Co-anomaly Event Detection in Multiple Temperature Series

Xue Bai[1], Yun Xiong[1], Yangyong Zhu[1], Qi Liu[2], and Zhiyuan Chen[3]

[1] Fudan University
[2] University of Science and Technology of China
[3] University of Maryland Baltimore County
{xuebai,yunx,yyzhu}@fudan.edu.cn, feiniaol@mail.ustc.edu.cn,
zhchen@umbc.edu

Abstract. Co-anomaly event is one of the most significant climate phenomena characterized by the co-occurrent similar abnormal patterns appearing in different temperature series. Indeed, these co-anomaly events play an important role in understanding the abnormal behaviors and natural disasters in climate research. However, to the best of our knowledge the problem of automatically detecting co-anomaly events in climate is still under-addressed due to the unique characteristics of temperature series data. To that end, in this paper we propose a novel framework $Sevent$ for automatic detection of co-anomaly climate events in multiple temperature series. Specifically, we propose to first map the original temperature series to symbolic representations. Then, we detect the co-anomaly patterns by statistical tests and finally generate the co-anomaly events that span different sub-dimensions and subsequences of multiple temperature series. We evaluate our detection framework on a real-world data set which contains rich temperature series collected by 97 weather stations over 11 years in Hunan province, China. The experimental results clearly demonstrate the effectiveness of $Sevent$.

Keywords: Event Mining, Co-anomaly Event, Time Series.

1 Introduction

Since climate events reveal seasonal or interannual variations in climate change from periodic weather behaviors, mining them from climate data recorded in temperature series has attracted much attention in the literature [1,2]. Among climate events, the co-anomaly event, which represents the co-occurrence of similar abnormal behaviors in different temperature series, is one of the most important events in climate research for understanding climate variability and analyzing the process of abnormal events.

For better understanding co-anomaly event, Fig. 1 illustrates the subsequences of six daily temperature series from Dec.13 to Dec.24, 1998 in Hunan, China, where Fig. 1 (a) presents three stations with normal temperature behaviors and Fig. 1 (b) presents three temperature series that have unusual higher values than expected. Though the six weather stations in Fig. 1 are similar to each other in magnitudes, we can see the ones in Fig. 1 (b) are suffering from a co-anomaly event represented by unusual warm in the middle of winter. As the abnormal temperature behaviors in one co-anomaly event (e.g., that in Fig 1 (b)) are much likely to be caused by the similar climatic factors, mining

M. Wang (Ed.): KSEM 2013, LNAI 8041, pp. 1–14, 2013.

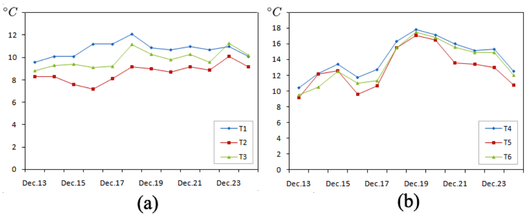

Fig. 1. Subsequences of six temperature series ($^{\circ}C$) from different weather stations. (a) T1, T2, and T3 are normal in winter. (b) T4, T5, and T6 are in a co-anomaly event.

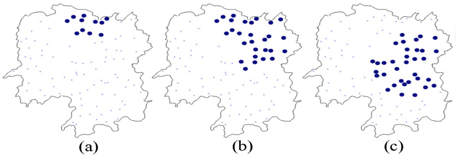

Fig. 2. A cold wave event moving from north to south of Hunan in adjacent three days

and identifying such co-anomaly events can provide a detailed exploration on these climate phenomena. For instance, it helps experts quickly identify whether a co-occurrent unusual phenomenon occurred by chance or not and the value of further analysis. Thus, capturing this co-occurrence of similar abnormal behaviors (co-anomaly event) is of growing interests in many real-world applications [3].

However, there are many technical and domain challenges inherent in detecting co-anomaly climate events in temperature series. First, temperature series in climate are relatively smooth curves, e.g., much smoother than stock price time series and vehicle sensor time series. In other words, the values of temperature series usually do not deviate far from the average. Thus, some co-anomaly events taking place at a limited number of cities are not that obvious with respect to the average temperature series or each single series, and this raises difficulties for traditional methods. Secondly, different from traditional anomaly events, the similar abnormal behaviors should co-occur in a number of series (i.e., a sub-dimension of the entire series set) before we can claim this is a co-anomaly event. However, we can not simply use the frequency as an evaluation to find interesting patterns because the frequent patterns in climate usually represent well-known normal phenomena. Thirdly, co-anomaly climate events often evolve with time, thus the associated groups of temperature series are changing too, i.e., they usually correlate with different sub-dimensional subsequences of multiple temperature series. For instance, Fig. 2 shows a cold wave moved from north to south during three days in spring, where we can see that the affected cities on the first day (blue points in Fig. 2 (a)) were quite different from the ones on the third day (blue points in Fig. 2 (c)). Since identifying co-anomaly events from temperature series data is not technically straightforward, the researchers and experts usually have to search and analyze these events

