

## Fuzzy filtering for robust bioconcentration factor modelling

Shefali Kumar<sup>a</sup>, Mohit Kumar<sup>b,\*</sup>, Kerstin Thurow<sup>b</sup>, Regina Stoll<sup>c</sup>, Udo Kragl<sup>a</sup>

<sup>a</sup> Institute of Chemistry, University of Rostock, Albert Einstein Strasse-3a, D-18059 Rostock, Germany

<sup>b</sup> Center for Life Science Automation, F-Barnewitz-Strasse 8, D-18119 Rostock, Germany

<sup>c</sup> Institute of Preventive Medicine, University of Rostock, St.-Georg-Strasse 108, D-18055 Rostock, Germany

### ARTICLE INFO

#### Article history:

Received 30 July 2007

Received in revised form 30 April 2008

Accepted 3 May 2008

Available online 20 June 2008

#### Keywords:

Bioconcentration factor

Fuzzy filter

Robustness

QSAR model

### ABSTRACT

This study introduces a fuzzy filtering based technique for rendering robustness to the modelling methods. We consider a case study dealing with the development of a model for predicting the bioconcentration factor (BCF) of chemicals. The conventional neural/fuzzy BCF models, due to the involved uncertainties, may have a poor generalization performance (i.e. poor prediction performance for new chemicals). Our approach to improve the generalization performance of neural/fuzzy BCF models consists of (1) exploiting a fuzzy filter to filter out the uncertainties from the modelling problem, (2) utilizing the information about uncertainties, being provided by the fuzzy filter, for the identification of robust BCF models with an increased generalization performance. The approach has been illustrated with a data set of 511 chemicals (Dimitrov, S., Dimitrova, N., Parkerton, T., Comber, M., Bonnell, M., Mekenyan, O., 2005. Base-line model for identifying the bioaccumulation potential of chemicals. SAR and QSAR in Environmental Research 16 (6), 531–554) taking different types of neural/fuzzy modelling techniques.

© 2008 Elsevier Ltd. All rights reserved.

### 1. Introduction

Fuzzy systems by virtue of their uncertainties' handling capabilities have much to offer in the field of environmental modelling. Several studies, applying fuzzy techniques in the environmental problems, can be found in the literature. Some of the recent studies applying fuzzy methods for environmental problems include Fisher (2006), Li et al. (2007), Schlüter and Rüger (2007), Fleming et al. (2007), Nguyen et al. (2007), Lad et al. (2008), Wieland and Mirschel (2008) and Nasiri and Huang (2008). A fuzzy logic framework was presented in Fisher (2006) for environmental decisions. An integrated fuzzy-stochastic modelling approach to quantify both probabilistic and fuzzy uncertainties associated with the problem of risk assessment of groundwater contamination was introduced in Li et al. (2007). The work of Fleming et al. (2007) applied a fuzzy model to predict the cholera outbreak risk potential in southern Africa. Fuzzy set theory was used in Nguyen et al. (2007) to test an integrated water systems model. Lad et al. (2008) outlined an approach using fuzzy multicriteria decision making for environmental pollution potential ranking of industries. Fuzzy models and neural networks were discussed in Wieland and Mirschel (2008) with application to an estimation of regional yield of agricultural crops. The authors in Nasiri and Huang (2008) presented a fuzzy based methodology for environmental performance assessment of waste recycling programs.

Bioconcentration refers to the process of accumulation of chemicals in an aquatic organism as a result of exposure of the organism to a chemical concentration in the water via non-dietary routes. The extent of chemical bioconcentration is expressed in terms of bioconcentration factor (BCF) defined as the ratio of the chemical concentration in the organism to that in water (Mackay and Fraser, 2000). The BCF is a measure of the tendency of a substance to bioconcentrate in aquatic organisms. For an assessment of the bioaccumulation potential of chemicals, BCF in marine or freshwater organisms is traditionally used as an indicator. A flow-through method (European Centre for Ecotoxicology and Toxicology of Chemicals, 1995) is used for an experimental determination of BCF. The guidelines for characterizing potential bioconcentration in fish under flow-through conditions are provided in Organization for Economic Cooperation and Development (1994). A method suitable for very hydrophobic chemicals has been outlined in Gobas and Zhang (1995).

The motivation for developing the computer models for predicting the BCF of chemicals is derived from the fact that the experimental measurements are time-consuming, expensive, and not feasible for many thousands of chemicals that are of potential regulatory interest. Another motivation of BCF modelling is due to the ethical issues involving animal testing. Many studies aiming at the prediction of BCF values, based on Quantitative Structure–Activity Relationship (QSAR) approach, have appeared in the literature (Dearden, 2004). Typically, the models that map the hydrophobicity ( $\log K_{ow}$ ) of the chemicals to their ( $\log$  BCF) values are developed. Several modelling approaches including linear BCF

\* Corresponding author. Tel.: +49 381 4949956; fax: +49 381 4949952.  
E-mail address: [mohit.kumar@uni-rostock.de](mailto:mohit.kumar@uni-rostock.de) (M. Kumar).

models (Veith et al., 1979; Veith and Kosian, 1983; Mackay, 1982), bilinear BCF model (Bintein et al., 1993), polynomial BCF model (Connell and Hawker, 1988), fragment based additive BCF model (Meylan et al., 1999), nonlinear empirical model (Dimitrov et al., 2002) can be found in the literature. The researchers, in addition to the  $\log K_{ow}$  based modelling, also examined the BCF models based on solubility in octanol (Banerjee and Baughman, 1991), models based on aqueous solubility (Kenaga and Goring, 1980; Davies and Dobbs, 1984; Isnard and Lambert, 1988), models based on linear solvation energy relationships (Park and Lee, 1993; Ivanciuc, 1998), models based on connectivity indices (Lu et al., 2000), models based on fragment constants (Tao et al., 2000), models based on quantum chemical descriptors (Wei et al., 2001), models based on diverse theoretical descriptors (Gramatica and Papa, 2003, 2005; Dearden and Shinnawei, 2004).

The most important concern in BCF modelling is to generate a model with good generalization capability (i.e. good prediction performance of the model for an unseen compound not included in the training data set). The generalization performance of a modelling technique would be affected by number of factors including the choice of descriptors (model inputs), model type (linear, neural, fuzzy, etc.), model structure (number of free parameters to be adjusted), noise present in experimentally measured training data, and the model training algorithm. A non-robust modelling method typically shows good performance on the training compounds and poor performance on the testing compounds. This study is meant to improve the generalization capabilities of a modelling technique via providing a robustness against aforementioned factors which may affect adversely the generalization performance of a modelling technique. In the text, these factors would be represented mathematically by introducing a variable (termed as uncertainty) in the input–output model mappings.

We consider the problem of developing a model  $M$  whose inputs  $(x_1, \dots, x_n)$  are the numerical values of the chosen descriptors and output is the  $\log$  BCF value. The modelling problem consists of identifying a set of model parameters  $(p_1, p_2, \dots)$  using available data such that

$$\log \text{BCF} = M(x, p) + n, \quad (1)$$

where  $x = [x_1 \dots x_n]^T \in R^n$  is the input vector,  $p = [p_1 p_2 \dots]^T$  is the parameters vector that characterizes the model  $M$ , and  $n$  is the underlying uncertainty due to the non-optimal choice of the chosen inputs, non-optimal structure of the model  $M$ , non-optimal number of model parameters, noise in the experimentally measured data, and so on. The modelling methods identify the parameters vector  $p$ , that matches the model output to the  $\log$  BCF value in some “optimal manner”, using the available data of  $N$  compounds (i.e.  $\{x(k), \log \text{BCF}(k)\}_{k=1}^N$ ). The different choices of model  $M$  and the parameters “identification criteria” lead to the different modelling techniques.

The uncertainty  $n$  is the root cause of the poor generalization performance of the model. To improve the generalization performance of QSAR models, Bayesian regularization has been suggested in Burden and Winkler (1999a,b), Burden et al. (2000) and Winkler (2004). Regularization is a general method of improving generalization of the identified model via converting the identification problem into a “well-posed” problem. However, the choice of regularization parameters is usually not obvious. Bayesian regularization makes some stochastic assumptions on the nature of uncertainties and provides an optimal value of regularization parameters (MacKay, 1992). If these assumptions are not met, the identification performance may not be optimal. Recently, we have proposed in Kumar et al. (2007a) a fuzzy based method that takes into account the uncertainties without making any assumptions

about their nature and thus leads to QSAR models with an improved generalization performance.

Given the training compounds data  $\{x(k), y(k)\}_{k=1}^N$ ,  $y(k) = \log \text{BCF}(k)$ , our approach to improve the generalization performance of the modelling techniques is based on the following ideas.

