mosaicked together and analyzed using training sites created from Quickbird images provided by the National Geospatial Intelligence Agency (NGA). The Quickbird images will also be recent images to coincide with the Landsat images. Quickbird images were subsetted randomly, classified as 0 – pervious or 1 - impervious, aggregated, and then applied to the Landsat images via the cubist regression tree algorithm already developed. This process will involve the utilization of LEADAPS software for atmospheric correction as well. The overall percentage of pervious/impervious cover of the Landsat images should be relational to that of the Quickbird training sites.

INTRODUCTION

New satellite-based mapping is becoming increasingly more popular now that imagery data, particularly Landsat data, is now free. Prior to free imagery, costs were a large limiting factor when it came to satellite-based mapping and along with the size of the area that could be mapped. Commercial satellite imagery is now also available; some of these commercial satellites have a much more frequent temporal resolution, which allows for more observations over time. With these advances national and global wide mapping efforts are taking place, such as that done by Masek, J.G., et al. (2006a) mapping North American surface reflectance values over time with Landsat; studies such as these can allow researchers to map and monitor land cover such as tree canopy or crop coverage of specific areas based on surface reflectance values and signatures. Being able to map land cover on the global scale leads to a greater understanding of our ever changing planet and the impact humans have.

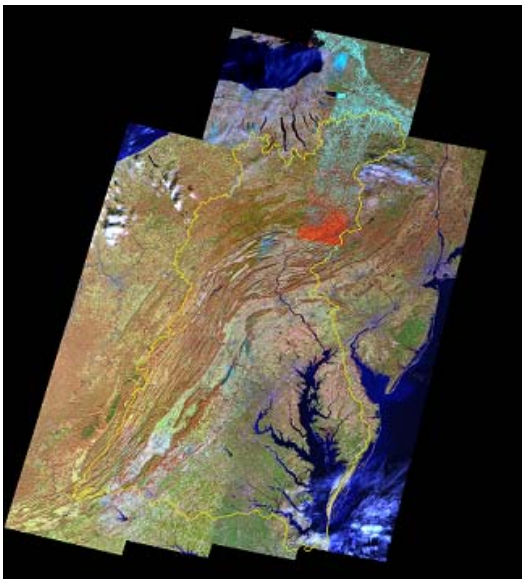


Figure 1: Leaf off Landsat 5 TM mosaic of the Chesapeake Bay Watershed. Images acquired from winter 2009 – 2010.

The CBW (Figure 1) is one such large area of interest when it comes to new satellite-based mapping; the watershed is the largest estuary in the United States and holds about 80 giga-tons of water (Hall 2010). It is a government initiative to better monitor and understand the intricate web the CBW supports upon realization of the large amounts of pollutants present. Much of the change in the CBW came about during the 18th and 19th centuries when a conversion to agriculture took place followed by expansive amount of commercial and residential development on what was once a very homogenous area of forest (Jantz, P., S. Goetz, and C. Jantz. 2005). Impervious land cover, such as urban and residential areas, is an important focus for the region; the more impervious cover, the higher the possibility of flooding, pollution from runoff, and greater amount of runoff.

Authors such as Jantz, Goetz, and C. Jantz (2005) have focused studies on the CBW as well, estimating the extent and rates of urbanization and the impacts this land use change development has on the area. Along with a study by Masek, J.G., et al. (2008), this focused on

forest disturbance in North America utilizing decadal Landsat data, for which the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) was developed. It was to provide a robust way for processing large quantities of remote sensing data in order to monitor forest change analyses (Masek, J.G., et al. 2008). Studies such as these utilized Multi-Resolution Landcover Consortium (MRLC) land cover map, Landsat 7 Enhanced Thematic Mapper (ETM+) images, and GeoCover images derived from Landsat data; the focus on this paper is to take this a step higher and utilize high resolution imagery as training sites to classify Landsat imagery which, in theory, will allow for a more accurate classification of the Landsat imagery.

Studies such as those by Huang, C. et al. (2001) have utilized Landsat 7 ETM+ high resolution data to monitor tree canopy density over large areas although the high resolution data utilized were digital orthoquads (DOQ) instead of high resolution satellite imagery. Unlike a satellite data, DOQs are computer generated images that are created by utilizing aerial photography, usually taken from an airplane. A DOQ is just a picture map where as a satellite image provides you with bands which in turn provide insight to the land cover types based on reflectance values. A method utilizing high resolution imagery focuses (focuses) on the percentage per pixel (ppp) of impervious cover over the CBW could lead to annual monitoring of land cover types versus those over a span of 10+ years, which in turn could lead to better knowledge of change throughout the area.

