THE CHESAPEAKE BAY WATERSHED CROPLAND DATA LAYER

Rick Mueller, Program Manager
USDA/NASS
3251 Old Lee Hwy, Suite 305
Fairfax VA 22030
rick_mueller@nass.usda.gov

ABSTRACT

The Center for Geographic Sciences (CGIS) at Towson University sponsored a collaborative project with the USDA/National Agricultural Statistics Service (NASS) to produce a 2002 Cropland Data Layer (CDL) over the Chesapeake Bay Watershed. This CDL project was intended to produce a mapping of the agricultural variety and extent of the crops grown in the Chesapeake Bay Watershed. The CDL exists in the public domain and contains categorized ortho-rectified data layer of the specific crops grown in a state. Classification training data was based on a statistical sampling methodology applied to ground collected survey data. The CDL program has continually expanded in scope from one county (Craighead County, Arkansas) in 1996, to three states in 1997; five in 1999; eight in 2001 and to nineteen for 2003. Program expansion would not be possible without the creation of federal, state and university partnerships. Currently, the CDL program annually inventories the following states, based on a variety of these partnerships: AR, IA, IL, IN, MO, MS, ND, NE and WI. The CGIS project scope was to produce an inventory of the major crops grown in each of the following states for the 2002 crop season: CT, DE, MD, NJ, NY, NC, PA, RI, VA and WV. The remainder of this report will focus on the CGIS project and the results of the Virginia 2002 CDL, released in October of 2005.

INTRODUCTION

This project was initially funded through the NASA/Raytheon/Towson Synergy Project, and NASS used available in-house expertise and contract personnel for the production of this project. The traditional remote sensing methodology used in the Midwest, Northern Great Plains and Delta States of pixel-based supervised modified ISODATA clustering (Bellow and Ozga 1991) and maximum likelihood classification was applied to develop the CDL products for the CGIS project. Peditor software (Ozga 1985, 1995 and Day 2002) was used for the image processing, as Peditor was comprised of a modular set of programs that were specifically designed to estimate crop acreage with measurable precision using satellite imagery in conjunction with Agency ground survey data. A categorized ortho-rectified image of each entire state was released in the public domain in October 2005 for the entire ten state project and is available at www.nass.usda.gov/research/Cropland/SARS1a.htm or at cgis.towson.edu.

Forty-three multi-temporal Landsat TM and ETM scene locations displayed in Figure 1 were processed, and ground truth information was utilized from both the 2002 NASS June Agricultural Survey (JAS) and Agriculture Coverage Evaluation Survey (ACES) segments. The NASS Area Sampling Frame, used a stratified random sampling method, included 954 JAS and 407 ACES segments of approximately one square mile in size over the extent of the study area. In addition, extra signatures were obtained for training sites not covered by the ground surveys, providing additional non-agricultural class signatures. The final product released on CD by state contained: categorized mosaics published in GeoTIFF format, accuracy assessment results and associated metadata for the ten-state project.

