A Fresh Approach to Agricultural Statistics: Data Mining and Remote Sensing

Darcy Miller, Jaki McCarthy, Audra Zakzeski

National Agricultural Statistics Service
3251 Old Lee Highway, Fairfax, VA 22030

Abstract

Agriculture has evolved from mule and plow into a high-tech business. Likewise, cutting-edge techniques to produce agricultural statistics used to guide policies and standards that bring our nation's food from the farm to shelf have sprouted from the innovation of statisticians at the USDA's National Agricultural Statistics Service (NASS). Recent use of remote sensing has been the crux of the development of a "Census by Satellite" (Cropland Data Layer Program), and data mining techniques have excavated statistics utilized in ensuring the quality of data from collection to estimation. The CDL Program integrates the traditional enumerator collected ground survey data with satellite imagery to produce indications of acreage for major commodities and state-wide mosaicked categorized maps. Data mining uncovers otherwise hidden information NASS uses to improve its own operations. With its ease of use and ability to illustrate complex results in simple terms, data mining proves to be fertile ground for future innovations in agricultural statistics.

Key Words: Data Mining, Remote Sensing, Cropland Data Layer, Estimation, Weighting

1. NASS Products and Growth

The USDA's National Agricultural Statistics Service (NASS) conducts hundreds of surveys every year and prepares reports covering virtually every aspect of U.S. agriculture. Production and supplies of food and fiber, prices paid and received by farmers, farm labor and wages, farm finances, chemical use, and changes in the demographics of U.S. producers are only a few examples. NASS reports are often used directly and indirectly by farmers, producer organizations, agribusinesses, researchers, policymakers, and government agencies.

The Agency was originally started in 1863 for the purpose of providing general production information to those interested in agriculture, chiefly on an annual basis. Today, NASS issues about 425 statistical reports from its Headquarters each year. Its Field Offices issue 9,000 reports and news releases annually, highlighting or expanding on information in the national reports, and the Agency conducts and releases the results of the census of agriculture every five years.

The practice of agricultural statistics has changed tremendously since George Washington's first small scale inquiry and report in 1791, through a letter to farmers in a small region. Complex survey designs and analysis techniques are applied using new

technologies for data collection such as Computer Assisted Telephone Interviewing (CATI), Computer Assisted Personal Interviewing (CAPI), and satellite data.

NASS remains at the top of the statistical methodology food chain by continually appraising old methods and then generating more methods and/or incorporating techniques used in other sectors to yield the best estimates. Recent innovative applications in the Agency include remote sensing to construct land cover maps and acreage estimates with the Cropland Data Layer program and data mining techniques to identify patterns and subgroups in large datasets to use in data collection and analytic operations. These techniques have not only made an impact on the process, quality, and timeliness of agricultural statistics but also opened statistical windows for other sectors.

2. A Taste of Data Mining and Remote Sensing

Data mining techniques are used to find patterns, classify records, and extract information from large data sets. These techniques, often used in the private sector for market research, fraud detection, and customer relationship management, can also be used by statistical agencies to analyze their large survey datasets. While large datasets are common in many statistical agencies, data mining techniques have not been widely used to improve the production of official statistics. With large datasets, information is often hidden, but data mining techniques can be used to distill or uncover it. Various techniques can be used to classify data into subsets, predict outcomes based on the data, cluster records into like subgroups, or assign propensity scores for some measure to records.

Remote sensing is the extraction of information about an object without coming into physical contact with it. Many countries have remote sensing programs providing direct or indirect support to official agricultural statistics programs including the EU-25, China, India, and some undeveloped countries in Africa, Southeast Asia, and Latin America. NASS uses a form of remote sensing, satellite imagery, to obtain unique, timely, detailed land cover classification. An Indian Advanced Wide Field Sensor (AWiFS) satellite takes snapshots of the Earth's landscape. NASS purchases these pictures from a centralized bank which is utilized by several other federal agencies. The Spatial Analysis Research Section (SARS), in NASS' Research and Development Division, uses a commercial suite to manage data, classify pixels, produce visual products, and estimate acreages. Continual research in the area of remote sensing methodology by SARS has increased the ability to deliver geospatial content annually to stakeholders.

Innovative applications of these techniques can be very effective in efforts to improve survey data collection, processing, estimation and dissemination.

