Spatio-Temporal Pattern Detection in Climate Data

Daniel Levy
University of Toronto
dlevy@cs.toronto.edu

ABSTRACT

In this paper, a unique approach to the problem of spatio-temporal pattern detection is discussed in relation to climate data; this can be understood as discovering dependent climate events that occur over space. An accurate solution in this domain will provide climate scientists with highly valuable data which can be used to improve climate models and add to the knowledge base of climate science. This will in turn be beneficial to policy makers to make better informed decisions that can be based on improved data. This problem is one in which their are many valid solutions, which all must take into account the general problems that arise when dealing with big data, such as scaling and data complexity among other issues. The approach taken here calls upon the many concepts from the field of data-mining, and applies those concepts to the field of climate science to yield new and interesting knowledge. The concepts that will be discussed in this paper can also be used to locate and identify *events* in climate data-sets, which can be applied to improve the value of the climate data-sets that we have today by cross-referencing with other knowledge bases and data-sets.

1. INTRODUCTION

In the past decade, many institutions have attempted to make climate data from the past century publicly available. These data include measurements of temperature, precipitation, humidity and more. Some of the data has originated from hand-written paper records and has gone through rigorous levels of digitalization and validation to ensure accuracy. Scientists around the world have only recently begun to analyze this data and to scrape the surface of the knowledge hidden within. However, there is still much, much more that we can learn of the climate and its processes.

One particular area of interest, and the focus of this paper, is in the detection of spatio-temporal relations in climate data; these can essentially be described as related climate processes that take place in multiple locations in the earth either simultaneously, or sequentially. A greater understanding of this area will provide climatologists with insight on complex climatic patterns, and help discover new climatic processes. This new-found logic can then be used to improve our understanding of the climate and increase the accuracy of climate models, which are responsible for predicting climate scenarios many years ahead of time. These models are vital to policy makers to make informed decisions to prepare for our planet's future.

In addition, by improving our pattern matching techniques, we can perform content-based queries on data-sets, allowing us to detect all instances of any specified occurrence. Applying this concept to a data-set such as one containing the weather history in Canada during the past twenty years, we will be able to automate the identification of all instances of user-specified weather occurrences such as extreme-rainfall or droughts. This will not only improve the value of these data-sets by adding knowledge, but it will also allow us to cross-reference other data-sets that have additional data regarding these events.

A well-known example of the type of spatio-temporal processes we will be mining for, is the La Nina/El Nino phenomenon which occur once every few years. During a La Nina year, the sea surface temperature across the equatorial Eastern Central Pacific Ocean becomes 3-5 degrees lower than the average for that period of time in that year. Soon after the changes occur in the Eastern Central Pacific Ocean, there are simultaneous drought conditions in the Northern Pacific, flooding in northern South America, mild wet summers in northern North America, drought in the southeastern United States, and a wet period in the western United States among other climate effects. These events are all interrelated and caused by the initial cooling in the ocean. See Figure 1 for a visual representation of El Nino/La Nina.

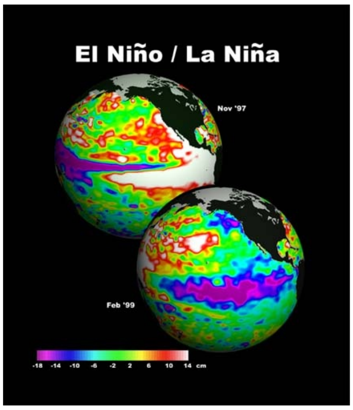

Figure 1: La Nina vs El Nino year. Image credit to http://sealevel.jpl.nasa.gov

There are many other spatio-temporal processes similar to La Nina/El Nino that are buried deeply under the complex layers of climate data. Currently, most of these types of processes are discovered by a physical observation of the phenomenon, which can be difficult when it is unknown exactly what is being looked for, and when the data is not fully understood by even our brightest climate scientists. Occasionally, discovering new patterns can even come down to pure luck. But what if we were to take a more automated and scientific approach of discovering these phenomena using data-mining algorithms and techniques that have been so successful in many discoveries in other fields such as bioinformatics and social analytics? The potential to enhance our knowledge about the inner-workings of the world we live in is very large.

