A framework to mine significant association patterns from climate time series

DENG Min[1], SHI Yan[1]

[1]Department of Geo-informatics, Central South University, Changsha 410083, China

Email: dengmin028@yahoo.com; csu_shiy@126.com

LIU Qi-liang[2]

[2]Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, China

Email: liuqiliang192@gmail.com

1. Introduction

In recent years, abnormal climatic events occur frequently, such as flood, drought, ice storm, etc, which have a negative impact on the natural and social environment. Earth scientists had found that the ocean climate had an effect on the land climate, so it is important to discover the mechanism of abnormal climatic events deeply. In order to find non-redundant and significant patterns, we first use a modified scale-space clustering method which considers the property of multi-scale to compress data and reduce the spatial autocorrelation, and we get effective climate zones. Second, we propose an algorithm called APMBTW to mine significant association patterns based on time windows and the related context. Finally, it is found that many patterns obtained by the algorithm, which are coincident with the priori knowledge, as also illustrated that our method is effective and reasonable. In fact, these new association patterns between abnormal sea surface pressure and abnormal land temperature of China, which are discovered by our method, will contribute to the domain of meteorology.

2. Mine significant association patterns from climate time series

To mine significant association patterns from climate time series, three steps are included in the following.

Step 1 Remove the temporal autocorrelation

Earth scientists had found that the long term climate time series had the temporal autocorrelation, which is embodied as the characteristics of period and seasonality. For example, there are four seasons, i.e. spring, summer, autumn and winter. However, we are not interested in these known patterns which disturb the process of mining other hidden knowledge, therefore it's necessary to remove the temporal autocorrelation. The common-used methods of dealing with the seasonality have the monthly Z core, the Discrete Fourier Transform, singular value decomposition and the moving average. The method of monthly Z core is employed in this study.

Step 2 Remove the saptial autocorrelation

The climate datasets record the climatic characteristics of a large number of spatial regions with the form of long term time series, so that there has spatial autocorrelation obviously. The closer several regions are to each other, the same the climatic characteristics of them are. If we make the cells that are not only close to each other, but also have similar climate characteristics into a big cell, the capacity of the data will be reduced by preserving the overall characteristics of the original data and eliminating the spatial autocorrelation. This process is quite important and necessary. In the field of data mining, this task can be completed by using the clustering methods.

Step 3 Mine significant patterns from interesting events

After removing the temporal and spatial autocorrelation, we will gain some new datasets which are more succinct and still efficient without unuseful interference. Time consumption and redundant rules will be largely reduced, because these new datasets are more suitable for mining significant patterns about climate problem. The work of optimizing data according to spatial and temporal properties is over and the next task is focused on dealing with the event property, which is so called extracting interesting events from time series. In the field of climate, we are interested in those unusual events, such as extremely high or low temperature events, the extremely precipitation events, etc. The percentile method is a simple and useful tool for our work, in which the top x% value is considered as the abnormally big and the last x% value is extremely small in the time series. And we only take these extreme values into count, while others are ignored. At this point, we have finished the task of constructing mining objects. Finally, the association rules mining methods from time series will be used for the mining objects with necessary improvement. Some representative methods are proposed, such as WINEPI, MINEPI and MOWCATL, etc. WINEPI and MINEPI define the concept of events in detail and propose the concept of time window that is proved very useful in the process of mining. MOWCATL firstly takes the time lags into account and it is very important for the climate problem. We summarize some great ideas of these methods and propose an algorithm that is more suitable for climate problem. The main idea of this algorithm is to consider the time lag based on the concept of time window. The more the number of same time windows where two events occur in sequentially, the more relevant the two events are.

3. Experiments

For the sea climate data, we use the global monthly sea surface temperature (SST) datasets and the global monthly sea level pressure (SLP) datasets. They are both from the National Oceanic and Atmospheric Administration (NOAA) using both situ and satellite. For the spatial attributes, they are both grid network data consisted of latitude and longitude, where the spatial resolution of the SST data and the SLP data is 1°× 1° and 2.5°× 2.5°, respectively. For the temporal attributes, they are both time series data, that is, each grid is a time series. The time range of the SST data is from January 1982 to December 2010, while the time range of the SLP data is from January 1982 to December 2004.

