Building Knowledge Discovery into a Geo-spatial Decision Support System*

Sherri K. Harms
Department of CSIS
University of Nebraska Kearney
Kearney NE
harmssk@unk.edu

Jitender Deogun
Department of CSE
University of Nebraska Lincoln
Lincoln NE
deogun@cse.unl.edu

Steve Goddard
Department of CSE
University of Nebraska Lincoln
Lincoln NE
goddard@cse.unl.edu

ABSTRACT

The emergence of remote sensing, scientific simulation, telescope scanning, and other survey technologies has dramatically enhanced our capabilities to collect spatio-temporal data. However, the explosive growth in data makes the management, analysis, and use of data difficult and expensive. In decision support applications with spatio-temporal data, it is important to study the temporal relationships of the parameters that influence the decision. Because multiple spatio-temporal data sets contain volumes of data, and often there is a delay between the occurrence of an event and its influence on the dependent variables, finding interesting patterns can be difficult.

A geo-spatial decision support system (GDSS) with data mining techniques is fundamental for effective decision-making on complex spatio-temporal issues. This paper presents a layered architecture for a distributed GDSS that uses temporal rule discovery to aid the decision-making process. Data mining algorithms are used to identify temporal relationships between multiple spatio-temporal data sets where time lags may exist between the related events. These algorithms allow the user to specify target events, to prune rules that are not of interest to the current decision-making problem. A geo-spatial decision support system for drought risk management is used to demonstrate the effectiveness of building knowledge discovery into a GDSS.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.2.8 [**Database Management**]: Database Applications—*Data mining, Spatial databases and GIS*; J.2 [**Physical Sciences and Engineering**]: Earth and atmospheric sciences

*This research was supported in part by NSF Digital Government Grant No. EIA-0091530 and NSF EPSCOR, Grant No. EPS-0091900.

SAC 2003 2003 Melbourne, Florida, USA

Keywords

Temporal Rule Discovery, Knowledge Discovery, Spatio-temporal Data Mining, Geo-spatial Decision Support, Drought Risk Management

1. INTRODUCTION

Making decisions that involve complex systems such as risk management require a cadre of domain experts to extract meaningful interpretations from large multidisciplinary databases. The emergence of remote sensing, scientific simulation, telescope scanning, and other survey technologies has dramatically enhanced our capabilities to collect spatio-temporal data. However, the explosive growth in data makes the management, analysis, and use of data difficult and expensive.

Data mining techniques provide automatic or semi-automatic means for data interpretation. Association rules are used to discover relationships between parameters that influence decisions. They are easy to understand, and most domain experts can relate to the rules that are expressed about their data and use them in their decision making process. The relationships are not always easy to find, and frequently there is a delay between the occurrence of an event and its influence on the dependent variables. Additionally, because multiple spatio-temporal data sets contain volumes of data, finding the patterns of interest to the problem can be difficult.

A geo-spatial decision support system (GDSS) is a collection of tools that can be used by analysts to assist the decision-making process. The tools in a GDSS combine data into pieces of information that can lead to domain knowledge useable by both experts and non-experts. A GDSS with data mining techniques is fundamental to solving the data interpretation problem and for effective decision-making on complex spatio-temporal issues.

1.1 Related Work

Literature on rule discovery algorithms has become extensive [5, 9, 14] since their introduction by Agrawal *et al.* in [1]. Recently there has been an increased interest to use temporal data mining techniques to index, cluster, classify and mine association rules from large databases [3, 8, 14]. Time-series data mining algorithms identify hidden patterns within the data in time-series analysis. These algorithms are designed to characterize and predict complex, non-periodic,

irregular, chaotic time-series.

There are many current approaches to temporal, spatial and spatio-temporal knowledge discovery. Researchers typically either approach the problem from a temporal approach first, and then apply spatial analysis, or vice versa. Tan *et al.* in [12] explored four categories of spatio-temporal patterns: 1) relationships between events at a given spatial location that ignore the temporal aspects of the data, 2) relationships between events across spatial locations that ignore the temporal aspects of the data, 3) temporal relationships among events occurring at the same location, and 4) temporal relationships among events occurring at different spatial locations. They transformed the data into market-basket type transactions, and applied existing algorithms to find spatio-temporal patterns in earth science data.

1.2 Overview

This paper provides an overview of an integrated, intelligent GDSS framework. Data mining algorithms are built into the GDSS framework to find temporal relationships among events occurring at different spatial locations, with user-specified targeted events, and with possible time lags embedded within the relationships. The National Agricultural Decision Support System (NADSS) [4] is used to demonstrate the effectiveness of building knowledge discovery into a GDSS.