In this paper, a new scalable data-mining algorithm is proposed for mining climate data-sets to detect interesting spatio-temporal patterns. The paper is organized as follows: in Section 2 we discuss how the algorithm works in its entirety from a relatively high-level perspective. Section 3 discusses how we can compare two climate regions to each other with a resultant similarity metric. Section 4 discusses the approach taken to cluster the climate data-points to reduce computational complexity. Section 5 discusses some of the preliminary testing results and how we plan to experiment on the algorithm in the future. Section 6 discusses related works and some features of the algorithm that we will be looking to improve in the near future. Finally, Section 7 provides final remarks and concludes the paper.

2. PATTERN ANALYSIS

2.1 Overview

Climate data generally comes in the form of sets of latitude and longitude coordinates accompanied with attributes such as temperature, precipitation, humidity, wind speed etc, all mapped to a time-stamp. Timestamps can vary from hours to days to months depending on the data-set. Many interesting insights can be made by having a comprehensive and in-depth understanding of this data-set, but it is simply too complex to fully comprehend without the aid of man's second best friend, the CPU. In order for a data mining algorithm to properly learn and understand patterns in a climate data-set, the data-set must (1) be large enough for patterns to appear frequently (2) have a small enough time-scale such that the patterns are not overlooked, i.e the patter does not occur on a time-scale smaller then what the data-set provides (3) have highly accurate data.

The chosen data-set that the initial implementation of the algorithm was tested on is publicly available, provided by the University of Washington. It contains highly accurate daily climate data from 1949 - 2000 on a 1/8-degree grid.

We will now discuss the algorithm by breaking it into four key steps:

2.2 First Step: Identifying an Anomaly

The first step in the algorithm is an external one. A climate scientist that is studying a specific region in the world may stumble upon a section of climate data that does not make sense, or one that stands out. He may ask questions such as 'why is this happening?' or 'what is the cause of this?'. Currently, his options to answer his questions are severely limited barring that he cannot extract an answer based on our current knowledge of climate science. By providing the region and date range of the anomaly into our algorithm, he can learn more about its causes and effects, and perhaps understand the science behind why the anomaly is occurring.

2.3 Second Step: Identifying Additional Occurrences of the Anomaly

What we have from the previous step: An anomaly contained in a region associated with a date range.

We now attempt to identify other regions in the world where the same anomaly has occurred in any valid time period from our data-set (years 1949 -2000 for the data-set we are using). A search is performed using the spatial similarity algorithm outlined in Section 3. We minimize our search time in two ways:

1. Focusing on the time of the yearly cycle on which the first anomaly has occurred
2. Leveraging data 'tags' that we have created in the preprocessing step (discussed in Section 3) to capture recurring anomalous and non-anomalous patterns. These are sequences of events stored in a cyclic data-structure, quickly found using identifying characteristics of the anomaly.

2.4 Third Step: Discovering Related Events

What we have from the previous step: A list of regions and associated dates where a specified anomaly has occurred.

We now wish to discover causes and effects for the given anomaly to obtain a deeper understanding of it. To accomplish this, we search for recurring events that have transpired adjacent to the region where the anomaly is located on and around the same time period in the yearly cycle.

Let's review this with an example. Imagine we have an anomaly that has occurred in the Toronto region in January 1963 and again in '72, '88 and '96. On and around the listed dates, we will search for additional recurring events in adjacent regions, say Waterloo for instance (i.e we will search in the years '63, '72, '88, '96). A positive result could be be a recurring event in Waterloo in February of 1972 and '88.

