Department of Informatics, University of Zürich

Self study project

Survey of Peaks/Valleys identification in Time Series

Roger Schneider

Matrikelnummer: 89-708-903

Email: roger.schneider@uzh.ch

August 23, 2011

supervised by Prof. Dr. M. Böhlen and M. Khayati

University of Zurich UZH

Department of Informatics

Abstract

The detection of peaks and valleys in time series is a long-standing problem in many applications. Peaks and valleys represent the most interesting trends in time series.

In the purpose to identify these trends in time series, we investigate two approaches of trends' detection. The first approach is based on a geometric definition of the trends. The second one uses a statistical definition of peaks and valleys. The two approaches are able to detect significant trends within time series. Nevertheless, only the statistical approach is able to find these trends in a global context.

In this report we describe, define and illustrate algorithms of the geometric approach and the statistical approach.

Contents

1 Introduction

1.1 Problem Definition

Time series data arise in a variety of domains, such as environmental, telecommunication, financial, and medical data. For example, in the field of hydrology, sensors are used to capture environmental phenomena including temperature, air pressure, and humidity at different points in time. This data is characterized by a big number of fluctuations. These fluctuations are mainly categorized into peaks and valleys as shown figure 1.1.

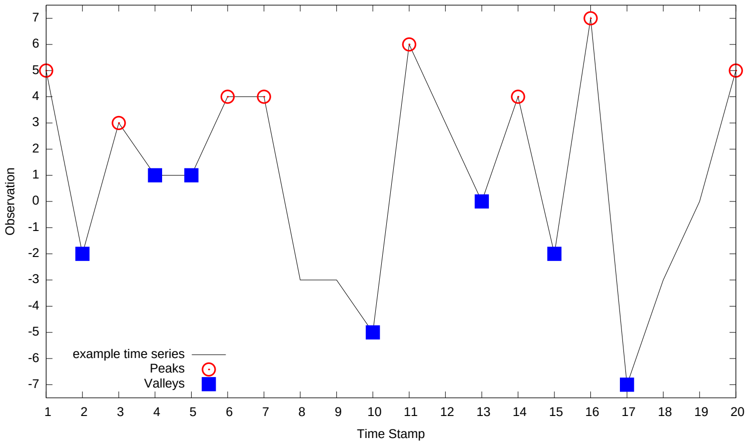

Figure 1.1: Examples of peaks and valleys in a time series

The detection of these fluctuations is of great importance for hydrologists. The identification of these fluctuations will make easy to apply time series analysis techniques e.g, sequence similarity, pattern recognition, missing values prediction. We are study in this report some of the most interesting peaks and valleys detection algorithms.

1.2 Motivation

Peaks and valleys denote significant events in time series. These events can be described as an abruptly increase on the heart rate or a sudden decrease in price on stock trading. Otherwise With these properties we can describe time series or their similarities to other time series.

A peak or a valley is an significant event within a mathematical function or a time series. A significant event is a point where the function graph changes from increasing (decreasing) behaviour to decreasing (increasing) behaviour. The identification of these behaviour is important in order to carry out analysis on the data.

1.3 Contribution

The main contributions of this work are the following:

- We propose a formal description of time series, peaks and valleys
- We describe some algorithms able to detect the most significant peaks and valleys.
- We evaluate the accuracy of these algorithms of real world hydrological data sets

2 Background

2.1 Notations

In order to state the problem and concepts clearly, we define some notations and terminologies in table 2.1.

Symbol	Description
(t_i, x_i)	representation of an observation by the time and observation pair
T	time series, set of *time* and *value* pairs
$f(x), g(x)$	user-defined continuous function
f_0, f_1	function values of function $f(x)$ at position x_0 and x_1
$f(x; \vec{\alpha})$	model function, for approximation with $\vec{\alpha}$ supporting points
$f'(x), f''(x)$	first and second derivation of function $f(x)$

Table 2.1: Notation of symbols used in the paper

2.2 Time Series

We define a time series as a sequence of observations on a specific attribute. Observations are measured variables such as temperature or relative moisture at a given time stamp. An observation is represented by the pair (time, value). Such a pair constitutes the smallest entity of a time series.

Without loss of generality we represent the pair (time,value) as (t_i, x_i) where t_i refers to the timestamp of the i^{th} observation and x_i refers to the value of the i^{th} observation. A time series is a sequence of n (t_i, x_i) pairs.

Formally, a time series T is described as follows:

$$\begin{aligned} T = \quad & \{(t_1, x_1), (t_2, x_2), \ldots, (t_n, x_n)\} \textit{ if and only if} \\ & \forall i, j : (t_i, x_i), (t_j, x_j) \in T \wedge i \leq j \Rightarrow t_i \leq t_j \end{aligned} \tag{2.1}$$

2.3 Peak and Valley

From a mathematical point of view, a peak and a valley represent respectively a local maxima and a local minima [1].

