NEW GENERATION OF DATA MINING APPLICATIONS

NEW GENERATION OF DATA MINING APPLICATIONS

Edited by

Jozef Zurada and Medo Kantardzic

A Wiley-Interscience Publication

JOHN WILEY & SONS

New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

Contents

1 Discovery of Patterns in Earth Science Data Using Data Mining

Pusheng Zhang, Michael Steinbach, Vipin Kumar, and Shashi Shekhar

Department of Computer Science & Engineering, University of Minnesota

Pang-Ning Tan

Department of Computer Science & Engineering, Michigan State University

Steven Klooster and Christopher Potter

NASA Ames Research Center

1.1 INTRODUCTION

NASA's Earth Observing System (EOS) consists of a series of satellites that generate global observations of the land surface, biosphere, solid Earth, atmosphere, and oceans. This remote sensing data, combined with historical climate records and predictions from ecosystem models, offers new opportunities for understanding how the Earth is changing, for determining what factors cause these changes, and for predicting future changes. Data mining techniques [4] have the promise to aid this undertaking by discovering interesting patterns that capture complex interactions among ocean temperatures, land surface meteorology, and terrestrial carbon flux.

[INSERT FIGURE B.1 ABOUT HERE]

Collaboration between Earth Scientists and data mining researchers has developed in two phases as illustrated by Figure B.1. In the first phase, data mining techniques are applied to discover some well-known patterns in Earth Science in order to build confidence in the use of data mining techniques. As an example, consider the El Nino climate pattern and its well-known effects on temperature and precipitation [47]. The second phase includes the exploration of novel patterns found by data mining, but not well-known by Earth Scientists. This phase also includes the exploration of patterns that are well-known to Earth scientists, but are not discovered by existing data mining techniques.

To explore Earth Science data, researchers have applied various data mining techniques [4], such as association rule mining for texture features in satellite images [34], classification of land cover types [23], and clustering of storm path trajectories [15]. Table A.1 shows some data mining techniques that can be used to address basic Earth Science questions.

[INSERT TABLE A.1 ABOUT HERE]

Although Earth Scientists have traditionally used statistical tools as their preferred method of data analysis, they are interested in using data mining tools to complement statistical tools for the following reasons. First, the statistical method of manually analyzing a single dataset via the hypothesize-and-test paradigm is extremely labor-intensive due to the extremely large and growing families of interesting spatio-temporal hypotheses and patterns in Earth Science datasets. Second, statistical methods are not designed to scale to large Earth Science datasets. Third, Earth Science datasets have selection bias in terms of being convenience or opportunity samples rather than traditional idealized statistical random samples from independent and identical distributions [8, 14, 16, 18, 19]. Data mining allows Earth Scientists to spend more time choosing and exploring interesting families of hypotheses derived from the data. More specifically, by applying data mining techniques, some of the steps of hypothesis generation and evaluation will be automated, facilitated, and improved, including steps involved in hypothesis generation, out-of-main-memory storage and manipulation of datasets, and the formation and evaluation of hypotheses from data with categorical attributes or data collected via opportunity sampling.

This chapter illustrates the application of data mining to the discovery of interesting and useful Earth Science patterns by describing some of the results [24, 41, 42, 43, 44, 46, 50, 51, 52] from our current project entitled *Discovery of Changes from the Global Carbon Cycle and Climate System Using Data Mining*. Section 1.2 describes the nature of the data, while Section 1.3 discusses data pre-processing techniques that are necessary before the data can be analyzed using data mining techniques. Sections 1.4–1.7 describe the data mining techniques used in Earth Science data, including clustering, association analysis, query processing, and other techniques. Section 1.8 concludes the chapter.

1.2 DATA DESCRIPTION AND DATA SOURCES

[INSERT FIGURE B.2 ABOUT HERE]

Earth Science data consist of a sequence of global snapshots of the Earth taken at various points in time, as shown in Figure B.2. Each snapshot consists of measurement values for a number of variables (e.g., temperature, pressure, and precipitation) collected globally. All attribute data within a global snapshot is represented using spatial frameworks. A spatial framework is a partitioning of the surface of Earth into a set of mutually disjoint regions which collectively cover the entire surface of Earth. Examples of spatial frameworks for land include the political boundaries of countries and latitude-longitude spherical grids at different resolutions, e.g., $0.5^\circ \times 0.5^\circ$ or $1^\circ \times 1^\circ$. Variables derived from global satellite data, e.g., Net Primary Production (NPP), are available at a resolution of $0.5^\circ \times 0.5^\circ$. (NPP is the net assimilation of atmospheric carbon dioxide into organic matter by plants.) Global snapshots, i.e., values of variables for each location in a spatial framework, are available for periodic, discrete points in time that span a range of twenty to one hundred years. These variable values can either be observations from different sensors, e.g., precipitation and sea surface temperature (SST), or the result of model predictions, e.g., NPP from the NASA-CASA model.

The primary focus of our work has been the development of algorithms and tools to help Earth Scientists to discover changes in the global carbon cycle and climate system, and we have focused on the datasets that are most relevant to that task. In particular, Earth Scientists who work at the regional and global scale have identified NPP as a key variable for understanding the global carbon cycle and the ecological dynamics of the Earth. Terrestrial NPP is driven by solar radiation and can be constrained by precipitation and temperature. Keeping track of NPP is important because it includes the food source of humans and all other organisms and, thus, sudden changes in the NPP of a region can have a direct impact on the regional ecology. An ecosystem model for predicting NPP, known as NASA-CASA (the Carnegie Ames Stanford Approach [30]), has been used for over a decade to produce a detailed view of terrestrial productivity. This project has made use of the multi-year output of NASA-CASA, as well as long term global sea surface temperature (SST) anomalies, to discover interesting patterns relating changes in NPP to land surface climatology and global climate. Predicting NPP based on sea surface temperature would be of great benefit given the near real-time availability of SST data and the ability of climate forecasting to anticipate SST El Nino/La Nina events.

1.3 DATA PREPROCESSING

Patterns derived from Earth Science data are often dominated by the presence of seasonal variations in the data. Although yearly patterns such as spring, summer, fall, and winter, or rainy season/dry season are important, they are already well-known. Earth Scientists are primarily interested in patterns that represent deviations from normal seasonal cycles. Examples of such patterns include anomalous climate events such as droughts, floods, heat waves, etc. Such anomalous events become apparent only if the seasonal components of the climate time series are removed. In the following, we describe a technique known as the “monthly" Z-score transformation for removing these components.

This transformation takes the set of values for a given month, e.g., all Januarys, calculates the mean and standard deviation for that set of monthly values, and then standardizes each value by calculating its Z-score, i.e., by subtracting off the mean and dividing by the standard deviation. Put another way, we express each data value in the time series in terms of its deviation from the mean value for its corresponding month, scaled by the volatility factor for that month. The month-by-month rescaling used in this transformation causes seasonal fluctuations to disappear. Figure B.3 shows the result of applying the monthly Z-score to a sample SST time series.

[INSERT FIGURE B.3 ABOUT HERE]

Other data preprocessing issues include handling trends and spatial/temporal autocorrelations in the data. The trends are long term upward or downward movements in the Earth Science time series data [5]. Spatial/temporal autocorrelation is the property by which measured values that are close in time and space tend to be highly correlated or similar [7]. Trends and spatial/temporal autocorrelations should be removed from data for two reasons. First, both have a direct impact on the statistical correlation computed between two time series. For example, temporal autocorrelation reduces the significance of a correlation by decreasing the degree of freedom in the time series. Second, removing trends and temporal autocorrelations makes the time series become stationary, a typical requirement of many statistical time series analysis techniques (e.g., ARIMA). For further details on these issues, we refer the reader to [46].