DATA AND METHODOLOGY

Study Area

The CBW is the largest estuary in the United States, stretching across 6 states, totaling more than 64,000 square miles, and provides a home for a vast number of species, both plant and animal (CBP 2008). Recently, an initiative was put forth by the CBP to conserve and preserve the CBW.

Data

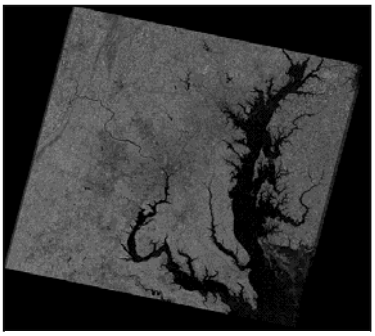


Figure 2: Band 4 Landsat 5 TM image acquired on April 26th, 2010.

Landsat-5 Thematic Mapper (TM) images (Figure 2) were the primary data source for determining the percentage of pervious/impervious cover for the CBW. Landsat-5 TM was chosen over Landsat-7 ETM+ due to the failure of the scan line corrector on the Landsat-7 satellite in 2003. This failure leads to gaps within the Landsat-7 imagery that need to be corrected before image analyses can take place. Therefore when wanting to make inferences of current land cover or change detection, usage of Landsat-5 saves time and is less prone to error that lies within the correction of Landsat-7 images. Spatial resolution of Landsat-5 TM is 30 meters, which is sufficient enough when looking at a large area such as the CBW, and the data is quantized to 8-bits. Landsat-5 TM data was acquired from the United States Geological Survey free of charge from GLOVIS.

Figure 3: Sample Quickbird scene acquired April 11th, 2006.

Quickbird data (Figure 3) was utilized as the training data to

be applied to the Landsat-5 TM data. Quickbird is a commercially owned satellite with very high spatial resolution, four 2.44 X 2.44 m multispectral bands and a 0.62 X 0.62 panchromatic band (Jensen, 2005). The four multispectral bands for Quickbird are blue, green, red, and near-infrared (nir). Interestingly, the Quickbird satellite is a non-sun-synchronous satellite, and the data is quantized to 11-bits. Quickbird data was acquired from the National Geospatial Intelligence Agency (NGA).

Methodology

The methodology for this study was largely based off the methods utilized by Yang (2003). Once the Quickbird imagery was preprocessed, methodology took place in three major steps: (i) Image preprocessing (ii) Training Data Creation and Data Processing (ii1) Regression Tree Modeling and Spatial Application (Appendix A).

Image Preprocessing

Landsat TM was preprocessed for atmospheric correction utilizing the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS). The Landsat TM images are atmospherically corrected to surface reflectance to help researchers track land-cover changes more easily using the GeoCover data (Masek, J.G. et. al. 2006a). LEDAPS is in-house software developed at the National Aeronautics and Space Administration (NASA) Goddard Space Flight Center (GSFC).

Quickbird preprocessing consisted of creating 20 small 2000 x 2000 pixel subsets of the 10 original images obtained from the NGA. This was done to speed up the preprocessing and processing of the training data. Once subsetted, the multispectral and panchromatic scenes were intertwined using Graham-Schmit pan-sharpening. This created one image of four bands (blue, green, red, and nir) overlain a “sharper” panchromatic image; this provides more detail when classifying an image. Lastly, the Quickbird data was re-projected to match the projection of the Landsat TM data.

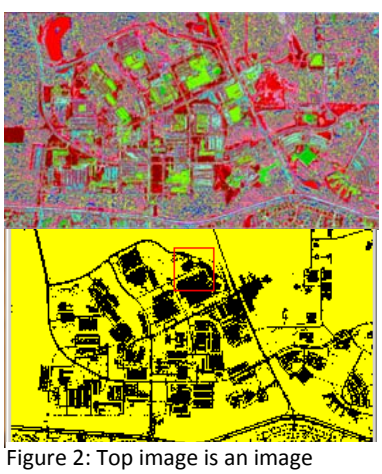


Figure 2: Top image is an image classified to 20 classes. Bottom image is of 2 classes; area is GSFC, Greenbelt, MD.