Let $f(x)$ be a function which transforms x from an user-defined subdomain $A \subseteq \mathbb{R}$ to the domain $\mathbb{R}$ as follows:

$$f : A \to \mathbb{R}$$

Let's consider an interval $\mathbf{I} = (a, b)$ and let's assume $\mathbf{I} \cap A \neq \emptyset$. A **local maximum** is detected at point $x_0 \in \mathbf{I}$ if

$$f(x_0) \geq f(x) \ , \ \forall x \in \mathbf{I}$$

The difference between a global maximum and a local maximum is the domain of **I**. If $\mathbf{I} \cap A = A$ than we obtain a **global maximum**.

$$f(x_0) \geq f(x) \ , \ \forall x \in A \tag{2.2}$$

Similarly, a **local minimum** is detected at point x_0 if:

$$f(x_0) \leq f(x) \ , \ \forall x \in \mathbf{I}$$

And the **global minimum** is detected at x_0 if

$$f(x_0) \leq f(x) \ , \ \forall x \in A \tag{2.3}$$

Based on the previous definitions, we consider a peak as local maximum and a valley as a local minimum.

Figure 2.1 illustrates a time series containing peaks and valleys.

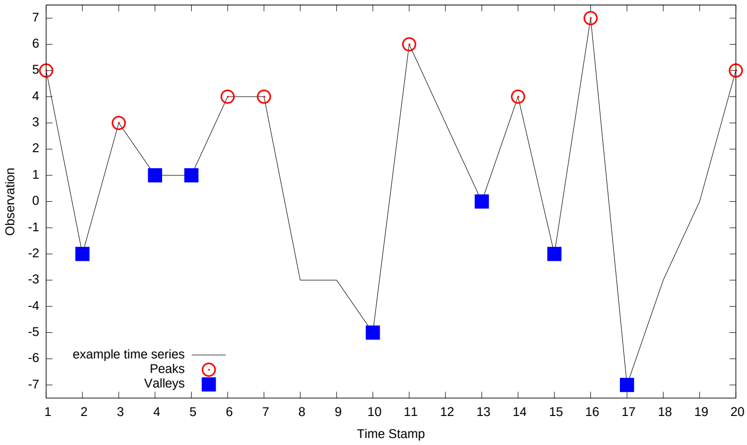

Figure 2.1: Examples of local and global peaks (valleys)

The application of the previous definitions on the example of figure 2.1 gives the following extrema:

Extremum	time/value pair
local peak	(1, 5), (3, 3), (6, 4), (7, 4), (11, 6), (14, 4), (16, 7), (20, 5)
global peak	(16, 7)
local valley	(2, -2), (4, 1), (5, 1), (8, -3), (9, -3), (10, -5), (13, 0), (15, -2), (17, -7)
global valley	(17, -7)

2.4 Continuity constraint

The algorithms of detection of peaks and valleys have to fulfill some requirements. The principal requirement is to assume that the time series is represented by a real function. The latter guarantees there exist points between any two given points of the function. This requirement defines the continuity principle.

A function f is continuous at a point x_0 if there exist for a given $\epsilon > 0$ a $\delta > 0$ and $\forall x \in dom(f)$ we have:

$$|x - x_0| < \delta \Rightarrow |f(x) - f(x_0)| < \epsilon$$

The variable δ depends on variable ϵ. A function $f(x)$ is denoted as continuous if the function at every $x_0 \in dom(f)$ is continuous.

A continuous function maps all points x with a distance $< \epsilon$ from x_0 to points $f(x_0)$ with a distance $< \delta$.

Any change in the area around x_0 will produce the same change in the domain of $f(x_0)$.

2.5 Derivation definition

We define in this subsection the concept of derivative using tangent lines. A tangent defines a linear slope which contacts a given function $f(x)$ in a given point x_0.

We use the angle between the tangent and the horizontal axis to describe the behaviour of the function $f(x)$ in point x_0. The value of the angle defines the slope of the tangent at $f(x_0)$. A positive value of the angle denotes an increase trend of the function $f(x)$ in point x_0. Negative angle denotes a decreasing trend. For an angle $= 0$, there is a flat trend in point x_0. Therefore, we have a local extrema in point x_0.

The definition of derivative can be described as followed: A derivative is the approximation of the tangent through the secant given by $f(x_0)$ and $f(x_0 + h)$ where $h \in \mathbb{R} \wedge h \to 0$.

A function $f(x)$ is **differentiable** at position x_0, if there exists a limit for $x \to x_0$ such that:

$$\lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} =: f'(x_0) \tag{2.4}$$

If such a limit exists we call it derivative. An example of derivative is shown in figure 4.4.

Figure 4.4 shows the approximation of $f'(x_0)$ with $x \rightarrow x_0$. The tangent in $f(x_0)$ is a linear slope. The slope of $f(x)$ in point x_0 is equal to $m = \dfrac{f(x_0+h) - f(x_0)}{(x_0+h) - x_0}$

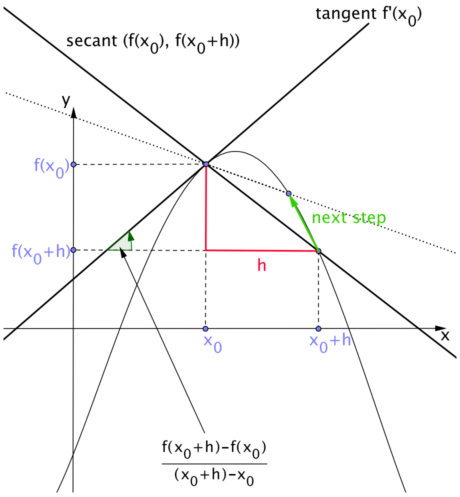

Figure 2.2: First deviation of the continuous function $f(x)$ in point x_0

If $f(x)$ is a differentiable function with an existing derivation function $f'(x)$ and at point $x_0 \in \mathbf{I} \subset \mathbb{R}$ exists $f'(x_0) = 0$ than $f(x)$ has a local maximum or a local minimum in point x_0.

