Remote sensing and US crop insurance program integrity: data mining satellite and agricultural data

B. Little[1], M. Schucking[1], B. Gartrell[1], B. Chen[1], S. Olson[1],
K. Ross[2], C. Jenkerson[3] & R. KcKellip[2]
[1]Texas Data Mining Research Institute (TDMRI) and Centre for Agribusiness Excellence (CAE), Tarleton State University, Stephenville, Texas, USA
[2]National Aeronautics and Space Administration (NASA) Stennis Space Centre, Stennis, Mississippi, USA
[3]Land Processes Distributed Active Archive Centre (LP DAAC), Sioux Falls, South Dakota, USA

Abstract

The objective of this investigation is to (1) integrate remote sensing data into an existing data warehouse of the US crop insurance program 1990 to 2007, (2) test remote sensing correlations with crop production, and (3) use remotely sensed time series data to assess variation in crop production. Previously (2000 to 2007) data mining of the data warehouse was based upon probabilistic and algorithmic approaches to identification of possible fraud, waste, or abuse. The value of adding satellite data warehouse to data mining resources is provision of (1) an additional empirical metric, (2) objective data on vegetative health, (3) measurable metrics for capturing the variance of plant health, and (4) a means for measuring the covariance of location and production. Refinement of data mining through the addition of satellite data for routine use to reduce fraud, waste, and abuse will ultimately diminish the frequency of false positives.
Keywords: GIS, NDVI, MODIS, data mining, satellite data, remote sensing.

WIT Transactions on Information and Communication Technologies, Vol 38, © 2007 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA070151

1 Data mining and detection of waste, fraud and abuse

In the United States federal government, data mining is being used in 199 programs in 2004 for several different areas: (1) intelligence/anti-terrorism (7%) (2) criminal analysis (7.5%), (3) human resources (8.5%), (4) scientific research (11.5%), (5) fraud, waste, and abuse (12%), and (6) improving service/ performance (32.5%) [1]. In the commercial sector data mining is used to detect fraud in the credit card industry, in banking, and other financial transactions such as insurance.

The United States Department of Agriculture (USDA) Risk Management Agency (RMA) administers the US crop insurance program through 16 private insurance companies that sell and service policies. The Federal Crop Insurance Corporation (FCIC) was created in 1938, and in 2007 insures more than 250 million acres, more than 140 crops, for more than 1.3 million producers, and has an estimated liability of $55 billion. More than 15,000 agents sell crop insurance in the US, and more than 7,500 crop insurance adjusters service policies on which a loss is claimed.

Since 2001, data mining the US crop insurance program for fraud, waste, and abuse at the Centre for Agribusiness Excellence (CAE) has resulted in cost avoidance of more than $450 million over a six year period, according to testimony before US House of Representatives Testimony, 2006, and US Senate, 2005. Cost avoidance was accomplished through identification of anomalies by data analytic algorithms and probabilistic models, with a high level of confidence ($P < 0.00001$). Anomalies identified through this method have an extremely high confidence level ($P < 0.0001$) of being anomalous.

The USDA Foreign Agriculture Service (FAS) has used satellite data to estimate crop production in countries other than the US for over a decade with reputedly high accuracy. USDA FAS estimates using remote sensing are held in such high regard that each US Embassy in other countries has an FAS attaché whose main mission is to monitor crop production. Remote sensing techniques for assessing crop production have not been systematically applied to US crop production or the US crop insurance program except on a case by case analysis, mainly for compliance and litigation functions. The exception is pasture-, forage-, and range-land biomass for the rangeland program, such as the USDA National Agricultural Statistics Service (NASS) crop land data layer.

For areas outside the US, more than 40 publications have reported correlations of satellite data with crop production in India, Pakistan, Brazil, China, and other non-US countries. It is generally accepted that remote imaging can easily distinguish between normal and catastrophic conditions. The ability to distinguish between normal and anomalous diminishes as differences decrease in magnitude. Robust detection of even subtle differences is needed for use of satellite data in the crop insurance program.

The research question is: Can NDVI distinguish between irrigated and non-irrigated when (1) aggregating to the county level, (2) including non-agricultural land, and (3) abnormally dry conditions exist throughout the growing season?

