Engineering Applications of Artificial Intelligence 41 (2015) 175–182

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

A new structure identification scheme for ANFIS and its application for the simulation of virtual air pollution monitoring stations in urban areas

Hamid Taheri Shahraiyni [a,b,*], Sahar Sodoudi [a], Andreas Kerschbaumer [c], Ulrich Cubasch [a]

[a] Institut für Meteorologie, Freie Universität Berlin, Carl-Heinrich-Becker-Weg 6-10, 12165 Berlin, Germany
[b] Faculty of Civil Eng., Shahrood University, Shahrood, Iran
[c] Senate Department for Urban Development and the Environment, Berlin, Germany

ARTICLE INFO

Article history:
Received 22 October 2014
Received in revised form
30 December 2014
Accepted 16 February 2015
Available online 11 March 2015

Keywords:
Structure identification
ANFIS
Virtual stations
Urban areas
Air pollution

ABSTRACT

Parameter and structure identifications are necessary in any modelling which aims to achieve a generalised model. Although ANFIS (Adaptive Network-based Fuzzy Inference System) employs well-known parameter-identification techniques, it needs to structure identification techniques for the determination of an optimum number of fuzzy rules and the selection of significant input variables from among the candidate input variables. In this study, a new structure identification scheme is developed and introduced, which is simultaneously capable of the selection of significant input variables and the determination of an optimum number of rules. This new structure identification was joined to ANFIS, and this joined modelling framework was applied to the simulation of virtual air-pollution monitoring stations in Berlin. In this study, 18 virtual particulate matter stations were simulated using the particulate matter data of some of the current stations. In other words, the particulate matter monitoring network of Berlin has been intensified. The evaluation of simulated virtual stations shows that, although the uncertainty of daily particulate matter measurement is about 10 percent, the simulated virtual stations can estimate the mean daily particulate matter with less than 10 percent of error. Mean absolute error and root mean square error of the simulations are less than 2.4 and 3.4 $\mu g/m^3$, respectively. The correlation coefficient of the simulation results was more than 0.94. In addition, the range of mean bias error is between -1.0 and 0.5 $\mu g/m^3$, and the range of factor of exceedance is between -14.8 and 10.8 percent. It means that the simulated virtual stations have a small bias. These results demonstrated the appropriate performance of the joined new structure identification scheme and ANFIS for development of a virtual air pollution monitoring network.

1. Introduction

The ability of fuzzy logic methodology to model complex non-linear systems has been proven. Up to now, different fuzzy modelling techniques have been developed (e.g., Mamdani, 1976; Takagi and Sugeno, 1985), and they have been widely used for the modelling of different systems.

System identification is the first step towards the modelling of a system using fuzzy approaches. Two types of identification (parameter identification and structure identification) are necessary for the fuzzy modelling (Sugeno and Yasukawa, 1993). Parameter identification is the determination of the parameters of the fuzzy model. In neuro-fuzzy systems, the antecedent and consequent parameters are optimised by a learning technique in the parameter identification step (Alizadeh et al., 2012). Structure identification can be divided into two sub-groups. The first sub-group is called input selection, and means the determination of the appropriate input variables from among all the possible input candidates. The second sub-group is the determination of the number of rules (Linkens and Chen, 1999).

As the number of input variables increases, the complexity of the model increases (Alizadeh et al., 2012). Thus, an input selection procedure is employed to select the appropriate input variables (Sindelar and Babuska, 2004). Up to now, many different input selection algorithms have been developed and utilised for the modelling, and the input selection of fuzzy systems (e.g., Chiu, 1994; Jang, 1996; Nakashima et al., 1997; Gaweda et al., 2001; Vieira et al., 2010).

Clustering techniques have been widely used to divide the space of the variables and to determine the significant rules in the neuro-fuzzy techniques. The clustering technique can be performed based upon the input, output, and the joint input–output datasets (Mascioli et al., 1997). Fuzzy C-mean Clustering (FCM) (Bezdek, 1973, 1981) is one of the most widely-used clustering techniques in the fuzzy modelling studies employed from among the different developed clustering techniques, (e.g., Wong and Chen, 1999; Yao et al., 2000; Panella et al.,

* Corresponding author at: Institut für Meteorologie, Freie Universität Berlin, Carl-Heinrich-Becker-Weg 6-10, 12165 Berlin, Germany. Tel.: +49 30 83854366; fax: +49 30 83871160.
E-mail address: hamid.taheri@met.fu-berlin.de (H. Taheri Shahraiyni).

http://dx.doi.org/10.1016/j.engappai.2015.02.010

2001; Angelov, 2004; Panella and Gallo, 2005). The most important problem in the application of the FCM technique in the fuzzy modelling techniques for the determination of the rules is that the number of rules must be known in advance. The determination of the optimum number of rules is very important in fuzzy modelling, and fuzzy modelling with an optimum number of rules can achieve a generalised model (Panella et al., 2001). There are two opportunities for this problem (1.utilizing of iterative techniques, 2.finding the optimal clusters) (Tsekouras et al., 2005). Many studies have been endeavoured to determine the optimum number of fuzzy-rules (the number of clusters) (e.g., Chiu, 1994; Emami et al., 1998; Chen et al., 1998; Chen and Linkens, 2000; Tsekouras et al., 2005; Dong and Wang, 2011; Panella, 2012).

In addition, some methods have been developed that not only select the significant input variables for fuzzy models, but also determine the optimum number of fuzzy rules (clusters) (e.g., Sugeno and Yasukawa, 1993; Lin and Cunningham, 1995; Yinghua and Cunningham, 1995; Linkens and Chen; 1999; Chen and Linkens, 2001; Min-You and Linkens, 2001).

A neuro-fuzzy system is a fuzzy system which is presented in the network architecture and the parameter identification of the fuzzy system is performed by an automatic learning technique of an adaptive network (Subasi, 2007). ANFIS (Adaptive Network-based Fuzzy Inference System) (Jang, 1993) is one of the well-known neuro-fuzzy approaches, and it has shown appropriate results in the modelling of complex non-linear problems (Alizadeh et al., 2012; Subasi, 2007). The parameter identification of ANFIS is performed by a learning technique in an adaptive network (Jang, 1993), but it needs to structure identification.

In this study, a new structure identification scheme for ANFIS is presented, which is able to determine not only the significant input variables, but also the number of fuzzy rules (clusters).

There are some major objectives for the installation of air pollution monitoring networks in urban areas, such as the description of the spatio-temporal concentrations of pollutants and the validation of the mechanistic models which describe the transporting and the conversion of the pollutants in the atmosphere (Van Egmond and Onderdelinden, 1981), investigation of the compliance of the concentration of air pollutants with air quality standards (Liu et al., 1986), the determination of critical air pollution conditions, and, consequently, decisions regarding the issuing of public warnings or the imposition of temporary immediate emission reductions (Chow et al., 2002), and the evaluation of the exposure of people and other vulnerable receptors to pollution, and the protection of public health (Trujillo-Ventura and Hugh Ellis, 1991; Lozano et al., 2009).

