



Vapour Pressure of Atmospheric Nanoparticles Using Genetic Algorithm-Partial Least Squares and Genetic Algorithm - Kernel Partial Least Squares

HADI NOORIZADEH* and ABBAS FARMANY

Ilam Branch, Islamic Azad University, Ilam, Iran

*Corresponding author: Tel: +98 918 3452507; E-mail: hadinoorizadeh@yahoo.com

(Received: 20 December 2010;

Accepted: 15 September 2011)

AJC-10420

The quantitative structure-property relationship (QSPR) of atmospheric nanoparticles against the comprehensive two-dimensional gas chromatography system coupled to time-of-flight mass spectrometry vapour pressure (P) was studied. A suitable set of molecular descriptors was calculated and the genetic algorithm (GA) was employed to select those descriptors that resulted in the best-fit models. The partial least squares (PLS) and the kernel partial least squares (KPLS) were utilized to construct the linear and nonlinear quantitative structure-property relationship models. The models were validated using leave-group-out cross validation (LGO-CV). The results indicate that genetic algorithm-kernel partial least squares can be used as an alternative modeling tool for quantitative structure-property relationship studies. This is the first research on the quantitative structure-property relationship of the nanoparticle compounds using the genetic algorithm-partial least squares and genetic algorithm-kernel partial least squares.

Key Words: Atmospheric nanoparticles, Vapour pressure, Comprehensive two-dimensional gas chromatography system, Time-of-flight mass spectrometry, Genetic algorithm-kernel partial least squares.

INTRODUCTION

Knowledge of the chemical composition of atmospheric nanoparticles (atmospheric aerosols) is essential to assess its impact on the environment, either by affecting the air quality in populated areas or by influencing the climate. Our understanding of the physical and chemical properties of aerosols is essential for properly assessing their effects on various issues such as human health, air quality and global climate and ultimately establishing effective control strategies. The effects of atmospheric aerosol particles on the environment and on human health are strongly dependent on their particle size and chemical composition^{1,2}. Atmospheric particles are also known to directly alter the Earth's radiative balance by scattering and absorbing solar radiation. Their heating or cooling effects depend on a number of properties such as number concentration, chemical composition and size³⁻⁵. The fine mode of the aerosol (diameter $d < 1 \mu\text{m}$) is of particular importance because fine particles are inhalable, they can interact with the solar radiation and provide the majority of cloud condensation nuclei. The major atmospheric nanoparticles include inorganic substances, such as sulphates and carbonaceous species. Carbonaceous aerosol, including elemental carbon (EC, a chemical structure similar to impure graphite) and organic carbon (OC,

a large variety of organic compounds), are important components of the atmospheric nanoparticles⁶.

For a time-series study, on the influence of organic aerosol compounds, it is necessary to have data of several compounds or groups of compounds at least with a daily resolution. Since most of the organic compounds occur in low concentrations in ambient aerosol, time-consuming analytical methods are required for their analysis. Comprehensive two-dimensional gas chromatography (GC \times GC) is a novel technique, whereby a sample is separated (in two dimensions) with two comprehensively coupled gas chromatographic columns. Two different chromatographic mechanisms (*i.e.* volatility and polarity) are used to separate the compounds in the two columns. A promising technique for analysis of air pollution research is GC \times GC coupled to time-of-flight mass spectrometry (TOFMS)^{7,8}. Due to the increased separation of GC \times GC with respect to one-dimensional GC, the mass spectra are of considerably increased quality (lower background level).

The problem of skewing of mass spectra in GC-MS experiments with scanning mass analyzers is also not present in time-of-flight mass spectrometry. Thus, TOFMS provide identical mass spectral patterns over a complete chromatographic peak for the same component. The TOFMS systems can readily

achieve the required spectral acquisition rates for reliable GC \times GC peak assignment and quantification^{9,10}. The combination of GC \times GC and TOFMS allowed detection of more than 10,000 individual organic compounds in aerosol samples¹¹.

