Data Mining in a Geospatial Decision Support System for Drought Risk Management

Steve Goddard
Stephen E. Reichenbach
William J. Waltman
Computer Science & Engineering
University of Nebraska - Lincoln
Lincoln NE 68588-0115
goddard@cse.unl.edu
reich@cse.unl.edu
wwaltman@unlnotes.unl.edu
http://www.cse.unl.edu/~goddard/
http://www.cse.unl.edu/~reich/

Sherri K. Harms
Computer Science & Engineering
University of Missouri-Columbia
Columbia, MO 65211
sharms@cse.unl.edu
http://www.cse.unl.edu/~sharms

Tsegaye Tadesse
National Drought Mitigation Center
University of Nebraska - Lincoln
Lincoln NE 68588
tadesse@unlserve.unl.edu

Topic Areas: Government and social policy; Drought Risk Management; Data integration and statistics

Abstract

We are developing an advanced **Geospatial Decision Support System (GDSS)** to improve the quality and accessibility of drought related data for drought risk management. This is part of a Digital Government project aimed at developing and integrating new information technologies for improved government services in the **USDA Risk Management Agency (RMA)** and the **Natural Resources Conservation Service (NRCS)**. Our overall goal is to substantially improve RMA's delivery of risk management services in the near-term and provide a foundation and directions for the future.

We integrate spatio-temporal knowledge discovery techniques into our GDSS using a combination of data mining techniques applied to rich, geospatial, time-series data. Our data mining objectives are to: 1) find relationships between user-specified target episodes and other climatic events and 2) predict the target episodes. Understanding relationships between changes in soil moisture regimes and global climatic events such as El Niño could provide a reasonable drought mitigation strategy for farmers to adjust planting dates, hybrid selection, plant populations, tillage practices or crop rotations.

This work highlights the innovative data mining approaches integral to our project's success and provides preliminary results that indicate our system's potential to substantially improve RMA's delivery of drought risk management services.

1. Introduction

This work is part of an NSF-funded Digital Government project[1] aimed at developing and integrating new information technologies for improved government services in the **USDA Risk Management Agency (RMA)** and the **Natural Resources Conservation Service (NRCS)**. The mission of the USDA is to enhance the quality of life of U.S. citizens by supporting farming and agriculture and by expanding global markets for agricultural products (USDA Strategic Plan 1997). USDA's vision is to help citizens

[1] This project is a multi-disciplinary collaboration between researchers at University of Nebraska - Lincoln and the USDA, with cooperation from the US Geological Survey (USGS) Earth Resources Observation Systems (EROS) Data Center and the Center for Rural Affairs at Walthill, NE. The University of Nebraska - Lincoln research team members are affiliated with the Computer Science and Engineering (CSCE) Department, the National Drought Mitigation Center (NDMC) and the High Plains Regional Climate Center (HPRCC).

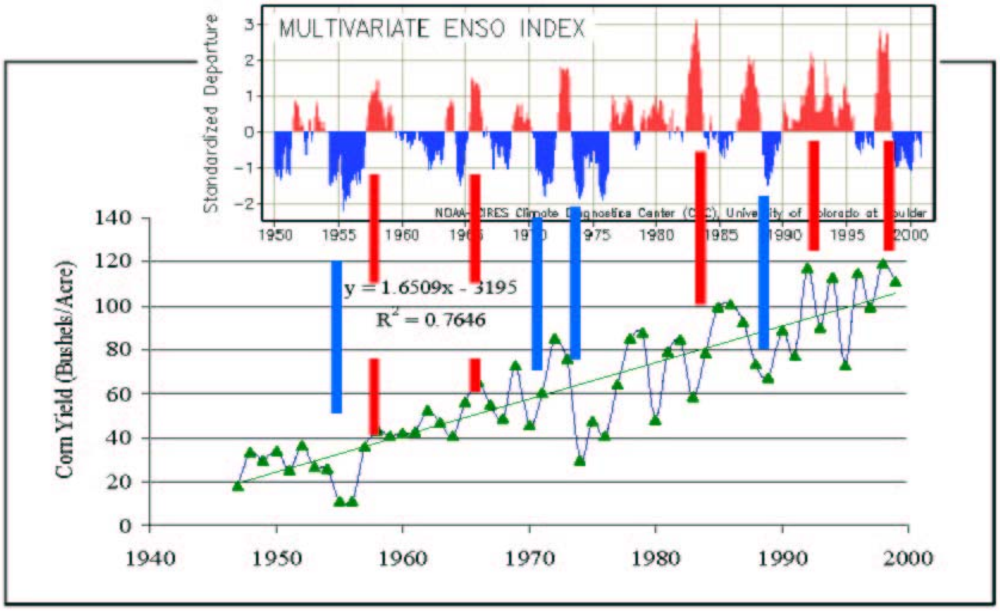

Figure 1. The relationship of corn yield in Nebraska to ENSO signatures through time.