Table 1. Mathematical notations

Notation	Description
$D = T_1, T_2, ...T_m$	The set of temperature series data
$T = t_1, t_2, ..., t_n$	A temperature series
$S = t_p, ..., t_{p+k-1}$	A subsequence of a temperature series T
$\overline{S} = \overline{s}_1, ..., \overline{s}_w$	A Piecewise Aggregate Approximation of a subsequence S
$\hat{S} = \hat{s}_1, ..., \hat{s}_w$	A symbol representation (word) of a subsequence S
$E = \hat{e}_1, ..., \hat{e}_u$	A co-anomaly event E
$B = \beta_1, ..., \beta_{\Psi-1}$	Breakpoints
w	The number of PAA elements
Ψ	Alphabet size. The total number of different symbols
ϕ	The number of common temperature series between words

manually. However, the volume and complexity of the data preclude the use of manual visualization to identify these co-occurrent patterns.

To address the above challenges, in this paper we propose a novel co-anomaly event detection framework called $Sevent$ which includes three phases: first, map multiple temperature series to symbolic representations based on data distributions; Second, apply statistical tests to extract interesting co-anomaly patterns; Third, the co-anomaly patterns are connected into co-anomaly events by their correlations. Thus, the co-anomaly events are finally generated and ranked. Our main contributions can be summarized as:

To the best of our knowledge, we are the first to solve the co-anomaly event mining problem in multiple temperature series. Meanwhile, we propose a symbolic representation framework that can differentiate group behaviors. Though we describe the work in a domain-depended way, worth noting that similar idea is generally applicable to mine co-anomaly events from other types of series data. We carry out extensive experiments on real-world data set [4] of temperature series collected from 97 weather stations over 11 years in Hunan province, China. The results show that the proposed $Sevent$ can successfully detect co-anomaly underlying events interested in meteorology.

2 Problem Statement and Data Description

In this paper, we focus on dealing with the problem of detecting and ranking significant co-anomaly climate events from a given set of temperature series, and meanwhile, identifying the corresponding cities(or sub-dimensions) and time-spans (or subsequence) affected. As have said our solutions can be generally applied to pattern mining problems for multiple time series, including but not limited to detecting climate co-anomaly events from temperature series.

We exploit a real-world temperature dataset [4] collected from 97 weather stations over a period of 11 years in Hunan province, China. Thus, each station stands for one temperature series, and each temperature series records the daily average temperatures of the corresponding weather station. In all, there are 365×11 data points for each temperature series to represent the temperature behaviors over time. The timestamps at February 29th are directly removed from the data set for simplicity. Here, we choose the temperature data because temperature is a well accepted important climate variable and many of the well known climate indices are based upon it. At last, worth noting that the spatial distances of stations, although important, are not taken into consideration for two reasons: First, the weather stations are not far from each other in our data set (all locating in one province); Second, in this work, we focus on the problem of detecting

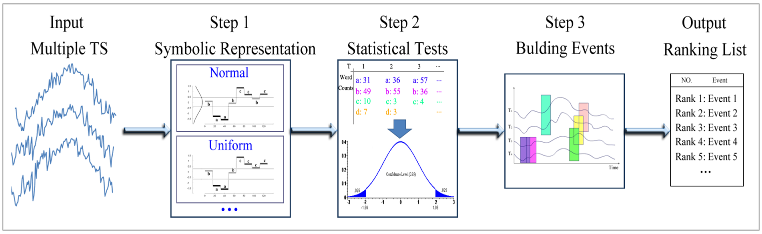

Fig. 3. The flowchart of Sevent

co-anomaly events from multiple temperature series and we would like to leave the detection of co-anomaly events from Geo-referenced time series as a future work.

3 Mining Co-anomaly Events

In this section, we present our framework $Sevent$ in detail. To facilitate understanding, the related important mathematical notations are illustrated in Table 1. To be specific, we define a ***temperature series*** $T = t_1, t_2, ..., t_n$, which records the temperature values over time, as an ordered set of n real-valued variables, where data points $t_1, t_2, ..., t_n$ are temporally ordered and spaced at equal time intervals (e.g., one day). Second, we define a ***temperature series data set*** D as a set of m temperature series. Moreover, a ***subsequence*** S of a temperature series $T = t_1, t_2, ..., t_n$ is a sampling of length $k \leq n$ of contiguous position from T, i.e., $S = t_p, ..., t_{p+k-1}$ for $1 \leq p \leq n - k + 1$.

