Broken line smoothing: A simple method for interpolating and smoothing data series

Demetris Koutsoyiannis

Department of Water Resources, Faculty of Civil Engineering, National Technical University, Athens

Abstract. A technique is proposed for smoothing a broken line fit, with known break points, to observational data. It will be referred to as "broken line smoothing". The smoothness term is defined by means of the angles formed by the consecutive segments of the broken line, and is given an adjustable weight. The roughness of the resulting broken line can then be controlled by appropriately tuning the weight of smoothness term and the number of straight-line segments. The broken line smoothing can be used for data analysis in several applications as an alternative to other methods such as locally weighted regression and smoothing splines. The mathematical background and details of the method as well as practical aspects of its application are presented and discussed. Also, several examples using both synthesised and real world (hydrological and climatological) data are presented to explore and illustrate the methodology.

Keywords: Curve fitting; data analysis; regression; interpolation; smoothing; scatterplots; piecewise linear regression; smoothing splines.

Software availability	
Program title:	**BLSmooth** – Broken line smoothing
Contact Address:	Demetris Koutsoyiannis, Department of Water Resources, Faculty of Civil Engineering, National Technical University, Heroon Polytechniou 5, GR-157 80 Zografou, Greece; e-mail dk@hydro.ntua.gr; URL: http://www.hydro.ntua.gr/faculty/dk
First available:	1994 (version 0); 1998 (version 1)
Program language:	Visual Basic for Applications (version 1), ready to use as an MS-Excel add-in (including the examples contained in this paper).
Cost:	Free

1. Introduction

The fitting of a curve to a set of paired measurements (x_i, y_i) is one of the most common problems in environmental science and engineering, as well as in other scientific fields involving data analysis. A fitted curve is a mathematical formulation of the dependence of the variable y on x, and it is utilised for many practical purposes such as interpolation between measurements, prediction, filling in missing values in time series, estimation and removal of the measurement errors, etc. In the case that the mathematical expression of the dependence of y on x is of an a-priori known type (e.g., linear, logarithmic, power, polynomial, etc.) the problem of curve fitting is simplified as the only requirement is the determination of the parameters of this expression, a task typically accomplished using regression (or least squares) techniques. The difficulty arises when such an expression is not known and cannot be approximated by a simple easily recognisable law.

Traditionally, the latter case has been remedied by graphical techniques such as drawing an "eyeball" curve on a scatterplot of points (x_i, y_i). This is apparently a non-parametric approach in the sense that it does not use any parameters of a specified law (in contrast to parametric regression techniques) but has the flaws of being subjective and unsusceptible to an algorithmic treatment, and thus it is not programmable to computers.

A modern alternative is the use of smoothing techniques in which the fitted value of y for any value of x is determined from the available data points (x_i, y_i) using weights to each one so that the weight for (x_i, y_i) is large if x is close to x_i and small otherwise. Among the smoothing techniques the one mostly widespread in environmental computing is the so-called *lowess* – locally weighted scatterplot smoothing (Cleveland, 1979; Cleveland and McGill, 1984; see also Hirsch *et al.*, 1993, pp. 17.45-47). According to this technique, at a first phase, known as locally weighted regression, the fitted value at x_k is the value of the polynomial fit to the data using weighted least squares with weights given for each x_i as a decreasing function of the distance $|x_i - x_k|$. In subsequent phases this procedure is repeated but with the weights being also dependent on the size of the estimation residuals (or errors) so that residual closer to zero are assigned a larger weight. In this manner, the outliers are down-weighted and the method acquires its robustness.

Another alternative is smoothing by spline functions (Reinsch, 1967, 1971; Wahba, 1990). A spline is a series of piecewise (typically cubic) polynomial functions that are joined together at each data point abscissa x_i in a smooth fashion (i.e., assuring continuity of the function as well as its first and second derivatives). The most well known spline is the so called cubic interpolating spline which passes through all data points (x_i, y_i) (thus leading to zero estimation errors) and therefore is appropriate for interpolation in between consecutive such points (e.g., de Boor, 1978; Bartels et al., 1987). The smoothing spline differs from the interpolating spline in that it does not pass through the data points, thus avoiding a very rough or wavy shape in case data points are subject to random noise. To acquire its smoothness the total curvature (or roughness) of the spline is considered and minimised under the constraint that the sum of the estimation square errors at the data points (x_i, y_i) (i.e., the amount of noise) does not exceed a specified value (Reinsch, 1967). With subsequent developments (Craven and Wahba, 1979; Hutchinson and de Hoog, 1985) the amount of noise need not to be specified in advance, but it can be assessed from the data in an objective manner.

The broken line smoothing (BLS) presented in this paper may be considered as a simple alternative to numerical smoothing and interpolating methods yet being close to the approach of the traditional graphical method. The method is also closely related to piecewise linear regression and to smoothing splines. The idea is to approximate a smooth curve that may be drawn for the data points (x_i, y_i) with a broken line (or open polygon) which can be numerically estimated by means of a least squares fitting procedure. The abscissae of the vertices of the broken line do not necessarily coincide with x_i's but they can form a series of points with some chosen, lower or higher, resolution. Although the broken line is a concatenation of straight-line segments, and therefore is not smooth in strict sense, it can be assigned a measure of smoothness (conversely, roughness) based on the angles formed by the consecutive segments. Thus, the smoothness is high (and the roughness low) if the angles are close to π radians.

If the only objective used for fitting the broken line is the minimisation of total square error then the result might be a very rough broken line, depending on the arrangement of data points (x_i, y_i). However, the roughness of the broken line can be controlled by introducing as a second objective the minimisation of the roughness. There is a trade-off between the two

objectives of minimising the fitting error and the roughness of the broken line. The larger the relative weight of the second objectives is, the smoother the broken line resulting by the fitting procedure will be. As the relative weight of the second term tends to infinity, the broken line will tend to the least-squares straight line.

This introductory description of the broken line smoothing gives a rough idea of its essential features and behaviour. Precise mathematical details are given in section 2. Section 3 is devoted to the choice of the method parameters, whereas the relation of broken line smoothing to other similar models is described in section 4. Examples of the method application are given in section 5 and conclusions are drawn in section 6. Some additional mathematical material is contained in an appendix.

