

Collusion in The U.S. Crop Insurance Program: Applied Data Mining

Bertis B. Little, Walter L. Johnston, Jr., Ashley C. Lovell, Roderick M. Rejesus, and Steve A. Steed

Center for Agribusiness Excellence
Tarleton State University
Stephenville, Texas

Abstract

This paper quantitatively analyzes indicators of Agent (policy seller), Adjuster (indemnity claim adjuster), Producer (policy purchaser/holder) indemnity behavior suggestive of collusion in the United States Department of Agriculture (USDA) Risk Management Agency (RMA) national crop insurance program. According to guidance from the federal law and using six indicator variables of indemnity behavior, those entities equal to or exceeding 150% of the county mean (computed using a simple jackknife procedure) on all entity-relevant indicators were flagged as “anomalous.” Log linear analysis was used to test (1) hierarchical node-node arrangements and (2) a non-recursive model of node information sharing. Chi-square distributed deviance statistic identified the optimal log linear model. The results of the applied data mining technique used here suggest that the non-recursive triplet and agent-producer doublet collusion probabilistically accounts for the greatest proportion of waste, fraud, and abuse in the federal crop insurance program. Triplet and agent-producer doublets need detailed investigation for possible collusion. Hence, this data mining technique provided a high level of confidence when 24 million records were quantitatively analyzed for possible fraud, waste, or other abuse of the crop insurance program administered by the USDA RMA. This data mining technique can be applied where vast amounts of data are available to detect patterns of collusion or conspiracy as may be of interest to the criminal justice or intelligence agencies.

Keywords: Log linear models, non-recursive, triplets, doublets, collusion, insurance fraud

Collusion in The U.S. Crop Insurance Program: Applied Data Mining

Bertis B. Little, Walter L. Johnston, Jr., Ashley C. Lovell, Roderick M. Rejesus, and Steve A. Steed

**Center for Agribusiness Excellence
Tarleton State University
Stephenville, Texas**

Introduction

The U.S. Congress is concerned that the United States Department of Agriculture (USDA) Risk Management Agency (RMA) crop insurance programs are vulnerable to waste, fraud, and abuse. Accordingly, the Agriculture Risk Protection Act (ARPA) of 2000 was passed. In that legislation it was stated that “data warehouse and data mining technologies should be employed to improve program compliance and integrity.” The first action mandated by ARPA, following construction of a data warehouse, was to identify crop insurance Agents (sellers of policies) and Adjusters (crop insurance indemnity claims adjusters) whose business activities were anomalous. The law further mandated that those to be considered anomalous were $\geq 150\%$ of the county area mean losses. The ARPA of 2000, Subtitle B, Section 121, (f) (1) (A) & (B) directs RMA, or its designees, to detect disparate performance of Agents and Adjusters at the 150% of the mean for all loss claims in the same area.

During the process of engineering the data warehouse and conducting preliminary data mining, our research team was briefed by USDA RMA Compliance investigators. These investigators indicated that many of the cases they had successfully investigated were configured as a classic conspiracy. The investigators characterized the conspiracy with a ‘cart wheel’ metaphor. The investigators’ observations subjectively indicated that crop insurance conspiracies were founded on the principle of linked actions from a central group of conspirators (cart wheel hub) through a spiraling network of conspiracy operatives (spokes in the wheel) relayed for action to many performing players (the rim). RMA Compliance Investigators further observed that Agents were probably the ‘hub’ of the cart wheel. Adjusters were probably the ‘spokes’ of the cart wheel, connecting Agents and Producers in the conspiracy model of crop insurance waste, fraud, and abuse. Finally, Producers were posited to be the ‘rim’ of the cart wheel, completing the common model of conspiracy in which collusion of few is connected to actions of many by some intermediary.

Typically, investigation of conspiracy originates from a qualitative analysis of individual instances. Such efforts are time consuming and identifying instances of fraud rely largely upon luck rather than objective analyses. In the current analysis, data mining was applied to the USDA RMA ‘national book of insurance business’ for the year 2000 to quantify conspiracy investigation. The working hypothesis was the ‘cart wheel metaphor’ of linked anomalous business transactions at the Agent, Adjuster, and Producer levels was performed seeking evidence for linked behaviors suggestive of collusion.

Data Mining and Insurance Fraud: A Review

The use of data mining tools and techniques for insurance fraud detection has been clearly documented in the literature. Yeo (2000), for example, documented its usefulness when he described a U.S. health insurance company using data mining techniques to compare an individual doctor’s claims against a larger historical base of data. The analysis identified several geographical areas where claims exceed the norm. The insurance firm

investigated and confirmed that a physician in an area was submitting false bills. As a result, the doctor was forced to pay restitution and fines. Data mining saved this health insurance company as much as \$4 million.

Grossman *et al.* (1999) have indicated that fraud detection is one of the areas where data mining is considered a “successful” tool. He *et al.* (1996) showed how data mining through neural networks could be used to detect medical fraud. Several recent publications have documented the usefulness of data mining techniques for fraud detection in health and automobile insurance claims (Johnson, 1997; Williams and Huang, 1997; Manchur, 1998; Panko, 1999; Fox, 2000; and Yeo, 2000). Most of these papers mention the large amount of potential savings in human investigative resources and indemnity payments when using data mining tools.