Adjacent regions are identified using the optimized Minimum Bounding Rectangle. When a positive result is found, we continue our search, recursively searching through the newly identified regions until the pattern trail is lost or until we have completed n iterations of searching (we will prune the path in Step Four and return later if it is likely that there are more than n iterations). We make no attempt to predict the movement of the pattern, to avoid introducing bias into the algorithm. Predicting the pattern assumes we understand the relations between them, which often is not the case.

2.5 Fourth Step: Pruning

What we have from the previous step: A specific anomaly with the regions and dates in which it has occurred, along with a set of other events that frequently happen in or around the same time period as the anomaly. Let these events be stored in set C.

We now ensure two things for the elements in set C:

1. They are truly associated with the anomaly, i.e these same events do NOT occur when the anomaly does not occur.
2. None of the events are due to the seasonal cycle.

This will prune set C, ensuring that irrelevant events are removed. Once C is pruned, we return to Step Three, until no new events are found.

3. SPATIAL SIMILARITY

3.1 Overview

The concept of similarity is paramount in many data-mining and clustering algorithms. We often need a metric of how similar or dissimilar two objects are to each other. In this paper, we require the concept of similarity for clustering purposes (Section 4) and for identifying sub-paths and regions with high similarity in climate trends.

Two regions are spatially similar if they exhibit 'similar' behavior within a specified tolerance in their measured climatic attributes (temperature, precipitation, humidity, wind speed etc). Yes, there is the dreaded 'using the word being defined in the definition itself' problem, however in this context, the meaning should be well understood.

Let us simplify the problem to 2 dimensions for an example: time and temperature. Assume our data-set S of temperatures over time consists of daily temperatures for the Toronto region in the year 2013. We have another much smaller data-set P that consists of temperatures for the Toronto region between February 2nd 1992, and February 13th 1992. We wish to find whether set P is contained within set S with an acceptable level of tolerance ϵ. To achieve this, we require some way of comparing P to sections of S. Set P could be contained within S numerous times (min: 0, max: length(S) / length(P)). See Figure 2 for an example of a pattern P being identified in a set S.

While defining an algorithm to detect spatial similarity, we need to be concerned with a number of things:

1. How do we handle spatial similarity when the event we are comparing occurs faster or slower during different years? For example, one year a specific phenomenon takes place over a 3 day period, however 4 years later, the same phenomenon takes place over a 6 day period?

2. What if we found a strong matching for P, however there exists some large outliers in a small portion of the data? (See Figure 3 for an example)

3. How do we deal with erroneous climate data that may throw off the algorithm?

4. How do we approach 'relative' similarity between the amplitudes of S and P? As an example (once again using temperature and time as the two dimensions for simplicity) if two events had the same spatial curvature, but one event has a mean value that is 10 degrees higher then the other.

3.2 The Algorithm

The spatial similarity algorithm begins with a preprocessing step in which the two dimensional graphs of all climate attributes versus time are transformed into linear representational segments (discussed in Section 3.3). This reduces the computational complexity of the comparison, while retaining the most vital information about the data. Additionally, this also allows us to incorporate a certain degree of logic into making the break points of the segments meaningful. See Figure 4 for an example.