2. Mathematical framework

Let (x_i, y_i) be a set of n points at the $x\,y$ plane for $i = 1, \ldots, n$. Let c_j, $j = 0, \ldots, m$, be $m + 1$ points at the x-axis so that the interval $[c_0, c_m]$ contain all x_i. For simplicity we will assume that the points are equidistant, i.e., $c_j - c_{j-1} = \delta$; the generalisation for any arrangement of points is direct. We wish to find the $m + 1$ values d_j in the y-axis so that the broken line defined by the $m + 1$ points (c_j, d_j) "fit" the set of points (x_i, y_i) (see Figure 1). This fitting is meant in terms of minimising the total square error among the set of original points (x_i, y_i) and the fitted broken line (c_j, d_j), i.e.,

$$p = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{1}$$

where $\hat{y}_i$ is the estimate of y_i given by the broken line for the known x_i. If x_i lies in the subinterval $[c_{j-1}, c_j]$ for some j $(1 \le j \le m)$ then obviously $\hat{y}_i$ is given by

$$\hat{y}_i = d_j \frac{x_i - c_{j-1}}{c_j - c_{j-1}} + d_{j-1} \frac{c_j - x_i}{c_j - c_{j-1}} = \frac{1}{\delta} [d_j (x_i - c_{j-1}) + d_{j-1} (c_j - x_i)] \tag{2}$$

The above equations can be more concisely be written in the form

$$p = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 \tag{3}$$

with

$$\hat{\mathbf{y}} = \mathbf{\Pi}\,\mathbf{d} \tag{4}$$

where $\mathbf{y} = [y_0, \ldots, y_n]^T$ is the vector of known ordinates of the given data points with size n (the exponent T denotes the transpose of a matrix or vector), $\hat{\mathbf{y}} = [\hat{y}_0, \ldots, \hat{y}_n]^T$ is the vector of estimates with size n, $\mathbf{d} = [d_0, \ldots, d_m]^T$ is the vector of the unknown ordinates with size $m + 1$, and $\mathbf{\Pi}$ is a matrix with size $n \times (m + 1)$ and ijth entry (for $i = 1, \ldots, n$; $j = 1, \ldots, m + 1$)

$$\pi_{ij} = \begin{cases} \dfrac{x_i - c_{j-2}}{\delta}, & c_{j-2} < x_i \le c_{j-1} \\ \dfrac{c_j - x_i}{\delta}, & c_{j-1} < x_i \le c_j \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

If $m = 1$ the broken line becomes a straight line and the problem of minimising p in (1) reduces to fitting the least-squares straight line. The problem is more interesting when $m > 1$. If the number of the broken line segments m is sufficiently less than the number of given data points n then the minimisation of (1) will lead to the best fit broken line which "summarises" the n points data set (x_i, y_i) using the $m + 1$ points (c_j, d_j). Moreover, if m is sufficiently greater than n then the broken line "expands" the available data set thus providing a means for a "detailed" interpolation through its points. However, in the last case, (1) is not sufficient to determine a unique set of values d_j as many combinations of d_j may lead to p equal to zero.

Now we add another requirement, aiming at avoiding a very rough broken line and also assuring a unique solution of the fitting problem. To acquire a measure of the roughness of the broken line we observe that the difference of slopes between two consecutive segments of the broken line (given that c_j are equidistant), will be

$$\frac{1}{\delta}(2\,d_j - d_{j-1} - d_{j+1}) \tag{6}$$

so that the following expression can be an appropriate measure for the roughness of the entire broken line

$$q = \sum_{j=1}^{m-1} (2\, d_j - d_{j-1} - d_{j+1})^2 \tag{7}$$

This can be written in matrix form as

$$q = \mathbf{d}^T \boldsymbol{\Psi}^T \boldsymbol{\Psi}\, \mathbf{d} \tag{8}$$

where $\boldsymbol{\Psi}$ is the $(m-1) \times (m+1)$ matrix whose ijth entry is

$$\psi_{ij} = \begin{cases} 2, & j = i + 1 \\ -1, & |j - i - 1| = 1 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

(Apparently, $\boldsymbol{\Psi}^T \boldsymbol{\Psi} = \mathbf{O}$ for the special case $m = 1$).

Combining (3) and (8) and introducing a dimensionless multiplier $\lambda \geq 0$ for q we get the more generalised objective function to be minimised

$$f(\mathbf{d}) := p + \lambda\, q = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \lambda\, \mathbf{d}^T \boldsymbol{\Psi}^T \boldsymbol{\Psi}\, \mathbf{d} \tag{10}$$

The multiplier λ controls the smoothness of the broken line. Indeed, if $\lambda = 0$ then (10) will lead to the broken line with the minimum square error $p_{\min}$. If $\lambda > 0$, then the broken line will be smoother, due to the term $\lambda\, q$ of the objective function; apparently, in that case the square error p will be greater than $p_{\min}$. And as $\lambda \to \infty$, the broken line will tend to the least squares straight line regardless of the number of segments m, as if it were $m = 1$ (see Figure 2 and subsequent figures); in that case we will have the maximum value of the square error p. Details about the choice of an appropriate value of λ are discussed in section 3.

Equation (10) can be written in the form

$$f(\mathbf{d}) = (\mathbf{y} - \boldsymbol{\Pi}\, \mathbf{d})^T (\mathbf{y} - \boldsymbol{\Pi}\, \mathbf{d}) + \lambda\, \mathbf{d}^T \boldsymbol{\Psi}^T \boldsymbol{\Psi}\, \mathbf{d} \tag{11}$$

The vector $\mathbf{d}$ that minimises $f(\mathbf{d})$ satisfies

$$\frac{df}{d\mathbf{d}} = 2\, \mathbf{d}^T \boldsymbol{\Pi}^T \boldsymbol{\Pi} + 2\, \lambda\, \mathbf{d}^T \boldsymbol{\Psi}^T \boldsymbol{\Psi} - 2\, \mathbf{y}^T \boldsymbol{\Pi} = \mathbf{O} \tag{12}$$

To derive (12) from (11) we have applied the typical rules of derivatives involving matrices (e.g., Marlow, 1993, p. 214). The solution of (12) for **d** is

$$\mathbf{d} = (\mathbf{\Pi}^T \mathbf{\Pi} + \lambda \mathbf{\Psi}^T \mathbf{\Psi})^{-1} \mathbf{\Pi}^T \mathbf{y} \tag{13}$$

We observe that both $\mathbf{B} := \mathbf{\Pi}^T \mathbf{\Pi}$ and $\mathbf{C} := \mathbf{\Psi}^T \mathbf{\Psi}$ are square matrices with size $(m + 1) \times (m + 1)$; the former is tridiagonal and the latter five-banded. **B** can be singular (not invertible) if one or more columns of **Π** have zero elements, that is, if at least two consecutive intervals $(c_{j-1}, c_j]$ contain no x_i's. **C** is always singular. However, for $\lambda > 0$, the sum $\mathbf{B} + \lambda \mathbf{C}$ is nonsingular and, thus, its inverse exists. In the Appendix we give detailed expressions for the items of the matrices **A** and **B** without reference to the original matrices **Π** and **Ψ**. These expressions are useful in the typical case where $m << n$ because they avoid the use of **Π** and **Ψ** whose memory requirements are much larger than those of **A** and **B**.

3. Choice of parameters

It is clear from the previous section that the model has two adjustable parameters: the number of intervals, *m*, and the smoothing parameter λ. If *m* is small, then it acts as a smoothing parameter, as well. The choice of parameters can be done by assessing the amount of data smoothing either graphically, or by using standard objective ways.

