Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 EECS Building
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 02-037

Correlation Analysis of Spatial Time Series Datasets: A Filter-and-Refine Approach

Pusheng Zhang, Yan Huang, Shashi Shekhar, and Vipin Kumar

December 03, 2002

Correlation Analysis of Spatial Time Series Datasets: A Filter-and-Refine Approach †

Pusheng Zhang*, Yan Huang, Shashi Shekhar, Vipin Kumar
Computer Science Department, University of Minnesota
200 Union Street SE, Minneapolis, MN-55455
[*pusheng*, *huangyan*, *shekhar*, *kumar*]@cs.umn.edu

December 3, 2002

Abstract

A spatial time series dataset is a collection of time series, each referencing a location in a common spatial framework. Correlation analysis is often used to identify pairs of interacting elements from the cross product of two spatial time series datasets. However, the computational cost of correlation analysis is very high when the dimension of the time series and the number of locations in the spatial frameworks are large. The key contribution of this paper is the use of spatial autocorrelation among spatial neighboring time series to reduce the computational cost. A filter-and-refine algorithm based on coning, i.e. group of locations, is proposed to reduce the cost of correlation analysis over a pair of spatial time series datasets. Cone-level correlation computation can be used to eliminate (filter out) a large number of element pairs whose correlation is clearly below (or above) a given threshold. Element pair correlation needs to be computed for remaining pairs. Using algebraic cost models and experimental studies with Earth science datasets, we show that the filter-and-refine approach can save a large fraction of the computational cost, particularly when the minimal correlation threshold is high.

Keywords: Spatial Time Series, Correlation Analysis, Filter-and-refine, Spatial Autocorrelation

†This work was partially supported by NASA grant No. NCC 2 1231 and by Army High Performance Computing Research Center contract number DAAD19-01-2-0014. The content of this work 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.

*The contact author. E-mail: pusheng@cs.umn.edu. Tel: (612) 626-7515

1 Introduction

Spatio-temporal data mining [14, 15, 17, 16, 18, 13, 20, 7] is important in many application domains such as epidemiology, ecology, climatology, or census statistics, where datasets which are spatio-temporal in nature are routinely collected. The development of efficient tools [1, 4, 8, 10, 11] to explore these datasets, the focus of this work, is crucial to organizations which make decisions based on large spatio-temporal datasets.

A spatial framework [21] consists of a collection of locations and a neighbor relationship. A time series is a sequence of observations taken sequentially in time [2]. A spatial time series dataset is a collection of time series, each referencing a location in a common spatial framework. For example, the collection of global daily temperature measurements for the last 10 years is a spatial time series dataset over a degree-by-degree latitude-longitude grid spatial framework on the surface of the Earth.

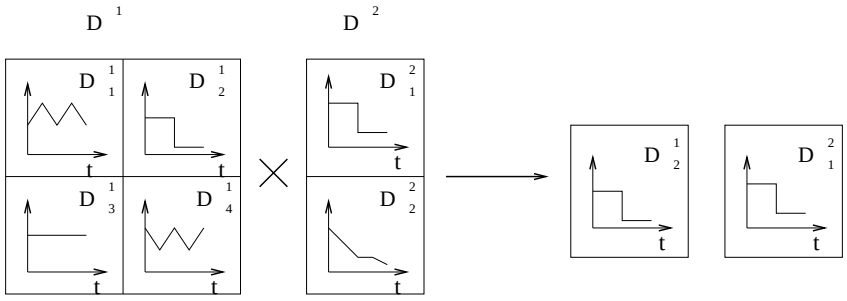


Figure 1: An Illustration of the Correlation Analysis of Two Spatial Time Series Datasets

Correlation analysis is important to identify interacting pairs of time series across two spatial time series datasets. A strongly correlated pair of time series indicates potential movement in one series when the other time series moves. For example, El Nino, the anomalous warming of the eastern tropical region of the Pacific, has been linked to climate phenomena such as droughts in Australia and heavy rainfall along the Eastern coast of South America [19]. Figure 1 illustrates the correlation analysis of two spatial time series datasets D^1 and D^2. D^1 has 4 spatial locations and D^2 has 2 spatial locations. The cross product of D^1 and D^2 has 8 pairs of locations. A highly correlated pair, i.e. (D^1_2, D^2_1), is identified from the correlation analysis of the cross product of the two datasets.

However, a correlation analysis across two spatial time series datasets is computationally expensive when the dimension of the time series and number of locations in the spaces are large. The computational cost can be reduced by reducing time series dimensionality or reducing the number of time series pairs to be tested, or both. Time series dimensionality reduction techniques include discrete Fourier transformation [1], discrete wavelet transformation [4], singular vector decomposition [6], etc.

The number of pairs of time series can be reduced by a cone-based filter-and-refine approach which groups similar time series within each dataset together. A filter-and-refine approach has two logical phases. First, the filtering phase groups similar time series as cones in each dataset and calculates the centroids and boundaries of each cone. These cone parameters

allow computation of the upper and lower bounds of the correlations between the time series pairs across cones. Many All-True and All-False time series pairs can be eliminated at the cone level to reduce the set of time series pairs to be tested by the refinement phase. We propose to exploit an interesting property of spatial time series datasets, namely spatial auto-correlation [5], which provides a computationally efficient method to determine cones. We use spatial auto-correlation measurement tools such as correlograms [5] to identify cone size. Experiments with Earth science data [12] and an algebraic cost model show that the filter-and-refine approach can save a large fraction of computational cost especially when the minimal correlation threshold is high. To the best of our knowledge, this is the first paper exploiting spatial auto-correlation among time series at nearby locations to reduce the computational cost of correlation analysis over a pair of spatial time series datasets.