For the land climate data, we use the global monthly land precipitation datasets and the monthly land temperature datasets of our country. The former is from NOAA and the latter is from China Meteorological Administration. For the spatial attributes, they are both grid network data that consist of latitude and longitude, and their spatial resolution both are 1°× 1°. For the temporal attributes, they are both time series data. The time range of the global monthly land precipitation data is from January 1982 to December 2010, while the time range of the monthly land temperature data of our country is from January 1982 to December 2004.

By using the above experimental data, we attempt to mine the association patterns between SST and the land precipitation, and the association patterns between the SLP and the land temperature. For the former, we will find the EI NINO regions that are highly consistent with the known ones and then mine how the EI NINO regions effect the land precipitation. It has been proven that our method is effective. For the latter, we find some interesting and unknown patterns between the SLP and the land temperature of our country.

Experiment 1: Mine the association patterns between the SST and the land precipitation

Based on our mining framework, first we remove the temporal autocorrelation and then cluster the processed data by the multi-scale clustering method to remove the spatial autocorrelation. And we gain a series of effective climate zones of sea and land. Then we need to find some proper results by statistical methods. Figure 1 shows suitable clustering results for SST and land precipitation. Next, we try to find the EI NINO regions in our best clustering results based on the correlation coefficients with EI NINO indices (showed in figure 2) and use our method to study how the founded EI NINO regions affect the land precipitation. We analyze four kinds of association patterns:

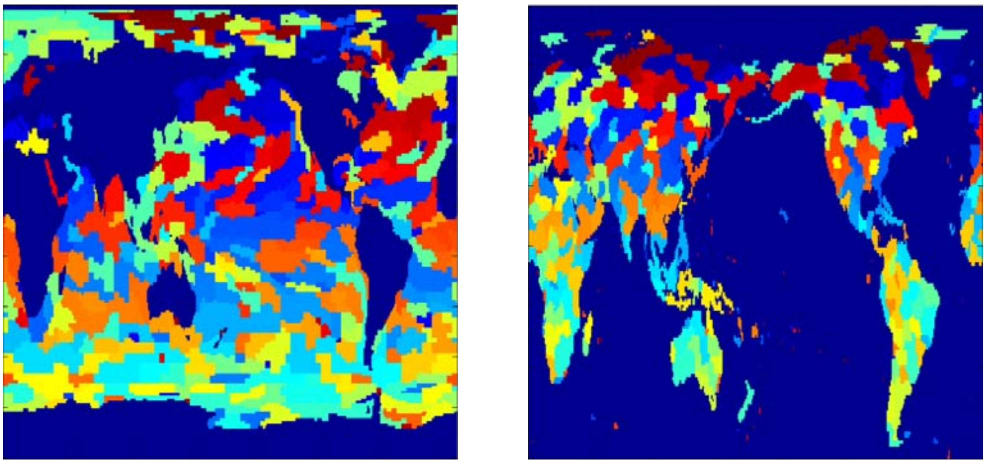
(I) The SST of EI NINO regions is abnormally high=>The land precipitation of some regions is abnormally high

(II) The SST of EI NINO regions is abnormally high=>The land precipitation of some regions is abnormally low

(III) The SST of EI NINO regions is abnormally low=>The land precipitation of some regions is abnormally high

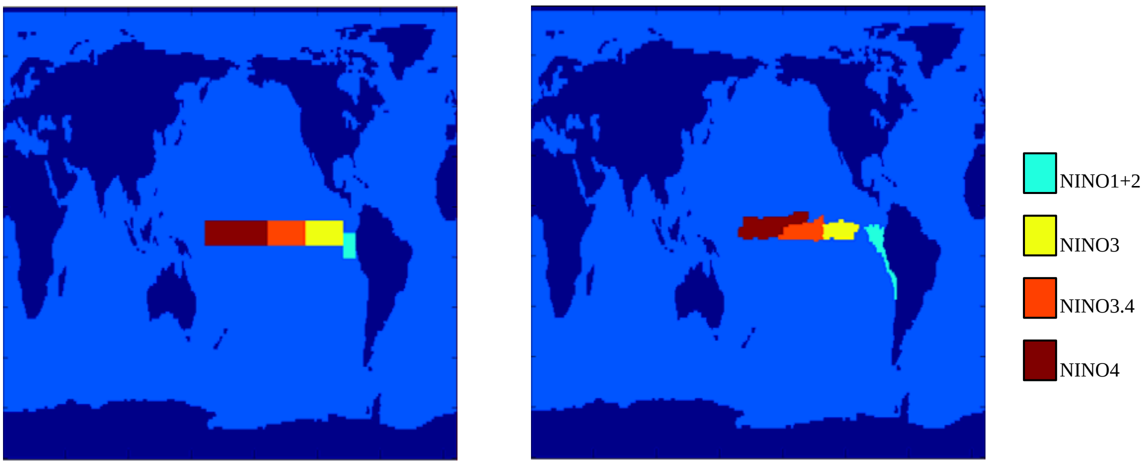
(IV) The SST of EI NINO regions is abnormally low=>The land precipitation of some regions is abnormally low