In the first case, the user utilises his or her experience about the examined problem, performing exploration fittings for different values of *m* and λ and assessing them graphically. The computational framework of the method implementation (MS-Excel, Excel Visual Basic) provides a direct means for data visualisation and graphical exploration. For a more convenient search of λ we provide a transformation of λ in terms of another number τ, whose values are restricted in the interval [0, 1). This transformation was established after a numerical investigation of the method on several examples and has the form

$$\lambda = \left(10\, m \frac{\ln \tau_{\mathrm{m}}}{\ln \tau}\right)^{\kappa} \tag{14}$$

where $\tau_{\mathrm{m}} = 0.99$ is the maximum allowed value of τ corresponding to an upper bound of λ, set for numerical stability equal to

$$\lambda_{\mathrm{m}} = \frac{\mathrm{trace}(\mathbf{B})}{\mathrm{trace}(\mathbf{C})}\, 10^{8} \tag{15}$$

The constant κ in (14) is determined by

$$\kappa = \frac{\ln \lambda_{\mathrm{m}}}{\ln (10\, m)} \tag{16}$$

which is a consequence of (14) and (15). The minimum allowed value of λ is 0 if the inverse of matrix **B** exists; otherwise it is estimated from (14) using a small value of $\tau = 1 - \tau_{\mathrm{m}} = 0.01$.

In the second case, selection of parameters is done in an objective manner. Combining (4) and (13) we obtain

$$\hat{\mathbf{y}} = \mathbf{A}\, \mathbf{y} \tag{17}$$

where **A** is a $n \times n$ symmetric matrix given by

$$\mathbf{A} = \mathbf{\Pi}\, (\mathbf{\Pi}^{T}\, \mathbf{\Pi} + \lambda\, \mathbf{\Psi}^{T}\, \mathbf{\Psi})^{-1}\, \mathbf{\Pi}^{T} \tag{18}$$

and depending on both adjustable parameters m and λ (or τ). In such a situation, parameter estimation can be done by minimising the so-called generalised cross-validation (GCV; Craven and Wahba, 1979), defined by

$$V = \frac{\frac{1}{n}\, \|(\mathbf{I} - \mathbf{A})\, \mathbf{y}\|^{2}}{\left[\frac{1}{n}\, \mathrm{trace}(\mathbf{I} - \mathbf{A})\right]^{2}} \tag{19}$$

For a given number of segments m the minimisation of V results in the optimum value τ. This can be repeated for several trial values of m until the global minimum V is reached. However, normally the mean square estimation error (the numerator in (19)), as well as the value of GCV decrease with the increase of m. As it will be demonstrated in the next section, the decrease becomes insignificant for moderate and large values of m, and this may help for the appropriate choice of m. Also, it is demonstrated that the objective method of parameter estimation may be not adequate in certain real world problems, and therefore, an exploratory

graphical assessment of the smoothing, combined with information relevant to the problem context, may be helpful.

4. Relation of broken line smoothing to other similar models

The broken line smoothing, with the mathematical formulation described in the previous sections, is closely related to other curve fitting models and, specifically, to broken lines (or piecewise linear regression models) and to cubic smoothing splines.

The main difference of the broken line smoothing to piecewise linear regression models with known break points (e.g., Ertel and Fowlkes, 1976) is the introduction of the smoothness term $\mathbf{\Psi}^T \mathbf{\Psi}$ in the problem formulation. As a direct consequence, the matrix of linear equation coefficients, which in piecewise linear regression models is tridiagonal, in (13) becomes five-banded diagonal. Piecewise linear regression models have also been studied for the case where the positions of the break points are not specified in advance (e.g., Bellman and Roth, 1969; Gallant and Fuller, 1973; Lerman, 1980). This problem which is highly nonlinear has not been considered in the above formulation of broken line smoothing.

On the other hand, the mathematical form of the broken line smoothing is quite similar to that of the cubic smoothing splines (Reinsch, 1967, 1971; Wahba, 1990). Both methods result in about the same final equation (13) and have the same computational cost as both involve five-banded linear systems to be solved. Apparently, however, the items of the involved matrices are determined in different ways, due to different considerations in the two methods. Specifically, the curvature of the broken line smoothing is considered in terms of the angles formed by the segments of the broken line, whereas in cubic smoothing splines it is considered in terms of the second derivative of the spline. Generally, the linear attitude of segments in the broken line smoothing renders the method simpler and more parsimonious with regard to the number of parameters determining each segment.

Due to similarity of equations of the broken line smoothing (particularly equations (13) and (17)) with those of the smoothing spline, several advanced numerical techniques devised for splines can be used in broken line smoothing as well. This is the case with GCV that was devised for splines (Craven and Wahba, 1979) and is also applicable in broken line

smoothing. The efficient (order n) ways for implementing GCV (Hutchinson and de Hoog, 1985) can be also transferred to broken line smoothing.

A remarkable property of the broken line smoothing is the fact that the resolution (length of consecutive segments of the broken line) δ does not necessarily coincide with that of the given data points, but it can be chosen either finer or coarser, depending on the specific requirements of the problem of interest. This is not met in typical spline models, whose joints coincide with the data points. However, Bates and Wahba (1982) have introduced splines whose joints are fewer than data points. The selectable number of joints, independent of the data points, may be an advantage in certain cases such as, for example, when the data points are too many, or they are quite non-uniformly distributed.

5. Applications

To demonstrate the method we present four applications, the first two being synthesised for exploration purposes and the last two corresponding to real world problems.

5.1 Exploration applications

As a first example we consider the problem of fitting a broken line to only four data points as shown in Figure 2. The four points have abscissae equidistant in the interval [0.5, 3.5] whereas the ordinates follow an irregular pattern. In Figure 2(a) we have fitted broken lines in the same interval [0.5, 3.5] with $m = 3$ segments and for three values of $\tau = 0$, 0.6, and 0.99. As expected, for $\tau = 0$ (corresponding to $\lambda = 0$; note that the inverse of matrix **B** exists in this case) the fitted broken line passes through (in fact is defined by) the four data points. For $\tau = 0.99$ (corresponding to $\lambda = \lambda_m$) the broken line coincides with the least-squares straight line. For an intermediate smoothing parameter $\tau = 0.60$ the fitted line has an intermediate shape in between these two extreme cases.

More interesting is the situation in Figure 2(b) where to the same four points we have fitted broken lines in the wider interval [0, 4] using a large number of segments $m = 50$. As it can be easily verified, in this case most rows of the matrix **B** have all their elements equal to zero and thus **B** is not invertible. Therefore, λ cannot be equal to zero. For the minimum allowed value $\tau = 0.01$ the resulting broken line in Figure 2(b) has a practically smooth shape

and passes through the four data points. For gradually increasing τ the shape becomes more and more smooth until the broken line becomes a straight line for $\tau = 0.99$. Note that in all cases, the broken line provides reasonable extrapolations on both sides of the given data set.

