# Using Data Mining to Detect Anomalous Producer Behavior: An Analysis of Soybean Production and the Federal Crop Insurance Program

**Stacey A. Olson**
Email: olson@tarleton.edu

**Bertis B. Little**
Email: little@tarleton.edu

**Ashley C. Lovell**
Email: lovell@tarleton.edu

**Center for Agribusiness Excellence**
**Tarleton State University**
Box T-0055
Stephenville, TX 76401

Phone Number: (254) 918-7676
Fax Number: (254) 918-7686

***Selected Paper prepared for presentation at the Southern Agricultural Economics Association Annual Meeting, Mobile, Alabama, February 1-5, 2003***

**Using Data Mining to Detect Anomalous Producer Behavior: An Analysis of Soybean Production and the Federal Crop Insurance Program**

## Introduction

United States agricultural policy has involved crop insurance programs since the era of the Great Depression and the Dust Bowl. Initially federal crop insurance was designed and implemented to help farmers overcome prevented planting losses and to minimize risk due to low prices. However in contemporary times, it is unclear whether the producer's primary benefit of the federal crop insurance program is risk reduction or income enhancement. Rather than using crop insurance to manage risk as intended, a limited number of producers have exploited the program. In April 2002, the Deputy Administrator for Compliance estimated that 25 percent of all claims in the federal crop insurance program contain some element of fraud, waste, and abuse. Overall, the crop insurance program provides positive gains for participating producers. In reinsurance year 2000, insured producers paid $2.54 billion in premium for $34.4 billion of coverage or liability and received $2.58 billion of indemnities for crop losses (RMAOnline).

Fraud, waste, and abuse are a continuing concern within the federal crop insurance program. Congress in 2000 passed legislation to expand and strengthen the program to address concerns regarding fraud, waste, and abuse. The Agricultural Risk Protection Act (ARPA) of 2000 was a major vote of confidence for the federal crop insurance program, and included $8.2 billion in funding over five years to enhance and improve the federal crop insurance program.

ARPA requires the USDA Risk Management Agency (RMA) to conduct program wide efforts (compliance, education, and insurance services) to prevent fraud, waste, and abuse. Growth in the size and complexity of the crop insurance program makes prevention and

detection of fraud, waste, and abuse increasingly more challenging. The lack of agronomic homogeneity across crops and across regions confounds identification of fraud, waste, or abuse. A major step toward preventing fraud, waste, and abuse is the development of tools to identify agricultural producers, insurance agents, loss adjusters, insurance companies, crops, and geographic areas that deviate from the norms established by the Risk Management Agency (H.R. 2559).

ARPA specifically states that RMA is to utilize data mining technology to reduce program vulnerabilities and waste, fraud, and abuse. The specific objective of this study is to develop a data-mining algorithm and to apply the algorithm to identify anomalous producers and counties within Land Resource Regions (LRR) based upon the percentage of acres harvested.

This research is the first to use data mining to identify anomalous county and producer behavior by mining data from the USDA/RMA Book of Business data warehouse at the Center for Agribusiness Excellence (CAE) and data from the Natural Resources Conservation Service (NRCS). The NRCS LRRs will be used to group spatially insured producers into agronomical homogeneous groups to account for the natural resource variability. However, weather variation was not addressed due to limitations regarding the scope of the project.

This paper proceeds with an overview of the data and methods. The following section is a discussion of the results with two subsections regarding anomalous producers and anomalous counties. The conclusion section suggests opportunities for expanding this frontier research.

**Data and Methods**

This study focused on U.S. soybean acres and includes irrigated and non-irrigated practices. Limiting the study to one crop facilitated the application of data mining to detect anomalous patterns of behavior in agricultural insurance. Data for this study are from USDA's RMA policy database available, through the cooperative agreement between Tarleton State University and USDA's Federal Crop Insurance Corporation (FCIC) (01-IE-0821-020). Land Resource Region (LRR) data are available through the NRCS.

