CREATING A CROPLAND DATA LAYER FOR AN ENTIRE STATE

Rick Mueller and **Martin Ozga**
USDA/National Agricultural Statistics Service
Spatial Analysis Research Section
3251 Old Lee Hwy, Suite 305
Fairfax, VA 22030
rmueller@nass.usda.gov
mozga@nass.usda.gov

PURPOSE

A collection of Landsat scenes corresponding to an entire state or a major portion of a state, are categorized based on ground truth information collected from farmers by USDA enumerators. However, no farmer reported data is revealed or derivable from the categorized Landsat scenes due to confidentiality protections. The individual categorized Landsat scenes need to be geo-referenced and stitched together to a common ortho-rectified base in order to be released as a public use GIS file. EarthSat Inc.'s GeoCover stock mosaic was chosen as the ortho base, because the GeoCover product offered accuracy and vast coverage over all of our project areas. The registration of the GeoCover mosaicked scene and the individual raw input scenes are used to get an approximate correspondence. A correlation procedure is used on the raw Landsat scenes and the mosaicked scene to get an exact mapping of each pixel from the input Landsat scenes to the mosaicked scene. The results of the correlation are used to remap the pixels from the individual input scenes into the coordinate system of the mosaicked scene. The image analyst then specifies the mosaic priorities for scene placement, and the mosaic process begins by using the polynomials from the correlation to place the categorized pixels. A classified ortho-rectified mosaicked image is then output for distribution into the public domain.

INTRODUCTION

The USDA National Agricultural Statistics Service (NASS) performs the June Agricultural Survey (JAS) annually where enumerators collect field and acreage data based on a probability survey derived from a stratified area frame sample. During the JAS approximately 11,000 samples sites are visited nationally, called segments, and each segment is approximately one square mile in size. Most of the states participating in this project contain between 150 to 400 segments. The segments usually provide adequate training sites for the major crops grown in a state. NASS has performed extensive research integrating the JAS with satellite data to perform acreage estimates (Allen and Hanuschak 1988, and May, Holko, and Jones 1986). The combination of using Landsat based imagery along with area frame ground based data, provides the opportunity to improve the precision of estimates for major commodities grown in a state (Allen 1990). NASS has used PEDITOR as the Agency's image processor since the mid 1970's, to provide supplementary county and state level estimates for the major crops grown within a state using a regression estimator. PEDITOR's major function up until a few years ago was to provide accurate acreage indications, however, in the late 1990's it became evident that distributing the categorized images would be a unique image and GIS product.

NASS tried a variety of ideas before determining the optimum solution for producing a quality categorized image mosaic in a timely basis (Mueller 2000). The two main issues involved mosaicking 6-15 full Landsat scenes per state each year, while repeating the process on each successive year, and providing a method to automate the ortho-rectification process. The process of mosaicking multiple categorized images across an entire state, combined with multiple state projects, provided incentive to develop an efficient alternative to massive scene stitching, masking and overlap prioritizing across all of these projects. Previous attempts to create these mosaics resulted in intensive manual efforts involving image registration, recoding all categorized image categories to a common category set, and then stitching the images together.

PEDITOR was developed to process digital satellite imagery to estimate crop acreage over a large area (Ozga & Craig 1995). The name PEDITOR was originally derived from Portable EDITOR, where the software was ported to a variety of computer hardware platforms, notably the Cray, Digital MicroVAX, and IBM mainframe, as those platforms were the only ones capable of handling the processing requirements at that time. However,

advances in PC computing hardware and software have made the porting of PEDITOR to the PC possible. PEDITOR has been maintained in-house and contains much of the functionality available in commercial image processing systems. However, program/process modifications are relatively easy to support in a research type environment, and the development/release cycle is faster. PEDITOR is deployed in all participating NASS State Statistical Field Offices to handle the ground truthing process and all image processing tasks, and is continuously tested with the Spatial Analysis Research Section (SARS) in Fairfax, Virginia. Currently, PEDITOR runs on most Microsoft Windows platforms; however, PEDITOR's batch processing system programs only runs under Windows NT or 2000.