- (1) A fuzzy filter is constructed using data  $\{x(k), y(k)\}_{k=1}^N$  that would filter out any uncertainties arising due to the compounds behaving differently from the input–output data trend. For a compound, described by descriptors values  $x(k)$ , the fuzzy filter is used to obtain a filtered  $y(k)$  value, denoted as  $y_f(k)$ . That is,  $N$  data pairs  $\{x(k), y_f(k)\}_{k=1}^N$  follow, without an exception, a trend of input–output mappings. The uncertainty associated to the compound is assessed as  $\hat{n}_k = y(k) - y_f(k)$ .
- (2) The uncertainties  $\{\hat{n}_k\}_{k=1}^N$  and filtered output values  $\{y_f(k)\}_{k=1}^N$  are assumed to have been produced by a set of random sources. We estimate the parameters of these random sources via modelling the  $N$  number of 2-dimensional data points  $\{z_k = [y_f(k) \hat{n}_k]^T \in R^2\}_{k=1}^N$  using finite mixture models McLachlan and Basford (1988) and McLachlan and Peel (2000). That is, we estimate the parameters of a set of probability density functions such that each data point  $z_k$  is modelled as having been generated by one of the probabilistic models in the set.
- (3) The finite mixture modelling leads to the clustering of the data via identifying which source (i.e. probabilistic model) produced each data point. Assume that  $C$  different sources, with the known probability density functions, have been identified producing the data  $\{z_k\}_{k=1}^N$ .
- (4) The data points associated to a source could be used to train (i.e. develop) a local model. A local model  $M_i$  (associated to the  $i$ th source), if trained using a non-robust algorithm conventionally with data  $\{x(k), y(k)\}$ , may lead to a poor generalization performance. The reason being that in the training of model  $M_i$ , the data points associated to a higher magnitude of uncertainties might act as outliers and adversely affect the training of the model. Therefore, we want to train the models with some penalized data  $\{x(k), y_p^i(k)\}$ .
- (5) For any  $k$ th data point used in the training of model  $M_i$ , the output value  $y(k)$  is penalized (in a context of the  $i$ th source) for the magnitude of the uncertainty associated to the  $k$ th data point. This is done via defining a penalized output value  $y_p^i(k)$  such that  $y_p^i(k)$  is closer to  $y(k)$  for the data points being treated as “regular” (typically characterized by a lower magnitude of estimated uncertainties), while  $y_p^i(k)$  is closer to  $y_f(k)$  for the data points being treated as outliers (typically characterized by a higher magnitude of estimated uncertainties). To define the penalized value  $y_p^i(k)$ , we make use of the  $i$ th probabilistic model that provided information about the uncertainties.
- (6) A model  $M_i$  (associated to the  $i$ th source) is not trained conventionally using data  $\{x(k), y(k)\}$ , however, trained using penalized data  $\{x(k), y_p^i(k)\}$ . Now, for the data points (might be acting as outliers),  $y_p^i(k)$  is closer to  $y_f(k)$  (i.e. closer to a point free from uncertainties) and thus training the model using  $y_p^i(k)$  values should not adversely affect the training method.
- (7) Finally, the  $C$  different local models  $M_1, \dots, M_C$  are combined to estimate the final output.

Roughly speaking, our approach renders robustness in the identification of local models  $M_1, \dots, M_C$  via penalizing the data. The local models operate in the predefined regions. To penalize the data, as will be explained, we make use of the information about uncertainties provided by a fuzzy filter. The design of the fuzzy filter is based on an “energy-gain bounding approach” (Kumar

et al., 2006a) that filters out the uncertainties without making any statistical assumptions about the nature of uncertainties. The filter design criterion is to minimize the maximum possible value of energy-gain from uncertainties to the filtering errors. The maximum value of energy-gain (that will be minimized) is calculated over all possible finite uncertainties (Kumar et al., 2006a). Our approach is closely related to the method presented in Kumar et al. (2007b), where the fuzzy filtering based approach has been introduced. This study is different from that of Kumar et al. (2007b) in the following.

- (1) In Kumar et al. (2007b), the local models are developed in the partitions of 1-dimensional real line of filtered values. In this study we partition the 2-dimensional space of filtered values and uncertainties, since the information about uncertainties will be used to penalize the data.
- (2) The method of Kumar et al. (2007b) trains the models with the filtered data  $\{x(k), y_f(k)\}$  and thus there are no uncertainties (in the training data) that could adversely affect the training procedure. However, here we use the penalized data  $\{x(k), y_p^i(k)\}$  for the training of the models, offering the flexibility of “smooth switching” between  $\{x(k), y_f(k)\}$  (for regular data points) and  $\{x(k), y_f(k)\}$  (for outliers).
- (3) The approach of Kumar et al. (2007b), unlike this study, penalizes all the data points (regular as well as outliers) and thus is over conservative.

This text is organized as follows. Section 2 presents the mathematical theory followed by the details of our method in Section 3. A case study concerned with the modelling of the bioconcentration factor of chemicals is provided in Section 4. Finally, the concluding remarks are given.

## 2. A review of the fuzzy filtering theory

This section reviews the mathematical theory of a clustering based fuzzy filter from our previous works (Kumar et al., 2007a,b). The fuzzy filter establishes the mappings between descriptor values and the corresponding output (i.e. log BCF value) via creating different clusters in the descriptor input space and associate to each cluster the output value. The mappings between input descriptors values (denoted by a vector  $x = [x_1 x_2 \dots x_n]^T \in R^n$ ) and output value (denoted by a scalar  $y$ ) are defined using different fuzzy rules:

$$R_1 : \text{If } x \text{ belongs to a cluster having centre } c_1 \text{ then } y = \alpha^1,$$

$$\vdots$$

$$R_K : \text{If } x \text{ belongs to a cluster having centre } c_K \text{ then } y = \alpha^K,$$

where  $c_i \in R^n$  is the centre of  $i$ th cluster, and the values  $\alpha^1, \dots, \alpha^K$  are real numbers. Such clustering based fuzzy mappings have been introduced in Kumar et al. (2006b) and applied to QSAR studies in Kumar et al. (2007a,b). The degree, by which an  $n$ -dimensional vector  $x$  belongs to the  $i$ th cluster, can be defined by a fuzzy set, say  $A_i$ . Given a universe of discourse  $X$ , a fuzzy subset  $A_i$  of  $X$  is characterized by a mapping:

$$A_i : X \rightarrow [0, 1]$$

where for  $x \in X$ ,  $A_i(x)$  is a value in the closed interval  $[0,1]$  that represents the degree to which  $x$  belongs to  $A_i$  (i.e.  $i$ th cluster). This mapping is called as membership function of the fuzzy set. For a given input vector  $x$ , the output of the filter is calculated by aggregating the rules as

$$F(x) = \frac{\sum_{i=1}^K \alpha^i A_i(x)}{\sum_{i=1}^K A_i(x)} \tag{2}$$

The membership function  $A_i(x)$  is chosen based on some fuzzy clustering criterion. By the method of fuzzy  $c$ -means (FCM), the membership function  $A_i(x)$  must satisfy (Bezdek, 1981)

$$\sum_{x \in X} \sum_{i=1}^K A_i^{\tilde{m}}(x) \|x - c_i\|^2 \rightarrow \text{Minimum}, \quad \sum_{i=1}^K A_i(x) = 1$$

where  $\tilde{m} > 1$  is the fuzzifier and  $\|\cdot\|$  denotes the Euclidean norm. The membership function that minimizes this objective function for a given choice of cluster centres  $\{c_i\}_{i=1}^K$  follows as

$$FCM_i(x, c_1, \dots, c_K) = \begin{cases} \frac{1}{\sum_{j=1}^K \left( \frac{\|x - c_j\|^2}{\|x - c_i\|^2} \right)^{\frac{1}{\tilde{m}-1}}} & \text{for } x \in X \setminus \{c_j\}_{j=1, \dots, K}, \\ 1 & \text{for } x = c_i, \\ 0 & \text{for } x \in \{c_j\}_{j=1, \dots, K} \setminus \{c_i\}. \end{cases} \tag{3}$$