live in harmony with the land while enjoying healthy and prosperous lives. The RMA, an important arm of USDA, "is responsible for helping keep America's farmers and ranchers in business as they face the uncertainties of weather and markets" (USDA Strategic Plan 1997). The mission of the RMA is to *strengthen the safety net for agricultural producers (farmers) through sound risk management programs and education.* RMA's mission requires understanding risks across the agricultural landscape, and communicating these risks with agriculture producers, the private sector crop insurance industry, and other government agencies. The problems in processing available data, developing information resources, and communicating with decision-makers are complex and difficult. Data comes from many sources, including other USDA agencies, the private sector crop insurance industry, and the US Geological Survey, with attendant problems in retrieval, reformatting, integration, and the creation and handling of meta-data. Developing information for risk management involves discovery of patterns in various types of crop and environmental data, modeling, and geospatial analyses. Finally, information involving complex concepts and relationships must be communicated to farmers and other decision-makers so that they can manage and mitigate risk.

We are developing an advanced **Geospatial Decision Support System (GDSS)** to support faster and better drought risk management. Our overall goal is to substantially improve RMA's delivery of drought risk management services in the near-term and provide a foundation and directions for the future. Helping RMA become a model "digital government" agency will directly assist those in agriculture and will have important economic and societal benefits.

Nationally, drought events are the dominant (47%) cause of crop loss, followed by excess moisture (22%) and cold or frost conditions (13%) (USDA RMA 1998). The Federal Emergency Management Agency (FEMA) reported the average annual cost of drought as $6-8 billion and the National Climatic Data Center (NCDC; Asheville, NC) reported the estimated cost of the 1988 drought alone was $40 billion. "Concerning drought impacts in the state of Nebraska, during the ten-year 1989-1998 period, the indemnity paid for drought losses in Nebraska totaled more than $92 million, or approximately $9 million per year (USDA RMA 1999)."

Crop productivity and risk are products of the environment and the grower's management practices. Crop yields can serve as integrated reflections or indicators of El Nino/La Nina signatures, as shown in

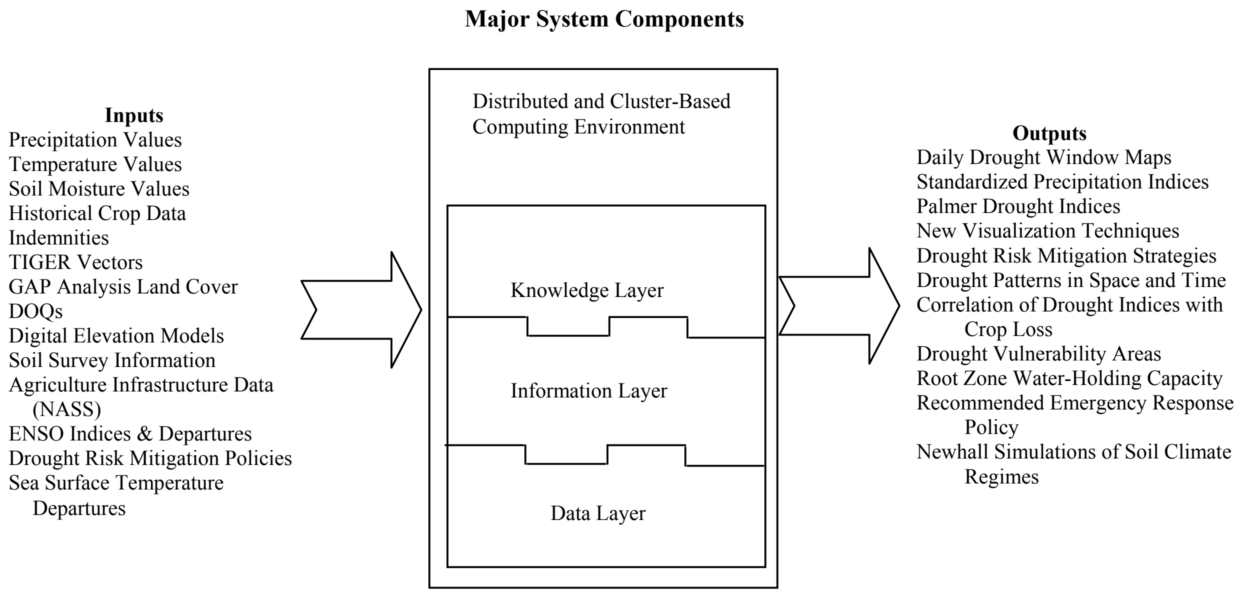
Major System Components
Inputs
Precipitation Values
Temperature Values
Soil Moisture Values
Historical Crop Data
Indemnities
TIGER Vectors
GAP Analysis Land Cover
DOQs
Digital Elevation Models
Soil Survey Information
Agriculture Infrastructure Data (NASS)
ENSO Indices & Departures
Drought Risk Mitigation Policies
Sea Surface Temperature Departures
Distributed and Cluster-Based Computing Environment
Knowledge Layer
Information Layer
Data Layer
Outputs
Daily Drought Window Maps
Standardized Precipitation Indices
Palmer Drought Indices
New Visualization Techniques
Drought Risk Mitigation Strategies
Drought Patterns in Space and Time
Correlation of Drought Indices with Crop Loss
Drought Vulnerability Areas
Root Zone Water-Holding Capacity
Recommended Emergency Response Policy
Newhall Simulations of Soil Climate Regimes