Generally, to detect co-anomaly events from multiple temperature series, we need to identify sub-group behaviors among these temperature series. However, a simple clustering-like method (e.g., based on Euclidean distance) would not be appropriate for this task: First, it is very time consuming to search for all possible subsequences in all sub-dimensions; Second, we focus on group abnormal behaviors mining rather than the abnormal behaviors of one temperature series, i.e., the subsequences in an co-anomaly event do not need to be abnormal if we only look into each single time series. Thus, for detecting co-anomaly events from multiple temperature series effectively and efficiently, we propose $Sevent$, a novel framework with three major steps. First, we represent temperature series by symbolic characters. In this way, behaviors of each temperature series can be easily represented by combinations of characters. Then, we apply statistical tests to identify co-anomaly patterns. Finally, correlated patterns are connected into co-anomaly events from different time-spans. The overall flowchart is illustrated in Fig. 3, and each step of $Sevent$ is introduced in the following subsections.

3.1 Symbolic Representation

Symbolic representation is a popular way for time series representation with the benefits of reducing the volume of data points and preserving the evolving trends of time series simultaneously. A general framework for symbolic representation usually includes three steps: First, apply Piecewise Aggregate Approximation (PAA) [5] to reduce the dimensions of temperature series; Second, determine a list of breakpoints, which are usually drawn from a pre-defined distribution (e.g., Uniform). Third, transform PAA results to symbolic characters by comparing their positions with breakpoints.

In the first step, the PAA representation of a temperature series T can be denoted by a vector $\overline{T} = \overline{t}_1, ..., \overline{t}_w$ (or $\overline{S} = \overline{s}_1, ..., \overline{s}_w$ for S). Specifically, the i^{th} element of $\overline{T}$ is calculated by the following equation,

$$\overline{t}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} t_j.$$

In the second step, the breakpoints can be drawn from the distribution of the specific data that we need to analysis. Here an alphabet size Ψ is required to be predefined, which is leveraged for determining breakpoints of symbolic representation. Since most of the time series can be approximately fitted by normal distribution or uniform distribution, in the following we take these two distributions as an example to illustrate the way to generate breakpoints. Specifically, the breakpoints for normal distribution can be defined as follows, which is the same with that used in Symbolic Aggregate Approximation (SAX) [6].

Definition 1 (Breakpoints of N(0,1) Distribution). *Breakpoints are a sorted list of numbers $B = \beta_1, ..., \beta_{\Psi-1}$ such that the area under a $N(0,1)$ normal curve from β_i to β_{i+1} is equal to $\frac{1}{\Psi}$ (β_0 and β_Ψ are defined as $-\infty$ and ∞, respectively).*

Similarly, if the data is drawn from uniform distribution, and the corresponding breakpoints can be defined as follows.

Definition 2 (Breakpoints of Uniform Distribution). *Breakpoints are a sorted list of numbers $B = \beta_1, ..., \beta_{\Psi-1}$ such that $\beta_{i+1} - \beta_i = \frac{\beta_\Psi - \beta_0}{\Psi}$ (β_0 and β_Ψ are defined as minimum and maximum value of the temperature series, respectively).*

Noting that the breakpoints for other data distributions can be defined in the same way. After we obtain a list of breakpoints (B), a subsequence can be mapped into symbolic representation which is defined as a **word** [6].

Definition 3 (Word). *A subsequence S of length k can be represented as a word $\hat{S} = \hat{s_1}, ..., \hat{s_w}$. Let α_i denote the i^{th} element of the alphabet, e.g., $\alpha_1 =$ **a** and $\alpha_2 =$ **b**. Then the mapping from a PAA approximation $\overline{S}$ to a word $\hat{S}$ is obtained as follows,*

$$\hat{s_i} = \alpha_i, \;\; iff. \;\; \beta_{j-1} < \overline{s}_i \leq \beta_j. \tag{1}$$

For example, the data points whose value locates between the first two breakpoints ($[\beta_0, \beta_1)$) are mapped to “a”, and the ones within $[\beta_1, \beta_2)$ are mapped to “ b”.

3.2 Detecting Co-anomaly Patterns

After transforming temperature series to words, we can calculate the number of each word at every timestamp. Words of different expressions represent different behaviors, e.g., the word *abcd* stands for a behavior of a rising temperature. Then by counting the number of the behaviors(words) at the same timestamp, we can find the frequent patterns(words) which are representations of group behaviors. However, frequency does not guarantee that pattern is interesting, and statistical tests are widely used to evaluate the importance of patterns. Specifically, co-anomaly patterns can be defined as follows.

Definition 4 (Co-anomaly pattern). *A word $\hat{S} = \hat{s}_1, ..., \hat{s}_w$ is a co-anomaly pattern if its count is statistically significant.*

The "co-anomaly pattern" is different from "anomaly pattern". The behavior of a co-anomaly pattern may not be abnormal if we only look into one single temperature series. It is abnormal and statistically significant in history only when we consider a group of consistent behaviors as a whole. For instance, in every year, there are always several cities experience extremely cold temperatures in winter. However, if dozens of the cities all have the same severe cold temperatures in one year's winter, it can be a co-anomaly event that may be caused by the same cold wave. For finding these co-anomaly patterns, a null hypothesis is defined and statistical hypothesis tests are used to calculate the P-value of each observed word.