1.4 CLUSTERING

It is well-known that ocean, atmosphere, and land processes are highly coupled, i.e., climate phenomena occurring in one location can affect the climate at a far away location. Indeed, understanding these climate teleconnections is critical for finding the answers to questions such as how the Earth's climate is changing and how ecosystems respond to global environmental change.

[INSERT FIGURE B.4 ABOUT HERE]

A common way to study such teleconnections is by using climate indices [21, 22], which distill climate variability at a regional or global scale into a single time series. For example, the NINO 1+2 index, which is defined as the average sea surface temperature anomaly in a region off the coast of Peru, is a climate index associated with the El Nino phenomenon, the anomalous warming of the eastern tropical region of the Pacific. El Nino has been linked to climate anomalies in many parts of the world such as droughts in Australia and heavy rainfall along the eastern coast of South America [47]. Figure B.4 shows the correlation between the NINO 1+2 index and land temperature anomalies, which are deviations from the mean. Observe that this index is highly correlated to the land temperature anomalies on the western coast of South America, which is not surprising given the proximity of this region to the ocean region defining the index. However, few outside the field of Earth Science would expect that NINO 1+2 is also highly correlated to land regions that are far away from the eastern coast of South America, e.g., Africa and South-East Asia.

Most commonly used climate indices are based on sea level pressure (SLP) and sea surface temperature in ocean regions. These indices can ease the discovery of relationships of SST and SLP to land temperature and precipitation. These variables in turn, impact plant growth, and are therefore important for understanding the global carbon cycle and the ecological dynamics of the Earth.

As a result, Earth Scientists have devoted a considerable amount of time to developing/discovering climate indices, such as NINO 1+2 and the other indices described in Table A.2. One of the approaches used to discover climate indices has been the direct observation of climate phenomenon. For instance, the El Nino phenomenon was first noticed by Peruvian fishermen centuries ago. The fishermen observed that in some years the warm southward current, which appeared around Christmas, would persist for an unusually long time, with a disastrous impact on fishing. In the early 20th century, while studying the trade winds and Indian monsoon, scientists noticed large scale changes in pressure in the equatorial Pacific region which they referred to as the 'Southern Oscillation.' Scientists developed a climate index called the Southern Oscillation Index (SOI) to capture this pressure phenomenon. In the mid and late 60's, the Southern Oscillation was conclusively tied to El Nino, and the impact of both on global climate was recognized. Needless to say, finding climate indices in this fashion is a very slow and tedious process.

[INSERT TABLE A.2 ABOUT HERE]

More recently, motivated by the massive amounts of new data being produced by satellite observations, Earth Scientists have been using eigenvalue analysis techniques, such as principal components analysis (PCA) and singular value decomposition (SVD), to discover climate indices [45]. While eigenvalue techniques do provide a way to quickly and automatically detect patterns in large amounts of data, they also have the following limitations:

(i) all discovered signals must be orthogonal to each other, making it difficult to attach a physical interpretation to them, and (ii) weaker signals may be masked by stronger signals.

We have developed an alternative clustering-based methodology for the discovery of climate indices that overcomes these limitations. The use of clustering [9] is driven by the intuition that a climate phenomenon is expected to involve a significant region of the ocean or atmosphere, and that we expect that such a phenomenon will be 'stronger' if it involves a region where the behavior is relatively uniform over the entire area. *Shared Nearest Neighbor* (SNN) clustering [10, 11, 12] has been shown to find such homogeneous clusters. Each of these clusters can be characterized by a centroid, i.e., the mean of all the time series describing the ocean points in the cluster, and thus, these centroids represent potential climate indices. This approach offers a number of benefits: (i) discovered signals do not need to be orthogonal to each other, (ii) signals are more easily interpreted, (iii) weaker signals are more readily detected, and (iv) an efficient way is proved to determine the influence of a large set of points, e.g., all ocean points, on another large set of points, e.g., all land points.

We applied SNN clustering on the SST data over the time period from 1958 to 1998. As shown in Figure B.5, SNN found 107 clusters or candidate indices. Note that many grid points from the ocean do not belong to any clusters (these are the points belonging to the white background), as these points come from regions that are not relatively uniform and homogeneous.

[INSERT FIGURE B.5 ABOUT HERE]

Some of the cluster centroids, i.e., candidate indices, that we found are very highly correlated to known indices. Figure B.6 shows clusters that reproduce some well-known climate indices. In particular, we were able to replicate the four El Nino SST-based indices: cluster 94 corresponds to NINO 1+2, 67 to NINO 3, 78 to NINO 3.4, and 75 to NINO 4. The correlations of these clusters to their corresponding indices are higher than 0.9. In addition, cluster 67 is highly correlated to the CTI index, which is defined over a wider area in the same region. Clusters 58 and 59 are very similar to the other El Nino indices, and correlate most strongly with NINO 3 and NINO 4, respectively, although their correlations to the El Nino indices are not as high as the other four clusters.

This rediscovery of well-known indices serves to validate our approach. In fact, we are able to rediscover most of the known major climate indices using our approach. In addition, some of the cluster centroids that have a high correlation to well-known indices may represent variants to well-known indices in that, while they may represent the same phenomena, they may be potentially better predictors of land behavior for some regions of the land. Finally, cluster centroids that have medium or low correlation with known indices may represent potentially new Earth Science phenomena. Further details on the application of clustering for discovering climate indices are available in [10, 11, 41, 42, 43, 44].

[INSERT FIGURE B.6 ABOUT HERE]

1.5 ASSOCIATION ANALYSIS

Association analysis can be used to derive spatio-temporal relationships hidden in Earth Science data. The goal of association analysis is to extract significant patterns, in the form of rules or sets of events, that will predict the occurrence of certain events based on the occurrence of other events. For example, the association rule $A \longrightarrow B$ suggests that event B is expected to occur whenever event A is observed.

Due to the spatio-temporal nature of Earth Science datasets, there are four types of association patterns that may be derived:

1. *Non-spatio-temporal patterns*. This type of pattern captures events that occur simultaneously in the same location. An example of such a pattern is “low solar radiation events where there are low rainfall events”.

2. *Spatial patterns*. This type of pattern captures events that occur simultaneously at different locations. An example of such a pattern is “surface ocean heating affects climate at the nearby coastal areas”.

3. *Temporal patterns*. This type of pattern predicts events that are expected to occur in the future at the same location. An example of such a pattern is “low rainfall events eventually lead to an increase in wildfires.”

4. *Spatio-temporal patterns*. This type of pattern captures time-lagged teleconnections, i.e., relationships among events that occur in geographically distant locations. An example of such a pattern is “surface ocean heating eventually affects regional wildfires and NPP.”