In our second application we have synthesised a hundred data points from the rather complicated generating function

$$y = g(x) := 14 + x\,(1 + 12\,e^{-0.08\,x}) \tag{20}$$

also incorporating a noise component generated from the normal distribution with mean 0 and standard deviation 7. In Figure 3 we have fitted to these data points broken lines with 50 segments. We observe that for $\tau = 0$ the broken line is too rough while for $\tau = 0.12$ the broken line almost coincides with the generating function, although this was not at all involved in the fitting procedure. The value $\tau = 0.12$ was obtained by minimising GCV (equation (19)). Even for $m = 12$ and $\tau = 0.20$ the broken line, also shown in Figure 3, provides a good approximation of the generating function, as it does not differ significantly from the broken line for $m = 50$ and $\tau = 0.12$. Again, the value $\tau = 0.20$ was obtained by minimising GCV for $m = 12$.

Figure 4 provides some more information for the same application regarding the variation of various indices of estimation error with the number of segments m. Specifically, we have plotted there the variation of (a) the mean square error with respect to data points given by $\varepsilon = (1 / n)\,\|\mathbf{y} - \hat{\mathbf{y}}\|^2 = (1 / n)\,\|(\mathbf{I} - \mathbf{A})\,\mathbf{y}\|^2$, (b) the mean square error with respect to the generation function (20) given by $e = (1 / n)\,\|\mathbf{y} - g(\mathbf{x})\|^2$, and (c) the generalised cross-validation given by (19). All these indices have been estimated for the optimum for each m value of τ (obtained by minimising GCV), which is also plotted in the same figure against m. All three error indices decrease with the increase of m, with a large slope for low values of m followed by a very low slope (almost zero) for moderate and large values of m (notably, τ follows an almost similar pattern). Thus, beyond m = 12-16, the errors ε and e, and GCV remain almost constant. This justifies the choice of a low value of m in this range (that is, about 1/6-1/8 of the number of points, $n = 100$) and explains why the optimal broken lines for $m = 50$ and $m = 12$ in Figure 3 almost coincide.

5.2 Real world applications

The first real world application concerns the establishment of river stage-discharge curves. The stage-discharge curves are a basic tool of hydrometry as they serve as the basis for extracting the river discharge time series at a daily or hourly basis using measurements of the river stage. The curves are constructed using simultaneous measurements of the river stage and discharge, which are rather sparse in time (e.g., one per week or month). Typically, the relationship between stage z and discharge Q can be approximated by a power function $Q = \lambda z^{\kappa}$ (see, e.g., Dingman, 1994, pp. 546-551). However, there are good reasons deterring this relationship from being exactly a power function and the fitting of the best curve is a rather complicating task. Therefore, the establishment of stage-discharge relationships has been the subject of several research studies (e.g., DeGagne et al., 1996), as well as the subject of international standards (ISO, 1982). Obviously, the greatest possible accuracy is expedient in the establishment of this relationship, because any inaccuracy is transferred to the river discharge series, which are extracted using this relationship. Therefore, the broken line smoothing may constitute a good non-parametric alternative to the parametric power relationship.

In Figure 5 we demonstrate the use of the broken line smoothing for establishing a stage-discharge curve. The data used are 44 simultaneous stage and discharge measurements at the location Avlaki of River Acheloos, Greece, for the period April 1970-May 1971 (published by Tsakalias and Koutsoyiannis, 1995). The coordinates x and y in Figure 5 are the natural logarithms of stage (in meters) and discharge (in cubic meters per second), respectively. Apparently, the power curve that is plotted as a straight line in Figure 5 is a good approximation for the data points. This straight line can be also obtained by fitting a broken line with $\tau = 0.99$. However, using a lower value of τ we obtain a broken line in closer agreement with the data points thus leading to a more accurate stage-discharge curve. Note that in this case, due to the logarithmic transformation of the data, the broken line is in fact a concatenation of power curves rather than a concatenation of linear segments. For a large $m = 50$, the value of τ that minimises GCV is $\tau = 0.01$. The resulting broken line is shown in Figure 5 along with the curves for $\tau = 0$ and $\tau = 0.99$. However, with the reasoning

demonstrated in the previous example we may choose a much lower $m = 5$ with resulting optimum $\tau = 0.18$. The resulting curve is also plotted in Figure 5; the obtained approximation of the data points is very satisfactory.

Our second real world example concerns modelling of tree-ring data used in dendroclimatology. Tree-ring widths are used to inspect variations of the past climate. However, it is known that the tree-ring data series are affected by non-climatic factors such as biological (the growth of trees declines in an orderly fashion with increasing age) and environmental (such as competition between trees for light and nutrients). Therefore, before the ring-width data can be used for climate studies, removal of non-climatic components is necessary. This process, known as standardisation, is typically performed by fitting linear or negative exponential functions to the data points. Cook and Peters (1981) introduced the use of the smoothing spline as a superior non-parametric approach to tree-ring width series standardisation.

The broken line smoothing may be used as another alternative for performing this standardisation. In the example given in Figure 6 we used a tree-ring data series from a study by Spethoyianni (1996). This is a 91-year record of tree-ring widths (denoted in Figure 6 as y and given in units of mm, whereas x is the corresponding year starting with 1 for the first year). As shown in Figure 6, for a large value of $m = 50$ the optimum τ obtained from minimisation of GCV is very small, i.e $\tau = 8 \times 10^{-8}$. Considering the problem context, the resulting broken line is too rough; in fact it does not perform an adequate smoothing and it differs substantially from the typically used negative exponential function. For a smaller $m = 12$ and for the optimum $\tau = 0.09$ the resulting curve (also shown in Figure 6) is more reasonable. Even more reasonable broken lines for the specific problem can be obtained by using larger values of the smoothing parameter τ. For example, the curve for $m = 25$ and $\tau = 0.25$ shown in Figure 6 may serve as a good basis for the standardisation process (comparable with that achieved by Cook and Peters (1981) with the use of splines, shown in their Figure 1). In conclusion, an objective assessment of the amount of data smoothing cannot be based on GCV in this case. Such an assessment can be done by introducing specific information relevant to the problem, e.g., by examining the correlation between standardised (by means of the chosen broken line smoothing) tree-ring data and available climatic data.

6. Conclusions

The proposed technique for smoothing a broken line fit to observational data is obtained by introducing a smoothness term in a typical piecewise linear regression model with known break points. The smoothness term is defined by means of the angles formed by the consecutive segments of the broken line, and is given an adjustable weight.