The literature indicates that there are two main data mining techniques used in insurance fraud detection. A predominant data mining technique used in the detection of fraudulent claims are straightforward statistical techniques. Weisberg and Derrig (1993) have used linear regression models to analyze the statistical significance of fraud indicators in bodily injury claims in Massachusetts. Other investigators developed a model and an expert system that can detect potentially fraudulent claims based on a probit model using a number of fraud indicators (Belhadji and Dionne, 1997; Belhadji, Dionne, and Tarkhani, 2000).

The second main data mining technique used in insurance fraud detection is clustering algorithms. Derrig and Oztaszewski (1995) used fuzzy set theories to classify and identify characteristics of fraudulent claims. They found that fuzzy techniques could efficiently classify claims based on the suspected fraudulent content. Brockett, Xia, and Derrig (1998), on the other hand, apply neural networks and back-propagation algorithms to cluster claims based on the degree of fraud suspicion. They demonstrated that this data mining technique performs better than the combination of an insurance adjuster’s fraud assessment and an insurance investigator’s fraud assessment with respect to consistency and reliability.

In summary, data mining techniques have been shown to be a methodology capable of helping detect insurance fraud in large databases. Thus, given its wide use in other insurance markets, it has potential as a tool to detect fraud, waste, and abuse in the U.S. crop insurance market where it has not been widely used. The detection of fraud, waste, and abuse in the U.S. crop insurance market is more important and challenging today because of the enormous growth and changes in the federal crop insurance program over the past few years. Therefore, data mining techniques can be seen as an efficient tool that can identify anomalous behavior of farmers participating in the crop insurance program, which may be worthy of further human investigation. This paper demonstrates how an applied data mining technique could be used in analyzing anomalous behavior suggestive of collusion in the crop insurance industry.

Methodology

Materials

A data warehouse was engineered and populated with RMA data for the decade from reinsurance years 1991 to 2000 using the Teradata System by NCR. In this pilot investigation we analyzed data from the year 2000 only. There were 157,180,000 acres insured under 1.002 million producers’ policies, sold by 13,434 crop insurance agents with \$2.49 billion in indemnities adjusted by 3,842 adjusters. Total liability was \$27.17 billion, and the total premium was \$2.26 billion. This represented the entire book of business for the USDA Risk Management Agency in year 2000. CAT (catastrophic) policy indemnities were excluded because they would bias the analysis

Hypotheses

The null hypothesis that no linkages existed between agent, adjuster, and producer nodes was tested using a log linear model (Fienberg, 1977; Gilbert, 1981; McCullagh and Nelder, 1983). The first three alternative hypotheses tested classic conspiracy linkages with node doublets (Agent-Adjuster-Producer, Agent-Producer-Adjuster, Adjuster-Agent-Producer) linked. The Agent-Adjuster-Producer represents the ‘cart wheel’ hypothesis discussed in the **Introduction**. In the fourth alternative, we tested the hypothesis that all three nodes were linked to one another non-recursively (Gilbert, 1981).

Table 1. Hypotheses Tested under the Collusion Model

Null Hypothesis:

H₀: No statistically significant relationship exists between the agent, adjuster, and producer nodes in indemnity payments.

Linked Doublet Hypotheses (Fig. 1):

H_{A1}: A statistically significant relationship exists linking agent to adjuster to producer nodes.*

H_{A2}: A statistically significant relationship exists linking agent to producer to adjuster nodes.

H_{A3}: A statistically significant relationship exists linking adjuster to agent to producer nodes.

Triplet Ring Hypothesis (Fig. 2):

H_{A4}: A statistically significant non-recursive relationships exist linking all nodes to all nodes: agent is linked to producer and adjuster nodes, producer is linked to agent and adjuster nodes, and adjuster is linked to producer and agent nodes.

* Cart Wheel Conspiracy Hypothesis

Methods

Flags that indicated unusual behavior in crop insurance claim premium, indemnity, and liability were computed using six derived measures indicating loss (Table 2). In accordance with the ARPA, county means were computed using the generalized jackknife procedure (Davison and Hinkley, 1997; Gray and Schucany, 1972). The generalized jackknife is used for the purpose of reducing bias in estimating means and confidence intervals, and its reliability is widely demonstrated (Schucany et al., 1971; Tukey, 1958).

Table 2. Derived Measures (Means Calculated by Jackknife Method within County)

| Derived Measure | Level |
|--|---------------------------|
| Ratio 1 = \$ Indemnity / \$ Premium | Adjuster, Agent, Producer |
| Ratio 2 = \$ Indemnity / \$ Liability | Adjuster, Agent, Producer |
| Ratio 3 = # Loss Policies / Total # Sold | Agent |
| Ratio 4 = # Loss Units / Total # Units Insured | Agent |
| Ratio 5 = \$ Adjuster _i / \$ County Indemnity | Adjuster |
| Ratio 6 = # Claims for Adjuster _i / Total County Claims | Adjuster |

Agents, Adjusters, and Producers who were $\geq 150\%$ of the county mean for a loss indicator were flagged on the derived measures (flagged = 1, not flagged = 0). Only entities (Agent, Adjuster, Producer) that were flagged on all indicator variables (Table 2) were classified as anomalous. Data manipulations were performed in SAS (SAS Institute, Cary, NC). The statistical program *R* was used to perform the log linear analysis (Ihaka and Gentleman, 1996).

Log linear analysis is appropriate to this research problem because it provides a flexible mechanism for the analysis of Poisson distributed count data (i.e., number flagged vs. number not flagged). The name *log linear* comes from the use of a logarithmic transform to convert a multiplicative model into a linear one. For example, the standard null hypothesis in Pearson's χ^2 test for a 2 dimensional contingency table tests for independence of row and column classifications and is given as:

$$H_0: P(r_i, c_j) = P(r_i) * P(c_j).$$

Taking the log of H_0 gives:

$$H_0: \log\{P(r_i, c_j)\} = \log\{P(r_i)\} + \log\{P(c_j)\}$$

Hence the label *log linear*.