Before applying association analysis, each time series is converted into a sequence of events. We define an event as an anomalously high or low value of the time series — specifically, if the value of the time series deviates by at least 1.5 standard deviations from its average. A standard association analysis algorithm, such as Apriori [2], is then applied to extract rules from the transformed data sets. The rules extracted by the Apriori algorithm are evaluated using the well-known support and confidence measures [2]. Rules with low support and low confidence tend to be statistically insignificant, and are pruned automatically by the Apriori algorithm. For example, some of the non-spatio-temporal patterns extracted by Apriori include:

R1: $\{$**FPAR-HI**$\} \longrightarrow \{$**NPP-HI**$\}$ (support = 4.6%, confidence = 51%)

Fraction of Photosynthetically Active Radiation (FPAR) measures the proportion of available radiation in the photosynthetically active wavelengths (400 to 700 nm) that a plant canopy absorbs [25]. For rule R1, an anomalously high FPAR implies that the vegetation in the region has generated more “light-harvesting” photosynthetic capability than average, which leads to higher than normal NPP. There is a 51% probability that a high NPP event would occur at a land point where a high FPAR event is observed.

R2: $\{$**PET-LO, FPAR-LO**$\} \longrightarrow \{$**NPP-LO**$\}$ (support = 3.0%, confidence = 68%)

The variable Potential EvapoTranspiration (PET) measures the potential loss of water to the atmosphere by evaporation and transpiration through plants. Whenever both low PET and low FPAR events are observed at a land point, there is a 68% chance that a low NPP event also occurs at the same location.

The association rules found using this approach are consistent with the predictions made by the NASA-CASA model. In the NASA-CASA model, NPP is a direct product of five input factors: the cloud-corrected solar irradiance, FPAR, maximum light use efficiency, temperature, and moisture stress scalars. Both example rules (R1 and R2) confirm the relationship between NPP and the input variables of the NASA-CASA model. However, since each geographical region has its own climate and topographical features, the primary drivers for low or high NPP events may be different at different locations. Earth Scientists are interested in knowing the primary drivers for anomalous NPP events associated with each land cover feature, but such information is not directly available from the output of the NASA-CASA model.

To that end, we have incorporated land cover information into the association analysis. For example, Figure B.7(a) shows the locations covered by the rule $R1$, i.e., locations where both FPAR-HI and NPP-HI events are observed simultaneously. It is not surprising to find that the rule covers almost all the regions on Earth because (i) NPP is a derivative of FPAR, and (ii) the association at a location could happen purely by chance. In the latter case, each Z-transformed time series of length 216 months (for a 17-year dataset) is expected to produce approximately 0.0668 × 216 = 11 HI and 11 LO events (where $P(|Z| \geq 1.5) = 0.0668$). If a pair of events, such as NPP-HI and FPAR-HI, are independent, then the probability of these events to co-occur together at least once is $1 - (1 - 0.0668^2)^{216} = 0.619$, which is better than random.

By increasing the support threshold at a location, the probability that these events co-occur together more than once will be significantly reduced. For example, Figure B.7(b) shows the locations covered by the rule $R1$, for which the events $\{$FPAR-HI, NPP-HI$\}$ co-occur at least 4 times. More importantly, we observe that these locations coincide with grassland and shrubland regions, as shown in Figure B.7(c). In other words, even though NASA-CASA is a global model for predicting NPP, the support for a pattern such as R1 depends strongly on its geographical locations. Regions that show a prominent R1 pattern correspond mainly to grassland and shrubland areas, a type of vegetation that is able to more quickly take advantage of periodically high precipitation (and possibly solar radiation) than forests.

[INSERT FIGURE B.7 ABOUT HERE]

We observe that the R2 pattern occurs frequently in the regions of evergreen forests (Figure B.7(d)). This leads us to believe that this pattern often appears in regions that are fire-prone (and thus, that have temporarily lost their photosynthetic capability) or have suffered other major disruptive events, but this needs to be verified by consulting historical records, which are not easily accessible through conventional sources. For further details, we refer the reader to [46].

1.6 QUERY PROCESSING

A spatial time series dataset [50] is a collection of time series [5], each referencing a location in a common spatial framework [49]. NASA Earth observation systems currently generate a large sequence of global snapshots of the Earth, including various atmospheric, land, and ocean measurements such as sea surface temperature, pressure, and precipitation. These data are spatial time series data in nature. Queries that find highly correlated time series are frequently used to discover interesting relationships among observations in spatial time series data. For example, such queries are used to identify the land locations whose climate is severely affected by El Nino [47]. However, correlation queries are computationally expensive due to large spatio-temporal frameworks containing many locations and long time sequences. Therefore, the development of efficient query processing techniques is crucial for exploring these datasets.

Previous work on query processing for time series data has focused on dimensionality reduction [1, 6, 13] followed by the use of low dimensional indexing techniques [17, 33, 35] in the transformed space. Unfortunately, the efficiency of these approaches deteriorates substantially when a small set of dimensions cannot represent enough information in the time series data. Many spatial time series datasets fall in this category. For example, finding anomalies is more desirable than finding well-known seasonal patterns. Therefore, the data used in anomaly detection is usually data whose seasonality has been removed. However, after transformations are applied to deseasonalize the data, the power spectrum spreads out over almost all dimensions. Furthermore, in most spatial time series datasets, the number of spatial locations is much greater than the length of the time series. This makes it possible to improve the performance of query processing of spatial time series data by exploiting spatial proximity in the design of access methods. We have proposed filter-and-refine query processing algorithms [50, 51] to exploit spatial autocorrelation [48] for facilitating correlation-based similarity queries on spatial time series data.

A normalized time series with *m* time measurements is a vector from the origin to the surface of an *m*-dimensional unit sphere [50]. The correlation of two time series is directly related to the angle between the two normalized time series vectors in the multi-dimensional unit sphere. We have proposed the concept of a cone [50], a set of normalized time series in a multi-dimensional unit sphere. A cone is characterized by two parameters, the center and the span of the cone. The center of the cone is the mean of all the time series in the cone. The span of the cone is the maximal angle between any time series in the cone and the cone center. For simplicity, Figure B.8 illustrates a cone in the two-dimensional case.

[INSERT FIGURE B.8 ABOUT HERE]

The proposed algorithms first divides the data into a collection of disjoint cells based on spatial proximity with a coarse resolution, where each cell corresponds to one cone in a multi-dimensional unit sphere. Each cell includes multiple time series, and the center and span are used to characterize each cone. Then each cell is divided recursively into quarters based on spatial autocorrelation to construct a cone-hierarchy search tree. The number of pairs of time series for correlation computations are substantially reduced by using a group-level join as a filtering step. Only the candidates which cannot be filtered are explored in the refinement step. The algorithms were proved to be correct and complete in [51], i.e., there were no false admissions or false dismissals.

[INSERT FIGURE B.9 ABOUT HERE]

We evaluated the performance of the proposed query processing algorithms using a NASA Earth Science dataset [51]. Correlation-based range queries and join queries were carried out on the SST data in the eastern tropical region of the Pacific Ocean and on the NPP data in the United States. The SST data contain 11556 ocean cells of the Pacific Ocean and the NPP data contain 2901 land cells of the United States. The records of SST and NPP were monthly data from 1982 to 1993. The experimental results showed that the proposed query processing algorithms often saved a large fraction of computational cost [51]. For example, 10 NPP time series from the United States were chosen to carry out the correlation-based similarity range queries with the SST data from the eastern tropical region of the Pacific Ocean respectively. The geographical locations of the 10 query time series were widely spread in the United States. Figure B.9 shows that the average computational savings for the queries range from 48 % to 89 % as the minimal correlation threshold increased from 0.3 to 0.9. For further details, we refer the reader to [50, 51, 52].

