Hydrological Sciences Journal – Journal des Sciences Hydrologiques, 2015
http://dx.doi.org/10.1080/02626667.2014.899703

Broken line smoothing for data series interpolation by incorporating an explanatory variable with denser observations: application to soil-water and rainfall data

Nikolaos Malamos[1] and Demetris Koutsoyiannis[2]

[1]*Department of Agricultural Technology, Technological Educational Institute of Western Greece, Amaliada, Greece*
nmalamos@teimes.gr

[2]*Department of Water Resources and Environmental Engineering, School of Civil Engineering, National Technical University of Athens, Zographou, Greece*

Received 30 June 2013; accepted 19 February 2014

Editor Z.W. Kundzewicz; **Associate editor** A. Carsteanu

Abstract Broken line smoothing is a simple technique for smoothing a broken line fit to observational data and provides a flexible means of interpolation. Here an extension of this technique is proposed, which can be utilized to perform various interpolation tasks, by incorporating, in an objective manner, an explanatory variable available at a considerably denser dataset than the initial main variable. The technique incorporates smoothing terms with adjustable weights, defined by means of the angles formed by the consecutive segments of two broken lines. The mathematical framework and details of the method as well as practical aspects of its application are presented and discussed. Also, examples using both synthesized and real-world (soil water dynamics and hydrological) data are presented to explore and illustrate the methodology.

Key words broken line smoothing; interpolation; explanatory variable; generalized cross-validation (GCV); hydraulic conductivity; rainfall

Lissage par ligne brisée pour l'interpolation de séries de données incorporant une variable explicative avec des observations plus denses: application à l'humidité du sol et aux précipitations
Résumé Le lissage par une ligne brisée est une technique simple pour ajuster une ligne brisée aux données observées et fournit des techniques souples pour l'interpolation. Nous proposons ici une extension de cette technique, qui peut être utilisée pour effectuer diverses tâches d'interpolation, en intégrant, de manière objective, une variable explicative dont l'observation est considérablement plus dense que celle de variable initiale principale. La technique introduit des termes de lissage de pondération réglable, définis à partir des angles formés par les segments consécutifs d'une ligne brisée. Nous présentons et discutons le cadre mathématique et les détails de la méthode ainsi que les aspects pratiques de son application. En outre, des exemples utilisant à la fois des données synthétiques et réelles (concernant la dynamique de l'eau du sol et l'hydrologie) sont présentés afin d'étudier et d'illustrer la méthodologie.

Mots clefs lissage de ligne brisée ; interpolation ; variable explicative ; validation croisée généralisée ; conductivité hydraulique ; pluie

INTRODUCTION

In numerous scientific and engineering applications the dependence of a variable y on another variable x, described by a fitted curve, is exploited for purposes such as interpolation between measurements, prediction, filling in missing values in time series, estimation and removal of the measurement errors, etc. Whenever 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 techniques. The difficulty arises when such an expression is not known and cannot be approximated by a simple, easily recognizable formulation.

Currently, many methods exist that can accomplish this task using appropriate computer codes and

they are mainly used for spatial interpolation in environmental studies. They fall into three categories (Li and Heap 2008):

(a) non-geostatistical methods such as: splines (Craven and Wahba 1978, Wahba and Wendelberger 1980) and regression methods (Davis 1986);
(b) geostatistical methods, including different approaches of kriging, such as: universal kriging, kriging with an external drift or co-kriging (Goovaerts 1997, Burrough and McDonnell 1998); and
(c) combined methods, such as: trend surface analysis combined with kriging (Wang *et al*. 2005) and regression kriging (Hengl *et al*. 2007).

Koutsoyiannis (2000) introduced the easy method of broken line smoothing (BLS) as a simple alternative to numerical smoothing and interpolating methods, 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_is, but they can form a series of points with some chosen (lower or higher) resolution. The main concept of the method is the trade-off between two objectives, i.e. minimizing the fitting error and the roughness of the broken line. The larger the relative weight of the second objective, the smoother the broken line resulting from the fitting procedure.

This study is focused on the combination of two broken lines into a piecewise linear regression model with known break points and adjustable weights. The first broken line is fitted to the available data points, while the second incorporates, in an objective manner, the influence of an explanatory variable available from a considerably denser dataset. The objective is to make the interpolation across the data points as accurate as possible.

The method is illustrated using three applications: (a) a theoretical–investigational; (b) the interpolation of hydraulic conductivity function using water retention data as explanatory variable and *vice versa*; and (c) the spatial interpolation of rainfall data using the surface elevation as explanatory variable.

METHODOLOGY

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 ($j = 0, \ldots, m, m + 1$) be points of the x-axis so that the interval $[c_0, c_m]$ contains all x_i. For simplicity we will assume that the points are equidistant, i.e. $c_j - c_{j-1} = \delta$ and that for every x value we know the value of an explanatory variable t. Therefore, for each point (x_i, y_i) there is a corresponding value $t(x_i)$, for $i = 1, \ldots, n$ and for a value c_j there is a corresponding value $t(c_j)$, for $j = 0, \ldots, m$.

We make the assumption that the dependent variable y in every position x can be expressed as a linear function of the variable t, i.e.

$$y = d + et \tag{1}$$

where d and e are coefficients, with their values changing according to x. This is not a global linear relationship but a local linear one, as the quantities d and e change with x. Their variation is expressed from two broken piecewise straight lines. At the vertices of the broken lines, the above relationship becomes:

$$y_j = d_j + e_j t_j \tag{2}$$

We wish to find the $m + 1$ values d_j and e_j, so that the curve which is defined by the $m + 1$ points $(c_j, d_j + t_j e_j)$, and consists of a combination of the two broken lines and of the $t(x)$ curve, 'fits' the set of points (x_i, y_i). This fit is defined in terms of minimizing the total square error among the set of original points (x_i, y_i) and the fitted curve, i.e.

$$p = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3}$$

where $\hat{y}_i$ is the estimate of y_i given by the broken lines for the known x_i.

In matrix form, this can be written as:

$$p = \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2 \tag{4}$$

where $\boldsymbol{y} = [y_1, \ldots, y_n]^{\mathrm{T}}$ 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) and $\hat{\boldsymbol{y}} = [\hat{y}_1, \ldots, \hat{y}_n]^{\mathrm{T}}$ is the vector of estimates with size n as given by application of equation (1), which for any x can be written as:

$$\hat{y}(x) = d(x) + t(x)e(x) \tag{5}$$

where $d(x)$, $e(x)$ are the ordinates of the corresponding broken lines at point x.

Therefore, if a point x_i lies in the subinterval $[c_{j-1}, c_j]$ for some j ($1 \le j \le m$), then the estimate $\hat{y}_i$ is given by:

$$\begin{aligned}\hat{y}_i(x_i, t(x_i)) &= \frac{1}{\delta}\{\left[d_j\left(x_i - c_{j-1}\right) + d_{j-1}\left(c_j - x_i\right)\right] \\ &\quad + t(x_i)\left[e_j\left(x_i - c_{j-1}\right) + e_{j-1}\left(c_j - x_i\right)\right]\}\end{aligned} \tag{9}$$

If we apply equation (9) for $i = 1, 2, \ldots, n$, we get:

$$\begin{aligned}\hat{y}_1 &= \frac{1}{\delta}\{[d_1(x_1 - c_0) + d_0(c_1 - x_1)] + t(x_1)[e_1(x_1 - c_0) + e_0(c_1 - x_1)]\} \\ &\vdots \\ \hat{y}_n &= \frac{1}{\delta}\{[d_m(x_n - c_{m-1}) + d_{m-1}(c_m - x_n)] + t(x_n)[e_m(x_n - c_{m-1}) + e_{m-1}(c_m - x_n)]\}\end{aligned} \tag{10}$$