Within the GDSS framework, the data mining approaches are well suited for sequential problems that have groupings of events that occur close together, but may occur relatively infrequently over the entire dataset. They are also well suited for problems that have periodic occurrences when the signature of one sequence is present in other sequences, even when the multiple sequences are not globally correlated or spatially co-located.

2. FRAMEWORK

Figure 1 shows a four-tier software architecture for an open, distributed GDSS [4]. An important aspect of this GDSS is accessibility of the tools to decision-makers, domain experts and the public. The decision-making process begins by combining and organizing data into pieces of information. Multiple pieces of information are then examined and combined to discover or create knowledge, which is the basis upon which a decision is made. The large vertical interface arrow at the right of the figure is meant to represent the ability of high-order layers to make requests to non-adjacent, low-order layers, via mechanisms such as HTTP, Internet Inter-ORB (Object Request Broker) Protocol (IIOP), Java Remote Method Invocation (RMI), or TCP. Each of the three lower layers (data, information, and knowledge) is associated with a cache for performance reasons. Strictly speaking, the cache is not needed. However, building the cache into the architecture provides performance benefits that outweigh the complexity it brings [4].

The *data layer* contains distributed spatial, constraint, and relational databases. The purpose of this layer is to provide transparent access to either local or remote data without concern for data formats. This layer provides a mechanism to encapsulate existing data inter-operability solutions such as Common Object Request Broker Architecture (CORBA)-based or Distributed Component Object Model (DCOM)-based Open GIS Consortium (OGC) objects, or data access via the Open Geographic Datastore Interface (OGDI).

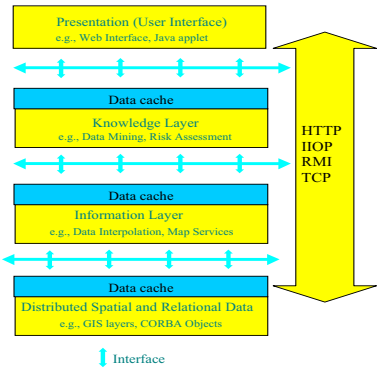

Figure 1: The Four-tier Architecture of a Distributed GDSS.

The *information layer* combines data and organizes it into information. It is organized around a collection of domain-specific servers that process data into information. Examples of servers in this layer are data interpolation servers and map servers, which may be either domain independent (e.g., spline interpolation) or domain specific (e.g., terrain regression). Depending on the domain, other servers can be added to this layer. For example, drought index servers are used in the NADSS to process current and historical climate data from weather stations. The resulting index reflects how dry or wet a site is for a given period of time relative to its historical record. Thus the drought index is domain specific information developed from climate data.

The *knowledge layer* builds on the information layer to create or discover knowledge. Servers that provide or discover domain-specific knowledge are implemented in the knowledge layer. The knowledge layer incorporates several sequential data mining techniques. Simulation models and other knowledge analysis algorithms may also be used. The intent is that decision-makers will interact with this layer, via the User Presentation interface, to build and gather domain-specific knowledge.

The *Presentation layer* provides the interface for the decision-makers to interact with the GDSS. The user interface can take many forms. The simplest interface is developed using web pages that interact with the lower layers via CGI requests.

3. RULE DISCOVERY PROCESS

The temporal knowledge discovery algorithms embedded within the knowledge layer of the GDSS framework are Representative Episodal Association Rules (REAR) [6], and Minimal Occurrences With Constraints and Time Lags (MOWCATL) [7]. Because rule discovery is an exploratory process, the framework allows for the iterative and interactive application of the data mining algorithms coupled with human interpretation of the rules. This process is more likely to lead to the more useful results than fully automating the approaches [3].

After data is combined and organized into information by tools in the information layer, it is discretized into meaningful categories for knowledge discovery, using transformations, normalization and clustering [3]. When multiple se-

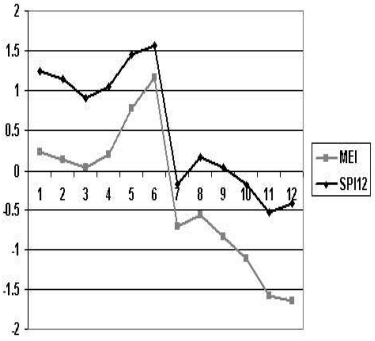

Figure 2: Sample Event Sequences.

quences are used, each data set is normalized and discretized independently. This step relies on domain-expert involvement for proper discretization. The time granularity is converted to a single (finest) granularity before the discovery algorithms are applied to the combined sequences as in [2].