Finally, we can find some patterns, as shown in figure 3. These patterns are quite consistent with the known knowledge. So it proves that our mining framework is feasible and it can be used to mine some unknown patterns.



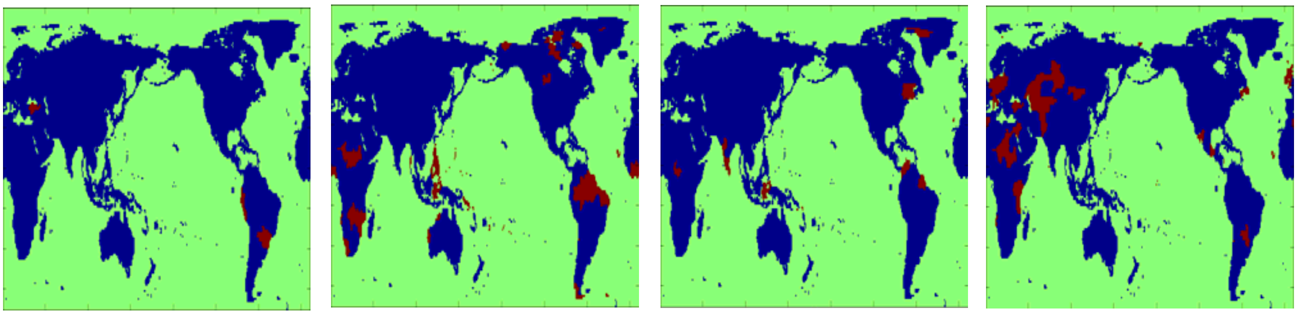
(a) the suitable clustering results for SST (b) the suitable clustering results for land precipitation

Figure 1. The suitable clustering results for SST and land precipitation



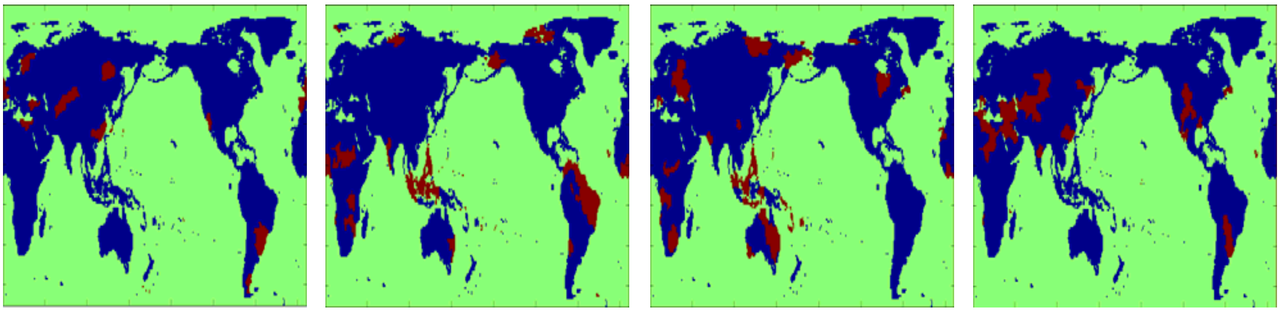
(a) the known EI NINO regions (b) the founded EI NINO regions