Nanoparticle analytes are first separated by their differences in boiling point through the use of a conventional dimension low polarity column, then by their differences in polarity (exploiting activity coefficient differences with the 2D stationary phase) by using a higher polarity column. Vapour pressure (P) is an important parameter for assessing transport and fate of organic pollutants in the environment, because it is a measure of volatility of those organic compounds. In addition, vapour pressure is required for determination of other environmentally relevant physicochemical properties, such as air-water partition coefficient, Henry's law constant, enthalpy of vapourization *etc.*^{12,13}.

Among all methods, quantitative structure-property relationships (QSPR) are most popular. In quantitative structure-property relationships, the property of given chromatographic system was modeled as a function of solute (molecular) descriptors. A number of reports, deals with QSPR vapour pressure (P) calculation of several compounds have been published in the literature¹⁴⁻¹⁶. The quantitative structure-property relationships models apply to partial least squares method often combined with genetic algorithms for feature selection^{17,18}.

In the recent years, nonlinear kernel-based algorithms as kernel partial least squares have been proposed^{19,20}. The kernel partial least squares can efficiently compute latent variables in the feature space by means of integral operators and nonlinear kernel functions²¹. Compared to other non-linear methods, the main advantage of the kernel based algorithm is that it does not involve nonlinear optimization. It essentially requires only linear algebra, making it as simple as the conventional linear partial least squares (PLS). In addition, because of its ability to use different kernel functions, kernel partial least squares (KPLS) can handle a wide range of non-linearities. In the present study, GA-PLS and GA-KPLS were employed to generate QSPR models that correlate the structure of nanoparticles; with observed vapour pressure (P). The present study is a first research on QSPR of the nanoparticle compounds against the vapour pressure, using GA-PLS and GA-KPLS.

EXPERIMENTAL

The methodology applied in this study involved the following five steps: (i) collecting experimental data and splitting the compounds, for which the data was available, into a training set and a validation set; (ii) calculating molecular descriptors for all congeners and selecting the optimal pool of the descriptors to be utilized in the QSPR model; (iii) calibrating and simultaneously, internal validating the QSPR model, (iv) externally validating the developed model with use of the validation set, (v) applying the model to predict the values of vapour pressure for new compounds (those congeners, for which the experimental data have been unavailable so far).

Data set: Vapour pressure of 29 nanoparticle compounds from coniferous forest were taken from the literature²² are presented in Table-1. Atmospheric particulate samples were collected onto quartz fibre filters (diameter 240 mm, Munktell,

Grycksbo, Sweden) with a high volume sampler, placed on the ground. A GC \times GC-TOF-MS system were carried out on an Agilent 6890 gas chromatograph equipped with a split/splitless injector and a Pegasus II time-of-flight mass spectrometer (LECO, St. Joseph, MI, USA). The injector was used in splitless mode at 300 °C. Helium (HoekLGOs, Schiedam, the Netherlands, purity 99.999 %) in constant flow mode was used as carrier gas with head pressure of 170 kPa at 60 °C. The GC oven was temperature programmed as follows: 60 °C (5 min) then 5 °C min⁻¹ to 300 °C (8 min). A 20 m \times 0.25 mm I.D. HP-5MS (Agilent Technologies) column with film thickness of 0.25 μ m was used as the first column and a 0.7 m \times 0.1 mm I.D. BGB-1701 column (BGB Analytik) with film thickness of 0.1 μ m as the second column. Most of the identified compounds have vapour pressures above 10 Pa and should be in the gas phase. Identified compounds of these atmospheric aerosols include acyclic alkanes, alkenes, ketones, aldehydes, a few alcohols and acids and aromatic compounds. In present studies, data was resulted by nonpolar capillary used for QSPR research. The vapour pressure of these compounds was decreased in the range of 275.9 and 13 for both octanal and 2-undecanone, respectively.

