Machine learning approaches for simultaneous spectrophotometric determination of heavy metal ions in water samples

Nguyen Thi Lan Anh[1*], Bui Phuong Thi[2], Do Thi Nhat Quyen[2], Vu Quynh Thu[2],Nguyen Thu Huong[1], Khuat Hoang Binh[1], Khong Manh Hung[1], Nguyen Chi Thanh[3], Ta Thi Thao[2]

[1]Institute of Chemistry and Materials, Academy of Military Science and Technology, 17 Hoang Sam, Cau Giay, Hanoi, Vietnam;
[2]Faculty of Chemistry, VNU University of Science, 19 Le Thanh Tong, Hoan Kiem, Hanoi, Vietnam;
[3]Institute of Information Technology, Academy of Military Science and Technology, 17 Hoang Sam, Cau Giay, Hanoi, Vietnam.
*Corresponding author: lananh.chemhus@gmail.com
Received: 19 Jan. 2024; Revised 13 Mar. 2024; Accepted 28 Mar. 2024; Published 20 May 2024.
DOI: https://doi.org/10.54939/1859-1043.j.mst.95.2024.47-54

ABSTRACT

In this study, the simultaneous determination of Co, Cd, Ni, Cu, and Pb was carried out as a color complex with 4-(2-pyridylazo) resorcinol in an aqueous solution under the assesting of machine learning. A partial least-squares multivariate linear regression and artificial neuron network for the analysis of mixtures of metals were developed. MATLAB is a powerful software machine learning program that was used to support matrix calculations and displays. The benefit of MATLAB in the construction of the machine learning model allows the development of a rapid and highly effective analysis of multiple components in the mixtures without separation and enrichment. For individual determinations, the working ranges were discovered as the important information for choosing the initial concentration of each heavy metal in a mixture, r. The results of analysis of Ni^{2+}, $Pb^{2+,}$ and Cd^{2+} by two methods Partial Least Squares - PLS and Artificial Neural Networks - ANN are sensitive and accurate for simultaneous determination of the concentration of these ions in the synthesis mixture with a high regression coefficient of 0.993, respectively, 0.997, 0.997 for Ni^{2+}, Pb^{2+} and Cd^{2+}. As for Cu^{2+} and Co^{2+}, the accuracy is higher when using the ANN method.

Keywords: Simultaneous determination; Heavy metal; ANN method; 4-(2-pyridylazo) resorcinol.

1. INTRODUCTION

The increasing of heavy metal content in the human body due to economic activities makes human health worse [1-4]. While mercury is the heavy metal with the greatest impact on the pollution of ecosystems worldwide [5, 6], lead affects children by leading to underdevelopment of the brain's gray matter, which in turn leads to a poor IQ [7]. Copper can destroy red blood cells due to its oxidizing properties and inhaling dust containing cadmium rapidly leads to problems with the respiratory system and kidneys, which can lead to death (usually from kidney failure) [8].

The determination of the metals is normally carried out on a variety of samples and by different methods [9-14]. These methods require very expensive equipment and reagents along with very specialized skills required [15]. Therefore, the use of UV-VIS spectroscopy for metal determination is increasingly popular because of its fast determination, low cost and simple technique [16-19]. However, when using this method, we will encounter difficulties due to overlapping absorption spectra [20]. So, partial least squares (PLS) is used [21, 22]. The PLS model is considered as a powerful and complete

multivariate linear statistical tool for quantitative spectrum analysis thanks to the characteristics and advantages of the remaining models. However, if in the mixture, there is a mutual interaction of components that causes the loss of additive properties of the measured signal, then a nonlinear multivariable regression must be used, such as artificial neural networks (ANN) [23-26]. Among organic reagents could be used for heavy metal determination by spectrophotometry, 4- (2-pyridylazo) resorcinol (PAR) which is a chelator used to determine the concentration of various heavy metal ions [27, 28].

Currently, one of the new research directions to simultaneously identify many components in the same mixture is to use nonlinear or linear multivariate regression models. In this paper, MATLAB is used to build multivariable regression algorithms to process data in the simultaneous determination of several heavy metal ions by UV-VIS spectroscopy in synthetic sample synthesis.