The true power of the technique is the ability to apply standard linear modeling techniques such as continuous as well as discrete covariates and the ability to specify a more complex model using a combination of discrete covariates, continuous covariates, and interactions between covariates. The log linear model results presented in this paper were performed using the `glm()` function of R to solve the Generalized Linear Model specified and the resulting measure of model adequacy, or goodness of fit, is the Deviance which is a $-2 \log$ likelihood measure and, therefore, χ^2 distributed. A detailed explanation of Generalized Linear Models is published (McCullagh and Nelder, 1983; Nelder and Wedderburn, 1972). Pearson correlation coefficient between entity pairings is the the correlation based upon data for both variables grouped in classes or categories (i.e., "flagged" vs. "not flagged")

Table 3. Frequencies of Agents, Adjusters, and Producers with Positive Indemnities

| <u>Entity</u> | <u>Anomalous</u> | <u>Not Anomalous</u> | <u>Total</u> |
|-------------------------|------------------|----------------------|--------------|
| Producers | 33,630 | 143,245 | 176,875 |
| Agents | 4,203 | 9,231 | 13,434 |
| Adjusters | 1,128 | 2,716 | 3,844 |
| <u>Doublets</u> | | | |
| Agent-Producer | 7,971 | 174,259 | 182,230 |
| Agent-Adjuster | 1,067 | 29,840 | 30,907 |
| Adjuster-Producer | 2,636 | 196,492 | 199,128 |
| <u>Triplet</u> | | | |
| Adjuster-Agent-Producer | 1,135 | 199,713 | 200,848 |

Total number of policies = 1,002,409; Percent with Indemnity = 26.8% (268,158/1,002,409)

Results

Log linear analysis of anomalous entities (Agents, Adjusters, and Producers) was employed to test various hypotheses of paths of association, not causation. Table 4 (a) is the model suggested by investigators who work in the crop insurance program, Agent-Adjuster-Producer. It was a statistically significant model for the observations. A model that is a statistically better fit is the Agent-Producer-Adjuster linkage (Table 4 (b)). However, of the three

hierarchical models possible in the classic conspiracy analysis, Adjuster-Agent-Producer is the optimal statistical fit (Best Fit) for the data. Thus, the associations shown in Figure 1 are not precisely those suggested by conspiracy investigators who work in the crop insurance program, but do follow the classically described conspiracy model. The path in Figure 1, Adjuster-Agent-Producer, was the log linear best fit for the hierarchical association model.

Table 4. Log Linear Analysis of Conspiracy Doublets (Agent-Adjuster, Producer-Agent, Producer-Adjuster) and Triplet

| <u>Model</u> | <u>Deviance</u> | <u>d.f.</u> | |
|--|-----------------|-------------|----------|
| Null | 484,957.7 | 7 | |
| <u>Linked by an Intermediary</u> | | | |
| (a) Agent-Adjuster-Producer | 7,501.9 | 2 | |
| (b) Agent-Producer-Adjuster | 1,790.5 | 2 | |
| (c) Adjuster-Agent-Producer (Figure 1) | 1,230.1 | 2 | Best Fit |
| <u>All Linked-No Intermediaries</u> | | | |
| (d) Triplet (Figure 2b) | 102.9 | 1 | Best Fit |
| <u>Doublets (Figure 3)</u> | | | |
| (e) Adjuster-Producer | 10,216.0 | 3 | |
| (f) Adjuster-Agent | 9,655.5 | 3 | |
| (g) Agent – Producer | 3,944.1 | 3 | Best Fit |

All results are significant at $P < 0.0001$. Statistic is Chi-square distributed $-2 \text{ Log Likelihood}$