The Cropland Data Layer (CDL) Program is processed in conjunction with the regression estimator, where the categorized scenes used to make an acreage estimate, are also used to create a categorized mosaic. The CDL is currently in production in seven States, including; Arkansas, Illinois, Iowa, Indiana, Mississippi, North Dakota, and New Mexico, while Missouri and Nebraska are in ongoing pilot programs (See Figure 1). The CDL is an offshoot from the acreage estimation program, where state and county estimates are made from the major crops grown within a state. The emphasis of the CDL is to provide the same categorized imagery that was used to provide the acreage estimates. Extensive research was performed on creating categorized mosaicked images of an entire state using Landsat 5 TM and Landsat 7 ETM+ imagery, including; using ERDAS Imagine to manually register and mosaic all scenes, using ESRI's ArcView to create registration tie points before stitching in Imagine, and using PEDITOR to reproject each categorized image to the GeoCover ortho-registered base. These methods all worked to some degree, but the human resource and capital requirements were too extensive to support the volume of work required on an annual basis for these ongoing projects. Designing a better way to accurately stitch these large mosaics was something that became a major focus within the SARS group. The following discussion provides the current methodology used to mosaic categorized scenes to a common ortho-rectified base image.

BACKGROUND

The mosaicking process proceeds through three major program steps. First, the mosaic information file is created with information solely from the GeoCover composite scene. Second, the registration of each scene is used to get an approximate mapping between the pixels of the composite and each raw scene to be used. A correlation procedure is performed with the raw data scenes against the GeoCover composite scene to get an exact mapping. This mapping is in the form of least squares polynomials which are stored in the mosaic information file. Third, the output mosaic scene is initialized to all zeroes and the pixels from each classified scene are mapped into the coordinate system of the mosaicked scene using the polynomials created in the second step. This involves remapping each scene's categories into a common class set for the composite scene, and also use of user supplied priorities where scenes overlap. The categorized pixel is then placed in the mosaic. Both the correlation and mosaic procedures are done in a batch system created for Windows NT/2000 (Ozga 2000) since several programs are involved and since both, particularly the mosaic procedure, can be quite time consuming. The image file is then imported into ERDAS Imagine and prepared for public distribution.

METHOD

Preparation

The stock mosaics were procured from EarthSat Inc.'s GeoCover program for an entire state. Each GeoCover mosaic covers approximately fifteen Landsat scenes. Approximately four to six GeoCover stock mosaics cover an entire state depending on the state size/boundary, and the mosaic cutline. A GeoCover mosaic covers up to five degrees latitude, and six degrees of longitude. Each mosaic contains Landsat bands 7,4,2, a spatial resolution of 28.5 meters, and is projected in the Universal Transverse Mercator (UTM) map projection. Band 2 is stripped out of each GeoCover mosaic and imported into ERDAS Imagine as the base layer to perform the correlation against. The GeoCover mosaic is resampled to 30 meters, and reprojected to a single UTM zone. Some of the participating CDL states are divided into two UTM zones. The UTM zone that predominates or occupies the largest area is chosen for that state. Each GeoCover composite scene is then mosaicked to cover the entire state, leaving some additional area to account for possible Landsat scene footprints outside of the state. The mosaicked scene is clipped outside of the state boundaries where necessary to account for all possible Landsat scene extents/footprints with respect to the state boundary. Finally the scene is exported as an ERDAS .LAN format file.

Initialization

A mosaic information file is initialized to contain information about the LAN file created from the GeoCover composite file. This includes the number of rows and columns, the UTM zone, and registration information. As scenes are correlated in the next step, polynomial and scene information is added. This program does not require batch type processing.