The spatio-temporal data is viewed as event sequences. An *event sequence* is a collection of time-ordered events that happen within a finite time period. For example, Figure 2 shows event sequences for the Multivariate El Niño Southern Oscillation (ENSO) Index (MEI) from the Pacific Ocean and the twelve-month Standardized Precipitation Index (SPI) values for 1999 at Clay Center, Nebraska. SPI values show rainfall deviation from normal for a given location at a given time [10]. As shown, MEI and SPI12 share the same trend of variation during this time period. One goal of the rule discovery process is to find rules that indicate this kind of relationship.

3.1 Representative Episodal Association Rule *(REAR)* Method

The Representative Episodes Association Rules (REAR) algorithm finds episodes of events that occur together in a relatively short time interval, called the *window width*. To process the data, a *sliding window* is used, by sequentially moving the window of width *win* one step at a time through the data. An *episode* in an event sequence is a partial order defined on a set of events [9]. It is said to occur in a sequence if events are consistent with the given order, within a given time bound. An episode is of type *parallel* if no order is specified and of type *serial* if the events of the episode have a fixed order. 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 window width and the minimum frequency value. These parameters constrain the set of target episodes to the episodes that occur frequently in time, and allow the user to control the closeness of the related event occurrences. REAR allows the user to constrain the search to user specified target episodes, to find these episodes quickly and without the distraction of the other non-interesting episodes.

The guiding principle of the algorithm lies in the "downward-closed" property of frequency, which means every subepisode is at least as frequent as its superepisode [9]. Based on this idea, candidate episode with k events are generated by joining frequent episodes that have $k-1$ events in common. Episodes that contain any subset that is not frequent are pruned. This algorithm uses only a subset of the set of frequent episodes, called *frequent closed episodes* [6]. A frequent closed episode X is equal to the intersection of all frequent episodes containing X. For example, the parallel episode with events *abc* is a frequent closed episode if no larger frequent episode contains it (such as *abcd*), and it meets the minimum frequency threshold. Closed episodes of size k can be generated in iterations prior to k. Once found, remaining subepisodes do not need to be generated. This results in a reduced input size and in a faster generation of the representative episodal association rules. Closed episodes are especially useful on datasets where events occur in clusters, even if the cluster occurs relatively infrequently over the entire dataset.

With the complete set of frequent closed episodes, association rule patterns within the episodes are found. An episodal association rule r is a rule of the form $X \Rightarrow Y$, where X is *antecedent* episode, Y is the *consequent* episode, and $X \cap Y = \emptyset$. *Support* and *confidence* are two widely used metrics in measuring the interestingness of association rules. The support of a rule $X \Rightarrow Y$ is denoted by $sup(X \Rightarrow Y)$. It indicates the percentage of episodes in the dataset that contain both X and Y. Support is simply a measure of its statistical significance [6]. The conditional probability that an episode contains Y given that it contains X is denoted $conf(X \Rightarrow Y)$. It is defined as $conf(X \Rightarrow Y) = sup(X \Rightarrow Y)/sup(X)$.

The number of potential rules grows quickly with the number of events in the antecedent [3]. REAR reduces 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,* 2) *using representative episodal association rules to find the minimal set of rules that cover the entire set of frequent closed episodes,* and 3) *using the antecedent and consequent constraints to keep track of the events of interest.*

The number of rules generated depends on the minimum frequency, the window width, and the minimum confidence values. In selecting these parameters one may have to consider the advantages and disadvantages of the parameters on the outputs (i.e., rules that are generated). For example, if a wider window width is selected, more relationships may be found but the analysis and interpretation of the rules may be difficult. If a smaller frequency is chosen, there could be more rules with high confidence but since the episodes are not frequent, they may be less meaningful.

3.2 Minimal Occurrences With Constraints and Time Lags (*MOWCATL*) Method

The Minimal Occurrences With Constraints And Time Lags (MOWCATL) algorithm is used to find relationships between sequences in the multiple data sets, where a lag in time exists between the antecedent and the consequent. In addition to the traditional frequency and support constraints in sequential data mining, MOWCATL uses separate antecedent and consequent inclusion constraints, and separate antecedent and consequent maximum window widths, to specify the antecedent and consequent patterns that are separated by a time lag. This approach is based on association rules combined with frequent episodes, time lags and event constraints [7].

The MOWCATL approach identifies minimal occurrences

of episodes along with their time intervals. Given an episode α and an event sequence S, the window $w = [t_s, t_e)$ is a *minimal occurrence* of α in S, if: (1) α occurs in the window w, and (2) α does not occur in any proper subwindow of w [8]. The maximal width of a minimal occurrence for both the antecedent and the consequent are fixed (separately) during the process, and measure the interestingness of the episodes. Instead of counting the frequency of the episodes, the number of minimal occurrences is counted as the *support* of the episode. Episodes that do not meet the minimal support threshold are pruned.

