PATTERN RECOGNITION ANALYSIS FOR THE PRESCHOOL CHILDREN SOIL AND DUST INGESTION

XIAORU WANG, PENGYUAN YANG, DONGXIN YUAN, BENLI HUANG and RAMON M. BARNES*

Department of Chemistry, Xiamen University, Xiamen 361005, P.R. China

*Department of Chemistry, University of Massachusetts Amherst, MA 01003, USA

Abstract - A method of improved principal component analysis is applied to the pattern recognition process for the correlation between the amount of home and school soil and dust ingested by the preschool children and the concentrations of eight marker elements selected in the biological and environmental samples. The pattern profiles for the environmental element concentrations and the children soil ingestion are illustrated.

Key words soil ingestion by preschool children, pattern recognition, principal component regression

In the previous work [1,2], eight marker elements (Al, Ba, Mn, Si, Ti, V, Y and Zr) known or assumed to be poorly digested or absorbed by humans are measured with inductively coupled plasma atomic emission spectrometry (ICP-AES) and inductively coupled plasma mass spectrometry (ICP-MS) for 64 preschool children's food, biological (feces and urine), and their environmental samples (soil and dust) to quantify the amount of soil ingested by those children during their normal play activity while at home and in the preschool period [3]. The amount of soil and dust ingested by the children studied were calculated based upon the difference between food consumed and feces recovered. To estimate the amounts of soil and dust ingested, respectively, pairs of simultaneous equations were solved [2].

In the present work, pattern recognition analysis is applied to the soil and dust ingestion by those preschool children studied previously. The method of improved principal component analysis is used to calculate the mean amount of home soil and dust, preschool soil and dust ingested daily by children. These results are then analyzed by the pattern recognition process to predict the correlation between the amount of soil and dust ingested by the children and the concentrations of eight marker elements in their biological and environmental samples.

THEORETICAL BASIS

In general, a system of linear equations can be written as

$$AX = M \quad (1)$$

where X is the independence function matrix, A is the coefficient matrix, and M is the objective function matrix. The equation (1) is considered as a consistent system.

For the study of children soil ingestion, we have the element A_{ij} within the matrix A (8 X 4 in dimension), representing the concentration of j 'th element ($j=1\ldots 8$, rows for 8 elements) at given soil source i ($i=1\ldots 4$, columns for soil sources). Similarly, we have the element X_l within the matrix X (1 X 4), representing for the amount of soil ingested from the soil source i , and the element M_j within the matrix M (1 X 8) which is the total amount of j 'th element absorbed by child from the non-food materials. There are total 64 Eqn. (1) for all children ($l=1\ldots 64$, matrix number for each child).

In the present study the concentration matrix A and the objective matrix M are calculated from the experimental determinations [1], however the independence matrix X is unknown. Therefore, the process for resolving X is an inversion operation on Eqn. (1) in principle. Unfortunately, there may exist a large noise in matrices A and M due to the analysis errors introduced from the analysis process. These data scatter result in a group of morbid (or ill) equations. The principal component regression [4] is generally considered as a sophisticated method which can efficiently reduce the noise of matrix A .

In an improve principal component regression, Eqn (1) is modified to Eqn (2), as:

$$Aa = C \quad (2)$$

where, $a_i = X_i / \sum X_i$, a factor with no unit; while, $C_j = M_j / \sum X_i$, a concentration function with the same unit as the elements in the matrix A . Generally, the normalization condition is given by

$$\sum_{i=1}^4 a_i = 1 \quad (3)$$

Because the coefficient matrix A are consisted of the element concentration in four soils (i.e. home soil, dust and preschool soil and dust), it naturally shows a source pattern. Therefore the principal component analysis method can be applied to the matrix consisted of all matrices A to yield score Y (8 X 8 in dimension) for any 1'th child:

$$Y_{ki}^1 = \sum_{j=1}^4 (A_{ji}^1 - \bar{A}_k) v_{kj} \quad (4)$$

The same operation can be applied to the objective matrix C , leading to a pseudo-score Z (1 X 8):

$$Z_k^1 = \sum_{j=1}^4 (C_j^1 - \bar{A}_k) v_{kj} \quad (5)$$

In Eqn. (4) and (5) v_{kj} is a loading vector in the principal component analysis, and $k=1..8$. With the relations shown in Eqn (2), (3) and (4), Eqn (5) can be developed to:

$$Z_k^1 = \sum_{i=1}^4 \left(\sum_{i=1}^4 a_i A_{ji}^1 - \bar{A}_k \sum_{i=1}^4 a_i \right) v_{kj} = \sum_{i=1}^4 a_i Y_{ik}^1 \quad (6)$$

Therefore, the final system of equations for the principal component regression is deduced as:

$$Z - aY = 0 \quad (7)$$

Because both matrix Z and a contain the implicit unknown variable X_i , Eqn (7) is a non-linear equation.