1.7 OTHER TECHNIQUES

Spatial and Temporal Outlier Detection Outliers have been informally defined as observations which appear to be inconsistent with the remainder of the data [3], or which deviate so much from other observations so as to arouse suspicions that they were generated by a different mechanism [20]. The identification of outliers can lead to the discovery of unexpected knowledge and has a number of practical applications in areas such as Earth Science, credit card fraud, voting irregularities, bankruptcy, weather prediction, and the performance analysis of athletes.

A spatial outlier is a spatially referenced object whose non-spatial attribute values are significantly different from those of other spatially referenced objects in its spatial neighborhood [39]. Informally, a spatial outlier is a local instability (in values of non-spatial attributes) or a spatially referenced object whose non-spatial attributes are extreme relative to its neighbors, even though these attributes may not be significantly different from those of the entire population. For example, a new house in an old neighborhood of a growing metropolitan area is a spatial outlier based on the non-spatial attribute house age. A temporal outlier is an object in a snapshot whose attribute values are significantly different from those of neighboring snapshots in a time series. For example, 1987 was an abnormal year in terms of El Nino activity in the world, and the values of sea surface temperature in the Eastern Pacific Ocean were very different from those of neighboring years when El Nino activity was not evident in the same locations. Outlier detection techniques are being applied to Earth Science data to identify abnormal and potentially useful spatio-temporal phenomenon.

Predictive Modeling Classical data mining algorithms often make assumptions (e.g., independent, identical distributions) that violate the first law of Geography, which says that everything is related to everything else but nearby things are more related than distant things. In other words, the values of attributes of nearby spatial objects tend to systematically affect each other. In spatial statistics, an area within statistics devoted to the analysis of spatial data, this is called spatial autocorrelation. Ignoring spatial autocorrelation may lead to residual errors that vary systematically over space exhibiting high spatial autocorrelation [36, 40]. The models derived may not only turn out to be biased and inconsistent, but may also be a poor fit to the dataset.

One way to model spatial dependencies is by adding a spatial autocorrelation term in the regression equation. This term contains a neighborhood relationship contiguity matrix. Such spatial statistical methods, however, are computationally expensive due to their reliance on contiguity matrices that can be larger than the spatial datasets being analyzed. We have developed an efficient algorithm, called PLUMS (Predicting Locations Using Map Similarity) [16, 36], to search the parameter space of classification models utilizing spatial autocorrelation. We have used PLUMS for building models to predict bird nest locations on a wetland dataset. The preliminary results show that PLUMS outperforms classical regression models substantially on this dataset. We plan to apply PLUMS to Earth Science data and further refine the algorithm to better handle the more complex spatial context present in this domain.

Co-location Mining The co-location pattern discovery process finds frequently co-located subsets of spatial event [37] types given a map of their locations (see Figure B.10). For example, the analysis of the habitats of animals and plants may identify the co-location of predator-prey species, symbiotic species, and fire events with fuel, ignition sources etc. Readers may find it interesting to analyze the map in Figure B.10 to find co-location patterns. In this example, finding a '+' implies a high chance of finding an '×' in its nearby region and vice versa. In fact, there are two co-location patterns of size 2 in this map.

[INSERT FIGURE B.10 ABOUT HERE]

The spatial co-location problem [38] looks similar to the association rule mining problem, but in fact is very different from it. Even though boolean spatial feature types (also called spatial events) may correspond to items in association rules over market-basket datasets, there is no natural notion of transactions. This makes it difficult to use traditional measures (e.g. support, confidence) and apply association rule mining algorithms which use support based pruning. In market basket datasets, transactions represent sets of item types bought together by customers. The purpose of mining association rules is to identify frequent item sets for planning store layouts or marketing campaigns. In many spatial application domains such as Earth Science, transactions are often not a natural concept. The transactions in market basket analysis are independent of each other. Transactions are disjoint in the sense of not sharing instances of item types. In contrast, the instances of boolean spatial features are embedded in a space and share a variety of spatial relationships (e.g., neighbor) with each other. We have developed one of the most natural formulations as well as one of first algorithms [38] for discovering co-location patterns from large spatial datasets and applying it to Earth Science data.

1.8 CONCLUSIONS

In this chapter, we provided an overview of our preliminary efforts to apply data mining techniques to the analysis of Earth Science data. We believe that our initial results are encouraging. For instance, we have been able to use clustering to find patterns that represent well-known climate indices. More importantly, we have also found new patterns that are not known to Earth Scientists, e.g., candidate climate indices and association patterns that relate land covers to rules that connect climate variables and NPP. While more evaluation is necessary to assess the Earth Science significance of these results, one of the major goals of our work is to produce new patterns and hypotheses for Earth Scientists to investigate, and we feel that we have made progress towards that goal.

Nonetheless, there are many tasks remaining both with respect to data mining and the application of data mining results. The main focus of the data mining work has been clustering and association analysis, and we have only lightly explored outlier detection, co-location mining, predictive modeling, and other data mining approaches. Also, while the Earth Science members of our team have evaluated and interpreted the data mining results that we have produced so far, which has led to publications in Earth Science journals [26, 27, 28, 29, 31, 32], there is much more to do within the scope of our current project. Furthermore, in the long term, we are hopeful that data mining can play an important role in helping Earth Scientists understand both global scale changes in biosphere processes and patterns, and the effects of widespread human activities. More broadly, improvements in data mining techniques made during the investigation of Earth Science data may have potential applications in other domains, such as transportation, business logistics, public health, and public safety.

Acknowledgments

We are particularly grateful to Prof. George Karypis and other colleagues on the NASA project for their helpful comments and valuable discussions.

This work was partially supported by NASA grant # NCC 2 1231 and by the Army High Performance Computing Research Center under the auspices of the Department of the Army, Army Research Laboratory cooperative agreement number DAAD19-01-2-0014, the content of which does not necessarily reflect the position or policy of the government and no official endorsement should be inferred. Access to computing facilities was provided by the AHPCRC and the Minnesota Supercomputing Institute.

References

1. R. Agrawal, C. Faloutsos, and A. Swami. Efficient Similarity Search In Sequence Databases. In *Proc. of the 4th Int'l Conference of Foundations of Data Organization and Algorithms*, 1993.

2. Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499, 1994.

3. V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley, New York, 3rd edition, 1994.

4. J. Behnke, E. Dobinson, S. Graves, T. Hinke, D. Nichols, and P. Stolorz. Final Report for NASA Workshop on Issues in the Application of Data Mining to Scientific Data, 1999.

5. G. Box, G. Jenkins, and G. Reinsel. *Time Series Analysis: Forecasting and Control*. Prentice Hall, 1994.

6. K. Chan and A. W. Fu. Efficient Time Series Matching by Wavelets. In *Proc. of the 15th Int'l Conference on Data Engineering*, 1999.

7. N.A. Cressie. *Statistics for Spatial Data (Revised Edition)*. Wiley, New York, 1993.

8. R. Duba, P. Hart, and D. Stork. *Pattern Classification, 2nd Edition*. Wiley-Interscience, 2000.

9. R. C. Dubes and A. K. Jain. *Algorithms for Clustering Data*. Prentice Hall, 1988.

10. L. Ertöz, M. Steinbach, and V. Kumar. Finding Topics in Collections of Documents: A Shared Nearest Neighbor Approach. In *Proc. of Text Mine'01, First SIAM International Conference on Data Mining, Chicago, IL, USA*, 2001.