The broken line smoothing can be useful for data analysis in several purposes such as interpolation, smoothing, prediction, filling in missing values, estimation and removal of the measurement errors, etc. In this context it may substitute other methods such as locally weighted regression and cubic smoothing splines. The broken line smoothing is closely related to smoothing splines, the main difference being the simpler mathematical expression of the former, as it comprises simply a concatenation of straight-line segments. It is also related to piecewise linear regression, the main difference being the control of the smoothness through the smoothness term. The weight of this term and the number of linear segments are the two adjustable parameters of the proposed method that give its flexibility. As the weight of the smoothing term increases, or the number of segments decreases, the broken line tends to the least-squares straight line corresponding to the given data points. A remarkable property of the broken line smoothing, is the fact that the resolution (i.e., the length of consecutive segments of the broken line) does not necessarily coincide with that of the given data points, but it can be chosen either finer or coarser, depending on the specific requirements of the problem of interest. The selectable resolution, independent of the data points (which has been implemented to splines, too), may be an advantage in certain cases such as, for example, when the data points are too many, or too few, or they are quite non-uniformly distributed.

The examples presented show that the method is appropriate for smoothing of data sets, performing interpolation between data points, either smooth or non-smooth, and adequately approaching complicated laws between the variables, even under the presence of significant random noise in measurements. Two real world applications presented indicate the method’s applicability for diverse tasks in environmental science and engineering, such as the

identification of river stage-discharge relationships used in hydrometry and the standardisation of tree-ring widths used in dendroclimatology.

Appendix: Direct estimation of the elements of square matrices involved in parameter estimation of the broken line smoothing

Using the definitions of the matrices **Π** and **Ψ** from equations (5) and (9), respectively, also defining $\mathbf{B} := \mathbf{\Pi}^T \mathbf{\Pi}$, $\mathbf{C} := \mathbf{\Psi}^T \mathbf{\Psi}$, and $\mathbf{h} = \mathbf{\Pi}^T \mathbf{y}$, and denoting

$$\Sigma_j\, \xi(x, y) := \sum_{c_{j-1} < x_i \le c_j} \xi(x_i, y_i) \tag{A1}$$

for any function ξ, we easily find the following results for the items of **B**, **C** and **h**.

a) The diagonal items of **B** are

$$b_{jj} = \frac{1}{\delta^2} \left(\Sigma_j\, x^2 + \Sigma_{j+1}\, x^2 - 2\, c_{j-1}\, \Sigma_j\, x - 2\, c_{j+1}\, \Sigma_{j+1}\, x + c_{j-1}^2\, n_j + c_{j+1}^2\, n_{j+1}\right) \tag{A2}$$

and the off-diagonal elements

$$b_{j,j-1} = -\frac{1}{\delta^2} \left[\Sigma_j\, x^2 - (c_{j-1} + c_j)\, \Sigma_j\, x + c_{j-1}\, c_j\, n_j\right] \tag{A3}$$

b) The diagonal elements of **C** are

$$c_{jj} = \varepsilon_{j,0}\, \varepsilon_{j,1} + \varepsilon_{j,m-1}\, \varepsilon_{j,m} + 4\, \varepsilon_{j,0}\, \varepsilon_{j,m} \tag{A4}$$

and the off-diagonal elements

$$c_{j,j-1} = -2\, \varepsilon_{j,0}\, \varepsilon_{j,1} - 2\, \varepsilon_{j,0}\, \varepsilon_{j,m}, \qquad c_{j,j-2} = \varepsilon_{j,0}\, \varepsilon_{j,1} \tag{A5}$$

where by definition

$$\varepsilon_{j,k} := \begin{cases} 0 & k = j \\ 1 & k \ne j \end{cases} \tag{A6}$$

c) The elements of the vector **h** are

$$h_j = \frac{1}{\delta}\left(\Sigma_j\, xy - \Sigma_{j+1}\, xy - c_{j-1}\,\Sigma_j\, y + c_{j+1}\,\Sigma_{j+1}\, y\right) \qquad \text{(A7)}$$

The quantity trace(**I** – **A**) that appears in the definition of GCV (equation (19)) can be also expressed in terms of **B** and **C**. Indeed, using an elementary property of trace, we obtain

$$\begin{aligned}\text{trace}(\mathbf{I}-\mathbf{A}) &= n - \text{trace}(\mathbf{A}) = n - \text{trace}[\mathbf{\Pi}\,(\mathbf{\Pi}^T\mathbf{\Pi} + \lambda\,\mathbf{\Psi}^T\mathbf{\Psi})^{-1}\,\mathbf{\Pi}^T] = \\ &n - \text{trace}[\mathbf{\Pi}^T\mathbf{\Pi}\,(\mathbf{\Pi}^T\mathbf{\Pi} + \lambda\,\mathbf{\Psi}^T\mathbf{\Psi})^{-1}] = n - \text{trace}[\mathbf{B}\,(\mathbf{B} + \lambda\,\mathbf{C})^{-1}] \qquad \text{(A8)}\end{aligned}$$

Alternatively,

$$\begin{aligned}\text{trace}(\mathbf{I}-\mathbf{A}) &= n - \text{trace}[(\mathbf{B} + \lambda\,\mathbf{C})\,(\mathbf{B} + \lambda\,\mathbf{C})^{-1}] + \text{trace}[\lambda\,\mathbf{C}\,(\mathbf{B} + \lambda\,\mathbf{C})^{-1}] = \\ &n - m - 1 + \lambda\,\text{trace}[\mathbf{C}\,(\mathbf{B} + \lambda\,\mathbf{C})^{-1}] \qquad \text{(A9)}\end{aligned}$$

Acknowledgements

The earlier version 0 of the computer program for broken line smoothing, based on a slightly different mathematical background (D. Koutsoyiannis, unpublished notes), was developed in the framework of the Hydroscope project (Creation of a National Data Base for Hydrological and Meteorological Information) funded by the European Union in the framework of the STRIDE programme (1992-94), and the Greek Government. Thanks are due to A. Manetas for the programming and testing of version 0 in C, and I. Nalbantis and A. Christofidis for their comments on this manuscript. The anonymous reviewers' comments, which resulted in substantial improvement of the paper, are gratefully appreciated. Visual Basic and MS-Excel are registered trademarks of Microsoft Corporation.

References

Bartels, R. H., J. C. Beatty, and B. A. Barsky, 1987. *An introduction to Splines for Use in Computer Graphics and Geometric Modeling*, Morgan Kaufmann, Los Altos, 476 pp.

Bates, D., and G. Wahba, 1982. Computational methods for generalized cross-validation with large data sets, in *Treatment of Integral Equations by Numerical Methods*, edited by C. Baker and G. Miller, Academic Press, London.

Bellman, R., and R. Roth, 1969. Curve fitting by segmented straight lines, *Journal of the American Statistical Association*, 1079-1084.

Cleveland, W. S., 1979. Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association*, 74(368), 829-836.

Cleveland, W. S., and R. McGill, 1984. The many faces of a scatterplot, *Journal of the American Statistical Association*, 79(388), 807-822.