Variables included in the data mart are: state and county code, LRR, reinsurance year, crop code, practice code, acres planted, acres harvested, liability, indemnity, producer premium, and risk premium. The original dataset contained 625,031 unique producers and over 2.58 million observations from reinsurance years 1994 to 2001.

The dependent variable for identifying anomalous behavior was the percentage of acres harvested, computed as the ratio of harvested acres to planted acres at the individual producer level. The percentage of acres harvested was selected because it was a reliable indication of anomalous behavior based upon prior investigation of the data warehouse. An exceptionally low percentage of acres harvested does not alone suggest that crop insurance fraud occurred. It does suggest that further investigation may be warranted including the evaluation of absolute indemnity, liability, and premium amounts and associated ratios.

Prior to computing the percent of acres harvested, producer-level planted acres and harvested acres were determined. Planted acres were calculated by summing all insured soybean acres for each unique producer and subtracting any acreage claimed as prevented planting. Harvested acres were calculated as the planted acres minus any acreage claimed as unharvested and/or ungleaned. If no claims (unharvested or ungleaned) were filed on the

planted acres it was assumed that the acres were 100% harvested. It is important to note that harvested acres may include claimed acres associated with lowered quality and/or quantity of production due to drought, heat, hail, frost, freeze, flood, wind, insects, mycotoxin, and plant disease. However, this study only identified anomalous harvested acres.

The producer data mart included only producers who planted soybeans three or more years between 1994 and 2001 to establish a reliable harvest history. The data mart was comprised of 2.15 million records. Five-year moving averages were computed within state, county, crop, and practice (irrigated and non-irrigated) from 1994 to 2001. The smoothing procedure was applied to de-emphasize spikes and emphasize trends for percent of acres harvested. When computing the five-year moving average, a grouping system was used to compensate for missing data. Producers may not have planted soybeans every year; thus it was inappropriate to compare them to producers who had planted all five years. Five groups of producers were identified during the moving-average process. The group to which a producer was assigned depended upon the number of years the producer planted within the time period. In addition to the grouping process, another criteria was established to handle missing data. A producer's percent of acres harvested had to exist for the last year within the five-year moving average period, e.g. if a producer did not plant in 1998, no moving-average percent was recorded in the time period ending in 1998. However, if that producer planted in 1999, a moving average was recorded for 1999.

Following the smoothing, the percentage was normalized by z-score within LRR: $(z = (\chi_i - \bar{x})/\sigma)$.

LRRs were used because they identify natural resource characteristics as compared to risk rating regions or administrative districts. LRRs do not conform to county lines. If a

county was divided among two or more LRRs, the county was assigned to the LRR that included the largest proportion of the county. Risk rating regions usually are designated by their crop yield capability. Boundaries are changed and new regions are established, as the RMA Regional Office deems necessary. Administrative districts are currently used by the National Agricultural Statistics Service (NASS), but documentation describing the criteria used to set up the administrative districts was unavailable.

After normalization, an outlier detection method was used to identify producers with anomalous harvest behavior. A *t*-value was calculated for each producer within the five groups discussed above. Only producers within the same group and LRR were compared. Producers were identified as anomalous if they were at or below the $5^{th}$ percentile and had a $p \leq 0.05$. Both assumptions had to be satisfied for the producer to be flagged. The same procedure was applied to the first percentile and $p \leq 0.01$.

Each county was regarded as a super producer. All producer level data, including those excluded from the smoothing process with one or two years of data between 1994 and 2001, were aggregated to the county level. The same outlier detection procedure was applied to the county level data mart to identify anomalous counties. Counties were flagged at two levels, $5^{th}$ percentile and $p \leq 0.05$, and $1^{st}$ percentile and $p \leq 0.01$.

**Results**

Federal law prohibits this study from releasing any information that may specifically identify a specific producer. All data presented here are aggregated either to the state or LRR level if a limited number of observations within the county may allow the identification of a specific producer.