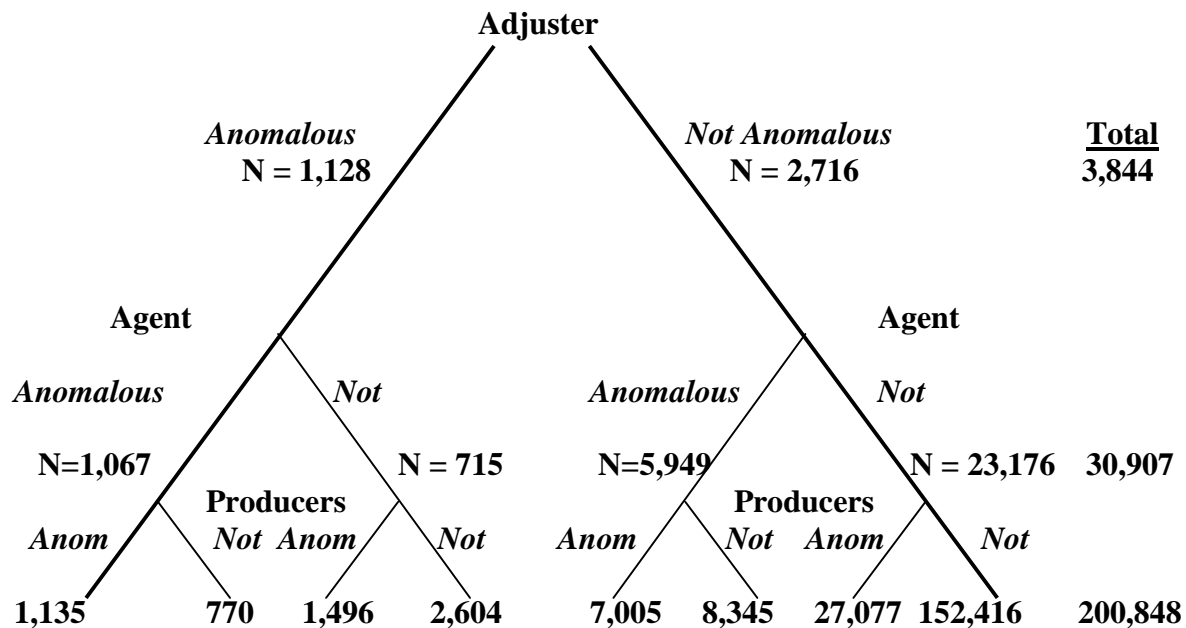


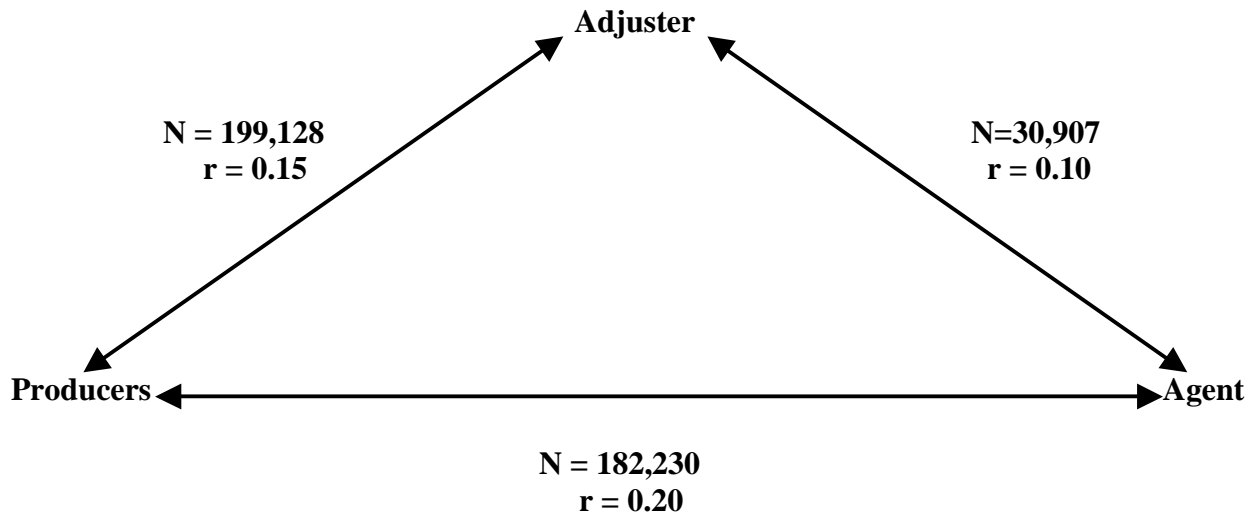
Figure 1. Hierarchical Model of Anomalous Loss Ratio Behavior at $\geq 150\%$ of County Mean in the Crop Insurance Program FY 2000

The best possible fit with the three entities, Adjuster-Agents-Producers, was a non-recursive triplet (Table 4 (d) and Figure 2). The non-recursive triplet indicates that there is not a true hierarchy as in a classic conspiracy, but there is significant sharing of information among all three entities. The strongest relationship is between Producers and Agents in the entire data set, whether or not they are anomalous ($r = 0.20$, Figure 2 (a)). This indicates an R^2 of 4% from the correlation of $r = 0.20$. The best overall fit for the model was as shown in Figure 2 (b) (Table 4 (d)). The second best fit was the Agent-Producer doublet. The biserial correlation of all Agent-Producers was, of course, at 0.20.

These findings are consistent with actual cases of collusion that has been litigated (Rose and Freivogel, 1995). In a particular case in the Texas High Plains, the agent was the central conspirator who submitted inflated claims in behalf of the producers for whom he wrote insurance. The producers involved sometimes knew of the deception, sometimes they did not. These claims were then falsely verified by unscrupulous adjusters who were paid by the collusive agent. Therefore, the pattern of collusion identified by our applied data mining procedure is in line with actual fraud behavior that has been observed in the past.

The geographic distribution of anomalous in Figure 4 (a) Producer, (b) Agent, and (c) Adjuster was also analyzed. This shows that most flagged producers and agents are located in the middle of the country, while flagged adjusters are more geographically dispersed. Similarly, geographic distribution of flagged doublets in Figure 5 (a) Adjuster-Producer (b) Agent-Adjuster, and (c) Agent-Producer revealed no obvious clustering of these anomalous pairs. Although majority of the flagged Agent-Producer pairs are in the Midwest and the Great Plains region. Finally, the flagged triplet (i.e., Adjuster-Agent-Producer) indicated no apparent visual geographic clusters (Figure 6). The absence of geographic clusters suggested that the jackknife procedure had adequately normalized means computed within counties.