To achieve the major objectives of air pollution monitoring network development in an appropriate manner, the number of air pollution monitoring stations in urban areas must be increased (Stalker and Dickerson, 1962; Beaulant et al., 2008). Although an increase in the number of stations in a monitoring network leads to better achievement of the main objectives of the monitoring network development, it is very expensive (Kanaroglou et al., 2005). The total cost of a monitoring network has a direct relation with the total number of stations in the network (Hickey et al., 1971; Modak and Lohani, 1985), and it is the main constraint on the development of a dense air pollution monitoring network (Trujillo-Ventura and Hugh Ellis, 1991).

There are about 0.32 cars and LDV (Light Duty Vehicles) per resident and about 100,000 HDV (Heavy Duty Vehicles) in Berlin. Berlin is situated approximately 200 km northwest of the industrialised area at Germany's borders with Poland and the Czech Republic, which is called the "Black Triangle" (Lenschow et al., 2001). Although Berlin had a dense air pollution monitoring network in 1990s, it now has a small number of monitoring stations and many of its station have been shut down because of the high total cost of the dense monitoring network. One of the techniques for the densification of the air pollution monitoring network is the simulation of a virtual station that is completely free-of-charge. The idea of a virtual station was introduced by Ung et al. (2001, 2002).

In this study, the shut-down stations in Berlin are virtually re-constructed by simulation technique, and these re-constructed (virtual) stations are added to the current monitoring network for the development of a dense air pollution monitoring network for Berlin. Here, a new structure identification scheme is developed and joined to the ANFIS technique for the simulation of some virtual air pollution monitoring stations in Berlin.

The European Union (EU) has set two limit values for PM10 (Particulate matter less than 10 μm in aerodynamic diameter) for the protection of human health. According to these limits, the mean daily PM10 concentration may not exceed 50 ($\mu g/m^3$) more than 35 times per year, and the mean annual PM10 concentration may not exceed 40 ($\mu g/m^3$) (European Union, 2008). The concentration of PM10 in Berlin sometimes exceeds the EU limit (Görgen and Lambrecht, 2007). Accordingly, particulate matter as an important pollutant in Berlin is selected as an important case study, and virtual particulate matter monitoring stations in Berlin are simulated in this study.

2. Algorithm of modelling

In this study, an input–output database is introduced to the new structure identification algorithm. The algorithm of this new structure identification scheme is described in the next section. This structure identification scheme determines the optimum number of rules and the significant input variables from among the candidate input variables.

Then, the input–output database is divided into two databases (training and testing databases). Then, fuzzy C-mean (FCM) clustering technique (Bezdek, 1973, 1981) is utilised for the clustering of the training database to the optimum number of clusters, determined by structure identification scheme.

An initial TS (Takagi-Sugeno) fuzzy inference system (Takagi and Sugeno, 1983) is developed by defining the input and output membership functions. In this fuzzy system, each cluster is considered as a rule; hence, the number of fuzzy rules is equal to the number of clusters, developed by FCM.

Next, the parameters of the initial fuzzy inference system is tuned in an ANFIS architecture.

Finally, the performance of trained model is evaluated using testing database.

In order to clarify the modelling procedure, a brief explanation is presented, given that the ANFIS modelling procedure is a well-known technique.

First, the input–output database is introduced to the structure identification scheme. Imagine that the X and Y are determined as significant input variables and the optimum number of clusters is determined equal to 2.

The FCM clustering technique divides the data into two clusters and the initial TS fuzzy inference system initially adjusts a rule for each cluster as below:

$$\text{If } X \in \mu_{A_1}(x) \quad \text{and} \quad Y \text{ is} \in \mu_{B_1}(y) \quad \text{then} \quad f_1 = p_1X + q_1Y + r_1$$

$$\text{If } X \in \mu_{A_2}(x) \quad \text{and} \quad Y \text{ is} \in \mu_{B_2}(y) \quad \text{then} \quad f_2 = p_2X + q_2Y + r_2$$

where $\mu_{A_i}(x)$ and $\mu_{B_i}(y)$ are the membership functions and p_1, q_1, r_1, p_2, q_2 and r_2 are the constant values.

Then, the ANFIS architecture is utilised to tune the values of p_1, q_1, r_1, p_2, q_2 and r_2 (consequent parameters) and the parameters of the membership functions (premise parameters).

ANFIS has five layers. The outputs of layer 1 are membership functions. In the second layer, the membership functions are multiplied and the output of layer 2 is w_i: $w_i = \mu_{A_i}(x)\mu_{B_i}(y), \quad i = 1, 2.$

In the third layer, the w_i values are normalised $(\overline{w}_i = (w_i/(w_1 + w_2)); \quad i = 1, 2)$.

The output of the next layer are $\overline{w_i}f_i = \overline{w_i}(p_iX + q_iY + r_i)$ $i = 1, 2$.

Finally, in the fifth layer, the output is calculated as: $F = \sum_i \overline{w_i}f_i$.

Different learning techniques have been developed for the tuning of the premise and consequent parameters in the ANFIS (e.g., Jang and Mizutani, 1996; Mascioli et al., 1997; Tang et al., 2005; Ho et al., 2009). In this study, the hybrid of the gradient descent and least squares technique, developed by Jang (1993), is employed to tune the premise and consequent parameters.

3. New structure identification scheme

In this new structure identification scheme, a heuristic partitioning method is utilised for the partitioning of the main MISO (Multi Inputs–Single Output) database to some MISO sub-databases in a successive manner. Each MISO sub-database is converted to some SISO (Single Input–Single Output) sub-databases. In each SISO sub-database, the non-linear relationship between the input and the output is determined by a fuzzy curve fitting technique, and a non-linear one-variable function is developed. Among the developed one-variable functions for the SISO sub-databases in each MISO sub-database, the best one for the output estimation is determined, and, consequently, the overall accuracy of the output estimation in the main MISO database is evaluated. The heuristic partitioning method is iterated to achieve the maximum overall accuracy of output estimation. When the heuristic partitioning method is terminated, the number of MISO sub-databases is utilised as the optimum number of fuzzy rules for ANFIS. The selected input variables for partitioning in the heuristic method are considered as the first group of candidate input variables and their relative importance are determined based upon the number of partitioning times of each candidate variable. In addition, the employed input variables in the non-linear one-variable functions are considered as the second group of candidate variables, and their relative importance are determined based upon the number of estimated output data by each variable. Finally, the suitable input variables for ANFIS are determined by the combination of the relative importance of two groups of candidate variables.

This structure identification technique uses the fuzzy curve fitting technique, hence, its results are completely compatible with fuzzy modelling techniques (e.g., ANFIS). In addition, the main database is randomly divided into training and testing datasets in this new structure identification scheme. Hence, the algorithm of structure identification is iterated according to a user-defined number of iterations, and, in each iteration, the main database is randomly divided again. Finally, the suitable input variables and number of fuzzy rules are determined based upon the combination of the results of all of the iterations. These iterations neutralise the effects of random dividing and generalise the results of structure identification scheme.

Now, the algorithm of the new structure identification scheme is described in detail, step by step.

Step 1. Input–output database preparation: A database (D) of input candidate variables and corresponding output values is prepared. Imagine that the database has n input variables ($X = \{X_1, X_2, \ldots, X_n\}$) and one output variable (Y). Thus, D can be expressed as Eq. (1).

$$D = \left\{ \left(x_k^m, y^m\right) \right\},\ m = 1, \ldots, M,\ k = 1, \ldots, n \tag{1}$$

where x_k^m is the mth member of the kth variable (X_k) ($x_k^m \in X_k$ and $X_k \in X$), y^m is the mth member of Y and M is the total number of observations.