2.6 Theorem of Rolle

If following conditions are satisfied:

1. $f(x)$ is continuous in interval $[a, b]$
2. $f(x)$ is in interval (a, b) differentiable
3. it obtains $f(a) = f(b)$

The theorem of Rolle states that there exists at least one position $x_0 \in (a, b)$ having $f'(x_0) = 0$.

Rolle's theorem give us the guarantee that under the three conditions above there exists one local maximum or local minimum at least between the point a and b.
In figure 2.3 there is a local peak with $f'(x_0) = 0$ at position x_0

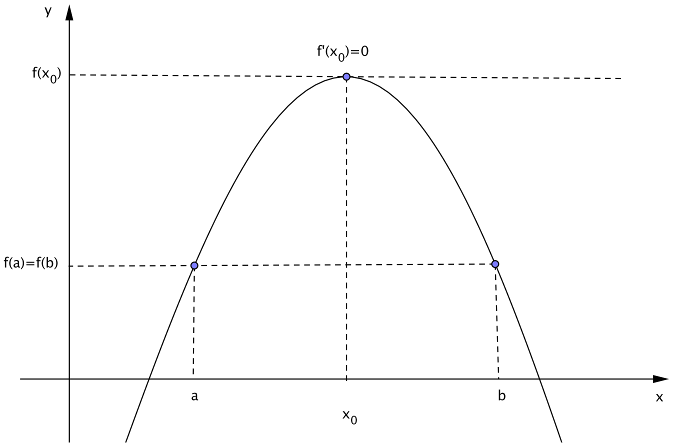

Figure 2.3: Illustration of the theorem of Rolle

2.7 Extrema in noncontinuous functions

Until now, we considered continuous functions $f(x)$ in which we can compute a local maxima or a local minima under certain conditions. A time series is function of following form:

$$\begin{aligned} T: N &\rightarrow \mathbb{R} \\ i &\rightarrow x_i \end{aligned}$$

T is a function which assigns any natural number i in $N := \{1, 2, 3, \ldots, n\} \subset \mathbb{N}$ exactly one real number $x_i \subset \mathbb{R}$. Function T is a discrete function. Therefore, T is not continuous and T is not differentiable.

If we connect every observation x_i of a time series T with its adjacent neighbours x_{i-1} and x_{i+1} through a linear slope then we get two linear functions which represent the three points. A local extrema is given by the intersection of two linear slopes. Each linear slope is represented through its own linear function $f(x), g(x)$. Each function contains its own slope $m = \frac{\triangle y}{\triangle x}$. In a local extrema we find two slopes m each has a different sign.

We define a local peak following:

There exist two linear functions $f(x)$ and $g(x)$ with $f(x_0) = g(x_0)$ and $x_0 \in \mathbb{R}$. The linear functions have the following form:

$$\begin{aligned} f(x) &= a + m_f x \\ g(x) &= b + m_g x \end{aligned}$$

Assumption: $f(x)$ connects the left neighbour point of x_0 with x_0 and $g(x)$ connects the right neighbour point of x_0 with x_0 . Then in point x_0 exists a **local peak** if $m_f > 0$ and $m_g < 0$. Contrary, $m_f < 0$ and $m_g > 0$ denotes a **local valley**.

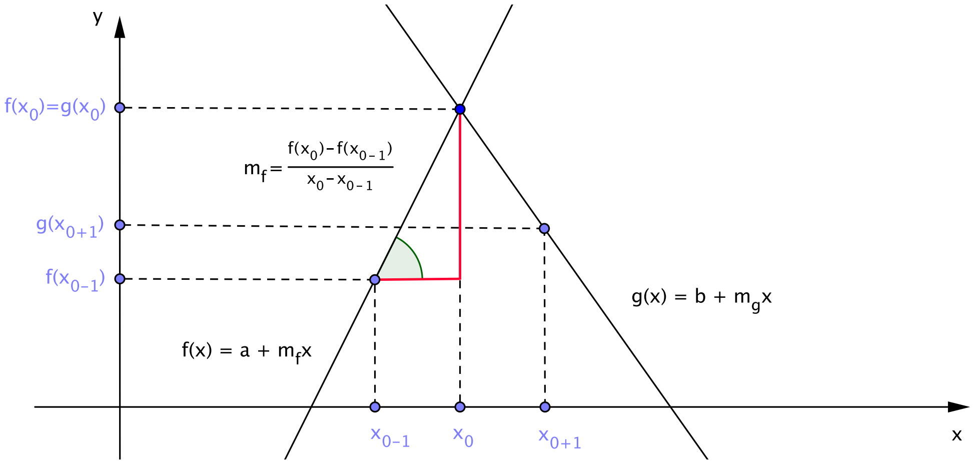

Figure 2.4: Example of a local peak

Figure 2.4 illustrates a local peak in a given time series T.