Definition 5. *For a given word $\hat{S}$ and a timestamp t we define hypotheses H_0 and H_1:*
H_0: $\hat{S}$ is uninteresting at t.
H_1: $\hat{S}$ has a frequency that is significantly greater than the expected count at t.

Here, the expected count of each word at timestamp t are learned from the historic data, and it is used as the baseline of each concurrent behavior. The probability of $\hat{S}$ is,

$$\mu^t(\hat{S}) = \frac{N(\hat{S})^t}{nN_y}, \tag{2}$$

where $N(\hat{S})^t$ is the count of $\hat{S}$ at timestamp t for all years in history, N_y is the year number, and n is the number of temperature series. The expected count of $\hat{S}$ is,

$$\hat{N}(\hat{S})^t = n\mu^t(\hat{S}). \tag{3}$$

Then, for a word frequency x, we use the normal approximation to calculate its P-value, i.e., $N(\hat{S})^t$ follows the normal distribution $N(\hat{S})^t \sim \mathcal{N}(n\mu,\, n\mu(1-\mu))$.

$$\mathbb{P}(\mathcal{N}(\mu,\, \sigma^2) \geq N^{obs}(\hat{S})^t) = 1 - \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right], \tag{4}$$

where $\text{erf}(x)$ is the Normal Error Function and the formula is as follows,

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt. \tag{5}$$

The P-value is then compared to a predefined critical value α. If $P < \alpha$, the null hypothesis H_0 is rejected and the word is accepted as a co-anomaly pattern. Noting that there are some other statistical methods for computing P-values, e.g., the Binomial trails, and Poisson approximation. Any of them can be used to test whether a word is statistically significant, while a detailed analysis of the pros and cons of these methods is beyond the scope of this paper.

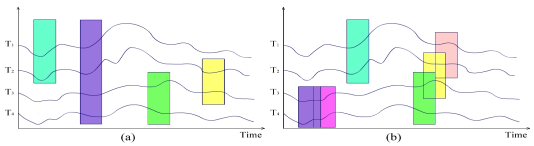

Fig. 4. Different ways for correlated significant words to form co-anomaly events

3.3 Building Co-anomaly Events

The co-anomaly patterns, which are adjacent in timestamps and correlated with the similar group of temperature series, are much likely to be involved in one co-anomaly event. Thus, the time-span of a co-anomaly event is not limited by the length of sliding windows. Here, a threshold ϕ is pre-defined, and two adjacent co-anomaly patterns will be connected if they have more than ϕ temperature series in common. Therefore, the co-anomaly events are able to have different durations(time-spans). Finally, we propose a ranking function $Pscore$ to evaluate a co-anomaly event. Generally speaking, one co-anomaly event with a higher $Pscore$ are likely to be formed by patterns with lower P-values and more affected temperature series. The co-anomaly event is defined as,

Definition 6 (Co-anomaly event). $E = \hat{e}_1, ..., \hat{e}_u$ *is a co-anomaly event if* $\forall \hat{e}_i \in E$ *is a co-anomaly pattern, and* $|\hat{e}_i \cap \hat{e}_{i+1}| > \phi$.

Under this definition, co-anomaly events are clusters of correlated significant words with various time-length, and those significant words that are connected by timestamps and temperature series should be put into one event. However, the temperature series associated with each word at different timestamps are not necessarily to be the same. For instance, Fig. 4 shows several possible ways for words to form events. In Fig. 4 (a), four separate words form four independent events. Fig. 4 (b) shows three events formed by different number of associated words. Specifically, the left one represents a event covering T_3 and T_4 for two timestamps. The right one displays an event moving from T_4 and T_3 to T_3 and T_2, and finally to T_2 and T_1, which could be a cold front or a typhoon flowing from west to east. Noting that threshold ϕ is defined to be the minimum of temperature series in common for adjacent words, i.e., if two words adjacent by timestamps and have more than ϕ common temperature series, then they can also be connected as a candidate event. Thus there may be multiple words at t_{k+1} qualified to be connected. It is natural in real word phenomenon for some events to change with time because they may have different kinds of evolutions. In this connection process, we are able to deal with this scenario by capturing every kind of the evolution record.

Due to the differences in word expressions and time-span, it is difficult to establish comprehensive evaluations for events. Generally, the anomaly of events are associated with the rareness of each behavior and the range of its influence: The lower the probability, the rarer the behavior, and the bigger the coverage, the more serious the behavior. As the P-value is the probability of each observed word count ranging from 0 to 1, we propose to compute $-\log$ of the P-value such that a rarer word can have a bigger positive value. In this way, we design a ranking function to evaluate each event according to the P-value of each word and the number of affected temperature series.