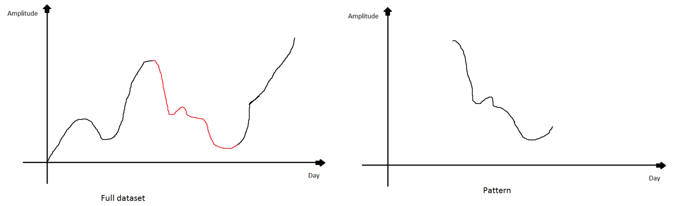

Figure 2: Example of pattern matching within a data-set

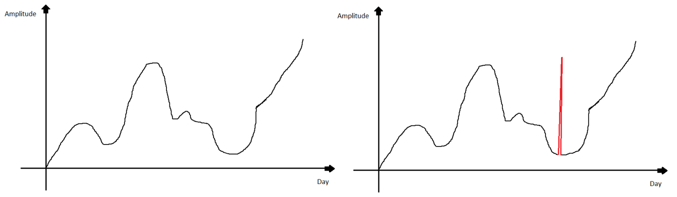

Figure 3: A High similarity with large outlier

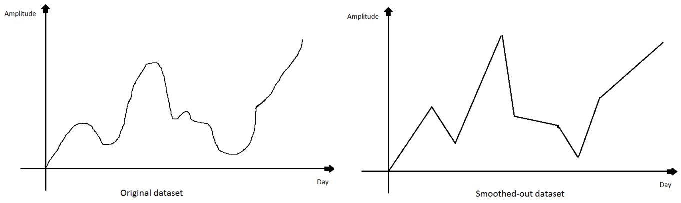

Figure 4: Moving from the original data-set to a a linear transformation

This transformation provides us with a clearer image of minimum and maximum points, and sharp slope changes, while removing 'noise' values that we wish to ignore (however we still keep track of noise in the event that it proves to be useful and we need to restore it). We store these values in the database as sets of slopes and lengths.

Once the transformation is completed on both our data-set S, and our input pattern P, we can iterate through the entirety of S to identify whether P is contained within. This is accomplished by sequentially mapping all slopes in P to slopes in S with a relativistic mapping of slope lengths within the accepted threshold of ϵ. We repeat this process across all dimensions of the data, and optimize by taking note of the events that highly contrast the base-line (discussed in Section 3.3) and ensuring these events are found in both set S and set P.

To further reduce computational complexity, we tag common sequences of patterns in the data during the preprocessing step so that we can quickly refer to or ignore them. Additionally, we use a cyclic data-structure that can quickly identify commonly reoccurring types of anomalies, quickly found using identifying characteristics of the anomaly.

3.3 Linear Transformation

We would like to have the capability to transform the graphs of climate attributes versus time into linear segments that better represent the graph, and in addition create a moving 'baseline' which will represent the most acceptable value for the attribute at any given moment. See Figure 5 for an example of a graph with two baselines.

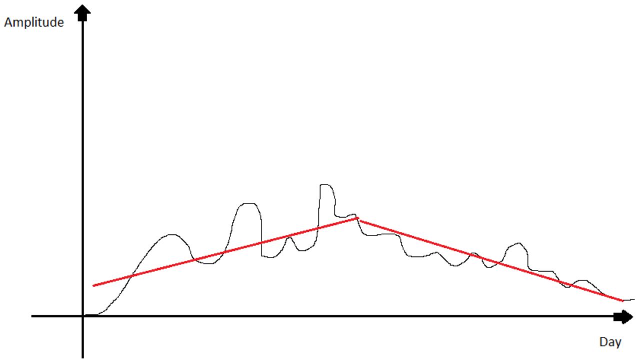

Figure 5: Example of a representative moving baseline

Once again, we will use temperature versus time to explain the procedure using an example. Let the set $S_i = (x_i, y_i)$ represent the amplitudes of the points in our graph. We begin by cycling through each point sequentially and creating a linear least square fit (represented by $f(x)$) by *minimizing* the following equation:

$$R^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{1}$$

where $f(x) = \alpha + \beta x_i$ is the resulting linear fit. We can minimize the equation by applying the following conditions:

$$\frac{\partial R^2}{\partial \alpha} = 0 \; where \; i = 1..n \tag{2}$$

$$\frac{\partial R^2}{\partial \beta} = 0 \; where \; i = 1..n \tag{3}$$

While we are calculating the best fit, we simultaneously calculate the integral between the best fit line and our graph (example: shaded area in Figure 6 A).

$$\int_{i=0}^{n} f(x) - \int_{j=0}^{n} S_j \tag{4}$$

When the value of the integral exceeds our pre-set threshold of ϵ we attempt to create a new baseline. However, we associate a cost with creating new baselines; therefore, this reduces to a problem of minimizing both the number of baselines and the value of the integral between the baseline and the integral.