Training Data Creation and Data Processing

Quickbird training data was created by utilizing Iterative Self-Organizing Data Analysis Technique (ISODATA), a method of unsupervised classification, which is a modification of the k-means clustering algorithm. The difference here is that instead of just making two passes through the dataset, ISODATA makes a large number of passes until specified results are obtained such as 10 iterations and 20 classes in this instance. Default values were assumed for the rest of the criteria such as maximum standard deviation, split separation value, etc. Using the overlay tool**,** the 20 classes were then identified and eventually broken down into two classes; pervious and impervious (Figure 4). Once the training data is broken down into pervious and impervious, the data needs to be changed from integer to float in order to depict a percentage value in the data

field. This is done by choosing band math and applying the equation FLOAT(B1) to each image. After the band math is applied, the training data then needs to be aggregated to 30 meters by 30 meters to overlay properly with the Landsat TM images. This is done by resizing the data, utilizing pixel aggregate. The resulting image will look fuzzy but it is to be expected. Once this is done, training data are then rectified using image to image analyses and finally layer stacked with the Landsat TM images (Figure 5).

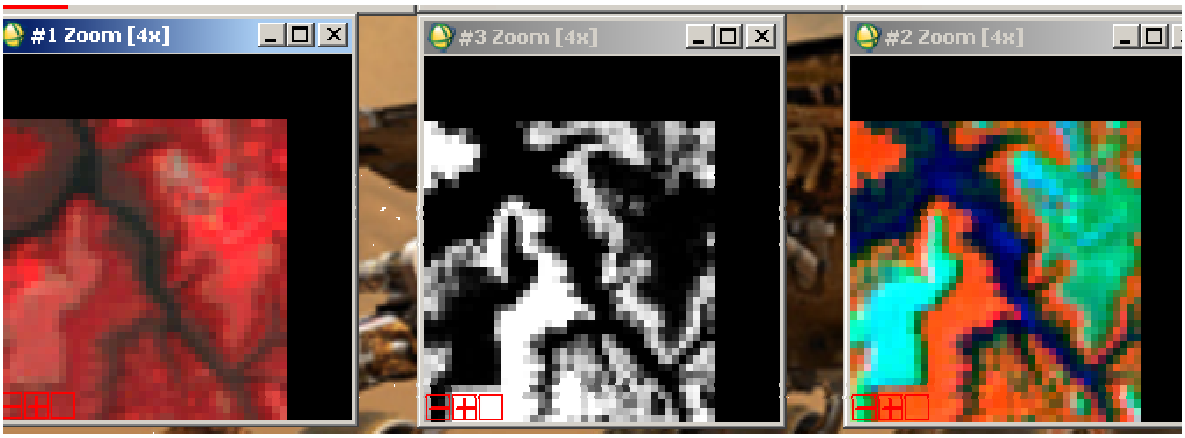


Figure 3: Image progression from Landsat 5 TM scene and aggregated Quickbird data to the layer stacked image.

Regression Tree Modeling and Spatial Application

Our training data is a continuous data set between 0 and 1 and classified as percentage per pixel. A regression tree algorithm is then applied to the layer stacked image which is composed of the Landsat TM bands 1-5, 7, and Quickbird training data band. The regression tree algorithm produces rule-based models for prediction of continuous variables and can account for a nonlinear relationship between predictive and target variables (Yang 2003). Much like Yang 2003, a regression tree algorithm from Cubist, commercial software available free from http://rulequest.com. Cubist utilizes the average error and relative error along with other statistical measures to evaluate model performance (Yang 2003); input data for Cubist must be in simple text format and there must be a .data and a .names file present for the regression tree to run (Appendix B). The regression tree algorithm is used to develop an initial model to apply to a larger area.

The initial model from the regression tree model will then be applied to the CBW as a whole, which consists of 18 Landsat TM images. The model is incorporated into the ENVI software and used as an unsupervised classification of the larger Landsat Images. The final classified Landsat images are then mosaicked together to provide an estimate of the total percentage of impervious cover.

RESULTS

Simplifying things, the training data matched up well with the Landsat imagery depicting areas of impervious and pervious coverage; showing properly processed, high resolution satellite imagery can provide a decent training site for lower resolution images. Our models also depicted; although as with all satellite-based mapping, there will be some underlying error. The average error ranges were from roughly 5% to 12%, and the correlation coefficients percent