3. Innovative Data Mining Techniques

3.1 Decision Trees

Decision tree models use an algorithm to segregate data based on the maximum difference found in a set of variables with respect to a target variable. Algorithms such as the chi-square automatic interaction detection (CHAID) algorithm can be used to determine how to split the segments. The segregation is done in a sequential manner. The splitting rule finds the variable and value to split on, and then observations are

carried down the appropriate branch to a node. At each of the new nodes, the splitting rule will repeat the procedure, sending the observations within that node down the appropriate branch to the next node. The final nodes are called leaves. The result is a figure that resembles a tree.

For example, Figure 1 is a set of characteristic data including the independent variables *outlook*, *temperature*, *humidity*, and *windy* as well as the dependent (target) variable *play*. The objective is to predict whether or not a person will play golf on a given day. A traditional approach may use logistic regression or a chi-square test. Developing methodology in the area of data mining, the decision tree, can also be used to predict the conditions that are most appealing for playing golf. Figure 2 is a decision tree that was formed using the data. The decision tree in Figure 2 shows that individuals are more likely to play on days that are *overcast*, *sunny & less humid* OR *rainy & not windy*.

Play golf dataset

Independent variables				Dep. var
OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	FALSE	Don't Play
sunny	80	90	TRUE	Don't Play
overcast	83	78	FALSE	Play
rain	70	96	FALSE	Play
rain	68	80	FALSE	Play
rain	65	70	TRUE	Don't Play
overcast	64	65	TRUE	Play
sunny	72	95	FALSE	Don't Play
sunny	69	70	FALSE	Play
rain	75	80	FALSE	Play
sunny	75	70	TRUE	Play
overcast	72	90	TRUE	Play
overcast	81	75	FALSE	Play
rain	71	80	TRUE	Don't Play

Figure 1: Example Data

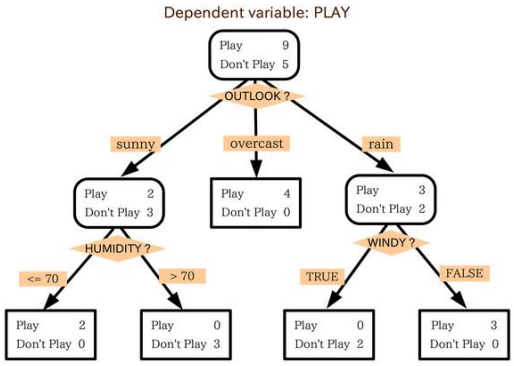


Figure 2: Example Tree

NASS employs data mining as a tool in a variety of research applications. Thus far, NASS has chosen to use the CHAID algorithm in determining splits over variables.

3.1.1 Census Non-response Weighting

Decision tree models were used to divide the 2007 Census records into response propensity groups representing weighting adjustment cells (Cecere, 2008). Variables such as operator race and gender, farm type and size were used to segment operations on the census mail list (CML) into subsets with homogeneous response propensities. Non-response weights were then generated for each of these groups individually.

3.1.2 Census Mail List Trimming

Decision tree models were also built to identify records on the initial CML (Garber, 2008) that were not likely to represent farming operations. A set of variables including the source of the record, the length of time the record had been on the NASS list frame, the location of the operation, and the previous gross receipts of the operation were used to identify operations with lower probabilities of qualifying for farms. For areas with larger than desired mailing lists, these operations with lower probabilities of qualifying as a farm were removed from the CML. This application improved the overall efficiency of census processing and reduced data collection costs.

3.1.3 Prediction of Survey Non-respondents

Decision tree models were used to indentify predictors of non-response (McCarthy and Jacob, 2009; McCarthy and McCracken, 2009) in several NASS surveys. Variables describing the operation and operator were included along with information concerning

the response history of the operation. Research concluded that the significant predictors of non-response are the response history variables (See Figures 3 & 4 below). Study results may be used to maximize data collection efforts through better direction of intensive data collection efforts or extra incentives.

Figure 3 shows two of the trees grown using the Quarterly Agricultural Survey data. How can these be interpreted? Two types of non-response are non-contact (respondent cannot be reached) and refusals (respondent refuses to answer when contacted). In the refusal tree on the left hand side, for example, the variables with the highest dichotomy in refusal/nonrefusal are *3-year response rate* history and *number of refusals in the current year*. The refusal rate for the subgroup with a *3-year response rate* less than 42% and *more than one refusal in the current year* are more than 2.5 times as likely to be a refusal for the current survey.