TABLE-1
DATA SET AND THE CORRESPONDING OBSERVED AND PREDICTED VAPOUR PRESSURE VALUES BY GA-KPLS

No	Name	P _{Exp}	P _{Cat}	RE
Training set				
1	2-Undecanone	13.0	14	7.69
2	3-Methyl-Dodecane	14.1	15.3	8.51
3	Alpha, alpha,4-trimethyl-benzenemethanol	20.8	22.9	10.10
4	Alpha, alpha-dimethyl-benzenemethanol	24.4	23.6	3.28
5	Decanal	27.6	28.4	2.90
6	1-Dodecene	31.2	29	7.05
7	2-Decanone	33.1	34.5	4.23
8	Acetophenone	39.9	40.2	0.75
9	2-Methyl-undecane	40.1	43.7	8.98
10	5-Methyl-undecane	42.1	44.2	4.99
11	Nonanal	70.9	67.3	5.08
12	2-Octenal	73.6	67.6	8.15
13	2-Nonanone	86.0	97	12.79
14	Trans-decahydro-naphthalene	98.0	89	9.18
15	2-Ethyl-1,3-dimethyl-benzene	101.3	106.8	5.43
16	4-Ethyl-1,2-dimethyl-benzene	103.4	113.9	10.15
17	2,2,7,7-Tetramethyloctane	131.8	120.1	8.88
18	1-Ethyl-3,5-dimethyl-benzene	132.8	136.4	2.71
19	Butyl-benzene	140.0	139	0.71
20	1,3-Diethyl-benzene	153.3	146.8	4.24
21	1-Methyl-2-(1-methylethyl)-benzene	181.3	160.7	11.36
22	1-Methyl-4-(1-methylethyl)-benzene	219.9	250.7	14.01
23	2-Octanone	229.3	238	3.79
24	Octanal	275.9	250.4	9.24
Validation t set				
25	Benzothiazole	15.9	17.8	11.95
26	1-Hexanol,	27.6	22.6	18.12
27	cis-Decahydro-naphthalene	98.0	110	12.24
28	1-Methyl-3-propyl-benzene	138.6	159.4	15.01
29	1-Methyl-3-(1-methylethyl)-benzene	209.3	163.7	21.79

Computer hardware and software: All calculations were run on a HP Laptop computer with AMD Turion64X2 processor with windows XP operating system. The optimizations of molecular structures were done by the HyperChem 7.0 (AM1

method) and descriptors were calculated by Dragon Version 3.0 software's. MINITAB software version 14 was used for the simple PLS analysis. Cross validation, GA-PLS, GA-KPLS and other calculation were performed in the MATLAB (Version 7, Mathworks, Inc.) environment.

Genetic algorithm: A detailed description of the genetic algorithm can be found in the literature²³⁻²⁵. Genetic algorithm is simulated methods based on ideas from Darwin's theory of natural selection and evolution (the struggle for life). In genetic algorithm a chromosome (or an individual) can be defined as an enciphered entity of a candidate solution, which is expressed as a set of variables. Genetic algorithm consist of the following basic steps: (1) A chromosome is represented by a binary bit string and an initial population of chromosomes is created in a random way; (2) A value for the fitness function of each chromosome is evaluated; (3) Based on the values of the fitness functions, the chromosomes of the next generation are produced by selection, crossover and mutation operations. The fitness function was proposed by Depczynski *et al.*²⁶. The parameters for the algorithm are reported in Table -2.

Population size	30 chromosomes
On average, five variables per chromosome in the original population	
Regression method	PLS, KPLS
Cross validation	Leave-group-out
Number subset	4
Maximum number of variables selected in the same chromosome	PLS, 30
Elitism	True
Crossover	Multi point
Probability of crossover	50 %
Mutation	Multi point
Probability of mutation	1 %
Maximum number of components	PLS, 10
Number of runs	100

