Detecting Spatio-temporal Outliers in Climate Dataset: A Method Study

Sun Yuxiang, *Xie Kunqing, Ma Xiujun, Jin Xingxing, Pu Wen, Gao Xiaoping
Department of Intelligent Science, Peking University, Beijing, China, 100871
National Laboratory on Machine Perception, Peking University, Beijing, China, 100871
{pekingsun, kunqing, maxj, jinxx, pwpro, gaoxp}@cis.pku.edu.cn

Abstract—**Outlier detecting is one of the most important data analysis technologies in data mining, which can be used to discover anomalous phenomena in huge dataset. Many literatures on spatial outlier detecting and time series outlier detecting have appeared, while the area of spatio-temporal outliers considering both spatial and temporal dimensions has still rarely been touched. Defining outliers in traditional dataset is more explicit because the data structure we need to focus on is very straightforward (e.g., a spatial point or a transaction record). However, it is much more difficult to give outlier a definite characterization in spatio-temporal lattice data, since there are so many data structures we can pay attention to. With the aim of detecting useful and meaningful outliers in climate dataset, we introduce a formalized way to define outliers in spatio-temporal lattice data, in which the importance of clarifying basic data structure (we call it basic element in our paper) is stressed. As a case study, we define two kinds of spatio-temporal outliers based on a global climate dataset, according to the three aspects we propose in defining an outlier. The introduction of basic element and the formulation of outlier definition process make it easier and clearer to define meaningful outliers. Thus outlier detecting in spatio-temporal lattice data will provide us with really interesting and useful knowledge.**

Keywords-Outlier detecting; Spatio-temporal outlier

I. INTRODUCTION

Outlier detection is one of the most important technologies in data mining, and several successful applications are always referred when talking about outliers, such as credit card fraud detection. Since complex data such as spatial data and temporal data can be acquired more and more easily, data mining techniques including outlier detection technique have to adjust themselves to adapt to the new situation. Shekhar provided a general definition of S-outliers for spatial outliers in [SLZ03], which in fact brings several existing outlier definitions into a common formalization framework. [ZLK03] uses wavelet method to detect region outliers, which has borrowed many things from image processing. However, in [SLZ03] and [ZLK03], only spatial dimension has been considered.

Time series and sequential data is another kind of complex data. The methods including statistics and temporal data mining have been applied and finding deviant points in them is also a relative mature field.

However, as the data collection technology and device have been dramatically improved, the data with both spatial and temporal dimensions are more and more popular. The climate data we are studying is just of this kind, which we will describe in detail later. Others, such as sensor data, data of moving objects and daily or hourly sent data from earth orbiting satellites, are also getting familiar by us. There must be many exceptional events or anomalous phenomena hidden in these spatio-temporal data, but the traditional methods, even the methods that can get fantastic results from spatial or temporal data respectively, are not able to be used directly on spatio-temporal data.

What is more serious, the previous definition of outlier is not sound any more, since all the definitions are all confined to some certain kinds of data. Thus what we can do is coming down to the original definition of outliers, in order to give reasonable definition of spatio-temporal outliers. We use the general definition of outliers motioned in [BKNS00] all through this paper, "An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism".

The rest of the paper is organized as follows. Section 2 introduces some background of the study of spatio-temporal outliers. In section 3, we give our first definition of spaito-temporal outlier in our climate dataset, which aims at finding locations that are distinct from its neighbors. In section 4, we define another kind of outliers with the aim to find anomalous time periods (e.g., a certain year) with distinct behavior compared with others in this dataset. We make a conclusion in section 5.

II. BACKGROUND

A. Related Work

[AJA04] uses distance-based method offered by [KN98] to distinguish outliers, what is treated carefully is the selection of

Supported by the Major Project of National Natural Science Foundation (40235056)
*Corresponding author. E-mail address: kunqing@cis.pku.edu.cn (Xie Kunqing) Tel.: 86-10-62756920

neighborhood and the spatial relation. The whole work substantially belongs to spatial method. And the basic data structure it focuses on is sensor node, which is sparse and irregular compared to lattice data. [CL04] provide a very straightforward way to define spatio-temporal outlier and offer a clear four-step method to detect these outliers. It takes the spatial outlier that is not always present in consecutive time frames as spatio-temporal outlier. We can see that the temporal dimension is poorly treated, since it can only compare spatial outliers between immediate time snapshots and the time dimension virtually has no metrics. [LL04] proposed a method based on [ZLK03], which introduces time dimension by linking the center region outliers in different time frames. The method can describe moving outliers well, which can detect some weather events effectively. But it is only good at finding transient spatio-temporal events. Our climate dataset is not suitable to detect such small events because of its coarse precision. So, we still need other methods to find large-scale outliers.