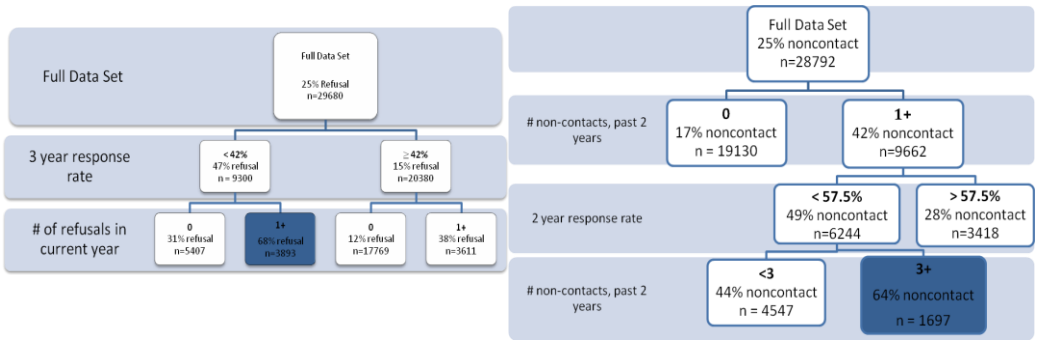


Figure 3: Decision Trees Grown Using the Quarterly Agricultural Survey Data. The tree on the left has refusals (respondent refused to answer the survey) as a target while the tree on the right shows non-contact (the respondent was not able to be reached) as a target variable.

3.1.4 Analysis of Reporting Errors

Decision tree models were also used to identify characteristics of operations with specific reporting errors (McCarthy and Earp, 2008). Models were constructed for several types of reporting errors such as subtotals not equal to their subparts. Location, type of commodities raised, size of operation, and operator demographics were the variables used as possible predictors. The resulting trees revealed that operations with certain types of land had a much higher reporting error rate. The fruit of this study can be used as a tool in questionnaire design or the programming of edits for these items.

3.1.5 Allocation of Survey Incentives

Decision tree models were used to identify the characteristics of sample units that were most likely to respond with or without incentives (Earp and McCarthy, 2009). The Agricultural Resource Management Survey (ARMS) uses several different types of incentives, both monetary and non-monetary, to encourage response. Several models that were constructed performed well based on an analysis of the number of respondents that were correctly classified as being respondents. Results from this study could possibly assist in targeting certain types of incentives to subgroups of sample units to maximize the efficacy of the incentives. Analysis was also done on potential savings. Using a model could allow NASS to reallocate between $6,290.08 and $119,370.08 towards enticing likely mail non-respondents currently not enticed by the monetary incentive.

3.2 Cluster Analysis

Cluster analysis groups records into clusters based on similarity among all designated variables. Although many methods can be utilized to classify the records, all methods lead to observations within each group being more similar to each other than to those in other groups.

3.2.1 2007 Census Donor Pool Screening

The 2007 Census of Agriculture imputation process required NASS to seed the donor pool with records from the 2002 Census of Agriculture. However, during the 2002 Census of Agriculture the use of an optical character recognition (OCR) program to capture the data led to some errant results. For example, the OCR program interpreted 0's with a slash through them to be 8's. Analysts had to manually correct this in the editing process. Due to the large number of records, a manual review became impossible. However, cluster analysis applied to outlier detection was used to segment the data. The cluster containing the OCR errors could then be screened out.

3.2.2 Questionnaire Design and Construction

Different commodities may be estimated at a variety of intervals across each of the 50 states. Therefore, 50 state versions of the Quarterly Agricultural Survey questionnaire exist. Hierarchical clustering techniques were used to propose how individual state's questionnaire versions for the Quarterly Agricultural Survey could be combined to reduce the number of questionnaire versions (Earp, Cox, McDaniel and Crouse, 2008). Implementing "regional" versions of the questionnaires highlighted in this research could reduce the amount of resources necessary to produce questionnaires and collect data while still recognizing the unique estimation needs of each state. An example of one region can be found in Figure 5.