2. PROBLEM

2.1. Experiment preparation

2.1.1. Instrumentation

Spectrophotometric measurements were performed on a UV-VIS 1601PC (Shimadzu) spectrophotometer equipped with a 1-cm glass cuvette. Measurement of pH was carried out on a HANNA Instrument 211 microprocessor pH meter. All weighing was made using an analytical balance (Scientech SA 120), to an accuracy of 0.0001 g.

All absorption spectra were saved and subsequently exported UV-Win PC software to the Microsoft Excel program for statistical manipulation. Chemometric assisted spectrophotometric measurements were performed using the software MATLAB R2020b.

2.1.2. Reagents and solutions

All chemicals were analytical-grade and double distilled water was used to make up all solutions. 1000 mg/L standards of single heavy metal solutions (Cu^{2+}, Ni^{2+}, Co^{2+}, Pb^{2+}, and Cd^{2+}) from Merck, Darmstadt, Germany. The borax buffer solution (pH 10) was prepared by mixing 3.092 g phosphorous acid with 3.728 g potassium chloride. Then, this mixture was transferred to a 1L volumetric flask, added double distilled water and shaken well. Finally, it was made up to the mark and adjusted with a pH meter. For the preparation of $7.5x10^{-4}$ M PAR reagent solution ($C_{11}H_8N_3NaO_2.H_2O$, molecular weight = 255 g/mol), 0.0404 g PAR *was* added to a 250 mL volumetric flask and diluted with double distilled water.

2.1.3. Apparatus

Different size beakers, measuring cylinders, micropipettes, volumetric flasks, funnel, oven, Erlenmeyer flasks (different sizes), and filter papers.

2.2. Procedures

2.2.1. Individual calibration

The stock solutions were diluted using the standard solution of heavy metals to get a suitable concentration range of 0.1 – 1.4 mg/L of Cu^{2+}, 0.1 – 1 mg/L of Ni^{2+}, 0.1 – 1.2

mg/L of Co^{2+}, 0.5 – 6 mg/L of Pb^{2+} and 0.1 – 1.6 mg/L of Cd^{2+} to investigate absorbance ranges and select some values for constructing calibration curves.

2.2.2. Multivariate calibration

To have a colored complex solution, 5 mL of PAR $7.5x10^{-4}$ M and 5 mL borax buffer solution (pH 10) were added. The different volume of standard heavy metal ions was added into the 25 ml volumetric flask to obtain solutions in the appropriate concentration range. After 20 minutes, the absorbance was measured by UV-VIS double beam spectrophotometer. The UV absorption spectra for building calibration curves were recorded at the respective wavelength. That of all mixtures over the wavelength range from 450 to 600 nm with 1 nm data interval at 151 wavelengths.

2.2.3. Data Processing Using Machine Learning Methods

The machine learning model has four data matrices including two concentration matrices and two signal matrices.

C (80 x 5) is a concentration matrix of 80 mixtures of 5 components (training set). C_{test} (20 x 5) is a concentration matrix of 20 mixtures of 5 components (test set). A (80 x 151) is an absorbance matrix of 80 mixtures (training set) at 151 wavelengths. Atest (20 x 151) is an absorbance matrix of 20 mixtures (test set) at 151 wavelengths. After the necessary training and test matrices had been built, these data matrices were imported into calculation programs corresponding to different machine learning methods in MATLAB software.

2.2.4. Validation of Machine Learning Methods

Statistical analysis was performed for the developed models. The accuracy of the method was evaluated by trueness and precision. Trueness is determined by the difference between the results obtained by the machine learning model and the concentrations in the training and test samples and is evaluated by the correlation coefficient R of these two results. Precision is evaluated through the root mean square error of the model. Parameters like regression coefficient (R^2) and root mean square error of calibration (RMSEC), root mean square error of cross-validation (RMSECV) were calculated as follows.

The correlation coefficient (R) between the results calculated by the model (x) and the actual contents of heavy metals in the training or test set (y) is indicated by the formula:

$$\boldsymbol{R} = \frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}} \quad (1)$$

(n is the number of the given dataset)

The regression coefficient (R^2) is the R-squared formula.