Anomalous Producers:

Following normalization of the producer data mart using z-scores, a *t*-value was calculated for each producer in each of the five groups (see methods). The outlier detection method flagged 29,567 producer observations, 2.6 percent of final data mart, resulting in 14,621 unique producers [i] at $p \leq 0.05$. At $p \leq 0.01$, there were 13,999 producer observations, 1.2 percent of final data mart, resulting in 7,786 unique producers. Approximately 27 percent of flagged producers were identified as anomalous in three or more years, $p \leq 0.05$.

In 2001, 7402 unique producers at $p \leq 0.05$ were identified as anomalous accounting for 2.2 percent of the 334,481 producers who insured soybeans. The flagged producers accounted for 2.8 percent of the total premiums for all insured soybeans and 1.6 percent of the total liability. Yet, they received a disproportionate 6.1 percent of the total soybean indemnities in 2001 (Table 1).

The loss ratio (indemnity/premium) and the loss cost ratio (indemnity/liability) are frequently used indicators in crop insurance analysis. These ratios further substantiate the anomalous nature of the flagged producers. Table 2 shows the loss ratios and loss cost ratios for 2001. Outliers at $p \leq 0.05$ have a loss ratio that is over twice as large as all insured soybean producers nationally. The loss cost ratio is also indicative of the outliers' claim records being inconsistent with the national book of business.

Anomalous Counties:

The same outlier-detection method used for producers was used for counties. Following normalization of the county data mart using z-scores, a *t*-value was calculated for each county within the five groups as described above. The outlier detection method flagged

404 county observations (4.7 percent of final data mart) at $p \leq 0.05$ and 132 county observations (1.5 percent of final data mart) at $p \leq 0.01$.

The maps show county outliers in LRR by practice and reinsurance year at $p \leq 0.05$ and $p \leq 0.01$. Visual inspection indicates that the majority of the outliers fall near LRR boundaries. Transitional soils near the boundaries may have caused the counties to be flagged. However, further investigation and additional research is necessary to test the hypothesis suggested by this anecdotal observation.

Without regard to the group designation and practice, approximately 30 percent of the counties at $p \leq 0.05$ were identified as anomalous for three or more years using the outlier-detection method (Figure 1). Even among the small number of soybean producing counties identified as anomalous three or more years, clustering occurs along LRR boundaries. Further investigation appears warranted to evaluate possible vulnerability of crop insurance programs to LRR variation. Counties identified as anomalous three or more years should be further scrutinized to determine why the percent harvested is so unusual. Adjustments in insurance program provisions for these counties may need to be considered.

**Conclusion**

Identification of producers and counties as anomalous is not *prima facie* evidence of fraud. However, the outlier detection of anomalous producers and anomalous counties leads to some interesting questions because both sets of outliers cluster at or near LRR boundaries. Although no publications were located that correlated production and LRR interfaces, it is agronomically intuitive that yield variations likely increase at soil transition zones. Additional investigation and research should examine and identify improved production practices that are appropriate at or near transitional soils. Future investigation of these patterns should consider

weather variation; for example it is an empirical observation that weather patterns in Iowa are different from those in Texas or Oklahoma, for example. Also, variation in the current investigation was dampened by the use of: (1) moving averages, (2) normalization, and (3) aggregation. Therefore, the results of the current investigation should be considered preliminary and not conclusive, but do identify an area rich with interesting research questions. Strategies may need to be devised specifically for soybean producers to adapt farming practices to accommodate the variation in plant performance in the transitional soil zones.

Producers identified as anomalous may not be committing fraud, but program waste may be occurring. The loss ratios and loss cost ratios for the outlier producers exceed the national book of business across the board and are resulting in a drain on taxpayer dollars. Analysis of insurance provisions and possible adjustments, including county T-yields along LRR boundaries, may deserve consideration.