Figure 2. A geospatial decision support system for drought risk management.

Figure 1. The vulnerability of croplands to drought events is strongly influenced by crop selection or rotation, time of planting, genetics of the crop, tillage and fertility management, diseases, insects, and the quality of underlying soils. Risks can be mitigated and losses reduced with improved knowledge development and dissemination. Automation and knowledge discovery in decision support systems is key to the future success of RMA and its responsiveness to the nation's farmers. The successful completion of this project will help RMA integrate advanced geospatial applications into their operations, and result in a government agency that is much more efficient and responsive to drought risk management.

One of our project's performance objectives is to report associations between weather station data, crop yields, and sea surface thermal properties. To meet this performance objective, we integrate spatio-temporal knowledge discovery techniques into our GDSS using a combination of data mining techniques applied to rich, geospatial, time-series data. Our data mining objectives are to: 1) find relationships between user-specified target episodes and climatic events and 2) predict the target episodes. Understanding relationships between changes in soil moisture regimes and global climatic events such as El Niño could provide a reasonable drought mitigation strategy for farmers to adjust planting dates, hybrid selection, plant populations, tillage practices, or crop rotations. This work highlights the innovative data mining approaches integral to our project's success and provides preliminary results that indicate the potential our system has to substantially improve RMA's delivery of drought risk management services.

2. A Geospatial DSS for Drought Risk Management

The project entails making improvements and innovations in RMA's government services by developing an advanced GDSS for drought risk management. The data layer, information layer, and knowledge layer and example inputs and outputs of the GDSS are illustrated in Figure 2. The **network-clustered server environment** allows for distributed and parallel computing that will be necessary in our advanced geospatial applications.

The **data layer** builds on the UNL MLPQ/GIS (Kanjamala 1998) system. Its roles are ingesting data and providing access to data that are fundamentally important. This layer will:

- Import and translate data relevant to drought risk management into constraint database representations.