2 Methods and materials

The present report is preliminary analysis of agricultural practices in the U.S. using remotely sensed data. The pilot study was a partnership between CAE, USDA RMA, LP DAAC, and NASA. The present pilot investigation is designed to (1) analyze remotely sensed data variance in vegetative health in two counties (one mainly irrigated, one mainly non-irrigated) in the transition zone (Abnormally dry, D0) between normal and Moderate Drought, D1 drought conditions, and (2) assess the ability to detect "boundary effects" in pre-drought conditions (Abnormally dry, D0) when the signal is weakened by several sources of interference, and (3) compare with actual production for the two areas.

2.1 Data warehouse resources

The CAE data warehouse is currently 2.5 terabytes in a Teradata® System. Satellite data are pre-processed and currently stored in the Satellite Data Warehouse, a 42 terabyte RAID 5 system from which data may be post-processed to the appropriate level of resolution (aggregation) and joined to other tables as needed for analysis.

2.2 Data mining approach

In the present pilot study of integrating satellite data into the CAE data warehouse previously described approaches to data mining are applied, including but not limited to multivariate correlational and predictive analyses. The goal of data mining in the present study is to demonstrate the added value (augmented information) from the use of satellite data to identify anomalous crop insurance claims.

2.3 Spatial referencing

USDA RMA data contain a data field termed LAND ID that had a relatively high rate of errors during the early years of the CAE research project. Over the ensuing years, effort and emphasis was placed on correcting LAND ID errors. Contemporary RMA data on LAND ID are reliable.

Digitization in a GIS of each farm field in the US is underway, and is completed and certified for 30 of 50 states. The digitization process from 1-meter spatial resolution aerial photographs provides a georeferenced polygon for each field that is called a Common Land Unit (CLU). Software developed for this pilot application is scalable to CLU level by design.

2.4 Satellite data

Moderate Resolution Imaging Spectroradiometer (MODIS) data were retrieved from the NASA archive (LP DAAC) 250-meter spatial resolution available for 2000 to 2007 for the US. Importantly, the common perception of satellite data as an attractive multi-color image is NOT the target for the current analysis of satellite data. Individual pixels that comprise the ‘pictures’ are actually a digital signal of surface reflectance of vegetation, soil, or other land covers reflectance of chlorophyll molecules recorded by the satellite [2]. The digital signal for each pixel is the unit of analysis instead of the ‘picture.’

2.4.1 Satellite data: pre-processing

As part of a partnership between CAE, USDA RMA, and NASA, pre-processing algorithms were developed in collaboration with scientists at NASA Stennis Space Centre. Briefly, of 36 bands of satellite data available, a ratio of two bands, near infrared (NIR) and red (R): (NIR - R) / (NIR + R), were the target of the present pilot study. Additional bands were retained in the Satellite Data Warehouse for future use.

2.4.2 Satellite data: post-processing

Post-processing of satellite data was done to match the level of crop production data that were available. MODIS data (37,021 MODIS pixels in Lynn County, 41,639 MODIS pixels in Hale County) were available to 250 m^2 per pixel per time period but crop production data for this pilot study were only available at the producer level. Texas does not have PLSS data, therefore the crop production data and the MODIS data were aggregated to the county level. In states that utilize the PLSS system, MODIS data is aggregated to the section level (average of 1 square mile) in the PLSS system (section, range, township).

2.5 Joining satellite data and existing data warehouse

In the future, the georeferenced CLU, which is at the sub-farm unit level, will be used. This will allow a much more accurate determination of what is happening at the actual farm level as it greatly improves the ability to locate a specific farm. MODIS data can be incorporated by farm in this case.