If R^2 is nearest to 1, the method has high accuracy.

The formula of the root mean square error of calibration (RMSEC) and root mean square error of cross-validation (RMSECV) are shown as follows. The lower the RMSEC and RMSECV, the more accurate the analysis results were obtained.

$$RMSEC = \sqrt{\frac{\sum(y-x)^2}{N-f-1}} \quad (2); \qquad RMSECV = \sqrt{\frac{\sum(y-x)^2}{M-1}} \quad (3)$$

N is the number of mixtures in the training set,

M is the number of mixtures in the test set, and f is LVs.

3. RESULTS AND DISCUSSION

Since the absorption spectra of Cu-PAR, Ni-PAR, Co-PAR, Pb-PAR, and Cd-PAR complexes (shown in figure 1) have extensive overlap, it is not possible to perform simultaneous determination by usual methods by using a calibration curve. Therefore, multiple regression should be applied to the simultaneous determination and analysis of five substances in a mixture.

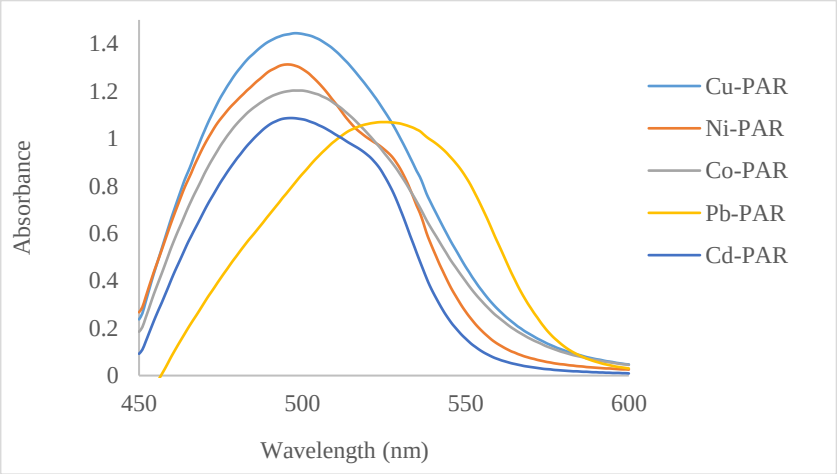


Figure 1. *Overlay absorption spectra of HM-PAR.*

The results of the absorbance ranges shown in table 1 indicated that the selected concentration ranges were suitable for the least error in the absorbance measurement. If the absorbance of the sample is higher than 2, the UV-VIS used will export the result with a big error.

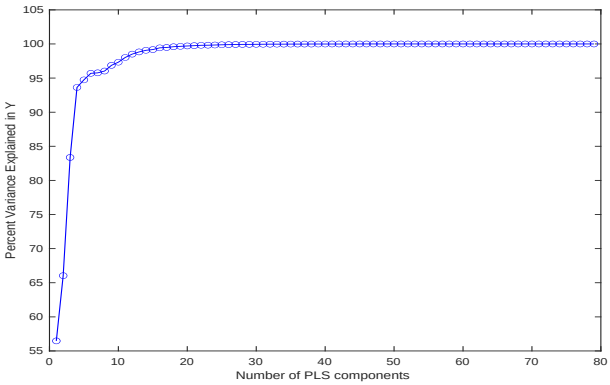
Table 1. *Absorbance ranges correspond to selected concentration ranges.*

Heavy metal	**Cu**	**Ni**	**Co**	**Pb**	**Cd**
Concentration range (mg/L)	0.1 – 1.4	0.1 – 1	0.1 – 1.2	0.5 – 6	0.1 – 1.6
Absorbance range	0.063 – 1.216	0.165 – 1.216	0.051 – 1.126	0.078 – 1.164	0.012 – 1.130

To select the number of factors in the PLS algorithm (model the system without overfitting the concentration data), a cross-validation method, leaving out one sample at a time, was used. Given the set of 80 calibration spectra, the PLS calibration on 79 spectra was performed, and using this calibration, the concentration of the compounds in the sample left out during calibration was predicted. This process was repeated 80 times until each calibration sample had been left out once. As can be seen in figure 2, it can be seen that nearly 95% of the variance of the calibration matrix was in the ninth PLS component. Thereby, this number was selected to investigate the optimal number of PLS components (as known latent variables – LVs). Compare the obtained content of heavy metals in the test sample through the selected LVs and the prepared concentration of that, then calculate quantities as R^2 and RMSECV. If these values show unsatisfactory, change different LVs from 9 to 15 to select the appropriate regression model and number of LVs. The results are depicted in table 2.