Non-linear models

Kernel partial least squares (KPLS): The KPLS method is based on the mapping of the original input data into a high dimensional feature space \mathfrak{S} where a linear PLS model is created. By nonlinear mapping $\Phi : x \in \mathfrak{X}^n \rightarrow \Phi(x) \in \mathfrak{S}$, a KPLS algorithm can be derived from a sequence of NIPALS steps and has the following formulation²⁷: (1) Initialize score vector w as equal to any column of Y ; (2) Calculate scores $u = \Phi\Phi^T w$ and normalize u to $\|u\| = 1$, where Φ is a matrix of regressors; (3) Regress columns of Y on u : $c = Y^T u$, where c is a weight vector; (4) Calculate a new score vector w for Y : $w = Yc$ and then normalize w to $\|w\|=1$; (5) Repeat steps 2-4 until convergence of w ; (6) Deflate Φ and Y matrices:

$$\Phi\Phi^T = (\Phi - uu^T\Phi) (\Phi - uu^T\Phi)^T \tag{1}$$

$$Y = Y - uu^TY \tag{2}$$

(7) Go to step 1 to calculate the next latent variable.

Without explicitly mapping into the high-dimensional feature space, a kernel function can be used to compute the dot products as follows:

$$k(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) \tag{3}$$

$\Phi\Phi^T$ represents the $(n \times n)$ kernel Gram matrix K of the cross dot products between all mapped input data points $\Phi(x_i), i = 1, \dots, n$. The deflation of the $\Phi\Phi^T = K$ matrix after extraction of the u components is given by:

$$K = (I - uu^T)K(I - uu^T) \tag{4}$$

where, I is an m -dimensional identity matrix. Taking into account the normalized scores u of the prediction of KPLS model on training data \hat{Y} is defined as:

$$\hat{Y} = KW(U^TKW)^{-1}U^TY = UU^TY \tag{5}$$

For predictions on new observation data \hat{Y}_t , the regression can be written as:

$$\hat{Y}_t = K_t W(U^TKW)^{-1}U^TY \tag{6}$$

where, K_t is the test matrix whose elements are $K_{ij} = K(x_i, x_j)$ where x_i and x_j present the test and training data points, respectively.

RESULTS AND DISCUSSION

Linear model

Results of the genetic algorithm-partial least squares (GA-PLS) model: To reduce the original pool of descriptors to an appropriate size, the objective descriptor reduction was performed using various criteria. Reducing the pool of descriptors eliminates those descriptors which contribute either no information or whose information content is redundant with other descriptors present in the pool. After this process, 1099 descriptors were remained. These descriptors were employed to generate the models with the GA-PLS and GA-KPLS program. The best model is selected on the basis of the highest multiple correlation coefficient leave-group-out cross validation (LGO-CV) (Q^2), the least root mean squares error (RMSE) and relative error (RE) of prediction and simplicity of the model. These parameters are probably the most popular measure of how well a model fits the data. The best GA-PLS model contains 5 selected descriptors in 3 latent variables space. These descriptors were obtained constitutional descriptors [(sum of atomic polarizabilities (scaled on carbon atom (SP))] and sum of atomic van der Waals volumes [scaled on carbon atom (Sv)], 3D-MoRSE descriptors [3D-MoRSE - signal 02 / weighted by atomic Sanderson electronegativities (Mor02e)], charge descriptors [Relative positive charge (RPCG)] and quantum chemical descriptors (polarizability). For this in general, the number of components (latent variables) is less than number of independent variables in PLS analysis. The obtained statistic parameters of the GA-PLS model were shown in Table-3. The PLS model uses higher number of descriptors that allow the

Model	Training set					Validation set				
	R ²	Q ²	RE	RMSE	N	R ²	Q ²	RE	RMSE	N
GA-PLS	0.887	0.881	11.98	18.40	24	0.782	0.784	26.87	38.07	5
GA-KPLS	0.958	0.958	6.84	10.63	24	0.868	0.875	15.82	23.17	5

model to extract better structural information from descriptors to result in a lower prediction error.