Lets take a look at a concrete example to put this into perspective. In Figure 6 A we have a single baseline representing the entire length of a graph. While adding additional baselines would greatly reduce the value of the integrals, the cost of creating multiple new lines for each peak would be greater then the cost of having a larger integral value. In Figure 6 B we have 2 baselines, because by adding a second baseline we can greatly reduce the value of the integral between the baselines and the graph.

4. CLUSTERING

The amount of calculations that are required to be performed to study climatic patterns are vast given the sheer number of data-points provided from the weather stations in which the climate data has been recorded. To simplify this problem, we create clusters of weather stations with similar climatic behavior, using the spatial similarity algorithm from section 3. These clusters will also be referred to as 'regions' throughout this paper.

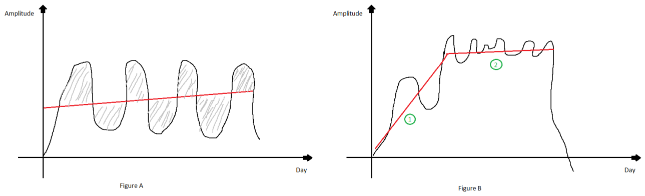

Figure 6: Example of a representative moving baseline

Besides reducing computation complexity, clustering is beneficial for two other reasons:

1. Reducing error - A single station might have erroneous data for a number of dates due to misreadings, miscalibrations, disruptive particles in the air etc. Clustering can help reduce the impact of such erroneous data.

2. Localizing the phenomenons - The spatio-temporal relationships we are seeking with this algorithm are more likely to exist over regions of space versus single points in space.

The most applicable algorithms for spatio-temporal clustering are PAM, CLARA, CLARANS, agglomerative clustering and k-means [SK12]. As is common with applying any clustering algorithm, each choice has advantages and disadvantages. For the purposes of this paper, we keep the decision simple with the use of agglomerative clustering as it does not require initial clusters to be selected to perform well, or the number of desired clusters; this is beneficial because it avoids the introduction of user bias into the selection of cluster locations, which could lead to unexpected results, and avoids poor selections of clusters such as near areas with high variance (mountains/flat-land, rivers/land etc).

To perform the clustering, we randomly select initial cluster seeds (note - the location and number of seeds should have minimal impact) and then proceed to identify adjacent stations through the use of an optimized Minimum Bounding Box search (using R-Tree). Adjacent stations with short distance and high spatial similarity are merged into the cluster of the original station. As the cluster grows, the distance is measured to the center of the cluster, and similarity is taken to the cluster as a whole. Different combinations are attempted recursively to optimize the cluster, and minimize its variance. Given that the variance of the stations in a cluster becomes too high, we attempt to split the cluster.

Clustering is performed only once, using climate data from approximately a sixty day period. To increase the accuracy of the clusters (as they remain constant afterwords), we randomly prod the data-set in other locations (i.e many years in the future) to ensure cluster similarity persists over time. This approach has flaws which are discussed in section 6 and will be improved upon in a future version of this paper.

5. PRELIMINARY RESULTS AND FUTURE EXPERIMENTATION AND VALIDATION

As this work is currently ongoing as of September 2013, the experimental results we can express at this point are limited. The experiment that has been completed at this time uses a cluster seed located at latitude/longitude points 46, 77 in the first half of January, 1949. A resulting related cluster was discovered at latitude/longitude points 52, 121. The temperature comparison can be seen in Figure 7. Ofcourse, the other climatic values (precipitation, wind speed etc) are also highly similar between these two sets of latitude/longitude coordinates. A visual representation of the latitude/longitude coordinates and their respective clusters can be viewed in Figure 8.

