Aerosol and Air Quality Research, 15: 743–748, 2015

ISSN: 1680-8584 print / 2071-1409 online
doi: 10.4209/aaqr.2014.12.0317

Technical Note

New Technique for Ranking of Air Pollution Monitoring Stations in the Urban Areas Based upon Spatial Representativity (Case Study: PM Monitoring Stations in Berlin)

Hamid Taheri Shahraiyni[1,2*], Sahar Sodoudi[1], Andreas Kerschbaumer[3], Ulrich Cubasch[1]

[1] *Institut für Meteorologie, Freie Universität Berlin, Carl-Heinrich-Becker-Weg 6-10, 12165 Berlin, Germany*
[2] *Faculty of Civil Engineering, Shahrood University, Shahrood, Iran*
[3] *Senate Department for Urban Development and the Environment, Berlin, Germany*

ABSTRACT

The spatial representativity of monitoring stations plays a major role for the reasonable estimation of air pollutants. The ranking of air pollution monitoring stations based upon their spatial representativity identifies the level of representativeness of the stations and is very useful for developing optimum monitoring networks. In this study, a new ranking method, named RTFI (Ranking Technique based upon Fuzzy Interpolation) is introduced. This ranking method is able to rank air pollution monitoring stations in the urban areas based upon their spatial representativity. Although spatial correlation techniques are often used in the ranking techniques in order to consider spatial representativity, in this ranking technique, the spatial representativity of a station is not limited to its surroundings and is measured independently of its location. RTFI was applied to airborne Particulate Matter (PM) at seven stations in Berlin, and ranked them according to their spatial representativity. The results showed that the Neukölln-Nanenstr station (MC 42) is the most spatially representative station among the studied stations.

Keywords: Airborne particulate matter; Spatial representative; Monitoring network; Ranking Technique based upon Fuzzy Interpolation (RTFI); Background stations.

INTRODUCTION

One of the major objectives of the installation of air pollution monitoring networks in urban areas is the description of spatio-temporal concentrations of pollutants (van Egmond and Onderdelinden, 1981) and the evaluation of the exposure of people and other vulnerable receptors to pollution (Trujillo-Ventura and Ellis, 1991).

Because of both the high intensity of turbulence in the atmosphere and the changes in emissions, air pollution distribution is a function of space and time (Liu *et al.*, 1986). Accordingly, the spatial representativity of the monitoring stations plays a major role in the realistic estimation of air pollutants. Spatial estimation of pollutants with reasonable accuracy implies that the monitoring network is able to present spatially-resolved information; if not, there are some unsuitable stations in the monitoring network (van Egmond and Onderdelinden, 1981).

Many studies have been undertaken to obtain an optimum design of air pollution monitoring networks by employing different techniques (Castelton, 1984; Modak and Lohani, 1985; Trujillo-Ventura and Ellis, 1991, Haas, 1992; Kanaroglu *et al.*, 2005). The spatial representativity of an air pollution monitoring station can be expressed as the degree to which the records of the station resolves the air pollution variation across an area. The spatial correlation around each station and its spatial coverage is often used to express the spatial representativity of stations, and is often employed in the ranking techniques (e.g., Liu *et al.*, 1986; Janis and Robeson, 2004). The ranking of the stations in an installed monitoring network can help in the determination of non-spatially-representative stations. These stations can be replaced at sites which are more appropriate.

The European Union (2008) has emphasised that the monitoring sites are to be representative of the exposure of the general population, and are to avoid measuring very small micro-environments. At street level, measurements are to be representative of street segments of no less than 100 m length, and, at urban background level, are to represent all sources upwind of the station, and, in general, cover several square kilometres. (European Union, 2008).