11. L. Ertöz, M. Steinbach, and V. Kumar. A New Shared Nearest Neighbor Clustering Algorithm and its Applications. In *Workshop on Clustering High Dimensional Data and its Applications, SIAM Data Mining 2002, Arlington, VA, USA*, 2002.

12. L. Ertöz, M. Steinbach, and V. Kumar. Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. In *Proc. of 3rd SIAM International Conference on Data Mining, San Francisco, CA, USA*, May 2003.

13. C. Faloutsos. *Searching Multimedia Databases By Content*. Kluwer Academic Publishers, 1996.

14. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. MIT Press, Cambridge, MA, 1996.

15. Scott Gaffney and Padhraic Smyth. Trajectory Clustering with Mixtures of Regression Models. In *Proc. of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999.

16. R. Grossman, C. Kamath, P. Kegelmeyer, V. Kumar, and R. Namburu. *Data Mining for Scientific and Engineering Applications*. Kluwer, 2001.

17. A. Guttman. R-Trees: A Dynamic Index Structure For Spatial Searching. In *ACM SIGMOD*, 1984.

18. D. Hand. Data Mining: Statistics and More. *The American Statistician*, 52:112–118, 1998.

19. D. Hand, H. Mannila, and Padhraic Smyth. *Principles of Data Mining*. MIT Press, 2001.

20. D. Hawkins. *Identification of Outliers*. Chapman and Hall, 1980.

21. http://www.cgd.ucar.edu/cas/catalog/climind/.

22. http://www.cdc.noaa.gov/USclimate/Correlation/ help.html.

23. S. Kumar, J. Ghosh, and M. Crawford. Hierarchical Fusion of Multiple Classifiers for Hyperspectral Data Analysis. *Pattern Analysis and Applications(Special Issue on Fusion of Multiple Classifiers)*, 5(2), 2002.

24. V. Kumar, M. Steinbach, P. N. Tan, C. Potter, S. Klooster, and A. Torregrosa. Mining Scientific Data: Discovery of Patterns in the Global Climate System. In *Joint Statistical Meeting*, 2001.

25. R. Myneni, R. Nemani, and S. Running. Algorithm for the Estimation of Global Land Cover, LAI and FPAR based on Radiative Transfer Models. *IEEE Transactions on Geoscience and Remote Sensing*, 35:1380–1393, 1997.

26. C. Potter, S. Klooster, R. Myneni, V. Genovese, P. Tan, and Vipin Kumar. Continental Scale Comparisons of Terrestrial Carbon Sinks. *Global and Planetary Change*, 39:201–213, 2003.

27. C. Potter, S. Klooster, M. Steinbach, P. Tan, V. Kumar, S. Shekhar, and C. Carvalho. Understanding Global Teleconnections of Climate to Regional Model Estimates of Amazon Ecosystem Carbon Flux. *Global Change Biology, accepted for publication*, 2003.

28. C. Potter, S. Klooster, M. Steinbach, P. Tan, V. Kumar, S. Shekhar, R. Nemani, and R. Myneni. Global Teleconnections of Ocean Climate to Terrestrial Carbon Flux. *Journal of Geophysical Research*, 108, 2003.

29. C. Potter, S. Klooster, P. Tan, M. Steinbach, V. Kumar, and V. Genovese. Variability in Terrestrial Carbon Sinks over Two Decades. Part I: North America. *Earth Interactions*, 7, 2003.

30. C. Potter, S. A. Klooster, and V. Brooks. Inter-annual Variability in Terrestrial Net Primary Production: Exploration of Trends and Controls on Regional to Global Scales. *Ecosystems*, 2(1):36–48, 1999.

31. C. Potter, P. Tan, M. Steinbach, S. Klooster, V. Kumar, R. Myneni, and V. Genovese. Major Disturbance Events in Terrestrial Ecosystems Detected using Global Satellite Data Sets. *Global Change Biology, accepted for publication*, 2003.

32. C. Potter, P. Zhang, S. Klooster, V. Genovese, S. Shekhar, and V. Kumar. Land Use - Understanding Controls on Historical River Discharge in the World’s Largest Drainage Basins. *Earth Interactions, accepted for publication*, 2003.

33. P. Rigaux, M. Scholl, and A. Voisard. *Spatial Databases: With Application to GIS*. Morgan Kaufmann Publishers, 2001.

34. J. Rushing, H. Ranganath, T. Hinke, and S. Graves. Using Association Rules as Texture Features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8), 2001.

35. H. Samet. *The Design and Analysis of Spatial Data Structures*. Addison-Wesley Publishing Company, Inc., 1990.

36. S. Shekhar and S. Chawla. *Spatial Databases: A Tour*. Prentice Hall, ISBN:0130174807, 2003.

37. S. Shekhar, S. Chawla, S. Ravada, A. Fetterer, X. Liu, and C.T. Lu. Spatial Databases: Accomplishments and Research Needs. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):45–55, 1999.

38. S. Shekhar and Y. Huang. Discovering Spatial Co-location Patterns: A Summary of Results . In *Proc.in 7th International Symposium on Spatial and Temporal Databases*, 2001.

39. S. Shekhar, C.T. Lu, and P. Zhang. Detecting Graph-Based Spatial Outlier: Algorithms and Applications(A Summary of Results). In *Proc. of the Seventh ACM-SIGKDD Int'l Conference on Knowledge Discovery and Data Mining*, Aug 2001.

40. S. Shekhar, P. Schrater, R. Vastsavai, W. Wu, and S. Chawla. Spatial Contextual Classification and Prediction Models for Mining Geospatial Data. *IEEE Transactions on Multimedia(Special Issue on Multimedia Database)*, 2002.

41. M. Steinbach, P. Tan, V. Kumar, S. Klooster, and C. Potter. Discovery of Climate Indices Using Clustering. In *Proc. of the 9th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining*, August 2003.

42. M. Steinbach, P. Tan, V. Kumar, C. Potter, and S. Klooster. Data Mining for the Discovery of Ocean Climate Indices. In *Proc of the Fifth Workshop on Scientific Data Mining*, 2002.

43. M. Steinbach, P. Tan, V. Kumar, C. Potter, and S. Klooster. Temporal Data Mining for the Discovery and Analysis of Ocean Climate Indices. In *Proc of the KDD Workshop on Temporal Data Mining*, 2002.

44. M. Steinbach, P. Tan, V. Kumar, C. Potter, S. Klooster, and A. Torregrosa. Clustering Earth Science Data: Goals, Issues and Results. In *Proc. of the Fourth KDD Workshop on Mining Scientific Datasets*, 2001.

45. H. V. Storch and F. W. Zwiers. *Statistical Analysis in Climate Research*. Cambridge University Press, July 1999.

46. P. Tan, M. Steinbach, V. Kumar, C. Potter, S. Klooster, and A. Torregrosa. Finding Spatio-Temporal Patterns in Earth Science Data. In *Proc. of KDD Workshop on Temporal Data Mining*, 2001.

47. G. H. Taylor. Impacts of the el nino/southern oscillation on the pacific northwest. Technical report, Oregon State University, Corvallis, Oregon, 1998.