Assuming that for some j, $c_{j-1} \le x \le c_j$ (see Fig. 1), the ordinate of the broken line $d(x)$ can be determined from:

$$\begin{aligned}d(x) &= d_j + \left(d_{j-1} - d_j\right)\frac{c_j - x}{c_j - c_{j-1}} \\ &= d_j + \left(d_{j-1} - d_j\right)\frac{c_j - x}{\delta}\end{aligned} \tag{6}$$

which can be written as:

$$d(x) = \tfrac{1}{\delta}\left[\left(x - c_{j-1}\right)d_j + \left(c_j - x\right)d_{j-1}\right] \tag{7}$$

Likewise, the corresponding expression for e is:

$$e(x) = \tfrac{1}{\delta}\left[\left(x - c_{j-1}\right)e_j + \left(c_j - x\right)e_{j-1}\right] \tag{8}$$

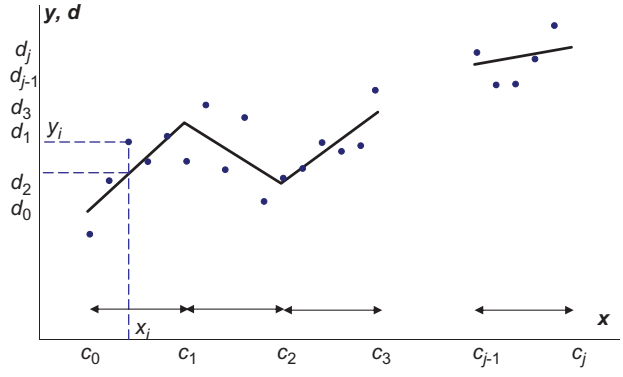

Fig. 1 Definition sketch for vector $\boldsymbol{d}$, adopted from Koutsoyiannis (2000) (a similar sketch can be drawn for vector $\boldsymbol{e}$).

in which we assumed that the point x_1 lies in the interval $[c_0, c_1]$ and point x_n lies in the interval $[c_{m-1}, c_m]$.

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

$$\hat{\boldsymbol{y}} = \boldsymbol{\Pi d} + \boldsymbol{T\Pi e} \tag{11}$$

where $\hat{\boldsymbol{y}} = [\hat{y}_1, \ldots, \hat{y}_n]^{\mathrm{T}}$ is the vector of estimates with size n; $\boldsymbol{d} = [d_0, \ldots, d_m]^{\mathrm{T}}$ is the vector of the unknown ordinates of the broken line d; $\boldsymbol{e} = [e_0, \ldots, e_m]^{\mathrm{T}}$ is the vector of the unknown ordinates of the broken line e; $\boldsymbol{T}$ is a diagonal matrix:

$$\boldsymbol{T} = \mathrm{diag}(t(x_1), \ldots, t(x_n)) \tag{12}$$

with its elements $t(x_1)$, …, $t(x_n)$ being the values of the altitude at the given data points; and $\boldsymbol{\Pi}$ is a matrix with size $n \times (m + 1)$ whose ijth entry (for $i = 1, \ldots, n$; $j = 0, \ldots, m$) is:

$$\pi_{ij} = \begin{cases} \frac{x_i - c_{j-1}}{\delta}, & c_{j-1} < x_i \le c_j \\ \frac{c_{j+1} - x_i}{\delta}, & c_j < x_i \le c_{j+1} \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

In addition to minimizing the fitting error defined in equation (4), we consider two requirements in order to avoid a very rough shape of both broken lines d and e, and also ensuring a unique solution irrespective of how large m is (see explanation below). To acquire a measure of the roughness of the broken

line, we consider the differences of slopes between two consecutive segments of the broken line d and the broken line e, so that the following expressions can be appropriate measures for the roughness of the entire broken line:

$$q_d = \sum_{j=1}^{m-1} \left(2d_j - d_{j-1} - d_{j+1}\right)^2 \tag{14}$$

and

$$q_e = \sum_{j=1}^{m-1} \left(2e_j - e_{j-1} - e_{j+1}\right)^2 \tag{15}$$

These can be written in matrix form as:

$$q_d = \boldsymbol{d}^{\mathrm{T}} \boldsymbol{\Psi}^{\mathrm{T}} \boldsymbol{\Psi} \boldsymbol{d} \tag{16}$$

and

$$q_e = \boldsymbol{e}^{\mathrm{T}} \boldsymbol{\Psi}^{\mathrm{T}} \boldsymbol{\Psi} \boldsymbol{e} \tag{17}$$

where $\boldsymbol{\Psi}$ is a matrix with size $(m-1) \times (m+1)$ and ijth entry:

$$\psi_{ij} = \begin{cases} 2, & j = i+1 \\ -1, & |j-i-1| = 1 \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

Apparently, $\boldsymbol{\Psi}^{\mathrm{T}} \boldsymbol{\Psi} = 0$ for the special case $m = 1$.

Combining equations (4), (11), (16) and (17), and introducing the dimensionless multipliers λ, $\mu \geq 0$ for q_d and q_e, respectively, we form the generalized objective function to be minimized:

$$\begin{aligned} f(\boldsymbol{d}, \boldsymbol{e}) &:= p + \lambda q_d + \mu q_e \\ &= \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2 + \lambda \boldsymbol{d}^{\mathrm{T}} \boldsymbol{\Psi}^{\mathrm{T}} \boldsymbol{\Psi} \boldsymbol{d} + \mu\, \boldsymbol{e}^{\mathrm{T}} \boldsymbol{\Psi}^{\mathrm{T}} \boldsymbol{\Psi} \boldsymbol{e} \end{aligned} \tag{19}$$

By differentiating equation (19) with respect to $\boldsymbol{d}$ and $\boldsymbol{e}$, and equating them to zero, we obtain, respectively:

$$\begin{aligned} \frac{\partial f_1}{\partial \boldsymbol{d}} = &- 2\boldsymbol{y}^{\mathrm{T}} \boldsymbol{\Pi} + 2\boldsymbol{d}^{\mathrm{T}} \boldsymbol{\Pi}^{\mathrm{T}} \boldsymbol{\Pi} \\ &+ 2\boldsymbol{e}^{\mathrm{T}} \boldsymbol{\Pi}^{\mathrm{T}} \boldsymbol{T}^{\mathrm{T}} \boldsymbol{T} \boldsymbol{\Pi} + 2\lambda \boldsymbol{d}^{\mathrm{T}} \boldsymbol{\Psi}^{\mathrm{T}} \boldsymbol{\Psi} = 0 \end{aligned} \tag{20}$$

$$\begin{aligned} \frac{\partial f_2}{\partial \boldsymbol{e}} = &- 2\boldsymbol{y}^{\mathrm{T}} \boldsymbol{T} \boldsymbol{\Pi} + 2\boldsymbol{d}^{\mathrm{T}} \boldsymbol{\Pi}^{\mathrm{T}} \boldsymbol{T} \boldsymbol{\Pi} \\ &+ 2\boldsymbol{e}^{\mathrm{T}} \boldsymbol{\Pi}^{\mathrm{T}} \boldsymbol{T}^{\mathrm{T}} \boldsymbol{T} \boldsymbol{\Pi} + 2\mu \boldsymbol{e}^{\mathrm{T}} \boldsymbol{\Psi}^{\mathrm{T}} \boldsymbol{\Psi} = 0 \end{aligned} \tag{21}$$

The solution of the set of equations (20) and (21), which minimizes (19), is obtained after applying the typical rules of derivatives involving matrices and has the following form:

$$\begin{bmatrix} \boldsymbol{d} \\ \boldsymbol{e} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Pi}^{\mathrm{T}} \boldsymbol{\Pi} + \lambda \boldsymbol{\Psi}^{\mathrm{T}} \boldsymbol{\Psi} & \boldsymbol{\Pi}^{\mathrm{T}} \boldsymbol{T} \boldsymbol{\Pi} \\ \boldsymbol{\Pi}^{\mathrm{T}} \boldsymbol{T} \boldsymbol{\Pi} & \boldsymbol{\Pi}^{\mathrm{T}} \boldsymbol{T}^{\mathrm{T}} \boldsymbol{T} \boldsymbol{\Pi} + \mu \boldsymbol{\Psi}^{\mathrm{T}} \boldsymbol{\Psi} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{\Pi}^{\mathrm{T}} \boldsymbol{y} \\ \boldsymbol{\Pi}^{\mathrm{T}} \boldsymbol{T}^{\mathrm{T}} \boldsymbol{y} \end{bmatrix} \tag{22}$$

The vector of estimates, $\hat{\boldsymbol{y}}$, is obtained from equation (11), once vectors $\boldsymbol{d}$ and $\boldsymbol{e}$ are calculated from equation (22). Also, from equation (9), we can estimate the ordinate $\hat{y}$ of any x that lies in the interval $[c_0, c_m]$ if we know the value of parameter t at that point.