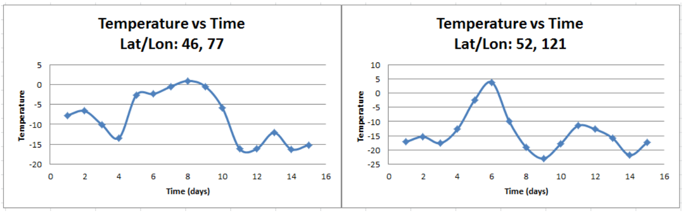

Figure 7: Temperature pattern comparison

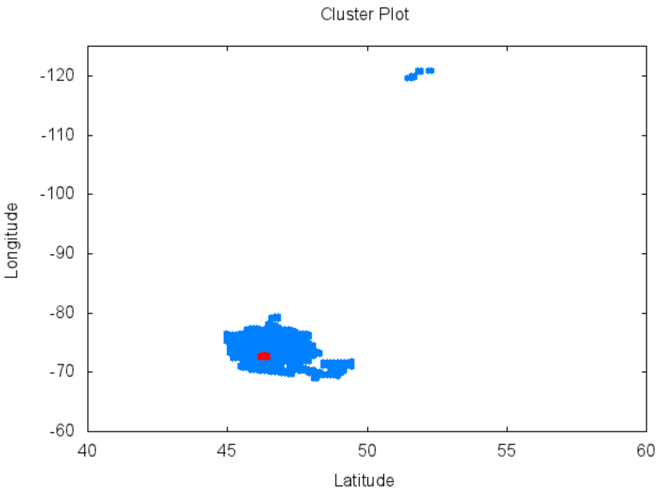

Figure 8: Results from an input cluster in 1949

In a future version of this paper, we will tackle experimentation and validation of results by examining the algorithm's accuracy against the initial seed condition of El Nina/La Nina and ensure that it mines the majority of the pattern. Next, we will choose 5-10 other well-known spatio-temporal patterns and compare results.

6. RELATED WORKS AND FUTURE WORKS

Anomaly pattern detection is not a new field, but is seeing a lot of research recently due to the popularity of the big data and biomedical fields (among others) that make use of these techniques. One such example from the biomedical world is from [WMCW02] which uses a Bayesian network to detect recurring anomalous patterns for disease outbreaks. One of the earlier attempts to mine association rules from climate data is the Geominer project [HKS97] from 1997. It unsuccessfully attempted to mine characteristic rules, comparison rules and association rules from a climate data set. The majority of the more recent papers on the topic of spatio-temporal pattern detection in climate data are focused or specialized on a specific area and cannot detect all type of pattern, just ones they have been calibrated and setup to understand. [DNK09] for example, searches for patterns in extreme rainfall events in India, however it somewhat simplifies the process by only studying 'extreme' events that can be easily identified in the data. Or [HLDT02], which mines association rules that are strictly related to drought conditions. [KSnT$^+$01] contains interesting and highly relatable work, however they only explore patterns that occur in a single region, whereas in this paper we are focusing on cross-region similarity.

The most vital feature to the accuracy of this paper is the spatial similarity measure. The algorithm for spatial similarity proposed in Section 3, despite having strong performance and accuracy, is not perfect by any means, and can miss certain cases of similarity. There are a large number of alternative algorithms and data transformations that are also applicable, and more work needs to be performed to see if they should be incorporated into this work in this paper. However, most of them, if not all have pitfalls of their own. Some examples include using the Discrete Fourier Transform (DFT) or Independent Component Analysis (ICA) as discussed in [KSnT$^+$01] and [BST$^+$04] respectively to remove seasonal data such that it is easier to detect patterns; [ZSM$^+$11]'s interesting work on detecting 'sub-patterns' using an enumeration and pruning approach defined by an interest measure, mainly to detect abrupt change; and a constraint-based similarity query proposed by [GK95]. Perhaps a dynamic choice of the most accurate similarity algorithm for the scenario at hand could be developed.