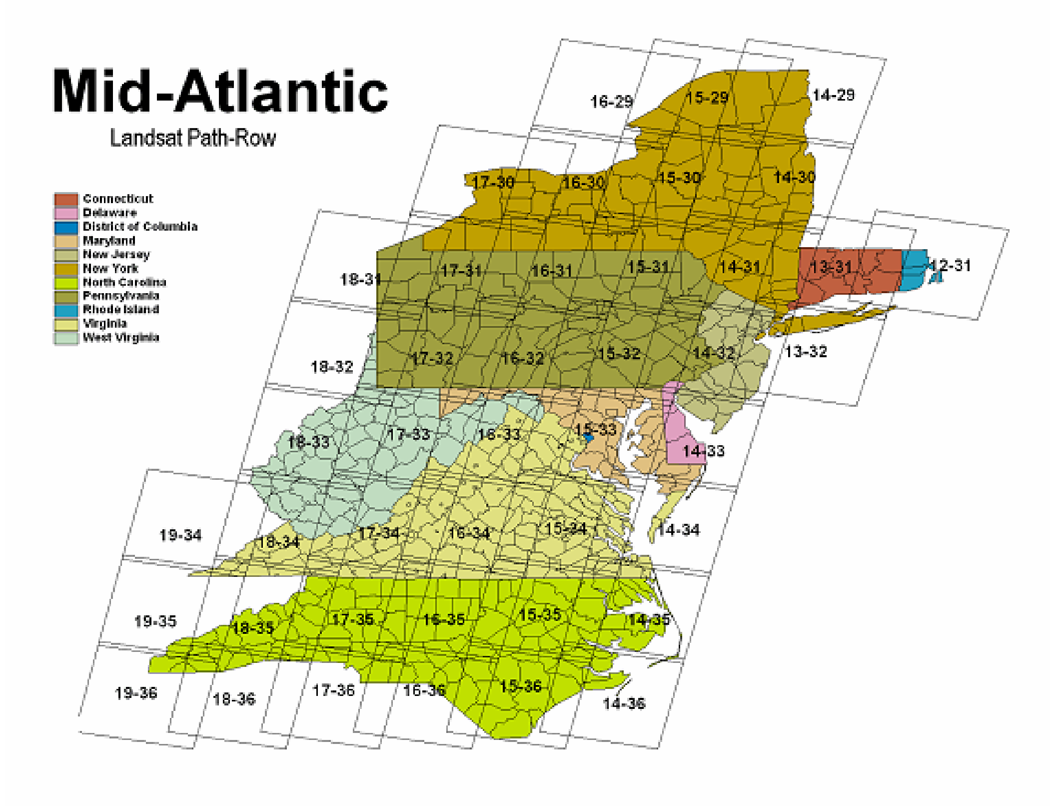

Figure 1. The Landsat TM and ETM path and row scenes used for the project.

Landsat Imagery

The timing of this project (i.e., crop year 2002) was fortunate enough to have both Landsat satellites in working order to cover a project of this magnitude allowing eight day repeat coverage, as the Landsat 7 ETM had the Scan Line Corrector failure during the next crop year (2003). Repeating this project in successive years would be nearly impossible with only one satellite. It was preferable to have multitemporal coverage to distinguish planted covers such as small grains versus row crops additionally from forested covers and non-agricultural/pasture and urban areas, as there exists a small window of optimum opportunity during the growing season to separate the different crop types (Allen 1988, Maxwell 1996). For this project, an early spring base scene of May/June and a secondary scene of July to early September was in most cases sufficient to separate the major crops, pastureland and non ag fields from one another.

A total of eighty unique Landsat scenes were processed for the entire project, with thirty-six being Landsat 5 TM and forty-four Landsat 7 ETM over the forty-three scene footprint project area. The multitemporal scene combinations were as follows; seven TM/TM, fourteen ETM/ETM, six TM/ETM, ten ETM/TM; and six unitemporal TM scenes. To cover Virginia, fifteen Landsat scene footprints were used, including the following multitemporal combinations; two TM/TM; three TM/ETM; four ETM/ETM; five ETM/TM and one TM unitemporal scene.

Visual comparisons of the categorized outputs between the TM and ETM sensor products provide similar results. However, for acreage estimation, determining the r-squared and slope statistics were more important as these measures were directly related to the standard error of the estimates (Craig 2002). Craig compared both TM and ETM sensors in North Dakota and Arkansas in 1999 and North Dakota and Iowa for 2000 in areas where each sensor overlapped each other by one day. The results showed that the regression slope statistics were not significantly different between the two sensors, but the r-squared regression statistics were significantly different, possibly because of better commission errors, with preference given to the ETM sensor over the TM. The satellite age may also play a determining factor in the results given that the ETM was between 6 months to 2 years in age versus the TM that was at least 15 years old at the time of the study.

Software

Three software components were used for this project; 1) Peditor, the image processor, regression estimator and mosaic creator, 2) RSP or Remote Sensing Project, managed the ground truth data collection and digitizing process, 3) XLNT a COTS software program developed by www.advsys.com that provided batch type processing on a XP device. These software applications were integrated to perform acreage estimation and mosaicking using satellite imagery with an Area Sampling Frame and stratified random sample segments.

NASS uses Peditor (Ozga 1995) for field digitizing/labeling, ground truth validation, image reformats and overlays, modified supervised clustering, maximum likelihood classification, percent correct and kappa coefficients based on classification accuracy assessment, crop acreage statistics based on ratio and regression estimators and full state ortho-rectified categorized mosaics (Mueller 2000). Peditor was originally written in Pascal and Fortran and is currently maintained in Delphi on the XP platform.

RSP was written in Visual FoxPro and RSP maintains the segment training set databases during the production of segment digitizing, labeling and ground truth validation. Each segment's production status was maintained by RSP, allowing an analyst to monitor the digitizing progress when there were between 200-300 segments in a state and it also allowed for the identification and rectification of problem segments. Additionally, RSP allows automated viewing and preliminary visual analysis of the raw and categorized training segments across each scene.

XLNT software is an commercial application that allows for batch type processing in the PC environment. Batch processing was utilized for scene reformatting, multitemporal scene overlays, clustering, classification (small and large scale), regression estimation and mosaic creation (Ozga 2000). The batch process was optimized to allow for linking of programs for processing efficiency. For instance, the clustering of a cover type was propagated across the network where separate cover types were identified and sent to available empty queues across the batch network where corn, soybean, cotton and pasture land will be processed by separate devices/processors across the network.

