CSIS Discussion Paper Series
No. 8

A Robust Exploratory Method for Qualitative Trend Curve Analysis against Poor Quality Data

Atsuyuki Okabe* and Atsushi Masuyama**

March 9, 1999

* Center for Spatial Information Science, University of Tokyo
** Department of Urban Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

Abstract

This paper proposes a theoretical method for qualitative trend curve analysis, supposing the situation that data quality is poor and the amount of data is very large. Here the 'trend curve' means, for example, a sequence of monthly temperature over twelve months at a location. The paper first defines 'qualitative' characteristics of a trend curve in terms of isomorphic relations. Second, these concepts are extended to macroscopic qualitative similarity. Third, using the macroscopic qualitative similarity, the paper proposes a method for categorical clustering and a method for measuring the magnitude of qualitative change in trend curves. Fifth, the proposed theoretical method is implemented in a GIS environment and an exploratory tool (a computer program called QuaT) is developed. Sixth, this tool is applied to the analysis of land cover change in the Persian Gulf Area between 1982 and 1993. Last, the paper discusses the limitations of the proposed method.

Acknowledgements

This paper has been developed from the theoretical part of Masuyama, Okabe, Sadahiro and Shibasaki (1998) presented at the 7th Annual Conference of the GIS Association, Japan. We express our thanks to Y. Sadahiro and R. Shibasaki for their valuable comments on an earlier development, and to K. Sakai for her interpretation of our numerical analysis in the Persian Bay Area. This study was partly supported by the National Science Funds, Japan, 09NP1301 (Islamic Area Studies) and 10202201 (Spatial Information Science for Human and Social Sciences).

Figure 1: An illustrative example of trend curves, categorical clustering and qualitative changes

1 Introduction

The objective of this paper is to propose a theoretical as well as practical method for analyzing qualitative characteristics of trend curves that often appear in spatio-temporal analysis, supposing the situation that data quality is poor and the amount of data is very large (such as remotely sensed data). Here the 'trend curve' means, for example, a sequence of monthly temperature over twelve months at a location (Figure 1). First, we develop a method for clustering trend curves with respect to 'qualitative' characteristics of the trend curves (the colored circles in Figure 1). Second, we develop a method for detecting 'qualitative' change in trend curves between two periods in time, say 1980, and 1990. The main concept used in the these methods is 'qualitative similarity' of trend curves, and this concept is fully discussed and formalized mathematically in this paper. Third, we implement these theoretical methods in a GIS environment and develop a user-friendly exploratory tool (a computer program, called QuaT).

We develop the above method as the first-phase analysis in the context of the two-phase analysis: the first-phase analysis is exploratory and the second phase analysis is explanatory. The two-phase analysis is motivated by the recent progress in modern data acquisition technologies, such as remote sensing, global positioning systems and mobile GIS. They bring us a huge amount of data every month or everyday. In such an 'excess' data situation, we are often buried in the data and have a difficulty in fixing a suitable explanatory model or an appropriate hypothesis. To overcome this difficulty,

we first carry out pre-analysis or data mining (Adriaans and Zantinge, 1996), in which we attempt to find potential explanatory models using crude data. The second-phase analysis is the ordinary (or classic) analysis, in which we develop an explanatory model based upon the result of the first-phase analysis and test it with well adjusted data.

In the first-phase analysis, data quality is usually very poor, because data are not adjusted enough, or sometimes the data quality is unknown. In this wild situation, it is useless to apply 'sophisticated' analysis that is sensitive to quantitative values. Rather, we need 'simple' analysis that is robust against poor quality data. We also need efficient computational methods, because a huge amount data requires much processing time and the first-phase analysis often requires fast processing. The method proposed in this paper is developed for the first-phase analysis to satisfy these requirements.

The related literature is numerous. First, the trend curve analysis is discussed in depth in the time-series analysis (Anderson, 1994) that is applied to meteorology (for example, Handcock and Wallis, 1994), hydrology (for example, Rouhani and Wackernagel, 1990) and so forth. We note, however, that those methods mostly deal with quantitative nature of time-series phenomena; less attention is paid to qualitative nature. Second, the exploratory analysis dates back Tukey (1977) in statistics, and it has been developed in geographical analysis since the late 1980's. Examples are Openshaw *et al.* (1987), Haslett *et al.* (1987), Walker and Moore (1988) and Anslein *et al.* (1993), among others. These methods (or tools), however, do not deal with trend curves. Third, trend curves often appear in the analysis of remotely sensed data, because they are periodically provided. A typical analysis of such data is the analysis of seasonal land cover change, for example, DeFries and Townshend (1994) and Millington *et al.* (1994). In these studies, the

major effort is address to how to adjust data with ground truth data. We consider that this type of analyses is carried out in the second-phase. As mentioned in the above, we are concerned with the trend curve analysis in the first phase and such an analysis is few in the related literature except for Eastman and Fulk (1993) and Samson (1993).

2 Qualitative similarity between trend curves

Consider a continuous curve, $f(t)$, defined on $[0, T]$, which is assumed to be second order differentiable and $\mathrm{d}^2 f(t)/\mathrm{d}t^2 \neq 0$ (these assumptions are made just for theoretical consideration; they will be relaxed in Section 5). We classify the characteristic of $f(t)$ in a sufficiently small neighborhood of t according to whether $f(t)$ is non-singular or singular at t, i.e. $\mathrm{d}f(t)/\mathrm{d}t \neq 0$ or $\mathrm{d}f(t)/\mathrm{d}t = 0$. When the former holds, we say that the local characteristic of $f(t)$ around t is a *slope*; when the latter holds, the local characteristic of $f(t)$ around t is a *flat* (Figure 2). Note that the *local characteristic* means the characteristic of $f(t)$ in a sufficiently small neighborhood around t, denoted by $N_\varepsilon(t)$.

We further classify the flats according to $\mathrm{d}^2 f(t)/\mathrm{d}t^2 < 0$ or $\mathrm{d}^2 f(t)/\mathrm{d}t^2 > 0$. When the former holds, we say that the local characteristic of $f(t)$ around t is a *peak*; when the latter holds, the local characteristic of $f(t)$ around t is a *bottom* (Figure 2). In special cases at boundaries $t = 0$ and $t = T$, we define a *peak* by $\lim_{t\to 0+} \mathrm{d}f(t)/\mathrm{d}t \leq 0$ and a *bottom* by $\lim_{t\to 0+} \mathrm{d}f(t)/\mathrm{d}t > 0$; similarly, a *peak* by $\lim_{t\to T-} \mathrm{d}f(t)/\mathrm{d}t \geq 0$ and a *bottom* by $\lim_{t\to T-} \mathrm{d}f(t)/\mathrm{d}t < 0$. For notational convenience we use $C(f(t)) = S, F, P, B$ if the local characteristic of $f(t)$ around t is a slope, flat, peak and bottom, respectively (Figure 2).

Figure 2: Two trend curves that are weakly isomorphic.

It should be noted that these local characteristics are *qualitative* in the sense that a slope, a flat, a peak and a bottom remain a slope, a flat, a peak and a bottom, respectively under a broad class of monotonically increasing transformations. Stated precisely, if $C(f(t)) = X$ in $N_\varepsilon(t)$, then $C(f(g(t))) = X$ in $N_\varepsilon(g(t))$ and $C(g(f(t))) = X$ in $N_\varepsilon(t)$ $(X = S, F, P, B)$ for any monotonically increasing function g.