We observe that the three matrices $\boldsymbol{B} := \boldsymbol{\Pi}^{\mathrm{T}} \boldsymbol{\Pi}$, $\boldsymbol{C} := \boldsymbol{\Psi}^{\mathrm{T}} \boldsymbol{\Psi}$ and $\boldsymbol{D} := \boldsymbol{\Pi}^{\mathrm{T}} \boldsymbol{T}^{\mathrm{T}} \boldsymbol{T} \boldsymbol{\Pi}$ appearing in (22) are square matrices with size $(m+1) \times (m+1)$. $\boldsymbol{B}$ and $\boldsymbol{D}$ are tridiagonal while $\boldsymbol{C}$ is five-banded. $\boldsymbol{B}$ can be singular (not invertible) if one or more columns of $\boldsymbol{\Pi}$ have zero elements' that is, if at least two consecutive intervals $[c_{j-1}, c_j]$ contain no x_is while $\boldsymbol{C}$ is always singular. However, for $\lambda, \mu > 0$, the sums $\boldsymbol{B} + \lambda \boldsymbol{C}$ and $\boldsymbol{D} + \mu \boldsymbol{C}$ are non-singular and, thus, their inverses exist.

Choice of parameters

It is apparent that the method has three adjustable parameters: the number of intervals, m, and the smoothing parameters λ and μ corresponding to vectors $\boldsymbol{d}$ and $\boldsymbol{e}$, respectively. The choice of parameters can be made by assessing the achieved data smoothing, either graphically in the case of a limited number of data points ($n \leq 3$), or by using standard objective ways as described by the following analysis.

In order to provide a convenient search of the two smoothing parameters, we selected a transformation of λ and μ in terms of what has been called tension parameters, τ_λ and τ_μ, whose values are restricted in the interval [0, 1). These were derived from the numerical investigation performed by Koutsoyiannis (2000), concerning the transformation of the smoothing parameter λ, and have the form:

$$\lambda = \left(10\, m \frac{\ln \tau_{\mathrm{m}}}{\ln \tau_\lambda}\right)^{K_\lambda}, \quad \mu = \left(10\, m \frac{\ln \tau_{\mathrm{m}}}{\ln \tau_\mu}\right)^{K_\mu} \tag{23}$$

where $\tau_m = 0.99$ is the maximum allowed tension, corresponding to the upper bound of λ and μ, set for numerical stability equal to:

$$\lambda_m = \frac{\text{trace}(\boldsymbol{B})}{\text{trace}(\boldsymbol{C})}\,10^8, \mu_m = \frac{\text{trace}(\boldsymbol{D})}{\text{trace}(\boldsymbol{C})}10^8 \quad (24)$$

The exponents in equations (23) are determined by the relationships:

$$\mathcal{K}_\lambda = \frac{\ln\lambda_m}{\ln(10m)}, \mathcal{K}_\mu = \frac{\ln\mu_m}{\ln(10m)} \quad (25)$$

which are obtained by combining equations (23) and (24). The minimum allowed value of λ, μ is 0 if the inverse of matrixes $\boldsymbol{B}$ and $\boldsymbol{D}$ exist; otherwise they are estimated from equations (23) using small values of τ_λ and τ_μ, such as: $\tau_\lambda = 1 - \tau_m = 0.01$ and $\tau_\mu = 1 - \tau_m = 0.01$.

Combining equations (11) and (22), we obtain:

$$\hat{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{y} \quad (26)$$

where $\boldsymbol{A}$ is a $n \times n$ symmetric matrix given by:

$$\boldsymbol{A} = [\boldsymbol{\Pi}\boldsymbol{T}\boldsymbol{\Pi}]\begin{bmatrix} \boldsymbol{\Pi}^T\boldsymbol{\Pi} + \lambda\boldsymbol{\Psi}^T\boldsymbol{\Psi} & \boldsymbol{\Pi}^T\boldsymbol{T}\boldsymbol{\Pi} \\ \boldsymbol{\Pi}^T\boldsymbol{T}\boldsymbol{\Pi} & \boldsymbol{\Pi}^T\boldsymbol{T}^T\boldsymbol{T}\boldsymbol{\Pi} + \mu\boldsymbol{\Psi}^T\boldsymbol{\Psi} \end{bmatrix}^{-1}[\boldsymbol{\Pi}\,\boldsymbol{T}\boldsymbol{\Pi}]^T \quad (27)$$

depending on all adjustable parameters: m, τ_λ and τ_μ.

The estimation of these adjustable parameters can be done by minimizing the generalized cross-validation (GCV; Craven and Wahba 1978), defined by:

$$\mathbf{GCV} = \frac{\frac{1}{n}\|(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{y}\|^2}{\left[\frac{1}{n}\text{trace}(\boldsymbol{I} - \boldsymbol{A})\right]^2} \quad (28)$$

For a given number of segments m, the minimization of GCV results in the optimum values of τ_λ and τ_μ. This can be repeated for several trial values of m until the global minimum of GCV is reached.

Relationship to broken line smoothing and other methods

The formalization of the above setting of the broken line smoothing interpolation (BLSI) method was derived from that of the single broken line method (Koutsoyiannis 2000), by adding a linear function of the explanatory variable t, along with the introduction of the smoothness term $\boldsymbol{\Psi}^T\boldsymbol{\Psi}$ in the corresponding problem formulation. This allows GCV to be implemented in the parameter selection procedure. The main difference is the fact that the present method uses two broken lines to obtain the vector of estimates $\hat{\boldsymbol{y}}$ from equation (11). Thus, the method does not provide the vertices of a single broken line, but the estimates of points (x_i, y_i) with available $t(x_i)$ values ($i = 1, \ldots, n$) fitted to the problem of interest. The above characteristics of the proposed method do not appear either in smoothing splines or in any other piecewise linear regression method.

It should be obvious from the above discourse that BLSI does not require linearity between the involved variables, namely y, x and the explanatory variable t, but local linearity is incorporated in the mathematical framework in a broken line approach. Also, the functional dependence, in terms of vectors $\boldsymbol{d}$, $\boldsymbol{e}$, the number of segments, m, and the tension parameters τ_λ and τ_μ, is neither constant nor known *a priori*, but in each case is determined through the procedure of minimizing the GCV.

Finally, the method retains the remarkable property of broken line smoothing (Koutsoyiannis 2000), in that the resolution (length of consecutive segments of the broken line) δ does not necessarily have to coincide with that of the given data points, but can be either finer or coarser, depending on the specific requirements of the problem of interest.

RESULTS AND COMMENTS

The exploration of the proposed method took place against synthesized and real-world data. To demonstrate the method we present three applications, the first being synthesized for exploration purposes and the last two corresponding to real-world problems. The computational framework of the method's implementation (Microsoft Excel) provides a direct means of data visualization and graphical exploration.