Scope and Outline: In this paper, the computation saving methods focus on reduction of time series pairs to be tested. Methods based on non-spatial properties (e.g. time-series power spectrum [1, 4, 6]) are beyond the scope of the paper and will be addressed in future work.

The rest of the paper is organized as follows. In Section 2, the basic concepts and lemmas about cone boundaries are provided, and Section 3 proposes our filter-and-refine algorithm. The cost model is proposed in Section 4, and the experimental design and results are presented in Section 5. We summarize our work and discuss future directions in Section 6.

2 Basic Concepts

In this section, we introduce the basic concepts of correlation calculation and the multi-dimensional unit sphere formed by normalized time series. We define the cone concept in the multi-dimensional unit sphere and prove two lemmas to bound the correlation of pairs of time series from two cones.

2.1 Correlation and Test of Significance of Correlation

Let $x = \langle x_1, x_2, \ldots, x_m \rangle$ and $y = \langle y_1, y_2, \ldots, y_m \rangle$ be two time series of length m. The correlation coefficient [3] of the two time series is defined as:

$$corr(x, y) = \frac{1}{m-1} \sum_{i=1}^{m} (\frac{x_i - \overline{x}}{\sigma_x}) \cdot (\frac{y_i - \overline{y}}{\sigma_y}) = \widehat{x} \cdot \widehat{y}$$

where $\overline{x} = \frac{\sum_{i=1}^{m} x_i}{m}$, $\overline{y} = \frac{\sum_{i=1}^{m} y_i}{m}$, $\sigma_x = \sqrt{\frac{\sum_{i=1}^{m} (x_i - \overline{x})^2}{m-1}}$, $\sigma_y = \sqrt{\frac{\sum_{i=1}^{m} (y_i - \overline{y})^2}{m-1}}$, $\widehat{x_i} = \frac{1}{\sqrt{m-1}} \frac{x_i - \overline{x}}{\sigma_x}$, $\widehat{y_i} = \frac{1}{\sqrt{m-1}} \frac{y_i - \overline{y}}{\sigma_y}$, $\widehat{x} = \langle \widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_m \rangle$, and $\widehat{y} = \langle \widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_m \rangle$.

A simple method to test the null hypothesis that the product moment correlation coefficient is zero can be obtained using a Student's t-test [3] on the t statistic as follows: $t = \sqrt{m-2} \frac{r}{\sqrt{1-r^2}}$, where r is the correlation coefficient between the two time series. The freedom degree of the above test is $m - 2$. Using this we can find a $p - value$ or find the critical value for a test at a specified level of significance. For a dataset with larger length m, we can adopt Fisher's Z-test [3] as follows: $Z = \frac{1}{2} \log \frac{1+r}{1-r}$, where r is the correlation coefficient between the two time

series. The correlation threshold can be determined for a given time series length and confidence level

2.2 Multi-dimensional Sphere Structure

In this subsection, we discuss the multi-dimensional unit sphere representation of time series. The correlation of a pair of time series is related to the cosine measure between their unit vector representations in the unit sphere.

Fact 1 (Multi-dimensional Unit Sphere Representation) *Let* $x = \langle x_1, x_2, \ldots, x_m \rangle$ *and* $y = \langle y_1, y_2, \ldots, y_m \rangle$ *be two time series of length* m. *Let* $\widehat{x}_i = \frac{1}{\sqrt{m-1}} \frac{x_i - \overline{x}}{\sigma_x}$, $\widehat{y}_i = \frac{1}{\sqrt{m-1}} \frac{y_i - \overline{y}}{\sigma_y}$, $\widehat{x} = \langle \widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_m \rangle$, *and* $\widehat{y} = \langle \widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_m \rangle$. *Then* $\widehat{x}$ *and* $\widehat{y}$ *are located in the surface of a multi-dimensional unit sphere and* $corr(x, y) = \widehat{x} \cdot \widehat{y} = \cos(\angle(\widehat{x}, \widehat{y}))$ *where* $\angle(\widehat{x}, \widehat{y})$ *is the angle of* $\widehat{x}$ *and* $\widehat{y}$ *in* $[0, 180^\circ]$ *in the multi-dimensional unit sphere .*

Because the sum of the $\widehat{x}_i^2$ is equal to 1: $\sum_{i=1}^{m} \widehat{x}_i^2 = \sum_{i=1}^{m} (\frac{1}{\sqrt{m-1}} \frac{x_i - \overline{x}}{\sqrt{\frac{\sum_{i=1}^{m} (x_i - \overline{x})^2}{m-1}}})^2 = 1$, $\widehat{x}$ is located in the multi-dimensional unit sphere. Similarly, $\widehat{y}$ is also located in the multi-dimensional unit sphere. Based on the definition of $corr(x, y)$, we have $corr(x, y) = \widehat{x} \cdot \widehat{y} = \cos(\angle(\widehat{x}, \widehat{y}))$.