48. W. R. Tobler. *Cellular Geography, Philosophy in Geography*. Gale and Olsson, Eds., Dordrecht, Reidel, 1979.

49. M. F. Worboys. *GIS - A Computing Perspective*. Taylor and Francis, 1995.

50. P. Zhang, Y. Huang, S. Shekhar, and V. Kumar. Correlation Analysis of Spatial Time Series Datasets: A Filter-and-Refine Approach. In *the Proc. of the 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2003.

51. P. Zhang, Y. Huang, S. Shekhar, and V. Kumar. Exploiting Spatial Autocorrelation to Efficiently Process Correlation-Based Similarity Queries. In *the Proc. of the 8th Int'l. Symp. on Spatial and Temporal Databases*, 2003.

52. P. Zhang, S. Shekhar, Y. Huang, and V. Kumar. Spatial Cone Tree: An Index Structure for Correlation-based Similarity Queries on Spatial Time Series Data. In *the Proc. of the Int'l Workshop on Next Generation Geospatial Information*, 2003.

Appendix: List of Tables

Table A.1 Connection of data mining techniques to Earth Science questions.

Earth Science Question	Examples of Data Mining Techniques
How is the Earth changing?	Principal Component Analysis (PCA), Cluster Analysis, Anomaly Detection, ARIMA time series modeling, Trend Detection, Change Point Detection
What factors cause these changes?	Correlation, Canonical Correlation Analysis, Association Analysis, Causal Analysis
Can we predict future changes?	Regression, correlation

Table A.2 Description of well-known climate indices.

Index	Description
SOI	(Southern Oscillation Index) Measures the SLP anomalies between Darwin and Tahiti
NAO	(North Atlantic Oscillation) Normalized SLP differences between Ponta Delgada, Azores and Stykkisholmur, Iceland
AO	(Arctic Oscillation) Defined as the first principal component of SLP poleward of 20° N
PDO	(Pacific Decadel Oscillation) Derived as the leading principal component of monthly SST anomalies in the North Pacific Ocean, poleward of 20°N
QBO	(Quasi-Biennial Oscillation Index) Measures the regular variation of zonal (i.e. east-west) stratospheric winds above the equator
CTI	(Cold Tongue Index) Captures SST variations in the cold tongue region of the equatorial Pacific Ocean (6°N-6°S, 180°-90°W)
WP	(Western Pacific) Represents a low-frequency temporal function of the ‘zonal dipole’ SLP spatial pattern involving the Kamchatka Peninsula, southeastern Asia and far western tropical and subtropical North Pacific
NINO1+2	Sea surface temperature anomalies in the region bounded by 80°W-90°W and 0°-10°S
NINO3	Sea surface temperature anomalies in the region bounded by 90°W-150°W and 5°S-5°N
NINO3.4	Sea surface temperature anomalies in the region bounded by 120°W-170°W and 5°S-5°N
NINO4	Sea surface temperature anomalies in the region bounded by 150°W-160°W and 5°S-5°N

Appendix: List of Figures

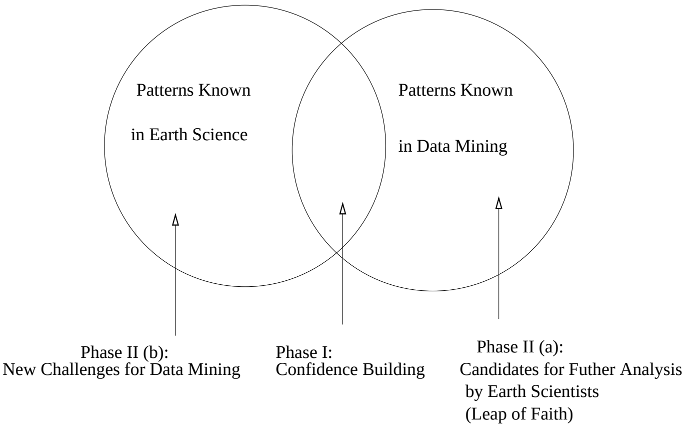


Fig. B.1 Using Data Mining to Find Earth Science Patterns

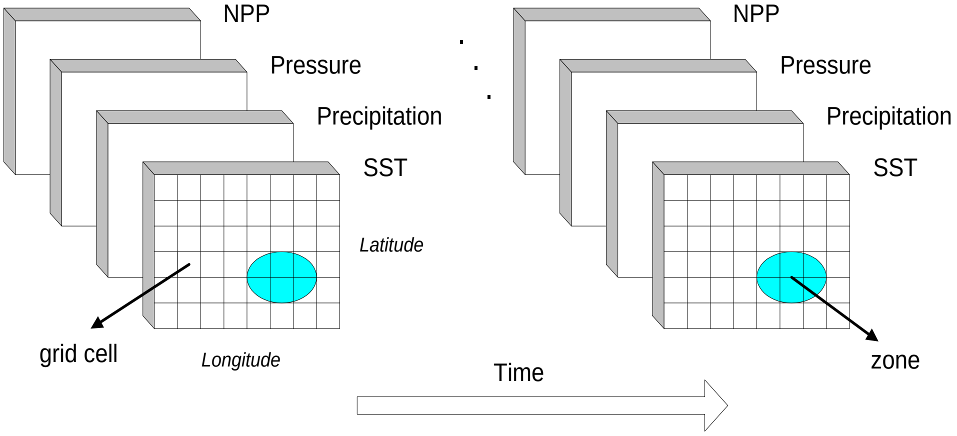
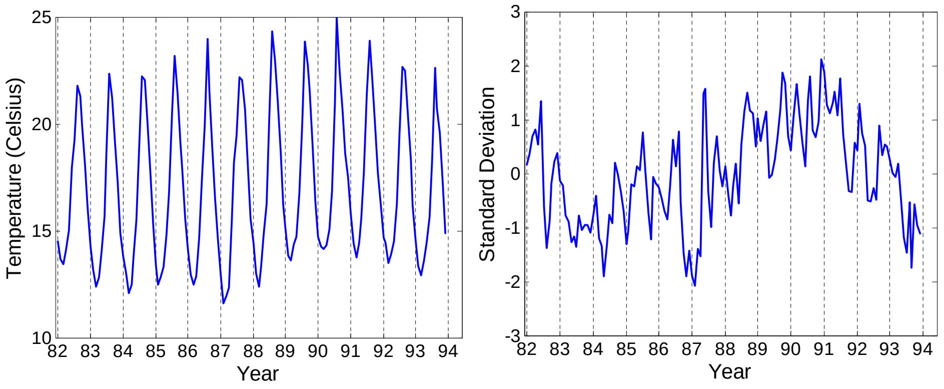


Fig. B.2 A simplified view of the problem domain



(a) Original SST Time Series

(b) Transformed SST Time Series

Fig. B.3 Monthly Z-score transformation applied to deseasonalize a sample SST time series