Exploration application

The first application was the implementation of the method in interpolation-fitting to 10 data points

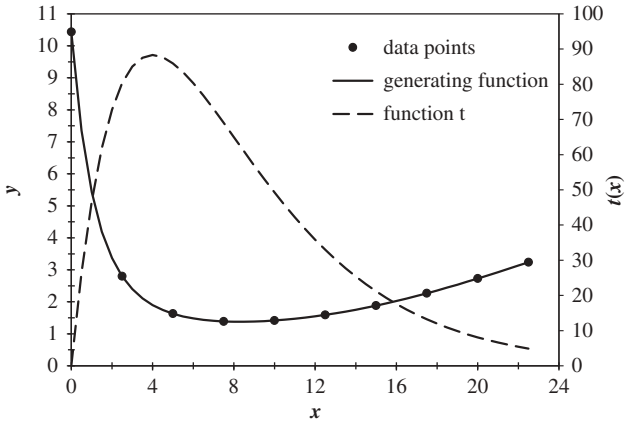

Fig. 2 Generating function $y(x)$ and explanatory function $t(x)$ for the purposes of the exploration application.

obtained from the rather complicated generating function (Fig. 2):

$$y(x) = \left(1 + 10e^{-0.01x - 0.1\sqrt{tx}}\right) - 0.57 \tag{29}$$

while the values of the explanatory variable t, which depends on x values, are given from the relationship (Fig. 2):

$$t(x) = 60xe^{-0.25x} \tag{30}$$

The main objective of this application, apart from illustrating how the proposed method performs in interpolation based on given data points, was the investigation of the variation of the three parameters: m, τ_λ and τ_μ, and the confirmation that there is a single global minimum for the generalized cross-validation (GCV).

In order to achieve this, we implemented the method for different numbers of segments m ($2 \le m \le 30$) using 10 data points ($i = 10$), in order to approximate the generating function of equation (29). The explanatory dataset, $t(x)$ comprised 46 points ($j = 1, \ldots, 46$) equally spaced along [0, 22.5], with $x_j - x_{j-1} = \delta = 0.5$. Consequently, for each case we obtained 46 point estimates of the generating function by applying equation (9).

Figure 3 presents the variation of the following indices *versus* the number of segments m:

(a) the minimum GCV given by equation (28);
(b) the GCV given by equation (28) when τ_λ and τ_μ are close to their lower limit ($\tau_\lambda = \tau_\mu = 0.01$) and when they take their maximum value ($\tau_\lambda = \tau_\mu = 0.99$);
(c) the mean square error (MSE) with respect to data points, provided by the numerator of equation (28), obtained by minimizing GCV;
(d) the MSE with respect to data points provided by the numerator of equation (28), obtained by minimizing GCV when τ_λ and τ_μ are close to their lower limit ($\tau_\lambda = \tau_\mu = 0.01$) and when they take their maximum value ($\tau_\lambda = \tau_\mu = 0.99$); and
(e) the optimum values for each m, of τ_λ and τ_μ obtained by minimizing GCV.

Figure 3 shows that, in the case of maximum τ_λ and τ_μ values (=0.99), both error indices are almost invariant and independent of the number of segments. In this case, the influence of the smoothness term $\boldsymbol{\Psi}^T\boldsymbol{\Psi}$ in equation (27) is much higher than the influence of the broken line segments, resulting in a single 'maximum smoothness' solution.

The global minimum value of GCV was 1.460×10^{-4}, corresponding to $m = 7$, $\tau_\lambda = 0.01$ and $\tau_\mu = 0.304$. Beyond $m = 11$, the minimum GCV, as well as the GCV for the case of $\tau_\lambda = \tau_\mu = 0.01$, remain almost constant. However, the existence of local minima should be taken into consideration during the parameter estimation procedure.

When GCV is minimized, MSE follows a similar pattern to the GCV variation, with its minimum value being 2.540×10^{-8} for $m = 7$,$\tau_\lambda = 0.01$ and $\tau_\mu = 0.01$. However, the global minimum value of MSE was 8.099×10^{-14} and occurred in the case of minimum tension values, i.e. $\tau_\lambda = \tau_\mu = 0.01$ and $m = 9$. This complies with the formulation of equation (19) concerning the roughness of the broken lines. Also, Fig. 3 shows that, in this case, the values of MSE tend to remain stable for larger values of m.

The optimum values of τ_λ and τ_μ achieved by minimizing GCV *versus* different numbers of segments m, are presented in Fig. 3. Even though they appear in a different scale, their pattern is similar to these of the error indices, being almost stable after $m = 6$ for τ_λ and $m = 9$ for τ_μ.

Figure 3 confirms that the proposed mathematical formulation ensures the presence of a single global minimum value of GCV according to equation (28) and therefore the applicability of the objective way to assess the optimum values of τ_λ and τ_μ, as previously noted.

Figure 4 presents the BLSI fit, using 10 data points and 46 values of the explanatory variable t ($i = 10, j = 46$), to the generating function of equation

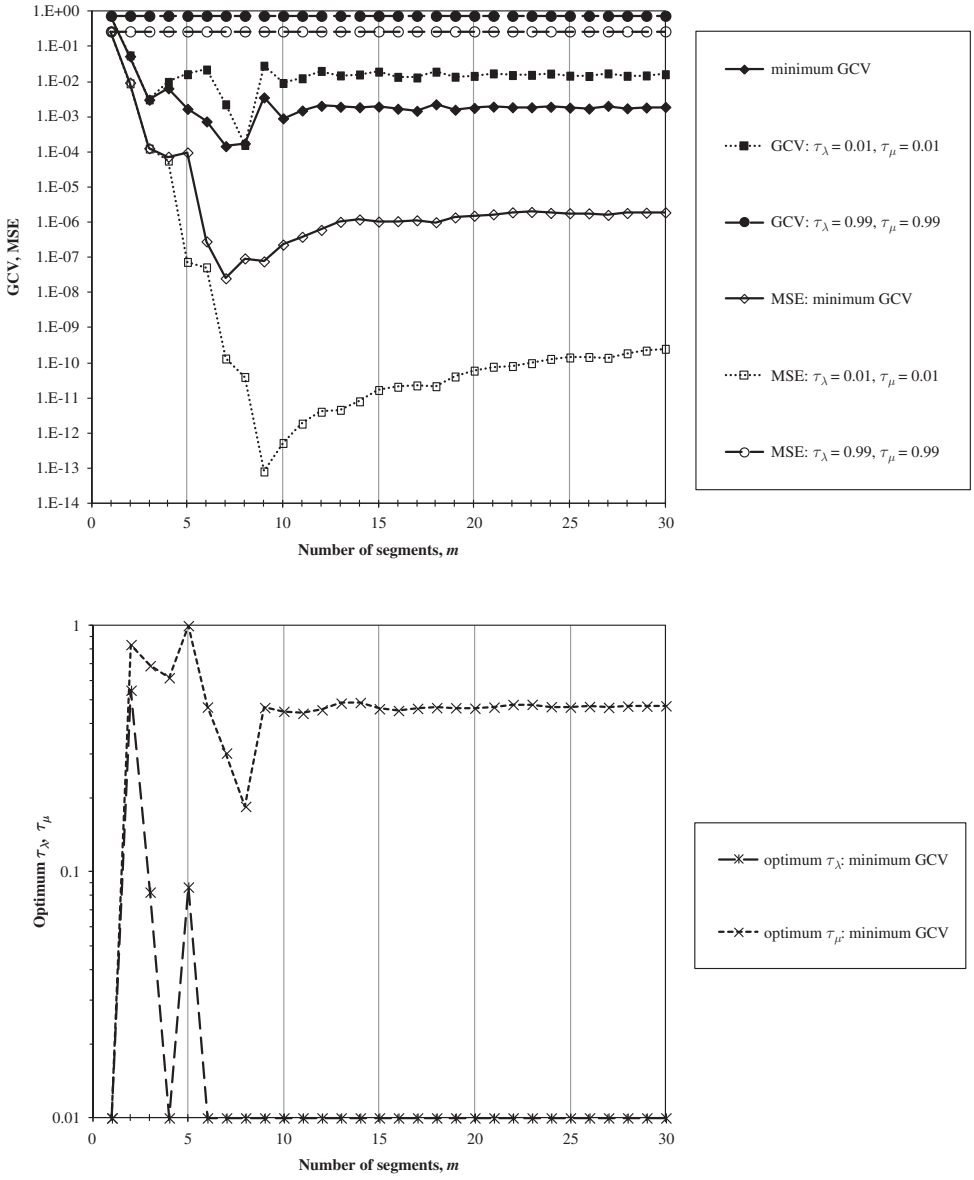

Fig. 3 Variation of the MSE and GCV values along with the corresponding smoothing parameters τ_λ and τ_μ *versus* the number of segments, *m*.