Lemma 1 (Correlation and Cosine) *Given two time series* x *and* y *and a user specified minimal correlation threshold* θ *where* $0 < \theta \leq 1$, $|corr(x, y)| = |\cos(\angle(\widehat{x}, \widehat{y}))| \geq \theta$ *if and only if* $0 \leq \angle(\widehat{x}, \widehat{y}) \leq \theta_a$ *or* $180^\circ - \theta_a \leq \angle(\widehat{x}, \widehat{y}) \leq 180^\circ$, *where* $\theta_a = \arccos(\theta)$ *and* $0 \leq \theta_a \leq 90^\circ$.

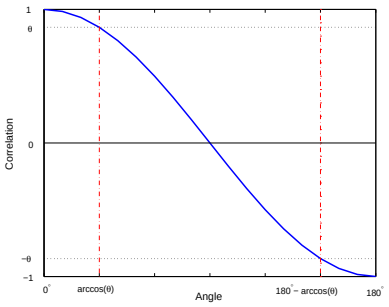


Figure 2: Cosine Value vs. Central Angle

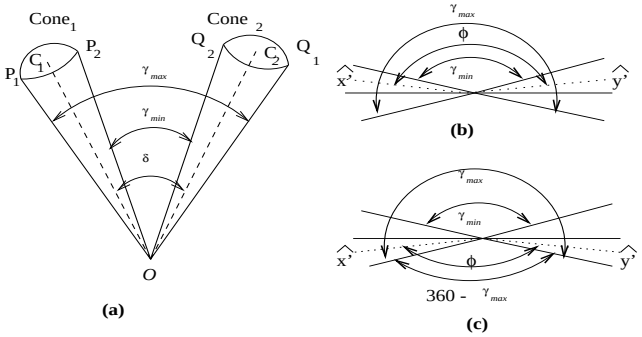


Figure 3: Angle of Time Series in Two Spherical Cones

Proof: Figure 2 shows that $|corr(x, y)| = |\cos(\angle(\widehat{x}, \widehat{y}))|$ falls in the range of $[\theta, 1]$ or $[-1, -\theta]$ if and only if $\angle(\widehat{x}, \widehat{y})$ falls in the range of $[0, arccos(\theta)]$ or $[180^\circ - arccos(\theta), 180^\circ]$. □

The correlation of two time series is directly related to the angle between the two time series in the multi-dimensional unit sphere. Finding pairs of time series with an absolute value of correlation above the user given minimal correlation threshold θ is equivalent to finding pairs of time series $\widehat{x}$ and $\widehat{y}$ on the unit multi-dimensional sphere with an angle in the range of $[0, \theta_a]$ or $[180^\circ - \theta_a, 180^\circ]$.

2.3 Cone and Correlation between a Pair of Cones

This subsection formally defines the concept of cone and proves two lemmas to bound the correlations of pairs of time series from two cones. The user specified minimal correlation threshold is denoted by θ and $\arccos(theta)$ is denoted by θ_a accordingly.

Definition 1 (Cone) *A cone is a set of time series in the multi-dimensional unit sphere, and it 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 τ of the cone is the maximal angle between any time series in the cone and the cone center.*

We now investigate the relationship of two time series from two cones in the multi-dimensional unit sphere as illustrated in Figure 3 (a). The largest angle($\angle P_1OQ_1$) between two cones C_1 and C_2 is denoted as γ_{max} and the smallest angle ($\angle P_2OQ_2$) is denoted as γ_{min}. We prove the following lemmas to show that if γ_{max} and γ_{min} are in specific ranges, the absolute value of correlation of any pair of time series from the two cones are all above θ (or below θ). Thus all pairs of time series between the two cones satisfy (or dissatisfy) the minimal correlation threshold.

Lemma 2 (All-True Lemma) *Let C_1 and C_2 be two cones from the multi-dimensional unit sphere structure. Let $\widehat{x}$ and $\widehat{y}$ be any two time series from the two cones respectively. If $0 \leq \gamma_{max} \leq \theta_a$, then $0 \leq \angle(\widehat{x}, \widehat{y}) \leq \theta_a$. If $180^\circ - \theta_a \leq \gamma_{min} \leq 180^\circ$, then $180^\circ - \theta_a \leq \angle(\widehat{x}, \widehat{y}) \leq 180^\circ$. If either of the above two conditions is satisfied, $\{C_1, C_2\}$ is called an All-True cone pair.*

Proof: For the first case, it is easy to see from Figure 3 that if $\gamma_{max} \leq \theta_a$, then the angle between $\widehat{x}$ and $\widehat{y}$ is less or equal to θ_a. For the second case, when $180^\circ - \theta_a \leq \gamma_{min} \leq 180^\circ$, we need to show that $180^\circ - \theta_a \leq \angle(\widehat{x}, \widehat{y}) \leq 180^\circ$. If this were not true, there exist $\widehat{x}' \in C_1$ and $\widehat{y}' \in C_2$ where $0 \leq \angle(\widehat{x}', \widehat{y}') < 180^\circ - \theta_a$ since the angle between any pairs of time series is chosen from 0 to 180°. From this inequality, we would have either $\gamma_{min} \leq \phi = \angle(\widehat{x}', \widehat{y}') < 180^\circ - \theta_a$ as shown in Figure 9 (b) or $360^\circ - \gamma_{max} \leq \phi = \angle(\widehat{x}', \widehat{y}') < 180^\circ - \theta_a$ as shown in Figure 9 (c). The first condition contradicts our assumption that $180^\circ - \theta_a \leq \gamma_{min} \leq 180^\circ$. The second condition implies that $360^\circ - \gamma_{max} < \gamma_{min}$ since $180^\circ - \theta_a \leq \gamma_{min}$. This contradicts our choice of γ_{min} as the minimal angle of the two cones. □

Lemma 2 shows that when two cones are close enough, any pair of time series from the two cones are highly positively correlated; and when two cones are far apart enough, any pair of time series from the two cones are highly negatively correlated.