3 Peak-Valley Algorithm

3.1 Identification of peaks and valleys

The Peak-Valley algorithm uses a geometrical approach to find local peaks and local valleys in a time series. The algorithm detects all local peaks and valleys in a time series T.

Given a time series T with n observations. We take the definition we have introduced at the beginning (2.1).

$$T = \{(t_1, x_1), (t_2, x_2), \ldots, (t_n, x_n)\}$$

A peak is defined as following:

$$x_{i-1} < x_i > x_{i+1} \text{ , } \forall i = 2, 3, \ldots, n-1 \tag{3.1}$$

The first and last observation in T have to be examined special. The first and last observation in T are peaks if

$$x_1 > x_2 \tag{3.2}$$

$$x_n > x_{n-1} \tag{3.3}$$

Let's define the set of peaks P with

$$P = \{(t_i, x_i) | (x_{i-1} < x_i > x_{i+1}) \vee (x_1 > x_2) \vee (x_n > x_{n-1})\} \text{ , } \forall i = 2, \ldots, n-1$$

In contrary we define the set of valleys V with

$$V = \{(t_i, x_i) | (x_{i-1} > x_i < x_{i+1}) \vee (x_1 < x_2) \vee (x_n < x_{n-1})\} \text{ , } \forall i = 2, \ldots, n-1$$

A peak cannot be a valley and vice versa. All other points that are not a local peak or a local valley will be ignored in the algorithm. Therefore, it has to be obtained

$$P \cap V = \varnothing$$

The algorithm contains our definitions of local peak and local valley from chapter (2).

$$\text{Peak} := x_{i-1} < x_i > x_{i+1} \text{ , Valley} := x_{i-1} > x_i < x_{i+1}$$

By controlling the definition (3.1), we state:

The algorithm can not detect local peaks or local valleys if on the left side of point x_i or on the right side of point x_i resides a horizontal straight line. In such a case, we detect another number of peaks and valleys. If we want to detect all peaks and valleys in time series T, than we have first to extract all horizontal lines from the curve before we detect local peaks and local valleys. In other words, we have to eliminate lines with equal starting and ending points.

Figure 3.1 shows the two different results of peak and valley detection with and without horizontal lines.

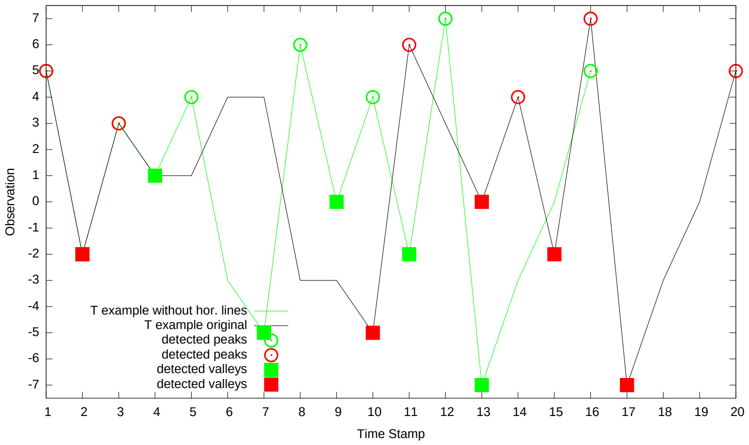

Figure 3.1: Peaks and valleys detection with described algorithm

The black curve shows time series T with horizontal lines. The green curve shows T with extracted horizontal lines. In the green curve there are detected more peaks and valleys.

4 Significant Peak-Valley Algorithm

4.1 Introduction

Significance denotes a concept in statistical methods. A result is called statistically significant if it is unlikely to have occurred by chance. The *Significant Peak-Valley* algorithm bases on the statistical approach. With significant peaks or valleys we denote local peaks and local valleys which are significant in a global sense. The construct significant peak and significant valley is expanded than the definitions of global peaks and global valleys in (2.2) and (2.3).

4.2 Detection of significant peaks

Let's take a time series T with n observations. We assume there exist a peak function and a valley function which find local peaks and local valleys. We will describe such a peak and a valley function in the next section.

The peak and the valley functions produce values $x_i \in \mathbb{R}$ with $i = 1, \dots, n \in \mathbb{N}$.

A peak function S produces for a local peak a positive value. We can define the set of ℓ local peaks as P with

$$P := \{(t_i, x_i) | S(x_i) > 0\} \text{ with } \; i = 1, \dots, \ell$$

The valley function is vice versa to the peak function. ℓ' local valleys in V:

$$V := \{(t_i, x_i) | S(x_i) \leq 0\} \text{ with } \; i = 1, \dots, \ell'$$

Statistically, the values of all elements in P and V build two univariate distributions. If we compute the **arithmetical mean value** of these distributions $\bar{x}$, we divide the values of P or V in two parts.

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) \tag{4.1}$$

There are values which are greater or smaller than the mean value $\bar{x}$. A peak $x_i > \bar{x}$ is a better candidate to be a significant peak than $x_i < \bar{x}$. We can divide the result further to be more precise. At first we define the **variance** v which describes the distances of x_i to the mean value $\bar{x}$.

$$v = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{4.2}$$

In the next step, we define the **standard deviation** s. Which is the square root of variance v.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{v} \tag{4.3}$$

The standard deviation is a measure of all values arranged around the mean value of the distribution.