In the previous studies on spatial representativity, each

* Corresponding author.
Tel.: +49-30-83854366; Fax: +49-30-83871160
E-mail address: hamid.taheri@met.fu-berlin.de

monitoring station has been evaluated individually. The air pollution at each point in an urban area is the result of the agglomeration of pollution of different sources. Each station can represent the effects of some pollution sources; hence, the combination of the stations (the network) can be very useful for the spatial representation of the different points in a given urban area. In this study, the interactions among stations are considered, and spatial representativity is expressed as a result of these interactions. Using this new approach, the share of each station in the general representation of the whole urban area is determined, and the stations are ranked based upon their spatial representativity. In this study, a new ranking technique, named RTFI (Ranking Technique based upon Fuzzy Interpolation) is presented and applied for the ranking of airborne Particulate Matter (PM) stations in Berlin based upon their spatial representativity.

BERLIN AND ITS MONITORING STATIONS

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 the beginning of the 1990s, Berlin had a high level of Total Suspended Particulate (TSP) (Lenschow *et al.*, 2001) and hence, a dense monitoring network with more than 40 stations (Fig. 1) was developed for the appropriate monitoring of the pollutants in Berlin (SenStadt, 1998). Both the concentration of TSP and the number of TSP monitoring stations decreased greatly until the end of 1990s (Lenschow *et al.*, 2001). In 1999, there were 18 TSP monitoring stations in Berlin. By 2013, there were only 12 stations (Fig. 1, circle and triangle stations) monitoring Particulate Matter less than 10 μm in aerodynamic diameter (PM_{10}). There are continuous PM data from 1990s until the present day in only 7 stations (Fig. 1, triangle stations) and the generation of the necessary input-output databases was possible using only these 7 stations as the input of the databases. The original representativity of these stations has been presented in Table 1. These 7 stations are ranked by RTFI in this study.

METHOD

In this section, first the algorithm of the ranking in RTFI is explained step by step, then the results of ranking by RTFI are presented and discussed.

Step 1. Database preparation: We tried to find old concurrent hourly particulate matter data for one-year from the 24 removed stations (squares in Fig. 1) (output variables) and the 7 stations still-operating (candidate input variables). A database of concurrent hourly PM data from the 7 stations (Fig. 1, triangle stations) and from each of the removed stations (Fig. 1, square stations) is generated as input variables and output variable, respectively. Thus, 24 (equal to the number of removed stations, presented by squares in Fig. 1) databases with seven-input variables and one-output variable are generated. This new ranking technique is applied to all the databases one by one. Thereafter, the ranking algorithm is explained for a single database.

A database (D) has n input variables ($X = X_1, X_2, \ldots, X_n$) and one output variable (Y). In our case study, the number of input variables (n) is 7. Thus D can be expressed as Eq. (1).

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

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 hourly PM observations.

Step 2. The database is randomly partitioned into a training database (two-thirds of the database) and a testing database (one-third of the database). 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 this ranking 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.

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

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

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

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 (D_k^{11} or D_k^{12}). Imagine D_k^{11} is selected for the division. D_k^{11} is divided into two smaller databases ($D_{k'}^{21}$, $D_{k'}^{22}$, $k' \in \{1, ..., n\}$). $D_{k'}^{21}$ 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 into three databases as below:

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

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

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

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

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.