Now let us consider two trend curves, $f_1(t)$ and $f_2(t)$ (the suffix i may indicate different locations or the same location but different points in time), and let $(t_{i1}, \ldots, t_{in_i})$, $(t_{i1} < \ldots < t_{in_i})$ be the sequence of points in $[0, T]$ at which either $C(f_i(t_{ij})) = P$ or B holds, $j = 1, \ldots, n_i, (t_{i1} = 0, t_{in_i} = T)$, $i = 1, 2$ (Figure 2). Then we can describe a *global characteristic* of $f_i(t)$ over $[0, T]$ by the sequence $\mathcal{C}(f_i) = (C(f_i(t_{i1})), \ldots, C(f_i(t_{in_i})))$, $i = 1, 2$. For example, in Figure 2, $\mathcal{C}(f_1) = (B, P, B, P, B, P, B)$.

For the two trend curves $f_1(t)$ and $f_2(t)$, we consider

Condition 1: $\mathcal{C}(f_1) = \mathcal{C}(f_2)$.

Condition 1 means that if the jth flat in $f_1(t)$ is a peak (bottom), then the jth flat in $f_2(t)$ is also a peak (bottom) for all $j = 1, \ldots, n_1 (= n_2)$. For example, in Figure 2, $C(f_1(t_{11})) = C(f_2(t_{21})) = B; C(f_1(t_{12})) = C(f_2(t_{22})) = P$ and so forth. When Condition 1 holds, we say that the two trend curves $f_1(t)$ and $f_2(t)$ are *weakly isomorphic*, and denote this relation by $f_1(t) \sim f_2(t)$. Since $\mathcal{C}(f_1) = (B, P, B, P, B, P, B) = \mathcal{C}(f_2)$ holds in Figure 2, these two trend curves are weakly isomorphic.

The theoretical notion of the weak isomorphism has practical implications. In the real world we often have data whose acquisition time is in-

accurately recorded. For instance, a watch equipped in a data acquisition device may be inaccurate, or researchers forget to record the exact acquisition date. In such a situation, even if we observe the same phenomenon, we obtain different trend curves $f_1(t)$ and $f_2(s)$ (t and s are used for indicating different time measures). Although they are different, it is quite likely that the sequence of events is preserved over different time measures t and s. Mathematically this implies that the time measure t of $f_1(t_{1j})$ and the time measure s of $f_2(s_{2j})$ have the relation $t_{1j} = g(s_{2j})$ where g is a monotonically increasing function. The weakly isomorphic trend curves imply that the sequence of local characteristics is the same for the two different time measures.

We next introduce a little stronger relation than the weakly isomorphic relation. We order the attribute values of $f_i(t_{ij})$, $j = 1, \ldots, n_i$ from the largest to the smallest, and let $R(f_i(t_{ij}))$ be the rank of $f_i(t_{ij})$ (i.e. the value of $f_i(t_{ij})$ is the $R(f_i(t_{ij}))$th largest value among the values of $f_i(t_{ij}), j = 1, \ldots, n_i$). Then we may describe a *global characteristic* of $f_i(t)$ over $[0, T]$ by the sequence $\mathcal{R}(f_i) = (R(f_i(t_{i1})), \ldots, R(f_i(t_{in_i})))$, $i = 1, 2$. For example, in Figure 3, $\mathcal{R}(f_1) = (6, 2, 3, 1, 5, 4, 7)$. In terms of this sequence, we state

Condition 2: $\mathcal{R}(f_1) = \mathcal{R}(f_2)$.

Note that Condition 2 includes Condition 1. Condition 2 implies that if the jth flat in $f_1(t)$ is the kth highest flat in $f_1(t)$, then the jth flat in $f_2(t)$ is also the kth highest flat in $f_2(t)$, and vice versa. For example, in Figure 3, $R(f_1(t_{11})) = R(f_2(t_{21})) = 6$, $R(f_1(t_{12})) = R(f_2(t_{22})) = 2$, and so forth. When the two trend curves $f_1(t)$ and $f_2(t)$ satisfy Condition 2, we say that $f_1(t)$ and $f_2(t)$ are *strongly isomorphic*, and denote this relation by $f_1(t) \approx f_2(t)$. The two trend curves in Figure 2 are not strongly isomorphic,

Figure 3: Two trend curves that are strongly isomorphic.

but those in Figure 3 are strongly isomorphic ($\mathcal{R}(f_1) = (6, 2, 3, 1, 5, 4, 7) = \mathcal{R}(f_2)$). Obviously, the strongly isomorphic relation implies the weakly isomorphic relation but not converse.

The theoretical notion of the strongly isomorphic relation does not remain a theoretical notion. In reality, it is likely that sensitivity of a sensor is different from device to device, or it may deteriorate over time. As a result, the observed attribute value $u = f_1(t)$ at t is not always comparable with the attribute value $v = f_2(t)$ at t; the measures of u and v are likely to be different or researchers observe attribute values in different manners. In such a situation, even if we observe the same phenomenon, the observed trend curves $u = f_1(t)$ and $v = f_2(t)$ are not identical. It is, however, often plausible to assume that the ranks of attribute values at peaks and bottoms are preserved for two different measures u and v. Mathematically this implies that the measures u and v have the relation $v = g(u)$ where g is a monotonically increasing function.

This assumption, however, might not be acceptable when data acquisition devices are very unstable. In such a wild case, it is safe to assume the weakly isomorphic relation. In this sense, the weakly isomorphic relation is a fundamental relation when we analyze poor quality data.

Both weakly and strongly isomorphic relations show *qualitative similarity* between two trend curves. In the following discussion, when distinction between the weakly isomorphic relation and the strongly isomorphic relation is not necessary or analysis applies to both relations, we just refer to those relations as *isomorphic relations*, or *qualitative similarity*, and denote it by $\simeq$.

Figure 4: Two trend curves that are not microscopically isomorphic but macroscopically isomorphic.

3 Macroscopic qualitative similarity between trend curves

When we defined the isomorphic relations, we assumed that a peak was a peak no matter how small it was, but this assumption is arguable. For example, consider two trend curves in Figure 4. These two curves are not isomorphic, because the curve in panel (a) has 4 peaks, whereas the curve in panel (b) has 3 peaks. We feel, however, that these two curves are 'macroscopically' isomorphic, because the number of 'distinct' peaks are the same for the both curves.

To represent 'macroscopic' isomorphism, we first define the height of a peak explicitly. Suppose that $C(f(t_j)) = P$. From the definition of a peak, it is obvious that $C(f(t_{j-1})) = B$ and $C(f(t_{j+1})) = B$ hold (Figure 5(a), that is, a peak is always in between two adjacent bottoms. In relation to these bottoms, we define the *height*, $h(t_j)$, of the peak at t_j by

$$h(t_j) = f(t_j) - \max\{f(t_{j-1}), f(t_{j+1})\} \tag{1}$$

(Figure 5(a); Okabe, 1982; Okabe and Masuda, 1984). Then we may define a *distinct peak* as the peak whose height is greater than a threshold height, h^*, and we refer to such a peak as an h^*-*distinct peak*. When we pay attention to a macroscopic nature of the trend curve $f_i(t)$, we ignore peaks whose height is less than h^*, or we pay attention to only h^*-distinct peaks. Mathematically this implies that: if $h(t_j) \leq h^*$, the curve $f(t)$ in $t_{j-1} \leq t \leq t_{j+1}$ is replaced with a monotone curve joining $f(t_{j-1})$ and $f(t_{j+1})$ (Figure 5(b)). The resulting curve is denoted by $f(t|h^*)$ and we say that $f(t|h^*)$ is the h^*-*distinct*

Figure 5: An h^*-distinct trend curve of $f(t)$ ((a) $f(t)$, (b) $f(t|h^*)$).

trend curve of $f(t)$.