Ground Truth Data

This project encompassed the entire Chesapeake Bay watershed for crop year 2002 utilizing the extra training segments from the Agricultural Census along with the June Ag Survey. The training sites were drawn from an Area Sampling Frame where land was stratified into broad land use categories. The land was further stratified into several strata based on the distribution of cultivation in the state. A sample segment, approximately one square mile in size was selected by stratum with more concentrated samples occurring in highly intensive agriculturally productive areas. Table 1 shows the state sample segments for the June Agricultural Survey (JAS) and the Agriculture Coverage Evaluation Survey (ACES). The JAS segments (954 total) were selected annually for this survey, while the ACES segments (407 total) were added for the 2002 Ag Census, providing over 1,361 square miles of training fields. These surveys were combined and used as areas for ground truthing and training fields. Performing this same project on a non Ag Census year would have been very difficult without the benefit of additional training segments.

NASS was approached by the CGIS group in late 2002/ early 2003 about the possibility of performing a CDL project over the Chesapeake Bay Watershed for crop year 2002. Once NASS and CGIS came to terms about the project scope, it was necessary to work expediently to capture the segment photos before the next year's JAS occurred. It's customary in NASS Fields Offices to erase the field boundaries and labels in March or April prior to the next ground survey. The photos were sent into the NASS Spatial Analysis Research Section in early January/February of 2003 from the NASS Field Offices. The CGIS group dedicated resources to scanning the ACES photos, while NASS captured the JAS photos with a digital camera. The timing was critical to getting the photos captured and archived so that they were readable and useful for referencing when digitizing field boundaries and labeling the fields using proper naming conventions. Once the photos were archived they were sent back to the NASS Field Offices and the field boundaries and labels were promptly erased and the materials were prepared for the next JAS survey.

Table 1. Area Sampling Frame

State	JAS	ACES	State Total
CT	8	14	22
DE	23	1	24
MD	61	9	70
NC	319	29	348
NJ	48	16	64
NY	96	87	183
PA	179	101	280
RI	8	1	9
VA	146	132	278
WV	66	17	83
Total	**954**	**407**	**1361**

The NASS enumeration method records the crop or land use during the survey in addition to the acreage. The enumerator collects the following land use information on the entire segment and categorizes each land use type into one of the following categories: crop type, farmstead, waste, woodland (pasture/not pasture), pasture (cropland or permanent), summer fallow, idle cropland, non ag, two crops or two uses in the same field, acres left to be planted and irrigated acres. Additionally, there were additional crop or commodities that were only grown in certain regions of the country, and each state has the ability to add these commodities as necessary (i.e., Christmas trees, orchards, groves, vineyards, nursery, fruit or nut trees, etc). Some of these land use classification categories can make clustering separability difficult; for instance, summer fallow and idle cropland were the same except summer fallow allows for chemical application or land tillage, and woodland pasture was different from woodland not pasture as woodland pasture allows for grazing animals. However, the definitions for crops were well defined and fields were delineated for only one crop type in a field at any one time.

Once the segments were enumerated during crop year 2002, the segment field boundaries were digitized and labeled according to the surveys and questionnaires. The grower reported acreage was compared to the digitized acreage, and corrections to field boundaries and acreage were made post survey called ground data editing. The final ground truth was prepared and fields were marked as bad for training, such as: acreage discrepancies greater that ten percent between reported and digitized acreage were excluded, fields smaller than eight acres were excluded, each field was buffered inside by one pixel with the intent of dropping all edge and field boundary pixels; additionally the analyst had the ability to drop additional fields where the imagery or field boundaries did not match the cover type or field layout. The small and bad fields were removed from the training set, fields buffered and the data was now ready for clustering.

Analysis Districts

The Analysis Districts (AD) were defined based on best available cloud free satellite imagery and the scenes that offered the best separability for crops vs. non crop ground cover. An AD can be defined by a scene edge or a county boundary. The scenes in an AD must be of the same image acquisition date, but the scenes do not have to be consecutive or contiguous, but they must have the same base and secondary acquisition dates to be defined as an AD, if multitemporal or the same date if unitemporal. The entire ten state project required twenty four AD's to cover the entire Chesapeake basin. Figure 2 shows the Analysis Districts and dates of image acquisition across Virginia, while Figure 3 shows the extent of the AD's throughout Virginia. AD 13, 17, 20 and 21 on the southern Virginia border expand into North Carolina, as well as AD 9, 16, 19, and 21 expand into West Virginia, Maryland and Delaware. Each of Virginia's ten AD's overlapped into the adjacent states, and the training data from these other states was appended and used with the Virginia training data to increase the amount of training data available for clustering. All of the segments or training fields that lie within the confines of an AD were clustered regardless of state boundaries. For instance; AD 13, 17, 20 and 21 used both the VA and NC 2002 JAS and ACES segments, additionally the adjoining AD's in North Carolina used the VA surveys for their clustering.