We can define significant peaks P' and valleys V' as following:

$$\begin{aligned} P' := P' \subset P \text{ with } P' &:= \{(t_i, x_i) | S(x_i) > 0 \land S(x_i) > (m' + h \cdot s')\} \\ &\forall i = 1, \ldots, \ell \text{ where } \ell = \text{number of observations in } P \\ V' := V' \subset V \text{ with } V' &:= \{(t_i, x_i) | S(x_i) \leq 0 \land S(x_i) \leq (m' + h \cdot s')\} \\ &\forall i = 1, \ldots, \ell' \text{ where } \ell' = \text{number of observations in } V \end{aligned}$$

with m' = mean value of all $S(x_i) > 0$ and s' = standard deviation of all $S(x_i) > 0$ for peaks and vice versa for valleys. h is an user defined parameter with $1 < h \leq 3 \in \mathbb{R}$.

The user defined parameter h defines the significance of a peak and a valley detection.

In set P' and V' we want retain only one peak and valley within distance k inside the given sequence $\{(t_{i-k}, x_{i-k}), \ldots, (t_i, x_i), \ldots, (t_{i+k}, x_{i+k})\}$. For every adjacent pair of peaks in P' and valleys in V' with index $|j - i| \leq k$ we remove the observation with the smaller value for peaks and greater value for valleys of $\{(t_i, x_i), (t_j, x_j)\}$ from P' and V'.

4.3 Peak functions S_1 to S_3

The first three peak function in the paper [4] to detect peaks are very similar. They are defines as following:

$$S_1(k, i, T) = \frac{\max\{x_i - x_{i-1}, \ldots, x_i - x_{i-k}\} + \max\{x_i - x_{i+1}, \ldots, x_i - x_{i+k}\}}{2} \tag{4.4}$$

$$S_2(k, i, T) = \frac{\frac{(x_i - x_{i-1} + \ldots + x_i - x_{i-k})}{k} + \frac{(x_i - x_{i+1} + \ldots + x_i - x_{i+k})}{k}}{2} \tag{4.5}$$

$$S_3(k, i, T) = \frac{\left(x_i - \frac{(x_{i-1} + \ldots + x_{i-k})}{k}\right) + \left(x_i - \frac{(x_{i+1} + \ldots + x_{i+k})}{k}\right)}{2} \tag{4.6}$$

k gives the size of subsequence from T with $(2 \cdot k + 1)$
i index which denotes the i^{th} observation in T
T time series with n observations

In the case where $k = 1$ then the 3 peak functions are equal for $\forall\,(t_i, x_i) \in T$:

$$\begin{aligned} S_1(1,i,T) &= \frac{\max\{x_i - x_{i-1}\} + \max\{x_i - x_{i+1}\}}{2} = \frac{\dfrac{(x_i - x_{i-1})}{1} + \dfrac{(x_i - x_{i+1})}{1}}{2} \\ &= S_2(1,i,T) = \frac{\left(x_i - \dfrac{x_{i-1}}{1}\right) + \left(x_i - \dfrac{x_{i+1}}{1}\right)}{2} = S_3(1,i,T) \end{aligned}$$

Based on the initialization $k = 1$, we simplify the previous peak functions definitions as follows:

$$S_1(1,i,T) = S_2(1,i,T) = S_3(1,i,T) = \frac{x_i - x_{i-1} + x_i - x_{i+1}}{2} = x_i - \left(\frac{x_{i-1} + x_{i+1}}{2}\right)$$

The values of the peak functions S_1 bis S_3 can be updated as follows:

$$\begin{aligned} S_1(1,i,T) = S_2(1,i,T) = S_3(1,i,T) > 0 \quad &\Leftrightarrow \quad x_i > \frac{x_{i-1} + x_{i+1}}{2} \\ S_1(1,i,T) = S_2(1,i,T) = S_3(1,i,T) \leq 0 \quad &\Leftrightarrow \quad x_i \leq \frac{x_{i-1} + x_{i+1}}{2} \end{aligned}$$

x_i is a peak, if the functions S_1, S_2 or S_3 produce a positive value. Figure 4.1 shows a peak in t_i with observation x_i

In the example, x_i is exactly a local peak if the following condition is satisfied

$$x_i > \max\{x_{i-1}, x_{i+1}\}$$

If we take the case:

$$\frac{x_{i-1} + x_{i+1}}{2} < x_i \leq \max\{x_{i-1}, x_{i+1}\}$$

The shown situation (marked red) in figure 4.1 injures in some cases the peak definition. We will demonstrate this case explicit.