In terms of $f(t|h^*)$, the statement that two trend curves $f_1(t)$ and $f_2(t)$ are *macroscopically (qualitatively) similar* is precisely written as $f_1(t|h^*) \simeq f_2(t|h^*)$; that is, two h^*-distinct trend curves are isomorphic. Note that $f(t) = f(t|0)$. Also note that deleting h^*-distinct peaks from an original trend curve is similar to filtering in the Fourier analysis. As will be shown in Section 5, however, our method is much simpler and faster than the Fourier method.

The macroscopic similarity changes according to the value of h^*. To illustrate this change, we depict two trend curves, $f_1(t|h^*)$ and $f_2(t|h^*)$ in Figure 6. We notice that $f_1(t|0) \sim f_2(t|0)$. As h^* increases, the isomorphic relation changes. For $0 \leq h^* < 0.2$, $f_1(t|h^*) \sim f_2(t|h^*)$ holds, but at $h^* = 0.2$, the isomorphic relation changes to $f_1(t|h^*) \not\sim f_2(t|h^*)$, and this relation holds for $0.2 \leq h^* < 0.7$. For $0.7 \leq h^* < 1.3$, $f_1(t|h^*) \sim f_2(t|h^*)$. For $1.3 \leq h^* < 1.4$, $f_1(t|h^*) \not\sim f_2(t|h^*)$. At $h^* = 1.4$, two trend curves are isomorphic (one h^*-distinct peak).

When we can fix a specific level h^* of macroscopic similarity, we analyze trend curves in terms of $f_i(t|h^*)$. Sometimes, however, we want to see overall qualitative similarity from a low level (the lowest is $h^* = 0$, i.e. microscopic similarity) to a high level. To measure this overall qualitative similarity, we may use the length of h^* in which two trend curves are isomorphic. For example, in Figure 6 (the lines on the right-hand-side), the heavy line segments shows the range in which the isomorphic relation $f_1(t|h^*) \simeq f_2(t|h^*)$ holds. If the heavy line segment is long, we may consider that two trend curves are qualitatively similar. We hence define the magnitude, $M(f_i, f_k)$, of the

Figure 6: h^*-distinct trend curves with respect to h^*.

overall qualitative similarity between $f_i(t|h^*)$ and $f_k(t|h^*)$ by

$$M(f_i, f_k) = \frac{1}{h^*_{\max}} \int_0^{h^*_{\max}} \delta(f_i(t|x), f_k(t|x))\mathrm{d}x, \tag{2}$$

where

$$\delta(f_i(t|x), f_k(t|x)) = \begin{cases} 1 & \text{if } f_i(t|x) \simeq f_k(t|x), \\ 0 & \text{if } f_i(t|x) \not\simeq f_k(t|x), \end{cases} \tag{3}$$

and $h^*_{\max}$ is the maximum value of h^*. $M(f_i, f_k)$ takes a value between 0 and 1; $M(f_i, f_k) = 0$ implies that the isomorphic relation does not hold at all for any value h^* in $[0, h^*_{\max}]$; $M(f_i, f_k) = 1$ implies that the isomorphic relation holds for all values in $[0, h^*_{\max}]$ (the two trend curves are qualitaitvely similar from the lowest level ($h^* = 0$) to the highest level ($h^* = h^*_{\max}$)). In the example of Figure 6, $M(f_1, f_2) = 0.57$.

4 Categorical classification and qualitative change

We can utilize the isomorphic relation to classify trend curves categorically. Let $f_{is}(t|h^*)$ be an h^*-distinct trend curve at a location $i = 1, \ldots, n$ in a time period $s = 1, \ldots, p$. We can classify the trend curves over space ($i = 1, \ldots, n$) into k classes, $\mathcal{K} = \{K_1, \ldots, K_k\}$ by the isomorphic relation $\simeq$. That is, $f_i(t|h^*) \simeq f'_i(t|h^*)$ holds if and only if $f_i(t|h^*)$ and $f'_i(t|h^*)$ belong to the same class K_j. In particular, when the isomorphic relation is the weakly isomorphic relation ($\sim$), trend curves are characterized by the number of peaks, and so K_j is the class of trend curves which has j peaks.

We make two remarks on $\mathcal{K}$. First, classes in $\mathcal{K}$ are *categorical* in the sense that trend curves are classified by the qualitative similarity (the isomorphic

Figure 7: Qualitative change occurs in a trend curve.

relations). Second, we can control the number k of classes through h^*. As the value of h^* increases, the number of classes decreases.

Now we are ready to define 'qualitative change' in the trend curve $f_{is}(t|h^*)$ at a location $i(=1,\ldots,n)$ in a time period $s(=1,\ldots,p)$. When $f_{is}(t|h^*) \not\approx f_{is+1}(t|h^*)$ holds, we say that *qualitative change* occurs in the trend curve at a location i in a time period s (Figure 7).

Sometimes we wish to measure the magnitude of this qualitative change. When $f_{is}(t|h^*) \not\sim f_{is+1}(t|h^*)$ holds, recalling that the class K_j is the class of trend curves which has j h^*-distinct peaks, the *magnitude of the qualitative change*, $D_1(f_{is},\ f_{is+1})$, may be measured by the difference between the numbers of peaks, i.e.

$$D_1(f_{is}, f_{is+1}) = j - j' \text{ for } f_{is}(t|h^*) \in K_j, f_{is+1}(t|h^*) \in K_{j'}. \tag{4}$$

In the example of Figure 7, $D_1 = +1$. Note that this measure is applicable only to the weakly isomorphic relation. Recalling the overall qualitative similarity defined by equation (2), we can define an alternative measure by

$$D_2(f_{is}, f_{is+1}) = 1 - M(f_{is}, f_{is+1}). \tag{5}$$

In the example of Figure 7, $D_2 = 0.7$. This measure is more general than D_1 because it can be defined not only for the weakly isomorphic relation but also for the strongly isomorphic relation. Using these measures, we can detect where qualitative changes occur in a region. An actual example is shown in Section 5.

5 Implementation: program QuaT

Having shown the theory of how to detect qualitative changes in trend curves over space and time, we now wish to implement this theory in a GIS environment and develop an exploratory tool, called QuaT.

In the above theory, we assume that the function $f(t)$ is continuous, second order differentiable and $\mathrm{d}f(t)/\mathrm{d}t \neq 0$. In practice, however, we usually observe the value of $f(t)$ at a finite number of points, $\hat{t}_1, \ldots, \hat{t}_m$ (filled circles in Figure 8; note that $\hat{t}_i$ used here should be distinguished from t_i or t_{ij} (the location of a peak or a bottom) in the above section). Since this function is discrete, the above theoretical analysis is not directly applicable. It should be noted, however, a few modifications make the above theoretical analysis applicable. First, we can construct a continuous curve from $f(\hat{t}_1), \ldots, f(\hat{t}_m)$ by joining $f(\hat{t}_j)$ and $f(\hat{t}_{j+1})$ by a straight line segment (Figure 8). This modification implies a linear interpolation. Second, we define a slope, a peak and a bottom by:

$$C(f(\hat{t}_j)) = S \text{ if and only if } f(\hat{t}_{j-1}) < f(\hat{t}_j) < f(\hat{t}_{j+1}) \text{ or } f(\hat{t}_j) > f(\hat{t}_{j+1}) > f(\hat{t}_{j+1}); \quad (6)$$

$$C(f(\hat{t}_j)) = P \text{ if and only if } f(\hat{t}_{j-1}) < f(\hat{t}_j) \text{ and } f(\hat{t}_j) > f(\hat{t}_{j+1}); \quad (7)$$

$$C(f(\hat{t}_j)) = B \text{ if and only if } f(\hat{t}_{j-1}) > f(\hat{t}_j) \text{ and } f(\hat{t}_j) < f(\hat{t}_{j+1}). \quad (8)$$

In special cases at boundaries $t = \hat{t}_1$ and $t = \hat{t}_m$,

$$C(f(\hat{t}_1)) = P \text{ if and only if } f(\hat{t}_1) > f(\hat{t}_2) \text{ or } f(\hat{t}_{m-1}) < f(\hat{t}_m); \quad (9)$$

$$C(f(\hat{t}_1)) = B \text{ if and only if } f(\hat{t}_1) < f(\hat{t}_2) \text{ or } f(\hat{t}_{m-1}) > f(\hat{t}_m). \quad (10)$$

Figure 8: A piece-linear trend curve.