Further analysis of agricultural production variation along LRR boundaries is needed. Additional research at CAE has been directed toward determining whether or not anomalous producers/anomalous counties along LRR boundaries would still be flagged in adjacent regions, in situations where one or more LRRs are represented within the county. In that analysis since LRRs do not conform to county lines, counties were assigned to the LRR with the smallest proportion of LRR. The reassignment of LRRs to counties did not significantly change the results. The same frequency distributions for flagged producers and flagged counties occurred after the change in LRR assignment. Furthermore, the same clustering effect along LRRs occurred in visual inspection of the outliers.

**Footnotes**

[i] A unique producer is determined by a distinct tax identification (social security number or employer identification number), tax id type code (type of tax id), and entity code (for of business organization).

**Table 1.** Total Premium, Indemnity, and Liability for U.S. and Anomalous Soybean Producers, 2001

| | U.S. Totals, Insured Soybeans | Outliers $p \leq 0.05$ | % of U.S. Total | Outliers $p \leq 0.01$ | % of U.S. Total |
|---|---|---|---|---|---|
| Producers | 334,481 | 7402 | 2.2 | 3169 | 1.1 |
| Premium | $ 509,317,484 | $ 14,370,174 | 2.8 | $ 4,523,429 | 0.9 |
| Indemnity | $ 317,216,531 | $ 19,206,378 | 6.1 | $ 7,715,366 | 2.4 |
| Liability | $6,986,028,281 | $111,460,031 | 1.6 | $32,573,703 | 0.5 |

**Table 2.** Loss and Loss Cost Ratios for 2001

| | U.S. Insured Soybeans | Soybean Outliers $p \leq 0.05$ | Soybean Outliers $p \leq 0.01$ |
|---|---|---|---|
| Loss Ratio[a] | 0.62 | 1.34 | 1.70 |
| Loss Cost Ratio[b] | 0.05 | 0.17 | 0.24 |

[a] Loss Ratio = Indemnity / Premium.
[b] Loss Cost Ratio = Indemnity / Liability.