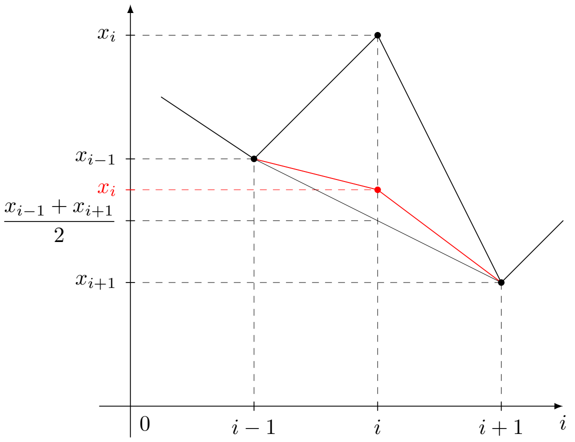

Figure 4.1: Example of a probably detected peak x_i

Figure 4.1 shows the geometrical interpretation of the peak function. $S_1(x_i)$ is the *signed distance* from the center of the secant $((i-1, x_{i-1}), (i+1, x_{i+1}))$ to the point (i, x_i).

We have shown that the functions S_1 to S_3 can be used to detect possible candidates for peaks and valleys. But the detection of peaks or valleys are unique in the following cases, only:

$$\begin{aligned} x_i > \max\{x_{i-1}, x_{i+1}\} \quad &\Leftrightarrow \quad x_i \text{ is a local peak} \\ x_i < \min\{x_{i-1}, x_{i+1}\} \quad &\Leftrightarrow \quad x_i \text{ is a local valley} \end{aligned}$$

In the case of $\frac{x_{i-1} + x_{i+1}}{2} < x_i \leq \max\{x_{i-1}, x_{i+1}\}$ there exists the possibility that the peak function accepts peaks which are not peaks (*false positives*). For example, we can show by setting $k = 1$, already.

We apply the peak function S_1 for every k where $k < i < n - k$. The same computations can be done for the other peak functions S_2 and S_3. We apply the maxima of the differences on the left and right side of x_i

$$\begin{aligned} M_{left} &:= \max\{x_i - x_{i-1}, \dots, x_i - x_{i-k}\} \\ M_{right} &:= \max\{x_i - x_{i+1}, \dots, x_i - x_{i+k}\} \end{aligned}$$

where M_{left} is the maximum of the differences with x_i and his left k neighbour and M_{right} is the same for the right side.

Therefore, there exists at least for each maxima a value with $x_{M_{left}} \in \{x_{i-1}, \dots, x_{i-k}\}$ and $x_{M_{right}} \in \{x_{i+1}, \dots, x_{i+k}\}$ with $M_{left} = x_i - x_{M_{left}}$ and $M_{right} = x_i - x_{M_{right}}$. Then

$$\begin{aligned} S_1(k, i, T) &= \frac{M_{left} + M_{right}}{2} = \frac{x_i - x_{M_{left}} + x_i - x_{M_{right}}}{2} \\ &= \frac{2x_i - x_{M_{left}} - x_{M_{right}}}{2} = x_i - \frac{x_{M_{left}} + x_{M_{right}}}{2} \end{aligned}$$

S_1 produces positive values only if x_i is greater then the arithmetic mean of $x_{M_{left}}$ and $x_{M_{right}}$. Additionally, it is essential for $x_{M_{left}}$ and $x_{M_{right}}$

$$x_i - x_{M_{left}} \geq x_i - x_{i-j} \,\forall\, j \in \{1, \dots, k\} \Rightarrow x_{M_{left}} \leq x_{i-j} \,\forall\, j \in \{1, \dots, k\}$$

$\Rightarrow x_{M_{left}} = \min\{x_{i-1}, \dots, x_{i-k}\}$ same for the right seide: $x_{M_{right}} = \min\{x_{i+1}, \dots, x_{i+k}\}$

If the application of peak function returns a positive result, x_i is a peak in the local sense. A negative result means that x_i is a local valley. If the result $= 0$, x_i is neither a peak nor a valley. Then x_i is locate on the secant between the two points.

$f(x_i)$ is a local peak and can be considered as a global peak if $f(x_i)$ is located above the secant which connects the points that constructs the maximum difference. In the inverse case, a local (or global) valley is detected.

If the signed distance is 0, then x_i is neither a peak nor a valley. In the case where $k = 1$, then the peak function S_1 works same as the peak and valley detection algorithm in section 2.1. We also need 3 points and then decide if x_i is greater then x_{i-1} and x_{i+1} or smaller or neither.

The application of S_1 on our example data gives the following result.

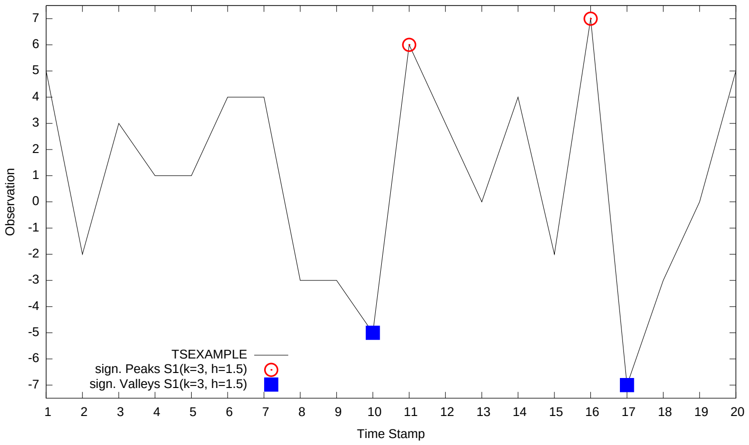

Figure 4.2: Detection of significant peaks and valleys with S_1

With $k = 1$, the peak functions S_1, S_2 and S_3 are all the same. With $k > 1$, S_1 produces other values than S_2 and S_3.