Definition 7 (Ranking function:$Pscore(E)$). *The overall Ranking value $Pscore(E)$ of a co-anomaly event E with regard to a set of relevant words $RW(E)$, the P-value of each word $Pvalue(\hat{S})$, and the observed count of each word $Count(\hat{S})$ is defined as,*

$$Pscore(E) = \sum_{\hat{S} \in RW(E)} -Count(\hat{S}) \cdot \log Pvalue(\hat{S}) \tag{6}$$

In summary, the connecting phase includes three steps: First, connect the words adjacent in timestamps if they share over ϕ common temperature series; Second, repeat the first step until no more new connections are formed; At last, rank events by $Pscore$.

The pseudocode of the proposed co-anomaly climate event detection framework $Sevent$ is shown in Algorithm 1. Specifically, the procedures in line 1, line 2 to line 5, and line 6 to line 15 can be mapped into Symbolic Representation(Step 1), Statistical Tests(Step 2) and Building Events(Step 3) of the flowchart in Fig. 3, respectively. The runtime complexity for symbolic representation is $O(m \cdot n)$,where m is the number of temperature series, and n is the length of one temperature series. Let Ψ denote the alphabet size, and w be the number of PAA elements . Then, the runtime complexity for calculating P-values and connecting words is $O(w \cdot \Psi)$+$O(w \cdot \Psi)$. Thus, the total runtime complexity for $Sevent$ is $O(m \cdot n)$+ $O(w \cdot \Psi)$.

Algorithm 1. Sevent(D, Ψ, w, α, ϕ)

```
Input: $D$: the m-dimensional temperature series dataset; $\Psi$: the alphabet size; $w$: PAA length;
    $\alpha$:significance level; $\phi$: the minimum number of common temperature series.
Output: The event set $E$ .
 1: Mapping $D$ into symbolic words using PAA and breakpoints;
 2: for each timestamp $T_i \in D$ do
 3:     Calculate the P-value of each words;
 4:     Delete the words with the P-value bigger than $\alpha$;
 5: end for
 6: for each timestamp $T_i \in D$ do
 7:     for each word $\hat{S}_j \in T_i$ do
 8:         if $\hat{S}_j$ can not connect with adjacent words then
 9:             $E = E \cup \hat{S}_j$;
10:         else
11:             Connect $\hat{S}_j$ with associated words;
12:         end if
13:     end for
14: end for
15: Sort $E$ by $Pscore$;
16: return $E$;
```

4 Experiments

In this section, we evaluate the proposed $Sevent$ on the real-world data set from Meteorology(Section 2). Specifically, we demonstrate: (1)The results of the co-anomaly events detection;(2)Two case studies of the detected co-anomaly events;(3) In-depth analysis on the generation of breakpoints(i.e., normal distribution and uniform distribution). We implemented our approaches in java, and all experiments were run on a personal computer with 2.0GB RAM and 2.26GHz CPU. In the following we fix the parameter settings as: $\Psi = 8$; $w = 122$ (122 timestamps in a year, i.e., 3 daily temperatures are combined to one timestamp) ; $\alpha = 0.01$; $\phi = 10$.

4.1 Co-anomaly Event Detection

One of our primary goals is to automatically detect the time-spans of co-anomaly events in multiple temperature series. The ground truth comes from The Climate Reports of Hunan Province [7] with the date, the magnitude of influences and other descriptions of some co-anomaly events.

Fig. 5 shows the time-length of co-anomaly events mining results of year 2001-2002. The reason of choosing year 2001 to 2002 for test is that there are plenty of detailed descriptions of the beginning and ending time in 2001-2002, while in other years, the ground truth of time-spans are not that clearly recorded in the report [7]. We demonstrate the time-spans of detected co-anomaly events based on the $Sevent$ with uniform distribution for generating breakpoints (Fig. 5 (b)) and that with normal distribution(Fig. 5 (c)), and the baseline [1] (Fig. 5 (d)), respectively. Meanwhile, Fig. 5 (a) shows the average temperature, and (e) shows the real time-spans of events recorded in Climate Reports (Ground truth). Specifically, there are eight recorded co-abnormally events from 2001 to 2002, and the detailed descriptions are listed in Table 2. From Fig. 5 we can see that the detected time-spans of normal distribution are the most similar ones to the ground truth. The $Sevent$ with uniform distribution detected most of the time-spans, however, it just found the beginning of event b (the warm spring in March), but the whole time durations. In contrast, the results of baseline are composed by lots of small fragments of timestamps, which has the least overlaps with the ground truth. Correspondingly, Table 3 lists the detailed information of top ten co-anomaly events detected by $Sevent$ along with the uniform distribution, and table 4 lists the results along with the normal distribution. From the results we can find that the two top ten rankings have approximately 9 events in common (Please also refer to Fig. 5), and most of the detected events can be found in annual report (i.e., Table 2). For example, the first event in Table 4 (the $7th$ event in Table 3, and **f** in Table 2 and Fig. 5) resulted in disasters for crops production, and the economic losses were 1.23 billion (RMB).

