Simultaneous quantitative estimation of four types of dissolved organic matters commonly seen in fresh water by using artificial neural network

Yuta Yamamoto[1], Masahito Yamamoto[2], Hiroshi Yamamura[3, *]

[1] Civil, Human and Environmental Science and Engineering Course - Graduate School of Science and Engineering - Chuo University - 1-13-27 Kasuga,Bunkyo-ku,Tokyo 112-8551, Japan

[2] Graduate School of Information Science and Technology - Hokkaido University - North 14, West 9, Kita-ku, Sapporo 060-0814, Japan

[3] Department of Integrated Science and Engineering for sustainable society - Faculty of Science and Engineering – Chuo University - 1-13-27 Kasuga,Bunkyo-ku,Tokyo 112-8551, Japan

* Corresponding author. Tel.: (+81)-3-3817-7257. E-mail address: yamamura.10x@g.chuo-u.ac.jp (H. Yamamura).

1. Introduction

Three-dimensional excitation emission matrix (EEM) analysis is effective as a simple and highly sensitive method for the characterization of dissolved organic matter (DOM). However, quantification of each DOM by EEM is difficult in the case where substances having close peaks or different spectral intensities are mixed.

Conventionally, N-way partial least squares (N-PLS) model which decomposes the independent and dependent variables into multilinear models so that the score of independent variable have a maximal covariance with that of dependent variable has been used for the simultaneous quantification of multiple fluorophores (Kumar et al 2011, Dinç et al 2017).

Besides N-PLS model, artificial neural network (ANN) is also used as a powerful regression model. ANN is a nonlinear model having a structure with three or more layers built with a set of nodes. A model is built by gradient descent method optimizing the weight values which represent the relationship between each layer. To our limited knowledge, no research has developed an ANN quantitative method which simultaneously quantifies the most common DOMs (humic acid, fulvic acid, tryptophan, tyrosine) containing in freshwater.

In this study, a method which quantifies four kinds of DOMs simultaneously from the EEM spectra using ANN is performed. EEM spectra was measured for 126 kinds of solutions with known concentrations mixed with four kinds of DOMs at an arbitrary ratio and N-PLS and ANN were performed on randomly selected data sets with the number of data 32, 64, and 126. The effectiveness of the developed method was examined by comparing the predictability of N-PLS model and ANN model.

2. Materials and Methods

Noise which was less than intensity 37.72 was replaced with 0 in all the measured spectra. Centering and scaling of the data was done. Root mean squared error (RMSE) was used in order to evaluate the predictability of the model. A trial which calculates the RMSE of the model built with an arbitrary number of data was conducted 10 times, and the average of the RMSE was used as the predictability of the model. The

minimum value of RMSE of N-PLS models built with the number of principal components 1 to 20 respectively using leave-one-out method was selected as a predictability of the model. The N-PLS model was built with N-way Toolbox of MatLab (2018 R2018a). The matrix (126×9951) unfolded from three-dimensional data set was fed into an input layer of the ANN model. Hyper parameters (number of hidden layers (1, 2, 3, 4) and nodes (32, 64, 128)) were optimized using leave-one-out method. Rectified linear unit was used as an activation function. Weight was estimated with Adam algorithm.

3. Result & Discussion

The RMSE of the N-PLS model built with the number of principal component 15 was the smallest, and the RMSE was 0.31 mg-C/L. The average values of RMSE of 10 trials defined in materials and methods with the number of data 32 and 64 were 0.35 mg-C/L, 0.29 mg-C/L, respectively in N-PLS model.

One hidden layer and the number of nodes 64 were selected as optimized hyper parameters of the ANN model built with the number of data 126, and the RMSE of the model was 0.41 mg-C/L. ANN model was built with the number of data 32 and 64 respectively using hyper parameters optimized with the number of data 126. The average values of RMSE of 10 trials with the number of data 32 and 64 were 0.78 mg-C/L and 0.51 mg-C/L, respectively in ANN model. The predictability of the N-PLS model was higher than that of the ANN model since RMSE of N-PLS model was lower than that of ANN model in all the number of data 32, 64, and 126.

In the trilinear tri-PLS model, a model is built using independent variables expressed as a product of a principal component vector and two weight vectors. Since tri-PLS model uses information in all dimensions for decomposition of independent and dependent variables, it is easier to interpret, and it is possible to build a model with less influence of noise than a model which unfolds those variables (Bro 1996).

When unfolding multidimensional independent variables, the structure and features of multidimensional data are lost in independent variables (Singh et al 2007). In the ANN model, it was assumed that the input data has lost the three-dimensional information since all the three-dimensional sample data matrices were unfolded into two-dimensional vectors. The data unfolded from three dimensions into two dimensions was used as an input of ANN model. Calculation of weight optimization becomes more complicated as the dimension of the input layer increases, and the complex model is more difficult to interpret than the simple model (Bowdena et al 2005). In the ANN model, the interpretation of the model was difficult since the dimension of the input layer was large (9951-dimensions). It is assumed that the dimension of the input layer should be decreased in order to facilitate the interpretation of the model.

4. Conclusion

In this study, a method which quantifies four kinds of DOMs simultaneously from the EEM spectra using ANN was performed. The conventional method N-PLS was performed in order to confirm the effectiveness of the developed model. The predictability of the N-PLS model was higher than that of the ANN model in the conditions of this study. Since three-dimensional EEM spectra data was unfolded into two-dimensional data, the structure and characteristics of the three-dimensional data has been lost in the input data of ANN model. On the other hand, the N-PLS model became simpler and easier to interpret than the ANN model since the N-PLS model was built using independent and dependent variables keeping the three-

dimensional structure. Moreover, as the dimension of the input layer became larger (9951-dimensions), the ANN model became complicated and interpretation became more difficult. In order to facilitate the interpretation of the ANN model, it is assumed that the dimensionality reduction of the fluorescence spectra data is required.

References :

Keshav Kumar & A. K. Mishra. (2011). Simultaneous quantification of dilute aqueous solutions of certain polycyclic aromatic hydrocarbons (PAHs) with significant fluorescent spectral overlap using total synchronous fluorescence spectroscopy (TSFS) and N-PLS, unfolded-PLS and MCR-ALS analysis. Anal. Methods, 2011, 3, 2616–2624

Erdal Dinç, Zehra Ceren Ertekin, Eda Büker. (2017). Multiway analysis methods applied to the fluorescence excitation-emission dataset for the simultaneous quantification of valsartan and amlodipine in tablets. Molecular and Biomolecular Spectroscopy 184 (2017) 255–261

RASMUS BRO. (1996). MULTIWAY CALIBRATION. MULTILINEAR PLS. JOURNAL OF CHEMOMETRICS,VOL. 10,47-61 (1996)

Kunwar P. Singh, Amrita Malik, Nikita Basant, Puneet Saxena. (2007). Multi-way partial least squares modeling of water quality data. Analytica Chimica Acta 584 (2007) 385–396

Gavin J. Bowden, Graeme C. Dandy, Holger R. Maier. (2005). Input determination for neural network models in water resources applications. Part 1—background and methodology. Journal of Hydrology 301 (2005) 75–92