4.4 Peak function S_4

The peak function S_4 uses the principle of entropy. Entropy is a measure for information in a given sequence A.

In the information theory the entropy of a sequence is defined after Shannon:

$$H(A) = -\sum_{i=1}^{M}(p(a_i)\log_2(p(a_i))) \tag{4.7}$$

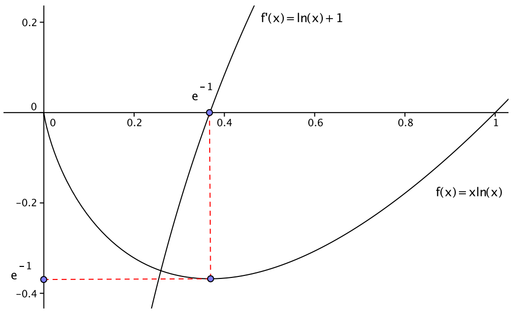

Figure 4.3: Graph of the entropy $p \cdot \log(p)$

In other words, the entropy of a given sequence is the measurement of disorder in this sequence. The calculation of entropy bases on the probability of the appearance of the values inside the given sequence. Therefore we have to compute the probability of the values inside the sequence of a given time series.

To compute the probability of the values inside the sequence we choose the kernel density technique after *E. Parzen* (also called "parzen window") [5]. With this technique we can compute the probability of the sequence with different kernels. An estimate of $f(x)$ can be given by:

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)$$

where h is an adapted positive number. Generally, h is a function of the number of elements inside the sequence. The kernel function $K(x)$ can be replaced by other kernel functions. The most used kernel functions are the *Epanechnikov* kernel function and the *Gaussian* kernel function.

The **Epanechnikov** kernel

$$K(x) = \begin{cases} \frac{3}{4}(1 - x^2) & \text{if } |x| < 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.8}$$

The **Gaussian** kernel

$$K(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}x^2} \tag{4.9}$$

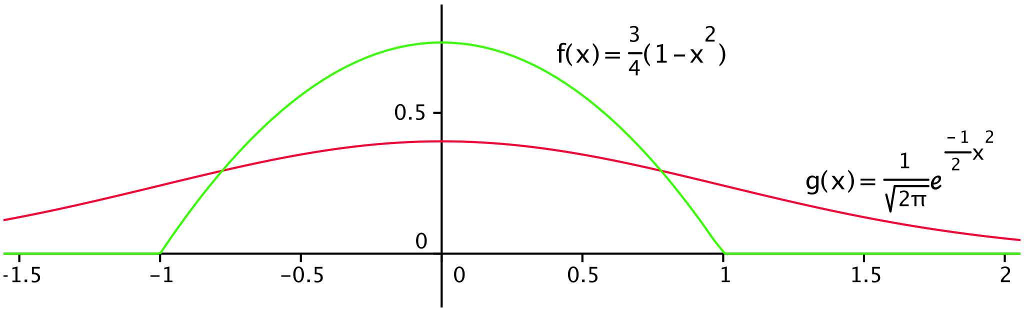

Figure 4.4: The two kernel functions

With this kernel technique we can estimate the probability of the occurrence of values a_i inside the sequence A.

In the S_4 peak function the estimation of the probability density at value a_i in a given sequence is defined as

$$p_w(a_i) = \frac{1}{M|a_i - a_{i+w}|}\sum_{j=1}^{M} K\left(\frac{a_i - a_j}{|a_i - a_{i+w}|}\right) \tag{4.10}$$

where M is the number of elements in the sequence and $|a_i - a_{i+w}|$ is the width of the parzen window. We define the width of the parzen window as follows:

$$|a_i - a_{i+w}| := \sqrt{(a_i - a_{i+w})^2 + w^2} \tag{4.11}$$

$|a_i - a_{i+w}|$ must never be 0. With the definition above, the term will be 1 at least.

The $H_w(A)$ and $p_w(a_i)$ indicate the width parameter used in the parzen window function. After updating the kernel density estimation $p_w(a_i)$, the entropy of the sequence is obtained as follows:

$$H_w(A) = -\sum_{i=1}^{M}(p_w(a_i)\log_2(p_w(a_i))) \tag{4.12}$$

where M is the number of elements in the sequence.

The last specification covers the case $p_w(a_i) = 0$. Then

$$\lim_{p \to 0} p \log_2 p = 0$$

because the $\log_2(0)$ is not defined.

In our implementation of the peak function S_4 we use the Gaussian kernel.

The principle of peak function S_4 is the difference in entropy from 2 sequences. We define the peak function S_4:

$$S_4(k, w, i, T) = H_w(N(k, i, T)) - H_w(N'(k, i, T)) \tag{4.13}$$

The difference in entropy gives a view how significant the point x_i is in the given sequences. If the difference is greater then 0, the peak function value of x_i is a candidate for a local maximum, it can be a global maximum too, and therefore it can be a significant peak.