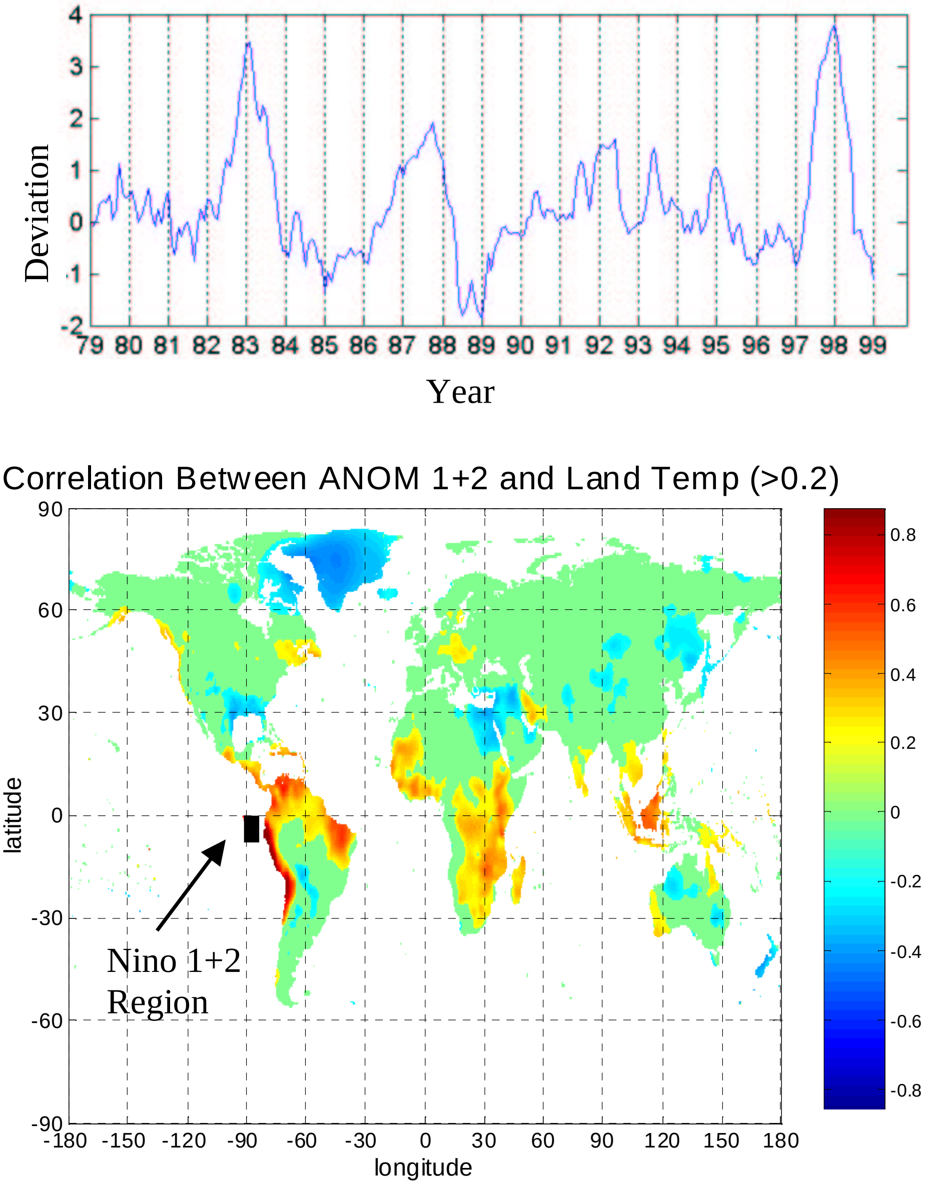


Fig. B.4 The NINO 1+2 climate index and its correlation to land temperature anomalies