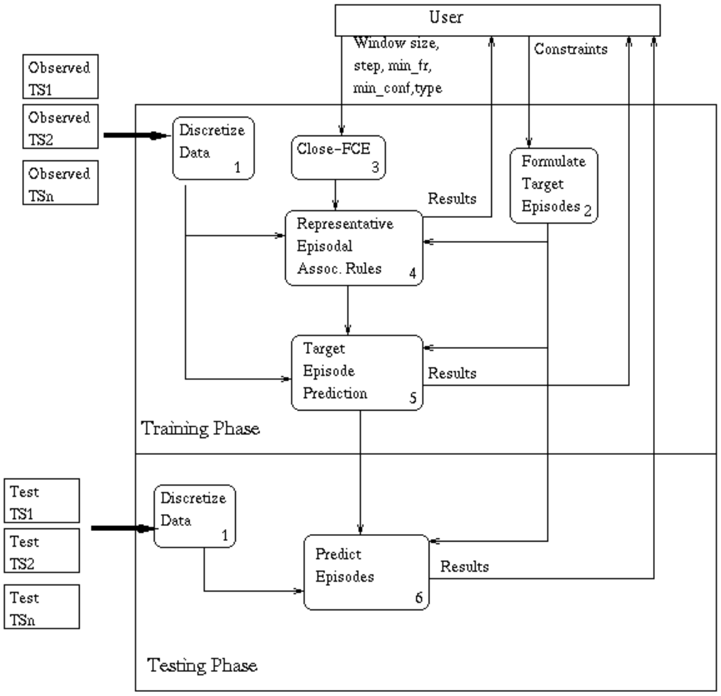

Figure 3. The information layer of a GDSS for drought risk management.

- Query the constraint database systems for analysis and forecasting of drought.
- Display and animate drought episodes.

The **information layer** uses **Targeted Episode Association, Learning and Prediction** for data mining and information retrieval. This layer integrates data into information products for manipulation in knowledge development. This layer will:

- Use the MLPQ/GIS system for data mining and information retrieval.
- Use data mining methodologies for associating episodes to targeted episodes.
- Use data mining methodologies for learning predictive algorithms for targeted episodes.

The **knowledge layer** builds on the information layer to provide Drought Risk Decision Support. This layer will:

- Use data mining and information retrieval tools to explore models for drought conditions, risk assessment, and prediction.
- Model the relationships of additional input variables with drought events and episodes.
- Generate dynamic vulnerability maps, drought forecasts, and economic and environmental impact assessments using decision support tools.

3. Targeted Episode Association, Learning, and Prediction

As a critical part of the proposed GDSS, the information layer incorporates several sequential data mining techniques shown in Figure 3. (See Harms 2001 for the detailed system.) Our overall goal is to find relationships with droughts and other climatic episodes and with agricultural outcomes, such as crop yield. The proposed techniques are intended as *exploratory* methods. Thus, iterative and interactive application of the approach coupled with human interpretation of the rules is likely to lead to the most useful results, rather than a fully automated approach (Das 1998). The basic idea is to:

1. Convert time series into discrete representations, using clustering (Das 1998), normalization (Goldin 1995), and transformations. Each data set will need its own discretization process, and thus this step relies on human (domain-expert) involvement.
2. Set the user specified target episodes as constraints (Srikant 1997) and set other system parameters as specified by the user, such as sliding window width, the minimum frequency and the minimum confidence value.
3. Record frequent episodes (Mannila 1995, Klemettine 1999) that meet the targeted episode constraints using closure techniques (Pasquier 1999, Saquer 2000) to increase the probability that the discovered episode is important.
4. Discover representative episodal association rules relating the frequent episodes using association rule techniques (Agrawal 1993), representative association rule techniques (Kryszkiewicz 1998a, Kryszkiewicz 1998b), and sequential association rule techniques (Das 1998).
5. Use highly significant rules where the consequent is restricted to the target episodes, to learn the target episodes using classification and rule mining prediction techniques (Liu 1998).
6. Validate the learning algorithm by predicting (new) test data.

We first discretize the data sets into sequences of events, using the same time measure for all data sets. We look for events that occur together in a relatively short time interval, called the *sliding window width*. A pattern of events that occurs within the sliding window width is called an *episode*. An episode may be repeated in several windows through time. The *frequency* of an episode is the number of windows in which the episode occurs. The domain-expert sets the minimum frequency value to the minimum number of windows in which episodes must occur to be of interest. This constrains the set of target episodes to the episodes that occur frequently in time.

After we have found the frequent episodes, we look for association rule patterns within each episode. An episodal association rule *r* is a rule of the form X => Y where both X and Y are nonempty subsets of events, and $X \cap Y = \emptyset$. We consider how often the episode occurs relative to how often a given subset of events occurs. For each episode Z, we look at each subset of events X, and let Y = Z\X. The events in X make up the antecedent of the rule *r*, and Y its consequent. The frequency of the episode Z divided by the frequency of the subset X is the *confidence* of the association rule X => Y. This gives the proportion of the time that all the events in Z occur given that the events in X occur. The *coverage* of a rule is the frequency of the subset X. The number of potential rules grows quickly with the number of events in the antecedent (Das 1998). We reduce this number while still maintaining rules of interest to the domain-expert, by: 1) considering only the association rules that meet the minimum confidence value, and 2) using representative episodal association rules to find the minimal set of rules that cover the entire set of frequent episodes. The representative association rules are the minimal set of association rule that can be generated from a given data set for the minimum frequency and confidence parameters (Kryszkiewicz 1998a). From this set, the entire set of association can be generated, so the user can mine around the set of rules. We then pick the set of association rules that are predictive of the target episodes, and validate the results on new data.