Cook, E. R. and K. Peters, 1981. The smoothing spline: A new approach to standardizing forest interior tree-ring width series for dendroclimatic studies, *Tree-Ring Bulletin*, 41, 45-55.

Craven, P., and G. Wahba, 1979. Smoothing noisy data with spline functions, *Numerische Mathematik*, 31, 377-403.

de Boor, C., 1978. *A Practical Guide to Splines*, Applied Mathematical Sciences, Vol. 27, Springer-Verlag, New York, 392 pp.

Ertel, J. E., and E. W. Fowlkes, 1976. Some algorithms for linear spline and piecewise multiple linear regression, *Journal of the American Statistical Association*, 71 (355), 640-648.

DeGagne, M. P. J., G. G. Douglas, H. R. Hudson, and S. P. Simonovic, 1996. A decision support system for the analysis and use of stage-discharge rating curves, *Journal of Hydrology*, 184, 225-241.

Dingman, S. L., 1994. *Physical Hydrology*, Prentice Hall, Englewood Cliffs, New Jersey.

Gallant A. R., and W. A. Fuller, 1973. Fitting segmented polynomial regression models whose join points have to be estimated, *Journal of the American Statistical Association*, 68(341), 144-147.

Hirsch, R. M., D. R. Helset, T. A. Cohn, and E. J. Gilroy, 1993. Statistical analysis of hydrologic data, in *Handbook of Hydrology*, D. R. Maidment (ed.), McGraw-Hill.

Hutchinson, M. F., and F. R. de Hoog, 1985. Smoothing noisy data with spline functions, *Numerische Mathematik*, 47, 99-106.

ISO, 1982. Liquid flow measurements in open channels – Part 2: Determination of the stage-discharge relationship, ISO Standard 1100/2-1982, in Measurement of Liquid Flow in Open Channels – ISO Standards Handbook 16, International Organization of Standards, Geneva, pp. 154-186.

Lerman, P. M., 1980. Fitting segmented regression models by grid search, *Applied Statistics*, 29(1), 74-84.

Marlow, W. H., 1993. *Mathematics for Operations Research*, Dover Publications, New York.

Mitasova, H., and L. Mitas, 1993. Interpolation by regularised spline with tension, I, Theory and implementation, *Mathematical Geology*, 25, 641-655.

Reinsch, C. H., 1967. Smoothing by spline functions, *Numerische Mathematik*, 10, 177-185.

Reinsch, C. H., 1971. Smoothing by spline functions, II, *Numerische Mathematik*, 16, 451-454.

Spethoyianni, M., 1996. Applications of dendrochronology in hydrology and climatology, diploma thesis (in Greek) , National Technical University, Athens.

Tsakalias, G., and D. Koutsoyiannis, 1995. Stage-discharge curves and derivation of discharges, vol. 19 of Evaluation and Management of Water Resources of Sterea Hellas (report in Greek), National Technical University, Athens.

Wahba, G., 1990. Spline models for observational data, in *CBMS-NSF Regional Conference Series in Applied Mathematics*, Society for Industrial and Applied Mathematics, Philadelphia, 169 pp.

List of Figures

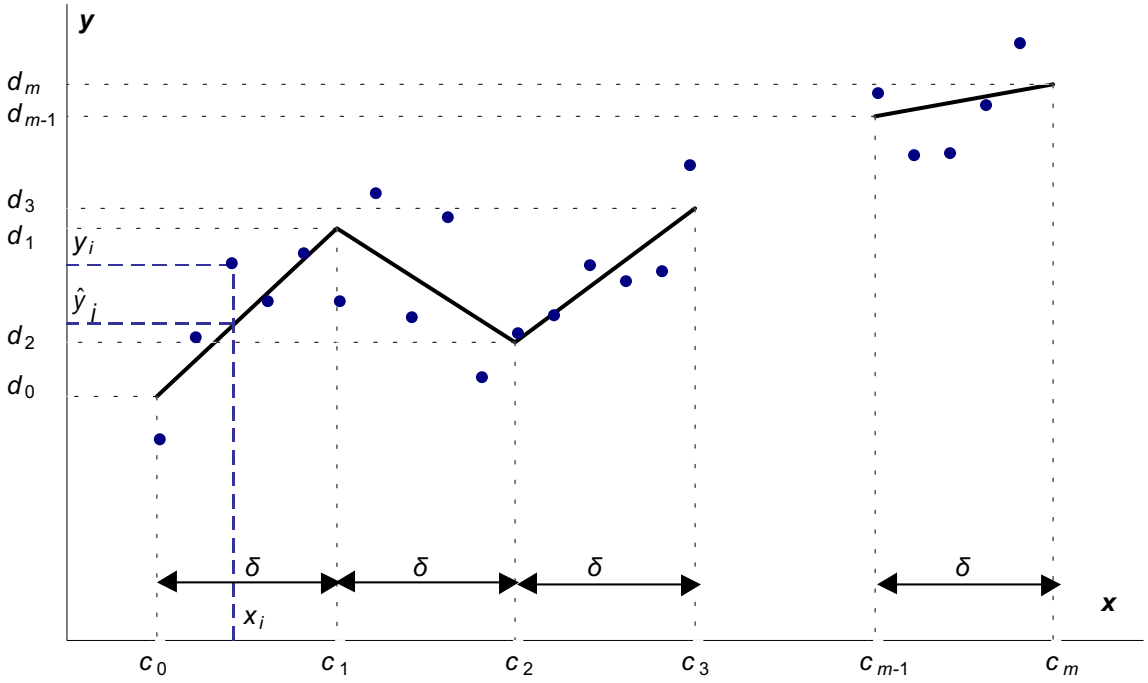

Figure 1 Definition sketch.