Correlation Procedure

The next step involves image correlation, and requires usage of batch processing because of the volume of processing. The correlation procedure assumes that the GeoCover composite scene and each additional input raw scene are from the same sensor or closely related sensors and have the same pixel size. The same band (2) is chosen from each raw input scene and the GeoCover scene. The creation of raw multi-temporal scenes (Allen 1990) uses the same methodology and shares similar PEDITOR code, as the scene correlation process. The registration of the GeoCover composite and each individual raw scene is used to find the block location within the raw scene for each block in the composite scene. The block pairs are used to generate the final overlay. All of the parameters including the channel used and the various threshold parameters are set by the software based on past experience and not by the user. There is no limitation to the number of scenes that can be correlated to the base scene, as long as the scene falls within the extent of a given state. Also, the program can be re-run later as additional scenes are obtained, and scene prioritizing is not a factor at this point.

The correlation proceeds as follows. A collection of 64 by 64 pixel blocks is selected from the GeoCover scene and a collection of 32 by 32 pixel blocks is chosen from the candidate raw scene to be overlaid. Currently, 60 rows of 60 blocks are chosen on a grid. The blocks of satellite data that have corresponding centers are created for both scenes. The centers of the blocks are selected to be at the same location based on each registration. For improved correlation performance, a gradient procedure is performed on each block. The correlation computation is performed at each position of the 32 by 32 block within the 64 by 64 block and the position yielding the highest result is saved.

Blocks are deleted that have too low a correlation value or too high a shift from the center. The remaining blocks are used to create least squares polynomials of degree three. These polynomials compute the row and column of the input scene from the row and column of the composite scene. Then, an iterative procedure begins with the polynomials applied to each block pair and the block pair with the highest residual eliminated until the residual goes below a value or the number of blocks falls below a threshold value. If the number of blocks is above the threshold value, new polynomials are generated from the remaining blocks. In practice, this procedure very rarely fails and if it does, it is usually due to a bad registration for a scene.

The correlations are done with the GeoCover image being treated as the primary scene and each raw scene selected being the secondary scene. This creates a mapping into each scene for each pixel in the GeoCover image. Of course, since the GeoCover image is larger than any scene, many pixels in the GeoCover image may map to outside of any particular individual scene. The polynomial coordinates from all the various correlations are stored in the mosaic information file for later retrieval during the mosaicking process.

Creating the Mosaic

The mosaicked categorized scene may be created once all raw scenes of interest have been correlated and the polynomial coordinates written to the mosaic information file. Before proceeding however, all scenes must have already been classified. A master crop category data set was previously established that allows for the remapping of certain class types into other predetermined class types (e.g., idle cropland and fallow cropland are collapsed into a single idle cropland class, or cropland pasture, permanent pasture and non agricultural classes are collapsed to single a non agricultural class), or it can collapse multiple categories of the same class (e.g., twenty classes of spring wheat are collapsed to one class of spring wheat). The data set was established in .DBF format, to be accessed as each pixel of each scene is processed. Each pixel is translated to the category from the crops .DBF file based on the category to cover assignment from the classification process. This is necessary since each of the input classified scenes may have several categories for a cover and these are not necessarily the same for all of the classified scenes. Each analysis district has a unique set of categories or classes, and mapping each class to a common set is necessary. PEDITOR maintains the pixel information in a separate statistics file, where the means and covariance matrices used for classification are stored.

The image analyst must also decide on a priority of scenes to be used in overlapping areas. Pixels from a lower priority scene will not overwrite those from a higher priority scene, unless those in the higher priority scene are in a cloud or filler category. Filler is the area of the scene rectangle that contains null data. Finally, the image analyst can select certain county scene pairs for which the pixels will overwrite those created from the full scenes, even if the scene is of a lower priority. If this option is used, the mask files created during the acreage estimation process are utilized. These mask files define the cutlines between counties and scenes, and contain the priority of scenes overlap/underlaps to one another. It is possible to prioritize scenes either by scene edges only, all county boundaries, a selected set of counties, or use a combination of scene edges and county boundaries for processing.