(a) All Claims with Positive Indemnities and Point Biserial Correlations



(b) Linked Triplets of Anomalous Indemnities

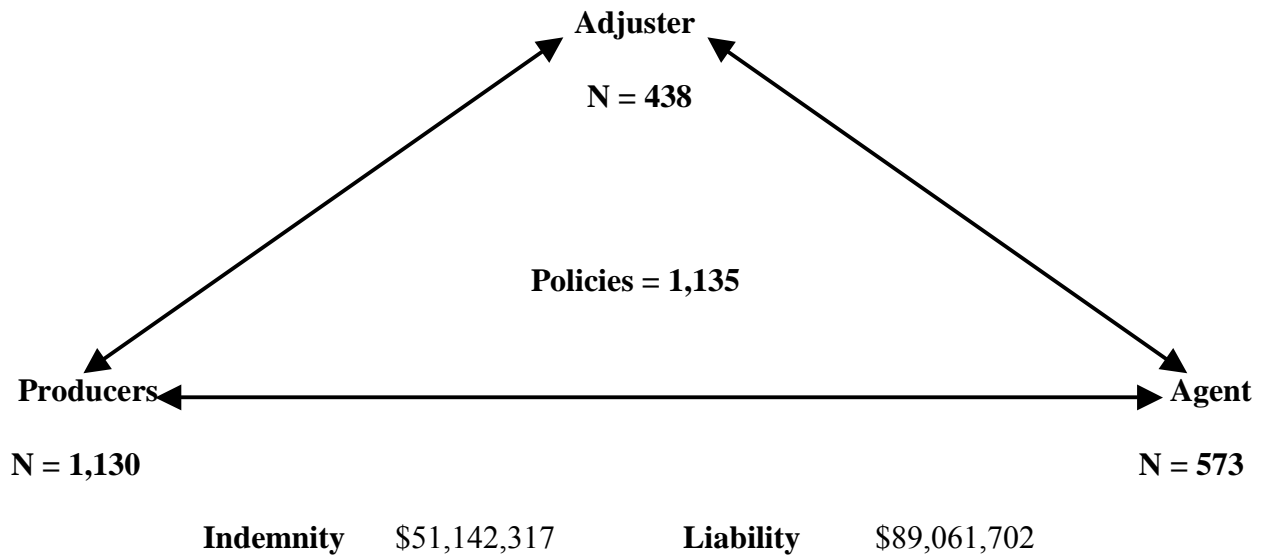


Figure 2. Non-Recursive Ring Triplet Analysis of Crop Insurance Collusion

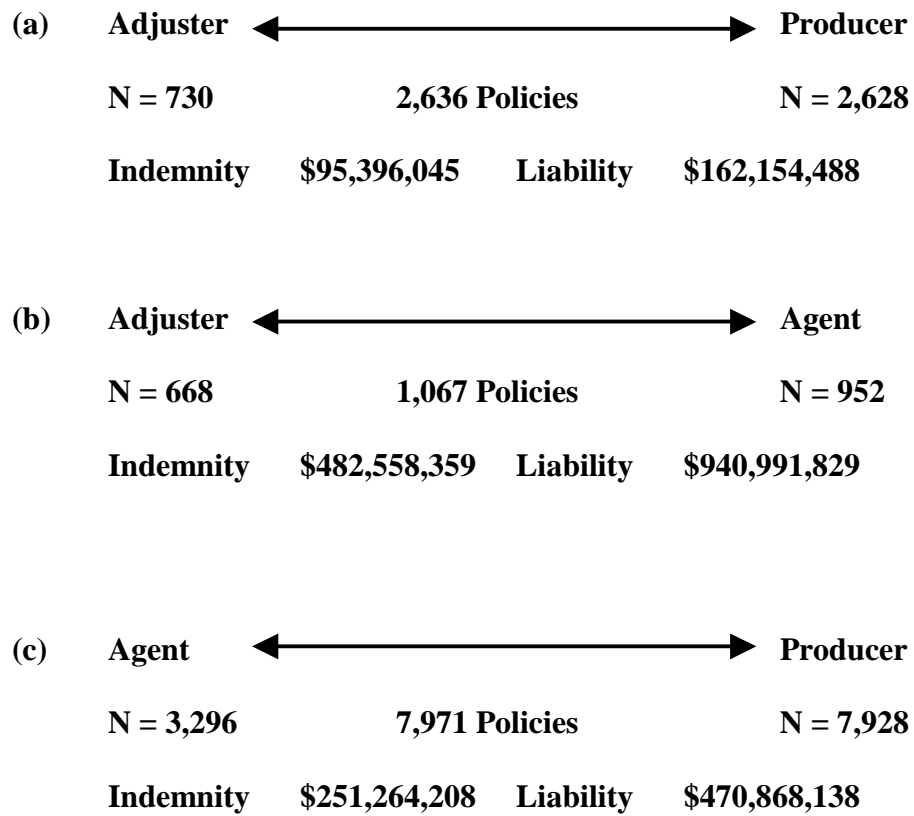


Figure 3. Doublet Links between Unique Pairings of Agent, Adjusters, and Producers