Lemma 3 (All-False Lemma) *Let C_1 and C_2 be two cones from the multi-dimensional unit sphere; let $\widehat{x}$ and $\widehat{y}$ be any two time series from the two cones respectively. If $\theta_a \leq \gamma_{min} \leq 180^\circ$ and $\gamma_{min} \leq \gamma_{max} \leq 180^\circ - \theta_a$, then $\theta_a \leq \angle(\widehat{x}, \widehat{y}) \leq 180^\circ - \theta_a$ and $\{C_1, C_2\}$ is called an All-False cone pair.*

Proof: The proof is straightforward from the inequalities. □

Lemma 3 shows that if two cones are in a moderate range, any pair of time series from the two cones is weakly correlated.

3 Cone-based Filter-and-Refine Algorithm

Our algorithm consists of four steps as shown in Algorithm 1: Pre-processing (step 1), Cone Formation (step 2), Filtering i.e. Cone-level Join (step 4), and Refinement i.e. Instance-level Join (steps 7-10).

Algorithm 1 Correlation Finder

Input: 1) $S^1 = \{s_1^1, s_2^1, \ldots, s_n^1\}$: n_1 spatial referenced time series where each instance references a spatial framework SF_1;
2) $S^2 = \{s_1^2, s_2^2, \ldots, s_n^2\}$: n_2 spatial referenced time series where each instance references a spatial framework SF_2;
3) a user defined correlation threshold θ;

Output: all pairs of time series each from S^1 and S^2 with correlations above θ;

Method:

```
Pre-processing(S^1); Pre-processing(S^2);                                   (1)
CN_1 = Cone_Formation(S^1,SF_1); CN_2 = Cone_Formation(S^2,SF_2);            (2)
for all pair c_1 and c_2 each from CN_1 and CN_2 do {                       (3)
   Filter_Flag = Cone-level_Join(c_1, c_2, θ);                              (4)
      if (Filter_Flag == ALL_TRUE) output all pairs in the two cones        (5)
      else if (Filter_Flag != ALL_FALSE) {                                  (6)
         for all pair s_1 and s_2 each from c_1 and c_2 do {                (7)
            High_Corr_Flag = Instance-level_Join(s_1,s_2, θ);               (8)
               if (High_Corr_Flag) output s_1 and s_2;                      (9)
         }                                                                 (10)
}                                                                          (11)
```

The first step is to pre-process the raw data to the multi-dimensional unit sphere representation. The second step, cone formation, involves grouping similar time series into cones in spatial time series datasets. Clustering the time series is an intuitive approach. However, clustering on time-series datasets itself may be expensive and sensitive to the clustering method and its objective function. For example, K-means approaches [9] find globular clusters while density-based clustering approaches [9] find arbitrary shaped clusters with user-given density thresholds. Spatial indexes, e.g. R^* trees, built after time series dimensionality reduction [1, 4] could be another approach to group similar time series together. In this paper, we explore spatial auto-correlation for the cone formation. First the space is divided into disjoint cells. The cells can come from domain experts, such as the El Nino region, or could be as simple as uniform grids. For uniform grids, we will discuss the cell size by using a correlogram in section 5.1. By scanning the dataset once, we map each time-series into its corresponding cell. Each cell contains similar time series and represents a cone in the multi-dimensional unit sphere representation. The center and span are calculated to characterize each cone.

Example 1 (Spatial Cone Formation) Figure 4 shows an illustrative example of the spatial cone formation for two datasets, namely land and ocean. Both land and ocean frameworks consist of 16 locations. The time series of length m in a location s is denoted as $F(s) = F_1(s), F_2(s), \ldots, F_i(s), \ldots F_m(s)$. Figure 4 only depicts a time series for $m = 2$. Each arrow in a location s of ocean or land represents the vector $< F_1(s), F_2(s) >$ normalized to the two dimensional unit sphere. Since the dimension of the time series is two, the multi-dimensional unit sphere reduces to a unit circle, as shown in Figure 4 (b). By grouping time series in each