Non-linear model

Results of the genetic algorithm-kernel partial least squares (GA-KPLS) model: In this paper a radial basis kernel function, $k(x,y) = \exp(-\|x-y\|^2/c)$, was selected as the kernel function with $c = rm\sigma^2$ where r is a constant that can be determined by considering the process to be predicted (here r set to be 1), m is the dimension of the input space and σ^2 is the variance of the data²⁸. It means that the value of c depends on the system under the study. The 4 descriptors in 1 latent variables space chosen by GA-KPLS feature selection methods were contained. These descriptors were obtained constitutional descriptors [number of hydrogen atoms (nH)], functional group counts [number of hydroxyl groups (nROH)], molecular properties [topological polar surface area using N, O, S, P polar contributions (TPSA (Tot))] and quantum chemical descriptors [lowest unoccupied molecular orbital (LUMO)]. For the constructed model, four general statistical parameters were selected to evaluate the prediction ability of the model for the pressure. The statistical parameters R^2 , Q^2 , RE and RMSE were obtained for proposed models. Each of the statistical parameters mentioned above were used for assessing the statistical significance of the QSPR model. Inspection of the results reveals a higher Q^2 and lowers other values parameter for the training and validation sets GA-KPLS compared with their counterparts for GA-PLS. The GA-PLS linear model has good statistical quality with low prediction error, while the corresponding errors obtained by the GA-KPLS model are lower. Plots of predicted pressure *versus* experimental pressure values by GA-KPLS for training and validation set are shown (Fig. 1). Obviously, there is a close agreement between the experimental and predicted pressure and the data represent a very low scattering around a straight line with respective slope and intercept close to one and zero. This clearly shows the strength of GA-KPLS as a nonlinear feature selection method. This result indicates that the vapour pressure of nanoparticles possesses some non-linear characteristics.

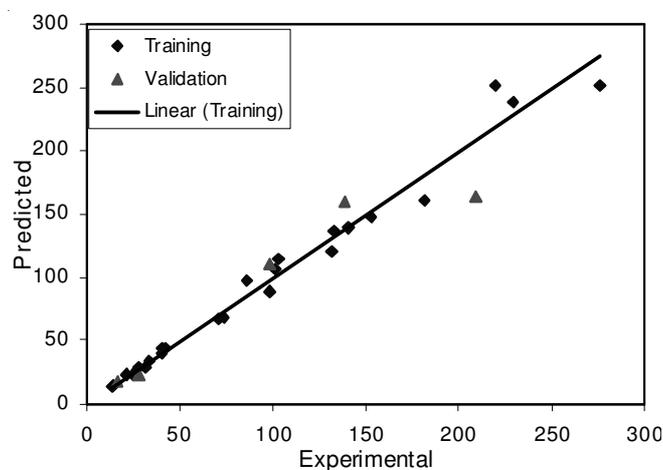


Fig. 1. Plot of predicted pressure obtained by GA-KPLS against the experimental values

Model validation: Validation is a crucial aspect of any QSPR modeling²⁹. The accuracy of proposed models was

illustrated using the evaluation techniques such as leave-group-out cross validation (LGO-CV) procedure.

Cross validation technique: Cross validation is a popular technique used to explore the reliability of statistical models. Based on this technique, a number of modified data sets are created by deleting in each case one or a small group (leave-some-out) of objects. For each data set, an input-output model is developed, based on the utilized modeling technique. Each model is evaluated, by measuring its accuracy in predicting the responses of the remaining data (the ones or group data that have not been utilized in the development of the model)³⁰. In particular, the leave-group out procedure was utilized in this study. A QSPR model was then constructed on the basis of this reduced data set and subsequently used to predict the removed data. This procedure was repeated until a complete set of predicted was obtained. The statistical significance of the screened model was judged by the correlation coefficient (Q^2). The predictive ability was evaluated by the cross validation coefficient (Q^2 or R_{cv}^2), which is based on the prediction error sum of squares (PRESS) and was calculated by following equation:

$$R_{cv}^2 \equiv Q^2 = 1 - \frac{\sum_{i=1}^n (y_i - y_i^{\wedge})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (7)$$