Figure 2. The 4 EI NINO regions



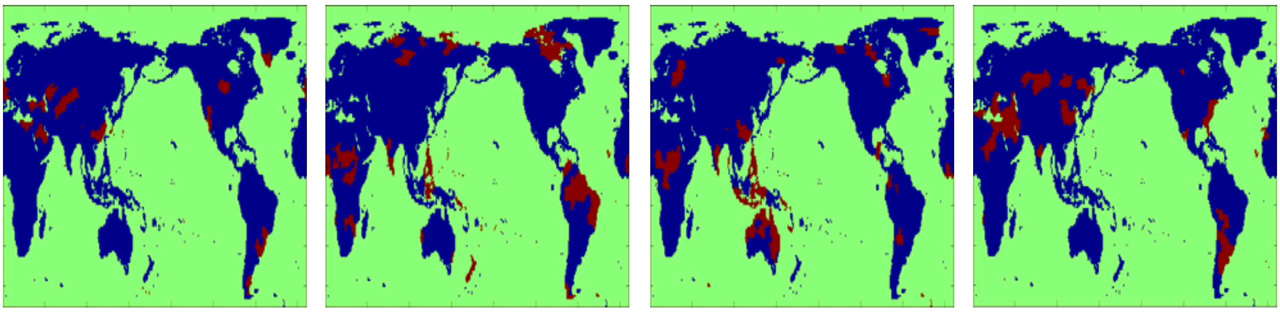
pattern_I, *width*=3, *min_SD*=0.4 pattern_II, *width*=6, *min_SD*=0.5 pattern_ III, *width*=6, *min_SD*=0.4 pattern_IV, *width*=9, *min_SD*=0.45

(a) the abnormal precipitation of land regions that the abnormal SST of EI NINO 1+2 affects



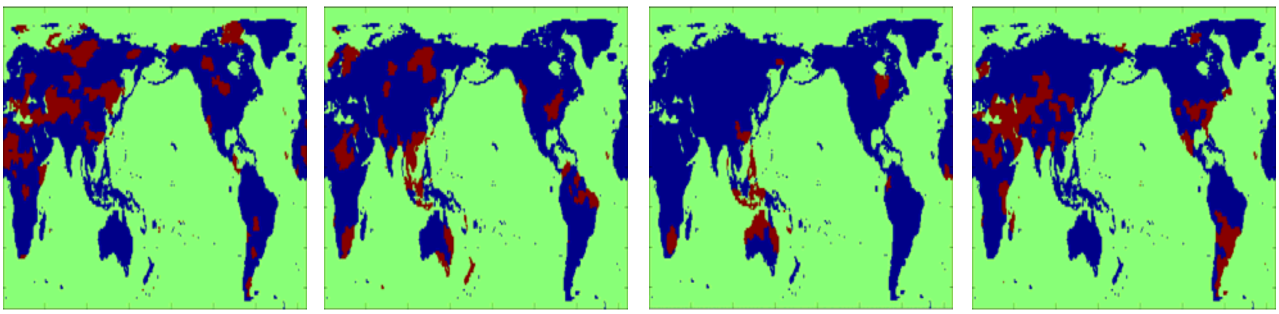
pattern_I, *width*=6, *min_SD*=0.45 pattern_II, *width*=6, *min_SD*=0.45 pattern_III, *width*=6, *min_SD*=0.4 pattern_IV, *width*=6, *min_SD*=0.4

(b) the abnormal precipitation of land regions that the abnormal SST of EI NINO 3 affects



pattern_I, *width*=6, *min_SD*=0.5 pattern_II, *width*=6, *min_SD*=0.5 pattern_III, *width*=6, *min_SD*=0.45 pattern_IV, *width*=6, *min_SD*=0.45

(c) the abnormal precipitation of land regions that the abnormal SST of EI NINO 3.4 affects



pattern_I, *width*=9, *min_SD*=0.45 pattern_II, *width*=9, *min_SD*=0.45 pattern_III, *width*=6, *min_SD*=0.5 pattern_IV, *width*=9, *min_SD*=0.5

(d) the abnormal precipitation of land regions that the abnormal SST of EI NINO 4 affects

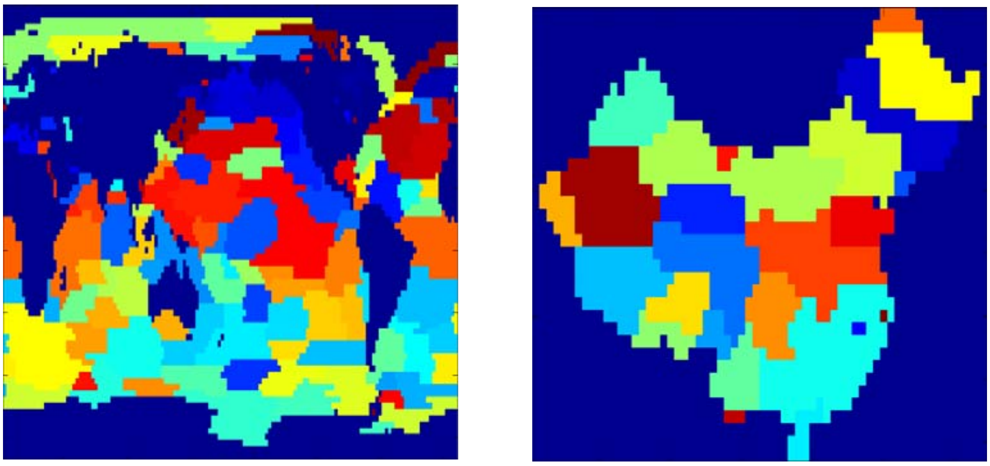
Figure 3. The abnormal precipitation of land regions that the abnormal SST of EI NINO regions affect