range, signifying a low average error between actual and predicted variables. Relative error also stays below one for all tests run; a relative error above one would indicate that the model created would not be very useful (Figure 6). It was also noted that the more training sites incorporated into the regression tree, the worse the output of average error and relative error became, yet these changes were minimal. The factor that can account for this will be discussed in the sources of error section; Yang (2003) also had a similar result of minimal changes which suggests the use of fewer input variables.

```
Rule 40: [35 cases, mean 0.80553, range 0.4431 to 0.9998, est err 0.12217]

  if
      Band1 > 87
      Band4 > 58
      Band4 <= 68
      Band5 <= 78
  then
      Band Agg = 1.66756 + 0.04383 Band5 - 0.04631 Band4 - 0.055 Band7
                 + 0.0194 Band2

Evaluation on training data (4791 cases):

    Average  |error|              0.06309
    Relative |error|                 0.39
    Correlation coefficient         0.83


        Attribute usage:
          Conds  Model

            72%    42%    Band2
            51%    59%    Band7
            42%    97%    Band5
            42%    85%    Band1
            21%    96%    Band4
             7%    33%    Band3

Evaluation on test data (252 cases):

    Average  |error|              0.06722
    Relative |error|                 0.42
    Correlation coefficient         0.81


Time: 0.3 secs
```

Figure 4: Sample results from modified regression tree analysis of the data.

When applied to a subset of Landsat TM imagery, results were fairly surprising with a correlation coefficient of .97 (Appendix B); when cross-validated though, the correlation coefficient dropped to as low as .93 depending on the amount of cases used. These findings are still acceptable in that our actual value and predicted values agree more than 90% of the time.

It should be noted that the Landsat TM subset estimated was from row 15 path 34. This was chosen because much of the training site data was located within this image; the full Landsat image was unable to be used due to the large size of the data. Based on the knowledge of these outputs though, one could assimilate that the errors and correlation coefficient would degrade slightly when applied to the full CBW but still within a range of acceptability that would allow us to conclude that our estimate of impervious coverage would be accurate.

Sources of Error

When utilizing satellites for mapping, one must take into account the various sources of error that can be produced. First and foremost, one large source of error lies within rectification; if two images are not properly rectified, layer stacking will become early impossible for aligning the images properly. When images are obtained, it is pertinent that the research must take note of the projection of the images that will be used. Utilizing images that are in the same projection could easily save time and effort as well as not add another source of error. Rectification is done in order to make data planimetric, which removes error found within most commercially available remote sensor data (Jensen 2005). Ground control points are commonly used when rectifying; although at times, images need to be rectified on a pixel to pixel basis, which, if not done properly, could produce more error.

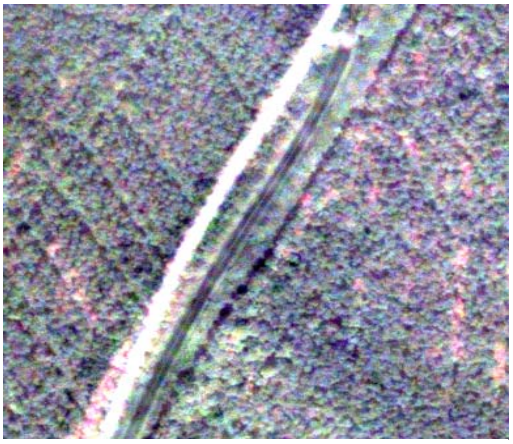


Figure 5: Example of artifact error in Quickbird image; artifact is white stripe seemingly running parallel to the road.

Another source of error lies within the pansharpen Quickbird training sites; when running classifications on the pansharpen imagery, it was noted that on occasion the classifications would not match up to the features within the image. It was as if the classification was reading just the multispectral image or the panchromatic image, instead of the pansharpen image which lead to artifacts within the image such as, a classified road nowhere to be found (Figure 7). There were also spots on the images where the processes seemed to not work, albeit small, these spots were still mis-classified (Figure 8). With the Quickbird imagery, subsetting was also preformed in a specific order with the panchromatic image subsetted first and then the multispectral imagery subsetted based on the panchromatic. This was done because when wanting a specific sized training site, 2000 pixels by 2000 pixels for example, if the multispectral imagery was subsetted first, the size of the training site was off when subsetting the panchromatic image.

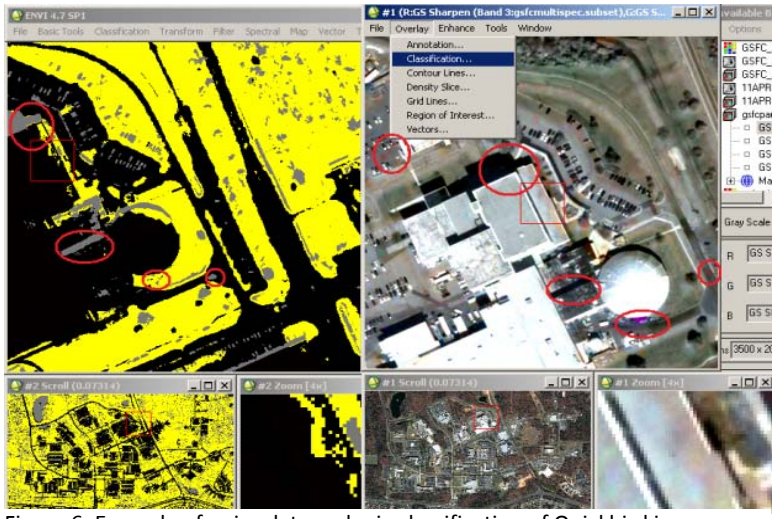


Figure 6: Example of noise data and mis-classification of Quickbird imagery.