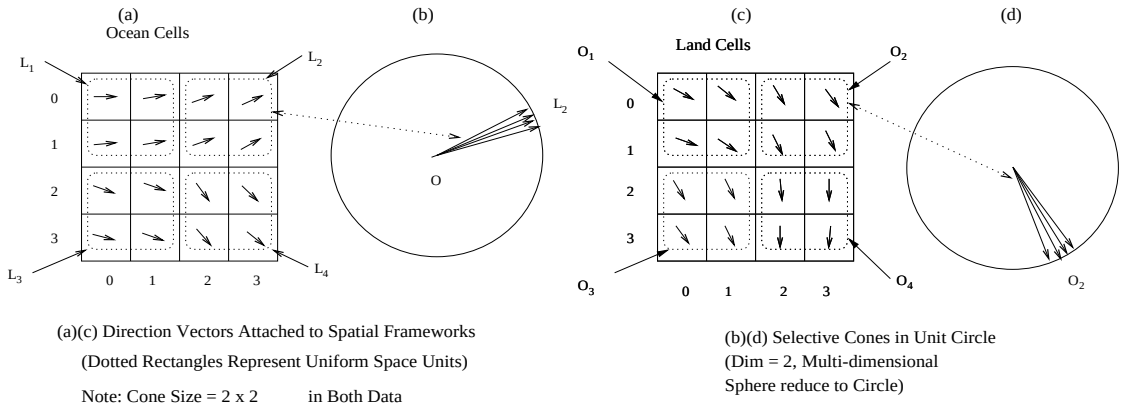


Figure 4: An Illustrative Example for Spatial Cone Formation

datasets into 4 disjoint cells according to their spatial proximity, we have 4 cells each for ocean and land. The ocean is partitioned to $L_1 - L_4$ and the land is partitioned to $O_0 - O_4$, as shown in Figure 4 (a). Each cell represents a cone in the multi-dimensional unit sphere. For example, the patch L_2 in Figure 4 (a) matches L_2 in the circle in Figure 4 (b).

After the cone formation, cone-based join is applied between the two datasets. The calculation of the angle between each pair of cone centers is carried out, and the minimum and maximum bounds of the angles between the two cones are derived based on the spans of the two cones. The All-False cone pairs or All-True cone pairs are filtered out based on the lemmas. Finally, the candidates which cannot be filtered are explored in the refinement step.

Example 2 (Filter-and-refine) The join operations between the cones in Figure 4 (a) are applied as shown in Table 1. The number of correlation computations is used in this paper as the basic unit to measure computation costs. Many All-False cone pairs and All-True cone pairs are detected in the filtering step and the number of candidates explored in the refinement step are reduced substantially. The cost of the filtering phase is 16. Only pairs (O_1, L_1), (O_3, L_4), and (O_4, L_4) cannot be filtered and need to be explored in the refinement step. The cost of the refinement step is 3×16 since there are 4 time series in both the ocean and land cone for all 3 pairs. The total cost of filter-and-refine adds up to 64. The number of correlation calculations using the simple nested loop is 256, which is greater than the number of correlation calculations in the filter-and-refine approach. Thus when the cost of the cone formation phase is less than 192 units, the filter-and-refine approach is more efficient.

Completeness and Correctness Based on the lemmas in Section 2, All-True cone pairs and All-False cone pairs are filtered out so that a superset of results is obtained after the filtering step. There are no false dismissal for this filter-and-refine algorithm. All pairs found by the algorithm satisfy the given minimal correlation threshold.

Ocean-Land	Filtering	Refinement	Ocean-Land	Filtering	Refinement
$O_1 - L_1$	No	16	$O_3 - L_1$	All-True	
$O_1 - L_2$	All-False		$O_3 - L_2$	All-True	
$O_1 - L_3$	All-False		$O_3 - L_3$	All-True	
$O_1 - L_4$	All-False		$O_3 - L_4$	No	16
$O_2 - L_1$	All-False		$O_4 - L_1$	All-True	
$O_2 - L_2$	All-False		$O_4 - L_2$	All-True	
$O_2 - L_3$	All-False		$O_4 - L_3$	All-True	
$O_2 - L_4$	All-False		$O_4 - L_4$	No	16

Table 1: Cone-based Join in Example Data

4 Analytical Evaluation and Cost Models

In this section, we provide simple algebraic cost models for the computation cost of correlation analysis in spatial time series datasets Suppose the two input datasets are D_1 and D_2, and the corresponding cone sets after the coning step are C_1 and C_2 respectively.

The cost model for the proposed algorithm can be divided into three parts: the cost of cone formation, the cost of cone-based correlation joins, and the cost of correlation calculations in the refinement step. The cost of the cone formation, M_1, consists of the cost of calculating cone center and cone angle for each cone and is determined by the number of time series in both datasets. Thus, $M_1 = |D_1| + |D_2|$, where $|D_1|$ and $|D_2|$ are the numbers of time series in D_1 and D_2 respectively. The second part, M_2, is the cost of the correlation join between two cone sets, and the cost is the number of correlation computations of their cross products. Given the number of cones in the two cone sets, M_2 is fixed as the product of the sizes of the two cone sets. Thus we get $M_2 = |C_1| \times |C_2|$. The third part, M_3, depends on the Filtering Ability Ratio, denoted as FAR, of the cone level join. The FAR is the fraction of time series pairs reduced in the filtering step, i.e. $FAR = \frac{N_{time\ series\ pairs-filtered}}{|D_1| \times |D_2|}$. The number of correlation computation after filtering is $|D_1| \times |D_2| \times (1 - FAR)$. The total cost model is denoted as follows:

$$\begin{aligned} Cost =& M_1 + M_2 + M_3 = |D_1| + |D_2| + |C_1| \times |C_2| + |D_1| \times |D_2| \times (1 - FAR) \\ =& |D_1| + |D_2| + \frac{|D_1|}{Cone - Size_1} + \frac{|D_2|}{Cone - Size_2} + |D_1| \times |D_2| \times (1 - FAR) \end{aligned} \quad (1)$$

From the equation, we see that the cost model is related to the sizes of the cones of the datasets and to the FAR. The FAR is determined by the cone sizes and minimal correlation threshold θ. Thus the cost model is sensitive to the cone sizes and the minimal correlation threshold. If we fix the cone sizes and increase the minimal correlation threshold, the FAR increases. The minimal correlation threshold does not affect the cost of M_1 and M_2. So increasing the minimal correlation threshold will decrease the overall cost. If we fix the minimal correlation threshold and increase the cone sizes, M_1 remains the same and M_2 monotonically decreases with increasing cone sizes. However, FAR stops increasing and starts to decrease after the cone sizes reach some value, which leads M_3 to stop decreasing and start to increase. So increasing the cone sizes does not necessarily decrease overall costs. The choice of the cone sizes depends on the datasets and more discussion about the selection of cone sizes is available

in Section 5.1.