A possibilistic approach for  $c$ -means clustering relaxes the unit sum constraint on the membership values so that  $A_i(x)$  better reflects the typicality of  $x$  to the  $i$ th cluster (Krishnapuram and Keller, 1993). Another approach called the noise clustering method has been introduced in Davé (1991) to deal with the noisy data. This approach considers noise a separate cluster such that membership of  $x$  to the noise cluster is defined as  $1 - \sum_{i=1}^K A_i(x)$  and the noise prototype is always at the same distance from every point in the data set. Another possible clustering criterion, assuming a noise cluster outside each data cluster, minimizes

$$J_c(A_i(x), c_1, \dots, c_K) = \sum_{x \in X} \sum_{i=1}^K \left[ A_i(x) \|x - c_i\|^2 + \{1 + A_i(x) \log A_i(x) - A_i(x)\} \delta_i \right]$$

where the second term in the objective function is intended to be a noise cluster. The term  $\{1 + A_i(x) \log A_i(x) - A_i(x)\}$  may be interpreted as the degree to which  $x$  does not belong to the  $i$ th cluster and thus the membership of  $x$  to the noise cluster. If the distance of  $x$  to the cluster centre  $c_i$  is greater than  $\sqrt{\delta_i}$ , then the minimization of  $J_c(\cdot)$  forces a small value of  $A_i(x)$  and a large value of membership of  $x_i$  to the noise cluster. Therefore, one of the strategies may be to set  $\delta_i$  equal to the distance of nearest cluster centre from  $c_i$ , i.e.  $\delta_i = \min \|c_j - c_i\|^2$ . Minimizing  $J_c(A_i(x), c_1, \dots, c_K)$  with respect to  $A_i(x)$  leads to the following expression for the membership function:

$$RC_i(x, c_1, \dots, c_K) = \exp\left(-\frac{\|x - c_i\|^2}{\delta_i}\right) \tag{4}$$

The membership functions of Eqs. (3) and (4) can be combined by adopting a mixed clustering criterion (Zhang and Leung, 2004; Pal et al., 2005). One way to do this is to assume that the membership function  $A_i$  has 2 components  $A_{1i}$  and  $A_{2i}$  such that

$$A_i = \frac{A_{1i}^{\tilde{m}}}{2} + \frac{A_{2i}}{2}$$

where  $A_{1i}, A_{2i}$  minimizes following constrained objective function:

$$\sum_{x \in X} \sum_{i=1}^K \left[ \left( \frac{A_{1i}^{\tilde{m}}(x) + A_{2i}(x)}{2} \right) \|x - c_i\|^2 + \left\{ 1 + A_{2i}(x) \log A_{2i}(x) - A_{2i}(x) \right\} \delta_i \right], \quad \sum_{i=1}^K A_{1i}(x) = 1.$$

Now,  $A_{1i}$  will be given by Eq. (3) and  $A_{2i}$  by Eq. (4). Thus,

$$A_i(x, c_1, \dots, c_K) = \frac{|\text{FCM}_i(x, c_1, \dots, c_K)|^{\bar{m}}}{2} + \frac{\text{RC}_i(x, c_1, \dots, c_K)}{2}. \quad (5)$$

For any membership function  $A_i(x)$ , defined by Eqs. (3), (4), or (5), if we define

$$G_i(x, c_1, \dots, c_K) = \frac{A_i(x, c_1, \dots, c_K)}{\sum_{i=1}^K A_i(x, c_1, \dots, c_K)},$$

then the output of the fuzzy filter follows from Eq. (2) as

$$F(x) = \sum_{i=1}^K \alpha^i G_i(x, c_1, \dots, c_K).$$

Introduce the notations:  $\alpha = [\alpha^i]_{i=1, \dots, K} \in \mathbb{R}^K$ ,  $\theta = [c_1^T \dots c_K^T]^T \in \mathbb{R}^{Kn}$ , and  $G(x, \theta) = [G_i(x, \theta)]_{i=1, \dots, K} \in \mathbb{R}^K$ , so that output of the fuzzy filter for an input  $x$  can be expressed as

$$F(x) = G^T(x, \theta)\alpha.$$

Thus, the fuzzy filter is characterized by 2 different parameters vectors:  $\alpha$  and  $\theta$ . For a development of the fuzzy filter, the parameters ( $\alpha, \theta$ ) must be identified using given input–output data set  $\{x(j), y(j)\}_{j=0}^k$ . Assume that there exist some true fuzzy filter, characterized by parameters ( $\alpha^*, \{\theta_j^*\}_{j=0}^k$ ), such that

$$y(j) = G^T(x(j), \theta_j^*)\alpha^* + n_j$$

where uncertainty  $n_j$  arises due to the non-optimal choice of the chosen inputs, non-optimal number of rules in the fuzzy filter, noise in the experimentally measured data  $y(j)$ , and so on. Let  $(\alpha_j, \theta_j)$  denote an estimate of ( $\alpha^*, \theta_j^*$ ) using data  $\{x(i), y(i)\}_{i=0}^j$  based on some recursive estimation strategy. The filtering error for  $j$ th-indexed data is given as

$$e_j = G^T(x(j), \theta_j^*)\alpha^* - G^T(x(j), \theta_j)\alpha_j.$$

Any estimation strategy will be considered performing good if it results in a small energy of filtering errors, being measured as  $\sum_{j=0}^k |e_j|^2$ . The performance of any estimation strategy will be affected by three kinds of unknown disturbances:

- the energy of uncertainties,  $\sum_{j=0}^k |n_j|^2$ ,
- deviation of initial guess  $\alpha_{-1}$  from true parameter  $\alpha^*$ , assessed as  $\|\alpha^* - \alpha_{-1}\|^2$ ,
- deviation of  $\{\theta_j^*\}_{j=0}^k$  from their initial guess  $\{\theta_{j-1}\}_{j=0}^k$ , assessed as  $\sum_{j=0}^k \|\theta_j^* - \theta_{j-1}\|^2$ . Here, we follow the approach of Kumar et al. (2006a), where the initial guess about  $\theta_j^*$  is taken equal to the estimate of  $\theta_{j-1}^*$ .

We are concerned with a robust identification method that is least sensitive to the disturbances. Our approach to the robust identification of fuzzy filter is based on energy-gain bounding criterion (Kumar et al., 2006a):

$$\min_{\{\alpha_j, \theta_j\}_{j=0}^k} \max_{\alpha^*, \{\theta_j^*\}_{j=0}^k, \{n_j\}_{j=0}^k} \times \frac{\sum_{j=0}^k |G^T(x(j), \theta_j^*)\alpha^* - G^T(x(j), \theta_j)\alpha_j|^2}{\mu^{-1}\|\alpha^*\|^2 + \mu_\theta^{-1} \sum_{j=0}^k \|\theta_j^* - \theta_{j-1}\|^2 \sum_{j=0}^k |n_j|^2}$$

where  $\mu$  and  $\mu_\theta$  are positive constants. The identification method minimizes the maximum possible value of energy-gain from disturbances to the filtering errors. Such an identification method will guarantee that *small disturbances cannot lead to large filtering errors*. The maximum value of energy-gain (that will be minimized) is

calculated over all possible finite disturbances without making any statistical assumptions about the nature of signals. It follows from Kumar et al. (2006a) that fuzzy filter parameters, based on energy-gain approach, are identified by performing for  $j=0, \dots, k$ , the recursions

$$\theta_j = \arg \min_{\theta} \left[ \frac{\|y(j) - G^T(x(j), \theta)\alpha_{j-1}\|^2}{1 + \mu\|G(x(j), \theta)\|^2} + \mu_\theta^{-1} \|\theta - \theta_{j-1}\|^2 \right],$$

$$\alpha_j = \alpha_{j-1} + \frac{\mu G(x(j), \theta_j) [y(j) - G^T(x(j), \theta_j)\alpha_{j-1}]}{1 + \mu\|G(x(j), \theta_j)\|^2}, \quad \alpha_{-1} = 0.$$

### 3. Methodology

Given the data of  $N$  training compounds  $\{x(k), y(k)\}_{k=1}^N$ , our approach to render robustness in a neural/fuzzy modelling technique consists of following steps.

#### 3.1. Identification of the parameters of a fuzzy filter

A fuzzy filter is identified based on the ideas outlined in Section 2. The identification method can be implemented using a Gauss–Newton based algorithm suggested in Appendix A. For a choice of the number of rules in the fuzzy filter (i.e. number of clusters  $K$ ) and initial guess about cluster centres  $\theta_{-1}$ , a clustering on input data (e.g. using finite mixture models (Figueiredo and Jain, 2002)) could be performed.