Step 2. Training and test databases: the database is randomly partitioned to the training (two third of database) and testing (one third of database) databases. Hereinafter, the training database is called the database.

Step 3. Dividing the database: each database must be divided into two smaller databases. In the first iteration of the system identification algorithm, there is only one database (D) and it is divided into two smaller databases. In general, a generated database is expressed as D_k^{ds} and it is the sth database in the dth iteration and has been generated by dividing the kth variable of a bigger database. The bigger database has been divided into two parts ($s \in \{1, 2\}$) and this database is the sth part.

When any of the input variables (X_k) are divided into two parts, then D is divided into two sub-databases (D_k^{11}, D_k^{12}). D_k^{11} and D_k^{12} are the databases generated by dividing the kth variable in the first iteration.

$$D = \left\{ D_k^{11},\ D_k^{12} \right\} \tag{2}$$

$$D_k^{11} = \left\{ \left(x_l^t, y^t\right) \right\},\ t = 1, \ldots, Q^1;\quad l = 1, \ldots, n \quad \text{if} \quad X_k \le T_k^1 \tag{3}$$

$$D_k^{12} = \left\{ \left(x_l^t, y^t\right) \right\},\ t = 1, \ldots, Q^1;\quad l = 1, \ldots, n \quad \text{if} \quad X_k > T_k^1 \tag{4}$$

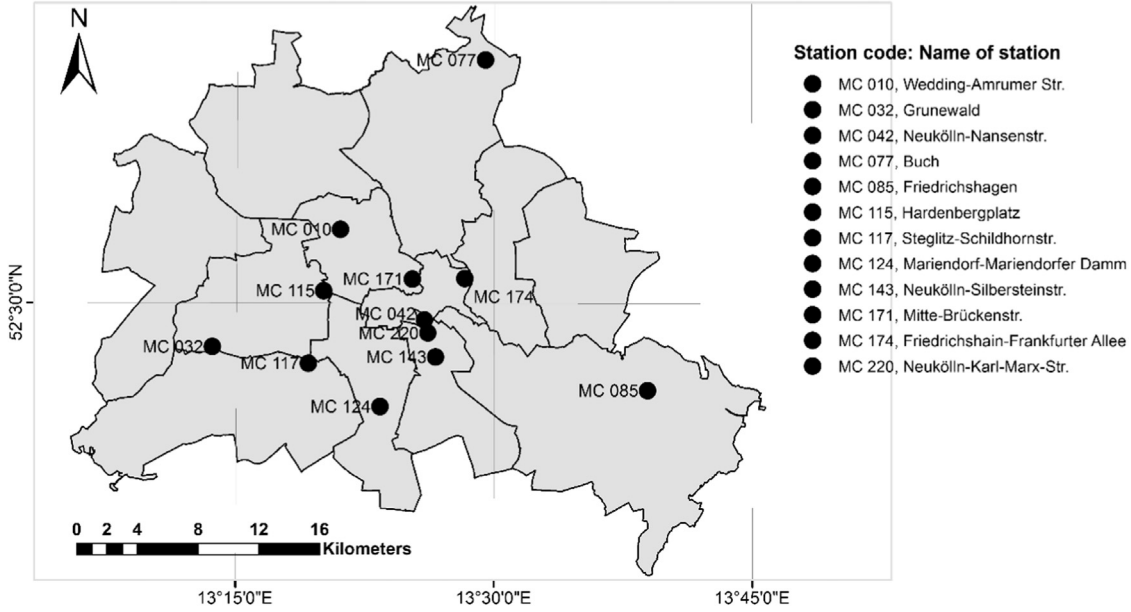

Fig. 1. Berlin map with its 12 particulate matter monitoring stations.

where T_k^1 is the median of X_k in database D. Q^1 is the number of observations in each database and it is equal to $M/2$.

In the second iteration, it is decided which database should be divided into two smaller databases $\left(D_k^{11} \text{ or } D_k^{12}\right)$. Imagine D_k^{11} is selected for the dividsion and it is divided to two smaller databases ($D_{k'}^{21}$, $D_{k'}^{22}$, $k' \in \{1,\ldots,n\}$). $D_{k'}^{21}$ and $D_{k'}^{22}$ are the databases, generated by dividing the k'th variable of D_k^{11} in the second iteration. In the second iteration, D has been divided to three databases as below:

$$D = \left\{ D_{k'}^{21},\ D_{k'}^{22},\ D_k^{12} \right\} \tag{5}$$

$$D_{k'}^{21} = \{(x_l^t, y^t)\},\ t = 1,\ldots,Q^2,\ l = 1,\ldots,n \quad \text{if} \quad (X_k \le T_k^1 \,\&\, X_{k'} \le T_{k'}^2) \tag{6}$$

$$D_{k'}^{22} = \{(x_l^t, y^t)\}, \quad t = 1,\ldots,Q^2, \quad l = 1,\ldots,n \quad \text{if} \quad (X_k \le T_k^1 \,\&\, X_{k'} > T_{k'}^2) \tag{7}$$

$$D_k^{12} = \{(x_l^t, y^t)\}, \quad t = 1,\ldots,Q^1; \quad l = 1,\ldots,n \quad \text{if} \quad X_k > T_k^1 \tag{8}$$

where Q^2 is the number of observations in each database and is equal to $Q^1/2$. $T_{k'}^2$ is the median of $X_{k'}$ in database D_k^{11}.

This algorithm is iterated and the D is divided into more small databases. In general, D in the dth iteration is divided into $d+1$ small databases.

Which database is selected for the division into two smaller databases in each step, and which X_k is the best one to be divided into the selected database?

First, the method for the determination of the appropriate X_k for dividing a database is explained here. Then, the method for the selection of a database for the division will be explained in Step 4.

For the determination of the best variable for the division, all of the possible dividing options are performed. Hence, for the division of the D_k^{ds} into two smaller databases, n possible options are performed and $2n$ databases are generated. The data in each generated database are divided into n one-variable databases. Thus, when the jth variable is divided into two small databases, $2n$ one-variable databases (S) are generated.

$$S_{js}^i = \left\{ (x_i^{t,j},\ y^t) \right\}, \quad i = 1,\ldots,n, \quad t = 1,\ldots,Q^d, \quad s = 1,2 \tag{9}$$

Q^d is the number of (x,y) points in the one-variable database.

The relationship between X_i and Y ($\hat{Y}_i = f_i^j(X_i)$, $i = 1,\ldots,n$) in all of S_{j1}^i is calculated by a fuzzy interpolation technique called IDS (Ink Drop Spread) (Bagheri Shouraki and Honda, 1999). Similarly, the relationship between X_i and Y $\left(\hat{Y}_i = g_i^j(X_i),\ i = 1,\ldots,n\right)$ in all of S_{j2}^i is calculated. The accuracy of f_i^j and g_i^j functions (one-variable functions) for the estimation of output (Y) is evaluated. Consequently, f_z^j and $g_{z'}^j$ ($z \,\&\, z' \in \{1,\ldots,n\}$) are determined as the

Table 1
The removed and current stations, with the time periods of the concurrent hourly particulate matter data of the current and removed stations (the codes of the stations are corresponding to Figs. 1 and 2).