Table 2. *Statistical parameters obtained by the PLS method.*

Statistical parameters	Cu	Ni	Co	Pb	Cd
Concentration range (mg/L)	0.1 – 1.4	0.1 – 1	0.1 – 1.2	0.5 – 6	0.1 – 1.6
No.of factors (LVs)	11	9	9	12	12
R^2	0.255	0.993	0.589	0.997	0.997
RMSECV	0.280	0.018	0.205	0.054	0.031



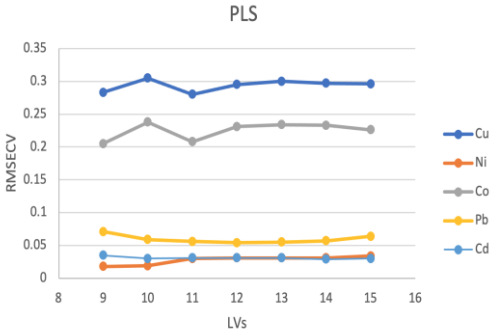


Figure 2. *Plot of % variance explained vs. LVs.* ***Figure 3.*** *RMSECV graph for PLS.*

Table 3. *Statistical parameters obtained by the ANN method.*

Heavy Metals	Cu	Ni	Co	Pb	Cd
R^2	0.724	0.958	0.781	0.997	0.995
RMSECV	0.176	0.044	0.162	0.056	0.038

The predicted concentration of the compounds in each test sample was compared with the known concentration of the compound in this sample and the root mean square error of cross-validation (RMSECV) was calculated. The plot of PLS components versus RMSECV is shown in figure 3. The correlation coefficients are 0.993, 0.997, and 0.997 for Ni^{2+}, Pb^{2+} and Cd^{2+}, respectively, which again verify the good performance of the PLS model in predicting the concentrations of these cations in mixture solutions. However, for the determination of Cu and Co, this result is still not good with R^2 values of 0.255 and 0.589 for Cu and Co, respectively. So, nonlinear model -artificial neural networks (ANN) methods were used. The training class is randomly divided into three parts at the ratio of 80:10:10 which are train, cross-validate, and test, respectively. The training dataset is used to calculate the gradient and continuously update the loadings and errors of the network being trained. The cross-validation dataset is used to monitor the training process.

The values of R^2 and RMSECV parameters were also used to evaluate the simultaneous quantity of heavy metals in 20 synthesis mixtures of the test set using the ANN method. Table 3 describes these quantities and the predictions of the test set involving nickel, cobalt, cadmium, lead, and copper through the training algorithms. The R^2 value of Cu and Co increased to 0.724 and 0.781, respectively, indicating the accuracy of the ANN method in separating Cu^{2+} and Co^{2+}.

4. CONCLUSIONS

An analytical method has been successfully developed for the multicomponent spectrophotometric determination based on their complexation with PAR in a borax buffer medium. The analytical results of Ni^{2+}, Pb^{2+}, and Cd^{2+} with all machine learning methods used (PCR, PLS, and ANNs) are sensitive and accurate for the simultaneous determination of these ion contents at self-made mixtures with high regression coefficients, approximate 0.993, 0.997, and 0.995 for Ni^{2+}, Pb^{2+}, and Cd^{2+}, respectively. For copper and cobalt, the accuracy was improved significantly when ANNs were replaced for PCR or PLS. The R^2 value had a significant increase to 0.724 for Cu^{2+} and 0.781 for Co^{2+}.