4. Preliminary Results for the Nebraska South Central Research & Extension Center

We have used this approach to find relationships between drought episodes at the South Central Research & Extension Center (SCREC) at Clay Center, NE, and other climatic episodes, from 1989-1999. There is a network of agricultural research stations in Nebraska with automated weather stations that can serve as long-term reference sites to search for key patterns and link to climatic events. We use data from a variety of sources:

- Satellite vegetation data from USGS's EROS Data Center (US National Oceanic and Atmospheric Administration (NOAA) Advanced Very High Resolution Radiometer AVHRR biweekly data set, 1989-1999)
- Standardized Precipitation Index (SPI) data from NDMC
- Precipitation and soil moisture data such as daily rainfall amount and the Crop Moisture Index (CMI) from HPRCC
- North Atlantic Oscillation Index (NAO) from Climatic Research Unit, University of East Anglia, UK[2]
- Pacific Ocean Southern Oscillation Index (SOI) (Climate Prediction Center, NOAA)[3]

The data for the satellite and climatic indices are grouped into seven categories, i.e. extremely dry, severely dry, moderately dry, near normal, moderately wet, severely wet, and extremely wet. In this preliminary study, the vegetation conditions are assessed using the Standardized Vegetation Index (SVI) based on the NOAA AVHRR satellite data. The SOI and NAO categories are based on the standard deviation from the normal and the negative values are considered to show the dry periods. The one month SPI values also are grouped into the same seven categories to show the precipitation intensity relative to normal precipitation for a given location and a given month.

After normalizing and discretizing each data set using the seven categories above, we specified droughts (the three dry categories) as our target episodes and used a sliding window of 3 months. Using these parameters, we looked for the drought episodes that occurred at least 10% of the time, since droughts are infrequent in nature. From the frequent drought episodes we found several representative association rules, which occur with at least 70% confidence including those shown in Table 1, explained below.

Table 1. Sample rules discovered for Nebraska South Central Research & Extension Center

Rule Number	Rule (X ==> Y)	Confidence	Frequency	Coverage
		(X∪Y)/X	|X∪Y|	|X|
1	29, 3, 31 ==> 59	0.86	12	14
2	29, 38 ==> 59	0.72	13	18
3	24, 3 ==> 59	0.80	12	15

1. If the SOI is moderately dry to extremely dry, then the SVI (vegetation) with one-month lag was found to be under moderately dry conditions, with 86% confidence.
2. If the SOI is extremely dry and NAO is moderately dry, then the SVI (vegetation) with one-month lag was found to be under moderately dry conditions, with 72% confidence.
3. If the one-month SPI is moderately dry and the SVI (actual vegetation) is under moderately dry conditions, then the SVI (vegetation) with one-month lag was found to be under moderately dry conditions, with 80% confidence.

The preliminary results show that the episodes of below normal Southern Oscillation Index values in the Pacific Ocean (El Niño signatures) and below normal North Atlantic Index values were associated with occurrences of dry vegetation conditions after one month at the SCREC weather station in Clay Center, NE. These rules indicate that the knowledge discovery decision support system will help to assess the local and global climatic conditions and identify the local drought conditions ahead of time. The third rule validates common sense, and basically states that if it is dry now, and if it does not rain, then in one month it will also be dry. Although this knowledge is intuitive, a knowledge discovery

[2] Available from: http://www.cru.uea.ac.uk/cru/data/nao.htm

[3] Available from: http://www.cpc.noaa.gov/data/indices/index.html

decision support system must be able to validate common knowledge, in addition to discovering the truly interesting patterns.

5. Conclusions & Future Work

By allowing the user to explore iteratively and interactively, our GDSS gives the user firm control on the knowledge discovered. We use a data-driven approach; using current and new data mining techniques to discovery association rules for target drought episodes. Our system will discover relationships in various types of crop and environmental data. It will then go a step further to learn and predict the target episodes. We plan to expand our information layer to support decisions about the spatial extent of drought and to discover time-delayed relationships in the data sets.