Removed stations (output variables)	Current stations (input variables)	Time periods of particulate matter data	Number of hourly concurrent data
MC 001	MC 10, 32, 42, 77, 85, 117, 174	1996.01.23–1997.01.22	5,952
MC 006	MC 10, 32, 42, 77, 85, 174	1994.01.31–1995.04.03	9,168
MC 007	MC 10, 32, 42, 77, 85, 117, 174	1994.12.01–1996.02.02	6,360
MC 009	MC 10, 32, 42, 77, 85, 117, 174	1994.12.01–1996.03.29	7,440
MC 017	MC 10, 32, 42, 77, 85, 117, 174	1995.10.16–1997.10.06	12,360
MC 018	MC 10, 32, 42, 77, 85, 117, 174	2009.01.18–2011.01.18	14,928
MC 020	MC 10, 32, 42, 77, 85, 174	1994.01.31–1995.11.24	6,480
MC 023	MC 10, 32, 42, 77, 85, 117, 174	1995.03.30–1996.03.29	4,992
MC 024	MC 10, 32, 42, 77, 85, 174	1994.08.28–1995.10.19	8,112
MC 025	MC 10, 32, 42, 77, 85, 117, 174	1995.03.28–1997.03.27	11,064
MC 027	MC 10, 32, 42, 77, 85, 117, 174	2009.01.10–2011.01.10	11,688
MC 028	MC 10, 32, 42, 77, 85, 117, 174	1994.10.19–1995.10.19	7,080
MC 030	MC 10, 32, 42, 77, 85, 117, 174	1994.10.19–1995.10.19	7,152
MC 078	MC 10, 32, 42, 77, 85, 117, 174	1995.01.30–1996.01.30	5,064
MC 080	MC 10, 32, 42, 77, 85, 117, 174	1995.03.27–1997.03.26	11,088
MC 081	MC 10, 32, 42, 77, 85, 117, 174	1995.04.04–1996.04.03	4,920
MC 083	MC 10, 32, 42, 77, 85, 117, 174	1995.01.30–1996.01.30	5,016
MC 145	MC 10, 32, 42, 77, 85, 117, 171, 174	2002.03.27–2004.03.09	10,272

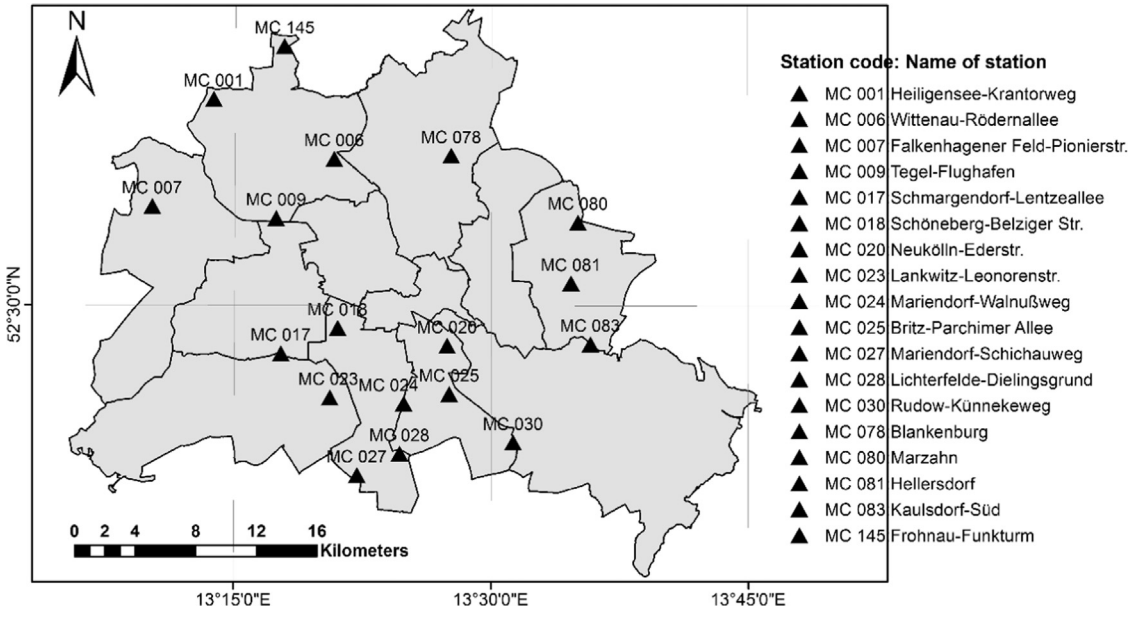

Fig. 2. The 18 old removed particulate matter monitoring stations (Virtual stations).

best one-variable functions with the lowest errors, respectively. If we consider e^j as the total error of the output (Y) estimation in D_k^{ds} by f_z^j and $g_{z'}^j$, then e^j for $j=1,\ldots,n$ is calculated and the minimum value in $\{e^1,\ldots,e^n\}$ is determined. Consider $e^{k'}$ as the minimum. Consequently, the input variable corresponding to the minimum error ($X_{k'}$) is the best variable for dividing D_k^{ds} into two smaller databases ($D_{k'}^{d+1,1}$ and $D_{k'}^{d+1,2}$) and $f_z^{k'}(X_z)$ and $g_{z'}^{k'}(X_{z'})$ are the best one-variable functions for the estimation of output in the two generated databases, and $e(f_z^{k'})$ and $e(g_{z'}^{k'})$ are their corresponding errors, respectively.

Step 4. Rule-base generation: in the first iteration of the dividing algorithm, D is divided into two databases (see Eqs. (2)–(4)). Then two one-variable functions ($f_z^k(X_z)$ and $g_{z'}^k(X_{z'})$) are determined and utilised for the output estimation in two databases. The error of the one-variable functions are $e(f_z^k)$ and $e(g_{z'}^k)$. Therefore, the rule-base can be expressed as Eq. (10).

$$\begin{cases} \text{If} \quad X_k \le T_k^1 & \text{then} \quad \widehat{Y_1} = f_z^k(X_z) \\ \text{If} \quad X_k > T_k^1 & \text{then} \quad \widehat{Y_2} = g_{z'}^k(X_{z'}) \end{cases} \tag{10}$$

Using the test database, the accuracy of generated rule-base (Eq. (10)) for the estimation of the output variable (Y) is evaluated. The error of output estimation in the first iteration is expressed as E_1.

In the second iteration of the dividing algorithm, the database with higher error is selected for dividing. Imagine $e(f_z^k) > e(g_{z'}^k)$, then, D_k^{11} must be divided to two smaller databases using the dividing method, explained in Step 3. Thus, two one-variable functions ($f_{z_1}^{k'}(X_{z_1})$ and $g_{z_2}^{k'}X_{z_2}$, $z_1 \,\&\, z_2 \in \{1,\ldots,n\}$) are determined and utilised for the output estimation in the two databases. Accordingly, D is divided to three databases (Eq. (5)), and a rule-base with three rules (Eq. (11)) is generated. The error of these one variable functions are $e\left(f_{z_1}^{k'}\right), e\left(g_{z_2}^{k'}\right)$ and $e\left(g_{z'}^{k}\right)$.

$$\begin{cases} \text{if} \quad (X_k \le T_k^1 \quad X_{k'} \le T_{k'}^2) & \text{then} \quad \widehat{Y_1} = f_{z_1}^{k'}(X_{z_1}) \\ \text{if} \quad (X_k \le T_k^1 \quad X_{k'} > T_{k'}^2) & \text{then} \quad \widehat{Y_2} = g_{z_2}^{k'}(X_{z_2}) \\ \text{if} \quad X_k > T_k^1 & \text{then} \quad \widehat{Y_3} = g_{z'}^{k}(X_{z'}) \end{cases} \tag{11}$$

Using the test database, the accuracy of generated rule-base (Eq. (11)) for the estimation of the output variable (Y) is evaluated. The error of the output estimation, in the second iteration is expressed as E_2.