The output of the identified fuzzy filter represents the filtered output value. If we denote the parameters of identified fuzzy filter by  $(\alpha^l, \theta^l)$ , then the filtered output value of  $k$ th-indexed compound is given as

$$y_f(k) = G^T(x(k), \theta^l)\alpha^l. \quad (6)$$

The uncertainty associated to the  $k$ th-indexed compound will be assessed as

$$\hat{n}_k = y(k) - y_f(k). \quad (7)$$

#### 3.2. Gaussian mixture modelling of filtered data and uncertainties

Assume that the vector  $z_k = [y_f(k) \ \hat{n}_k]^T$  represents one particular outcome of a 2-dimensional random variable  $\mathbf{Z} \in \mathbb{R}^2$  whose probability density function can be written as a mixture of the Gaussian distributions:

$$p(z) = \sum_{i=1}^C a_i p(z|m_i, \Sigma_i), \quad (8)$$

such that

- the mixing probabilities  $a_1, \dots, a_C$  satisfy  $a_i \geq 0$  and  $\sum_{i=1}^C a_i = 1$ ,
- the parameters  $m_i \in \mathbb{R}^2$ ,  $\Sigma_i$  (a  $2 \times 2$  positive definite matrix) characterize fully the  $i$ th Gaussian component:

$$p(z|m_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^2 |\Sigma_i|}} \exp\left\{-\frac{1}{2}(z - m_i)^T \Sigma_i^{-1} (z - m_i)\right\}. \quad (9)$$

An approach to the clustering of data  $\{z_k\}_{k=1}^N$  is to fit finite mixture models (8) to the data, where a component distribution is used to model a specific cluster. That is,  $i$ th cluster (with mean  $m_i$  and covariance  $\Sigma_i$ ) is mathematically represented by Gaussian distribution  $p(z|m_i, \Sigma_i)$ . ‘‘Expectation–maximization’’ (EM) is the standard algorithm (McLachlan and Krishnan, 1997; McLachlan and Peel, 2000) used to fit finite mixture models to data. In this study, however, we use the algorithm of Figueiredo and Jain (2002) for estimating the parameters of the mixture (8). This algorithm is capable of automatically selecting the number of components  $C$ . The algorithm, unlike EM, is less sensitive to initialization and avoids the possibility of algorithm convergence to the boundary of the parameter space. As an illustration, Fig. 1 shows the Gaussian mixture modelling of an example data where the drawn ellipses are the level-curves of component distributions. The data points in Fig. 1 could be clustered via associating each point to 1 of the 5 components. The matrix  $\Sigma_i$  in Eq. (9) could be chosen to be a diagonal matrix (i.e. the 2 random variables are

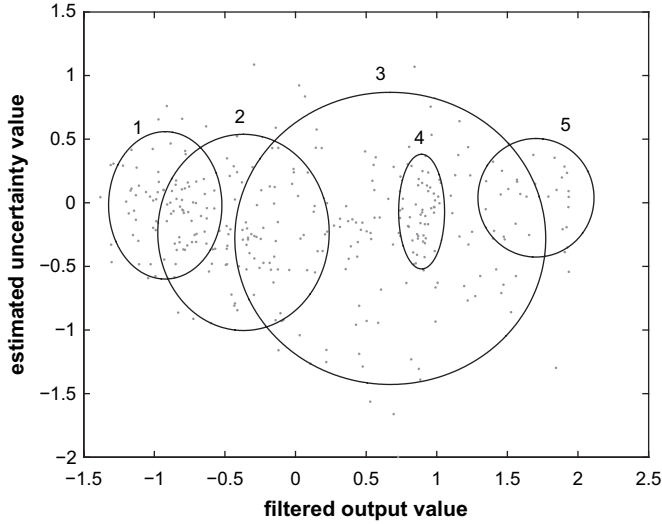


Fig. 1. Gaussian mixture modelling of data: data points and level-curves (solid line) for the different components.

independent). If  $m_i = \begin{bmatrix} m_i^1 \\ m_i^2 \end{bmatrix}$  and  $\Sigma_i = \begin{bmatrix} \Sigma_i^1 & 0 \\ 0 & \Sigma_i^2 \end{bmatrix}$ , then

$$p(z|m_i, \Sigma_i) = p(y_f|m_i^1, \Sigma_i^1) p(\hat{n}|m_i^2, \Sigma_i^2), \quad (10)$$

where

$$p(y_f|m_i^1, \Sigma_i^1) = \frac{\exp\left\{-\frac{(y_f - m_i^1)^2}{2\Sigma_i^1}\right\}}{\sqrt{2\pi\Sigma_i^1}}, \quad p(\hat{n}|m_i^2, \Sigma_i^2) = \frac{\exp\left\{-\frac{(\hat{n} - m_i^2)^2}{2\Sigma_i^2}\right\}}{\sqrt{2\pi\Sigma_i^2}} \quad (11)$$

The data points in Fig. 1 are taken from a case study to be discussed in Section 4.

### 3.3. A combination of local models

Given the knowledge of component distributions  $p(z|m_i, \Sigma_i, \dots), i = 1, \dots, C$ , we want to utilize this information in the development of neural, fuzzy, or of any other type local models ( $M_1, \dots, M_C$ ) valid in the predefined operating regions. The operating regions can be represented by fuzzy sets and the local models can be combined using a fuzzy rule base:

$R_1$ : For input  $x$ , if the filtered value  $y_f = G^T(x, \theta^1)\alpha^1$  is  $A_1(y_f)$ ,  
then output =  $M_1(x)$ ,  $[w_1]$

⋮

$R_C$ : For input  $x$ , if the filtered value  $y_f = G^T(x, \theta^C)\alpha^C$  is  $A_C(y_f)$ ,  
then output =  $M_C(x)$ ,  $[w_C]$

Here,  $(A_1(y_f), \dots, A_C(y_f))$  are the membership functions,  $M_i(x)$  denotes the  $i$ th model output for the input  $x$ , and  $w_i \in [0, 1]$  is the weight of the rule that represents the belief in the accuracy of the  $i$ th rule  $R_i$ . The degree of fulfillment of the  $i$ th rule is given by  $\beta_i(y_f) = w_i A_i(y_f)$ . The overall output  $y$ , for input  $x$ , is estimated by taking the weighted average of the output provided by each rule:

$$y = \frac{\sum_{i=1}^C w_i A_i(y_f) M_i(x)}{\sum_{i=1}^C w_i A_i(y_f)}$$

We want to define the membership function  $A_i(y_f)$  in such a way that the data points, belonging to the region covered by  $A_i(y_f)$ , are most likely to be generated by the  $i$ th probabilistic model  $p(y_f|m_i^1, \Sigma_i^1)$ . This is done by defining  $A_i(y_f)$  as follows

$$A_i(y_f) = k_i p(y_f|m_i^1, \Sigma_i^1), \quad i = 1, \dots, C \quad (12)$$

where  $k_i$  is a normalizing constant that ensures that  $A_i(y_f) \in [0, 1]$ . In view of this choice of the membership functions, the natural choice of the rule weight  $w_i$  is the prior probability of observing a data point from  $i$ th source i.e.  $w_i = a_i$ . Thus, the overall output by combining the local models is given as

$$y = \frac{\sum_{i=1}^C a_i p(y_f|m_i^1, \Sigma_i^1) M_i(x)}{\sum_{i=1}^C a_i p(y_f|m_i^1, \Sigma_i^1)} \quad (13)$$

### 3.4. The development of local models

One would normally expect to train a local model  $M_i$  (associated to fuzzy set  $A_i(y_f(k))$ ) with input–output data set  $D_i$  defined as

$$D_i = \left\{ x(k), y(k), 1 \leq k \leq N, A_i(y_f(k)) \geq \epsilon \right\}, \quad 0 \leq \epsilon \ll 1. \quad (14)$$

The data set  $D_i$  contains all those training compounds whose filtered output value belongs to fuzzy set  $A_i$  at least by a degree of  $\epsilon$ . As an illustration, the output values of set  $D_i$  have been displayed (marked as “.”) in Fig. 2. However, as stated earlier, the training of  $M_i$  with data set  $D_i$  using a non-robust algorithm may lead to a poor generalization performance of the model. The reason being that in the training of model  $M_i$ , the data points lying far away from  $i$ th cluster centre along the estimated-uncertainty-axis might act as outliers and adversely affect the training of the model. Therefore, we want to train the models with some penalized data  $\{x(k), y_p^i(k)\}$ . A penalized value  $y_p^i(k)$  is defined such that  $y_p^i(k)$  is closer to  $y(k)$  for the data points being treated as “regular” (lying closer to the  $i$ th cluster centre), while  $y_p^i(k)$  is closer to  $y_f(k)$  for the data points being treated as outliers (far away from  $i$ th cluster centre along the estimated-uncertainty-axis). Now, for the data points (might be acting as outliers),  $y_p^i$  is closer to  $y_f$  (i.e. closer to a point free from uncertainties) and thus training the model  $M_i$  using  $\{x(k), y_p^i(k)\}$  values should not adversely affect the training method.

Fig. 2 displays an example of the penalized values (marked as “o”), shifting from  $\{y(k)\}$  (marked as “.”) to the  $\{y_f(k)\}$  (marked as “+”), as we move away from the cluster centre along the estimated-uncertainty-axis. To define the penalized values, we make use of the information (provided by  $i$ th probabilistic model) on uncertainties. One of the possible methods for defining the penalized values is as follows:

$$y_p^i(k) = \bar{w}_k^i y_f(k) + (1 - \bar{w}_k^i) y(k), \quad (15)$$

where

$$\bar{w}_k^i = \left( 1 - \frac{p(\hat{n}_k|m_i^2, \Sigma_i^2)}{p_{\max}^i} \right)^{s_p}, \quad p_{\max}^i = \max_k p(\hat{n}_k|m_i^2, \Sigma_i^2), \quad s_p > 0. \quad (16)$$

Here,  $s_p$  is a “switching parameter” that controls the rate at which the switching of  $y_p^i$  from  $y$  to  $y_f$ , with a decrease in  $p(\hat{n}_k|m_i^2, \Sigma_i^2)$  (i.e. while moving away from  $i$ th cluster centre along the estimated-uncertainty-axis), takes place. A lower value of  $s_p$  results in a faster switching and vice-versa. Let  $D_i^p$  denotes the penalized training data set for  $M_i$ :

$$D_i^p = \left\{ x(k), y_p^i(k), 1 \leq k \leq N, A_i(y_f(k)) \geq \epsilon \right\}, \quad 0 \leq \epsilon \ll 1 \quad (17)$$

Finally, the data sets  $D_1^p, \dots, D_C^p$  could be used to train the local models  $M_1, \dots, M_C$ , respectively.