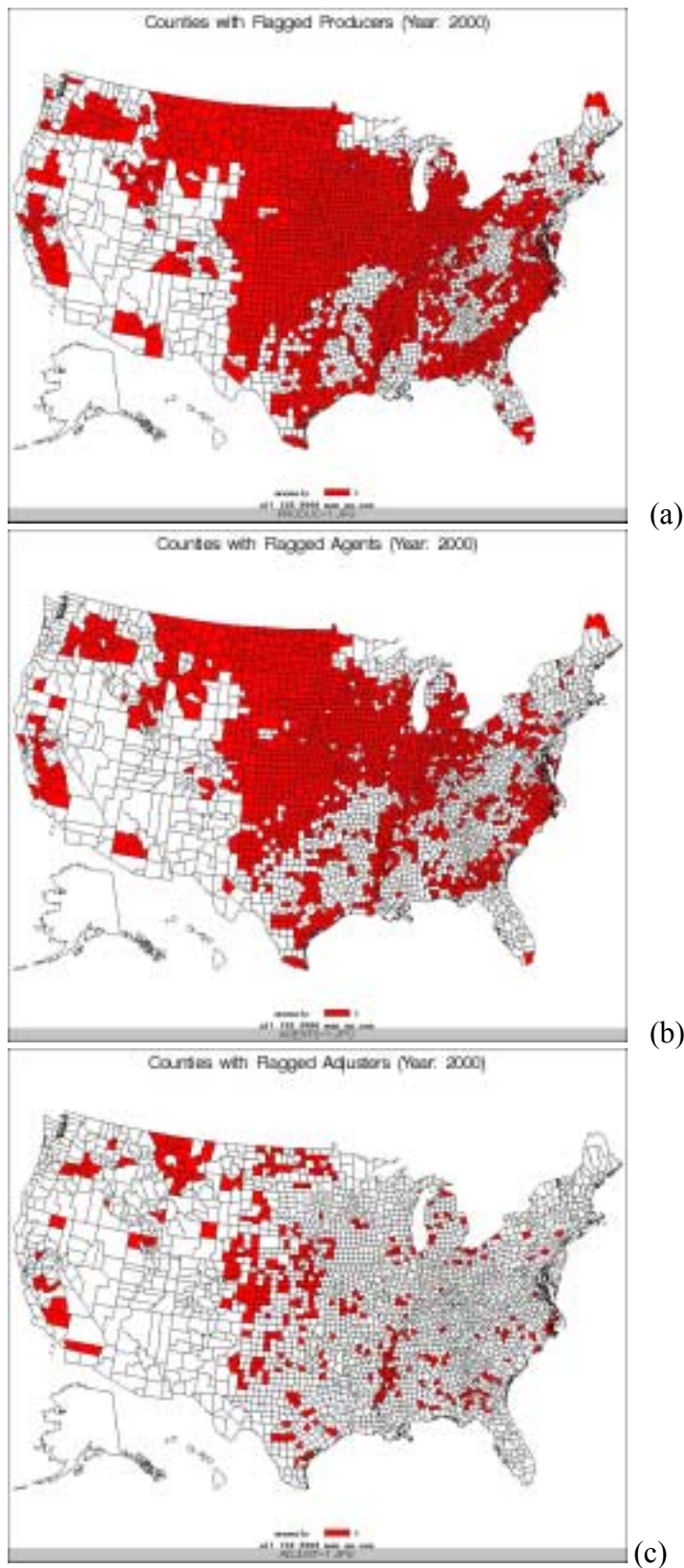


Figure 4. Geographic Distributions of “Flagged” Entities: (a) Producers, (b) Agents, and (c) Adjusters.