By these modifications, we can apply the analysis of a continuous trend curve $f(t)$, $0 \leq t \leq T$ to the analysis of a discrete trend curve $(f(\hat{t}_1), \ldots, f(\hat{t}_m))$.

Given these modifications, we can implement the method by the following procedure.

Program QuaT

Step 0 (initial setting). Input data are: discrete trend curves $f_{is}(\hat{t}_j)$ $(j = 1, \ldots, m)$ at locations $i = 1, \ldots, n$ in time periods $s = 1, \ldots, p$.

Step 1. Find peaks and bottoms in $f_{is}(\hat{t}_j|0), j = 1, \ldots, m$ by equations (6) - (10) $(i = 1, \ldots, n; s = 1, \ldots, p)$.

Step 2. Compute the height of each peak in $f_{is}(\hat{t}_j|0), j = 1, \ldots, m$ by equation (1) $(i = 1, \ldots, n; s = 1, \ldots, p)$.

Step 3. Construct $f_{is}(\hat{t}_j|h^*), j = 1, \ldots, m$ $(i = 1, \ldots, n; s = 1, \ldots, p)$.

Step 4. Classify $f_{is}(\hat{t}_j|h^*), j = 1, \ldots, m$ $(i = 1, \ldots, n; s = 1, \ldots, p)$ according to an appropriate isomorphic relation ($\sim$ or $\approx$), and display the resulting classes over space.

Step 5. Compute the magnitude of qualitative change in f_{is} $(i = 1, \ldots, n)$ over time periods $s = 1, \ldots, p$ by equation (4) or (5), and display the magnitude over space.

This program runs fast. The order of computational time in Step 1 is O(mnp), and that in Step 2 is also O(mnp). In Step 3, we order the heights of the peak from the lowest to the highest, and this ordering requires O($m \log m$) time for one trend curve. Thus the order in Step 3 is

$O(npm \log m)$. Step 4 requires $O(kn)$ time where k is the number of classes. Step 5 requires $O(kn)$. In theory, the order of the total computational time is $\max\{O(npm \log m), O(kn)\}$. In practice, we suppose that the attribute values in a region are given by raster type data (such as remotely sensed data), the order of the number n of pixels representing a region is higher than that of m, k and p. Thus, the total computational time is dominated by n, implying that QuaT runs with the linear time order of the number of locations.

6 Application of QuaT to the analysis of seasonal land cover (NDVI) change in the Persian Gulf Area between 1982 and 1993

We applied QuaT to the analysis of seasonal land cover change in the Persian Gulf Area (E40° − 50°, N27° − 37°) between 1982 and 1993. The data source was Pathfinder AVHRR Land Data Set (NOAA-7, -9 and -11) in 1982 and 1993 (Smith *et al.*, 1997). The area consists of 100 by 100 pixels; consequently, one pixel represents a 0.1° by 0.1° region (approximately 8 km by 8 km). We used the NDVI (Normalized Difference Vegetation Index; roughly speaking, NDVI indicates the amount of vegetation) over twelve months, which gave the trend curves $f_{is}(\hat{t}_j)$, $j = 1, \ldots, 12$, $i = 1, \ldots, 10000$, $s = 1982, 1993$.

The quality of the data was not satisfactory. First, we tried to remove the effect of the clouds by taking the maximum value of NDVI among the NDVI values obtained at three time-points in a month. This treatment resulted in unequal time intervals because the maximum value was achieved at a different time-point in each month (note that the order of observation times

Figure 9: The frequency distribution of peak heights.

Figure 10: Categorical classification of NDVI trend curves in the Persian Gulf Area.

over twelve months are preserved). Second, it is quite likely that the data quality changed between 1982 and 1993. Third, usually remotely sensed data should be adjusted by the ground truth data, but it was difficult to do so because such data were difficult to obtain in the Persian Gulf Area. Fourth, the amount of data was huge. Although we finally used only 1982 and 1993, but on the way of the analysis, we used three sets of data in every month over twelve months over 12 years; consequently the total data amounted to 10000 (pixel) $\times$ 3 (times in a month) $\times$ 12 (months) $\times$ 12 (years) $= 4.32 \times 10^6$. Considering these factors, we considered that QuaT was appropriate for the first-phase analysis.

The input data were the NDVI values $f_{is}(\hat{t}_j)$, $j = 1, \ldots, 12$ (twelve months), $i = 1, \ldots, 10000$ (locations), $s = 1983, 1992$ (years) (Step 0). QuaT computed peaks and bottoms (Step 1), and gave the heights of the peaks (Step 2). To determine an appropriate value of h^*, QuaT computed the distribution of peak heights (Step 3). The result is shown in Figure 9. From this distribution, QuaT chose the value $h^* = 0.08$ at which the frequency drastically changed (we also examined several values and realized that this value was an appropriate value) (Step 3).

For $h^* = 0.08$, QuaT classified the NDVI trend curves in the Persian Gulf Area. The result is shown in Figure 10 (Step 4). Finally, QuaT computed qualitative change in the NDVI trend curves between 1982 and 1993 using two measures (Step 5). Figure 11 shows the magnitude of the qualitative

Figure 11: Qualitative change in NDVI trend curves between 1982 and 1993 in the Persian Gulf Area measured by D_1 given by equation (4).

Figure 12: Qualitative change in land cover (NDVI trend curves) between 1982 and 1993 measured by D_2 given by equation (5).

change in terms of D_1 given by equation (4), and Figure 12 shows that in terms of D_2 given by equation (5).

7 Concluding discussion

In addition to the above application, we tested the performance of QuaT in several empirical examples. From these tests, we found the performance quite satisfactory.

First, QuaT runs very fast as was theoretically examined in Section 4.

Second, QuaT is robust against poor quality data. This robustness results from the property that QuaT deals with qualitative characteristics.

One might question, however, that the isomorphic relations, in particular the weakly isomorphic relation assumed in QuaT is too crude. At first glance, it looks so, but it is not so crude as one might feel. One should recall the overall similarity $D_2(f_{is}, f_{is+1})$ shown in Figure 6. This similarity considers the height of peaks from the lowest peak to the highest peak, implying that the height of the peaks are implicitly taken into account even if we use the weakly isomorphic relation.