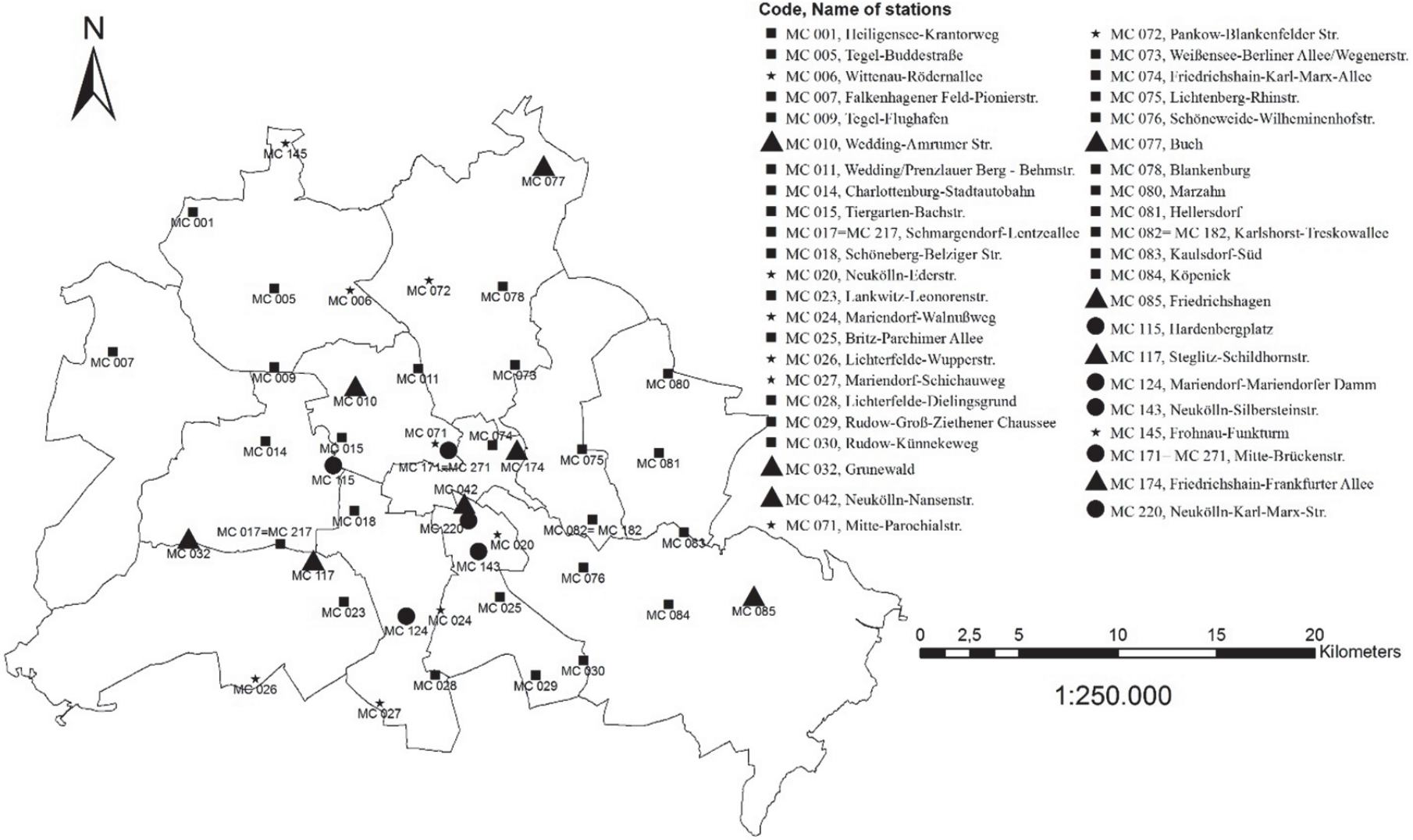

Fig. 1. Berlin map with the location of PM monitoring stations at the end of 2013 (Triangles and circles), utilized stations as input variables (7 Triangles) and output variables (24 Squares) of the ranking technique and the location of PM stations at the beginning of 1990s (Triangles, squares, stars).

Table 1. *FI* values of 7 PM stations with their ranks.

Station	MC 10	MC 32	MC 42	MC 77	MC 85	MC 117	MC 174
Original representativity	Background	Background	Background	Background	Background	Traffic	Traffic
FI	0.192	0.118	0.260	0.132	0.128	0.066	0.104
Rank based upon spatial representativity	2	5	1	3	4	7	6

The only remaining questions are which database is selected for the division into the two smaller databases in each step, and which X_k is the best one for the division of the selected database.

In order to determine the best X_k for the division of a database, all of the possible division options are performed. Hence, in order to divide the D_k^{ds} database 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.