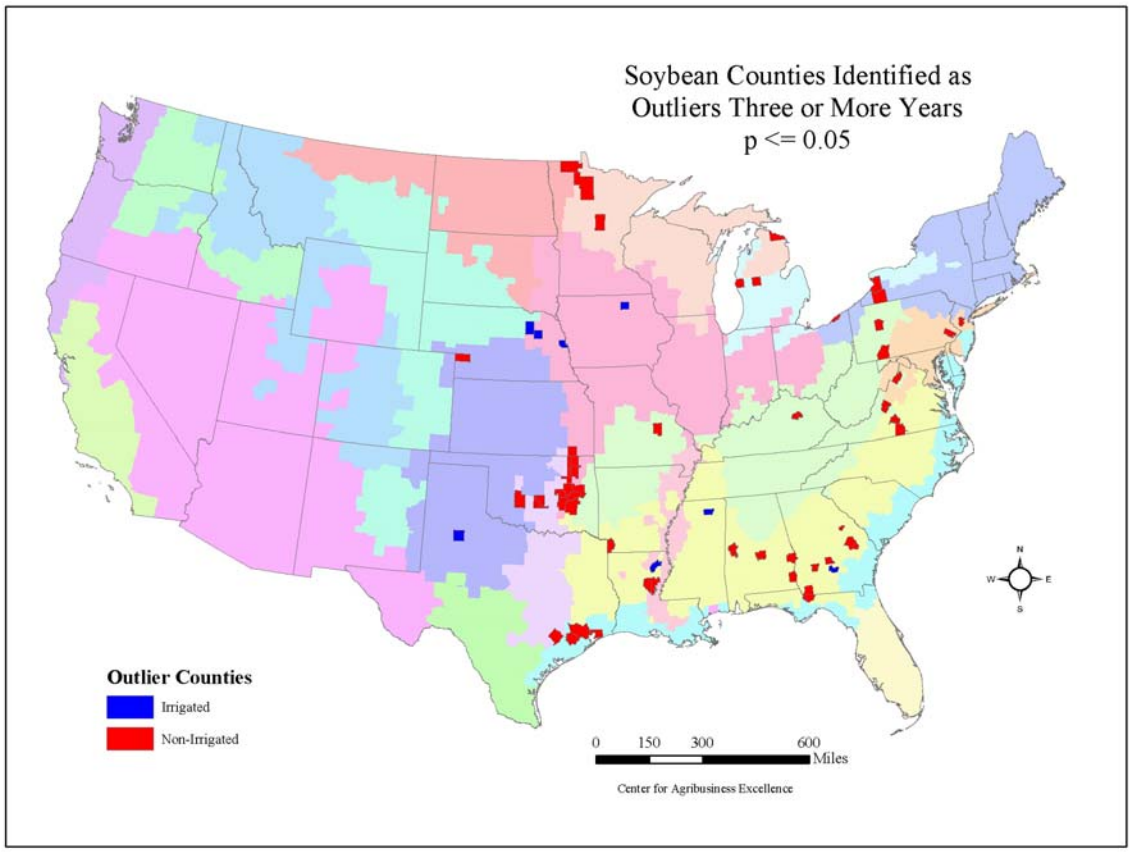


**Figure 1.** Soybean Counties Identified as Outliers Three or More Years, $p \leq 0.05$

**References**


Brockett, P.L., X. Xia, and R.A. Derrig. "Using Kohonen's Self-Organizing Feature Map to Uncover Automobile Bodily Injury Claims Fraud." *The J. of Risk and Insurance.* 65(June 1998): 245-274.

Crop Insurance Industry. *Combating Fraud, Waste, and Abuse in the Crop Insurance Program.* A report by the Crop Insurance Industry, January 2001. In: http://www.amag.com/library/pdf/fraudreport_jan01.pdf (Last accessed: October 24, 2001).

Federal Crop Insurance Corporation. *Loss Adjustment Manual (LAM) – 2001 and Succeeding CropYears.* FCIC-25010, Washington DC, Feb. 2001.

Goodwin, B.K. and J. Deal. *The Federal Crop Insurance Program: An Empirical Analysis of Regional Differences In Acreage Response and Participation.* Selected paper at the 2001 AAEA Meetings, Chicago, IL, August 6-8, 2001.

Katsaras, N., P. Wolfson, J. Kinsey, and B. Seanuer. "Data Mining: A Segmentation Analysis of U.S. Grocery Shoppers." Working Paper 01-01, Retail Food Industry Center, University of Minnesota, 2001.

Knutson, R.D., J. Penn, and B. Flinchbaugh. *Agricultural and Food Policy,* 4th. ed. New Jersey: Prentice-Hall, Inc., 1998.

Little, B.B., W. Johnston, A. Lovell, S. Steed, V. O'Conner, G. Westmoreland, and D. Stonecypher. "Data Mining U.S. Corn Fields." In *The Proceedings of the First Society of Industrial and Applied Mathematics (SIAM) International Conference on Data Mining,* Issue 1, Chicago IL, 2001: 99-104.

Rejesus, R.M., B. Little, W. Johnston, A. Lovell, and S. Steed. *Data Mining U.S. Corn Fields: The Application of a Tool to Detect Anomalous Farmer Behavior in the U.S. Crop Insurance Program.* Selected paper presentation at the 2002 SAEA Meetings in Orlando, FL (February 3-6, 2002).

RMAOnline. *A History of the Crop Insurance Program*. A report by the Risk Management Agency/ U.S. Department of Agriculture. In: http://www.rma.usda.gov/aboutrma/history.html (Last accessed: Feb. 2, 2002).

Smith, V.H. and B. Goodwin. "Crop Insurance, Moral Hazard, and Agricultural Chemical Use." *American Journal of Agricultural Economics* 81(May 1996): 428-38.

U.S. Congress, House of Representatives. *Agricultural Risk Protection Act of 2000.* Washington DC: H.R. 2559, 106th Cong., 2nd sess., 24 January 2000.