where, y_i , y_i^{\wedge} , and \bar{y} were the experimental, predicted and mean vapour pressure values of the samples, respectively. The accuracy of cross validation results is extensively accepted in the literature considering the Q^2 value. In this sense, a high value of the statistical characteristic ($Q^2 > 0.5$) is considered as proof of the high predictive ability of the model. However, this assumption is in many cases incorrect and can be that exist the lack of the correlation between the high LGO-CV Q^2 and the high predictive ability of QSPR models has been established and corroborated³¹. Thus, the high value of LGO-CV Q^2 appears to be necessary but not sufficient condition for the models to have a high predictive power. These authors stated that an external set is necessary. As a next step, further analysis was also followed for pressure of the new set of compounds using the developed QSPR model.

Validation through the external validation set: Validating QSPR with external data (*i.e.* data not used in the model development) is the best method of validation. However the availability of an independent external validation set of several compounds is rare in QSPR. Thus, the predictive ability of a QSPR model with the selected descriptors was further explored by dividing the full data set. The predictive power of the models developed on the selected training set is estimated on the predicted values of validation set chemicals. The data set was randomly divided into training (calibration and prediction sets) and validation sets after sorting based on the vapour pressure values. The training set consisted of 24 molecules and the validation set, consisted of 5 molecules. The training set was used for model development, while the validation set in which its molecules have no role in model building was used for evaluating the predictive ability of the models for external set. Table-3 shows the statistical parameters for the compounds obtained by applying the four models to the training and validation sets. The whole of these data clearly displays a significant improve-

ment of the QSPR model consequent to non-linear statistical treatment and a substantial independence of model prediction from the structure of the test molecule. In the above analysis, the descriptive power of a given model has been measured by its ability to predict pressure of unknown molecules. For instance, as to prediction ability, it can be observed in Fig. 1 that scattering of data points from the ideal trend in validation set is poor.

Interpretation of descriptors: The vapour pressure of a compound is determined by different interactions between molecules. These interactions include dispersion interaction, dipole-dipole interaction, dipole-induced interaction and hydrogen bonding interaction.

Constitutional descriptors are most simple and commonly used descriptors, reflecting the molecular composition of a compound without any information about its molecular geometry. The most common constitutional descriptors are number of atoms, number of bond, absolute and relative numbers of specific atom type, absolute and relative numbers of single, double, triple and aromatic bond, number of rings, number of rings divided by the number of atoms or bonds, number of benzene ring, number of benzene ring divided by the number of atom, molecular weight and average molecular weight.

The hydrogen bonding a measure of the tendency of a molecule to form hydrogen bonds. This is related to number of hydrogen atoms (nH). Hydrogen-bonding may be divided into an electrostatic term and a polarization/charge transfer term. A particularly strong type of polar interaction occurs in molecules where a hydrogen atom is attached to an extremely electron-hungry atom such as oxygen, nitrogen or fluorine.

3D-MoRSE (3D-Molecule representation of structures based on electron diffraction) descriptors are based on the idea of obtaining information from the 3D atomic coordinates by the transform used in electron diffraction studies. These descriptors are calculated by summing atom weights viewed by a divergent angular scattering function.

Although constitutional descriptors, 3D-MoRSE descriptors and functional group are often successful in rationalizing vapour pressure of nanoparticles, they cannot account for conformational changes and they do not provide information about electronic influence through bonds or across space. For that reason, quantum chemical descriptors are used in developing QSPR.

Quantum chemical descriptors can give great insight into structure and reactivity and can be used to establish and compare the conformational stability, chemical reactivity and inter-molecular interactions. They include thermodynamic properties (system energies) and electronic property (LUMO energy). Quantum chemical descriptors were defined in terms of atomic charges and used to describe electronic aspects both of the whole molecule and of particular regions, such atoms, bonds and molecular fragments. Electronic properties may play a role in the magnitude in a biological activity, along with structural features encoded in indexes. LUMO as an electron acceptor represents the ability to obtain an electron. The energy of the LUMO is directly related to the electron affinity and characterizes the susceptibility of the molecule toward

attack by nucleophiles. The LUMO energy can be interpreted as a measure of charge transfer interactions and/or of hydrogen bonding effects. Electron affinity was also shown to greatly influence the chemical behaviour of compounds, as demonstrated by its inclusion in the QSPR.