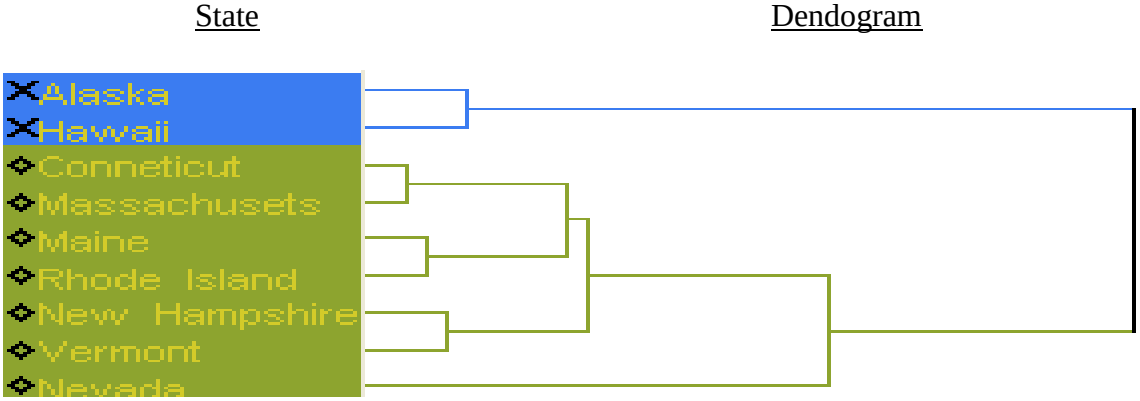


Figure 5: One of the "5 Region" Clusters formed through cluster analysis. This region is comprised of 2 of the "20 Region" Clusters.

3.3 Association Analysis

Association analysis generates association rules that describe which items within records tend to occur together. In market research, this is known as "market basket" analysis. These item associations find items that tend to appear together in consumer purchases. Generating these associations is a function of the strength of the association, the frequency of occurrence, and the predictive utility of the relationship.

3.3.1 Survey Data Edit Design

Association analysis can be applied to a data set such that the data within each record acts as items in a basket. Known relationships will appear; however, it is possible that this type of analysis will reveal unknown associations, relationships on missingness, or rare associations as a result of imputation. Although this type of research has not been yet been completed, it is possible that the outcome of such an effort could be used to assist in designing survey edits.

4. Remote Sensing and the Cropland Data Layer

According to the American Society of Photogrammetry and Remote Sensing, remote sensing can be defined as "the extraction of information about an object without coming into physical contact with it." NASS uses remote sensing through several different satellites for several of its new and on-going projects, including the Cropland Data Layer (CDL) program. The CDL is a "census by satellite," which depicts accurate field crop locations. This provides an estimate of acreage for targeted crops. Geographic information systems (GIS) software can be used to process the data and build geospatial snapshots of cropland (See Figure 6). These crop-specific maps provide land cover information that is used not only by farmers and agribusiness, but also by other entities with interests such as assessing urban sprawl, watershed analysis, or deforestation.