Lastly, error was also noted with the Landsat 5 imagery in that there was noise along the end of the images that disrupted the classification. There was also a large gap between the various bands within the imagery which also led to some mis-classifications. This was easily dealt with through by just clipping off the edges of the imagery; the scenes overlap enough where this is possible. As mentioned previously, results suggested that for regression tree models, the use of fewer variables was suggested at from the results. This error could arise from having two people classify the training sites separately without strict guidelines of the process that both would like to take. Slight differences in the classification methods between two people can also account for sources of error when combining the training site data. Imagery analyses are also somewhat subjective; what one person may classify as pervious, another may classify as impervious. Take a well traveled dirt road or manmade ponds, for example. Other minor sources of error could possibly arise from imagery that was not properly processed, forgetting to assign wavelengths to images that need pansharpening, not utilizing enough classes, or shadows overlapping pervious/imperious cover.

CONCLUSIONS

Similar to Yang 2003 we were able to develop a method of mapping impervious surfaces as continuous variables, ppp, using both low and high resolution data. Our results also show that software is capable of developing a method to accuracy and consistently map land cover features, in our case impervious percentage of the CBW. A future endeavor of this study is field validation work, which will be done at a later date, to truly estimate the accuracy of the method. One field validation work took place locally just to be sure that equipment worked. The data for the local area was accurate although only one pixel was completed due to time constraints; therefore, no conclusion can come from that. Overall, the final product will yield numerous

applications for the CBW, such as comparative studies of the hydrology of the land in comparison to elevation and impervious coverage. This method could also be applied to the land cover types such as the seasonal change of vegetation within the area and tree canopy cover. Bird migration patterns could also be applied, and therefore, one could make inferences about where water fowl and other migratory birds could possibly inhabit.

Numerous applications can also arise from land cover percentage analyses such as this, not only on a local scale but a global scale as well. Studies by Masek, J.G. et al. (2006a) indicate that free Landsat data classifications can also be applied to the United States and incorporated with MODIS data, although utilizing high resolution data as training sites for areas as large as these would be somewhat time consuming. The results via cubist suggest though that training sites from high resolution data can be utilized to classify lower resolution images such as Landsat TM, though the accuracy depends on the amount of care taken when classifying the training image. So hypothetically with well dispersed training sites of high resolution data throughout the United States, anything mapping wise at least, is possible. With a method such as this, annual mapping could take place, no matter what the area of focus, since much of the process is automated by ENVI and Cubist software packages. ENVI also provides an extension, ENVI EX, which would allow a user to create a classified image by rules; these rules can then be saved and applied to subsequent imagery as it is gathered. The only downfall is that the imagery would have to be of a specific area, such as the CBW, or the model may not have similar outputs for a different region. For example, if utilizing a rule classification for the United States, the soils and ground in Utah will differ spectrally from those in the more saturated areas such as Florida.

REFERENCES

CBP. 2008. Chesapeake 2008. Available at http://www.chesapeakebay.net Accessed July 2010.

Hall, Forrest "Remote Sensing of the Land Surface Carbon, Water and Energy Cycles: What do we know? What's missing? Where are we headed?" Goddard Space Flight Center, Exploring Earth Seminar Series. Greenbelt, MD. 26th July, 2010

Huang, C et al. (2001) At-Satellite Reflectance: A first order normalization of Landsat 7 ETM+ images. Retrieved from: http://landcover.usgs.gov/pdf/huang2.pdf

Jantz, P., S. Goetz, and C. Jantz. 2005. Urbanization and the loss of resource lands in the chesapeake bay watershed. *Environmental Management.* 36(6):808-825

Jensen, John R. (2005) Introductory Digital Image Processing: A remote sensing perspective. Upper Saddle River, NJ: Pearson Prentice Hall