This dividing procedure and the rule-base generation (Steps 3 and 4) are continued until $E_d > E_{d-1}$.

Step 5. The rule-base with $d-1$ rules is considered as the best rule-base in the first iteration of the algorithm, and it is expressed as $R_1 = d-1$. In this rule base, the number of dividing times of each input variable is calculated, and, consequently, is expressed as the dividing vector ($\overrightarrow{DV}_1 = [dv_1, dv_2, \ldots, dv_n]$). In addition, the number of the estimated data provided by the different variables can be calculated using the one-variable functions in the rule-base and the number of data in the $d-1$ databases. Consequently, the results of these calculations can be presented as the function vector $\left(\overrightarrow{FV}_1 = [fv_1, fv_2, \ldots, fv_n]\right)$.

Step 6. Combine the training and testing databases to generate the original database, and then proceed to Step 2. Steps 2–6 are iterated according to the user defined number of iterations (I). These iterations neutralise the effects of the random dividing in the second step, and generalise the results. Steps 2–6 are iterated I times. Thus, I different dividing vectors ($\overrightarrow{DV}_1, \ldots, \overrightarrow{DV}_I$), function vectors ($\overrightarrow{FV}_1, \ldots, \overrightarrow{FV}_I$) and vector of number of rules ($R_1, \ldots, R_I$) are generated.

Step 7. Input selection: the average of the I dividing and function vectors are calculated ($\overline{DV}$, $\overline{FV}$). Then $\overline{DV}$ and $\overline{FV}$ are normalised as the sum of the elements in each vector as equal to 1. Afterwards, the average of two normalised vectors is calculated and called importance vector $\left(\overrightarrow{IV} = (IV_1, \ldots, IV_n)\right)$. The elements of this vector show the relative importance of the different input variables for the modelling of the output variable. Finally, the variables with low importance are removed from the database and the modelling is performed by the remained variables.

Step 8: Optimum number of rules: the average of vector of number of rules is calculated ($\overline{R}$). $\overline{R}$ is considered as the optimum number of rules or clusters.

4. Experiment

4.1. Case study, data and software

Berlin (Fig. 1) is the capital city of Germany and is located in the North-eastern part of Germany. It has a population of 3.4 million residents and covers an area of about 900 km^2. At present, Berlin has only 12 PM10 monitoring stations (Fig. 1). At the beginning of the 1990s, it had a high level of airborne particulate matter concentration (Lenschow et al., 2001), and it had a dense monitoring network with more than 40 stations, developed for the appropriate monitoring of the pollutants in the city (SenStadt, 1998). The number of monitoring stations decreased greatly until the end of 1990s (Lenschow et al., 2001). Fig. 2 shows some of these removed stations that are used as virtual stations in this study.

We tried to find some old concurrent hourly particulate matter data from these stations, represented in Figs. 1 and 2. Table 1 shows the time period and the number of the concurrent data of each removed station (the stations in Fig. 2) and some of the current stations (the stations in Fig. 1). The stations in Table 1 are either far from the main traffic lanes or the current traffic level around them has no significant difference with the traffic level in the time period of the simulations (Table 1). In this study, the current stations and each removed station in Table 1 are considered as input and output variables for simulation, respectively. Finally, each simulated removed station is considered as a virtual station, and, consequently, the particulate matter monitoring network of Berlin is densified (12+18=30 stations).

Table 2
The appropriate input variables (current stations) for the simulation of the output variables (removed or virtual stations) with the optimum number of rules (clusters).

Output variables	Appropriate input variables	Optimum number of rules
MC 001	MC 10,32,77	16
MC 006	MC 10,77	4
MC 007	MC 10,32,77,85	16
MC 009	MC 10,32,42,77	32
MC 017	MC 10,32,42	8
MC 018	MC 10,42,117	8
MC 020	MC 10,42,77	16
MC 023	MC 10,32,42,85	16
MC 024	MC 32,42,77,85	24
MC 025	MC 42,85	6
MC 027	MC 10,32,42	8
MC 028	MC 10, 32,42,77	16
MC 030	MC 42,77	4
MC 078	MC 77,85	8
MC 080	MC 10,42,85,117	16
MC 081	MC 10,42,77,85	24
MC 083	MC 10,42,85	8
MC 145	MC 10,32,77	12

A computer programme was developed in MATLAB (R2013b) for the implementation of our new structure identification scheme, and, consequently, ANFIS was also implemented in MATLAB.

4.2. Results and discussion

For each row in Table 1, an input–output database was created. The input and output variables in each database were the hourly particulate matter concentration of the stations of second and first columns of Table 1, respectively. Finally, 18 input–output databases were created. The structure identification scheme was applied to these databases, and, consequently, the appropriate input variables and the optimum number of rules in the databases were determined (Table 2).

Then, initial TS fuzzy inference systems were generated, after which the parameters of the fuzzy inference systems were tuned using ANFIS. After each training step (epoch), the performance of the model is evaluated by the testing dataset. If the error of the model is less than the previous step, then the training procedure is continued and the next training step is performed, otherwise the training procedure is terminated. The EU limit for PM10 is expressed as daily and annual scales, and, hence, the results of the hourly simulations of 18 virtual stations are converted to daily scales.

The results of the training of the 18 ANFIS model for the simulation of the 18 virtual stations have been presented in Table 3. Then, the test dataset was employed for the evaluation of the performance of the developed models. The evaluation results of the 18 developed models for the estimation of the mean daily PM10 concentration have been presented in Table 4. The R, MAE, RMSE, MBE, FOEX and MAPE values in Tables 3 and 4 represent the correlation coefficient, the Mean Absolute Error $(\mathrm{MAE}=(1/n)\sum_{i=1}^{n}|O_i-S_i|)$, the Root Mean Square Error (RMSE $=\sqrt{(1/n)\sum_{i=1}^{n}(O_i-S_i)^2}$), the Mean Bias Error $\left(\mathrm{MBE}=(1/n)\sum_{i=1}^{n}(S_i-O_i)\right)$, the Factor Of Exceedance $(\mathrm{FOEX}=((n_{S>O}/n)-0.5)\times 100)$, and the Mean Absolute of Percentage Error $(\mathrm{MAPE}=(1/n)\sum_{i=1}^{n}(|O_i-S_i|/|O_i|)\times 100)$, respectively. n is the number of observation data, O_i and S_i are the observed and simulated PM10 concentration of the ith data, respectively. $n_{S>O}$ is the number of data whose simulated values is higher than their observation values.