Then, we compare the recall of results on the whole data set. For simplicity, we only select the following events as the ground truth for the comparison: (1) January 12 to February 8, 2008. The most severe snow storm disaster since 1949, with the direct economic losses of over 680 billion RMB. (2) the abnormal spring in 1998, (3) the warm winter in early 1999, (4) late spring coldness and hailstorm in 1999, (5-8) the 4 events in 2002 (Fig. 5 (e) $f - i$), (9) the late spring coldness in 2006,(10) the warm winter in 2007. One reason for choosing these events is that they are all important climate events, and most of them cause significant economic losses. The other reason

[1] Which is calculated by comparing the difference between the average temperature of multiple temperature series of each year and the total average temperature series, and then the timestamps that have a gap greater than 3 are chosen as candidates.

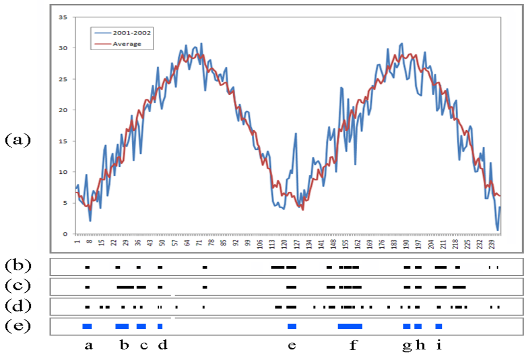

Fig. 5. Co-anomaly events mining results in 2001-2002. (a) The average temperatures. (b) The time-spans results when applying uniform distribution, (c) The time-spans results when applying normal distribution. (d) The baseline. (e) The time-spans recorded in the climate reports.

is that there are precise date records for each event and thus easy for us to compare. To test the effectiveness of $Sevent$, we only consider top 50 detected events ranked by $Pscore$, i.e., if these important events are not in top 50, the recall value will be low. The final results are illustrated in Table 5, where we can see that $Sevent$ based on both normal distribution and uniform distribution performs much better than the baseline. Specifically, we find that the snow storm in 2008 is detected as the most rare event(the same as the evaluation in the climate reports) by uniform distribution with the total $Pscore$ 6002.16. However it is not reported in the top 50 of normal distribution. That is because in normal distribution, the partitions in the lowest(or the highest) range are much coarser than that of in the middle range. Thus, although the snow storm in 2008 is very severe, it is not significant under the normal distribution. In contrast, all of the important events are successfully detected by the method with uniform distribution.

Table 2. The co-anomaly events in 2001-2002

Label	Duration	Description
a	Jan. and Feb.	Warm winter
b	Early Mar.	Warm spring
c	Late Apr.	Early summer
d	Late May.	Early heat waves
e	Jan.	Warm winter.
f	Apr. 1 - May. 10	Hailstone, coldness
g	Jul. 18 - Jul. 27	Low temperature
h	Aug. 6 - Aug. 15	Low temperature
i	Sep. 14 - Sep. 16	Cold dew wind

Table 3. Top ten co-anomaly events (uniform)

NO.	Durations	Score
1	Aug. 8 - Aug. 17, 2002	2430.80
2	Jan.1 - Jan. 12, 2002	2361.28
3	Oct. 16 - Oct. 22, 2002	2194.96
4	Apr. 8 - Apr. 17, 2002	1804.53
5	Mar.27 - Apr. 3, 2002	1775.56
6	Apr. 19 - Apr. 22 , 2001	1531.90
7	Apr. 23 - May. 4, 2002	1504.58
8	Dec. 6 - Dec. 18, 2001	1466.07
9	May. 2 - May. 11, 2002	1321.16
10	Dec. 24 - Dec. 27, 2002	1230.57

Table 4. Top ten co-anomaly events (normal)

NO.	Durations	Score
1	Apr. 23- May. 7, 2002	3453.85
2	Oct. 13 - Oct. 30, 2002	3244.40
3	Aug. 8 - Aug. 17, 2002	3228.31
4	Jan.1 - Jan. 12, 2002	3089.23
5	Sept. 13 - Oct. 7, 2002	2972.07
6	Apr. 8 - Apr. 17, 2002	2575.32
7	Mar.27 - Apr. 3, 2002	1708.21
8	Mar.8 - Mar. 23, 2002	1533.34
9	Apr.19 - Apr. 28, 2001	1378.92
10	Nov. 3 - Nov. 12, 2001	1232.93

Table 5. The Recall result

Alg.	Normal	Uniform	Baseline
Recall	0.9	1.0	0.6

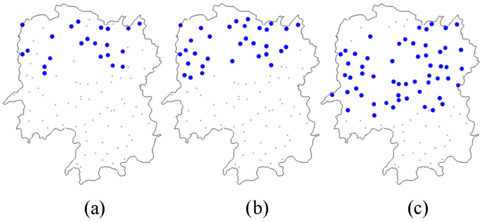

Fig. 6. A cold-wave event started from north and then expanded to the middle of Hunan province in late June, 1999. Only sub-dimensions of temperature series are involved in this events.