Figure 2 displays the analysis districts for North Dakota for crop year 2001. An analysis district (AD) can be defined by either scene boundaries or geo-political boundaries. An analysis district has to have the same date of observation, throughout all scenes whether primary or secondary scenes. North Dakota had nine AD's for crop year 2001. Analysis district 1 was comprised of two scenes, and used a combination of scene edges, and county boundaries. It was a good overall classification with observation dates of July 18th and August 27th, 2001, perfect for North Dakota's growing season. There were some cloud problems on the western edge of AD01, and it was preferable to use the counties associated with AD02 and AD03 to fill in the holes. Analysis district 4 had some cloud problems also, and it was necessary to use a combination of scene edge cutting between adjacent scenes, and county boundary clipping, allowing AD02 and AD05 to take priorities in pre-defined areas. Analysis district 9 was observed on May 12th and June 21st, 2001, and defined by the under-lap between Landsat paths 32 and 34. The classification was not good enough for acreage estimation, but is being evaluated for inclusion in the full statewide mosaic. These are some of the issues faced by the image analyst when determining when and where to prioritize images, and where to place the scene cutlines.

The mosaic is created within the batch processing environment. A copy is made of the GeoCover image, and all pixel locations are set to zero, as there is never a class zero in a PEDITOR categorized image. Then, each scene is processed. For each pixel in the newly created mosaicked image, the polynomials from the mosaic output file are used to find the pixel location in the classified file for that scene using the nearest neighbor rule. If such a pixel exists, and it agrees with the priority scheme, it is placed in the mosaicked file. Finally, if any county scene pairs were selected, they are processed, overriding any other priority but using only pixels from within the county.

2001 Preliminary North Dakota Mosaic

Figure 3 shows the 2001 CDL for North Dakota. The North Dakota mosaic took approximately 10 hours to stitch together and process on a 550 MegaHertz Windows NT Workstation. There were lots of cloud problems over North Dakota for the 2001growing season. First of all, there was not an abundance of cloud free images available, combined with the very near shutdown of the Landsat 5 TM sensor. The transfer of ownership from SpaceImaging to the EROS Data Center appeared to create delivery and image availability problems when trying to access potential scenes from the TM sensor during the summer of 2001. Figure 4 shows Cass County, North Dakota, the most intensely cultivated county in the State. A quality classification usually results as long as optimum scene dates are obtainable, and in this case July 18th and August 27th were optimum. The field sizes in Cass are exceptionally large, well defined and provided good quality training sites, while minimizing spectral confusion between the crops.

Preparation For Distribution

The statewide CDL mosaics are not available for general distribution until the county estimates are released for a particular state. The majority of states are available for distribution at the end of March from the previous crop year. The county estimates for rice are not published until June, and the mosaics for the States of Mississippi and Arkansas are held until they are released.

The finalized mosaicked image is edited in ERDAS Imagine. The image statistics and projection information are rebuilt, and the image is colorized to maximize the color separation between the crops grown in any given state. Since the categories or classes were standardized in the PEDITOR mosaicking process, a text file is imported in ERDAS to populate each state's categories. The class colors do not change between years for a particular state; however, the colors of similar classes between neighboring state may be changed to enhance the contrast between the various crop types grown in the neighboring states. The images are exported to ERDAS .GIS image format and bundled into an ESRI ArcExplorer format project. This enables potential users who don't have access to a GIS or image processor, to view the categorized imagery, along with some ancillary vector data, also bundled on the CD-ROM. The SARS group also provide extensive metadata on the methodology used, frequently

anticipated questions (FAQs), as well as accuracy statistics by analysis district, such as percent correct, commission error, kappa coefficient, regression r-squared, and regression slope on the CD-ROM. The CDL products are available on the NASS website at http://www.nass.usda.gov/research/Cropland/SARS1a.htm.