(29) for the values of the smoothing parameters presented during the analysis of Fig. 3. The 46 point estimates of the generating function were obtained by applying equation (9). It is apparent that, in the case of the global minimum value for the GCV indicator (m = 7, τ_λ = 0.01, τ_μ = 0.304), the estimates are indistinguishable from the generating function, which suggests that the error is negligible. In the case of small tension values (m = 8, $\tau_\lambda = \tau_\mu$ = 0.01), the estimates are also very close to the generating function, but the overall appearance is somewhat rough with deviations between the data points. This characteristic was expected, since the GCV values for both cases were very close. However, the maximum tension case of $\tau_\lambda = \tau_\mu$ = 0.99 resulted in a smoother curve but a worse approximation of the generating function.

The fitted broken lines in terms of the vectors ***d*** and ***e*** that satisfy equation (22), for the above mentioned cases, are presented in Fig. 5.

Even though the final result as presented in Fig. 4 shows small differences between the cases of the optimum solution and small tension values, there is significant variation between the corresponding broken lines, i.e. vectors ***d*** and ***e***, as they appear in Fig. 5. Also, the scale and gradient differences between the two groups of broken lines indicate that the ordinates of the broken line *e*, are the

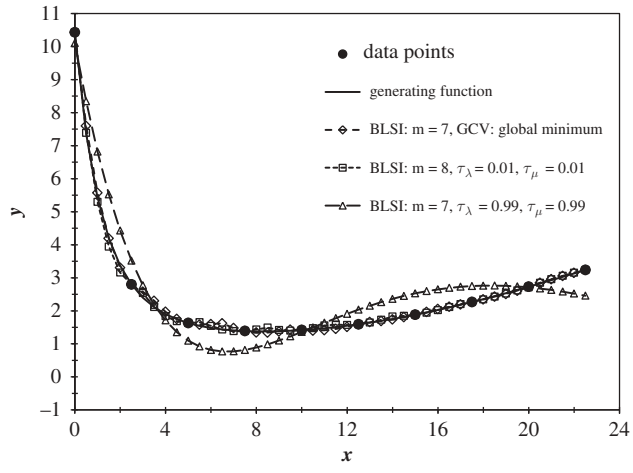

Fig. 4 BLSI fit, using 10 data points, to the generating function of equation (29) for various values of the smoothing parameters.

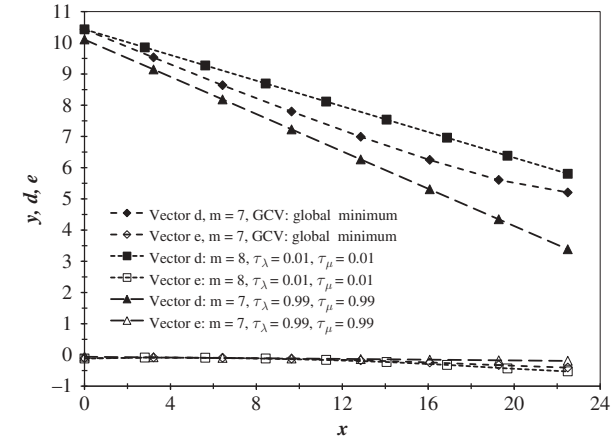

Fig. 5 Fitted broken lines (vectors ***d*** and ***e***) to the generating function of equation (29) for various values of the smoothing parameters.

adjustment coefficients–weights of the explanatory variable t. This complies with the mathematical framework of the method as expressed by equation (1).

In brief, the combination of the two broken lines by means of equation (11) results in a very satisfactory fit of the complex mathematical expression as described by equation (29) and depicted in Fig. 2.

Real-world applications

The first real-world application concerns interpolation between measurements of the hydraulic conductivity function $K(h)$ (h being the soil pressure head), mainly used in numerical methods for the simulation and prediction of mass transport phenomena in the vadose zone. Many different closed-form expressions have been widely employed to describe the unsaturated hydraulic properties of soils (Leij *et al*. 1997), but all of them need experimental data to be fitted upon.

For the experimental determination of the hydraulic conductivity function, a number of methods have been developed. Direct methods for measuring the $K(h)$ functions in a laboratory can be classified according to the flow mode as steady state (conventional constant head, constant flow, centrifuge) or unsteady state methods (outflow–inflow, instantaneous profile, thermal method) (Masrouri *et al*. 2008). Most of these methods are time-consuming and laborious, leading scientists to consider other methods such as conceptual models that could predict $K(h)$ from data obtained from the soil moisture retention curve and supportively coupled by K_s, measured independently at saturation, with the use of permeameters (Argyrokastritis *et al*. 2009).

For the needs of this application, we used the experimental data for Hygiene sandstone (Brooks and Corey 1964) adopted from van Genuchten (1980), in terms of relative hydraulic conductivity, K_r ($K_r = K/K_s$, K_s = 108 cm/d), as dependent variable for every h. The soil's moisture retention curve, $\Theta(h)$, was set as the explanatory function t. Therefore, for each point $K_r(h_i)$ for i = 1, …, 11, there was a corresponding value $\Theta(h_i)$ (Fig. 6).

The method was applied using as input the entire K_r dataset, as a general performance indicator; however, in order to examine the method's capability, we applied a cross-validation procedure by creating two additional subsets of the available K_r data and implementing the method for each case. The first subset comprised four data points numbered 1, 4, 8 and 11 (Fig. 6), which is the minimum amount of data points needed to implement the above mentioned procedure for obtaining a robust solution by minimizing the GCV (equation (28)). The second subset comprised only two points, the first and last of the dataset, points 1 and 11. The latter is an extreme case, since interpolating such complex variables with only the lower and higher boundaries known is a challenging task.

The explanatory dataset, $\Theta(h)$, was obtained by using the BLS method (Koutsoyiannis 2000), for m = 70 and τ = 0.01, to obtain 70 points from the 11 initial data points. Therefore, in each case the outcome of the BLSI method was 70 point estimates of K_r.