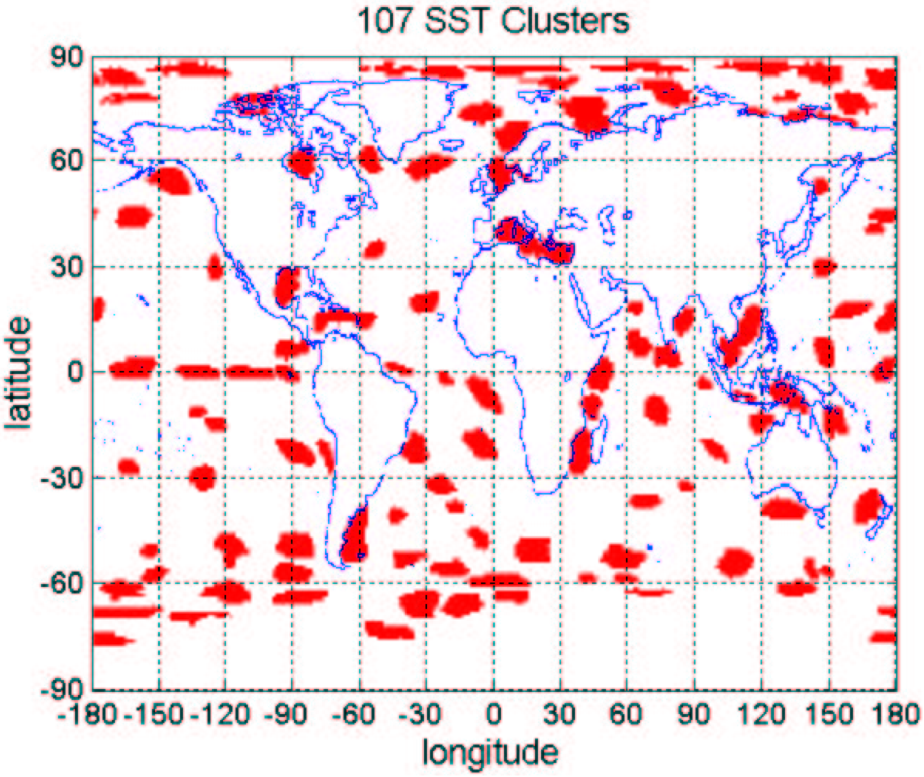


Fig. B.5 107 SST clusters

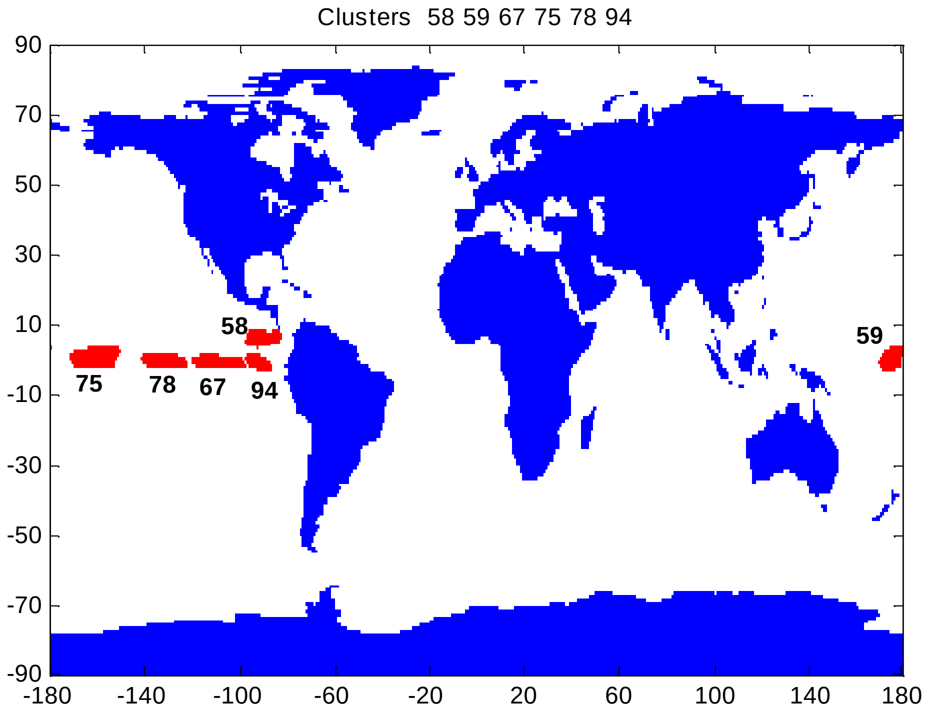
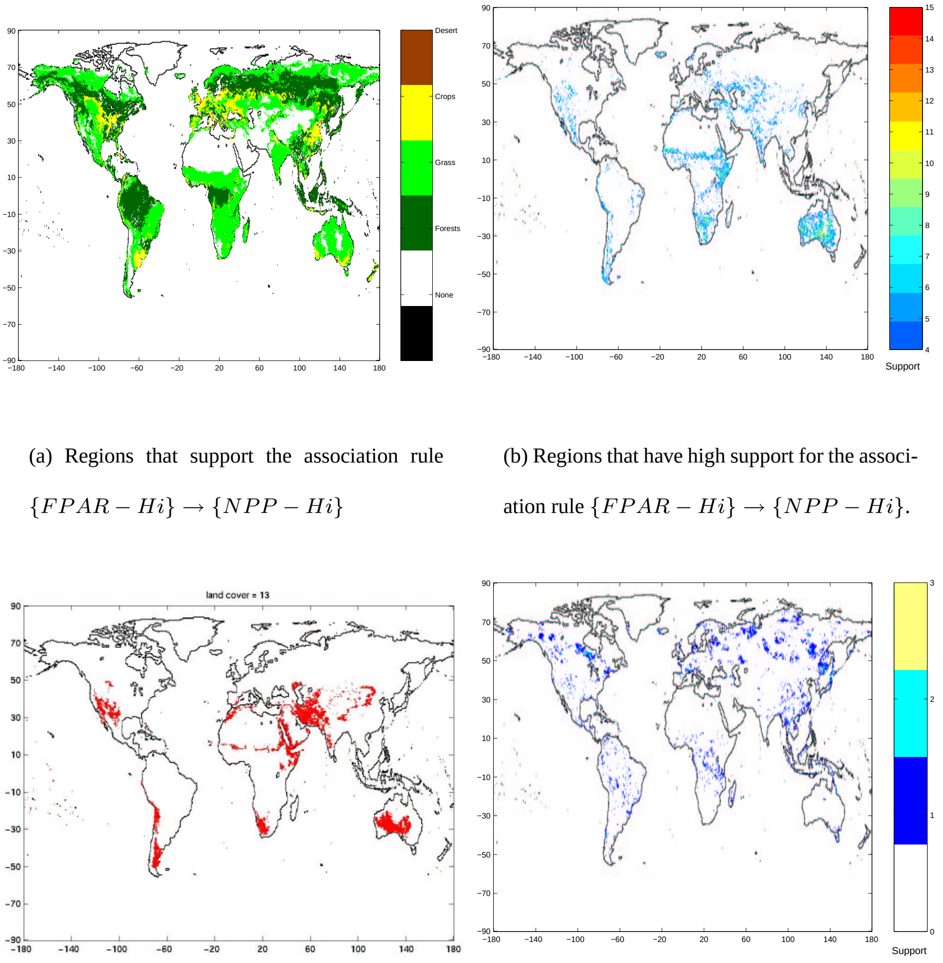


Fig. B.6 Clusters with correlation to known indices ≥ 0.8.



(a) Regions that support the association rule $\{FPAR-Hi\} \rightarrow \{NPP-Hi\}$

(b) Regions that have high support for the association rule $\{FPAR-Hi\} \rightarrow \{NPP-Hi\}$.

(c) Grassland and Shrubland Areas

(d) Regions that support the association rule $\{FPAR-Lo, PET-Lo\} \rightarrow \{NPP-Lo\}$

Fig. B.7 Visualizing regions that support the association rules

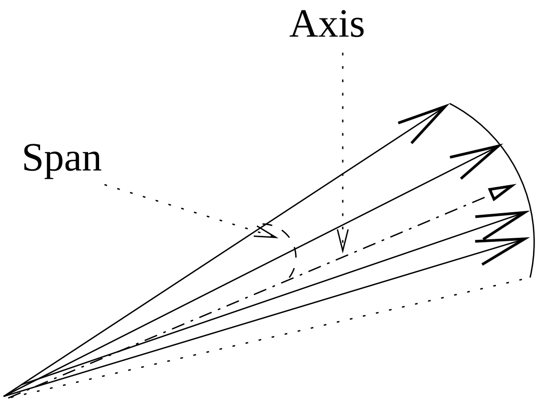


Fig. B.8 Illustration of a Cone

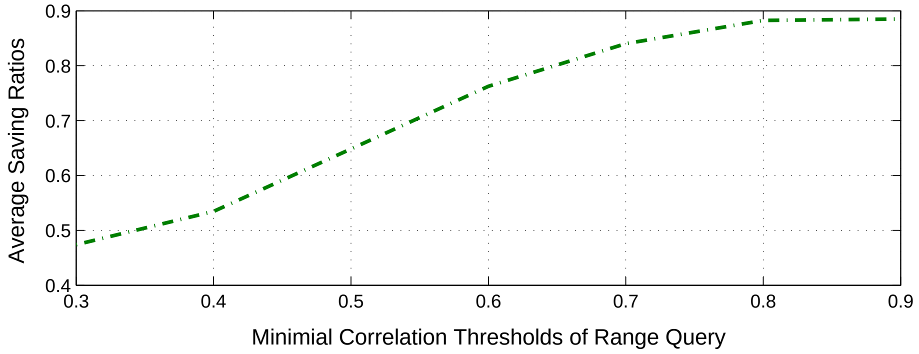


Fig. B.9 Savings for Query Processing

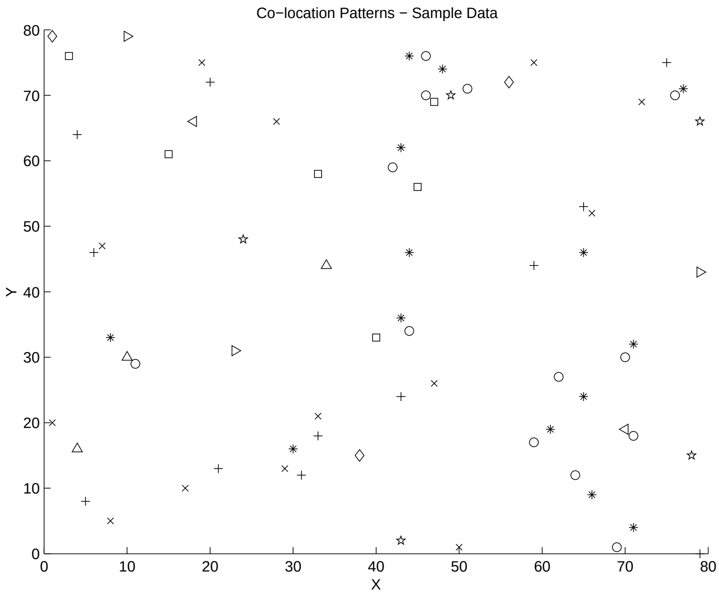


Fig. B.10 Illustration of Spatial Co-location Patterns. Shapes represent different spatial feature types. Spatial features in sets {‘+’, ‘×’} and {‘o’, ‘*’} tend to be located together.