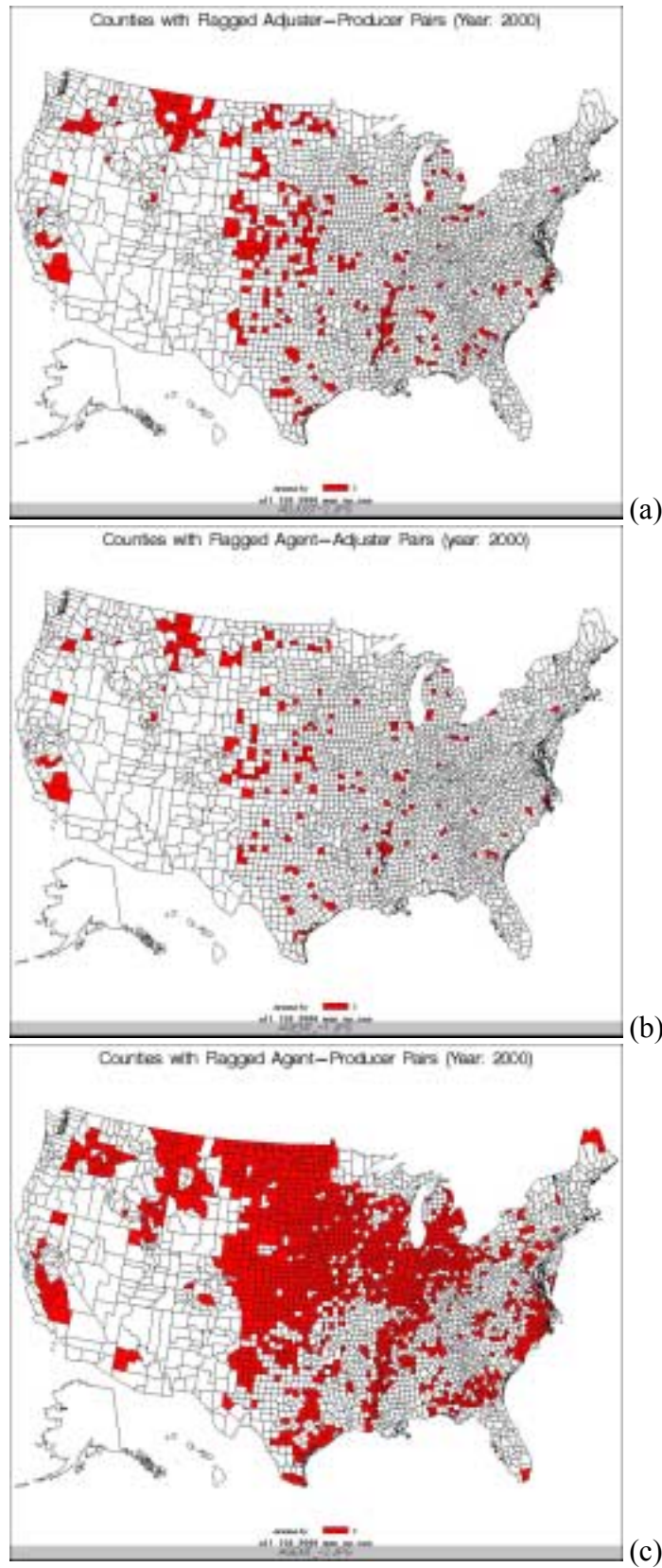


Figure 5. Flagged Doublets: (a) Adjuster-Producer (b) Agent-Adjuster, and (c) Agent-Producer

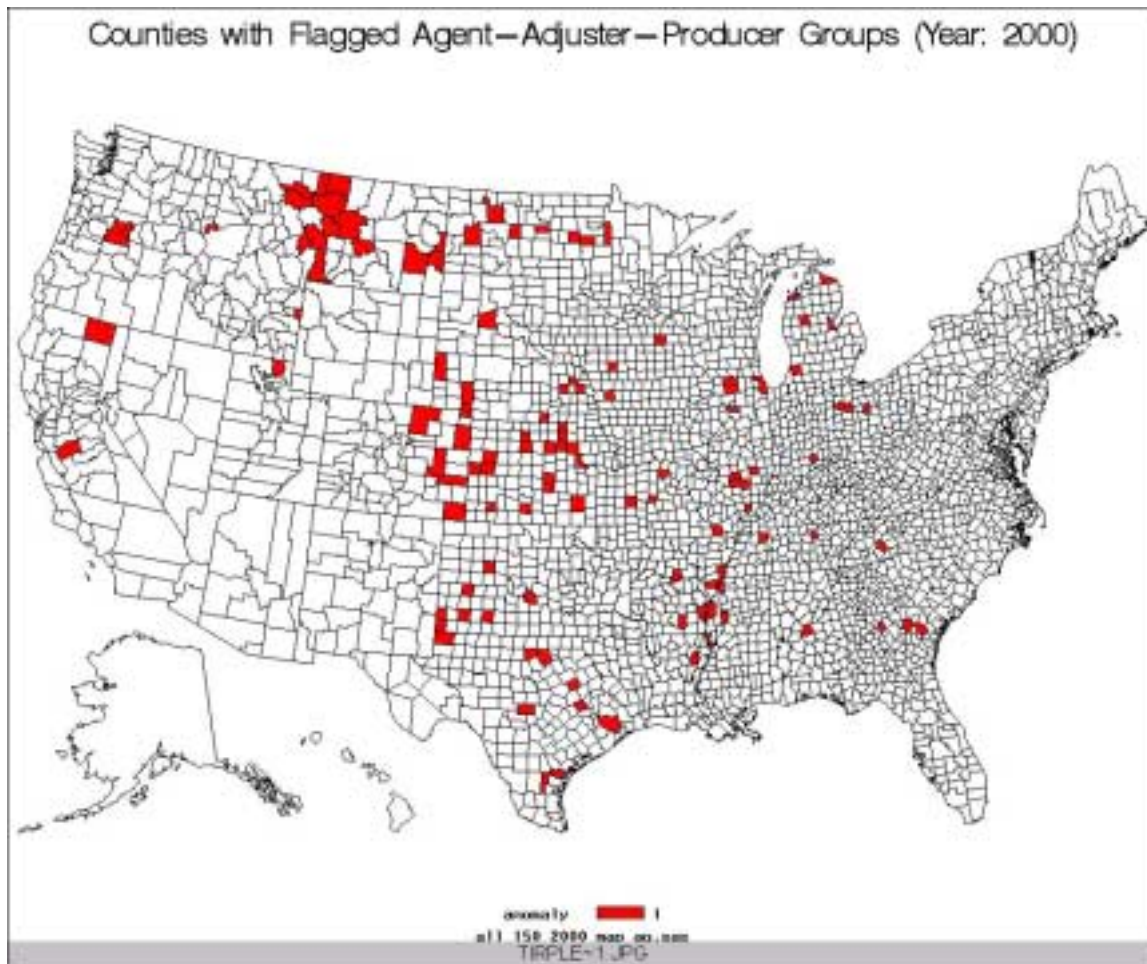


Figure 6. Adjuster-Agent-Producer Triplets Flagged

Discussion and Conclusions

Linkage of the triplets and a specific doublet (Agent-Producer) indicated need for detailed investigation for possible collusion among two or more of the entities. This provides a high level of confidence when screening cases for investigation.

Table 5. Best Statistical Fit Models

| <u>Model</u> | <u>Rank</u> | <u>Deviance</u> |
|--|-------------|-----------------|
| Triplet (No intermediaries) | 1 | 102.9 |
| Adjuster-Agent-Producer (Hierarchical) | 2 | 1,230.1 |
| Agent-Producer (Doublet) | 3 | 3,994.1 |

The next step in the screening process for investigation in this specific application of data mining is to further analyze the Indemnity-Liability (I-L) ratio for the linked triplets, hierarchical and non-recursive, and the Agent-Producer doublet as shown in Table 5. In this example of using log linear models to establish links (possible collusion) between nodes, the number of probable entities for investigation was reduced from more than 200,000 to 1,135. This is with the highest level of confidence afforded under these statistical models. It is recommended that I-L ratios which approach 1.0 will be prioritized for investigation.