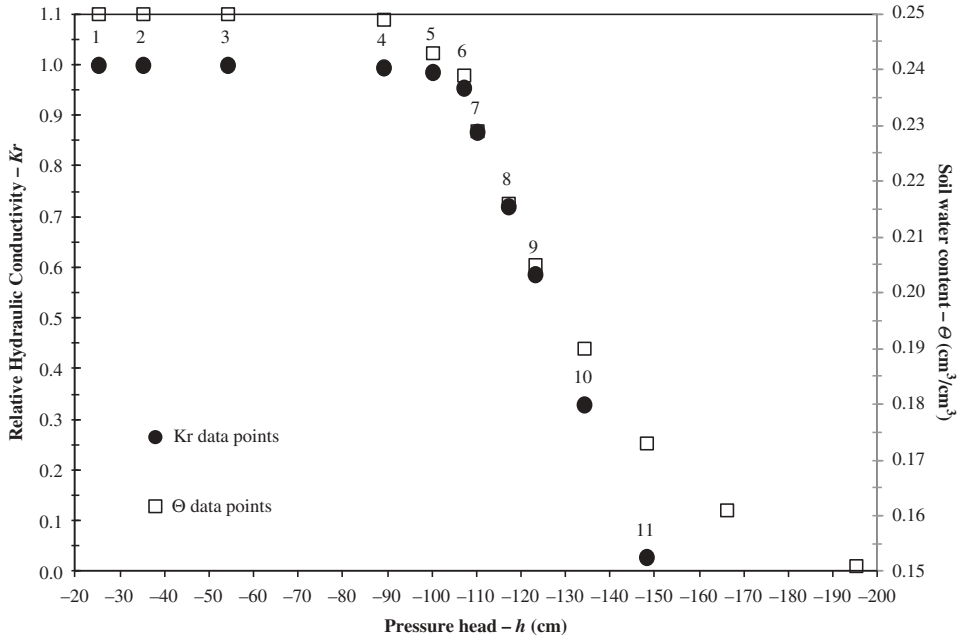

Fig. 6 Soil hydraulic properties of Hygiene sandstone (after van Genuchten 1980).

In the case of the entire K_r dataset and four available data points, the method was implemented for different numbers of segments, m ($2 \leq m \leq 30$) for each of which the GCV was minimized by altering τ_λ and τ_μ values. In this way, the global minimum GCV was reached. To obtain the optimum fit for the case of two available data points, we varied the m, τ_λ and τ_μ parameters and graphically assessed the results until the outcome was acceptable. The results of this procedure are presented in Fig. 7 where the 70 point estimates of $K_r(h)$ are presented as lines and in Table 1 together with the corresponding performance indices. Figure 7 demonstrates overall concurrence when using the entire dataset, but also an almost perfect fit is acquired in the case of four available data points.

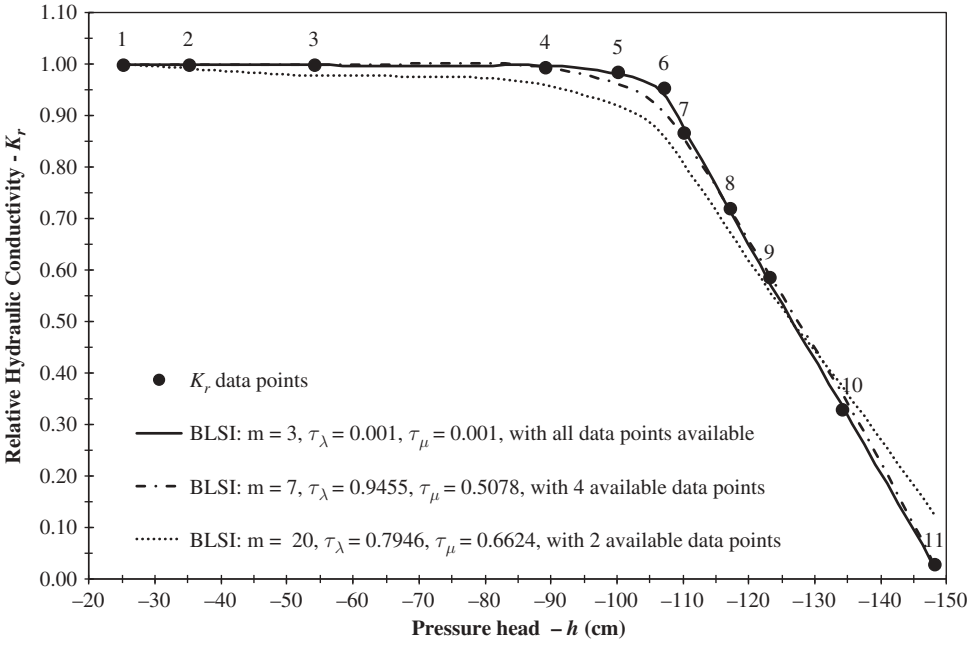

Fig. 7 BLSI fit to $K_r(h)$ data points, for different numbers of available data points.

Table 1 BLSI parameters and performance indices, for the $K_r(h)$ interpolation example.

Available data points	Optimum number of segments, m	Optimum τ_λ	Optimum τ_μ	MSE	Global minimum GCV	R^2*
All	3	0.001	0.001	2.06×10^{-5}	1.17×10^{-4}	1.000
Four (1, 4, 8, 11)	7	0.9455	0.5078	3.99×10^{-22}	5.041×0^{-5}	0.997
Two (1, 11)	20	0.7946	0.6624	4.72×10^{-3}	1.71×10^{-1}	0.971

* With respect to the entire dataset.

The performance indices presented in Table 1 confirm the efficiency of the BLSI method. Notable is the performance of the method considering the coefficient of determination, R^2, which was obtained for each case with respect to the entire dataset. In all three cases, the value of R^2 exceeded 0.97 and especially when all available data points were used in the interpolation process, where R^2 obtained its maximum possible value of 1.

In the case of two available points, i.e. the first and last of the dataset, BLSI performed amazingly well considering the limited amount of input data, presenting slight deviation, with acceptable magnitude, from the observed values.

Apart from the above mentioned example, the inverse problem was also studied, namely, the interpolation of the $\Theta(h)$ dataset by using $K_r(h)$ as the explanatory function t. As Table 2 and Fig. 8 show,

Table 2 BLSI parameters and performance indices, for the $\Theta(h)$ interpolation example.

Available data points	Optimum number of segments, m	Optimum τ_λ	Optimum τ_μ	MSE	Global minimum GCV	R^2*
All	8	0.01	0.9897	1.52×10^{-8}	1.02×10^{-6}	1.000
Four (1, 4, 8, 11)	16	0.9711	0.7498	8.51×10^{-26}	8.34×10^{-10}	0.987
Two (1, 11)	26	0.4307	0.8955	1.18×10^{-7}	5.45×10^{-2}	0.969

* With respect to the entire dataset

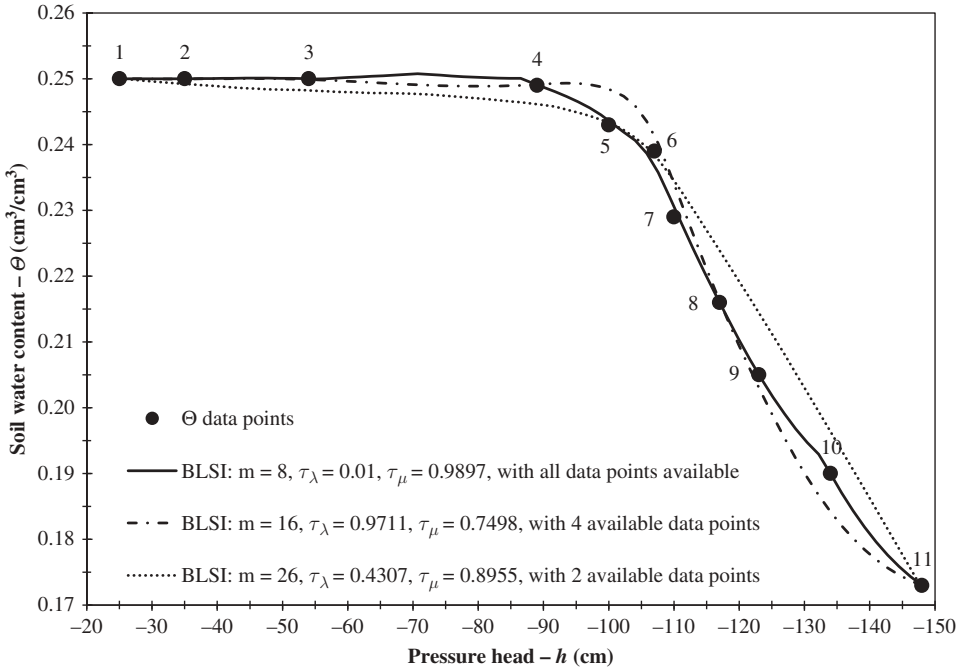

Fig. 8 BLSI fit to $\Theta(h)$, for different number of available data points.

the results were similar and very satisfactory, confirming the method's ability to interpolate scarce datasets of variables with complex relationships by utilizing denser, physically-related to them, explanatory datasets.