Polar functional groups account for many of the dipole-dipole, dipole-induced dipole and hydrogen bond interactions. Topological polar surface area (TPSA) also accounts for the steric shape of a molecule. The topological polar surface area is a surface descriptor, defined as the part of the surface area of a molecule contributed by nitrogen, oxygen and connected hydrogen atoms. As such, it is clearly related to the capacity of a drug to form hydrogen bonds.

Charge descriptors were defined in terms of atomic charges and used to describe electronic aspects both of the whole molecule and of particular regions, such atoms, bonds and molecular fragments. Electrical charges in the molecule are the driving force of electrostatic interactions and it is well known that local electron densities or charge play a fundamental role in many physico-chemical properties and receptors-ligand binding affinity. Thus, charge based descriptors have been widely employed as chemical reactivity indices or as measures of weak intermolecular interactions. Many quantum chemical descriptors are derived from the partial charge distribution in a molecule or from the electron densities on particular atoms. Relative positive charge (RPCG) is the quotient between maximum atomic positive charge in the molecule and positive atomic charge in the molecule. It contains electronic information to describe the molecule and therefore it encodes features responsible for interaction between molecules and the modified reversed stationary phase³².

From the above discussion, it can be seen that the particle size, hydrogen bonding and electrostatic interactions are the likely three factors controlling the pressure of these nanoparticles. All the descriptors involved in the model, which have explicit physical meaning, may account for the structure responsible for the vapour pressure of these compounds.

Conclusion

In this research, an accurate QSPR models for estimating the vapour pressure of atmospheric nanoparticles were developed by employing the linear model (GA-PLS) nonlinear model (GA-KPLS). The most important molecular descriptors selected represent the constitutional descriptors, functional group, molecular properties, charge descriptors and quantum chemical descriptors that are known to be important in the vapour pressure of nanoparticles. Two models have good predictive capacity and excellent statistical parameters. A comparison between these models revealed the superiority of the GA-KPLS to GA-PLS model. It is easy to notice that there was a good prospect for the GA-KPLS application in the QSPR modeling. This indicates that vapour pressure of nanoparticles possesses some nonlinear characteristics. The results showed that the GA-KPLS model can be effectively used to describe the molecular structure characteristic of these compounds. It can also be used successfully to estimate the vapour pressure for new compounds or for other compounds whose experimental values are unknown.