5 Performance Evaluation

We want to answer two questions: (1)How does the spatial auto-correlation based inexpensive grouping algorithm affect filtering efficiency? In particular, how do we identify the proper cone size to achieve better overall savings? (2) How does the minimal correlation threshold influence the filtering efficiency? These questions can be answered in two ways: algebraically as discussed in section 5.1 and experimentally as discussed in section 5.2.

Figure 5 describes the experimental setup to evaluate the impact of parameters on the performance of the algorithm. We evaluated the performance of the algorithm with a dataset from NASA Earth science data [12]. In this experiment, a correlation analysis between the East Pacific Ocean region (80W - 180W, 15N - 15S) and the United States was investigated. The time series from 2901 land cells of the United States and 11556 ocean cells of the East Pacific Ocean were obtained under a 0.5 degree by 0.5 degree resolution.

Net Primary Production (NPP) was the attribute of the land cells, and Sea Surface Temperature (SST) was the attribute for the ocean cells. NPP is the net photo-synthetic accumulation of carbon by plants. Keeping track of NPP is important because NPP 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. The records of NPP and SST are monthly data from 1982 to 1993.

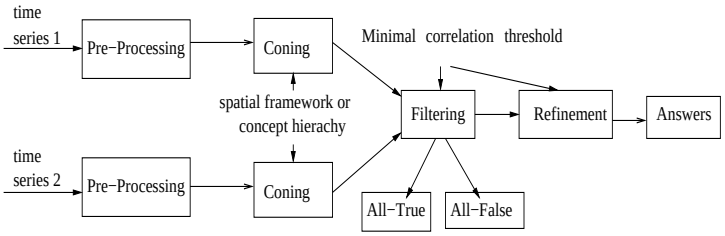


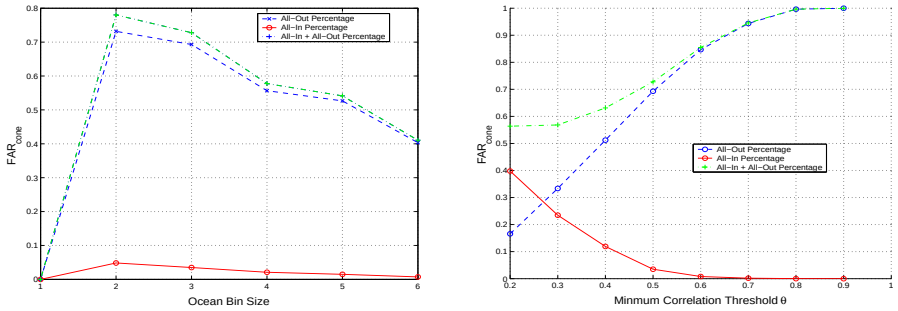
Figure 5: Experiment Design

5.1 Parameter Selections

In this section we investigate the selective range of the cone spans to improve filtering efficiency. Both All-False and All-True filtering can be applied in the filtering step. Thus we investigate the appropriate range of the cone spans in each of these filtering categories.

Given a minimal correlation threshold θ $(0 < \theta < 1)$, $\gamma_{max} = \delta + \tau_1 + \tau_2$ and $\gamma_{min} = \delta - \tau_1 - \tau_2$, where δ is the angle between the centers of two cones, and the τ_1 and τ_2 are the spans of the two cones respectively. For simplicity, suppose $\tau_1 \simeq \tau_2 = \tau$.

Lemma 4 *Given a minimal correlation threshold θ, if a pair of cones both with span τ is an All-True cone pair, then $\tau < \frac{\arccos(\theta)}{2}$.*



(a) Different Ocean Cone Size (Land Cone Size 1×1 and $\theta = 0.5$)

(b) Different θ (Land Cone Size 1×1, Ocean Cone Size 3×3)

Figure 6: All-True and All-False Filtering Percentages for Different Parameters

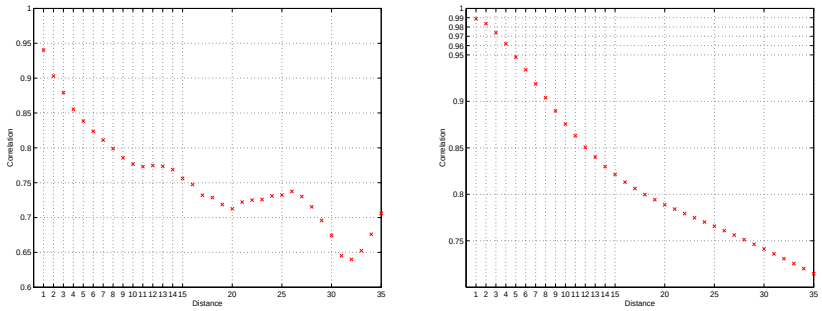
Proof: Assume that a cone pair satisfies the All-True Lemma, i.e., either $\gamma_{max} < \arccos(\theta)$ or $\gamma_{min} > 180^\circ - \arccos(\theta)$ is satisfied. In the former scenario, the angle δ is very small, and we get $\delta + 2\tau < \arccos(\theta)$, i.e., $\tau < \frac{\arccos(\theta)-\delta}{2}$. In the latter scenario, the angle δ is very large, and we get $\delta - 2\tau > 180^\circ - \arccos(\theta)$, i.e., $\tau < \frac{\arccos(\theta)+\delta-180^\circ}{2}$. The τ is less than $\frac{\arccos(\theta)}{2}$ in either scenario since $\tau < 180^\circ$. □