All of the papers above have offered a definition of spatio-temporal outlier, some of which has semantic meaning as in [LL04], while others are only definitions extending upon existing definitions without considering the practical meaning. None of these papers have pay attention to the problem of basic element selection, since it is very obvious. It is not so clear in spatio-temporal lattice data any more, however.

B. Introduction to the Climate Dataset We Use

The dataset "CRU TS 2.0" is offered by Climate Research Unit, UK and thanks to Dr. Tim Mitchell who made this dataset. [Mit03] gives a detail description of the data. It is a 0.5 degree grid dataset with five variables, which are cloud cover percentage, diurnal temperature range, precipitation, temperature, vapor and pressure, covering global land surface and with the time-step of one month from the year 1901.

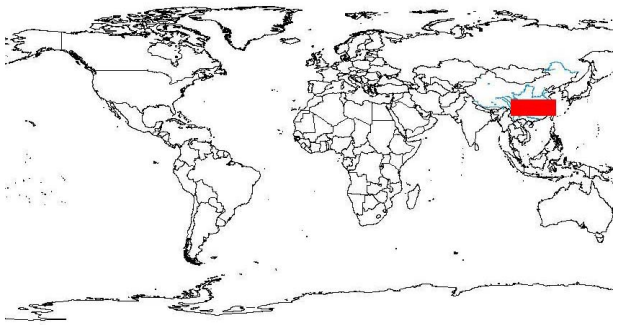

Figure 1. The area of our research data

The data we used is a subset of "CRT TS 2.0", and the area we chose is south china area which is illustrated in the red region in Figure 1. The time period we select is a close time period from 1992 to 2002, from which we hope our result can be verified by our experience.

C. Research Goals

[HK00] points out that outlier mining problem can be decomposed into two sub problems: the first is to define what kind of data can be thought as outliers and the second is to find an efficient method to find such outliers. The fact is that the solution method will come out simultaneously as the definition comes out. So, the definition is the most important step in outlier mining.

The definition of outlier needs to consider 3 aspects. The first is what kind of data structure you care, which we give a name "basic element"; second, what do you want the "basic element" to compare with, i.e., compare element; third, what do you want to compare and by what measurement, i.e., the compare function. Most of the existing definitions only consider the latter two. But basic element in spatio-temporal lattice data is not as obvious as in transaction data, spatial data, and temporal data, etc. In transaction data, we will certainly take a record of a transaction as a basic element; and in spatial data, we are interested in each location with its non-spatial attributes. If we imitate the selection of basic element as in spatial data, the element we should consider would be a spatio-temporal point, which stands for an observation of a certain location at a certain time. But in most occasions, especially in dataset of rough precision (e.g., the dataset we use), what we care is large-scale phenomenon rather than a particular weather event. Thus, we must choose appropriate basic element when we want to find such kind of outliers.

With these considerations, we define two large-scale outliers based on this spatio-temporal climate dataset. These outliers also can be considered as spatio-temporal events and used in a higher granularity level, e.g., mining association rules among the anomalous years of different areas.

III. LOCATION OUTLIERS GIVEN A TIME PERIOD

As in practical life, what we can see is just spatial locations, neglecting the time flying out. We usually ask such question as "which locations always have different temperature from their neighborhoods" (Notice that the location in our dataset in fact a $0.5^o \times 0.5^o$ grid). To make this question more reasonable, we modify the question into "which locations always have different temperature from their neighborhoods in recent 10 years", adding a time limitation. This kind of question is quite meaningful and can find locations with extremely high temperature, and earth scientists can probe further according to this clue.

Following the three aspects of outlier definition process we proposed above, we will give the definition of spatio-temporal outliers of this kind. To make the statement clearer, we take temperature attribute as an example.

A. The basic element

We focus on the spatial location in this dataset, so the basic element is just location or grid. The attributes of the basic element are the whole observations of the temperature time series at this location. We use $< i, L_i, T_i >$ to represent the element with the ID of i, in which L_i stands for its location and T_i stands for its temperature time series.

B. *The compare element*

We are interested in the difference between the location and its neighbors. The neighborhood is defined as 5×5 neighborhood. The compare element is defined as some aggregation functions on the neighborhood.