The algorithm first stores the occurrences of the single events in the antecedent and consequent separately. Larger episodes are built from smaller episodes by joining episodes with overlapping minimal occurrences, which occur within the specified window width. This procedure finishes when there are no more candidates to look through.

After finding the supported episodes for the antecedent and the consequent independently, they are combined to form an episode rule. An episodal rule occurrence is recorded when the antecedent episode occurs within a given maximum window width win_a, the consequent episode occurs within a given maximum window width win_c, and the start of the consequent follows the start of the antecedent within a given maximum time lag. The *confidence* of an episode rule $r = \alpha[win_a] \Rightarrow_{lag} \beta[win_c]$ in a sequence S with given windows win_a, win_c, and lag is the conditional probability that β occurs, given that α occurs, under the time constraints specified by the rule. The *support* of the rule is the number of times the rule holds in the database. This allows MOWCATL to easily find rules such as "if A and B occur within 3 months, then within 2 months they will be followed by C and D occurring together within 4 months."

For serial episodes, the starting time of the consequent must be greater than or equal to the ending time of the antecedent, and must be less than or equal to the starting time of the antecedent plus the time lag. For a zero time lag, the REAR algorithm can be used instead of MOWCATL. Also, the consequent ending time must be greater than the ending time of the antecedent.

For parallel episodes, the starting time of the consequent must follow the starting time of the antecedent and can differ at most by the time lag. The order of the events in the parallel episodes is not important. Parallel episodes are used to see if the events in one episode occur "close" to the events in the other episode.

In the MOWCATL algorithm, the time lag constraint can be either a fixed time lag constraint or a maximum time lag constraint. With fixed time lag, the antecedent and the consequent episodes are separated with fixed time. It may be used to monitor parameters that occur an exact number of timestamps prior to the consequent. With the maximal time lag constraint, the start of the consequent follows the start of the antecedent after at least one time stamp, and at most lag time steps. It may be used to monitor parameters that occur within a range of time prior to the consequent.

Both rule discovery methods use the J-measure [11] as an objective measure. The formulation of J-measure takes into consideration both frequencies of left and right sides of a rule. Therefore, it not only favors rules that occur more frequently, but also provides a more complex metric for ranking rules in a manner such that the user can trade-off rule support and rule confidence.

4. RULE DISCOVERY FOR DROUGHT RISK MANAGEMENT

Managing risk is an important task in many domains, especially in Drought Risk Management (DRM). Drought affects virtually all regions of the world and results in significant economic, social, and environmental impacts. Even though droughts occur infrequently and are difficult to detect, they are a normal feature of climate and their occurrence is inevitable [13]. Given the complexity of drought, where the impacts from a drought can accumulate gradually over time and vary widely across many sectors, a well-designed decision support system is critical for effective proactive management of drought risk efforts. The Federal Emergency Management Agency estimates the annual losses because of drought in the United States at $6–8 billion, which is more than any other natural hazard. Yet, the United States does not have a comprehensive drought monitoring system or a national drought policy that emphasizes risk management by promoting the development of drought plans at all levels of government. Congress recently enacted the Agricultural Risk Protection Act of 2000 to encourage the United Stated Department of Agriculture (USDA) Risk Management Association (RMA) and farmers to be more proactive in managing drought risks.

The NADSS is being developed to substantially improve RMA's delivery of risk management services in the near-term and provide a foundation and directions for the future [4]. Drought analysts will be able to use the NADSS system to look for drought episodes and their relationships to other climatic events. Achieving NADSS's goal will require the discovery of patterns in drought indices, climatological data and social characteristics. Unfortunately, traditional methods of pattern discovery have been slow and relatively ineffective due to the vast amounts of data in various databases maintained throughout the world, the spatial extent of the data, and the extended temporal lag between related events. Data mining techniques are applied to find the relationships between parameters that are useful to make improved analysis of drought based on the patterns derived from the data. The knowledge discovery objectives for the NADSS are: 1) find relationships between climatic episodes and user-specified target drought episodes and 2) find relationships with time delays between climatic episodes and the target drought episodes.

As an experiment, relationships are found between drought episodes at several automated weather stations in Nebraska, and other climatic episodes, from 1950-1999. 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. Data from a variety of sources is used, including: 1) Standardized Precipitation Index (SPI) data from the National Drought Mitigation Center (NDMC); 2) Precipitation, temperature and soil moisture data from the High Plain Regional Climate Center (HPRCC); 3) Palmer Drought Severity Index (PDSI), calculated using station data from the HPRCC; 4) The North Atlantic Oscillation Index (NAO) from the Climatic Research Unit at the University of East Anglia, UK; 5) The Pacific Decadal Oscillation (PDO) Index and Pacific/North American (PNA) Index from the Joint Institute for the Study of the Atmosphere and Ocean (JISAO), NOAA and the University of Washington; and 6) The Pacific

Ocean Southern Oscillation Index (SOI) and the Multivariate ENSO Index (MEI) from the Climate Prediction Center, NOAA.