### 3.5. Implementation of the methodology for prediction

The given training data  $\{x(k), y(k)\}_{k=1}^N$  is used to estimate the parameters  $(\alpha^i, \theta^i)$ ,  $\{(m_i^1, \Sigma_i^1), (m_i^2, \Sigma_i^2), a_i, i = 1, \dots, C\}$  and thus the training of local models  $M_1, \dots, M_C$

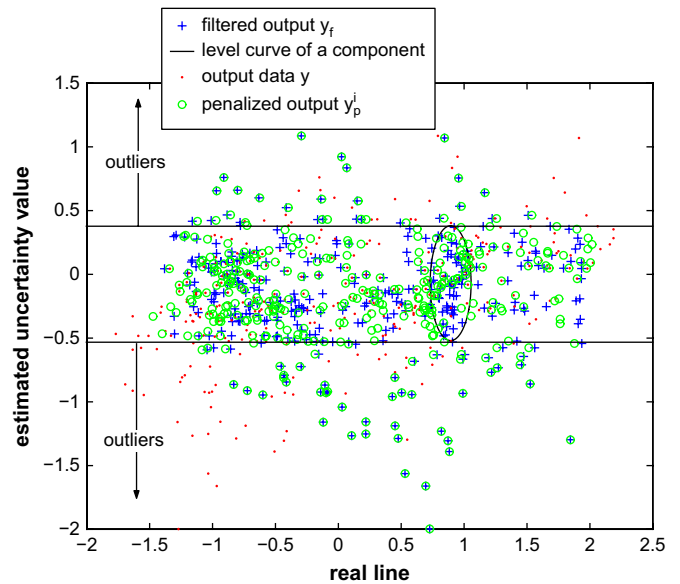


Fig. 2. Display of data output  $y$ , filtered output  $y_f$ , and penalized output  $y_p^i$ .

is accomplished. Now, the prediction of the output value for a given input (i.e. prediction of log BCF value of a compound that may or may not be included in training set) follows as

- for an input  $x$ , compute the filtered output  $y_f = G^T(x, \theta^l) \alpha^l$ ,
- the outputs of the local models could be combined to predict the output according to Eq. (13):

$$y = \frac{\sum_{i=1}^C a_i p(y_f | m_i^1, \Sigma_i^1) M_i(x)}{\sum_{i=1}^C a_i p(y_f | m_i^1, \Sigma_i^1)}, \quad p(y_f | m_i^1, \Sigma_i^1) = \frac{\exp\left\{-\frac{(y_f - m_i^1)^2}{2\Sigma_i^1}\right\}}{\sqrt{2\pi\Sigma_i^1}} \quad (18)$$

#### 4. Bioconcentration factor modelling

To illustrate our methodology, we consider the modelling of a BCF data set of 511 chemicals taken from Dimitrov et al. (2005). The data set includes following chemical classes: alkanes, alkenes, mono- and diaromatic hydrocarbons, polycyclic aromatic hydrocarbons (PAH), polychlorinated dibenzofuranes (PCDF), polychlorinated dibenzodioxines (PCDD), polychlorinated biphenyls (PCB), cycloalkanes and cycloalkenes, chloraromatic chemicals, perfluorinated acids (PFA) with 6–13 difluoromethylene functions in the chain, chlorinated biphenyl esters, aliphatic esters, chlororganic chemicals, aliphatic and aromatic N-containing compounds, polycyclic aromatic N-containing compounds, organotin compounds, and sulphur-containing heterocyclic compounds.

There is an uncertainty regarding the experimentally observed BCF values of chemicals in the data set. It was stated in Dimitrov et al. (2005) that the variations of 95% of the replicated experimentally observed BCF values are within 1.5 log units. Moreover, there are multiple BCF values associated to some chemicals in the data set. For example, the entry 24 (with log BCF = 2.83) and entry 76 (log BCF = 2.27) in the data set of Dimitrov et al. (2005) refer to the same chemical: 1,3,5-trimethylbenzene. Similarly entries 30 (log BCF = 2.72) and 150 (log BCF = 3.26) refer to the same chemical: anthracene. Also, hexachlorobenzene has been reported in the data set as entries 174 (log BCF = 4.21) and 218 (log BCF = 4.26). The uncertainties, as we will observe, affect adversely the performance of the BCF model.

##### 4.1. Generation of training and testing data

A large number of descriptors are available in the literature for QSAR studies. We calculate for our analysis several molecular descriptors of the compounds using E-DRAGON (Tetko et al., 2005). Out of the large number (several hundreds) of descriptors, a few descriptors, that serve as the inputs of the models for predicting the log BCF values, were chosen as in Kumar et al. (2007b):

- (1) descriptors with a standard deviation less than  $10^{-6}$  were rejected;
- (2) a pool of 20 descriptors, which showed highest absolute correlation with the log BCF values, were created for consideration for possible QSAR model inputs. This was done simply by calculating the values of correlation coefficients among the variables;
- (3) the method of “Principal Feature Analysis” (Cohen, 2000) was used to choose 5 descriptors out of the 20, which retain most of the information, both in the sense of maximum variability of the descriptors in the lower dimensional space and in the sense of minimizing the reconstruction error;
- (4) these 5 descriptors are
  - *H1v* (GETAWAY descriptor Consonni et al., 2002a,b): H autocorrelation of lag 1/weighted by atomic van der Waals volumes;
  - *MATS4v* (2D autocorrelation descriptor Moran, 1950): Moran autocorrelation - lag 4/weighted by atomic van der Waals volumes;

- *BLTD48* (molecular property): Verhaar model of Daphnia baseline toxicity for Daphnia (48 h) from MLOGP (mmol/l);
- *R5p* (GETAWAY descriptor Consonni et al., 2002a,b): R autocorrelation of lag 5/weighted by atomic polarizabilities;
- *TPSA(NO)* (molecular property Ertl et al., 2000): topological polar surface area using N, O polar contributions.

Our concern is not to optimize the choice of descriptors but to provide, for the chosen molecular descriptors, the improvements in the modelling performance using fuzzy filtering techniques.

The aim is to develop a QSAR model with these 5 descriptors as inputs and log BCF value as the output. The model will be trained with the data of around 2/3 of the total compounds and the remaining 1/3 compounds will be used for the testing of the model. The training and testing sets have been created as in Kumar et al. (2007b).

- (1) All descriptors and log BCF values are normalized to have zero mean and unit variance.
- (2) The point in the 6-dimensional space, whose coordinates correspond to the minimum values of 5 descriptors and log BCF value has been taken as the reference point.
- (3) The Euclidean distance of each compound from the reference point is calculated and all the compounds are arranged in the ascending order of their distances from the reference point.
- (4) Every third compound in the series of ascending order arranged compounds is taken as the testing compound and the remaining compounds as the training compounds.

This division of compounds into training and testing is meant for sandwiching of testing compounds between training ones in the sense of Euclidean distance.

##### 4.2. The issue of uncertainties

The BCF modelling problem is studied using a neural network and a fuzzy model. Let us first consider the training of a 3-layer feed-forward neural network. The first layer has 6 “tansig” (i.e. with hyperbolic tangent sigmoid transfer function) neurons, the second layer has 4 “tansig” neurons and the third layer 1 “purelin” (i.e. with linear transfer function) neuron. The network was initialized with random values of weights and biases. The network was trained using 2 different training algorithms: “scaled conjugate gradient backpropagation” (MATLAB Neural Network Toolbox command “trainscg”) and “Levenberg–Marquardt backpropagation” (MATLAB Neural Network Toolbox command “trainlm”). The training of the network stops if the number of epochs exceeds 10,000.

Also, a Sugeno type fuzzy model was trained using an in-built training algorithm in MATLAB Fuzzy Logic Toolbox (“anfis” command). The “anfis” algorithm combines the least-squares and backpropagation gradient descent method to identify the parameters of the fuzzy model. The structure of the fuzzy model was generated from the training data using subtractive clustering (MATLAB Fuzzy Logic Toolbox command “genfis2”). The fuzzy model was trained till 1000 epochs.