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

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 Ink Drop Spread (Bagheri Shouraki and Honda, 1999). Similarly, the relationship between X_i and Y ($\hat{Y}_i = g_i^j(X_i)$, $i = 1, \ldots, n$) in all of S_{j2}^i is calculated. The accuracy of f_i^j and g_i^j functions for the estimation of output (Y) is evaluated. Consequently, f_z^j and $g_{z'}^j$ (z & $z' \in \{1, ..., n\}$) are determined as the best one-variable functions with the lowest errors, respectively. 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 $j = 1, ..., n$ is calculated and the minimum value in $\{e^1, ..., e^n\}$ 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 ($D_{k'}^{d+1,1}$, $D_{k'}^{d+1,2}$) 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 utilized 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).

$$\begin{cases} \text{If } X_k \le T_k^1 \text{ Then } \hat{Y}_1 = f_z^k\left(X_z\right) \\ \text{If } X_k > T_k^1 \text{ Then } \hat{Y}_2 = g_{z'}^k\left(X_{z'}\right) \end{cases} \qquad (10)$$

Using the testing 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 into 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 & $z_2 \in \{1, ..., n\}$) are determined and utilized for the output estimation in the two databases. Accordingly, D is divided into three databases (Eq. (5)) and a rule-base with three rules (Eq. (11)) is generated. The error of these one-variable functions are $e(f_{z_1}^{k'})$, $e(g_{z_2}^{k'})$ and $e(g_{z'}^{k'})$.

$$\begin{cases} \text{If } (X_k \le T_k^1 \ \wedge\ X_{k'} \le T_{k'}^2) \text{ Then } \hat{Y}_1 = f_{z_1}^{k'}\left(X_{z_1}\right) \\ \text{If } (X_k \le T_k^1 \ \wedge\ X_{k'} > T_{k'}^2) \text{ Then } \hat{Y}_2 = g_{z_2}^{k'}\left(X_{z_2}\right) \\ \text{If } X_k > T_k^1 \text{ Then } \hat{Y}_3 = g_{z'}^{k}\left(X_{z'}\right) \end{cases} \qquad (11)$$

Using the testing 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 rule-base generation are continued until $E_d > E_{d-1}$.

Step 5. The rule-base with d–1 rules is considered to be the best rule-base. In this rule base, the number of dividing times of each input variable is calculated, and is consequently expressed as the dividing vector ($\overrightarrow{DV}_1 = [dv_1, dv_2, \ldots, dv_n]$). In addition, the number of the estimated data 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 ($\overrightarrow{FV}_1 = [fv_1, fv_2, \ldots, fv_n]$).

Step 6. The training and testing databases can be combined to generate the original database, and then proceed to Step 2. Steps 2–6 are iterated according to the user defined number of iterations (n'). These iterations neutralise the effects of the random divisions in the second step and generalise the results. Thus, n' dividing vectors ($\overrightarrow{DV}_1, ..., \overrightarrow{DV}_{n'}$) and n' function vectors ($\overrightarrow{FV}_1, ..., \overrightarrow{FV}_{n'}$) are generated.

Step 7. The average of the n' dividing and function vectors are calculated ($\overrightarrow{DV}$, $\overrightarrow{FV}$). Then $\overrightarrow{DV}$ and $\overrightarrow{FV}$ are normalised as the sum of the elements in each vector as equal to 1. Finally, the average of two normalised vectors is calculated and called the initial importance vector ($\overrightarrow{IIV}$), and, consequently, each element of $\overrightarrow{IIV}$ is called the initial importance value (*IIV*). The range of *IIV*s is between 0.0 and 1.0 and *IIV*s are dimensionless.

Step 8. Steps 1–7 are performed for all 24 databases and 24 $\overrightarrow{IIV}$ are calculated. Final Importance Vector ($\overrightarrow{FIV}$) is calculated by averaging of the 24 $\overrightarrow{IIV}$. This vector shows the

Final Importance (*FI*) values of different input variables. The range of the *FI* values is between 0.0 and 1.0, and *FI* values are dimensionless. It is clear that the sum of the elements in $\overrightarrow{FIV}$ is equal to 1.