Risks can be mitigated and losses reduced with improved knowledge development and dissemination. Our system provides the automation and knowledge discovery in decision support systems that is key to the future success of RMA and its responsiveness to the nation's farmers. The successful completion of this project will help RMA integrate advanced geospatial applications into their operations, and result in a government agency that is much more efficient and responsive. RMA will have improved effectiveness and greater impact in communicating about drought risk management.

6. References

R. Agrawal, T. Imielinski and A. Swami. Mining association rules between sets of items in large databases. In *Proc. of the ACM SIGMOD 1993 Int'l Conf. on Management of Data* {SIGMOD 93], Washington DC, 1993, pp. 207-216.

G. Das, K-I. Lin, H. Mannila, G. Ranganathan, and P. Smyth. Rule discovery from time series. In *Proc. of the 4th Int'l Conf. on Knowledge Discovery and Data Mining* [KDD 98], New York, NY Aug. 1998, pp. 16-22.

D. Q. Goldin and P. C. Kanellakis. On similarity queries for time-series data: Constraint specification and implementation. In *Proc. of the 1995 Int'l Conf. on the Principles and Practice of Constraint Programming*, Marseilles, France, Sept. 1995.

S. Harms. PhD Dissertation Proposal: *Associating and Predicting Episodes of Events in Multiple time Series for Supporting Policy Decision Making*, Department of Engineering & Computer Science, University of Missouri - Columbia, Columbia MO, February 6, 2001.

Hurrell, J.W., 1995: Decadal trends in the North Atlantic Oscillation and relationships to regional temperature and precipitation. *Science* 269, 676-679.

P. Kanjamala, P.Z. Revesz, and Y. Wang. MLPQ/GIS: a GIS using linear constraint databases. In C.S. R. Prabhu, editor, *Proc. of the 9th COMAD Int'l Conf. on Management of Data*, Tata McGraw Hill, 1998, pp. 389-393.

M. Kryszkiewicz. Fast discovery of representative association rules. In *Lecture Notes in Artificial Intelligence: Proc. of the RSCTC 98*, vol. 1424, Springer-Verlag, 1998, pp. 214-221.

M. Kryszkiewicz. Representative association rules. In *Lecture Notes in Artificial Intelligence: Proc. of PAKDD 98*, vol. 1394, Springer-Verlag, 1998, pp. 198-209.

B. Liu, W. Hsu, and Y. Ma. Pruning and summarizing the discovered associations. In *Proc. of the 5th Int'l Conf. on Knowledge Discovery and Data mining* [KDD99], San Diego, CA, Aug. 1999.

H. Mannila and H. Toivonen. Discovering frequent episodes in sequences. In *Proc. of the 1st Int'l Conf. on Knowledge Discovery and Data Mining* [KDD 95], Montreal, Canada, Aug. 1995, pp. 210-215.

Jones, P.D., Jónsson, T. and Wheeler, D., 1997: Extension to the North Atlantic Oscillation using early instrumental pressure observations from Gibraltar and South-West Iceland. *Int. J. Climatol*. 17, 1433-1450.

Monthly Atmospheric & SSI Indices, Climate Prediction Center, National Oceanic and Atmospheric Administration, U.S. Department of Commerce http://www.cpc.noaa.gov/data/indices/index.html

N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices. *Information Systems,* vol. 24, 1999, pp. 25-46.

J. Saquer and J.S. Deogun. Using closed itemsets for discovering representative association rules. In *Proc. of the 12th Int'l Symposium on Methodologies for Intelligent Systems* [ISMIS 2000], Charlotte, NC, Oct. 2000.

R. Srikant, Q. Vu, and R. Agrawal. Mining association rules with item constraints. In *Proc. of the 3rd Int'l Conf. on Knowledge Discovery and Data Mining* [KDD 97], 1997, pp. 67-93.

USDA. *USDA Strategic Plan 1997-2002*. U.S. Government Printing Office, Washington DC, 1997.

USDA Risk Management Agency. *Building a Risk Management Plan: Risk-Reducing Ideas that Work*. U.S. Dept. of Agriculture, Washington DC, 1998.

USDA Risk Management Agency. *USDA National RMA Data Base.* Risk Management State Office, Billings, MT, 1999.

USDA Soil Conservation Service. *Land Resource Regions and Major Land Resource Areas of the United States.* Number 296 in Agric. Handbook. U.S. Gov't Printing Office, Washington DC, 1981.