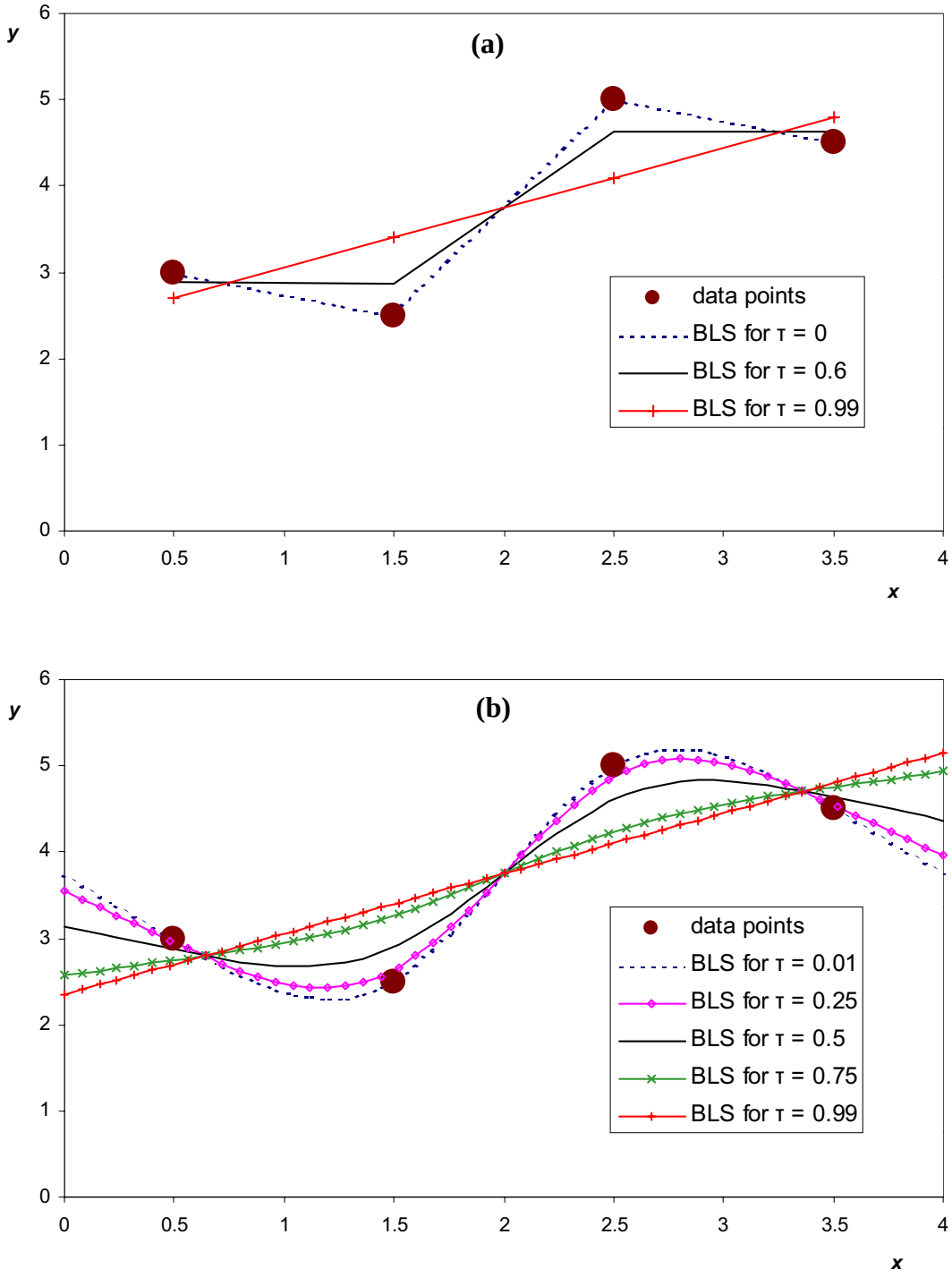

Figure 2 Fitted broken lines (BLS) to four data points for various values of the smoothing parameter τ and for number of segments: (a) $m = 3$, and (b) $m = 50$.

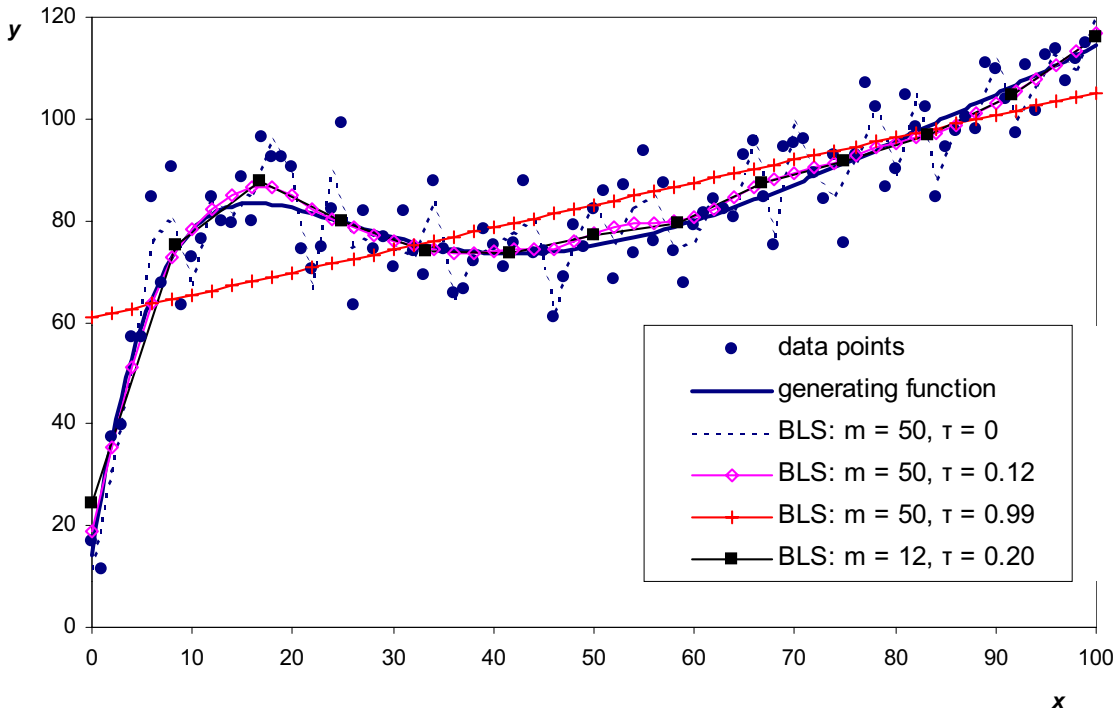

Figure 3 Fitted broken lines (BLS) to data points synthesised by the generating function (20) (plus a noise term) for various values of the smoothing parameters.