In this study, Eqn (7) is solved by the method of non-linear least square regression with the consideration of eigenvalues in the weight factor. In addition, A_1 is chosen as a monitoring element, because of its proper concentration in all types of the samples and relatively less contamination from the analysis process. To do so, the results obtained based on Eqn (7) are verified by Eqn (1) for A_1 (e.g., $j=1$) only, which should satisfy the condition of minimal error. Finally, the pattern of soil ingested by each child will be predicted based on Eqn (4) [5]. Fortran 77 is adopted to compile the program for all calculations.

RESULTS AND DISCUSSION

Based upon the improved principal component analysis and calculation described in the previous sections, the two dimension patterns of home soil and dust, preschool soil and dust are illustrated in Fig.1. Because the principal components resulted from the original information of eight elements, which contain all informations contributed by each individual element, the most two important principal components (component 1 and component 2) containing the most informations for the overall pattern are plotted on the pattern profile in Fig.1, which serves as a training set. The results shown in Fig.1 indicate that the pattern profile of home dust is smaller than that of home soil, and is covered by that of home soil, which

means that the home dust might seriously contaminated by the home soil. The dusts from the individual homes show a certain generality due to the contact of the people with each other in the public area, which results in the patterns of home dust are more concentrated.

Fig 1 Overall pattern for all training sets

The predicted patterns of soil or dust ingested by the children studied are shown on the Fig.2 based on the training set. The figure illustrates that the pattern of soil or dust ingested by the children are also covered by the pattern profile of home soil, which obviously indicates that the children mostly ingested the home soil and home dust. Namely, the soil ingested by the children are correlated with the corresponding home soil and home dust.

Fig. 2 Pattern for children soil ingestion test set

The correlation relationship between the home soil, home dust and their activity time at home is given in Fig.3. Generally speaking, the longer the children stay at home, the larger amount of home soil and dust they ingested. However, there exist some accidental cases in which the children stay home but eat less or do not eat any soil at all. Those children are believed to be a little bit older, and can follow the advice (e.g., avoid from eating soil)

given by their parents. It seems that there barely is any relationship between the preschool soil and dust ingested by the children and the preschool activity time.

Fig. 3 Correlation between soil ingestion and home time weight factor

The home soil, dust, preschool soil and dust ingested by each individual child studied have been calculated. The results further verify that obtained previously [2]. Some results obtained in the present work, for instance the amount of soil ingested by child with sample number 0851 and 0864, are almost the same as that reported in the previous paper [2].

CONCLUSION

The method of modified principal component analysis is applied to the pattern recognition process for the correlation between the amount of soil and dust ingested by the children studied and the concentrations of eight marker elements in their biological and environmental samples. The results reply that there is a correlation between the amount of soil and dust ingested by the children studied and their home soil and dust. There are a few children who ingested extremely high amount of soil and dust.

The method described in the present work has the advantages over the previous method [2]: With the pattern profiles obtained by the principal components regression the correlation relationship between the amount of soil and dust ingested by the children studied and their environmental samples can apparently be concluded, and the results calculated supposed to be more accurate and reasonable.

REFERENCES

1. X. Wang, A. Lasztity, M. Viczian, Y. Israel and R. M. Barnes, J. of Anal. Atom. Spectr., 4, 727, 1989.
2. A. Lasztity, X. Wang, M. Viczian, Y. Israel and R. M. Barnes, J. of Anal. Atom. Spectr., 4, 737, 1989.
3. E. Calabrese, R. M. Barnes, E. J. Stanek, H. Pastides, C. Gilbert, P. Veneman, X. Wang, A. Lasztity and P. T. Kostecki, Regul. Toxic. Pharm. 10, 123, 1989.
4. N. Draper and H. Smith, Applied Regression Analysis, 2nd Ed. Wiley, New York, 1981.
5. M.A. Sharaf, D.L. Illman and B.R. Kowalski, Chemometrics, Wiley, New York, 1986.