The oceanic and climatic data were converted into discrete representations and classified into seven categories. These seven categories are extremely dry, severely dry, moderately dry, normal, moderately wet, severely wet, and extremely wet. The thresholds for each classification were based on expert knowledge. The drought episodes (extremely dry, severely dry, and moderately dry) were identified as constraints based on the climatic drought indices i.e., SPI and PDSI values. The oceanic parameters were used as the rule antecedents and the droughts episodes were used as the rule consequents. To record the frequent episodes and generate rules, a variety of window widths, minimum frequency values, and minimum confidence values were selected for analysis.

Both the REAR and MOWCATL methods were used to find the relationships between the oceanic indices and the drought episodes. To demonstrate the use of these algorithms, five weather stations in Nebraska were selected. These stations were Ainsworth in Brown county, Alliance in Box Butte county, Clay Center in Clay county, Hayes Center in Hayes county, and West Point in Cuming county. The stations were selected randomly but show different geographical locations in Nebraska. In addition to these stations, the state wide average data for Nebraska was used to get the general overview of the state in comparison to the stations. Oceanic indices that are based on sea surface temperatures and atmospheric pressure are used since they change relatively slower that surface climatic parameters such as precipitation and temperature.

4.1 Discovering association rules using the REAR method

Rules generated using the REAR algorithm demonstrate the importance and potential use of the data-mining algorithms in monitoring drought using the oceanic and atmospheric indices. Using the REAR algorithm with parallel episodes, sample rules generated within a one-month time window are shown in Table 1, and within a two-month window are shown in Table 2. Table 3 shows sample rules generated within a three-month window for serial episodes, when using the REAR algorithm for these weather stations.

For example in Table 1 Rule 1 at Clay Center, Nebraska, if the SOI value was between 1 and 1.5, the MEI value was less than -1.5, and the PDO value was less than -2, then the PDSI value was extremely dry (less than -4) with 83% minimum confidence. The PDSI value was in an extremely dry condition when both the MEI value was less than -1.5 and the PDO value was less than -2 with 86% minimum confidence as shown in Rule 2. Rule 3 shows that if the NAO value was between -3 and -2, and the PDO value was less than -2, then the PDSI value was extremely dry and the twelve-month SPI value was severely dry with 100% confidence. Even though the confidence is 100%, the J-measure value was smaller (0.02) than the other rules described above.

A few of the association rules discovered for serial episodes (which consider the time-order of the parameters as a necessary condition), with a window size of 3 months, are shown in Table 2. Sample interesting rules that were generated for Clay Center include: Rule 1. If the MEI value was less than -1.5 followed by PDO values less than -2, then the PDSI value was extremely dry with more than 64% confidence and with a J-measure of 0.08; and Rule 2. If the NAO value was between -3 and -2, followed by the PDO value less than -2, then the PDSI value was extremely dry and the twelve-month SPI was severely dry with more than 60% confidence and with a J-measure of 0.04.

4.2 Discovering association rules using the MOWCATL method

Table 3 shows sample rules generated with the MOWCATL algorithm with parallel episodes for the five selected stations in Nebraska and the statewide average for Nebraska. As shown, the antecedent and consequent window sizes are 1 month, and a maximum of three-month time lag between the start of the antecedent oceanic parameters and the start of the consequent drought episodes is used. It can be observed from the generated rules (Rules 1-6) that if a MEI value less than -1.5, a PDO value less than -2, and a SOI value greater than 1 occurred close together, then there was high likelihood of an upcoming drought in all selected stations as well as for state of Nebraska.

Sample rules generated using the MOWCATL algorithm with a fixed time lag are shown in Table 4 for serial episodes. Because a fixed time lag is used, fewer rules are generated than when a maximal lag is used. The advantage of using a fixed time lag with the MOWCATL algorithm is that it provides the rules where the consequent occurs exactly after the specified lag, whereas the maximal lag provides the rules where the consequent occurs within the specified time lag. Thus, the rules that are generated using a fixed time lag are a subset of the rules generated using a maximal time lag.

5. ANALYSIS OF THE EXPERIMENTS

The study showed that most occurrences of drought based on the SPI and PDSI categories for each of the selected stations as well as for the Nebraska state wide average data in Nebraska, had strong associations with dry SOI, MEI, and PDO Pacific Ocean conditions, and with dry NAO Atlantic Ocean conditions. The combinations of negative MEI values (La Niña), positive values of SOI (La Niña), and negative PDO values implied occurrences of droughts in most cases for all selected stations as well as for state average data of Nebraska with different combinations of the indices and confidence values. These rules show that throughout the past 50 years, there has been a strong relationship between dry oceanic conditions in the Pacific and Atlantic Ocean and drought in Nebraska.