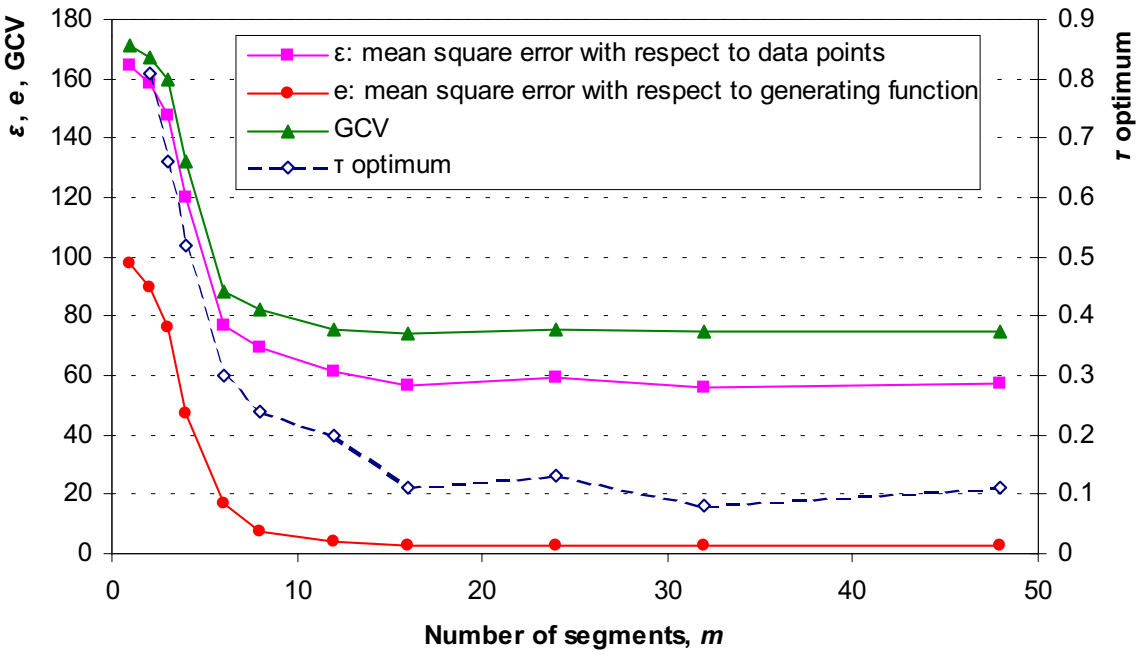

Figure 4 Variation of the optimum value of the smoothing parameter τ and the corresponding estimation errors, with the number of segments, m, in the application of Figure 3.

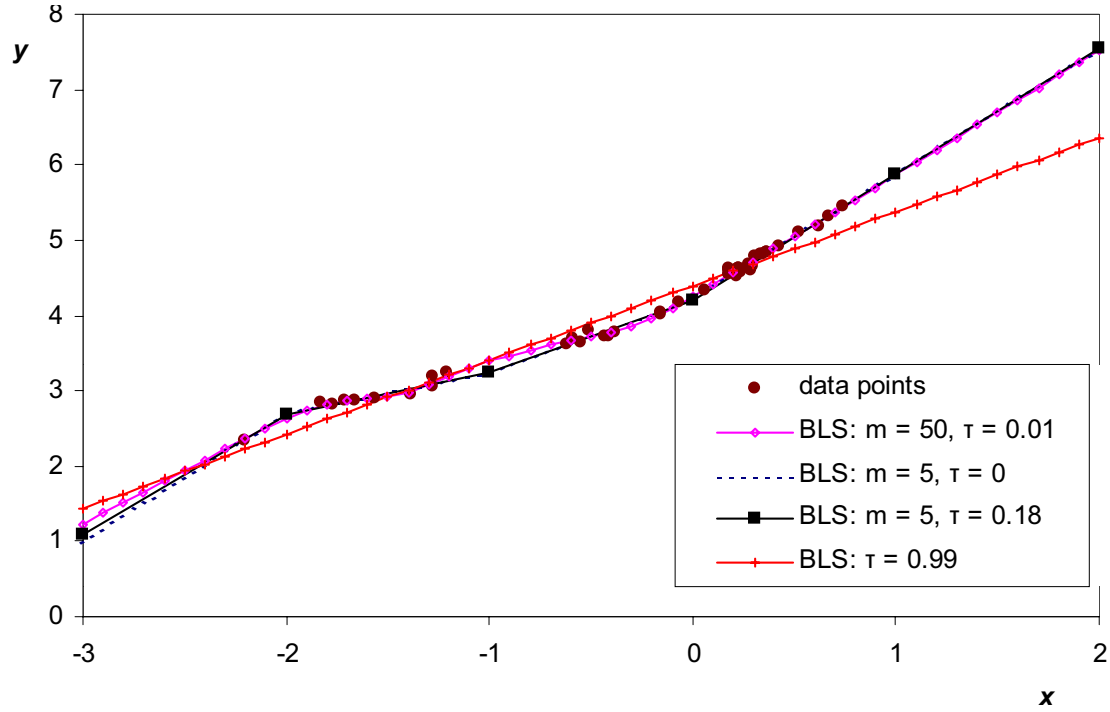

Figure 5 Fitted broken lines (BLS) to a series of river stage and discharge measurements for various values of the smoothing parameters τ and m. Coordinates x and y are the natural logarithms of stage (in meters) and discharge (in cubic meters per second), respectively.

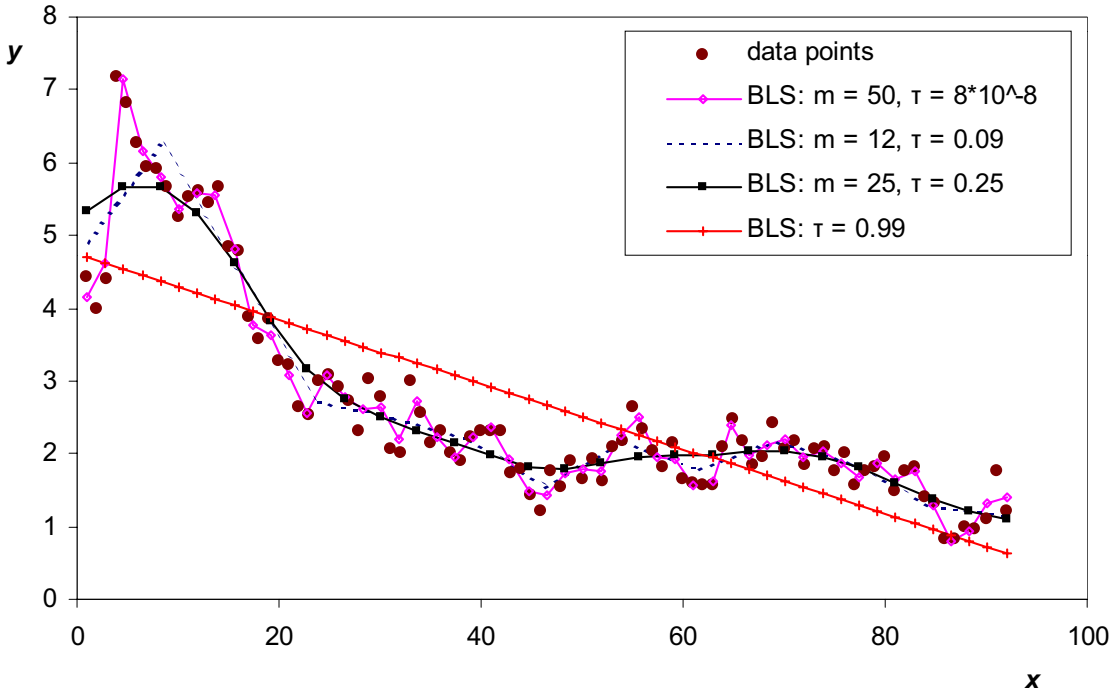

Figure 6 Fitted broken lines (BLS) to a tree-ring width series (y in millimetres) for various values of the smoothing parameters τ and m.