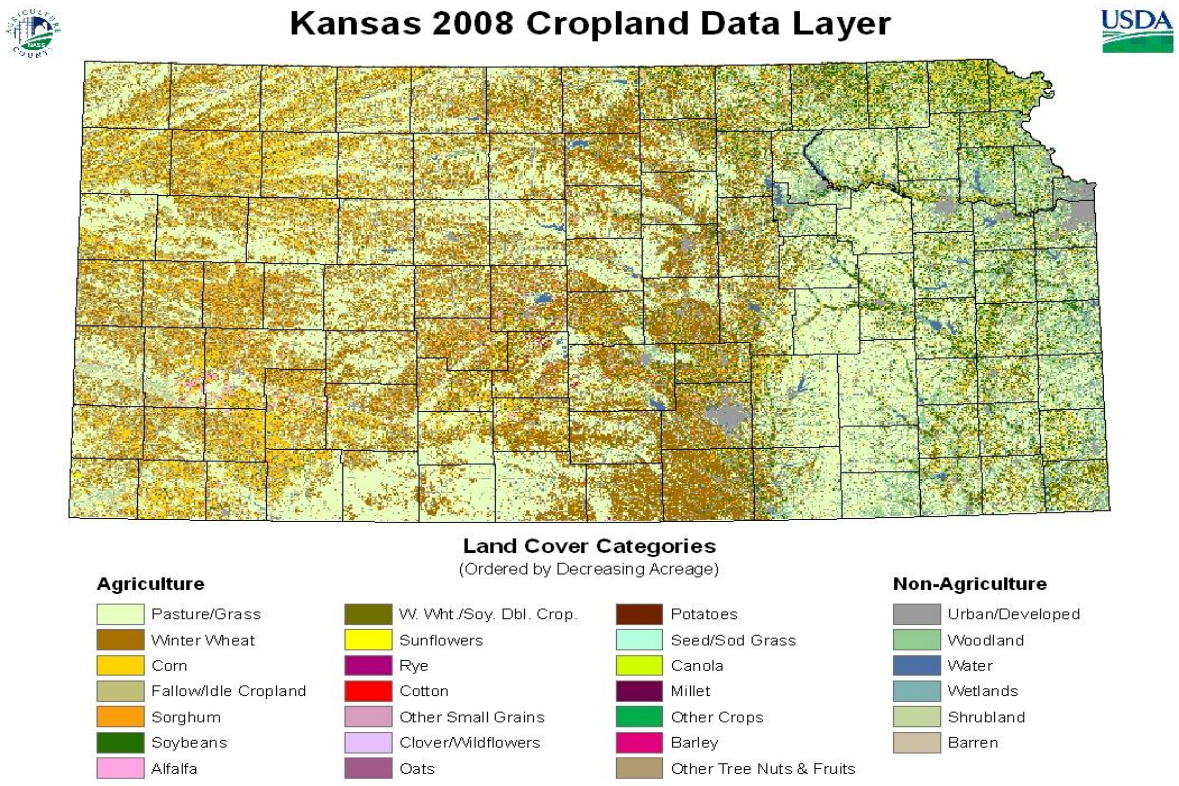


Figure 6: Kansas 2008 Cropland Data Layer

The Spatial Analysis Research Section (SARS) in Fairfax, VA has been performing remote sensing research and development activities since the early 1970's. CDL estimates were provided in tabular format, with limited pictures or outputs to demonstrate the results. In 1997, the CDL program began to deliver geospatial content annually to stakeholders and now provides the visual product and associated meta-data five times a year.

Today, NASS utilizes imagery from an Indian AWiFS satellite. Images that meet quality guidelines are acquired by the USDA's Foreign Agricultural Service and stored in a centralized location. From here, it is utilized by several federal agencies based upon need. The coordination of federal agencies in building and using this image bank allows NASS to have access to a plethora of sensory data over many time periods at reduced cost. A commercial software suite is then used to produce the CDL products. Each software program has a specific purpose as outlined below.

- ERDAS Imagine – Imagery Preparation
- ESRI ArcGIS –Ground Truth Information Preparation
- See5 – Decision Tree Software Used for Image Classification
- SAS/IML Workshop - Acreage Estimation

4.1 Imagery Preparation and Ground Truth Information Preparation

A 740KM swath is acquired by the AWiFS satellite daily throughout the year and is stacked. The AWiFS satellite orbits the Earth every five days. Figure 7 shows one 340KM swath over the State of Kansas that will be stacked with others taken over the same area of land. Each pixel is 56 square meters (.77 acre). Pixels are shown in Figure 8, a small magnified selected area of Figure 7.

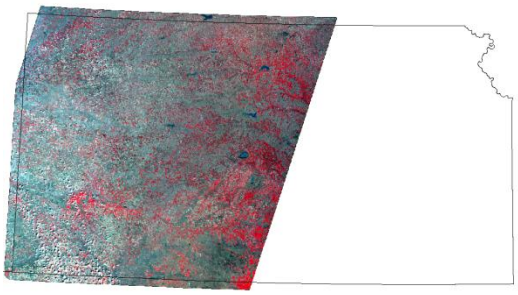


Figure 7: 340KM Swath over Kansas

Figure 8: Small Selected Area of Kansas.

In addition to the stack, the following layers of data are combined to form the CDL and the acreage estimate:

- Farm Service Agency (FSA)/Common Land Unit (CLU) data – Most farm operators sign up for farm program benefits. The FSA 578 data contain crop acreages for producers by the administrative county. FSA CLU's are the smallest land units with permanent boundaries, common land cover and management, common owner (tract) and producer association (farm). These data are used to train the program to assign crops to the unknown pixels.