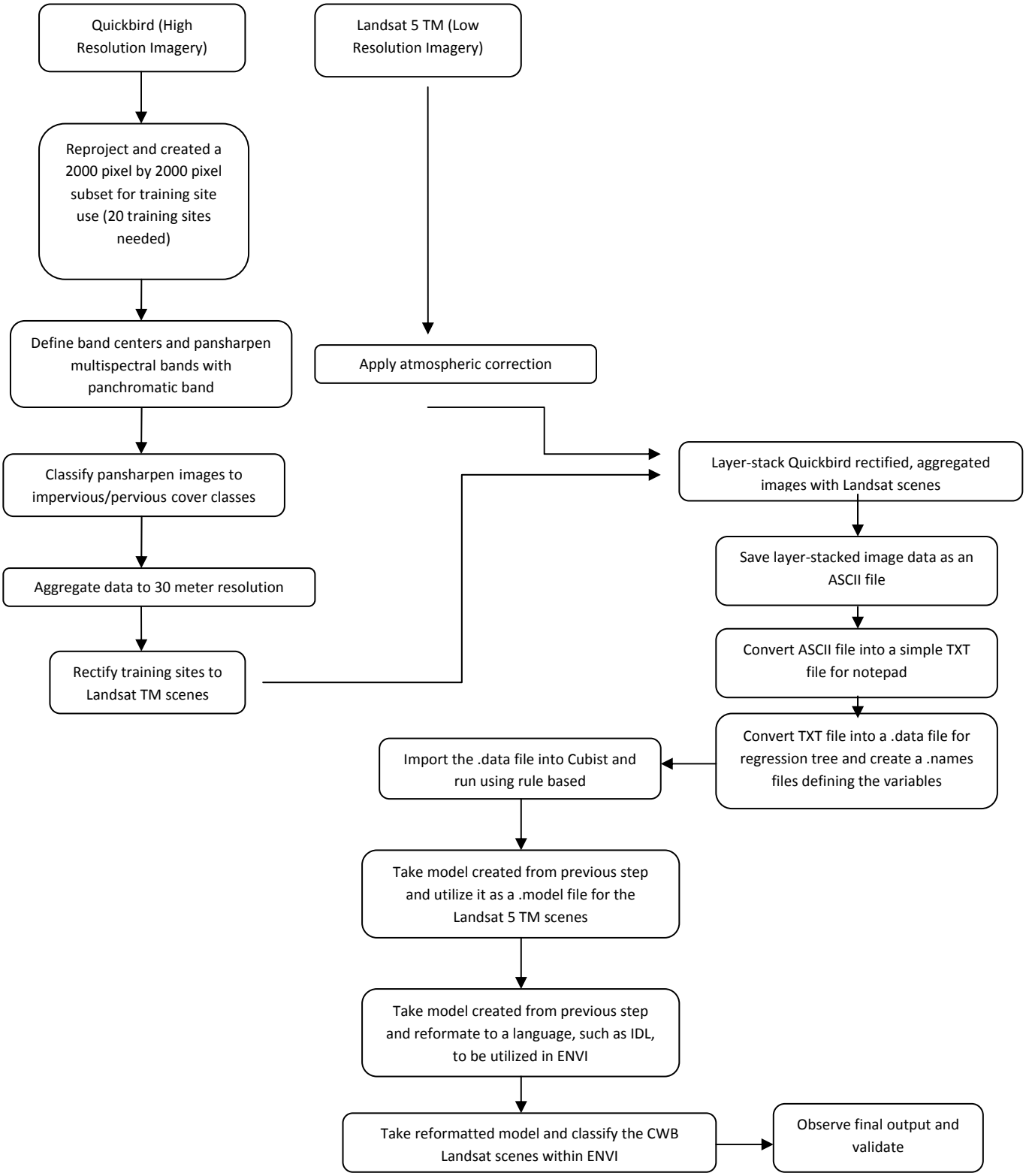
Masek, J.G., et al. (2006a), A landsat surface reflectance dataset for North America, 1990-2000, IEEE Geosci. Rem. Sens. Let., 3:68-72.

Masek, J.G., et al. (2008), North American forest disturbance mapped from a decadal Landsat record, Remote Sensing of Environment, 112:2914-2926.

Yang, L., C. Huang, C.G. Homer, B.K. Wylie, and M.J. Coan. (2003b). An approach for mapping large-area impervious surfaces; synergistic use of Landsat-7 ETM+ and high spatial resolution imagery. *Canadian Journal of Remote Sensing* 29:230-240.