These rules are influenced by the type of episodes (parallel or serial), the window width, the algorithm choice (REAR or MOWCATL), the frequency or support threshold, the confidence threshold, and the time lag in the case of the MOWCATL algorithm. In general, even though fewer serial rules are generated for a given minimum confidence level, the J-measure for serial rules is generally higher ($.04 - .09$) than for parallel rules ($.02 - .06$). This indicates that even though the ordered episodes do not occur as frequently as the unordered episodes, they are more likely to be interesting when they do occur.

The window width also influences the rules discovered. For example, Rule 3 in Table 1, has a 1-month window, 100% confidence and a J-measure of 0.2, whereas the same rule with a 2-month window has 86% confidence and a J-

Table 1: Sample parallel *REAR* rules with a 1 month window.

Number	Location	Rule	Confidence	J-Measure
1	Clay Center	SOIsd, MEIed, PDOed⇒ PDSIed	0.83	0.0355
2	Clay Center	MEIed, PDOed⇒ PDSIed	0.86	0.0414
3	Clay Center	NAOmd, PDOed⇒ SPI12sd, PDSIed	1.00	0.0265
4	Ainsworth	SOIsd, MEIed, PDOed⇒ PDSIed	0.83	0.0355
5	Ainsworth	SOIed, MEIed, PDOmd⇒ SPI9md, PDSIed	0.75	0.0228
6	Ainsworth	MEIed, PDOed⇒ PDSIed	0.86	0.0414
7	Alliance	SOIed, MEImd, PDOmd⇒ SPI12md	0.75	0.0228
8	Hayes Center	SOIsd, MEIed, PDOed⇒PDSIed	0.83	0.0355
9	Hayes Center	MEIed, PDOed⇒PDSIed	0.86	0.0414
10	Hayes Center	NAOmd, PDOmd⇒SPI12md, PDSIed	1.00	0.0265
11	West Point	NAOmd, PDOed⇒PDSIed	1.00	0.0265
12	State Average	SOIsd, MEIed, PDOed⇒SPI9sd, PDSIed	0.67	0.0271
13	State Average	MEIed, PDOed⇒PDSIed	0.71	0.0330
14	State Average	NAOmd, PDOed⇒SPI12sd	1.00	0.0265

Table 2: Sample serial *REAR* rules generated with a 3 month window.

Number	Location	Rule	Confidence	J-Measure
1	Clay Center	MEIed, PDOed⇒ PDSIed	0.64	0.0802
2	Clay Center	NAOmd, PDOed⇒ (SPI12sd, PDSIed)	0.60	0.0419
3	Ainsworth	MEIed, PDOed⇒ SPI9md	0.64	0.0802
4	Ainsworth	NAOmd, PDOed⇒ SPI9sd	0.60	0.0419
5	Hayes Center	MEIed, PDOed⇒PDSIed	0.64	0.0802
6	West Point	NAOmd, PDOed⇒PDSIed	0.60	0.0419
7	State Average	NAOmd, PDOed⇒PDSIed	0.60	0.0419

Table 3: Sample parallel *MOWCATL* rules with 1 month windows and a maximum lag of 3 months.

Number	Location	Rule	Confidence	J-Measure
1	Clay Center	SOIsd, MEIed, PDOed⇒ SPI9sd, SPI12sd, PDSIed	0.83	0.03552
2	Ainsworth	SOIsd, MEIed, PDOed⇒ SPI9md, SPI12md	0.83	0.03552
3	Alliance	SOIed, MEIed, PDOmd⇒ SPI6md, SPI12md	0.75	0.02282
4	Hayes Center	SOIsd, MEIed, PDOed⇒ SPI12md, PDSIed	0.83	0.03552
5	West Point	SOIsd, MEIed, PDOed⇒SPI6md, SPI9sd, PDSIed	0.86	0.03552
6	State Average	SOIsd, MEIed, PDOed ⇒SPI9md, SPI12sd, PDSIed	0.83	0.03552

Table 4: Sample serial *MOWCATL* rules with 2 month windows and a fixed lag of 3 months.