The results of Tables 3 and 4 demonstrate that the appropriate virtual stations have been simulated. According to Table 4, the MAPE of the developed models is less than 10 percent, and the correlation coefficients of the models are more than 0.94. In addition, the uncertainty of the daily PM10 measurement in Berlin is about 10 percent (K. Grunow, personal communication, 2014) and the percent of error for all of the simulations is less than 10 percent (Table 4). These results demonstrate that the simulated stations have excellent performance for the estimation of the daily PM10 concentration. The MAE and RMSE of the simulations are less than 2.4 and 3.4 $\mu g/m^3$, respectively. In addition, the range of the MBE is between -1.0 and 0.5 $\mu g/m^3$ and the range of the FOEX is between -14.8 and 10.8 percent. It means that the simulated virtual stations have a small bias, and that some of the simulations have been over-estimated while others have been under-estimated.

The scatter plot of the mean daily test data for 3 stations (MC 020, MC 027 and MC 080) has been presented in Fig. 3. In addition, the time series which simulated and measured the daily PM10 concentration ($\mu g/m^3$) in the test dataset of MC 020, MC 027 and MC 080 stations have been presented in Fig. 4. Figs. 3 and 4 demonstrate the high accuracy of the developed models and imply the appropriate

Table 3
Evaluation of the training results of the 18 ANFIS models for the simulation of the virtual stations for the estimation of the mean daily PM10 concentration.

Virtual station	Nr. of daily train PM10 data	Mean of daily train PM10 data ($\mu g/m^3$)	Standard deviation of daily train PM10 data ($\mu g/m^3$)	R	MAE ($\mu g/m^3$)	RMSE ($\mu g/m^3$)	MAPE (%)
MC 001	164	43.4	20.7	0.986	2.3	2.9	7.2
MC 006	246	37.5	18.4	0.976	2.3	3.2	7.5
MC 007	174	33.0	17.5	0.993	1.1	1.6	5.4
MC 009	203	36.9	22.9	0.992	1.7	2.7	6.6
MC 017	342	35.4	19.6	0.995	1.3	1.8	6.4
MC 018	411	24.0	12.1	0.996	0.8	1.2	3.5
MC 020	177	40.4	18.2	0.993	1.4	1.9	4.8
MC 023	139	40.2	19.1	0.990	1.6	2.1	6.3
MC 024	223	32.6	14.3	0.983	1.3	1.7	6.6
MC 025	303	36.8	19.8	0.990	1.4	2.3	6.0
MC 027	322	23.5	11.7	0.991	1.1	1.6	5.5
MC 028	194	33.2	14.5	0.998	1.1	1.3	5.2
MC 030	194	33.0	14.0	0.978	1.6	2.0	7.2
MC 078	139	38.8	16.3	0.965	2.3	3.2	7.4
MC 080	307	43.4	20.8	0.980	2.3	3.2	6.7
MC 081	136	50.9	22.4	0.984	2.4	3.4	5.6
MC 083	139	35.5	17.0	0.992	1.2	1.7	5.6
MC 145	283	27.8	18.1	0.997	1.3	2.0	4.7

Table 4
Evaluation of the testing results of 18 ANFIS models for the simulation of virtual stations for the estimation of mean daily PM10 concentration.

Virtual station	Nr. of daily test PM10 data	Mean of daily test PM10 data ($\mu g/m^3$)	Standard deviation of daily test PM10 data ($\mu g/m^3$)	R	MAE ($\mu g/m^3$)	RMSE ($\mu g/m^3$)	MBE ($\mu g/m^3$)	FOEX (%)	MAPE (%)
MC 001	82	43.7	20.2	0.943	2.9	3.7	−0.04	1.2	7.6
MC 006	136	37.3	16.4	0.962	3.5	4.7	0.00	8.0	9.5
MC 007	90	32.6	15.1	0.978	2.3	3.2	−0.27	4.5	7.8
MC 009	106	35.5	18.8	0.98	3.0	4.0	0.51	10.2	9.0
MC 017	173	35.5	16.9	0.985	2.5	3.1	−0.21	1.1	8.2
MC 018	208	24	10.6	0.992	1.0	1.3	0.05	6.7	4.5
MC 020	90	31.3	16.0	0.991	1.6	2.2	−0.16	−1.1	5.3
MC 023	68	34.1	15.6	0.982	2.5	3.1	0.22	−0.75	8.4
MC 024	114	28	11.9	0.98	1.8	2.4	0.02	6.1	6.9
MC 025	158	32.8	17.1	0.985	2.2	3.1	−0.14	1.9	7.5
MC 027	162	23	9	0.985	1.3	1.7	−0.03	0.0	6.5
MC 028	101	31.3	13.5	0.98	2.3	2.8	−0.42	−6.9	8.2
MC 030	104	30.6	12.4	0.972	2.5	3.1	−0.21	−2.0	8.9
MC 078	69	28	14	0.985	1.7	2.4	−0.49	0.72	6.5
MC 080	155	38.5	19.8	0.979	2.9	4.0	0.16	6.8	7.9
MC 081	68	38.8	18.6	0.984	2.9	3.6	0.26	4.2	8.4
MC 083	71	34.4	16.8	0.977	2.6	3.6	−1.08	−14.8	8.0
MC 145	145	27.9	14.6	0.986	1.9	2.5	−0.07	4.8	7.4

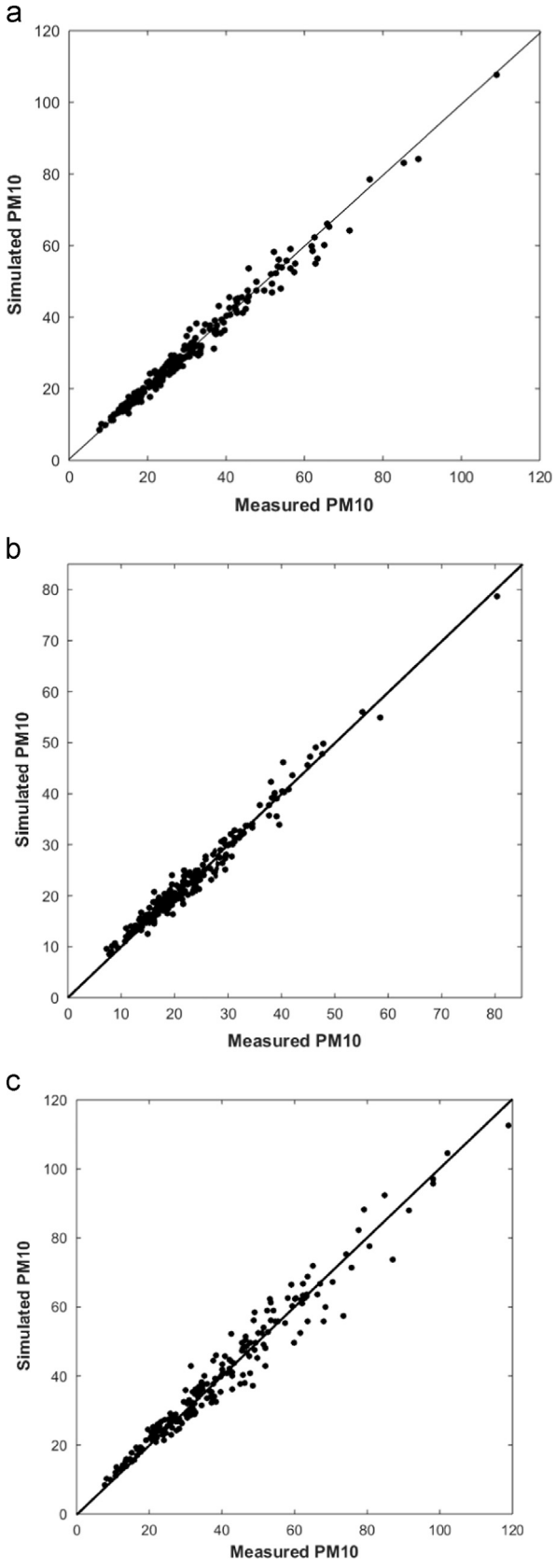

Fig. 3. The scatterplots of the daily test PM10 data (μg/m^3) of (a) MC 020, (b) MC 027 and (c) MC 080 stations.