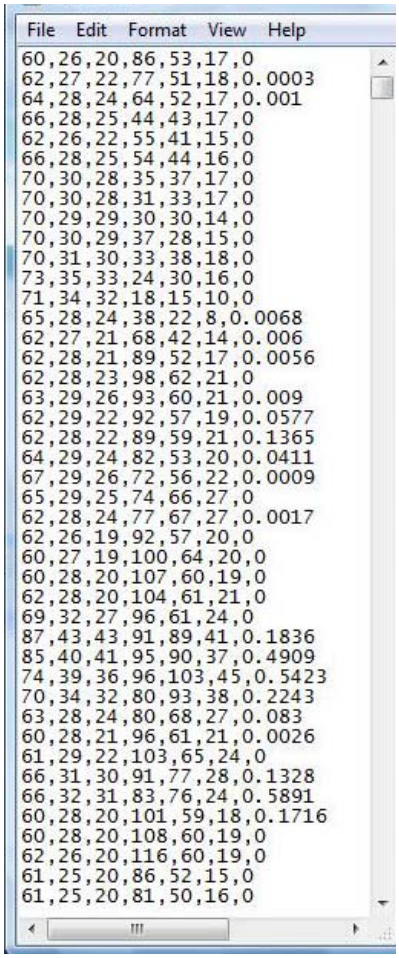
Appendix A

Logical model of methodology developed through the study.



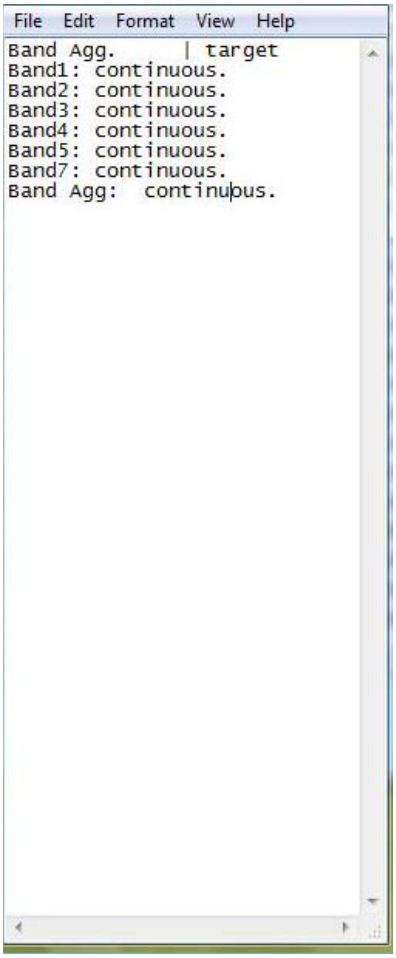
Appendix B

Example .data format and .names format for Cubist



```
60,26,20,86,53,17,0
62,27,22,77,51,18,0.0003
64,28,24,64,52,17,0.001
66,28,25,44,43,17,0
62,26,22,55,41,15,0
66,28,25,54,44,16,0
70,30,28,35,37,17,0
70,30,28,31,33,17,0
70,29,29,30,30,14,0
70,30,29,37,28,15,0
70,31,30,33,38,18,0
73,35,33,24,30,16,0
71,34,32,18,15,10,0
65,28,24,38,22,8,0.0068
62,27,21,68,42,14,0.006
62,28,21,89,52,17,0.0056
62,28,23,98,62,21,0
63,29,26,93,60,21,0.009
62,29,22,92,57,19,0.0577
62,28,22,89,59,21,0.1365
64,29,24,82,53,20,0.0411
67,29,26,72,56,22,0.0009
65,29,25,74,66,27,0
62,28,24,77,67,27,0.0017
62,26,19,92,57,20,0
60,27,19,100,64,20,0
60,28,20,107,60,19,0
62,28,20,104,61,21,0
69,32,27,96,61,24,0
87,43,43,91,89,41,0.1836
85,40,41,95,90,37,0.4909
74,39,36,96,103,45,0.5423
70,34,32,80,93,38,0.2243
63,28,24,80,68,27,0.083
60,28,21,96,61,21,0.0026
61,29,22,103,65,24,0
66,31,30,91,77,28,0.1328
66,32,31,83,76,24,0.5891
60,28,20,101,59,18,0.1716
60,28,20,108,60,19,0
62,26,20,116,60,19,0
61,25,20,86,52,15,0
61,25,20,81,50,16,0
```

.data file example



```
Band Agg.      | target
Band1: continuous.
Band2: continuous.
Band3: continuous.
Band4: continuous.
Band5: continuous.
Band7: continuous.
Band Agg:  continuous.
```