Extra Signatures

The CDL training categories were determined using a combination of the JAS and ACES enumerative surveys and a manual technique to derive extra training signatures where the NASS categories were missing or lacking adequate training data. For the clustering/classification processing fields smaller than eight acres were removed from the training set, non ag, waste and other waste fields were marked as bad or unknown and removed from training, and all other training fields were buffered in by one (30 meter) pixel to minimize field edges and capture the interior of the training field. The main reason for the removal of the non ag, waste and other waste fields was that the land usage of these fields were very similar and indistinguishable from other non ag fields established using the NASS survey specifications, as these covers may contain a combination of forest, pasture, fallow and non ag covers and can be best separated using a separate set of extra signatures.

The extra signatures were captured using a manual photo interpretative process, capturing rectangular image chips for the extra signature areas of interest. Five broad land use categories were used to further separate the non ag signatures. These broad land use classes were defined as follows: water, urban, clouds, wetlands and woods. These categories were further subdivided into additional categories such as; water was divided into shallow vs. deep, clouds: shadows and haze, urban: bright vs. dark and urban grassland, woods: hardwood versus softwood. For each scene footprint, approximately 500 image chips were used per scene and each chip was clustered and appended together with the Peditor signature editor called the Statistics Editor. Each category was appended to an extra signature cluster file and each cover type was renamed to its proper cluster name. The resultant extra signature clusters contained approximately 12-18 signatures or categories per Analysis District, and were then appended to the master training sets including the JAS and ACES segments. The extra signatures increase the separability and accuracy among non ag fields and further helped refine the classification.

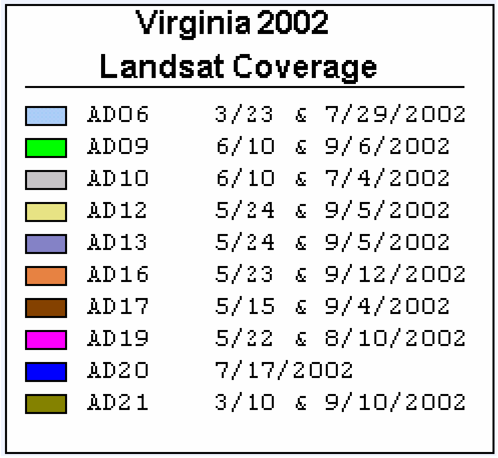

Figure 2. Landsat coverage dates for Virginia 2002.

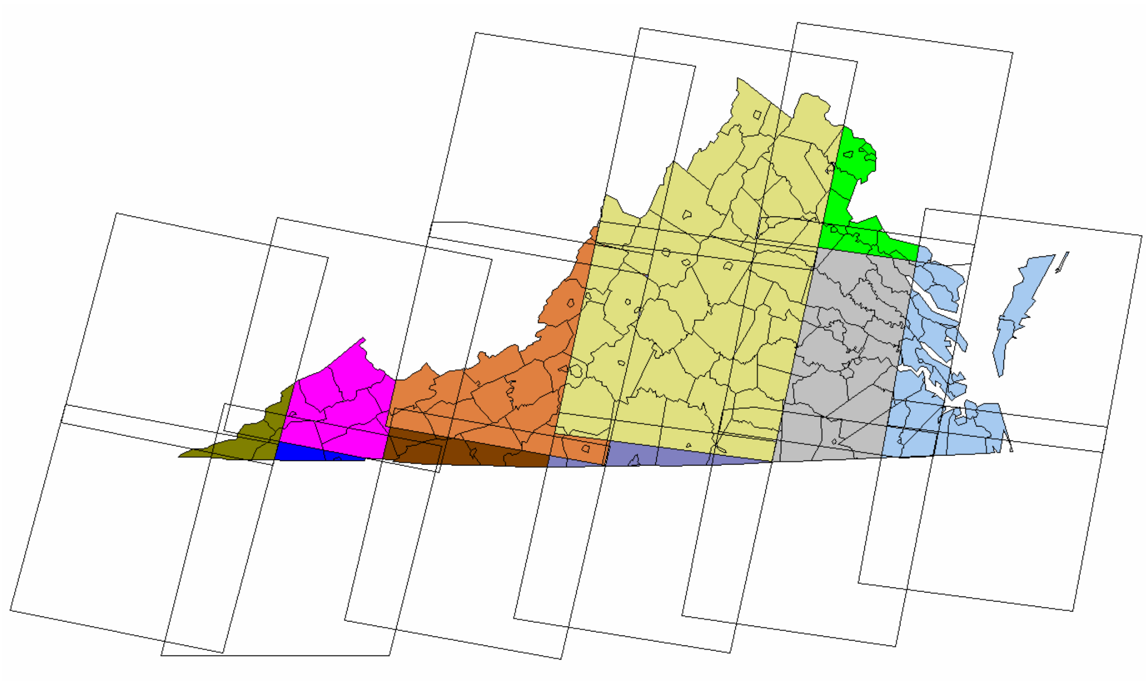

Figure 3. 2002 Virginia Analysis Districts.