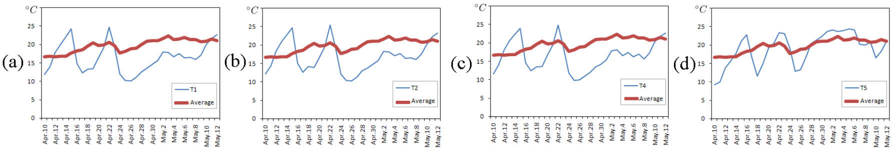

Fig. 7. Detail subsequences of four temperatures series. (a) - (c) are temperature series of the same event. (d) does not have continuous low temperature patterns.

4.2 Case Studies

To further explore the extracted co-anomaly events, we present two case studies to show how $Sevent$ can capture the evolutions of co-anomaly events, and trace the involved sub-dimensions as well. First, we present the evolution of one cold-wave event captured in parts of the stations(i.e., the sub-dimensions of the temperature series in June, 1999). Fig. 6 (a)-(c) shows the evolution of this event in adjacent timestamps. Here we can see that $Sevent$ successfully detected and traced the evolution of sub-dimensional events. In Fig. 6 (a), only 22 stations suffer from this cold-wave. Then the event expands to 32 stations (Fig. 6 (b)). Finally the event spreads to the middle of the province and 56 stations are affected (Fig. 6 (c)). From this result we can easily trace a cold wave movement, which is from north to the middle of Hunan province, and then blocked by the mountains in the middle and the south. Thus the evolutions of co-anomaly events can well support the further research of cold-wave abnormal activities for domain experts.

Then, we show a more detailed example of involved sub-dimensions in Fig. 7, where the original temperature series are from 10th April to 12th May in 2002. We find the corresponding event description from the climate report: "late spring coldness and low temperature in May". As shown in Fig. 7, station (a) (b) (c) all have similar behaviors during this period, while station (d) does not have "low temperature in May". This co-anomaly event does not span full dimensions or full durations (station (d) only joined the first half of the duration). From these two case studies, we can conclude that various co-anomaly events of sub-dimension and sub-durations can be detected by our $Sevent$.

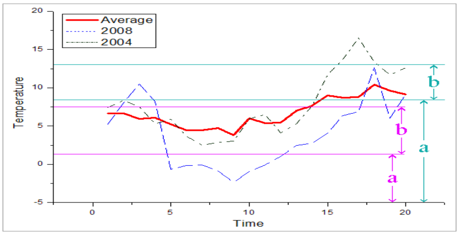

Fig. 8. The breakpoints under uniform and normal distribution

4.3 Normal Distribution *vs* Uniform Distribution

Here we give a detailed analysis on the generation of breakpoints, i.e., based on normal distribution or uniform distribution, and it provides more insights for future application. Along this line, we show a case study in Fig. 8. We choose the year 2008, when a severe snow disaster stroke many cities in Hunan province, and another year 2004 with correspondingly normal temperature as examples (the time is from Jan 1 to Feb 28 of each year). Fig. 8 shows the breakpoints under different distributions, uniform and normal, respectively. The purple lines are breakpoints under uniform distribution, where the gaps between lines are of the same. The blue lines are breakpoints under normal distribution, where the gaps in the middle (medium temperatures, e.g., the temperatures of springs and autumns) of the whole temperature series are much smaller than the up and low parts (high and low temperatures, e.g., the temperatures of summers and winters). Thus when mapping the subsequences of the winter temperatures in Fig. 8, the breakpoints under normal distribution has a much coarser differentiations than that of uniform distribution, where most of the temperature points in winters, no matter severe coldness or not, are mapped to the symbol “a”. In this way, the severe cold disaster is not that significant when we adopt normal distribution. In contrast, the uniform distribution maps most of the normal temperature points to the symbol “b”, and when it comes to severe cold event in 2008, a significant number of “a” co-occurrent together and lasting a long period of time forms a severe co-anomaly event. Based on the discussion, we can capture the pros and cons of each distribution, and they can guide us design more effective co-anomaly event detection framework.

5 Related Work

In the past several decades, there has been numerous work on finding abnormal patterns, change-points, and events in time series data. Due to the space limit, we just present a brief survey on major research directions most relevant to ours.

Event Detection in Time Series have been proposed in [2,8,9,10,11]. For instance, Guralnik and Srivastava [8] proposed an iterative algorithm that used a likelihood criterion for time series segmentation. Preston et al. [10] proposed a method to search for subintervals that are statistically significantly different from the underlying noise. Ihler et al. [11]proposed a time varying Poisson process to model periodic count data that can detect bursty events embedded in time series. Cho et al. [9] proposed a framework based on episode rules for event prediction over event streams. Minaei et al. [2] explored the correlation between time series streams and events stream, where event streams are logs from domain experts. Anomaly detection in time series is also similar to the problem of searching for events in time series. E.g., Keogh et al. [12] presented an anomaly detection method that searched for subsequences that differ most from the rest of its subsequences in one time series.