In the second real-world application we spatially interpolate annual rainfall using as explanatory variable the surface elevation. Spatial variability of precipitation is influenced by many factors, some of them connected to the chaotic nature of the atmospheric processes. At the annual scale, proximity to the sea and orography have significant effects (Goovaerts 2000). Hevesi *et al*. (1992a, 1992b) reported a significant correlation of 0.75 between average annual precipitation and elevation.

The objective of the application was: (a) to verify the applicability of the method against a hydrological variable with significant correlation to an easily measurable (hence available at considerably higher resolution) explanatory variable and (b) to verify the versatility of the method in terms of handling extensive datasets.

The study area was the region of Central Greece (Sterea Hellas) (Fig. 9(a)). The data consist of the mean rainfall at a network of 71 meteorological stations in the specified area, derived from all available measurements until the year 1992 (Christofides and Mamassis 1995). The surface elevation of the study area was obtained from the digital elevation model (DEM) SRTM Data Version 4.1 (Jarvis *et al*. 2008) and aggregated to a 2 km × 2 km grid (Fig. 9(a)) for practical and computational reasons, covering an area of approximately 25 620 km^2. The result was 6405 points of known elevation, which constituted the explanatory variable dataset.

Since the BLSI method is one-dimensional in terms of the independent variable x, the points' spatial coordinates (x_i, y_i) were projected onto a x' axis, according to the following expression:

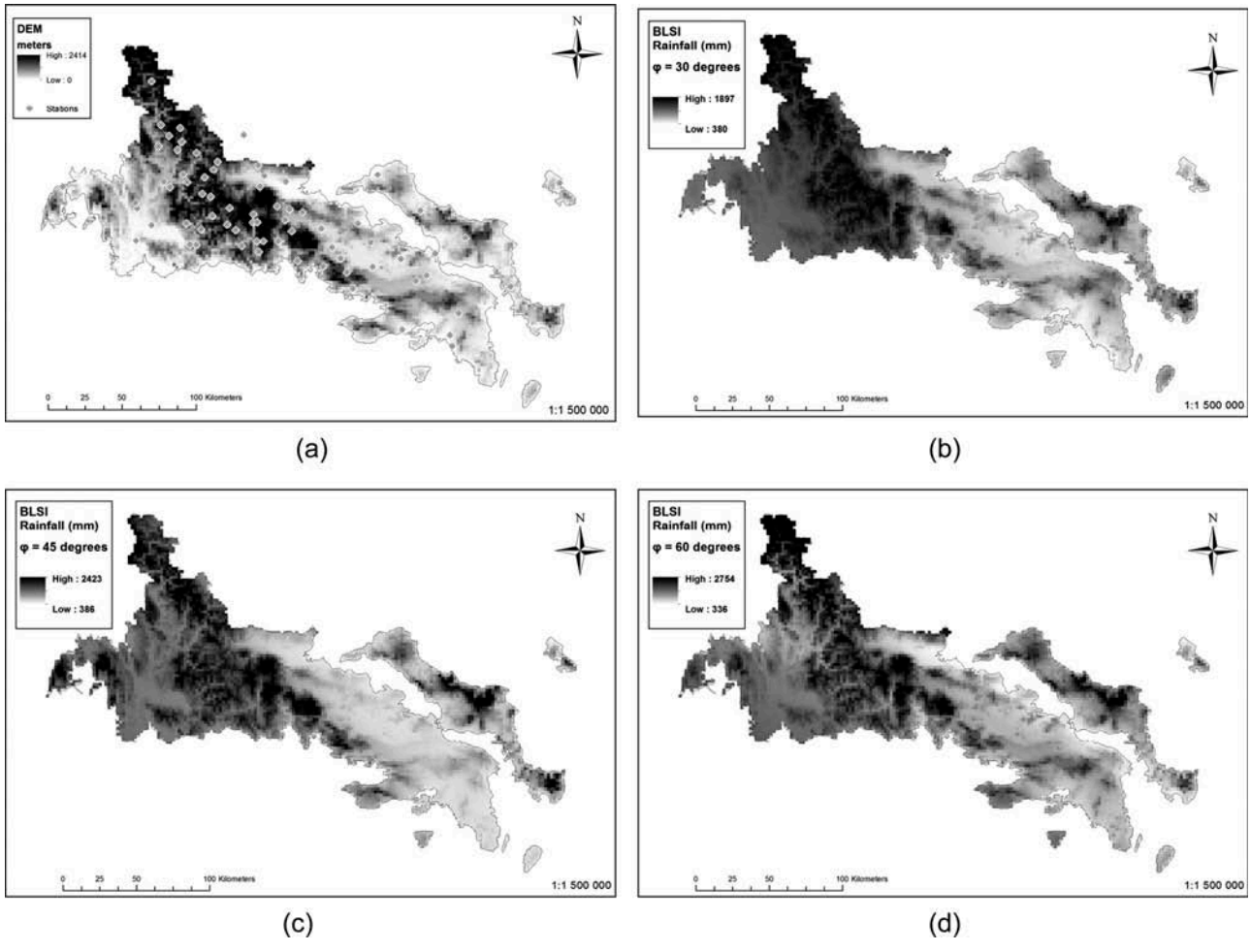

Fig. 9 (a) Elevation map and meteorological stations; (b)–(d) rainfall maps produced for the three cases of projection angles.

$$g(x_i, y_i) = g(x'_i), x'_i = x_i + \tan\varphi y_i \quad (31)$$

for alternative angles φ, namely:

$$\varphi = 30°, \tan\varphi = \frac{\sqrt{3}}{3}, x'_i = x_i + \frac{\sqrt{3}}{3} y_i \quad (32)$$

$$\varphi = 45°, \tan\varphi = 1, x'_i = x_i + y_i \quad (33)$$

$$\varphi = 60°, \tan\varphi = \sqrt{3}, x'_i = x_i + \sqrt{3} y_i \quad (34)$$

The global minimum of GCV for all three cases was obtained by implementing the method for a different number of segments, m ($2 \le m \le 30$) and minimizing GCV for each one by altering τ_λ and τ_μ. The results of the above procedure are presented in Table 3.

As a quality measure for the evaluation of the efficiency of the method, we utilized the minimum and maximum rainfall from the available meteorological stations, along with their corresponding elevation. Those values, compared to the minimum and maximum rainfall obtained from implementing BLSI for each of the three cases, are presented in Table 4.

Figure 9 and Table 4 indicate that the result of the method respects the dependence of rainfall on elevation (with increased elevation, rainfall increases, as happens in reality), confirming the efficiency of the BLSI method against the objectives set above. The validation of the obtained values, to conclude whether they could be regarded as acceptable for spatial interpolation of rainfall, exceeds the scope of the present study. Further investigation regarding the use of the specified methodology for spatial interpolation of rainfall in two dimensions, along with comparisons with other methods, will be reported in a future study.