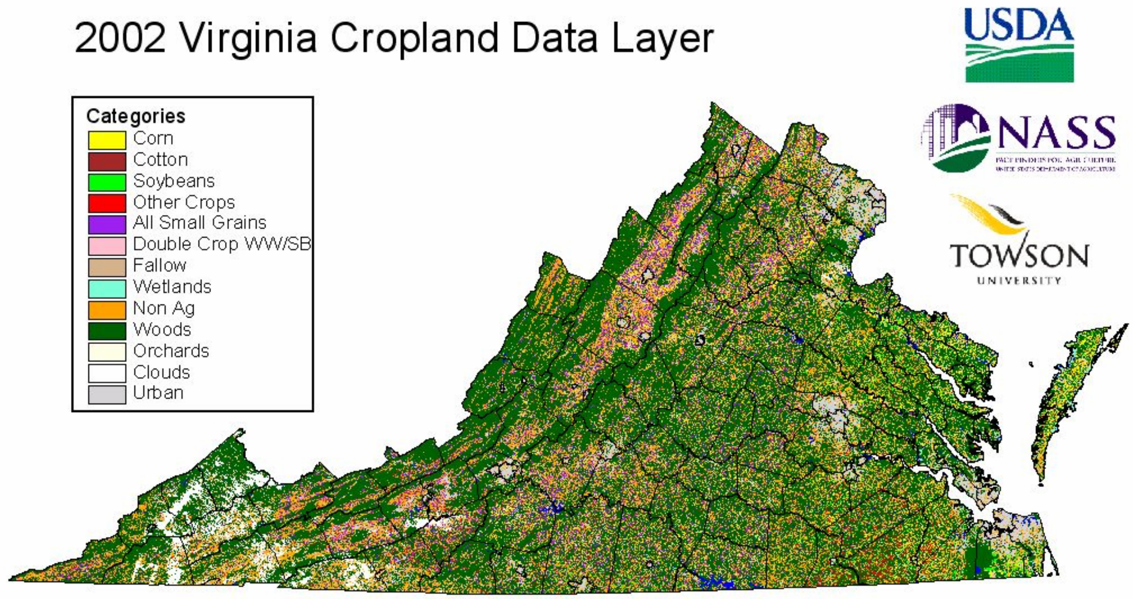

Figure 4. The 2002 Virginia Cropland Data Layer.

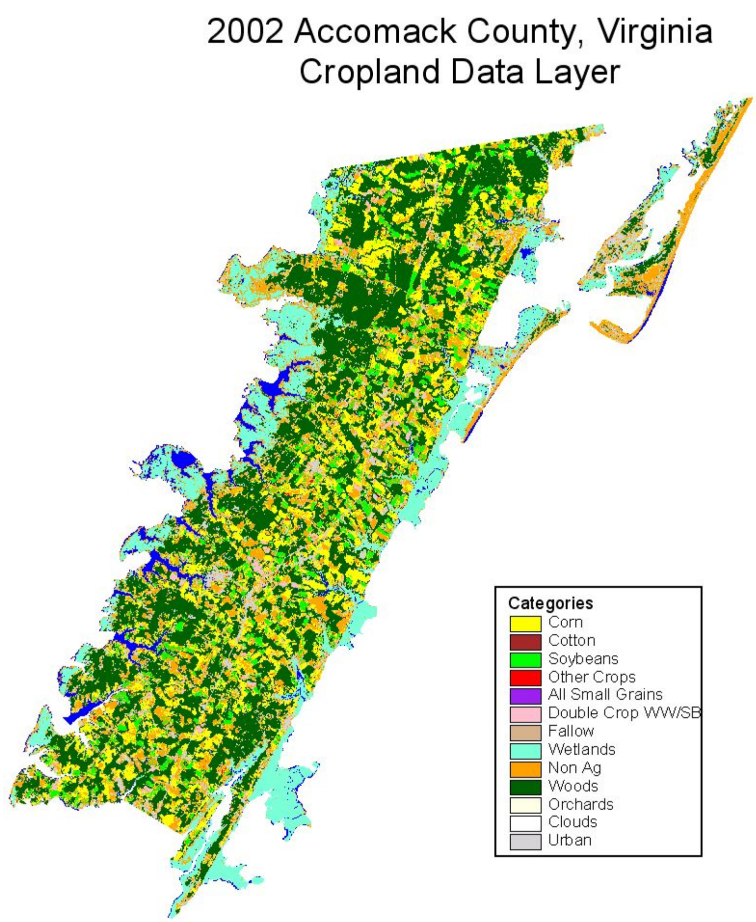

Figure 5. The 2002 Accomack County Cropland Data Layer.