The modelling performance is assessed by computing the coefficient of determination ( $R^2$ ) and root mean squared error (RMSE) on training and testing data. Table 1 shows the performance of some of the standard neural/fuzzy modelling methods. We observe from Table 1 that the modelling techniques show good performance on the training data, however, poor performance on the testing data. This indicates the presence of uncertainties in the modelling problem for the chosen molecular descriptors, chosen model type and structure, training algorithms related chosen parameters, and so on. These uncertainties resulted in the over-training of the model and thus a poor generalization performance

**Table 1**  
The performance of some neural/fuzzy modelling methods

Method	$R^2$ -training	RMSE-training	$R^2$ -testing	RMSE-testing
“trainscg”	0.8924	0.4297	0.5596	0.9975
“trainlm”	0.9144	0.3831	0.5859	0.9298
“anfis”	0.8691	0.4739	0.4721	1.1631

(as shown by a poor performance on the testing data). One could argue for a decrease in the number of training compounds, however, our aim here is to highlight the issue of overtraining and a method for dealing with the overtraining issue.

#### 4.3. Rendering robustness in modelling methods via fuzzy filtering

We demonstrate that the proposed fuzzy filtering based methodology could be used for rendering robustness in the modelling methods against uncertainties. It will be seen that the training algorithms of Table 1, if used to train the local models with penalized data (as suggested in Section 3), would result in an improvement in the generalization performance.

We employed a fuzzy filter, with membership functions defined by Eq. (5) for  $\tilde{m} = 2$ , for filtering out the uncertainties from the modelling problem. The fuzzy filter parameters were identified based on the energy-gain bounding approach described in Section 2. The identification method was implemented in MATLAB 6.5 using a Gauss–Newton based algorithm proposed in Appendix A. The number of rules in the fuzzy filter (i.e.  $K$ ) and initial guess about cluster centres ( $\theta_{-1}$ ) were chosen via performing clustering on the 5-dimensional input training data using finite mixture models (Figueiredo and Jain, 2002). The identification algorithm was run till 100 epochs taking  $\mu = \mu_0 = 0.1$ . The identified fuzzy filter was used to obtain for the training compounds the filtered values (6) and the underlying uncertainties (7).

The Gaussian mixture modelling of the 2-dimensional data (filtered and uncertainties values) identified 5 different component distributions describing the behavior of the data. These 5 component distributions have been displayed in Fig. 1. Associated to these components, the penalized data sets  $D_1^p, \dots, D_5^p$  (obtained using Eq. (17)) could be used to train the local models  $M_1, \dots, M_5$ , respectively. The local models are finally combined using Eq. (18) to predict the overall output. Table 2 shows the system performance when the local models  $M_1, \dots, M_5$  are neural networks trained with the “trainscg” algorithm. Here,  $M_1, \dots, M_5$  have the same structure, initial conditions, and training parameters (e.g. number of epochs) as of the network trained with “trainscg” in Table 1. The parameter  $\epsilon$  in Eq. (17), to define the penalized data sets for different values of switching parameter  $s_p$ , was chosen as  $\epsilon = 0$ . In this text, we made no comment on the choice of switching parameter  $s_p$ , thus we consider the different values of switching parameter  $s_p$  ranging from 0.01 to 2.

Similarly, the Tables 3 and 4 show the performance of the “trainlm” and “anfis” algorithms, respectively, via proposed fuzzy filtering based technique.

**Table 2**  
The performance of “trainscg” network training algorithm via proposed technique

$s_p$	$R^2$ -training	RMSE-training	$R^2$ -testing	RMSE-testing
0.01	0.7916	0.6355	0.6725	0.8542
0.03	0.7933	0.6309	0.6835	0.8405
0.05	0.7970	0.6230	0.6682	0.8598
0.1	0.8036	0.6084	0.6734	0.8433
0.2	0.8072	0.5962	0.6915	0.8188
0.4	0.8139	0.5793	0.7017	0.8009
0.5	0.8196	0.5681	0.7329	0.7485
0.75	0.8269	0.5541	0.6918	0.8022
1	0.8277	0.5516	0.7109	0.7790
2	0.8420	0.5253	0.7180	0.7627

**Table 3**  
The performance of “trainlm” network training algorithm via proposed technique

$s_p$	$R^2$ -training	RMSE-training	$R^2$ -testing	RMSE-testing
0.01	0.7905	0.6366	0.6856	0.8371
0.03	0.7928	0.6306	0.6854	0.8346
0.05	0.7935	0.6282	0.6854	0.8314
0.1	0.8032	0.6077	0.6475	0.8721
0.2	0.8129	0.5879	0.7213	0.7752
0.4	0.8158	0.5747	0.7279	0.7484
0.5	0.8202	0.5672	0.7190	0.7676
0.75	0.8333	0.5444	0.6187	0.8920
1	0.8405	0.5315	0.7021	0.7821
2	0.8457	0.5198	0.6789	0.8127

A comparison of Tables 2–4 with Table 1, shown in Fig. 3, verifies that the generalization performance (i.e. testing data performance) of the modelling methods improved considerably via proposed approach. The type, structure, and training conditions of the local models in the studies are the same as of the models in Table 1. However, none of the modelling method resulted in the overtraining of the model via proposed fuzzy filtering based technique. This indicates that the robustness offered to the modelling methods is obviously a result of

- (1) penalizing the data;
- (2) combining the local models using a fuzzy rule base that has been carefully designed, based on Gaussian mixture modelling of filtered data and uncertainties.

#### 4.4. Robust training algorithm

If the chosen training algorithm is robust towards uncertainties, then the local model could be trained with data sets  $D_1, \dots, D_C$  defined by Eq. (14). Since the training algorithm is robust, there is no need of penalizing the training data. In this case, an improvement in the modelling performance could be still expected as a result of the fuzzy combination of local models. As an illustration, we consider the Bayesian regularized neural networks that have been accepted as a robust method of QSAR modelling (Burden and Winkler, 1999a,b; Burden et al., 2000; Winkler, 2004). The local models are trained with data sets  $D_1, \dots, D_5$  defined by Eq. (14) for  $\epsilon = 0.01$  using Bayesian regularized neural network training algorithm (MATLAB Neural Network Toolbox command “trainbr”). Table 5 illustrates the performance of a Bayesian regularized neural network on BCF modelling problem and an improvement (although slightly) to this as a result of the fuzzy combination of local models.

## 5. Concluding remarks

Several modelling methods have been proposed in the literature aiming at the good generalization performance of the models. This work, unlike many studies, doesn't propose a new modelling

**Table 4**  
The performance of “anfis” training algorithm via proposed technique

$s_p$	$R^2$ -training	RMSE-training	$R^2$ -testing	RMSE-testing
0.01	0.7966	0.6293	0.6802	0.8421
0.03	0.7976	0.6248	0.6812	0.8383
0.05	0.7987	0.6208	0.6825	0.8343
0.1	0.7958	0.6197	0.6824	0.8257
0.2	0.8068	0.5961	0.6860	0.8152
0.4	0.8115	0.5830	0.6677	0.8352
0.5	0.8130	0.5789	0.6851	0.8067
0.75	0.8155	0.5715	0.7050	0.7805
1	0.8196	0.5636	0.6947	0.7938
2	0.8294	0.5459	0.7017	0.7841