One current issue, and the main focus of the next iteration of this paper is in the way clustering is performed. Clusters are created using the similarity measure with data from the time span of a sixty day period (Section 4), while incorporating adjustments by randomly prodding the rest of the data-set. As one might imagine, the clusters *should* evolve over time as the similarity measures change, but for simplicities sake and calculation cost, we have assumed that the clusters remain constant in time. [NH94] discusses a new and efficient method for clustering spatial data called CLARANS which is based of a randomized search and requires a predefined number of clusters, but it is very fast, meaning its use might make live clustering a possibility without hindering performance to much.

A summary on some of the leading anomaly detection algorithms can be found in [CBK09]. These ideas will be looked at more in depth for automating the anomaly identification process (Section 2.2) in a future version of this paper.

More work is also required to incorporate ideas from recent advances in association rule mining for large data-sets such as those discussed in [AS94] which uses a modified and very fast version of the classic Apriori algorithm.

7. CONCLUSIONS

This paper has laid out the groundwork for a novel algorithm in the area of spatio-temporal data mining for climate patterns by leveraging the sheer amount of climate data that we have available today. The goal is that this work will aid climate scientists to improve climate models by discovering new patterns and learning about patterns that are not well understood. A second significant usage will be to provide the capabilities for public and private climate data-sets to catalog events that are already well understood from a scientific perspective through pattern matching, and have multiple data-sets refer and cross-reference. This will improve the value and enrich the climate data that we have today. While this work is still far from proving that interesting results can be found, the preliminary work shows promise.

8. REFERENCES

Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

Jayanta Basak, Anant Sudarshan, Deepak Trivedi, M. S. Santhanam, Te won Lee, and Erkki Oja. Weather data mining using independent component analysis. *Journal of Machine Learning Research*, 5:239–253, 2004.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, July 2009.

C. T. Dhanya and D. Nagesh Kumar. Data mining for evolution of association rules for droughts and floods in india using climate inputs. *Journal of Geophysical Research: Atmospheres*, 114(D2):n/a–n/a, 2009.

Dina Q. Goldin and Paris C. Kanellakis. On similarity queries for time-series data: Constraint specification and implementation. In *Proceedings of the First International Conference on Principles and Practice of Constraint Programming*, CP '95, pages 137–153, London, UK, UK, 1995. Springer-Verlag.

Jiawei Han, Krzysztof Koperski, and Nebojsa Stefanovic. Geominer: A system prototype for spatial data mining. pages 553–556, 1997.

Sherri Harms, Dan Li, Jitender Deogun, and Tsegaye Tadesse. Efficient rule discovery in a geo-spatial decision support system. In *Proceedings of the 2002 annual national conference on Digital government research*, dg.o '02, pages 1–7. Digital Government Society of North America, 2002.

Vipin Kumar, Michael Steinbach, Pang ning Tan, Steven Klooster, Christopher Potter, and Alicia Torregrosa. Mining scientific data: Discovery of patterns in the global climate system. In *In Proceedings of the Joint Statistical Meetings (Athens, GA, Aug. 59). American Statistical Association*, 2001.

Raymond T. Ng and Jiawei Han. Efficient and effective clustering methods for spatial data mining. pages 144–155, 1994.

Sudheer Reddy Santhosh Kumar, Sitha Ramulu. Spatial data mining using cluster analysis. *International Journal of Computer Science and Information Technology*, 4(4), 2012.

Weng-Keen Wong, Andrew Moore, Gregory Cooper, and Michael Wagner. Rule-based anomaly pattern detection for detecting disease outbreaks. In *Eighteenth national conference on Artificial intelligence*, pages 217–223, Menlo Park, CA, USA, 2002. American Association for Artificial Intelligence.

Xun Zhou, Shashi Shekhar, Pradeep Mohan, Stefan Liess, and Peter K. Snyder. Discovering interesting sub-paths in spatiotemporal datasets: a summary of results. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pages 44–53, New York, NY, USA, 2011. ACM.