Experiment 2: Mine the association patterns between the SLP and the land temperature

Be similar with experiment 1, we first use the clustering methods to gain some climate zones of sea and land, then we select the best results to explore the association patterns between the SLP regions(as shown in figure 4) and the land temperature of China. Likewise, there are 4 kinds of association patterns:

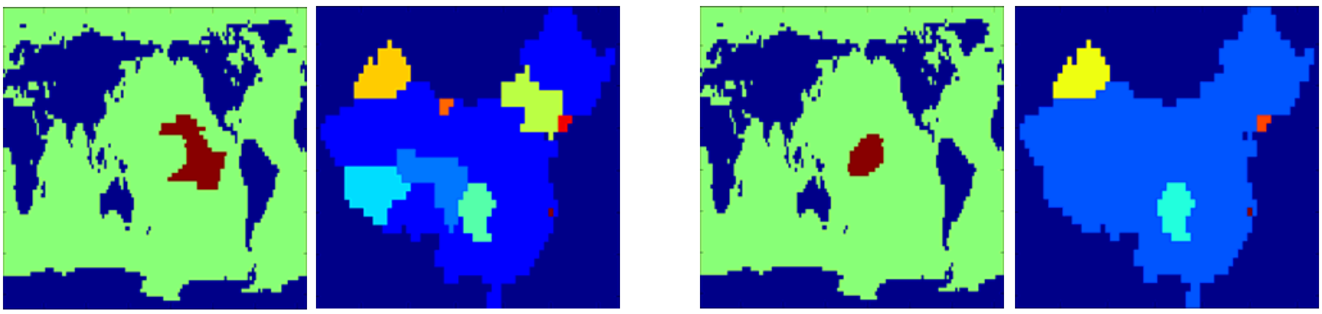
(I) The SLP of some regions is abnormally high=>The land temperature of some regions is abnormally high
(II) The SLP of some regions is abnormally high=>The land temperature of some regions is abnormally low
(III) The SLP of some regions is abnormally low=>The land temperature of some regions is abnormally high
(IV) The SLP of some regions is abnormally low=>The land temperature of some regions is abnormally low

We can find some patterns, as shown in figure 5. They are unknown, so they are more worthy than those known ones. And we can study them further to find the mechanism between the sea and the land climate. Finally, these patterns can be converted to useful knowledge to help the government make decisions.



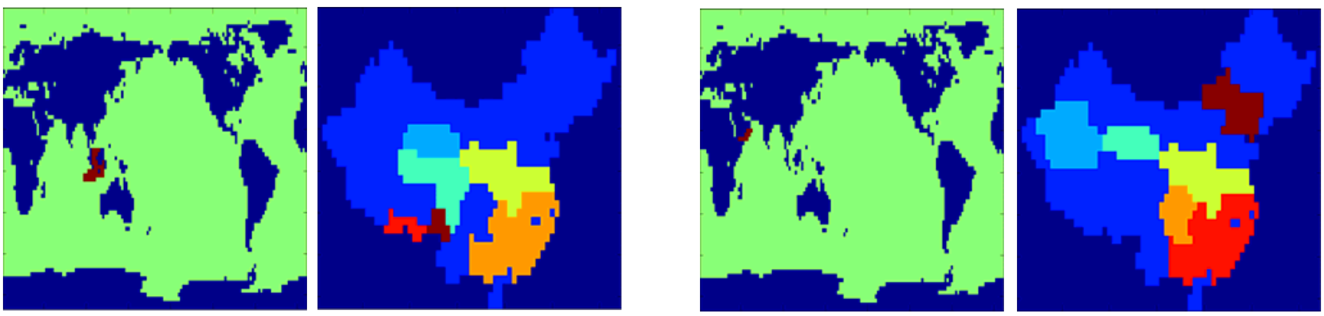
(a) the suitable clustering results for SLP (b) the suitable clustering results for land temperature