Obviously no methods are almighty and QuaT has some limitations. To see them, we depict Figure 13 which shows a good contrast between QuaT and an ordinary quantitative method. Figure 13(a) shows three clusters obtained by applying the K-means method (Hartigan, 1975) to the multivariate

Figure 13: (a) Clusters obtained by the ordinary clustering method (the K-means method) and (b) those obtained by QuaT.

data $(f_i(\hat{t}_{i1}), \ldots, f(\hat{t}_{i12}))$, $i = 1, \ldots, 10000$. Figure 13(b) shows three clusters obtained by QuaT with $h^* = 0.08$. In both figures, the flat trend curves seem to indicate sandy soil areas. Figure 13(a) seems to show two kinds of crops areas (probably wheat and rice). On the other hand, Figure 13(b) seems to show single-cropping areas and double-cropping areas. This contrast shows an advantage and a disadvantage of QuaT and the ordinary method. QauT cannot distinguish one-peak trend curves in Figure 13(a) but the K-means method can; the K-means method cannot detect double-cropping areas but QuaT can.

This comparison suggests that QuaT should be applied in the first-phase analysis. In the second-phase analysis, the K-means method should be applied to one-peak trend curves obtained by QuaT. Then we can detect quantitative as well as qualitative characteristics of trend curves.

In conclusion, QuaT shows good performance in the first-phase analysis.

References

Adriaans, P. and Zantinge, D. (1996) *Data Mining*, London: Longman.

Anderson,T.W. (1994) *The Statistical Analysis of Time Series*, Chichester: John Wiley and Sons.

Anslein, L., Dodson, R.F. and Hudak, S. (1993) Linking GIS and spatial data analysis in practice, *Geographical systems*, **1**, 3-23.

DeFries, R.S., and Townshend, J.R.G (1994) Global Land Cover: Comparison of Ground Based Data Sets to Classifications With AVHRR Data, in *Enviromental Remote Sensing from Regional to Global Scales*, G. Foody

and P. Curran, Editors, 84-110, John Wiley and Sons, Chichester.

Eastman, J.R. and Fulk, M. (1993) Long Sequence Time Series Evaluation Using Standard Principal Components, *Photogrammetric Engineering and Remote Sensing*, **59(6)**, 991-996.

Handcock, M.S. and Wallis, R.W. (1994) An Approach to Statistical Spatial-Temporal Modeling of Meteorological Fields, *Journal of the American Statistical Association*, **89**, 368-378.

Hartigan, J.A. (1975) *Clustering Algorithms*, New York: John Wiley and Sons.

Haslett, J., Willis, G. and Wise, S. (1990) SPIDER-an interactive statistical tool for the analysis of spatially distributed data, *International Journal of Geographical Information Systems*, **6**, 407-423.

Masuyama, A., Okabe, A., Sadahiro, Y. and Shibasaki, R. (1998) A Robust Method for analyzing Trend Curves, *Papers and Proceedings of the Geographic Information Systems Association*, **7**, 103-106 (in Japanese).

Millington, A.C., Wellens, J., Settle, J.J. and Saull, R.J. (1994) Explaining and Monitoring Land Cover Dynamics in Drylands Using Multi-temporal Analysis of NOAA AVHRR Imagery, in *Enviromental Remote Sensing from Regional to Global Scales*, G. Foody and P. Curran, Editors, 16-43, John Wiley and Sons, Chichester.

Okabe, A. (1982) A qualitative method of trend curve analysis, *Environment and Planning A*, **14**, 623-627.

Okabe, A. and Masuda, S. (1984) Qualitative Analysis of Two-dimensional Urban Population Distributions in Japan, *Geographical Analysis*, **16(4)**, 301-312.

Openshaw, S., Charlton, M., Wymer, C. and Craft, A. (1987) A Mark 1 Geographical Analysis Machine for the automated analysis of point data

sets, *International Journal of Geographical Information Systems*, **1(4)**, 335-358.

Rouhani, S. and Wackernagel, H. (1990) Multivariate geostatistical approach to space-time data analysis, *Water Resources Research*, **26(4)**,585-591.

Samson, S.A (1993) Two Indices to Characterize Temporal Patterns in the Spectral Response of Vegetation, *Photogrammetric Engineering and Remote Sensing*, **59(4)**, 511-517.

Smith, P.M., Kalluri, S.N.V., Prince, S.D. and DeFries, R. (1997) The NOAA/ NASA Pathfinder AVHRR 8-Km Land Data Set, *Photogrammetric Engineering and Remote Sensing*, **63(1)**, 12-32.

Tukey, J.W. (1977) *Exploratory Data Analysis*, Massachusetts: Addison-Wesley.

Walker, P. A. and Moore, D.M. (1988) SIMPLE: an inductive modelling and mapping tool for spatially-oriented data, *International Journal of Geographical Information Systems*, **2**, 347-363.

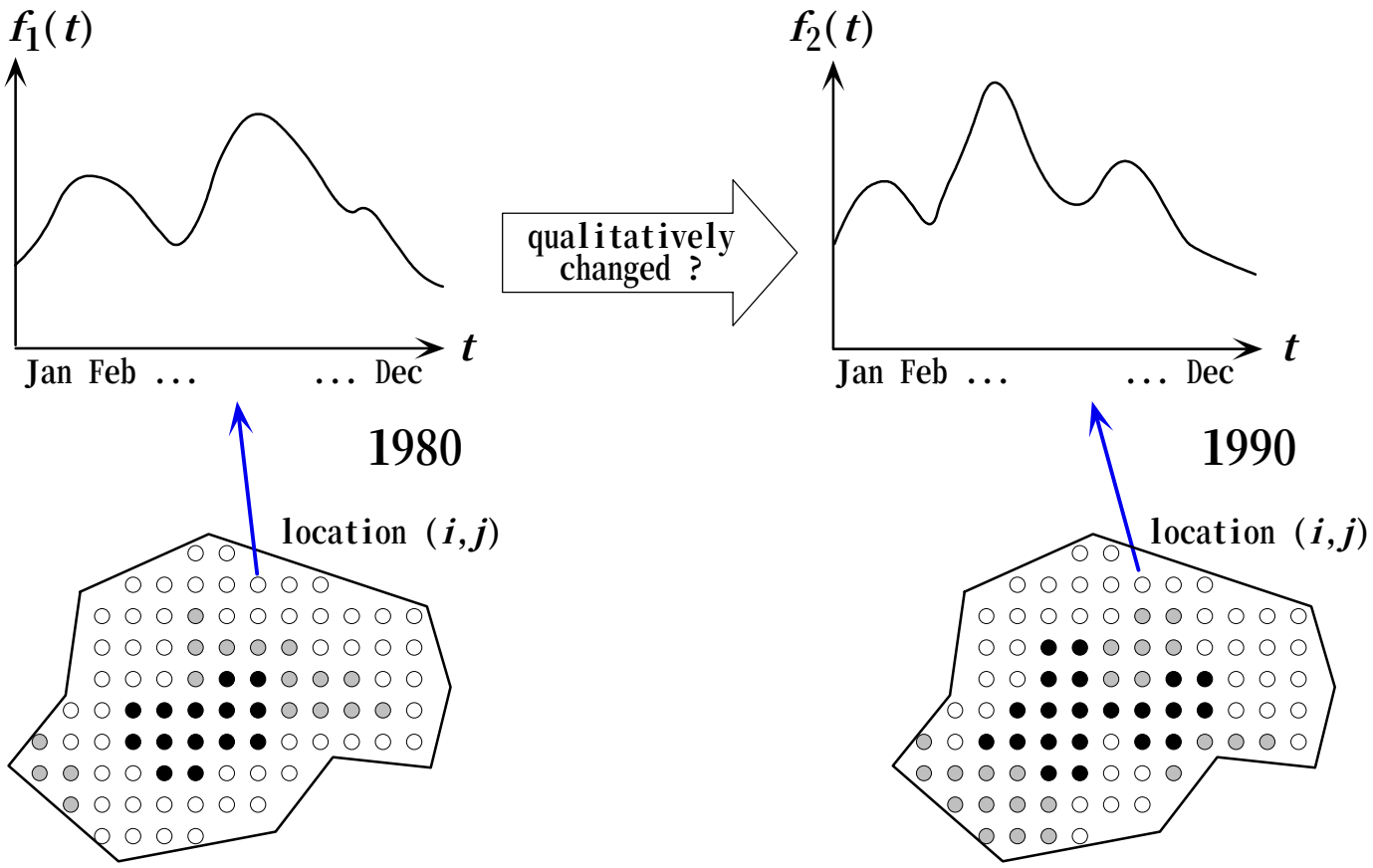
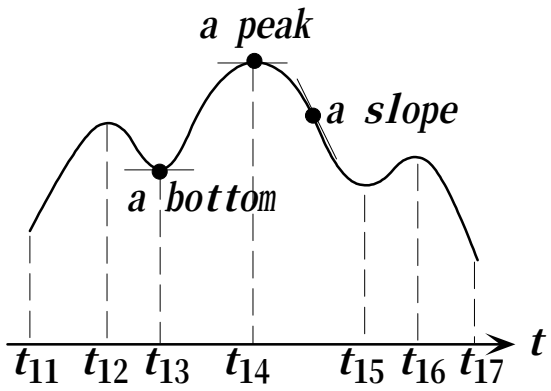
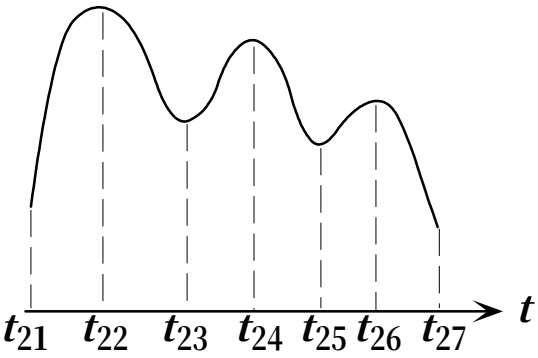