These results were constrained by the federal mandate to analyze and report on 150% of the county mean. Further constraining the model by more tightly defining the outlier (e.g., those at or above the 95th or 99th percentile) in the models described in Table 5 would greatly enhance yield during investigation.

In summary, the analytical results presented here are unique because they demonstrate that collusion or conspiracy can be investigated through automated techniques (i.e., data mining) (Bay and Pazzani, 2001; Cabena et al., 1998; Westphal and Blaxton, 1998; Weiss and Idurkha, 1998; Witten and Frank, 2000). The results further indicate that nearly any type of data may be analyzed for evidence of conspiracy or collusion because the actual data analyzed was not metric, and had no statistical assumptions regarding adherence to a specific distribution or other parameters. These data were simply frequency counts of events. These “events” could be counts of e-mails, words, combinations of words, business transactions, or an entire litany of things that normally require human intervention for analysis. When events number in the hundreds of thousands, millions, billions, or even trillions, the log linear technique of linkage can automate the detection of linkages that may point to collusion, conspiracy, or terroristic plots.

Bibliography

- Bay SD, Pazzani MJ: Detecting group differences: Mining contrast set. *Data Mining and Knowledge Discovery* 5:213-246, 2001.
- Belhadji, D.B. and G. Dionne. "Development of an Expert System for the Automatic Detection of Automobile Insurance Fraud." Working Paper 97-06, Risk Management Chair, HEC-Montreal, 1997.
- Belhadji, D.B., G. Dionne, and F. Tarkhani. "A Model for Detection of Insurance Fraud." *The Geneva Papers on Risk and Insurance*. 25(October 2000): 517-538.
- 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.
- Cabena P, Hadjinian P, Stadler, R, Verhees J, Zanasi A: *Discovering Data Mining*. Prentice Hall: Upper Saddle River, NJ, 1998.
- Davison AC, Hinkley DV: *Bootstrap Methods and Their Applications*. Cambridge University Press: Cambridge, 1997.
- Derrig, R.A. and K.M. Ostaszewski. "Fuzzy Techniques of Pattern Recognition in Risk and Claim Classification." *The J. of Risk and Insurance*. 62(Sept. 1995): 447-482.
- Fienberg SE: *The Analysis of Cross-Classified Categorical Data*. The MIT Press: Cambridge, Mass., 1977.
- Fox, B.R. "Technology: The New Weapon in the War on Insurance Fraud." *Defense Counsel Journal*, 67(April 2000): 237-244.
- Gilbert GN: *Modelling Society: An Introduction to Loglinear Analysis for Social Researchers*. George Allen and Unwin: London, UK, 1981.
- Gray HL, Schucany WR: *The Generalized Jackknife Statistic*. Marcel Dekker, Inc.: New York, 1972.
- Grossman, R., S. Kasif, R. Moore, D. Roche, and J. Ullman. 1999. *Data Mining Research: Opportunities and Challenge*. A report of three NSF workshops on Mining Large, Massive and Distributed Data. In: <http://www.ncdm.uic.edu/dmr-v8-4-52.htm> (Last Accessed: September 28, 2001).
- He, H., J. Wang, W. Graco. and S. Hawkins. "Application of Neural Networks to Detection of Medical Fraud." *Exp. Sys. with Applic.* 13(1996): 329-336.
- Ihaka R, Gentleman R: R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 5:299-314, 1996.
- Johnson, R.A. "Digging the Dirt." *Best's Rev.: Property/Casualty-Insurance-Edition*. 98(1997): 108-110.
- Manchur, D. "Mining for Fraud." *Canadian Insurance*. 103(Sept. 1998): 24-27.
- McCullagh P, Nelder JA: *Generalized Linear Models*. 2nd Ed. Chapman and Hall, 1983.
- Nelder JA, Wedderburn RWM: Generalized linear models. *Journal of the Royal Statistical Society A* 135:370-384, 1972.
- Panko, R. "Getting a Jump on Crime." *Best's Rev.(Property/Casualty)*. 100(Oct. 1999): 73-75.

Rose LJ and WH Freivogel. "Fertile for Fraud Farmers Routinely Collect Federal Insurance and Disaster Payments Prone to Fail." *St. Louis Post Dispatch*. January 15, 1995. pp. 01B.

Schucany WR, Gray HL, Owen DB: On bias reduction in estimation. *Journal of the American Statistical Association* 66: 524-533, 1971.

Tukey JW: Bias and confidence in not quit large samples. *Annals of Mathematical Statistics* 29: 614, 1958.

Westphal C, Blaxton T: *Data Mining Solutions*. Wiley: New York, 1998.

Weisberg, H.I. and R.A. Derrig. "Quantitative Methods for Detecting Fraudulent Automobile Bodily Insurance Claims." *AIB Cost Containment/Fraud Filing*, pp. 49-82, 1993.

Weiss SM, Idurkha N: *Predictive Data Mining*. Morgan Kaufman Publishers: San Francisco, 1998.

Williams, G.J. and Z. Huang. "Mining the Knowledge Mine. The Hotspots Methodology for Mining Large Real World Databases." In: Sattar, A. (ed.), *Advanced Topics in Artificial Intelligence*, Springer Verlag: Berlin, Germany. p. 340-348. 1997.

Witten IH, Frank E: *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Academic Press: San Diego, 2000.

Yeo, D. "Better Insuring With Information." *Canadian Underwriter*. 67(August 2000): 46-47.