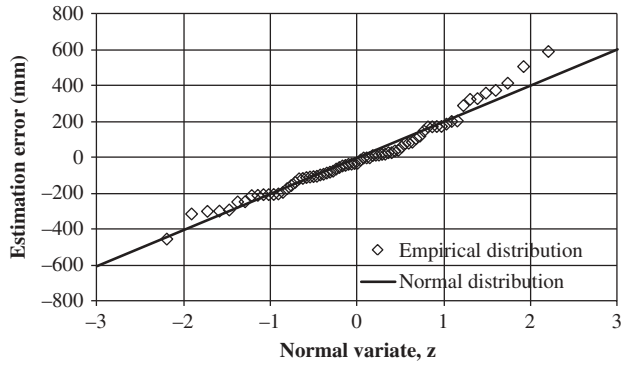

Fig. 10 Normal probability plot, in the case of $\varphi = 30°$, of the empirical distribution function of the estimation errors using Weibull plotting positions against normal distribution function N (0, 201.6).

Since the method's mathematical framework provides a direct means of evaluating interpolation errors across the available data points, an assessment of their distribution function could be of interest. Figure 10 demonstrates the normal probability plot of the empirical distribution function of the rainfall estimation error for $\varphi = 30°$. For comparison, the theoretical normal distribution function N (0, 201.6) was also plotted.

As can bee seen, Fig. 10 indicates that for the specific case of annual rainfall, the normal distribution is a good approximation of the interpolation errors produced by the implementation of BLSI. This remark does not constitute a generally valid conclusion. Nonetheless, the method provides a direct means to assess the distribution function of error, and hence the interpolation uncertainty, in a nonparametric manner without the need to hypothesize a specific distribution function. Similar plots

Table 3 BLSI parameters and performance indices, for the rainfall interpolation example.

Projection angle, φ	Optimum number of segments, m	Optimum τ_λ	Optimum τ_μ	MSE	Global minimum GCV
30°	13	0.141	0.986	4.01×10^4	5.63×10^4
45°	17	0.435	0.796	4.86×10^4	6.71×10^4
60°	8	0.282	0.984	5.74×10^4	7.41×10^4

Table 4 Comparison between meteorological stations data and BLSI results.

	Station data	Projection angle, φ		
		30°	45°	60°
Minimum rainfall (mm)/Elevation (m)	339/5	380/0	386/0	336/1
Maximum rainfall (mm)/Elevation (m)	1990/1420	1897/2414	2423/1310	2754/2012

were made for the other two cases studied, namely $\varphi = 45°$ and $\varphi = 60°$ (not shown here) and the results were found to be analogous to those for $\varphi = 30°$.

CONCLUSIONS

An innovative method is described which can be utilized to perform various interpolation tasks, by incorporating, in an objective manner, an explanatory variable available at a considerably denser dataset than the main variable. The technique incorporates smoothing terms with adjustable weights, defined by means of the angles formed by the consecutive segments of two broken lines into a piecewise linear regression model with known break points.

Apart from the demonstration of the mathematical framework, the method was illustrated and tested against three applications; a theoretical one with synthetic data from a known generating function, and two real-world examples: the interpolation of hydraulic conductivity function using water retention data as explanatory variable and *vice versa*, and the spatial interpolation of rainfall data using the surface elevation as explanatory variable. In every case, the method's efficiency to perform interpolation as well as smoothing between data points that are interrelated in a complicated manner, by incorporating the explanatory variable, was confirmed, indicating its applicability for diverse scientific and engineering tasks. A notable property of the proposed method is the fact that the resolution (length of consecutive segments of the broken line) does not necessarily have to coincide with that of the given data points, but it can be either finer or coarser, depending on the specific requirements of the problem of interest. This is an important property that makes the method applicable and reliable even in the case of scarce datasets (e.g. with as few as two points, as in the second case study, the method gave amazingly good results).

The third application showed that the method can be useful in spatial interpolation. However, the current formulation is not fully two-dimensional, although the general methodology allows extension in many dimensions. The extension of the methodology for spatial (two-dimensional) interpolation of variables will be reported in future studies.

Acknowledgements We thank the Associate Editor Alin Carsteanu, the eponymous reviewer Alberto Montanari and an anonymous reviewer for their comments and suggestions, which helped us to improve the presentation of the study.

REFERENCES

Argyrokastritis, I., Kargas, G., and Kerkides, P., 2009. Simulation of soil moisture profiles using K(h) from coupling experimental retention curves and one-step outflow data. *Water Resources Management*, 23 (15), 3255–3266. doi:10.1007/s11269-009-9432-3

Brooks, R.H. and Corey, A.T., 1964. Hydraulic properties of porous media, *Hydrology Papers*, 3, Colorado State University, Fort Collins, CO.

Burrough, P.A. and McDonnell, R.A., 1998. *Principles of geographical information systems*. Oxford: Oxford University Press, 333 pp.

Christofides, A. and Mamassis, N., 1995. Hydrometeorological data processing, *Evaluation of Management of the Water Resources of Sterea Hellas - Phase 2*. Athens: Department of Water Resources, Hydraulic and Maritime Engineering - National Technical University of Athens, Report 18, 268 pages. September 1995.

Craven, P. and Wahba, G., 1978. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31 (4), 377–403. doi:10.1007/BF01404567

Davis, C.J., 1986. *Statistics and data analysis in geology*. 2nd ed. New York: John Wiley & Sons.

Goovaerts, P., 1997. *Geostatistics for natural resources evaluation*. New York, NY: Oxford University Press, 483 pp.

Goovaerts, P., 2000. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology*, 228 (1–2), 113–129. doi:10.1016/S0022-1694(00)00144-X

Hengl, T., Heuvelink, G.B.M., and Rossiter, D.G., 2007. About regression-kriging: From equations to case studies. *Computers & Geosciences*, 33 (10), 1301–1315. doi:10.1016/j.cageo.2007.05.001

Hevesi, J.A., Istok, J.D., and Flint, A.L., 1992a. Precipitation estimation in mountainous terrain using multivariate geostatistics. Part I: structural analysis. *Journal Applied Meteor*, 31, 661–676. doi:10.1175/1520-0450(1992)031<0661:PEIMTU>2.0.CO;2

Hevesi, J.A., Flint, A.L., and Istok, J.D., 1992b. Precipitation estimation in mountainous terrain using multivariate geostatistics. Part II: Isohyetal maps. *Journal Applied Meteor*, 31, 677–688. doi:10.1175/1520-0450(1992)031<0677:PEIMTU>2.0.CO;2

Jarvis, A., *et al*., 2008. *Hole-filled SRTM for the globe Version 4* [online]. International Centre for Tropical Agriculture (CIAT). Available from: http://srtm.csi.cgiar.org [Accessed 16 June 2011].

Koutsoyiannis, D., 2000. Broken line smoothing: a simple method for interpolating and smoothing data series. *Environmental Modelling & Software*, 15 (2), 139–149. doi:10.1016/S1364-8152(99)00026-2

Leij, F.J., Russell, W.B., and Lesch, S.M., 1997. Closed-form expressions for water retention and conductivity data. *Ground Water*, 35 (5), 848–858. doi:10.1111/j.1745-6584.1997.tb00153.x

Li, J. and Heap, A.D., 2008. *A review of spatial interpolation methods for environmental scientists*. Geoscience Australia, Record 2008/23, 137 pp.

Masrouri, F., Bicalho, K.V., and Kawai, K., 2008. Laboratory hydraulic testing in unsaturated soils. *Geotechnical and Geological Engineering*, 26 (6), 691–704. doi:10.1007/s10706-008-9202-7

Van Genuchten, M.T., 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal*, 44, 892–898. doi:10.2136/sssaj1980.03615995004400050002x

Wahba, G. and Wendelberger, J., 1980. Some new mathematical methods for variational objective analysis using splines and cross validation. *Monthly Weather Review*, 108 (8), 1122–1143. doi:10.1175/1520-0493(1980)108<1122:SNMMFV>2.0.CO;2

Wang, H., Liu, G., and Gong, P., 2005. Use of cokriging to improve estimates of soil salt solute spatial distribution in the Yellow River delta. *Acta Geographica Sinica*, 60 (3), 511–518.