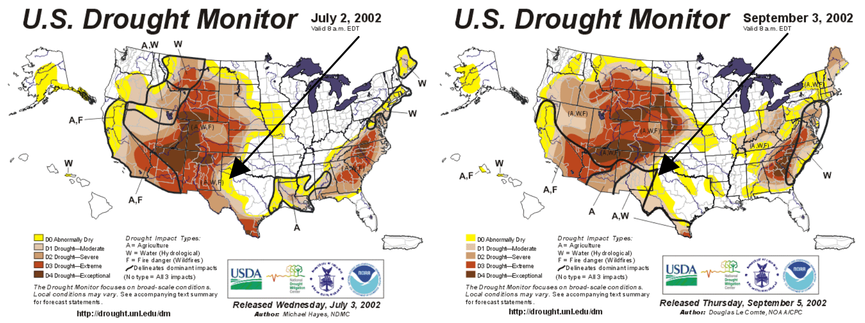
U.S. Drought Monitor
July 2, 2002
Valid 8 a.m. EDT
A,W
W
W
W
A,F
(A,W,F)
W
(A,W,F)
(A,W,F)
A,F
A
D0 Abnormally Dry
D1 Drought—Moderate
D2 Drought—Severe
D3 Drought—Extreme
D4 Drought—Exceptional
Drought Impact Types:
A = Agriculture
W = Water (Hydrological)
F = Fire danger (Wildfires)
Delineates dominant impacts
(No type = All 3 impacts)
The Drought Monitor focuses on broad-scale conditions. Local conditions may vary. See accompanying text summary for forecast statements.
http://drought.unl.edu/dm
USDA
National Drought Mitigation Center
NOAA
Released Wednesday, July 3, 2002
Author: Michael Hayes, NDMC
U.S. Drought Monitor
September 3, 2002
Valid 8 a.m. EDT
(A,W,F)
(A,W,F)
(A,W,F)
(A,W,F)
(A,W,F)
A,F
W
A,F
W
A
A,W
D0 Abnormally Dry
D1 Drought—Moderate
D2 Drought—Severe
D3 Drought—Extreme
D4 Drought—Exceptional
Drought Impact Types:
A = Agriculture
W = Water (Hydrological)
F = Fire danger (Wildfires)
Delineates dominant impacts
(No type = All 3 impacts)
The Drought Monitor focuses on broad-scale conditions. Local conditions may vary. See accompanying text summary for forecast statements.
http://drought.unl.edu/dm
USDA
National Drought Mitigation Center
NOAA
Released Thursday, September 5, 2002
Author: Douglas Le Comte, NOAA/CPC

Figure 1: USDA Drought Monitor in July and September 2002. Study area (indicated by →) is in the transition zone between white and the darker shades throughout the growing season.

2.6 Geographic unit of analysis and location

The location for the analysis was chosen because it remained in the “abnormally dry” boundary throughout the growing season in the 2002 drought (Figure 1), labelled the worst in 30 years. One of the more difficult problems is loss assessment in transition areas.

Two counties in west Texas (Hale and Lynn) were chosen, one in which the majority of acres were irrigated cotton, and another in which the majority of acres were non-irrigated cotton (Figure 2).

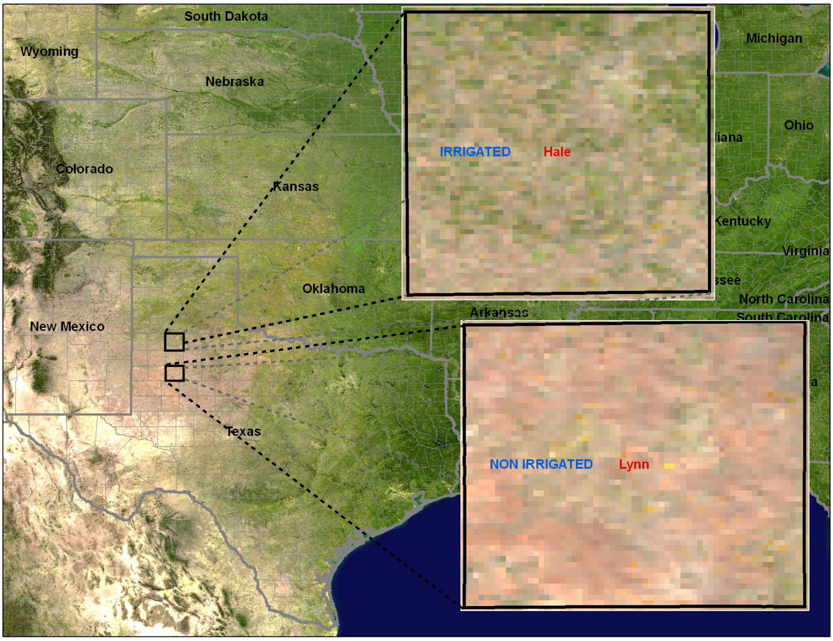

Figure 2: Counties in Texas Contrast: ~58.3% of Hale County land area (412,023 acres) is irrigated cotton agriculture (top), elevation 3200-3600 feet, 19.4” annual rainfall; and ~51.1% of Lynn County land area (570,835 acres) is dry land cotton agriculture (bottom), elevation 2881-3300 feet, 17.5” annual rainfall.