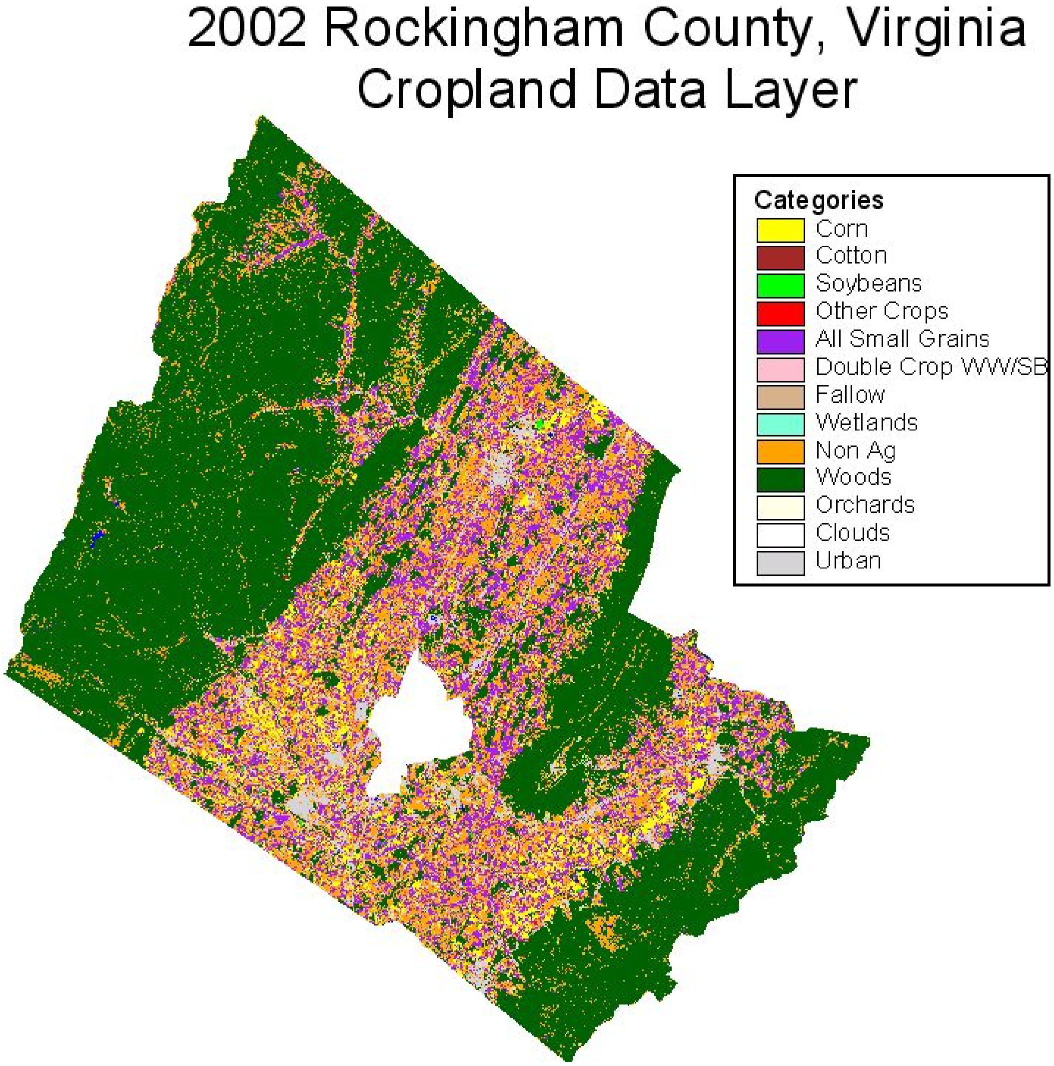

Figure 6. The 2002 Rockingham County Cropland Data Layer.

CDL Creation

Once the final full scene image classifications were processed, efforts began on creating a CDL for each of the ten states. The EarthSat Inc. GeoCover ortho stock mosaics were used as the base to correlate all of the processed raw scenes in the project. Band two of the GeoCover and band 2 of each raw scene in the project were correlated against each other determining the polynomial coefficients, and creating a pixel mapping between the scenes (Mueller and Ozga 2002). The categorized imagery was then prioritized by unitemporal or multitemporal classification with low accuracies or bad observation dates and were placed in the lowest priority, and images with better accuracies and image dates were prioritized higher and overlaid onto the lower priority images where they were clipped by scene edge or county boundary. The categorized imagery prioritization process allowed for the removal of clouds if a scene of lower priority was cloud free, otherwise the cloud will remain. The image categories were then recoded to a master catalog where the signatures were placed into previously determined categories, and the mosaic was exported to ERDAS Imagine .LAN format. The images were attributed and colorized in Imagine, and exported to GeoTIFF format. The accuracy assessments, Analysis District maps, supporting statistical information on the regression and slope coefficients and image metadata were then published as the Cropland Data Layer or CDL.