C. *The compare function*

Now, what we want to compare is just time series, so any compare function used to compare time series or methods to compare high-dimensional vectors can be borrowed, such as correlation and Euclidean distance. But as the time series will get very long, 132 observations in 11 years, the problem comes into a very high dimensional spatial outlier problem. [AY01] has pointed that it is really time-consuming to detect outliers in high dimensional data when the dimensions are not properly handled. We utilize the character of time series to reduce the dimensions, with the permission of the large-scale precision of this dataset. We distill the seasonality and tendency of each time series to represent it. Average temperature of every year is used to interpret tendency, so 11 numeric values are generated. And 12 numeric values of month averages are used to represent the seasonality. The 132-dimension vector thus is reduced to 23 dimensions, and can still be reduced further if we can bare less accuracy. We use V_i to represent the vector of the location i, and $V_{avg(N_i)}$ to represent the vector of the neighborhood average, which can be evaluated by just using the averages of the 24 locations.

Other symbols:

- $Diff_{(N_i)_j}$, the distance of vector between the jth neighbor of location i and $V_{avg(N_i)}$. We use Euclidean distance here.
- $Diff_i$, the distance between the vector of location i and its neighborhood average $V_{avg(N_i)}$.
- μ_i, the mean of $Diff_{(N_i)_j}$.
- σ_i, the standard deviation of $Diff_{(N_i)_j}$.

The compare function is defined as: $f(i) = \dfrac{Diff_i - \mu_i}{\sigma_i}$.

If $f(i) \geq \theta$, we classify location i as a location outlier in the given time period. θ is a parameter that can be adjusted.

The definition of location outlier can be extended to region outliers easily by only replacing the location in the basic element with region. This region can be man-defined area or result of some clustering methods. In this case, the compare element can be any other region rather than the neighbors.

IV. TIME PERIOD OUTLIERS GIVEN AN REGION

In other cases, we would like to find the anomalous time period in a given area. For instance, find the years that with too much precipitation. This problem can be easily solved using simple statistics. However, in most of the time, what we want is more complex. In a certain region, years with drought or flood can't be detected by only considering the average precipitation of the year. As illustrated in Figure 2, although the average precipitation in 1994 and 2002 are larger than the year 1998, flood happened only in 1998.

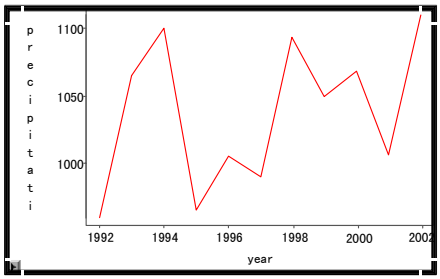

Figure 2. Average precipitaion of 1992-2002 in south China.

What makes the year 1998 so deviant from other years? If we compare the two contour maps of the same month June of 1998 and 2002, we can find that the two maps are so different from each other (see Figure 3 and Figure 4). The truth is that although the average precipitation of the two years is more or less the same, the distribution of the precipitation in space and time is rather different.

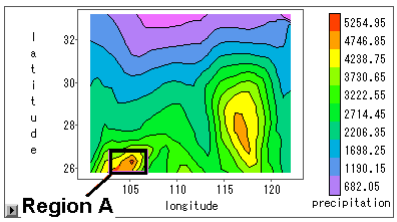

Figure 3. The contour map of precipitation of June, 1998, in south China.

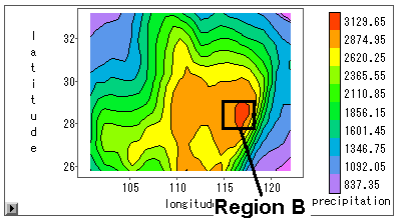

Figure 4. The contour map of precipitation of June, 2002, in south China

According to the analysis above, we give the time period outlier definition as follows.

A. *The basic element*

The emphasis element this time is the time period. Any time period we are interested in is fine, e.g., years, months, and seasons. The attributes of each time period is the spatio-temporal distribution of what we interested, e.g., precipitation in our case. We denote the basic element as $< i, T_i, STDistri_i >$, in which i means the id number, T_i

means the time period with the id number i, and $STDistri_i$ means the spatio-temporal distribution of this time period.

B. *The compare element*

We compare each time period with every other time periods, since we generally don't just compare a certain year with its fore-and-aft years, but compare it with most of the other years. The compare element is defined as some aggregation functions on all the time periods.

C. *The compare function*

The natural way to represent $STDistri_i$ of each time period is using all the observations in this time period i. As the dimension is extremely large (687 locations×12 months in our case), it is really hard to handle. A simple method to solve the problem is just dividing the area into several regions, such as 8×3, and dividing time into 4 seasons. However, the number of dimension is still as many as 96.