Lemma 5 *Given a minimal correlation threshold θ, if a pair of cones both with span τ is an All-False cone pair, then $\tau \geq \frac{180^\circ}{4} - \frac{\arccos(\theta)}{2}$.*

Proof: Assume that a cone pair satisfies the All-False Lemma, i.e., the conditions $\gamma_{min} > \arccos(\theta)$ and $\gamma_{max} < 180^\circ - \arccos(\theta)$ hold. Based on the two in-equations above, $\gamma_{max} - \gamma_{min} < 180^\circ - 2\arccos(\theta)$ and $\gamma_{max} - \gamma_{min} = 4\tau < 180^\circ - 2\arccos(\theta)$ are true. Thus when the All-False lemma is satisfied, $\tau < \frac{180^\circ}{4} - \frac{\arccos(\theta)}{2}$. □

The range of τ is related to the minimal correlation thresholds. In this application domain, the pairs with absolute correlations over 0.3 are interesting to the domain experts. As shown in Figure 6, All-False filtering provides stronger filtering than All-True filtering for almost all values of cone sizes and correlation thresholds. Thus we choose the cone span τ for maximizing All-False filtering conditions. The value of $\arccos(\theta)$ is less than 72.5° for $\theta \in (0.3, 1]$, so the cone span τ should not be greater than $\frac{180^\circ}{4} - \frac{\arccos(\theta)}{2} = 8.75^\circ$.

An empirical correlogram [5] is often used to demonstrate the spatial autocorrelation of spatial data in spatial statistics. As shown in Figure 7, the correlograms of samples from ocean and land are presented, and the relationships between pairwise distances and correlations among samples are illustrated. The x-axis represents the distances of the ocean-land pairs in the unit of degree, and the y-axis represents the correlations of the time series of the ocean-land pairs. According to this figure, the ocean demonstrates higher spatial autocorrelation than the land. This is because the maximum cone angle should be less than 8.75°, and the cone size should keep the correlations between any time series in the cone and the cone center less than 0.988. According to this cutting line, land cannot satisfy this criterion, and the distance in the ocean correlogram is between 1 and 2. Thus the cone size of land is chosen as 1×1, and the cone size of ocean is chosen as 3×3.



(a) Correlogram for Land (b) Correlogram for Ocean

Figure 7: Empirical Correlograms for Land and Ocean Samples

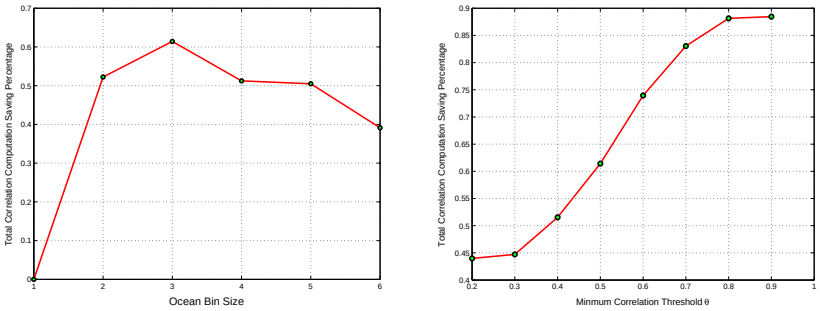
5.2 Experimental Results

Experiment 1: Effect of Coning The purpose of the first experiment was to evaluate under what coning sizes the savings from filtering outweighs the overhead. When the cone is small, the time series in the cone are relatively homogeneous, resulting in a small cone span τ. Although it may result in more All-False and All-True pairs of cones, such cone formation incurs more filtering overhead because the number of cones is substantially increased and the number of filtered instances in each All-False or All-True pair is small. When the cone is large, the value of the cone span τ is large, resulting in a decrease in the number of All-False and All-True pairs. The effects of the All-False and All-True filtering in the given data are investigated.

Experiment 2: Effect of Minimal Correlation Thresholds In this experiment, we evaluated the performance of the filtering algorithm when the minimal correlation threshold is changed. Various minimal correlation thresholds were tested and the trends of filtering efficiency was identified with the change of minimal correlation thresholds.

5.2.1 Effect of Coning

This section describes a group of experiments carried out to show the net savings of the algorithm for different cone sizes. For simplicity, we only changed the cone size for one dataset. According to the analysis of the previous section, the land cone size is fixed at 1×1. We carried out a series of experiments using the fixed minimal correlation threshold, the fixed land cone size, and various ocean cone sizes. The minimal correlation threshold θ was fixed as 0.5. Figure 8 (a) shows the net savings as a percentage of the computational cost of the nested loop join algorithm for different ocean cone sizes. The x-axis represents the different cone sizes ranging from 1×1 to 6×6, and the y-axis represents the net savings in computational cost as a percentage of the costs using the simple nested loop join algorithm. The net savings range from 40 percent to 62 percent, which is consistent with the empirical correlogram of the ocean samples.