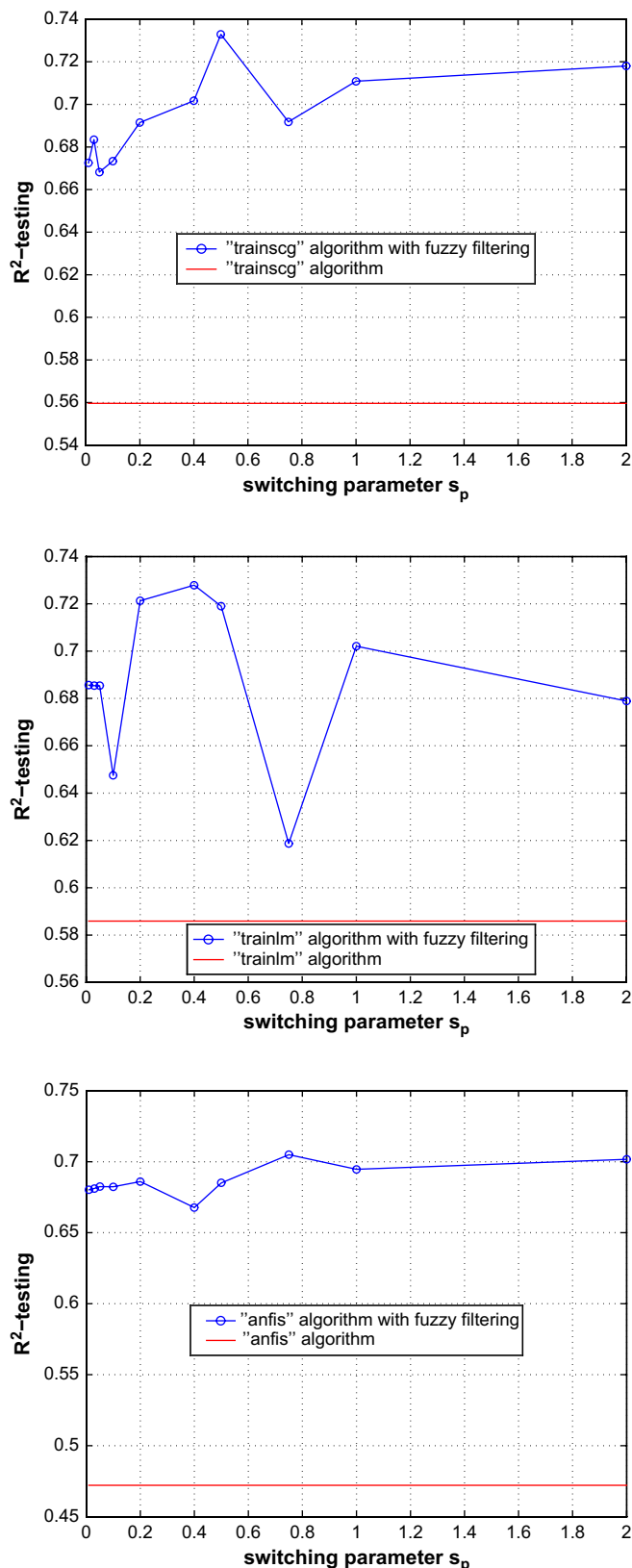


Fig. 3. An improvement in the generalization performance of the modelling methods via proposed approach.

method but provides a tool for rendering robustness in any modelling method. A case study dealing with the bioconcentration factor modelling of chemicals was provided to illustrate the effectiveness of our technique. The choice of the inputs (i.e. molecular

Table 5

The performance of Bayesian regularized neural network training algorithm

Method	$R^2$ -training	RMSE-training	$R^2$ -testing	RMSE-testing
"trainbr"	0.8731	0.4666	0.7112	0.7604
"trainbr" Via proposed method	0.8787	0.4579	0.7466	0.7129

descriptors), model type, model structure, and training algorithm are the critical issues that need to be addressed in solving a modelling problem. Our concern in this text was to improve the modelling performance once the descriptors, model type, model structure, and training algorithm have been chosen.

The uncertainties, affecting adversely the generalization capabilities of the modelling methods, are filtered using a fuzzy filter. Based on the available information about uncertainties, the local models are developed in a manner that uncertainties are not allowed to affect the training of the local models. This improves the generalization performance of a modelling technique. The combination of the local models using a fuzzy rule base (that has been carefully designed based on Gaussian mixture modelling of filtered data and uncertainties) provides additional tolerance towards uncertainties. One could observe in Fig. 3 a considerable improvement in the performance of the different modelling methods via proposed technique. However, there are some issues which remain to be addressed in our future work. The automatic selection of the value of switching parameter  $s_p$  is a part of our future work. Fortunately, the effectiveness of our approach has been observed at all considered values of  $s_p$  ranging from 0.01 to 2. For a choice  $s_p = 0$  (i.e. training of local models with filtered data), our technique becomes close to the method of Kumar et al. (2007b).

The aim of this study is to provide to the researchers a piece of software that would improve the robustness performance of their favourite modelling methods. A website (<http://www.fuzzymodeling.com>) is available to provide the users an online service for (1) a fuzzy filtering of the uncertainties from the data, (2) building a robust data model based on fuzzy filtering approach.

## Acknowledgements

We acknowledge "Deutsche Bundesstiftung Umwelt" for the financial support of Shefali Kumar. The work was partially financed by German Research Foundation (DFG) within the graduate school 1213: "Sustainability in Catalysis and Technique". We thank Prof. Ovanes Mekenyan (Laboratory of Mathematical Chemistry, "Prof. As. Zlatarov" University, 8010 Bourgas, Bulgaria) for providing us the BCF data of chemicals.

## Appendix A. A Gauss-Newton based algorithm

Given  $N$  input-output data pairs  $\{x(j), y(j)\}_{j=0}^{N-1}$ , to compute the parameters

$$\begin{aligned} \theta_j &= \arg \min_{\theta} \left[ \frac{[y(j) - G^T(x(j), \theta)\alpha_{j-1}]^2}{1 + \mu \|G(x(j), \theta)\|^2} + \mu_{\theta}^{-1} \|\theta - \theta_{j-1}\|^2 \right] \\ &= \arg \min_{\theta} \|r(\theta)\|^2, \text{ where } r(\theta) = \begin{bmatrix} \frac{[y(j) - G^T(x(j), \theta)\alpha_{j-1}]^2}{1 + \mu \|G(x(j), \theta)\|^2} \\ (\mu_{\theta}^{-1})^{1/2} (\theta - \theta_{j-1}) \end{bmatrix}, \end{aligned}$$

$$\alpha_j = \alpha_{j-1} + \frac{\mu G(x(j), \theta_j) [y(j) - G^T(x(j), \theta_j)\alpha_{j-1}]}{1 + \mu \|G(x(j), \theta_j)\|^2},$$

we suggest a Gauss–Newton based algorithm. The algorithm consists of following steps:

(1) Choose initial guess about cluster centres  $\theta_{-1}$ , number of maximum epochs  $E_{\max}$ ,  $\alpha_{-1} = 0$ , epoch count  $EC = 0$ , and data index  $j = 0$ .

(2) If  $EC < E_{\max}$ ,

(a) if  $j \leq (N - 1)$ ,

(i) define  $r(\theta) = \left[ \frac{[y(j) - G^T(x(j), \theta)\alpha_{j-1}]^2}{1 + \mu \|G(x(j), \theta)\|^2} \right]^{1/2}$  and let  $s^*(\theta)$  be the

unique solution of following linear least-squares problem:

$$s^* = \arg \min_s [\|r(\theta) + r'(\theta)s\|^2],$$

where  $r'(\theta)$  is the Jacobian matrix of vector  $r$  with respect to  $\theta$ , determined by the method of finite-differences. The Jacobian  $r'(\theta)$  is a full rank matrix, as a result of using regularization.

(ii) compute  $\theta_j = \theta_{j-1} + s^*(\theta_{j-1})$ .

(iii) compute

$$\alpha_j = \alpha_{j-1} + \frac{\mu G(x(j), \theta_j) [y(j) - G^T(x(j), \theta_j)\alpha_{j-1}]}{1 + \mu \|G(x(j), \theta_j)\|^2}.$$

(iv)  $j := j + 1$  and go to step 2(a).

(b)  $EC := EC + 1$ ,  $\alpha_{-1} := \alpha_{N-1}$ ,  $\theta_{-1} := \theta_{N-1}$ ,  $j = 0$ , and go to step 2.