Our method is to pick out the sub regions that are so called key regions and the relative more important sub time periods. These key regions and time periods can be told by the domain experts or automatically get by using data mining method.

According to our experience, we choose two sub regions A and B as illustrated in Figure 3 and Figure 4 respectively and the sub time period from June to July of each year. Region A belongs to the upriver area of Yangzi River of China, and B belongs to the downriver area. And the flood of Yangzi River valley only happens in summer period. Thus the attribute vector has only 4 dimensions now.

We use the similar compare function as provided in last section.

- V_{avg}, the vector of averages of 4 attributes that we pick out of all the time periods.
- $Diff_i$, the distance between the vector of time period i and the average vector V_{avg}.
- μ, the mean of $Diff_i$.
- σ, the standard deviation of $Diff_i$.

The compare function is defined as: $f(i) = \frac{Diff_i - \mu}{\sigma}$. If $f(i) \geq \theta$, we take time period i as a time period outlier in the given area. The time period in this case is year.

Some detail parts of this framework can be replaced by others, and the representation vector of the spatio-temporal distribution is one of them.

Given the parameter θ = 1.3, year 1998 will be detected as outlier (see Figure 5), when a big flood indeed happened in Yangzi River Valley. If we set θ = 1.2, two outliers will be found, which are year 1998 and 1995 respectively.

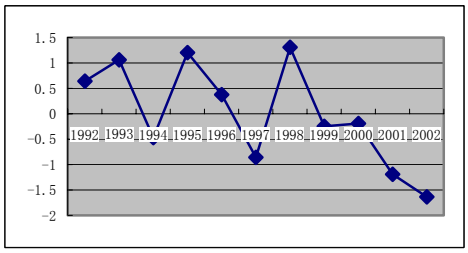

Figure 5. Compare function.

V. CONCLUSION

Outliers in spatio-temporal lattice data haven't got a reasonable definition so far, but they are so helpful in finding interesting phenomena in spatio-temporal dataset. And they can also be used as spatio-temporal events to mine further knowledge using other mining techniques such as spatio-temporal association. In this paper, we point out the importance of clarifying the basic element when defining outliers in spatio-temporal lattice dataset. The introduction of basic element would make the definition clearer and more meaningful. We propose three aspects when defining an outlier, the latter two of which is based on the first one. According to the three aspects, we define two kinds of the most important and useful spatio-temporal outliers based on climate dataset, considering the implement efficiency as well. Experiment shows that the definition is quite helpful in finding outliers with practical utility.

REFERENCES

[1] [AJA04] N.R. Adam, V.P. Janeja, and V. Atluri. Neighborhood based detection of anomalies in high dimensional spatio-temporal Sensor Datasets. SAC'04, pages 576-583, 2004.

[2] [AY01] C.C. Aggarwal and P.S. Yu. Outlier detection for high dimensional data. In ACM SIGMOD Conference, 2001.

[3] [BKNS00]M.M. Breunig, H.P. Kriegel, R.T. Ng, J. Sander. LOF: Identifying Density-Based Local Outliers. ACM SIGMOD Int. Conf. on Management of Data, 2000.

[4] [CL04] T. Cheng and Z. Li. A Hybrid Approach to Detect Spatial-temporal Outliers. In *Proc.* GeoInformaticas 2004, pages 173-178, 2004.

[5] [HK00] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2001.

[6] [KN98] E.M. Knorr and R.T. Ng. Algorithms for Mining Distance-Based Outliers in Large Datasets. VLDB, pages 392-403, 1998.

[7] [KNT00] E.M. Knorr, R.T. Ng, V. Tucakov. Distance-Based Outliers: Algorithms and Applications. The VLDB Journal, Volume 8, Numbers 3-4, pages 237-253, 2000.

[8] [LL04] C. Lu and L.R. Liang. Wavelet Fuzzy Classification for Detecting and Tracking Region Outliers in Meteorological Data. GIS'04, pages 258-265, 2004.

[9] [Mit03] T.D. Mitchell. CRU TS 2.0: Introduction. Tyndall Centre for Climate Change Research, School of Environmental Sciences, 2003.

[10] [PF03] S. Papadimitriou and C. Faloutsos. Cross-Outlier Detection. In *Proc.* SSTD 2003, pages 199-213, 2003.

[11] [SLZ03] S. Shekhar, C.T. Lu, and P. Zhang. A Unified Approach to Spatial Outliers Detection. In *Proc.* GeoInformatica 7(2), pages 139-166, 2003.

[12] [ZLK03] J. Zhao, C. Lu, and Y. Kou. Detecting Region Outliers in Meteorological Data. GIS'03, pages 49-55, 2003.