(a) Testing Different Ocean Cone Size

(b) Testing θ

Figure 8: Testing Different Cone Sizes and Minimal Correlation Threshold θ

5.2.2 Effect of Minimal Correlation Thresholds

In this experiment, we investigated the effects of minimal correlation threshold θ on the savings in computation cost for correlation analysis. The land and ocean cone sizes were fixed as 1×1 and 3×3 respectively, and a series of experiments was carried out for different θs. Figure 8 (b) shows the total savings as a percentage of the computational cost of the nested loop join algorithm for different θs. The x-axis represents the different cone sizes ranging from 1×1 to 6×6, and the y-axis represents the total savings as a percentage of the computational cost of the nested loop join algorithm. The net savings percentages range from 44 percent to 88 percent with the higher savings at higher values of correlation thresholds. Thus when other parameters are fixed, the filtering algorithm generally achieves better performance as the minimal correlation threshold is increased.

6 Conclusion and Future Work

In this paper, a filter-and-refine correlation analysis algorithm for a pair of spatial time series datasets is proposed. A cost model and experimental evaluations using a NASA Earth science dataset are presented. The total savings of correlation analysis computation range from 40 percent to 88 percent.

In future work, we would like to explore other coning methods, which are listed in Figure 9. Cluster and spatial methods using other schemes may provide higher filtering capabilities but possibly with higher overheads. Time series dimensionality reduction and indexing methods [1, 4, 6] (e.g., F-index [1]) will also be explored to determine the tradeoff between filtering efficiency and overhead.

Acknowledgment

We are particularly grateful to Spatial Database Group members for their helpful comments and valuable discussions. We would also like to express our thanks to Kim Koffolt for improving the readability of this paper.

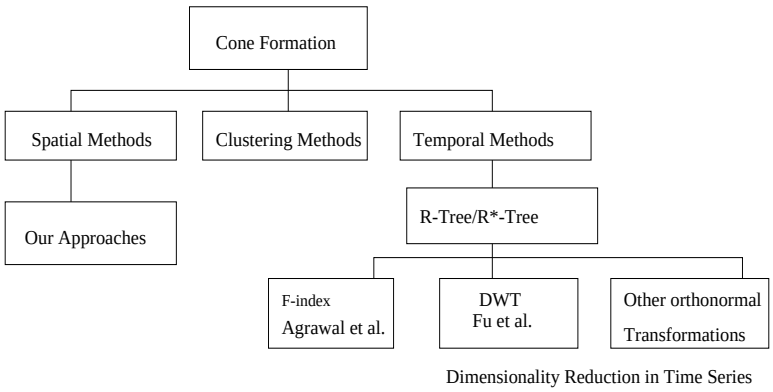


Figure 9: Cone Formation Candidates in Spatial Time Series Data

This work is partially supported by NASA grant No. NCC 2 1231, and it is also sponsored in part 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 the policy of the government, and no official endorsement should be inferred.

eferences

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

[3] B.W. Lindgren. *Statistical Theory (Fourth Edition).* Chapman-Hall, 1998.

[4] K. Chan and A. W. Fu. Efficient Time Series Matching by Wavelets. In *Proc. of the 15th ICDE*, 1999.

[5] N. Cressie. *Statistics for Spatial Data.* John Wiley and Sons, 1991.

[6] Christos Faloutsos. *Searching Multimedia Databases By Content.* Kluwer Academic Publishers, 1996.

[7] R. Grossman, C. Kamath, P. Kegelmeyer, V. Kumar, and R. Namburu, editors. *Data Mining for Scientific and Engineering Applications.* Kluwer Academic Publishers, 2001.

[8] D. Gunopulos and G. Das. Time Series Similarity Measures and Time Series Indexing. *SIGMOD Record*, 30(2):624–624, 2001.

[9] J. Han and M. Kamber. *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers, 2000.

[10] E. Keogh and M. Pazzani. An Indexing Scheme for Fast Similarity Search in Large Time Series Databases. In *Proc. of 11th Int'l Conference on Scientific and Statistical Database Management*, 1999.

[11] Y. Moon, K. Whang, and W. Han. A Subsequence Matching Method in Time-Series Databases Based on Generalized Windows. In *Proc. of ACM SIGMOD*, Madison, WI, 2002.

[12] C. Potter, S. 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.

[13] J. Roddick and K. Hornsby. Temporal, Spatial, and Spatio-Temporal Data Mining. In *First Int'l Workshop on Temporal, Spatial and Spatio-Temporal Data Mining*, 2000.

[14] J. Roddick and Brian G. Lees. Paradigms for spatial and spatio-temporal data mining. In *Geographic Data Mining and Knowledge Discovery.* H. Miller and J. Han (Eds), Taylor & Francis, 2001.

[15] S. Shekhar and S. Chawla. *Spatial Databases: A Tour.* Prentice Hall, 2002.

[16] 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.

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

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

[19] G. H. Taylor. Impacts of the El Nio/Southern Oscillation on the Pacific Northwest. http://www.ocs.orst.edu/reports/enso_pnw.html.

[20] J. Wijsen and R.T. Ng. Temporal Dependencies Generalized for Spatial and Other Dimensions. In *Int'l Workshop on Spatio-Temporal Database Management*, 1999.

[21] Michael F. Worboys. *GIS - A Computing Perspective.* Taylor and Francis, 1995.