CONCLUSIONS

The mosaicking process used by the SARS groups encompasses three major steps: first initialization of a mosaic information file to hold the least squares polynomials, next correlating the GeoCover composite image against each individual raw image to get an exact mapping between the two images, and finally mapping the pixels from the categorized image into the coordinate system of the mosaicked scene, by using the previously created polynomials. The image analyst has the ability to choose scene priorities by scene edges only, individual or groups of counties, all counties or a combination of scene edges and counties. The methodology attempts to account for the most common types of image overlay scenarios encountered. Additional or alternative mosaicking scenarios will be dealt with on an as needed basis with primary concern being of maintaining image integrity throughout the entire process. The GeoCover composite image proved it's worth as a viable product for image co-registration over a large sized area, such as a state. PEDITOR solved a difficult issue reducing the burden of manually registering each individual image, and then stitching it together, and uses batch processing methods to mosaic the computational intensive programs that are required to run each state. The future plans for the CDL are to continue production in the seven project states, and eventually expand the program using the current methodology into additional states as program partners are found, or as additional resources become available.

REFERENCES

Allen, J. Donald, (1990). "Remote Sensor Comparison for Crop Area Estimation Using Multitemporal Data," U.S. Department of Agriculture, *NASS Staff Report No. SRB-90-03*.

Allen, J. Donald, George A. Hanuschak (1988). "The Remote Sensing Applications Program of the National Agricultural Statistics Service: 1980-1987," U.S. Department of Agriculture, *NASS Staff Report No. SRB-88-08*.

May, George, Martin Holko, and Ned Jones Jr. (1986). "Landsat Large-Area Estimates for Land Cover," *IEEE Transactions on Geoscience and Remote Sensing*, GE-24, No. 1 (January 1986).

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

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

Ozga, Martin, and Michael Craig, (1995). "PEDITOR – Statistical Image Analysis for Agriculture," *Washington Statististical Society Seminar,* April 5th 1995.

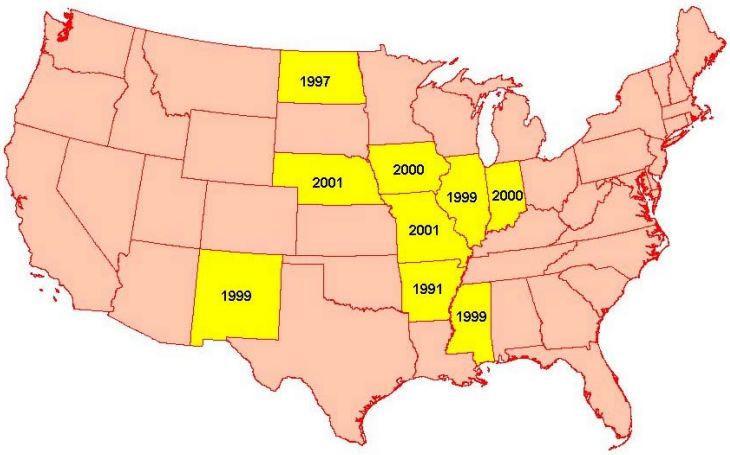

Figure 1. Cropland Data Layer participating states. The year the CDL program began is highlighted for each state. Note that Missouri and Nebraska are currently in pilot phases of the program.