Number	Location	Rule	Confidence	J-Measure
1	Clay Center	MEIed, PDOed⇒ PDSIed, SPI12sd	0.88	0.07772
2	Alliance	PDOed, PDOmd⇒SPI3md	0.71	0.06589
3	Hayes Center	MEIed, PDOed⇒ PDSIed	0.88	0.09385
4	Hayes Center	MEIed, PDOed⇒ PDSIed, SPI12md	0.75	0.06338
5	West Point	MEIed, PDOed⇒ SPI12sd	0.75	0.07717
6	State Average	MEIed, PDOed ⇒SPI12sd	0.75	0.07717

measure of .4. In this case, the wider window has more occurrences of the antecedent and consequent individually. As the J-measure indicates, when the antecedent and consequent do occur together, even though the confidence is not as high, the rule is still of interest. In general, the J-measure values were higher for rules with wider windows, because there are more occurrences of the combinations that exist in the datasets. However, if too wide of a window is selected, more relationships may be found but the analysis and interpretation of the rules may be difficult.

The MOWCATL algorithm, with its time lag between the rule antecedent and rule consequent, provides valuable information regarding what oceanic conditions precede drought conditions in Nebraska. As shown in the experiments (Tables 3-4), if the MEI was less than -1.5, and followed by the PDO value between 1 and 1.5, then within three months, the weather stations would record drought conditions. Additionally, the occurrences of a MEI value less than -0.5 followed by a PNA value less than -1 implied drought conditions within three months in all selected stations as well as the state of Nebraska, using 2 month windows. It can be concluded that these oceanic conditions have preceded long-term drought conditions in Nebraska over the past 50 years.

As shown in the experiments, using both the REAR and the MOWCATL data mining algorithms, more rules were generated between the 9-month and 12-month SPI drought values with dry conditions and the SOI, MEI, and PDO oceanic indices than between the 1-month, 3-month and 6-month SPI drought values with dry conditions and the oceanic indices. These results show that the oceanic parameters are more predictive for long-term droughts than short-term droughts. These associations of long-term drought with oceanic parameters may be justified by the fact that the oceanic parameters are relatively stable and change slowly, resulting in gradual changes in atmospheric circulation and its impacts on precipitation. The results of the data mining experiments show that it may be possible to identify drought episodes using oceanic parameters for a given month with a certain value of confidence. In general, most rules indicate the oceanic and atmospheric parameters could serve as a precursor to long-term drought defined by the PDSI and SPI drought indices.

6. COMPARISON OF DATA MINING TO CORRELATION

The traditional statistical correlation method was used to determine the correlation between the oceanic and climatic parameters. These correlations were compared and contrasted with the rules that were generated by the time series data mining algorithms. Given a pair of oceanic and climatic parameter values (i.e., X and Y respectively), the correlation coefficient ($\rho_{X,Y}$) provides an index of the degree to which the paired measures co-vary in a linear fashion. Based on this traditional statistical technique, the values of the correlation coefficient were calculated between the climatic and oceanic parameters.

One interesting pattern that can be confirmed with both data mining and traditional methods is the association of the climatic drought indices (both SPI and PDSI values) with the oceanic indices (SOI, MEI and PDO) for weather stations in Nebraska. The correlation of the SPI and PDSI drought indices with the MEI oceanic index was relatively higher than with the other oceanic parameters. For example, for Clay Center, NE, the correlations of SPI indices with the SOI index range between -0.15 and -0.20, while the correlations of the SPI indices with the MEI index range between 0.21 and 0.25, and the correlations of the SPI indices with the PDO index range between 0.08 and 0.15. Moreover, there was greater correlation between the PDSI index and the oceanic indices than between the SPI indices and the oceanic indices. The correlations of the PDSI index at Clay Center, NE with the oceanic indices were: -0.27 with the SOI index, 0.31 with the PDO, and 0.37 with the MEI. The Nebraska state-wide data had similar correlation values between the local drought indices and the oceanic indices, with the largest correlation again between the PDSI index and the oceanic indices. For the Nebraska state-wide data, the correlations of the PDSI index with the oceanic indices were: -0.28 with the SOI index, 0.30 with the PDO, and 0.40 with the MEI. Little correlation (less than .04) was found between any of the local drought indices and the NAO index for the selected weather stations and the Nebraska state-wide average. Also, correlation coefficients do not indicate what values of the oceanic indices are associated with specific values of the local drought indices.

Although the correlation coefficients are useful to drought risk management, clearly the spatio-temporal data mining algorithms provide a much more detailed analysis of the problem. There are three main advantages of the spatio-temporal data mining algorithms as compared to population correlation: (1) *instead of global correlation of the data, they specify focused temporal relationships between targeted episodes at specific spatial locations*, (2) *they allow for the discover of time lagged relationships between the parameters*, and (3) *they handle large volumes of data and complicated computations within a reasonable amount of time.* This shows that the data mining algorithms can identify the target episodes and generate rules that are robust tools in monitoring drought. However, the data mining tools are intended to complement, rather that replace, existing drought monitoring methods.