- National Land Cover Dataset (NLCD) – Non-agricultural land such as water and wetlands is included in this dataset used for classification.

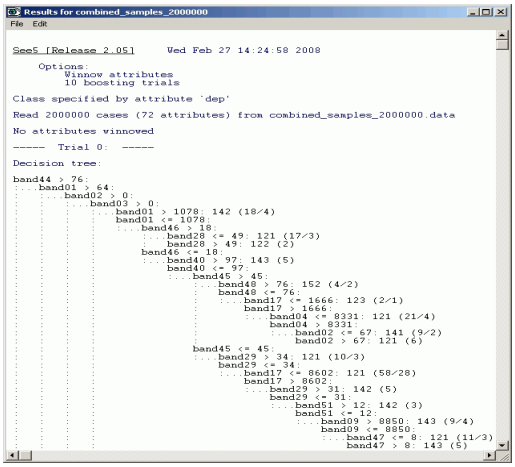
- Ancillary Data – Elevation, forest canopy, infrastructure, and other variables are also used to assign crops to unknown pixels.

- June Area Survey (JAS) data – NASS conducts a survey sampled from an Area Frame each June. This survey identifies all agricultural activity within the selected segments and collects information about crops, operator households, animals, grain storage facilities, and environmental factors. Data from the JAS are used for estimation.

4.2 Decision Tree Classification

The See5 decision tree software is the driver behind developing the CDL products. Decision trees are advantageous in that large data sets and missing data values can be handled. In addition to the advantages of using decision trees, See5, in particular, is also able to easily interface with ERDAS Imagine, due to a National Landcover Database (NLCD) Extension developed by the United States Geological Survey (USGS).

For each state, a random sample of pixels that can be matched with the ground truth is drawn. Every 56 meter square pixel drawn is classified through the decision tree program based on its spectral signature and its available classified and reported data. The decision tree results are used to create the CDL visual products using ERDAS Imagine. (See Figures 7 & 8).



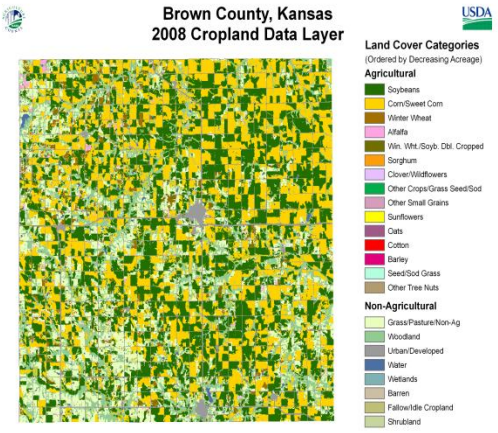


Figure 7: See5 Decision Tree Results

Figure 8: Map of Classified Cropland

SARS continually conducts pixel classification accuracy assessments. Analysts assess the producer's accuracy and the user's accuracy. The producer's accuracy is the probability that a ground truth pixel will be correctly mapped; this measures errors of omission (pixels are excluded from the correct category). The user's accuracy indicates

the probability that a pixel from the classification actually matches the ground truth data; this measures the errors of commission (pixels included in an incorrect category).

4.3 Acreage Estimation

After the CDL visual product has been created, it is filtered through ESRI ArcGIS to extract JAS intersecting pixels that will be used for acreage estimation. The remote sensing acreage estimator is more than a simple pixel count, it is a linear regression of the JAS reported acres on the CDL classified acres (See Figure 9). Essentially, this approach corrects the JAS sample estimate based on the relationship found between the reported data and classified pixels in each stratum where it is used. The regression adjusts the direct expansion estimate based on pixel information. Usually, this regression estimate leads to an estimate with a much lower variance than a direct expansion alone. If errors are found, they are corrected or that segment may be removed from consideration in the regression analysis. Segments that do not fit the linear relationship estimated by the regression are reviewed.