.names file example

Appendix C

Partial example of output model from Cubist software to be applied to Landsat 5 TM scenes within ENVI

```
Cubist [Release 2.07]  Fri Aug 06 02:00:36 2010

  Options:
    Use 85% of data for training


  Replacing unknown attribute values:
    `Band1' by 71.7

Read 7189 cases (7 attributes) from Cubist(6sites).data

Model:

  Rule 1: [74 cases, mean 0.02117, range 0 to 0.1866, est err 0.02537]

    if
      Band1 <= 89
      Band2 <= 48
      Band3 > 54
    then
      Band Agg = 0.30556 + 0.008 Band7 - 0.00367 Band5 - 0.007 Band1
               + 0.00184 Band4 + 0.0024 Band2

  Rule 2: [3031 cases, mean 0.02140, range 0 to 1, est err 0.02230]

    if
      Band7 <= 23
    then
      Band Agg = -0.00775 + 0.0015 Band7 - 0.00022 Band4 + 0.0007 Band2
               - 0.0004 Band1 + 0.00015 Band5

  Rule 3: [261 cases, mean 0.03532, range 0 to 0.417, est err 0.04016]

    if
      Band1 > 70
      Band1 <= 78
      Band5 <= 57
    then
      Band Agg = -0.48472 + 0.0103 Band7 - 0.0045 Band5 + 0.0081 Band1
               - 0.0056 Band3 + 0.00222 Band4

  Rule 4: [110 cases, mean 0.05941, range 0 to 0.6461, est err 0.06030]

    if
      Band1 > 74
      Band2 <= 37
      Band4 <= 75
      Band5 > 77
      Band7 > 28
      Band7 <= 60
    then
      Band Agg = 1.79324 - 0.0559 Band2 - 0.01148 Band5 + 0.0203 Band7
               + 0.005 Band4 + 0.0079 Band3 - 0.0012 Band1

  Rule 5: [187 cases, mean 0.06569, range 0 to 0.6597, est err 0.06170]

    if
      Band2 <= 37
      Band4 <= 75
      Band5 > 57
      Band7 <= 28
    then
      Band Agg = -0.16288 - 0.01227 Band5 + 0.0162 Band7 + 0.0153 Band3
               - 0.0173 Band2 + 0.00382 Band4 + 0.0063 Band1

  Rule 6: [67 cases, mean 0.06987, range 0 to 0.6756, est err 0.04760]

    if
      Band1 > 70
      Band3 <= 54
      Band4 > 123
    then
      Band Agg = -0.53359 + 0.011 Band1 - 0.00328 Band5 + 0.0029 Band7
               + 0.00023 Band4 + 0.0003 Band3

  Rule 7: [1271 cases, mean 0.08265, range 0 to 0.9757, est err 0.07872]

    if
      Band1 <= 70
      Band7 > 23
    then
      Band Agg = -0.32195 - 0.00295 Band5 + 0.0041 Band7 + 0.0049 Band1
```

(Appendix C cont.)

```
            + 0.005 Band2

Rule 8: [72 cases, mean 0.08892, range 0 to 0.8742, est err 0.09424]

  if
     Band1 > 70
     Band1 <= 72
     Band3 <= 54
     Band4 > 88
     Band4 <= 123
     Band5 > 106
  then
     Band Agg = -0.30825 - 0.00972 Band5 + 0.0129 Band7 + 0.00397 Band4
              + 0.0047 Band1 + 0.0047 Band3 - 0.0018 Band2

Rule 9: [302 cases, mean 0.11207, range 0 to 0.9382, est err 0.10436]

  if
     Band1 <= 89
     Band2 > 37
     Band4 <= 75
     Band5 > 82
  then
     Band Agg = -0.14806 - 0.01194 Band5 + 0.0134 Band7 + 0.0167 Band3
              + 0.0006 Band1 + 0.00024 Band4

Rule 10: [73 cases, mean 0.12972, range 0 to 0.7071, est err 0.09858]

  if
     Band1 > 70
     Band3 <= 39
     Band4 <= 88
     Band5 > 92
  then
     Band Agg = -2.05484 + 0.01637 Band4 + 0.00841 Band5

Evaluation on training data (7189 cases):

   Average  |error|            0.07266
   Relative |error|               0.42
   Correlation coefficient        0.82


      Attribute usage:
        Conds  Model

         71%    97%    Band7
         58%    90%    Band1
         36%    78%    Band4
         27%    97%    Band5
         21%    43%    Band3
         17%    77%    Band2


Evaluation on test data (1269 cases):

   Average  |error|            0.08121
   Relative |error|               0.47
   Correlation coefficient        0.77


Time: 0.4 secs
```