2.7 NDVI and analytical approach

Density of green vegetation on the Earth’s surface is measured by satellite sensors that record distinct wavelengths of visible and near-infrared sunlight that is reflected by plants, and by inference, absorbed. Calculating the ratio of the visible and near-infrared light reflected to the remote sensor results in a number that ranges from negative one (-1) to positive one (+1). This number is the

Normalized Difference Vegetation Index, or NDVI. NDVI values close to zero indicate no green vegetation, while values close to +1 (0.8 - 0.9) indicate very high density of green leaves [2].

2.8 Analytical and utility software

Data were drawn from the RMA and NASA data warehouses at CAE, and aggregated to the county level for high level analysis. SAS (SAS Institute, Cary, NC), SPSS (SPSS Inc., Chicago, Ill), ArcGIS (ESRI, Redlands, CA) were used in the analyses, in addition to tools developed for this project (See 2.4.).

3 Results: satellite data (NDVI) and agricultural production

Results of the pilot analysis indicate the MODIS satellite data is robust, and the signal is strong despite high level aggregation, inclusion of non-agricultural land, and being in a drought boundary or transition zone.

3.1 Pilot analysis

The pilot analysis is a comparison of MODIS NDVI between two geographically close counties in west Texas with predominately either irrigated or non-irrigated agricultural practices. The NDVI for the irrigated county was significantly different from the non-irrigated county early in the growing season, and reach their maximum in the last week of July 2002 and during August. By August the NDVI for irrigated was 37% higher (unadjusted values) than non-irrigated. Post-processed signal value differences between irrigated and non-irrigated are lower (11%), but still significant. The soils and temperatures were the same –water availability (irrigation) was the only differential, limiting factor. Production differences were significantly different despite excessive variation that masks differences in statistical analysis, typical of chaos in boundary transition areas.

3.2 Satellite data and agricultural production

In the past, transition areas have been difficult to analyze for losses. In the present analysis, irrigated NDVI was significantly different ($P < 0.01$) from non-irrigated NDVI even early in the growing season. Differences become apparent within 30 days of planting (Figure 3). Production was significantly different between irrigated and non-irrigated practices (Table 1), despite high variance.

3.3 Satellite data and anomalous agricultural production

NDVI signals analyzed in the present investigation range from nearly zero to more than 0.40. NDVI regressed on time indicate high direct relationship with irrigated ($R^2 = 0.98$) and non-irrigated ($R^2 = 0.97$) during the growing season (Figure 3).

Table 1: Cotton production.

	Hale (irrigated) County			Lynn (non-irrigated) County			
2002	n acres	Mean*	SD	n acres	Mean*	SD	P
Mean in Pounds	237,448	890.2	326.7	218,652	265.6	95.1	0.001
Irrigated	208,000	842	--	76,000	594	--	
Non-irrigated	16,700	209	--	201,000	241	--	
% Irrigated	256,700	94.8%	--	292,000	27.4%	--	0.01

*per acre

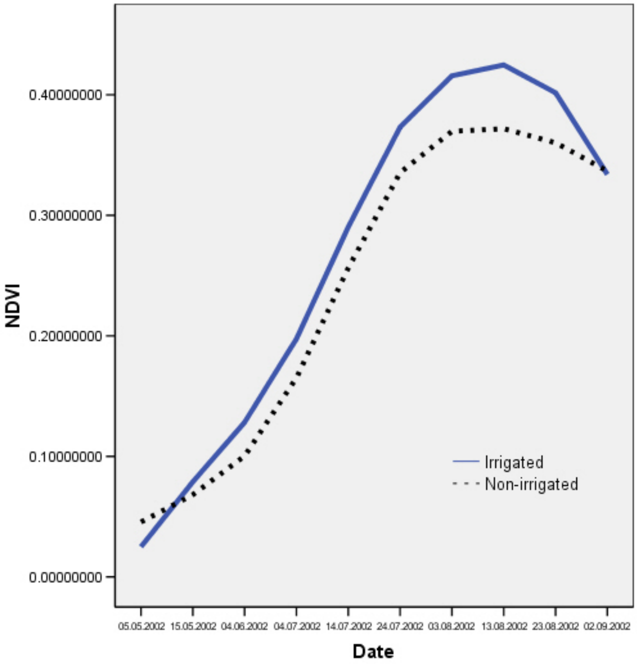

Figure 3: MODIS NDVI from May 5 2002 to September 2 2002 by Irrigated vs. Non-irrigated (P < 0.01) for Two Counties in West Texas (.95 CI Range: ± 0.01 to 0.05). Note convergence of NDVI lines when irrigation ceases.