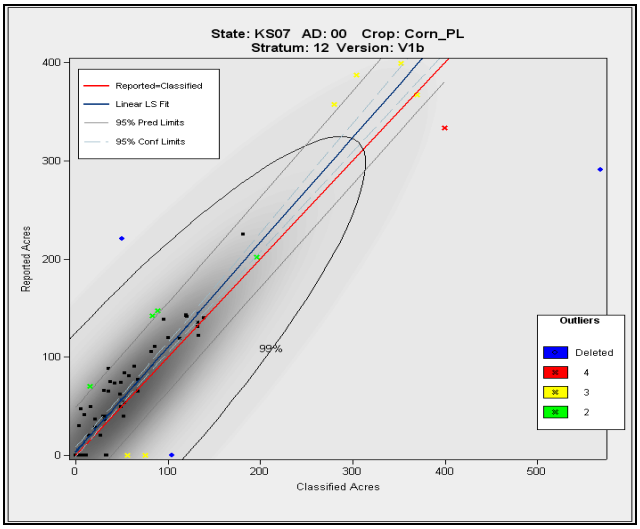


Figure 9: The graph above shows the approximately linear relationship between corn acres reported during the ground survey and acres classified to corn in the process of producing a CDL.

After the remote sensing (RS) acreage estimate is calculated using the CDL and JAS, NASS also gathers results of farmer reported data from surveys, Farm Service Agency data where available, agri-business data and the census of agriculture data as inputs into an analysis to set the official estimate. The official estimate is the single best number that NASS can derive from all of the available inputs.

The RS acreage estimate also performs well in comparison to the official estimate. An analysis of statistical relationships of all of NASS' estimates available for input has revealed that the RS acreage generally has a better correlation to the official state estimate than does the survey estimate. The RS acreage estimate performs almost as well and sometimes better than the FSA indication; however, this is expected since the indication of choice is the FSA estimate when it becomes available. On the county level, a majority of RS estimates are within 10% of the official estimates.

4.4 Cropland Data Layer Uses and Users

The CDL provides more than just an acreage estimate or literal lay-of-the-land for farmers and agri-businesses such as fertilizer companies. Water quality assessments, monitoring of watersheds, transportation, urban sprawl, deforestation, crop rotation patterns, wildlife habitat analysis, pesticide applications, and determination of crop stress locations are a few of the additional uses of the CDL. Agri-businesses; libraries; federal and government agencies; farm grower's associations; crop insurance, seed and fertilizer, and farm chemical companies; and universities all benefit from CDL data products. The CDL, along with related meta-data can be easily accessed and downloaded for FREE on the NASS Research and Development Division Website Homepage or Geospatial Data Gateway (See Appendix).

5. Future Seasons of Research Growth

The ability to collect data, find patterns, classify records, and extract information from large data sets is an integral part of statistical practices in large survey organizations that service the federal government. Data mining techniques and remote sensing are two new innovative approaches NASS uses to enhance its statistical practices. Data mining has extracted otherwise hidden information in existing datasets to improve NASS operations, and research in the area of remote sensing methodology by SARS has increased the ability to deliver geospatial content annually to stakeholders (See Appendix). Future applications of these techniques by NASS will be very effective in efforts to improve survey data collection, processing, estimation, and dissemination.

NASS' RDD will strive to continuously research and develop techniques and methodology such as data mining and remote sensing to remain the freshest and most thorough source of timely and accurate agricultural statistics.

References

Allen, Rich (2007) "Agriculture Counts - The Founding and Evolution of the National Agriculture Statistics Service 1957-2007" US Department of Agriculture, National Agricultural Statistics Service, http: // www.nass.usda.gov/About_Nass/index.asp

Beard, L. (2009) "Using Remote Sensing-Based Acreage Indications in NASS Operations: An Evaluation of CDL Acreage Indications," US Department of Agriculture, National Agricultural Statistics Service, RDD Presentation to the National Agricultural Statistics Service staff, Washington, D.C.

Boryan, C. (2009) "Remote Sensing of Agriculture: NASS' Cropland Data Layer Program," US Department of Agriculture, National Agricultural Statistics Service, RDD Presentation to George Mason University Digital Remote Sensing Seminar, Fairfax, VA.

Cecere, W. (2008) US Department of Agriculture, National Agricultural Statistics Service, RDD Report 08-in preparation, Fairfax, VA.

Cohen, S.B., DiGaetano, R., and Goksel, H. (1999) "Estimation Procedures in the 1996 Medical Expenditure Panel Survey Household Component," Agency for Health Care

Policy and Research, MEPS Methodology Report No. 5, AHCPR Publication No. 99-0027, Rockville, MD.