Finding Patterns in Multiple Time Series are explored in [1,13,14,15]. McGovern et al. [1] introduced a multi-dimensional motif mining approach to predict severe weather. They used labeled time series data to build the trie structure [12] to find subsequences that are relevant to severe weather, and then grow motifs into longer patterns. Minnen et al. [13] formulated multivariate motifs as regions with high estimated density via k-nearest neighbor search. An expected linear-time algorithm [14] was proposed to detect subdimensional motifs in multivariate time series. Tanaka et al. [15] used principal component analysis to project the multi-dimensional time series into one dimension signal. However, the frequent patterns are not necessarily the most interesting ones. To find significant patterns, many work on significant motif mining [10,11,16] have been proposed, while most of them do not detect significant subdimensional motifs. Xiong et al. [3] detected peculiar groups in day-by-day behavior datasets that are similar to the co-anomaly events problem in our work. However, they assume that most objects are dissimilar with each other, which was difficult to satisfy in real-word datasets.

Although these approaches are related to our work, they are fundamentally different and are not particularly well suited for our application. In summary, our work differs from them in four aspects: 1) We focus on detecting group abnormal behaviors, rather than single abnormal behaviors; 2) We consider periodic calendar time constrains for multiple temperature series modeling; 3) We propose a connection method based on correlation between objects and timestamp adjacency to form events with different time-span and even evolving with time; 4) We propose an abnormal ranking function based on statistical significance to evaluate the abnormal degree of events.

6 Conclusion

In this paper, we provided a focused study of exploiting multiple temperature series data for co-anomaly climate event detection. Specifically, we first map the original temperature series to symbolic representations based on data distributions. Then, we detect the co-anomaly patterns by statistical tests and finally generate the co-anomaly events that span different sub-dimensions and subsequences of multiple temperature series. Meanwhile, this proposed detection framework $Sevent$ also captures the evolutions of the co-anomaly events in multiple temperature series. The experimental results on real-world data of temperature series demonstrate that our $Sevent$ can successfully detect co-anomaly events interested in meteorology. In the future, we plan to apply and evaluate our framework in the co-anomaly event detection from other types of series data.

Acknowledgments. This work is supported in part by the National Science Foundation Project of China (Grant No. 61170096) and the National High Technology Research and Development Program (“863” Program) of China (Grant No. 2011AA100701).

References

1. McGovern, A., Rosendahl, D.H., Brown, R.A., Droegemeier, K.K.: Identifying predictive multi-dimensional time series motifs: an application to severe weather prediction. Data Mining and Knowledge Discovery 22(1), 232–258 (2011)
2. Minaei-Bidgoli, B., Lajevardi, S.B.: Correlation mining between time series stream and event stream. In: NCM 2008 (2008)
3. Xiong, Y., Zhu, Y.: Mining peculiarity groups in day-by-day behavioral datasets. In: ICDM 2009 (2009)
4. Liao, Y., Wang, K., Zhao, F., Bai, S.: Modern agro-climatic zoning of Hunan Province. Hunan University Press, Changsha (2010)
5. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Dimensionality reduction for fast similarity search in large time series databases. Knowledge and Information Systems 3(3), 263–286 (2001)
6. Lin, J., Keogh, E., Patel, P., Lonardi, S.: Finding motifs in time series. In: the 2nd Workshop on Temporal Data Mining (July 2002)
7. The climate reports of hunan province, `http://www.hnqx.gov.cn`
8. Guralnik, V., Srivastava, J.: Event detection from time series data. In: KDD 1999 (1999)
9. Cho, C.-W., Wu, Y.-H., Yen, S.-J., Zheng, Y., Chen, A.: On-line rule matching for event prediction. The VLDB Journal 20, 303–334 (2011)
10. Preston, D., Protopapas, P., Brodley, C.: Event discovery in time series. Arxiv preprint arXiv:0901.3329 (2009)
11. Ihler, A., Hutchins, J., Smyth, P.: Adaptive event detection with time-varying poisson processes. In: KDD 2006 (2006)
12. Keogh, E., Lin, J.: Hot sax: Efficiently finding the most unusual time series subsequence. In: ICDM 2005 (2005)
13. Minnen, D., Essa, I., Isbell, C.: Discovering multivariate motifs using subsequence density estimation. In: AAAI Conf. on Artificial Intelligence (2007)
14. Minnen, D., Essa, I., Isbell, C.L., Starner, T.: Detecting subdimensional motifs: An efficient algorithm for generalized multivariate pattern discovery. In: ICDM 2007 (2007)
15. Tanaka, Y., Iwamoto, K., Uehara, K.: Discovery of time-series motif from multidimensional data based on mdl principle. Machine Learning 58(2-3), 269–300 (2005)
16. Castro, N., Azevedo, P.J.: Time series motifs statistical significance. In: SDM 2011 (2011)