RESULTS

For the determination of suitable number of iterations (n'), one of the stations is randomly selected, and the *IIV*s for the 7 input stations (MC 10, MC 32, MC 42, MC 77 MC 85, MC 117 and MC 174) are calculated under a different n'. MC 78 was randomly selected and the *IIV*s for the 7 stations under $n' = \{1, 3, 6, 10, 15, 22, 30\}$ were calculated. The results of the changes of the *IIV*s for the 7 stations under the different n' have been presented in Fig. 2. This figure implies that, after the iteration of the algorithm 10 times, the *IIV*s converge to a relatively constant value and there is no significant variations in the *IIV*s for $n' \geq 10$. Hence, $n' = 10$ seems to be a suitable value and the Steps 2–6 of the algorithm were iterated 10 times.

In our case study, $\overrightarrow{FIV}$ shows the importance of the 7 stations based upon the spatial representativity of the 24 studied points and shows the relative degree of the 7 stations to resolve the PM in the 24 studied points. Consider that we are going to estimate or simulate the PM concentration in different parts of the urban area (i.e., the 24 stations, presented by the squares in Fig. 1) using the 7 studied stations. The multi-variate non-linear functions were developed using the 7 studied stations as input variables for the estimation of the output variables (24 stations). In fact, the *FI* values express the share of each input variables (the 7 studied stations) in the estimation of the 24 stations using the developed multi-variate non-linear functions. These 24 stations have been distributed in the urban area, and hence it can be roughly expressed that the $\overrightarrow{FIV}$ shows the spatial representativity of the 7 studied stations for PM estimation in Berlin.

The *FI* values of the seven studied stations with their corresponding ranks have been presented in Table 1. A primary particulate matter source at ground level can influence the surrounding areas within a radius of less than 100 m (Hewitt and Jackson, 2003), and the traffic has an immediate influence on the coarse particulate matter in the immediate vicinity of the stations. In addition, the European Union (2008) has pointed out that traffic stations are to be representative for a street segment of no less than 100 m in length, but urban background stations must be installed in the positions that are representative for several square kilometres. Hence, the RTFI technique, which ranks the stations based upon spatial representativity, is only suitable for the ranking of the urban background stations, and is not suitable for ranking the traffic stations, because the traffic sites are not spatially representative and are only locally (street pollution) representative. Therefore, when the spatial representativity of these stations is evaluated, these stations show the lowest spatial representativity values (*FI* values).

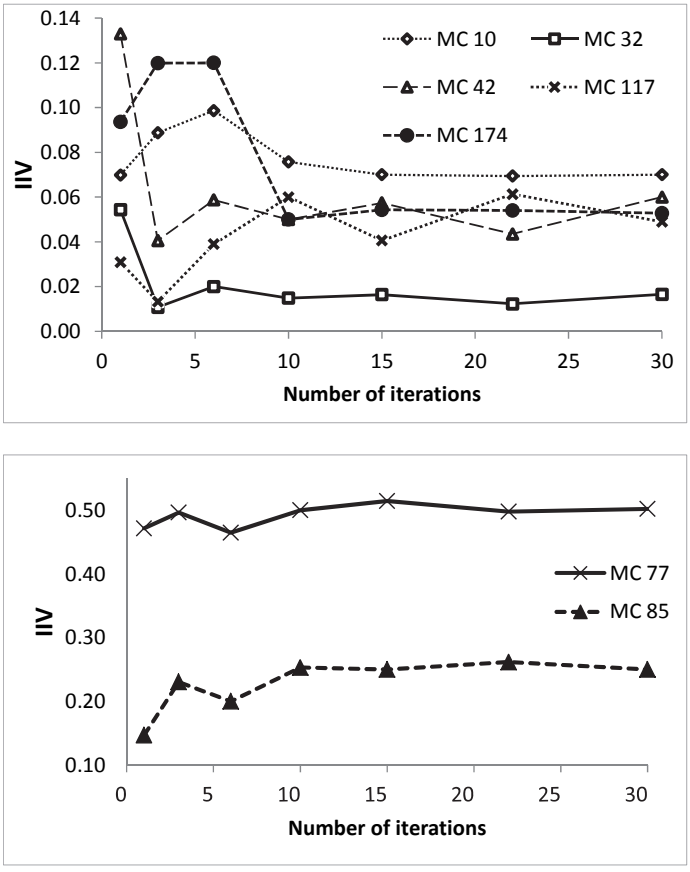

Fig. 2. The changes of *IIV*s of seven stations versus number of iterations for a randomly selected station (MC 78).