Earp, M., Cox, S., McDaniel, J., and Crouse, C. (2008) "Exploring Quarterly Agricultural Survey Questionnaire Version Reduction Scenarios," US Department of Agriculture, National Agricultural Statistics Service, RDD Report 08-11, Fairfax, VA.

Earp, M. and McCarthy, J. (2009) "Using Respondent Prediction Models to Improve Efficiency of Incentive Allocation," US Department of Agriculture, National Agricultural Statistics Service, Research and Development Division, Presentation at the 2009 American Association for Public Opinion Research Annual Conference, Hollywood, FL.

Garber, C. (2009) "Census Mail List Trimming using SAS Data Mining," US Department of Agriculture, National Agricultural Statistics Service, RDD Report 09-02, Fairfax, VA.

McCarthy, J. and Atkinson, D. (2008) "Innovative Uses of Data Mining Techniques in the Production of Official Statistics," US Department of Agriculture, National Agricultural Statistics Service, RDD, Paper for United Nations 2009 Statistical Commission Session on Innovations in Official Statistics.

McCarthy, J. and Earp, M. (in press) "Who Makes Mistakes? Using Data Mining Techniques to Analyze Reporting Errors in Total Acres Operated," Journal of Official Statistics.

McCarthy, J. and Jacob, T. (2009) US Department of Agriculture, National Agricultural Statistics Service, RDD Report 09-in preparation, Fairfax, VA.

McCarthy, J. and Jacob, T. (2009) "Who are You – Data Mining Approach for Predicting Non-respondents," US Department of Agriculture, National Agricultural Statistics Service, Research and Development Division, Presentation at the 2009 American Association for Public Opinion Research Annual Conference, Hollywood, FL.

McCarthy, J., Jacob, T. and McCracken, A. (2009) "Modeling NASS Survey Non-response using Classification Trees," US Department of Agriculture, National Agricultural Statistics Service, RDD Presentation to National Agricultural Statistics Service staff, Fairfax, VA.

Mueller, R. (2009) "NASS Remote Sensing Acreage Program for County Estimates," US Department of Agriculture, National Agricultural Statistics Service, RDD Presentation to National Agricultural Statistics Service staff, Washington, D.C.

Seffrin, B. (2009) "County Estimates from CDL," US Department of Agriculture, National Agricultural Statistics Service, RDD Presentation to National Agricultural Statistics Service staff, Washington, D.C.

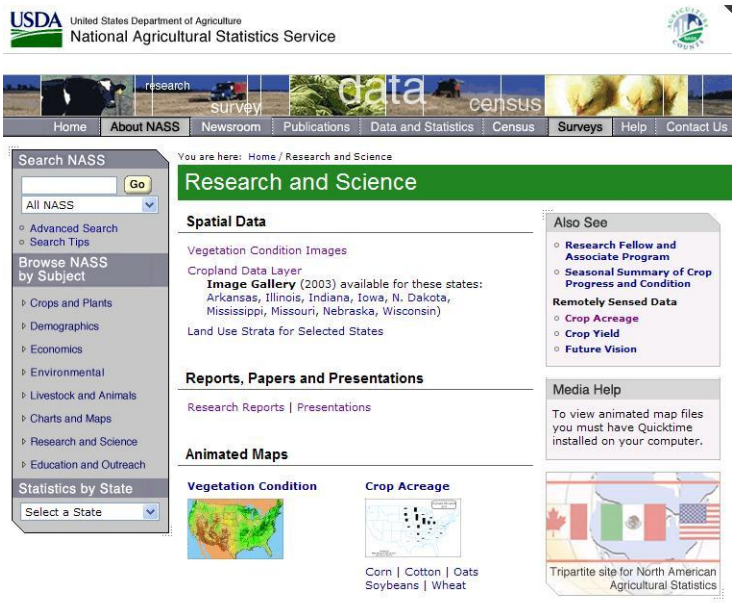
Wyland, J. (2008) "Agribusiness Grows with Crop-Specific Maps," ESRI Arc Watch, September 2008.

Unknown Author (2005) "An Evolving Statistical Science," US Department of Agriculture, National Agricultural Statistics Service,http://www.nass.usda.gov?About_NASS/index.asp

Appendix

NASS Research and Development Division Homepage

http://www.nass.usda.gov/Research_and_Science/index.asp



Geospatial Data Gateway

http://datagateway.nrcs.usda.gov