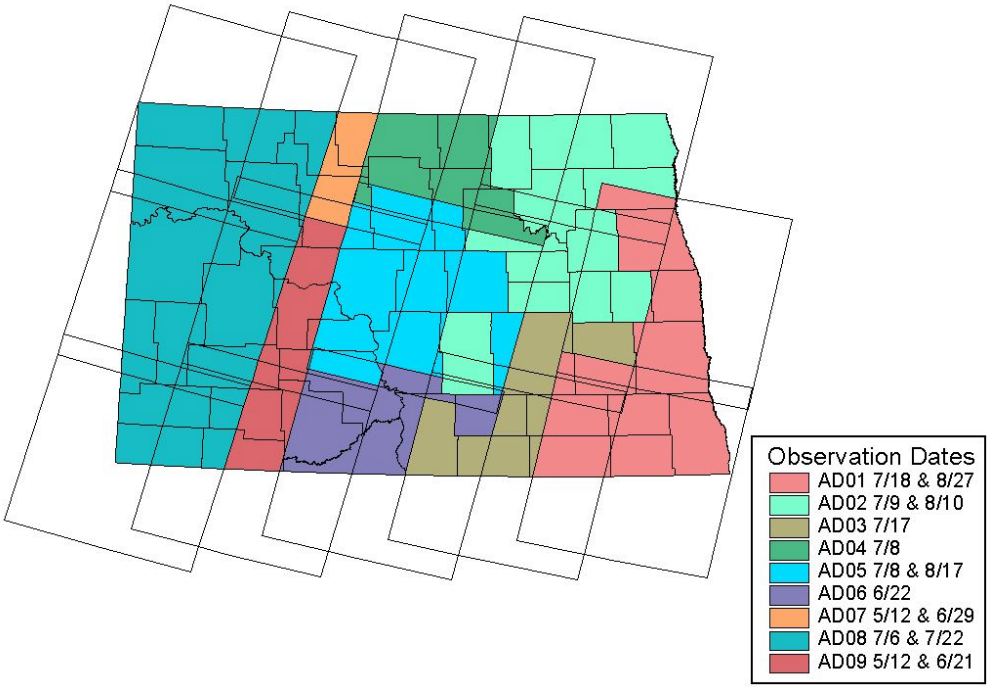

Figure 2. The North Dakota 2001 Scene Analysis Districts. These were the cutlines used to make the mosaic.

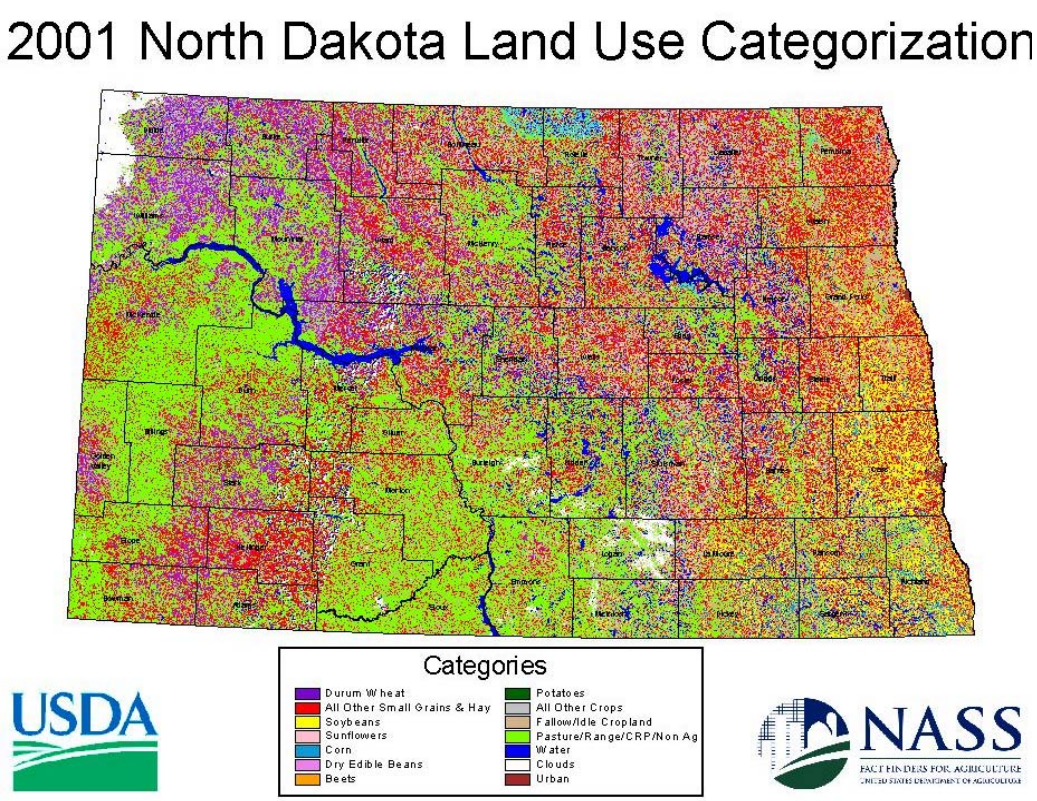

Figure 3. The 2001 North Dakota Cropland Data Layer. Note the visible scene line between AD06 & AD09