$f_1(t)$
t
Jan Feb ...
... Dec
1980
location (i,j)
qualitatively changed ?
$f_2(t)$
t
Jan Feb ...
... Dec
1990
location (i,j)

Figure 1

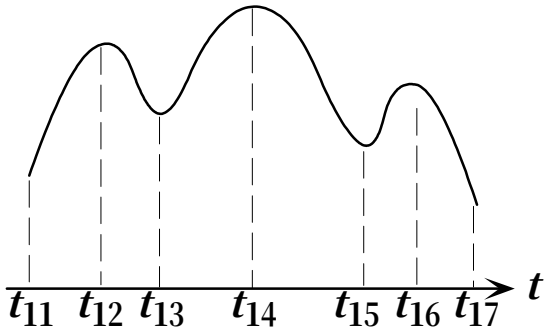


$C(t_{1j})$=(B P P B P B)

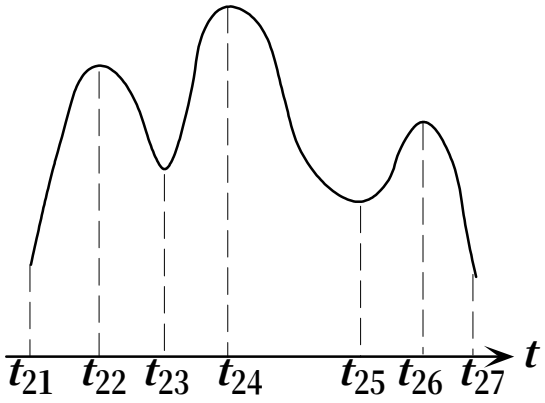


$C(t_{2j})$=(B P B P B P B)

Figure 2



$R(t_{1j})$=(6 2 4 1 5 3 7)



$R(t_{1j})$=(6 2 4 1 5 3 7)

Figure 3

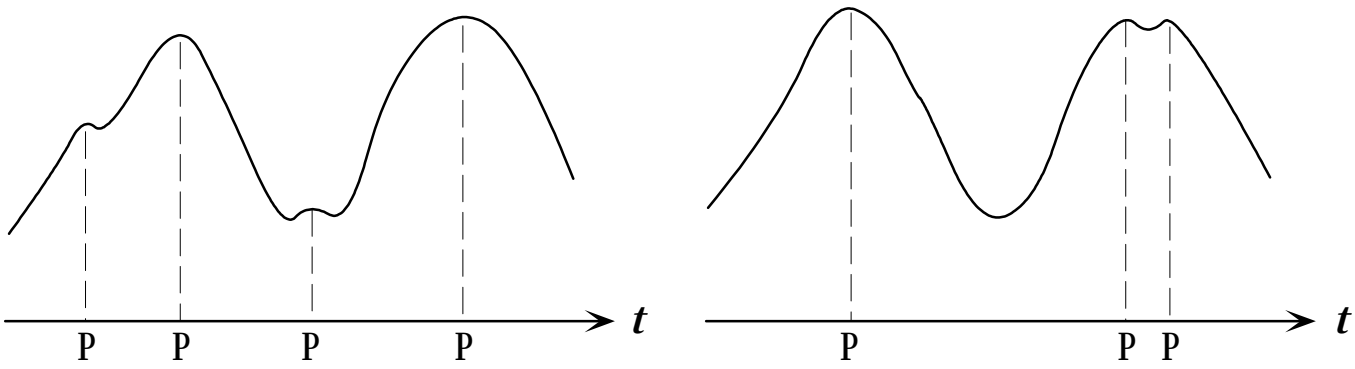
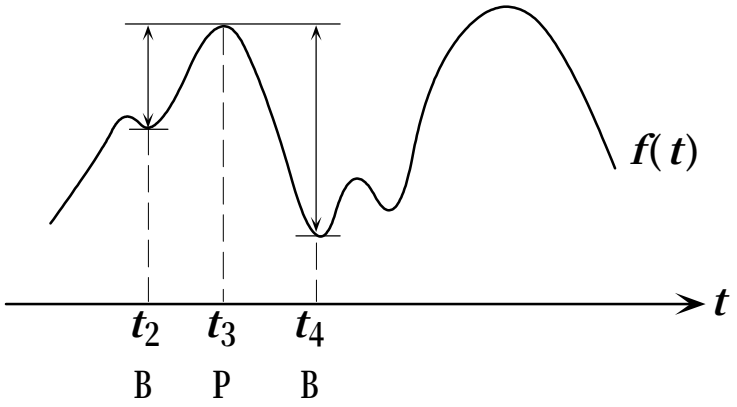
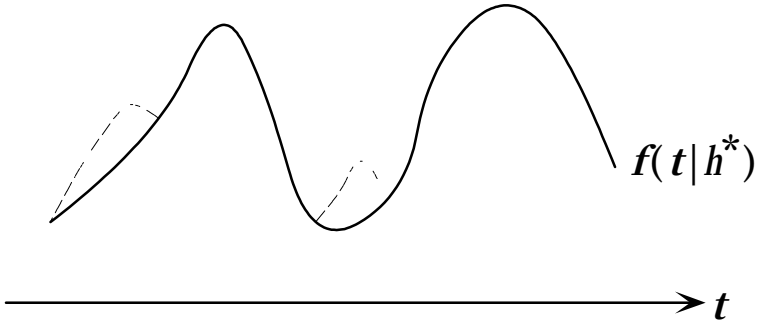