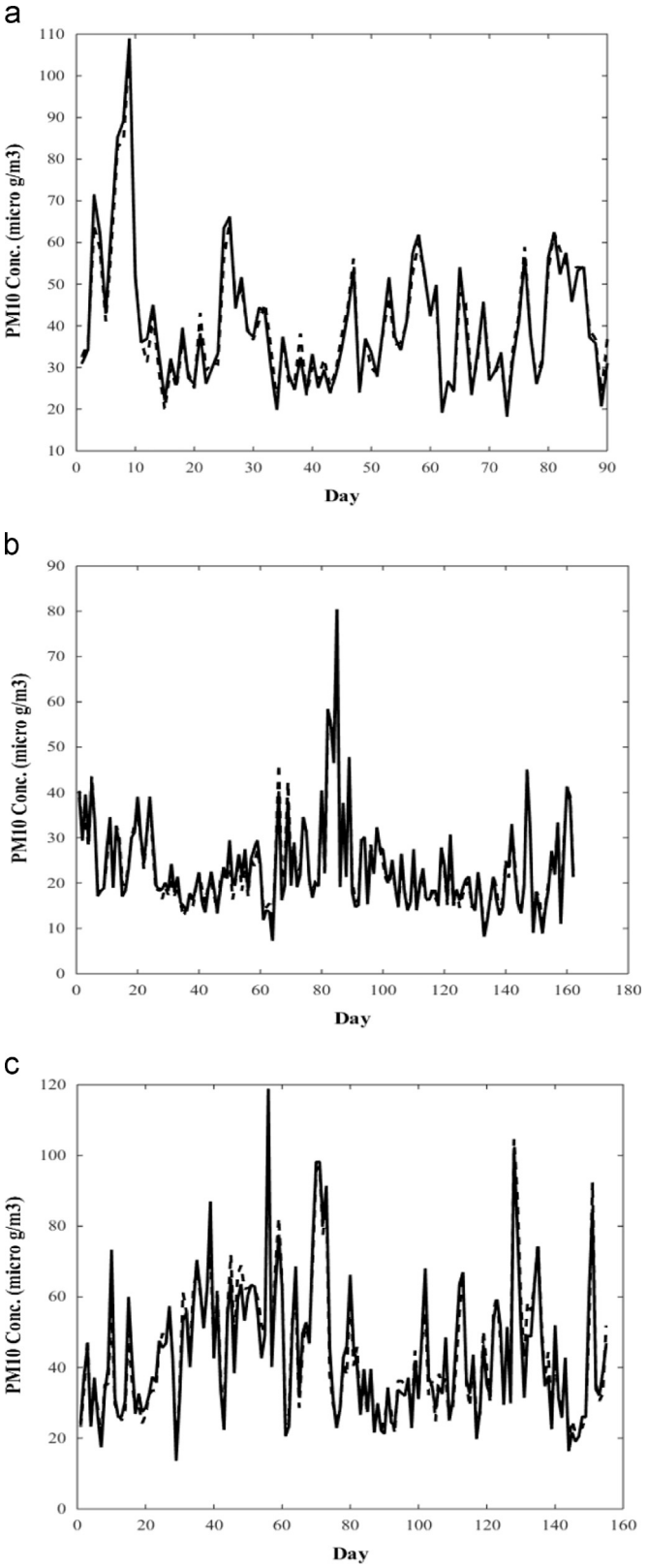

Fig. 4. The time series of simulated (dash line) and measured (solid line) daily PM10 concentration (μg/m^3) in the test dataset of (a) MC 020, (b) MC 027 and (c) MC 080 stations.

performance of the joined new structure identification scheme and ANFIS for development of a virtual particulate matter monitoring network.

In total, the results demonstrated that the new developed structure identification scheme provides the appropriate performance, and combining this scheme with ANFIS lead to the

development of high accuracy and free-of-charge virtual stations for the daily monitoring of particulate matter in Berlin.

5. Conclusions

In this study, a new structure identification scheme has been developed. This structure identification technique is able to determine the optimum number of fuzzy rules and select the significant input variables. This technique was joined with ANFIS in a modelling framework, and was applied for the simulation of virtual air pollution monitoring station in Berlin. The results of simulation of 18 virtual particulate matter stations for Berlin ($R > 0.94$ and MAPE < 10 percent) demonstrated the capabilities of this new structure identification technique.

Acknowledgements

The authors are grateful to the Alexander von Humboldt Stiftung/Foundation for funding this work under Humboldt ID 1149622.

The authors thank Chris Engert for his valuable proof-reading of this paper.

References

Alizadeh, M., Jolai, F., Aminnayeri, M., Rada, R., 2012. Comparison of different input selection algorithms in neuro-fuzzy modeling. Expert Syst. Appl. 39, 1536–1544.

Angelov, P., 2004. An approach for fuzzy rule-base adaptation using on-line clustering. Int. J. Approximate Reasoning 35, 275–289.

Bagheri Shouraki, S., Honda, N., 1999. Recursive fuzzy modeling based on fuzzy interpolation. J. Adv. Comput. Intell. 3, 114–125.

Beaulant, A.L., Perron, G., Kleinpeter, J., Weber, C., Ranchin, T., Wald, L., 2008. Adding virtual measuring stations to a network for urban air pollution mapping. Environ. Int. 34, 599–605.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers.

Bezdek, J.C., 1973. Cluster validity with fuzzy sets. J. Cybern. 3, 58–73.

Chen, J.-Q., Yu-Geng, X., Zhong-Jun, Z., 1998. A clustering algorithm for fuzzy model identification. Fuzzy Sets Syst. 98, 319–329.

Chen, M.Y., Linkens, D.A., 2000. A fuzzy modelling approach using hierarchical neural networks. Neural Comput. Appl. 9, 44–49.

Chen, M.Y., Linkens, D.A., 2001. A systematic neuro-fuzzy modeling framework with application to material property prediction. IEEE Trans. Syst. Man Cybern. 31, 781–790.

Chiu, S.L., 1994. Fuzzy model identification based on cluster estimation. J. Intell. Fuzzy Syst. 2, 267–278.

Chow, J.C., Engelbrecht, J.P., Watson, J.G., Wilson, W.E., Frank, N.H., Zhu, T., 2002. Designing monitoring networks to represent outdoor human exposure. Chemosphere 49, 961–978.

Dong, M., Wang, N., 2011. Adaptive network-based fuzzy inference system with leave-one-out cross-validation approach for prediction of surface roughness. Appl. Math. Model. 35, 1024–1035.

Emami, M.R., Turksen, I.B., Goldenberg, A.A., 1998. Development of a systematic methodology of fuzzy logic modeling. IEEE Trans. Fuzzy Syst. 6, 346–361.

European Union, 2008. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. Off. J. Eur. Union L 152, 1–44.