For the VA 2002 CDL only the scene lines were used as a basis to prioritize the placement of scenes into the mosaic, as shown in Figure 3. For instance, AD12 had priority over AD's 9, 10, 13 and 16, while AD10 had the lowest priority as AD 6, 9, 12 and 13 had priority and the extent of AD10 was minimized. Although efforts were made to create the best possible classification for each scene and AD, it was inevitable to have noticeable scene lines running across the image. NASS's regression estimator compensates for the inaccuracies in the classification (Winings 1990), as the pixel-based methodology does not eliminate the speckling effect on categorized imagery. There were no smoothing operations performed on the mosaic or CDL, as NASS prefers to leave the output unmodified, in its original form. However, object-based processing methods or smoothing algorithms might improve the appearance of the mosaic, but not necessarily improve the accuracy of the estimates.

Virginia Results

The accuracies for Analysis District 10 are displayed in Table 2. All of the row and wheat crops have a very high percent correct and Kappa, and low commission errors, with the exception of cotton; this had a low percent correct 40.54 and Kappa 39.88, and a high commission error of 40.59. AD10 was a multitemporal scene district defined by two scenes one on June 10, and the other on July 4. It was evident that the cotton fields were not easily separable, as an August observation is usually necessary to separate this crop from soybeans. Additionally, a winter wheat crop signature as well as a winter wheat followed by soybean signature was created allowing for the analysis of double cropped and single cropped fields. Other non ag and minor fields were not included in this table. However, the overall percent correct accuracy nearly 86 percent and an overall Kappa of 82 percent include all major and minor fields that were considered good for training.

The Virginia Area Sampling Frame was built in 1978 and was defined as follows: Stratum 13 was land use greater than 50% cultivated with four JAS segments and eight ACES segments contained in stratum 13 across the state. Stratum 20 was cultivated land use between 15 – 50% with 88 JAS segments and 154 ACES segments spread across the state. The other stratums were < 15 % cultivated; agri-urban > 20 homes per square mile; dense urban > 20 homes per square mile; resort > 20 homes per square mile and non ag; all of which were not included in the production of the CDL, as the inclusion of these segments did not contain additional field crops, and tended to produce mixed non ag and ag categories/signatures that were better defined just using extra signature collects. Virginia's land use stratum is currently being updated and will be released in time for the 2006 JAS.

Table 3 summarizes the major crops of corn, cotton, soybean, winter wheat and alfalfa for the 2002 Virginia CDL, displaying the r-squares and slope coefficients. AD10 performed well in corn, soybeans and winter wheat, while cotton under performed as expected because of the early scene observations. Winter wheat performed well with an r-squared at .991 and a .2294 slope coefficient with .2224 being the optimum slope. The slope coefficient was derived by dividing the pixel square meters by the total number of meters in one acre. This accuracy can be attributed to the early scene observation of June 10, catching the winter wheat in the ground.

Figure 4 is the finalized categorized mosaic of Virginia, note the variety and location of crops spread throughout the state. It was impossible to remove the cloud covers from the classification, as was evident in western Virginia. Accomack County on the eastern shore is one of the highest producing counties for corn and soybean acreage in Virginia and is displayed in Figure 5. The Accomack classification was dominated by corn, soybeans and woods, also note the detailed delineation of the wetlands areas. Figure 6 shows Rockingham County, with the town of Harrison removed from the center, note the dominance of all small grains and corn in the classification.

Table 2. Accuracy assessment of Analysis District 10 by major cover type.

ANALYSIS DISTRICT AD10 LANDSAT TM/ETM+ PATH: 15, ROW(S): 34 & 35 - (06/10 + 07/04/2002) 253 CROP / COVER TYPE SIGNATURES, 14 CHANNELS						
Category	**Crop Cover**	**Orig. # Category**	**Orig. # Pixels**	**Percent Correct**	**Commission Error**	**Kappa Coefficient**
1	Corn	14	1309	95.03	4.09	94.22
2	Cotton	2	148	40.54	40.59	39.88
5	Soybeans	27	825	94.06	4.9	93.48
10	Peanuts	4	220	96.36	24.56	96.25
11	Tobacco	4	154	94.81	9.32	94.71
24	Win Wheat	2	126	99.21	2.34	99.2
26	WW/Soybean	7	593	98.92	4.4	98.74
44	Other Crop	3	59	98.31	17.14	98.29
62	Perm Past	8	635	60.00	30.22	57.49
63	Woods	48	3932	83.6	0.72	74.43
63	Wood Past	4	440	81.59	50.55	80.22
	Overall	253	9235	85.86		82.28

Table 3. Accuracy assessment of the 2002 Virginia Cropland Data Layer by crop.