We define the two sequences $N(k, i, T)$ and $N'(k, i, T)$ as follows:

$$N^-(k, i, T) = \{x_{i-1}, \ldots, x_{i-k}\}; N^+(k, i, T) = \{x_{i+1}, \ldots, x_{i+k}\}$$

These are the k left and the k right temporal neighbours of x_i in a subsequence of time series T.

$$N(k, i, T) = N^-(k, i, T) \cup N^+(k, i, T) \tag{4.14}$$

This is a subsequence with $(2 \cdot k)$ elements of time series T (*without* x_i)

$$N'(k, i, T) = N^-(k, i, T) \cup \{x_i\} \cup N^+(k, i, T) \tag{4.15}$$

This is a subsequence with $(2 \cdot k + 1)$ elements of time series T (*with* x_i)

If we compute the entropy of $H_w(N)$ and $H_w(N')$ the result is always > 0. Further, if we compute $H_w(N)$ and $H_w(N')$ with the same window width for the parzen window, it obtains always $H_w(N) < H_w(N')$. Therefore, $S_4 < 0$ obtains always. To get a $S_4 > 0$, we have to define for sequence $N(k, i, T)$ another window width for the parzen window.

$$|a_{i-1} - a_{(i-1)+w}| := \sqrt{(a_{i-1} - a_{(i-1)+w})^2 + w^2} \tag{4.16}$$

If we apply the S_4 peak function to time series 137 of hydrological measurement, we get the following result:

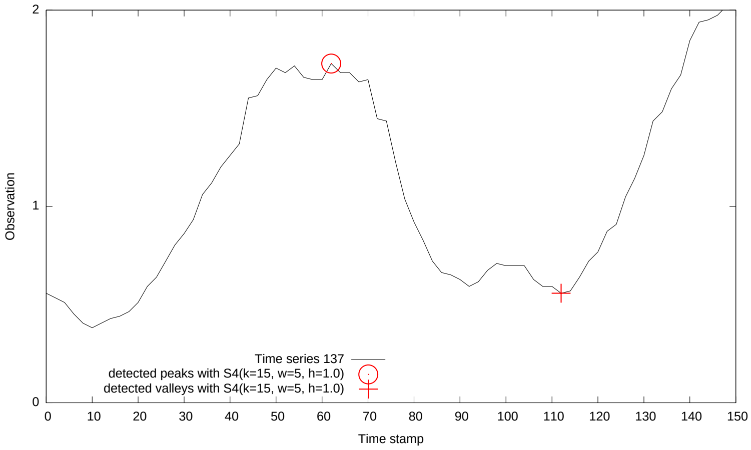

Figure 4.5: Detected significant peak and significant valley with S4 peak function

Testing the implementation, the choice of the user defined parameters k and w is very important for the outcome. The best results we get as follows:
k > 5, k has to be odd
w >3

5 Conclusion

At the beginning of this paper we introduced the mathematical background of peaks and valleys in continuous functions $f(x)$. If $f(x)$ is differentiable we can compute the local extrema. To detect such extrema in a discrete sequence like a time series we have to compute the peak function in every point of the time series.

We implemented the *Peak-Valley* algorithm described in chapter 3 and the *Significant Peak-Valley* algorithm described in chapter 4. Each algorithm has its own principle and its own problems. None of the two algorithms can be applied without preparation in the time series or dedicated choice in the user defines parameters.

5.1 Remarks to the Peak-Valley algorithm

- detects all peaks and valleys if the time series doesn't contain horizontal lines
- if the time series contains horizontal lines, the algorithm doesn't detect peaks and valleys which are at the beginning or ending of a horizontal line
- if we apply the algorithm on a time series with extracted horizontal lines, we don't get all peaks and valleys. In detection we loss the peaks and valleys at the ending of a horizontal line

5.2 Remarks to the Significant Peak-Valley algorithm

- under certain circumstances, the peak functions could produce *false positives*
- the choice of the user defined parameters k and w is very important

Bibliography

[1] C. Blatter. *Analysis 1*. Number v. 1 in Springer-Lehrbuch. Springer, 1991.

[2] Eamonn J. Keogh, Selina Chu, David Hart, and Michael J. Pazzani. An online algorithm for segmenting time series. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, ICDM '01, pages 289–296, Washington, DC, USA, 2001. IEEE Computer Society.

[3] Z. M. Nopiah, M. I. Khairir, S. Abdullah, and C. K. E. Nizwan. Peak-valley segmentation algorithm for fatigue time series data. *WSEAS Trans. Math.*, 7:698–707, December 2008.

[4] Girish K. Palshikar. Simple Algorithms for Peak Detection in Time-Series. In *Proc. 1st Int. Conf. Advanced Data Analysis, Business Analytics and Intelligence*, 2009.

[5] E. Parzen. On the estimation of probability density function and the mode. *The Annals of Math. Statistics,*, vol. 33:pp. 1065–1076, 1962.

[6] Luciana A. S. Romani, Ana Maria Heuminski de Ávila, Jurandir Zullo Jr., Caetano Traina Jr., and Agma J. M. Traina. Mining relevant and extreme patterns on climate time series with clipsminer. *Journal of Information and Data Management (JIDM)*, 1(2):245–260, 06/2010 2010.