7. CONCLUSION

This paper presented a layered architecture for a distributed GDSS that uses spatio-temporal rule discovery to aid the decision-making process. It demonstrated the application of this approach using the NADSS for drought risk management. The REAR and MOWCATL spatio-temporal rule discovery methods, described in [6, 7], were used to analyze the spatio-temporal data.

The REAR approach with parallel episodes is used to find rules for events that occur close together in time. When used with serial episodes, the REAR approach finds rules for time-ordered events. The MOWCATL approach with a maximal time lag and parallel episodes, finds rules where the antecedent events occur close together in time, the consequent events occur close together in time, and the consequent follows shortly after the antecedent. Similarly, using MOWCATL with a maximal time lag and serial episodes finds rules where the antecedent events are time-ordered, the consequent events are time-ordered, and the consequent follows shortly after the antecedent. When using the MOWCATL approach with a fixed time lag, the consequent follows the antecedent at exactly the number of time steps specified

by the lag.

In the experiments, sample rules were shown for the relationships between oceanic climatic conditions and weather stations in Nebraska. These results were compared with population correlation. These rules demonstrated that these weather stations are each affected by the oceanic parameters, but in different ways. Armed with this knowledge, drought-risk management decision makers will be able to proactively prepare for drought at these locations.

This work is being expanded to interpolate the rule discovery between weather stations, and to effectively visualize the rules discovered. Rule discovery for seasonal influences of the climatic parameters on crop yields is also being explored. These enhancements will make the spatio-temporal data mining more interesting and meaningful to the overall GDSS.

to acknowledge grants, funding, editing assistance and what have you. In the present case, for example, the authors would like to thank Gerald Murray of ACM for his help in codifying this *Author's Guide* and the **.cls** and **.tex** files that it describes.

8. ADDITIONAL AUTHORS

Additional author: Tsegaye Tadesse (The National Drought Mitigation Center, email: `tadesse@unlserve.unl.edu`).

9. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD 1993 International Conference on Management of Data [SIGMOD 93]*, pages 207–216, Washington D.C., 1993.

[2] C. Bettini, X. S. Wang, and S. Jajodia. Discovering temporal relationships with multiple granularities in time sequences. *IEEE Transactions on Knowledge and Data Engineering*, 10(2):222–237, March 1998.

[3] G. Das, K.-I. Lin, H. Mannila, G. Ranganathan, and P. Smyth. Rule discovery from time series. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining [KDD 98]*, pages 16–22, New York, NY, August 1998.

[4] S. Goddard, S. K. Harms, S. E. Reichenbach, T. Tadesse, and W. J. Waltman. Geospatial decision support for drought risk management. *Communications of the ACM*, to appear January 2003.

[5] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM-SIGMOD International Conference on Management of Data*, Dallas, Texas, USA, May 2000.

[6] S. K. Harms, J. Deogun, J. Saquer, and T. Tadesse. Discovering representative episodal association rules from event sequences using frequent closed episode sets and event constraints. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 603–606, San Jose, California, USA, November/December 2001.

[7] S. K. Harms, J. Deogun, and T. Tadesse. Discovering sequential association rules with constraints and time lags in multiple sequences. In *Proceedings of the 2002 International Symposium on Methodologies for Intelligent Systems*, Lyon, France, June 2002.

[8] H. Mannila and H. Toivonen. Discovering generalized episodes using minimal occurrences. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining [KDD 96]*, pages 146–151, Portland, Oregon, August 1996.

[9] H. Mannila, H. Toivonen, and A. I. Verkamo. Discovering frequent episodes in sequences. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining [KDD 95]*, pages 210–215, Montreal, Canada, August 1995.

[10] T. B. McGee, N. J. Doeskin, and J. Kleist. The relationship of drought frequency and duration to time series. In *Proceedings of the 8th Conference on Applied Climatology*, pages 179–184, Boston, MA, January 1993. American Meteorological Society.

[11] P. Smyth and R. M. Goodman. An information theoretic approach to rule induction from databases. *IEEE Transactions on Knowledge and Data Engineering*, 4(4):301–316, August 1992.

[12] P. Tan, C. Potter, M. Steinbach, S. Klooster, and V. K. A. Torregrosa. Finding spatio-temporal patterns in earth science data. In *KDD-2001 Workshop on Temporal Data Mining*, San Francisco, CA, August 2001.

[13] D. A. Wilhite. Drought as a natural hazard: concepts and definitions. In D. A. Wilhite, editor, *Drought Volume II: A Global Assessment, Routledge Hazards and Disaster Series*, pages 3–8. Routledge Publishers, New York, 2000.

[14] M. Zaki. Generating non-redundant association rules. In *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining [KDD 2000]*, pages 34–43, Boston, MA, USA, August 20-23 2000.