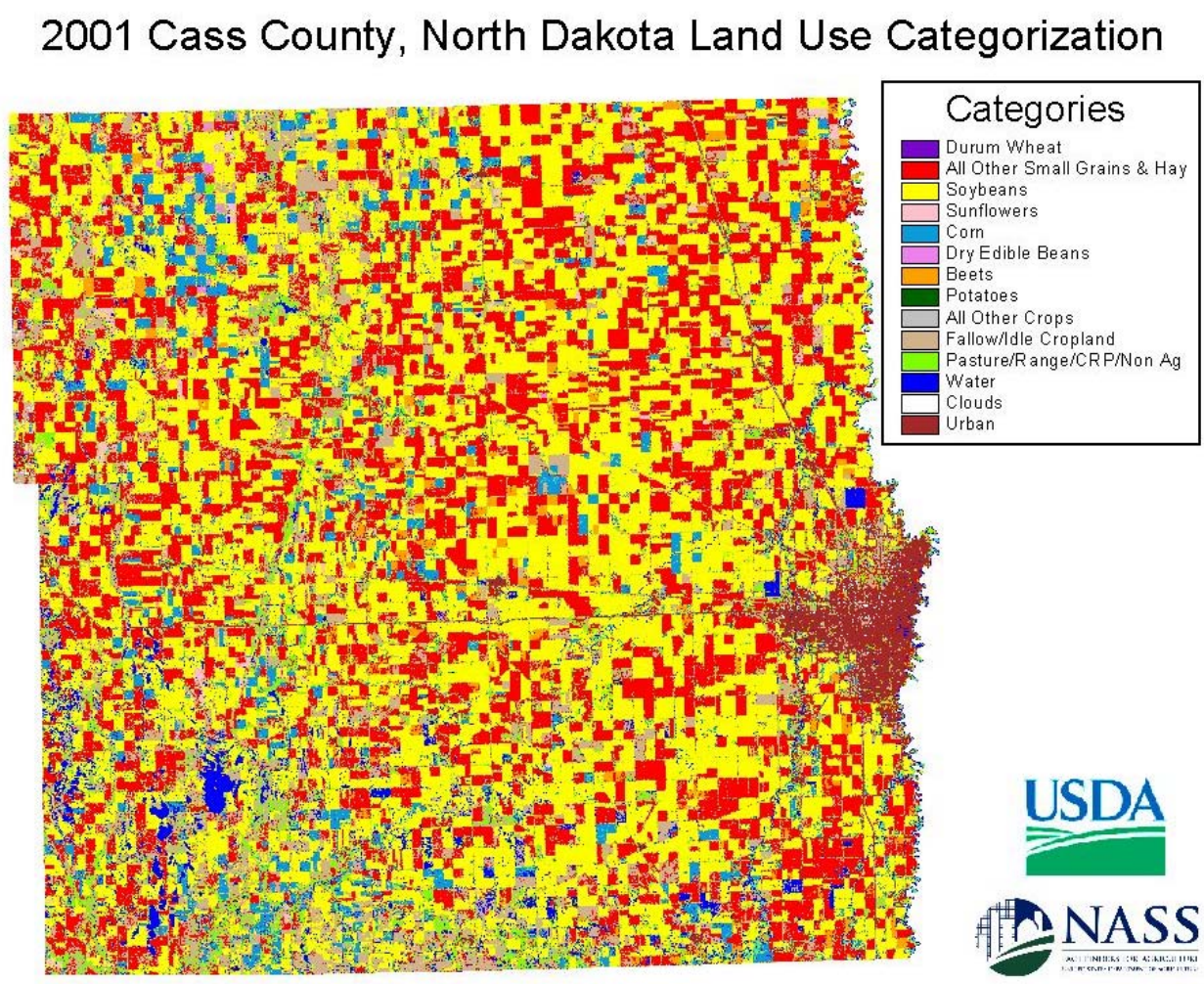

Figure 4. The 2001 Cass County, North Dakota Cropland Data Layer. Note the large field sizes, and the crisp boundaries between fields. Fargo is on the eastern edge.

.