CATEGORIZATION OF VIRGINIA 2002 CROPS					
Regression Analysis By Crop R-Squared and Slope Values From Sample Estimation					
R-Squared			Slope Coefficients		
Analysis District	Strata 13	Strata 20	Analysis District	Strata 13	Strata 20
CORN					
AD06	0	0.688	AD06	0	0.1614
AD10	0.972	0.972	AD10	0.1898	0.1898
AD12	0.735	0.735	AD12	0.2242	0.2242
SOYBEANS					
AD06	0	0.743	AD06	0	0.2251
AD10	0.908	0.908	AD10	0.2511	0.2511
COTTON					
AD10	0	0.714	AD10	0	0.2116
WINTER WHEAT					
AD10	0.991	0.991	AD10	0.2294	0.2294
ALFALFA					
AD12	0.631	0.631	AD12	0.2204	0.2204
AD16	0	0.984	AD16	0	0.2574

CONCLUSIONS

The project sponsorship by the Towson University, Center for Geographic Sciences allowed for the creation of a ortho-rectified categorized ag specific CDL over the Chesapeake Bay Watershed for crop year 2002. The project results were quite successful, as the timing allowed for the usage of both Landsat TM and ETM satellites as well as the benefit of the Ag Census provided nearly fifty percent more segments over the project area. Detailed image analysis using Peditor, a pixel-based public domain software application over the study area allowed for multiple states to be analyzed using Analysis Districts that crossed over state boundaries. The results of the Virginia 2002 analysis show the diversity of the crops grown in Virginia, including; corn cotton soybeans, small grains and other crops. Considerable effort was extended into creating extra signatures to separate the woodland from pastureland from urban areas, essentially the non ag cover types, future projects will need these extra signatures in order to provide useful classifications for the public to disseminate these land use types. The successful completion of this project will allow the public the opportunity to view this ag specific land cover mapping over the Chesapeake Bay Watershed in addition to the associated accuracy assessments and metadata available at www.nass.usda.gov/research/Cropland/SARS1a.htm or cgis.towson.edu.

ACKNOWLEDGMENTS

The author would like to thank the Towson CGIS, especially Drs. Rick Kutz and Jay Morgan for their support and confidence in the NASS remote sensing methodology for performing this project. Additionally, thanks go out to Mr. Patrick Willis, for his outstanding sustained efforts in processing this ten state project. Mr. Willis not only tackled a project of extremely large magnitude utilizing his outstanding analytical skills and capabilities, he continually refined and polished the image categorizations and his commitment to detail for the extra signature collections made the results of this project exceed expectations.

REFERENCES

Allen, J. D. and G. Hanuschak (1988). The Remote Sensing Applications Program of the National Agricultural Statistics Service: 1980-1987. *USDA/NASS SRB Staff Report No. SRB-88-08*.

Bellow, M.E. and M. Ozga (1991). Evaluation of Clustering Techniques for Crop Area Estimation Using Remotely Sensed Data. *1991 Proceedings of the Section on Survey Research Methods, American Statistical Association*, Atlanta, GA. pp 466-471.

Craig, M. (2002). Comparing 30 Meter Imagery From Landsat 5 And 7 For Crop Area Estimation. *American Society of Photogrammetry and Remote Sensing*, *Proceedings,* April 2002.

Day, C.D. (2002). A Compilation of PEDITOR Estimation Formulas. *RDD Research Paper RDD-02-03, USDA/NASS*, Washington, D.C. January, 2002.

Maxwell, S.K. and R. M. Hoffer (1996). Mapping agriculture crops with multi-date Landsat data. *Proceedings of the National ASPRS/ACSM 1996 Annual Convention*, Baltimore, MD, p433-443.

Mueller, R. (2000). Categorized Mosaicked Imagery from the National Agricultural Statistics Service Crop Acreage Estimation Program. *American Society of Photogrammetry and Remote Sensing, Proceedings,* May 2000.

Mueller, R. and M. Ozga (2002). Creating a Cropland Data Layer for an Entire State. *American Society of Photogrammetry and Remote Sensing*, *Proceedings,* April 2002.

Ozga, M. (1985). USDA/SRS Software for Landsat MSS-Based Crop-Acreage Estimation. *InternationalGeoscience and Remote Sensing Symposium*, IGARSS 85, Amherst, MA. September 1985.

Ozga, M. (2000). Batch Processing of Remote Sensing Jobs on the PC. *American Society of Photogrammetry and Remote Sensing, Proceedings,* May 2000.

Ozga, M. and M. E. Craig (1995). PEDITOR - Statistical Image Analysis for Agriculture. *Washington Statistical Society (WSS) Seminar*. April 5th 1995.

Winings, S. B., M. Ozga, and J. Stakenborg (1990). The Effects of the Application of Smoothing and Orthogonal Transforms to SPOT and TM Data on Regression Based Crop Acreage Estimates. *Proceedings of the IGARSS'90 Symposium*, College Park, MD. Vol. 1. pp. 647-650. February 1990.