REFERENCES

1. G. Buzorius, A. Zelenyuk, F. Brechtel and D. Imre, *Geophys. Res. Lett.*, **29**, 1974 (2002).
2. B.J. Finlayson-Pitts and J.J.N. Pitts, *Chemistry of the Upper and Lower Atmosphere*, Academic Press, San Diego (2000).
3. R. Arimoto, *Earth Sci. Rev.*, **54**, 29 (2001).
4. S.P. Harrison, K.E. Kohfeld, C. Roelandt and T. Claquin, *Earth Sci. Rev.*, **54**, 43 (2001).
5. P. Penttinen, K.L. Timonen, P. Tiittanen, A. Mirme, J. Ruuskanen and J. Pekkanen, *Eur. Respirat. J.*, **17**, 428 (2001).
6. C.-S. Zhu, C.-C. Chen, J.-J. Cao, C.-J. Tsai, Ch.C.-K. Chou, S.-C. Liu and G.-D. Roam, *Atmos. Environ.*, **44**, 2668 (2010).
7. F. Adam, F. Bertoincini, N. Brodusch, E. Durand, D. Thiebaut, D. Espinat and M.-C. Hennion, *J. Chromatogr. A*, **1148**, 55 (2007).
8. T. Hyölyläinen, M. Kallio, M. Shimmo, K. Saamio, K. Hartonen and M.-L. Riekkola, in: *Presentation at the First International Symposium on Two-Dimensional Gas Chromatography*, Volendam, The Netherlands (2003).
9. C. von Muhlen, C.A. Zini, E.B. Caramao and P.J. Marriott, *J. Chromatogr. A*, **1200**, 34 (2008).
10. M. Adahchour, M. Brandt, H.-U. Baier, R.J.J. Vreuls, A.M. Batenburg and U.A.Th. Brinkman, *J. Chromatogr. A*, **1054**, 57 (2004).
11. J.F. Hamilton, P.J. Webb, A.C. Lewis, J.R. Hopkins, S. Smith and P. Davy, *Atmos. Chem. Phys.*, **4**, 1279 (2004).
12. T.E. Daubert and D.K. Jones, *AIChE Symp. Ser.*, **86**, 29 (1990).
13. A.D. Site, *J. Phy. Chem. Ref. Data*, **26**, 157 (1997).
14. A. Gajewicz, M. Haranczyk and T. Puzyn, *Atmos. Environ.*, **44**, 1428 (2010).
15. D. Ravindranath, B.J. Neely, R.L. Robinson and Kh. A.M. Gasem, *Fluid Phase Equilibria*, **257**, 53 (2007).
16. X. Zeng, Z. Wang, Zh. Ge and H. Liu, *Atmos. Environ.*, **41**, 3590 (2007).
17. H. Noorizadeh and A. Farmany, *Chromatographia*, **72**, 563 (2010).
18. K. Bodzioch, A. Durand, R. Kaliszan, T. Baczek and Y.V. Heyden, *Talanta*, **81**, 1711 (2010).
19. A. Niazi, S. Jameh-Bozorghi and D. Nori-Shargh, *J. Hazard. Mater.*, **151**, 603 (2008).
20. S.H. Woo, Ch. O. Jeon, Y.S. Yun, H. Choi, Ch. S. Lee and D.S. Lee, *J. Hazard. Mater.*, **161**, 538 (2009).
21. N. Krämer, A.L. Boulesteix and G. Tutz, *Chemom. Intell. Lab. Syst.*, **94**, 60 (2008).
22. M. Kallio, M. Jussila, T. Rissanen, P. Anttila, K. Hartonen, A. Reissell, R. Vreuls, M. Adahchour and T. Hyötyläinen, *J. Chromatogr. A*, **1125**, 234 (2006).
23. D.E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley-Longman, Reading, MA, USA (2000).
24. S. Riahi, E. Pourbasheer, R. Dinarvand, M.R. Ganjali and P. Norouzi, *Chem. Biol. Drug Des.*, **72**, 205 (2008).
25. J.A.D. Sousa, M.C. Hemmer and J. Gasteiger, *Anal. Chem.*, **74**, 80 (2002).
26. U. Depczynski, V.J. Frost and K. Molt, *Anal. Chim. Acta*, **420**, 217 (2000).
27. R. Rosipal and L.J. Trejo, *J. Mach. Learning Res.*, **2**, 97 (2001).
28. K. Kim, J.M. Lee and I.B. Lee, *Chemom. Intell. Lab. Syst.*, **79**, 22 (2005).
29. J. Acevedo-Martinez, J.C. Escalona-Arranz, A. Villar-Rojas, F. Tellez-Palmero, R. Perez-Roses, L. Gonzalez and R. Carrasco-Velaz, *J. Chromatogr. A*, **102**, 238 (2006).
30. A. Afantitis, G. Melagraki, H. Sarimveis, P.A. Koutentis, J. Markopoulos and O. Igglessi-Markopoulou, *Bioorg. Med. Chem.*, **14**, 6686 (2006).
31. A. Golbraikh and A. Tropsha, *J. Mol. Graphics Modell.*, **20**, 269 (2002).
32. R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, Germany (2000).