## References

- Banerjee, S., Baughman, G.L., 1991. Bioconcentration factors and lipid solubility. *Environmental Science and Technology* 25, 536–539.
- Bezdek, J.C., 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York.
- Bintein, S., Devillers, J., Karcher, W., 1993. Nonlinear dependence of fish bioconcentration on *n*-octanol/water partition coefficients. *SAR and QSAR in Environmental Research* 1, 29–39.
- Burden, F.R., Ford, M.G., Whitley, D.C., Winkler, D.A., 2000. Use of automatic relevance determination in QSAR studies using Bayesian neural networks. *Journal of Chemical Information and Computer Sciences* 40, 1423–1430.
- Burden, F.R., Winkler, D.A., 1999a. New QSAR methods applied to structure–activity mapping and combinatorial chemistry. *Journal of Chemical Information and Computer Sciences* 39, 236–242.
- Burden, F.R., Winkler, D.A., 1999b. Robust QSAR models using Bayesian regularized neural networks. *Journal of Medicinal Chemistry* 42, 3183–3187.
- Cohen, I., 2000. *Automatic Facial Expression Recognition from Video Sequences Using Temporal Information*. Masters thesis, Univ. of Illinois at Urbana Champaign.
- Connell, D.W., Hawker, D.W., 1988. Use of polynomial expressions to describe the bioconcentration of hydrophobic chemicals by fish. *Ecotoxicology and Environmental Safety* 16, 242–257.
- Consonni, V., Todeschini, R., Pavan, M., 2002a. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *Journal of Chemical Information and Computer Sciences* 42, 682–692.
- Consonni, V., Todeschini, R., Pavan, M., Gramatica, P., 2002b. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. *Journal of Chemical Information and Computer Sciences* 42, 693–705.
- Davé, R.N., 1991. Characterization and detection of noise in clustering. *Pattern Recognition Letters* 12 (11), 657–664.
- Davies, R.P., Dobbs, A., 1984. The prediction of bioconcentration in fish. *Water Research* 18, 1253–1262.
- Dearden, J.C., 2004. QSAR modelling of bioaccumulation. In: Cronin, M.T.D., Livingstone, D.J. (Eds.), *Predicting Chemical Toxicity and Fate*. CRC Press LLC, Boca Raton, Florida.
- Dearden, J.C., Shinawei, N.M., 2004. Improved prediction of fish bioconcentration factor of hydrophobic chemicals. *SAR and QSAR in Environmental Research* 15, 449–455.
- Dimitrov, S., Dimitrova, N., Parkerton, T., Comber, M., Bonnell, M., Mekenyan, O., 2005. Base-line model for identifying the bioaccumulation potential of chemicals. *SAR and QSAR in Environmental Research* 16 (6), 531–554.
- Dimitrov, S.D., Mekenyan, O.G., Walker, J.D., 2002. Non-linear modeling of bioconcentration using partition coefficients for narcotic chemicals. *SAR and QSAR in Environmental Research* 13, 177–188.
- Ertl, P., Rohde, B., Selzer, P., 2000. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of Medicinal Chemistry* 43, 3714–3717.
- European Centre for Ecotoxicology and Toxicology of Chemicals, 1995. *The Role of Bioaccumulation in Environmental Risk Assessment: the Aquatic Environment and Related Food Webs*. Technical Report 67, Brussels, Belgium.
- Figueiredo, M.A.T., Jain, A.K., 2002. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (3), 381–396.
- Fisher, B.E.A., 2006. Fuzzy approaches to environmental decisions: application to air quality. *Environmental Science and Policy* 9 (1), 22–31.
- Fleming, G., Merwe, M., McFerren, G., 2007. Fuzzy expert systems and GIS for cholera health risk prediction in southern Africa. *Environmental Modelling and Software* 22 (4), 442–448.
- Gobas, F.A.P.C., Zhang, X., 1995. Measuring bioconcentration factors and rate constants of chemicals in aquatic organism under condition of variable water concentration and short exposure time. *Chemosphere* 25, 1961–1971.
- Gramatica, P., Papa, E., 2003. QSAR modelling of bioconcentration factor by theoretical molecular descriptors. *QSAR and Combinatorial Science* 22, 374–385.
- Gramatica, P., Papa, E., 2005. An update of the BCF QSAR model based on theoretical molecular descriptors. *QSAR and Combinatorial Science* 24, 953–960.
- Insnard, P., Lambert, S., 1988. Estimating bioconcentration factors from octanol–water partition coefficient and aqueous solubility. *Chemosphere* 17, 21–34.
- Ivanciuc, O., 1998. Artificial neural networks applications. Part 7. Estimation of bioconcentration factors in fish using solvatochromic parameters. *Revue Roumaine de Chimie* 43, 347–354.
- Kenaga, E.E., Goring, C.A.L., 1980. Relationship between water solubility and soil sorption, octanol–water partitioning and bioconcentration of chemicals in biota. In: *Aquatic Toxicology. Special Technical Publication 707*. American Society for Testing and Materials, Philadelphia, PA, pp. 78–115.
- Krishnapuram, R., Keller, J.M., May 1993. A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems* 1, 98–110.
- Kumar, M., Stoll, N., Stoll, R., May 2006a. An energy-gain bounding approach to robust fuzzy identification. *Automatica* 42 (5), 711–721.
- Kumar, M., Stoll, R., Stoll, N., Apr 2006b. A min–max approach to fuzzy clustering, estimation, and identification. *IEEE Transactions on Fuzzy Systems* 14 (2), 248–262.
- Kumar, M., Thurow, K., Stoll, N., Stoll, R., 2007a. Robust fuzzy mappings for QSAR studies. *European Journal of Medicinal Chemistry* 42, 675–685.
- Kumar, S., Kumar, M., Stoll, R., Kragl, U., 2007b. Handling uncertainties in toxicity modelling using a fuzzy filter. *SAR and QSAR in Environmental Research* 18 (7–8), 645–662.
- Lad, R.K., Desai, N.G., Christian, R.A., Deshpande, A.W., 2008. Fuzzy modeling for environmental pollution potential ranking of industries. *Environmental Progress* 27 (1), 84–90.
- Li, J., Huang, G.H., Zeng, G., Maqsood, I., Huang, Y., 2007. An integrated fuzzy-stochastic modeling approach for risk assessment of groundwater contamination. *Journal of Environmental Management* 82 (2), 173–188.
- Lu, X., Tao, S., Hu, H., Dawson, R.W., 2000. Estimation of bioconcentration factors of nonionic organic compounds in fish by molecular connectivity indices and polarity correction factors. *Chemosphere* 41, 1675–1688.
- Mackay, D., 1982. Correlation of bioconcentration factors. *Environmental Science and Technology* 16, 274–278.
- Mackay, D., Fraser, A., 2000. Bioaccumulation of persistent organic chemicals: mechanisms and models. *Environmental Pollution* 110, 375–391.
- MacKay, D.J.C., 1992. Bayesian interpolation. *Neural Computation* 4, 415–447.
- McLachlan, G., Basford, K., 1988. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York.
- McLachlan, G., Krishnan, T., 1997. *The EM Algorithm and Extensions*. John Wiley & Sons, New York.
- McLachlan, G., Peel, D., 2000. *Finite Mixture Models*. John Wiley & Sons, New York.
- Meylan, W.M., Howard, P.H., Boethling, R.S., Aronson, D., Printup, H., Gouchie, S., 1999. Improved method for estimating bioconcentration/bioaccumulation factor from octanol/water partition coefficient. *Environmental Toxicology and Chemistry* 18, 664–672.
- Moran, P.A.P., 1950. Notes on continuous stochastic phenomena. *Biometrika* 37, 17–23.
- Nasiri, F., Huang, G., 2008. A fuzzy decision aid model for environmental performance assessment in waste recycling. *Environmental Modelling and Software* 23 (6), 677–689.
- Nguyen, T.G., de Kok, J.L., Titus, M.J., 2007. A new approach to testing an integrated water systems model using qualitative scenarios. *Environmental Modelling and Software* 22 (11), 1557–1571.
- Organization for Economic Cooperation and Development, 1994. *Bioconcentration: flow-through fish test*. In: *OECD Guide-line for Testing of Chemicals: Draft Guideline* 305, Paris, France.
- Pal, N.R., Pal, K., Keller, J.M., Bezdek, J.C., Aug 2005. A possibilistic fuzzy *c*-means clustering algorithm. *IEEE Transactions on Fuzzy Systems* 13 (4), 517–530.
- Park, J.H., Lee, H.J., 1993. Estimation of bioconcentration factor in fish, adsorption coefficient of soils and sediments and interfacial tension with water for organic nonelectrolytes based on the linear solvation energy relationships. *Chemosphere* 26, 1905–1916.

- Schlüter, M., Rüger, N., 2007. Application of a GIS-based simulation tool to illustrate implications of uncertainties for water management in the Amudarya river delta. *Environmental Modelling and Software* 22 (2), 158–166.
- Tao, S., Hu, H., Xu, F., Dawson, R., Li, B., Cao, J., 2000. Fragment constant method for prediction of fish bioconcentration factors of non-polar chemicals. *Chemosphere* 41, 1563–1568.
- Tetko, I.V., Gasteiger, J., Todeschini, R., Mauri, A., Livingstone, D., Ertl, P., Palyulin, V. A., Radchenko, E.V., Zefirov, N.S., Makarenko, A.S., Tanchuk, V.Y., Prokopenko, V. V., 2005. Virtual computational chemistry laboratory – design and description. *Journal of Computer-Aided Molecular Design* 19, 453–463.
- Veith, G.D., DeFoe, D.L., Bergstedt, B.V., 1979. Measuring and estimating the bioconcentration factor of chemicals on fish. *Journal of Fisheries Research Board of Canada* 36, 1040–1048.
- Veith, G.D., Kosian, P., 1983. Estimating bioconcentration potential from octanol/water partition coefficients. In: Mackay, D., Paterson, S., Eisenreich, S.J., Simons, M.S. (Eds.), *Physical Behavior of PCBs in the Great Lakes*. Ann Arbor Sciences Publishers, Ann Arbor, pp. 269–282.
- Wei, D., Zhang, A., Wu, C., Han, S., Wang, L., 2001. Progressive study and robustness test of QSAR model based on quantum chemical parameters for predicting BCF of selected polychlorinated organic compounds (PCOCs). *Chemosphere* 44, 1421–1428.
- Wieland, R., Mirschel, W., 2008. Adaptive fuzzy modeling versus artificial neural networks. *Environmental Modelling and Software* 23 (2), 215–224.
- Winkler, D.A., 2004. Neural networks as robust tools in drug lead discovery and development. *Molecular Biotechnology* 27, 139–167.
- Zhang, J.S., Leung, Y.W., Apr 2004. Improved possibilistic c-means clustering algorithms. *IEEE Transactions on Fuzzy Systems* 12 (2), 209–217.