Gaweda, A.E., Zurada, J.M., Setiono, R., 2001. Input selection in data-driven fuzzy modeling, Fuzzy Systems, 2001. The 10th IEEE International Conference on, 1251–1254.

Görgen, R., Lambrecht, U., 2007. Particulate matter in ambient air. J. Eur. Environ. Plan. Law 4, 278–288.

Hickey, H.R., Rowe, W.D., Skinner, F., 1971. A cost model for air quality monitoring systems. J. Air Pollut. Control Assoc. 21, 689–693.

Ho, W.H., Tsai, J.T., Lin, B.T., Chou, J.H., 2009. Adaptive network-based fuzzy inference system for prediction of surface roughness in end milling process using hybrid Taguchi-genetic learning algorithm. Expert Syst. Appl. 36, 3216–3222.

Jang, J.S., 1993. ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans. Syst. Man Cybern. 23, 665–685.

Jang, J.S.R., 1996. Input selection for ANFIS learning, In: Proceedings of the Fifth IEEE International Conference on Fuzzy Systems, pp. 1493-1499.

Jang, J.S.R., Mizutani, E., 1996. Levenberg–Marquardt method for ANFIS learning. In: Fuzzy Information Processing Society, Biennial Conference of the North American, , pp. 87-91.

Kanaroglou, P.S., Jerrett, M., Morrison, J., Beckerman, B., Arain, M.A., Gilbert, N.L., Brook, J.R., 2005. Establishing an air pollution monitoring network for intra-urban population exposure assessment: a location-allocation approach. Atmos. Environ. 39, 2399–2409.

Lenschow, P., Abraham, H., Kutzner, K., Lutz, M., Preusz, J., Reichenbacher, W., 2001. Some ideas about the sources of PM10. Atmos. Environ. 35, 23–33.

Lin, Y., Cunningham III, G.A., 1995. A new approach to fuzzy-neural system modeling. IEEE Trans. Fuzzy Syst. 3, 190–198.

Linkens, D.A., Chen, M.-Y., 1999. Input selection and partition validation for fuzzy modelling using neural network. Fuzzy Sets Syst. 107, 299–308.

Liu, M.K., Avrin, J., Pollack, R.I., Behar, J.V., McElroy, J.L., 1986. Methodology for designing air quality monitoring networks: I. Theoretical aspects. Environ. Monit. Assess. 6, 1–11.

Lozano, A., Usero, J., Vanderlinden, E., Raez, J., Contreras, J., Navarrete, B., El Bakouri, H., 2009. Design of air quality monitoring networks and its application to NO_2 and O_3 in Cordova, Spain. Microchem. J. 93, 211–219.

Mamdani, E.H., 1976. Advances in the linguistic synthesis of fuzzy controllers. Int. J. Man Mach. Stud. 8, 669–678.

Mascioli, F.M.F., Varazi, G.M., Martinelli, G., 1997. Constructive algorithm for neuro-fuzzy networks, In: Proceedings of the Sixth IEEE International Conference on Fuzzy Systems, pp. 459-464 vol. 451.

Min-You, C., Linkens, D.A., 2001. A systematic neuro-fuzzy modeling framework with application to material property prediction. IEEE Trans. Syst. Man Cybern. 31, 781–790.

Modak, P., Lohani, B.N., 1985. Optimization of ambient air quality monitoring networks. Environ. Monit. Assess. 5, 1–19.

Nakashima, T., Morisawa, T., Ishibuchi, H., 1997. Input selection in fuzzy rule-based classification systems. In: Proceedings of the Sixth IEEE International Conference on Fuzzy Systems, pp. 1457–1462 vol. 1453.

Panella, M., 2012. A hierarchical procedure for the synthesis of ANFIS networks. Adv. Fuzzy Syst. 12.

Panella, M., Gallo, A.S., 2005. An input-output clustering approach to the synthesis of ANFIS networks. IEEE Trans. Fuzzy Syst. 13, 69–81.

Panella, M., Rizzi, A., Mascioli, F.M.F., Martinelli, G., 2001. ANFIS synthesis by hyperplane clustering. IFSA World Congress and 20th NAFIPS International Conference vol. 341, 340–345.

SenStadt, 1998. Air quality management in Berlin 1997. In: Department of Urban Development, Environmental Protection and Technology (Ed.). Air Quality Management Series, Brochure No. 22; Berlin.

Sindelar, R., Babuska, R., 2004. Input selection for nonlinear regression models. IEEE Trans. Fuzzy Syst. 12, 688–696.

Stalker, W.W., Dickerson, R.C., 1962. Sampling station and time requirements for urban air pollution surveys: Part II: Suspended particulate matter and soiling index. J. Air Pollut. Control Assoc. 12, 111–128.

Subasi, A., 2007. Application of adaptive neuro-fuzzy inference system for epileptic seizure detection using wavelet feature extraction. Comput. Biol. Med. 37, 227–244.

Sugeno, M., Yasukawa, T., 1993. A fuzzy-logic-based approach to qualitative modeling. IEEE Trans. Fuzzy Syst. 1, 7–31.

Takagi, T., Sugeno, M., 1985. Fuzzy identification of systems and its applications to modeling and control. IEEE Trans. Syst. Man Cybern. 15, 116–132.

Tang, A.M., Quek, C., Ng, G.S., 2005. GA-TSKfnn: parameters tuning of fuzzy neural network using genetic algorithms. Expert Syst. Appl. 29, 769–781.

Trujillo-Ventura, A., Hugh Ellis, J., 1991. Multiobjective air pollution monitoring network design. Atmos. Environ. Part A: Gen. Top. 25, 469–479.

Tsekouras, G., Sarimveis, H., Kavakli, E., Bafas, G., 2005. A hierarchical fuzzy-clustering approach to fuzzy modeling. Fuzzy Sets Syst. 150, 245–266.

Ung, A., Wald, L., Ranchin, T., Weber, C., Hirsch, J., Perron, G., Kleinpeter, J., 2002. Satellite data for air pollution mapping over a city-virtual stations. In: Proceedings of the 21st EARSeL Symposium, Paris, France.

Ung, A., Weber, C., Perron, G., Hirsch, J., Kleinpeter, J., Wald, L., Ranchin, T., 2001. Air pollution mapping over a city–virtual stations and morphological indicators. In: Tenth International Symposium “Transport and Air Pollution”, Boulder, Colorado USA.

Van Egmond, N.D., Onderdelinden, D., 1981. Objective analysis of air pollution monitoring network data; spatial interpolation and network density. Atmos. Environ. 15, 1035–1046.

Vieira, S.M., Sousa, J.M.C., Runkler, T.A., 2010. Two cooperative ant colonies for feature selection using fuzzy models. Expert Syst. Appl. 37, 2714–2723.

Wong, C.-C., Chen, C.-C., 1999. A hybrid clustering and gradient descent approach for fuzzy modeling. IEEE Trans. Syst. Man Cyber. Part B 29, 686–693.

Yao, J., Dash, M., Tan, S.T., Liu, H., 2000. Entropy-based fuzzy clustering and fuzzy modeling. Fuzzy Sets Syst. 113, 381–388.

Yinghua, L., Cunningham III, G.A., 1995. A new approach to fuzzy-neural system modeling. IEEE Trans. Fuzzy Syst. 3, 190–198.