Although the evaluation of traffic stations based upon spatial representativity will show very low representativity, in this study, two stations in the traffic sites (MC 117 and MC 174) were utilised with urban background stations only for the evaluation of RTFI technique. If the RTFI technique ranks the stations appropriately, traffic stations must show the lowest spatial representativity (*FI* values). According to Table 1, the station with higher *FI* value has more importance for spatial representation and the MC 117 and MC 174 (traffic stations) have the lowest *FI* values, and this result implies the appropriate performance of the RTFI technique. Consequently, the RTFI technique can be employed as a new technique for the ranking of urban background stations based upon their spatial representativity. Among the installed stations in the background sites, the most spatial representative station in Berlin is Neukölln-Nanenstr station (MC 42) and the Grunewald station (MC 32) shows the least representativity.

Although spatial correlation techniques are often used in the ranking techniques to consider spatial representativity, in this new ranking approach, the ranking is calculated based upon the spatial representativity of the background stations roughly throughout the whole urban area.

CONCLUSIONS

Here, a new approach to the spatial representativity of background stations was presented. In this new approach, the non-linear interaction among background stations was considered for their ranking based upon their spatial representativity. RTFI (Ranking Technique based upon Fuzzy Interpolation) as a new ranking approach was introduced, and it is able to consider the interactions among all of the stations in a monitoring network. This ranking technique is capable of ranking the air pollution monitoring stations based upon spatial representativity measures the spatial representativity of each station throughout the whole urban area.

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 proofreading of this paper.

REFERENCES

Bagheri Shouraki, S. and Honda, N. (1999). Recursive Fuzzy Modeling Based on Fuzzy Interpolation. *J. Adv. Comput. Intell.* 3: 114–125.

Caselton, W.F. and Zidek, J.V. (1984). Optimal Network Monitoring Designs. *Stat. Probab. Lett.* 2: 223–227.

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. *Official J. Eur. Union* L 152: 1–44.

Haas, T.C. (1992). Redesigning Continental-scale Monitoring Networks. *Atmos. Environ.* 26A: 3323–3333.

Hewitt, C.N. and Jackson, A.V. (2008). *Handbook of Atmospheric Science: Principles and Applications*, John Wiley & Sons.

Janis, M.J. and Robeson, S.M. (2004). Determining the Spatial Representativeness of Air-temperature Records Using Variogram-nugget Time Series. *Phys. Geog.* 25: 513–530.

Kanaroglou, P.S., Jerrett, M., Morrison, J., Beckerman, B., Arain, M.A. and Gilbert, N.L. (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. and Reichenbacher, W. (2001). Some Ideas about the Sources of PM_{10}. *Atmos. Environ.* 35: 23–33.

Liu, M.K., Avrin, J., Pollack, R.I., Behar, J.V. and McElroy, J.L. (1986). Methodology for Designing Air Quality Monitoring Networks: I. Theoretical Aspects. *Environ. Monit. Assess.* 6: 1–11.

Modak, P.M. and Lohani, B.N. (1985). Optimization of Ambient Air Quality Monitoring Networks. *Environ. Monit. Assess.* 5: 21–38.

SenStadt (1998). *Air Quality Management in Berlin 1997.* Air Quality Management Series, Brochure No. 22, Department of Urban Development, Environmental Protection and Technology, Berlin.

Trujillo-Ventura, A. and Ellis, J.H. (1991). Multiobjective Air Pollution Monitoring Network Design. *Atmos. Environ.* 25A: 469–479.

van Egmond, N.D. and Onderdelinden, D. (1981). Objective Analysis of Air Pollution Monitoring Network Data; Spatial Interpolation and Network Density. *Atmos. Environ.* 15: 1035–1046.

Received for review, December 8, 2014
Revised, January 23, 2015
Accepted, January 23, 2015