4 Discussion

The results indicate that NDVI with a moderate resolution remote sensor signal such as MODIS can distinguish between irrigated and non-irrigated counties, suggesting that remote sensing may be useful in the US crop insurance program for routine use comparing NDVI within and between CLUs over the growing season. Localised catastrophes could be verified, and variance-covariance analyzed components.

In prior investigations, NDVI has a moderate-to-high accuracy in forecasting crop production [3–11], but boundary areas (transition from drought to normal conditions) have not been analyzed. In the present study, the NDVI is stable in the transition zone, and strong enough to detect statistically significant differences in plant growth (irrigated vs. non-irrigated), even early in the plant growth cycle (Figure 3). Notably, production was statistically significantly different ($P < 0.001$) between irrigation and non-irrigated groups (Table 1).

The important application of remotely sensed data will occur with the integration of the CLU. The algorithms developed in partnership with NASA are scalable to allow use of: (1) higher resolution satellite data, and/or (2) specific geo-referenced polygons such as the CLU. Identification of specific farm units by GIS will allow temporal assessment of plant growth using the NDVI. The NDVI is added to the RMA data warehouse to strengthen the capability of data mining to discover anomalies with high sensitivity and specificity that suggest waste, fraud, or abuse. Satellite data can complement existing techniques and procedures. False positives can be reduced through analyses of concordance between probabilistic models and data analytic algorithms, and remotely sensed data; the model is fine-tuned by algorithmic subtraction of non-agricultural areas.

In summary, integration of NDVI and other satellite data into data mining with CLUs provides (1) another empirical metric, (2) data on vegetative health, (3) metrics on the variance of plant health, and (4) measurement of covariance of location and production. False positives can be reduced using remote sensing and GIS referenced land area, with CLUs being a key parameter. Further analyses of the relationship between NDVI and claims are underway.

References

[1] US Government Accountability Office (GAO), Data Mining. Covering a wide Range of Uses. Government Printing Office. Report 04-0548, 2004.

[2] Xiao, X.: Light Absorption by Leaf Chlorophyll and Maximum Light Use Efficiency. IEEE Transactions on Geoscience and Remote Sensing 44 (7): 1933-1935, 2006.

[3] Dutta, S., S.A. Sharma, A.P. Khera, Ajai, M. Yadav, R.S. Hooda, K.E. Mothikumar, M.L. Manchanda: Accuracy assessment in cotton acreage estimation using Indian remote sensing satellite data. ISPRS Journal of Photogrammetry and Remote Sensing 49 (6): 21-26, 1994.

[4] Hall-Beyer, M.: Comparison of Single-Year and Multiyear NDVI Time Series Principal Components in Cold Temperate Biomes. IEEE Transactions on Geoscience and Remote Sensing 41 (11): 2568-2574, 2003.

[5] Hartmann, T., C.D. Bella, P. Oricchio: Assessment of the possible drought impact on farm production in the SE of the province of Buenos Aires,

Argentina. ISPRS Journal of Photogrammetry & Remote Sensing 57: 281-288, 2003.

[6] Lobell, D.B., G.P. Asner: Cropland distributions from temporal unmixing of MODIS data. Remote Sensing of Environment 93: 412-422, 2004.

[7] Panigraphy, S., S.A. Sharma: Mapping of crop rotation using multidate Indian Remote Sensing Satellite digital data. ISPRS Journal of Photogrammetry & Remote Sensing 52: 85-91, 1997.

[8] Rizzi, R., B.F.T. Rudorff, Y.E. Shimabukuro, P.C. Doraiswamy: Assessment of MODIS LAI retrievals over soybean crop in Southern Brazil. International Journal of Remote Sensing 27 (10): 4091-4100, 2006.

[9] Sedano, F., P. Gong, M. Ferrao: Land cover assessment with MODIS imagery in southern African Miombo ecosystems. Remote Sensing of Environment 98: 429-441, 2005.

[10] Xavier, A. Candido, B.F.T. Rudorff, Y.E. Shimabukuro, L.M.S. Berka, Mauricio A. Moreira: Multi-temporal analysis of MODIS data to classify sugarcane crop. International Journal of Remote Sensing 27 (4): 755-768, 2006.

[11] Xiao, X., S. Boles, J. Liu, D. Zhuang, S. Frolking, C. Li, W. Salas, B. Moore III: Mapping paddy rice agriculture in southern China using multi-temporal MODIS images. Remote Sensing of Environment 95: 480-492, 2005.