Figure 4. The suitable clustering results for SLP and land temperature



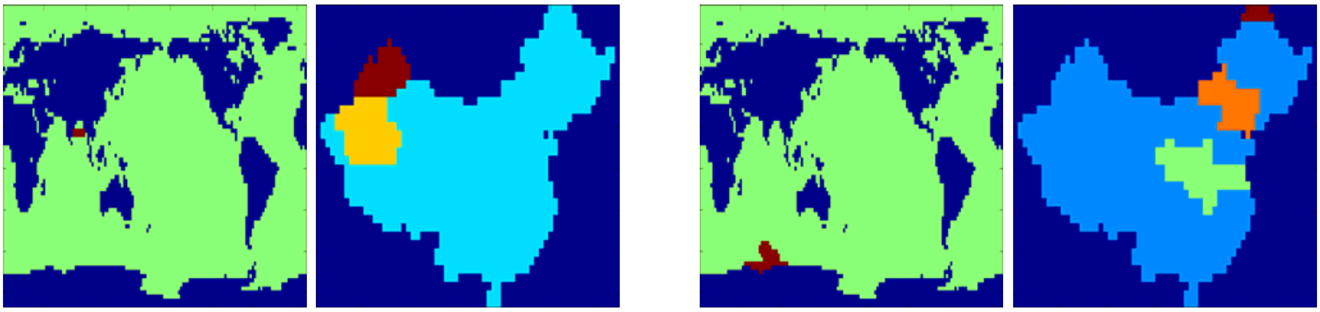
width=9, *min_SD*=0.45 *width*=9, *min_SD*=0.45

(a) Pattern_I



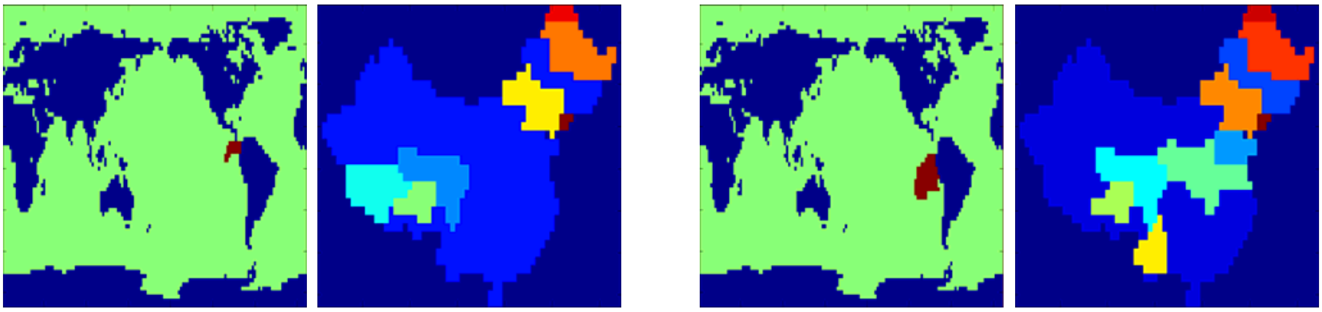
width=9, *min_SD*=0.45 *width*=9, *min_SD*=0.45

(b) Pattern_II



width=9, *min_SD*=0.45 *width*=9, *min_SD*=0.45

(c) Pattern_III



width=9, *min_SD*=0.5 *width*=9, *min_SD*=0.5

(d) pattern (IV)

Figure 5. The mined interesting association patterns between abnormal SLP and land temperature

4. Conclusion

This paper discusses a framework to mine significant association patterns from climate time series, where three important steps are involved. The first step is to remove the temporal autocorrelation of data; the second is to remove the spatial autocorrelation of data to get a series of effective climate zones; and the third is to mine significant association patterns between sea and land climate zones. In the second step, we take the multi-scale effect into account, so that the clustering results are more consistent with the objective law. In the third step, we first extract abnormal events which are interesting from time series, then an association patterns mining algorithm based on time windows is proposed to do the mining work. The experiments have illustrated that our method is able to mine some interesting and significant patterns. But our method also has some limitations. For example, the extracting of interesting events is too simple to find those hidden and more significant events. The selection of thresholds is very subjective, so that the efficiency of mining is reduced. Indeed, our future research work will be focused on these aspects.