Figure 4



(a)



(b)

Figure 5

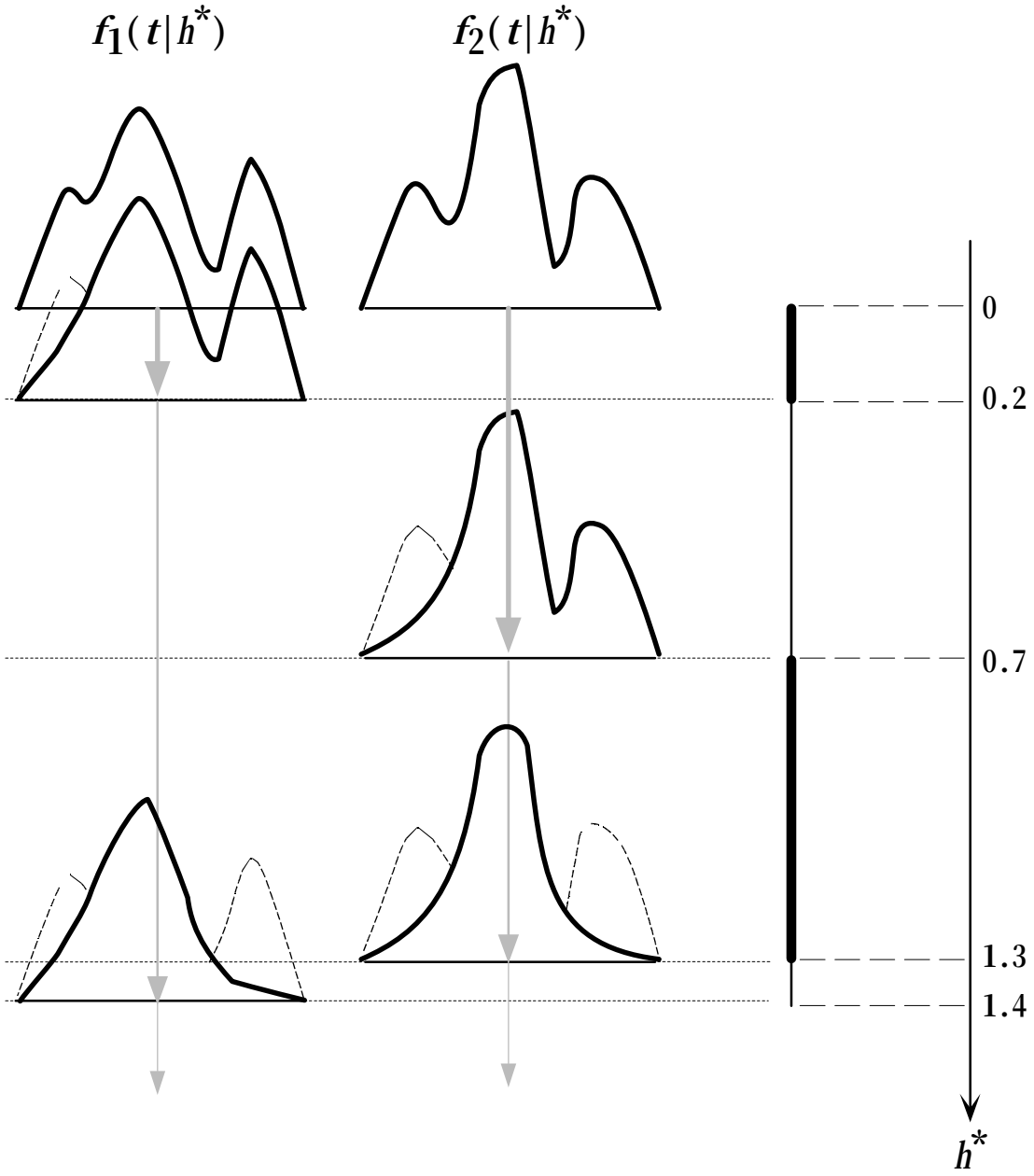
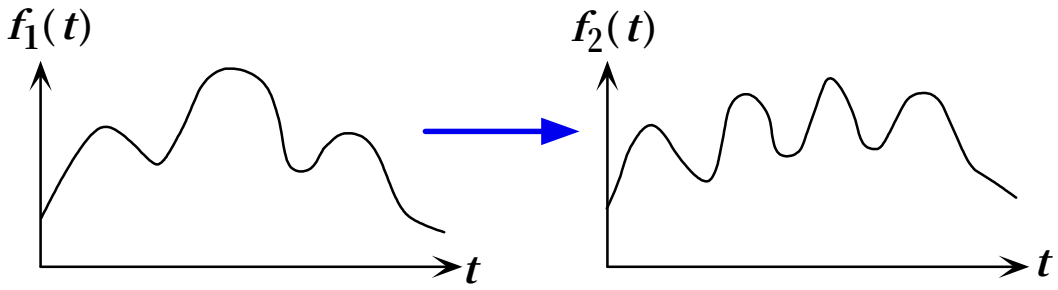


Figure 6



D_1	D_2
+1	0.7

Figure 7

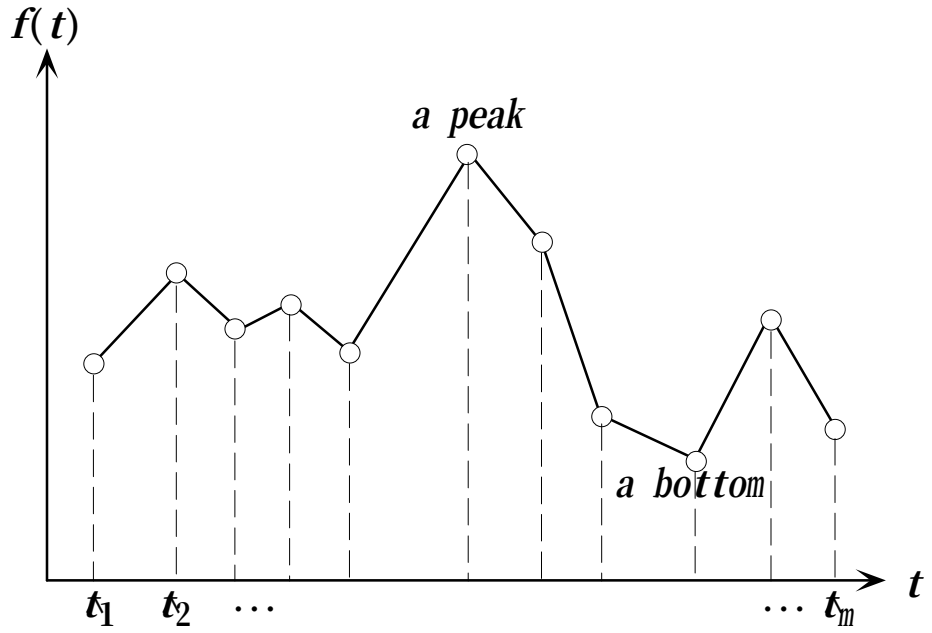


Figure 8

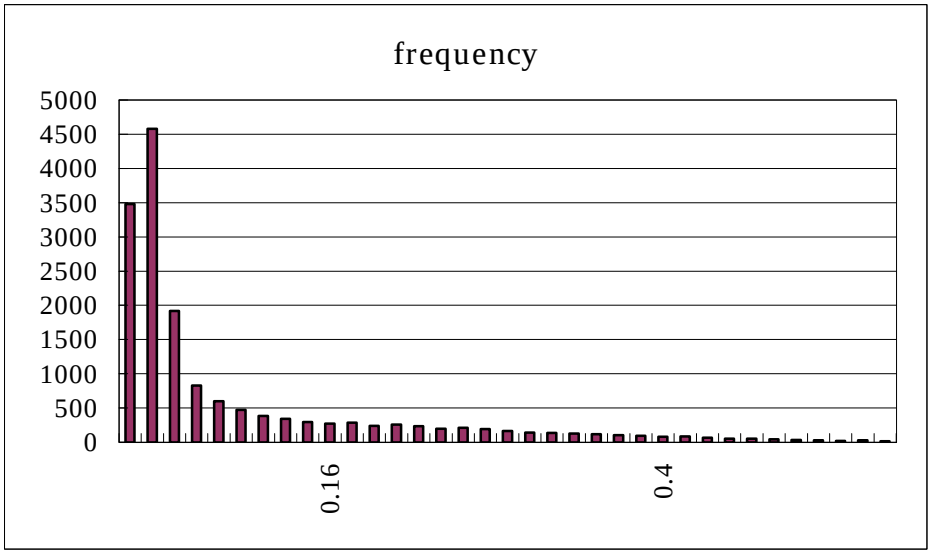

frequency
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0.16
0.4

Figure 9

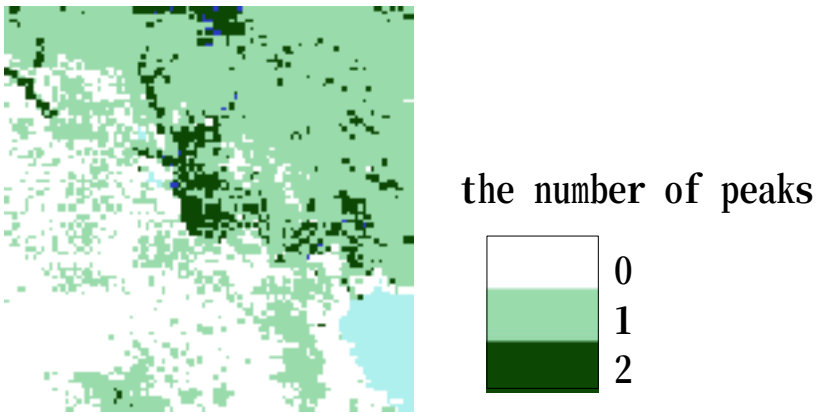


Figure 10

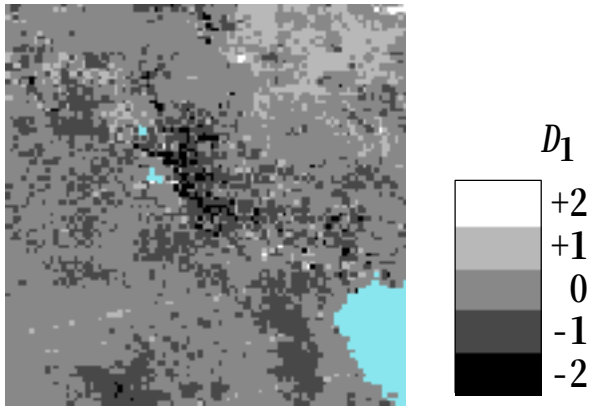


Figure 11

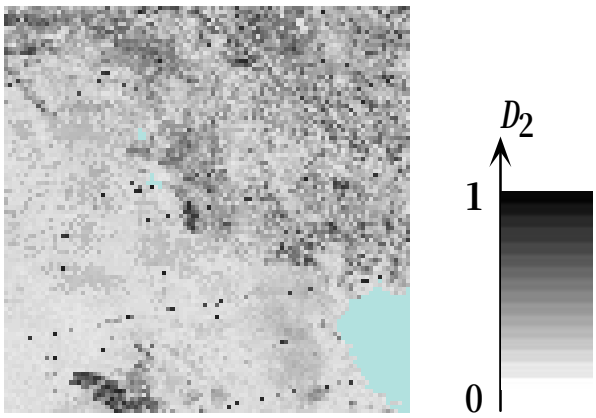
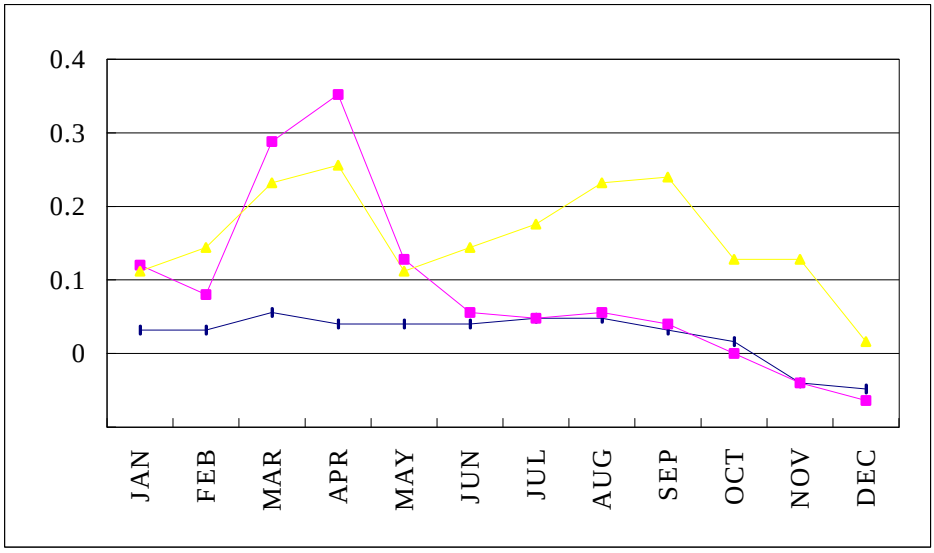
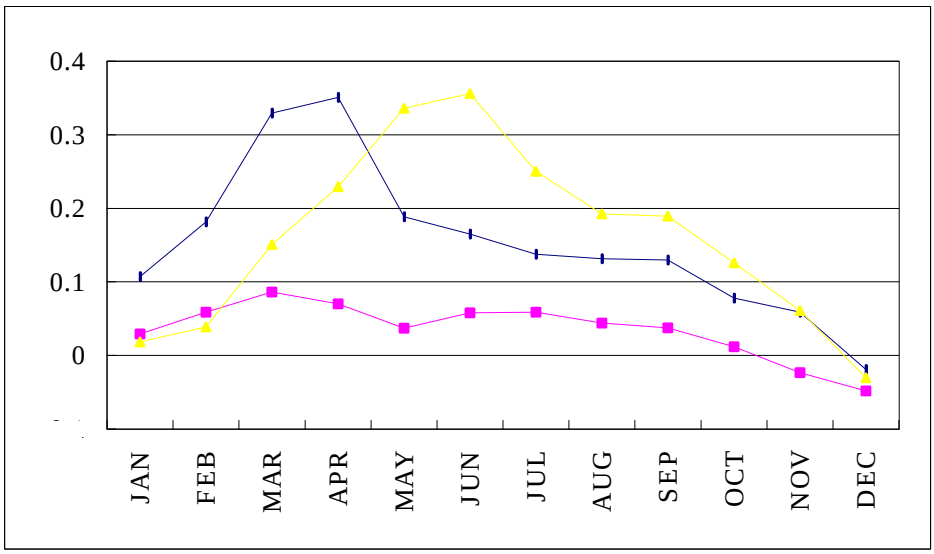


Figure 12



QuaT



K-means

Figure 13