This study has demonstrated that artificial neural networks (ANNs) can be employed in environmental monitoring, particularly different water samples. ANNs have successfully and simultaneously determined qualitatively and quantitatively some heavy metals that could be found commonly in natural water sources. The concentrations of three heavy metals, namely cadmium, lead, and nickel, contained in synthesis water samples were successfully simultaneously determined, although their absorption spectra seriously overlapped under the experimental conditions. Hence, the results can become the key concepts of future procedures to overcome the main drawback related to the significant degree of spectral overlap of the constitutive ions in various real samples.

Acknowledgment: *This work was financially supported by the Institute of Chemistry and Materials. The authors would like to thank Professor Ta Thi Thao and the laboratory at HUS (Hanoi University of Sciences) for their support.*

REFERENCES

[1]. Duruibe, J. O., Ogwuegbu, M. O. C., Egwurugwu, J. N., *"Heavy metal pollution and human biotoxic effects"*, International Journal of Physical Sciences, Vol. 2 (5), pp.112-118, (2007).

[2]. Muhammad Aqeel Ashraf et al., *"Speciation of heavy metals in the surface waters of a former tin mining catchment"*, Chemical Speciation & Bioavailability, 24, pp.1-12, (2015).

[3]. Herawati N, Suzuki S, Hayashi K, Rivai If, Koyoma H. Cadmium, *"Copper and zinc levels in rice and soil of japan, indonesia and china by soil type"*, Bulletin of Environmental Contamination and Toxicology, 64, pp.33-39, (2000).

[4]. He Zl, Yang Xe, Stoffella Pj., *"Trace elements in agroecosystems and impacts on the environment"*, Journal of Trace Elements in Medicine and Biology, 19(2–3), pp.125-140, (2005).

[5]. Korbas, M.; O'donoghue, J.L.; Watson, G.E.; Pickering, I.J.; Singh, S.P.; Myers, G.J.; Clarkson, T.W.; George, G.N., *"The chemical nature of mercury in human brain following poisoning or environmental exposure"*, Acs Chem. Neurosci., 1, pp.810–818, (2010).

[6]. Zhou, Y.; Vaidya, V.S.; Brown, R.P.; Zhang, J.; Rosenzweig, B.A.; Thompson, K.L.; Miller, T.J.; Bonventre, J.V.; Goering, P.L., *"Comparison of kidney injury molecule-1 and other nephrotoxicity biomarkers in urine and kidney following acute exposure to gentamicin, mercury, and chromium"*, Toxicol. Sci., 101, pp.159–170, (2007).

[7]. Strong FC, Martin NJ, *"Rapid determination of zinc and iron in food by flow – injection analysis with flame atomic – absorption spectrophotometry and slurry nebulization"*, Talanta 7:11-718, (1990).

[8]. Šmirjákova, S., Ondrašovičová, O., Kašková, A., Laktičová, *"The effect of cadmium and lead pollution on human and animal healths"*, 49, 3: — Supplementum, S31—S32, (2005).

[9]. Davidson, C. M. *"Methods for the Determination of Heavy Metals and Metalloids in Soils"*. Heavy Metals in Soils, pp.97–140, (2012).

[10]. Khamms AA, Al-Ayash AS, Jasin F. *"Indirect electrothermal atomization AAS Spectrometric determination of drugs, desferrioxamine in some pharmaceutical preparations using Vanadium (V) as a mediatory element elestial"*. J. Anal. Chem. 3, pp.257- 269, (2009).

[11]. Davidson, C. M. *"Methods for the determination of heavy metals and metalloids in soils"*. Heavy metals in soils: Trace metals and metalloids in soils and their bioavailability, 97-140, (2013).

[12]. Makedonski, L., Peycheva, K., & Stancheva, M. *"Determination of heavy metals in selected black sea fish species"*. Food Control, *72*, 313-318, (2017).

[13]. Ugulu, I. *"Determination of heavy metal accumulation in plant samples by spectrometric techniques in Turkey"*. Applied Spectroscopy Reviews, 50(2), 113-151, (2015).

[14]. Adebayo, I. A. *"Determination of heavy metals in water, fish and sediment from Ureje water reservoir"*. Journal of Environmental & Analytical Toxicology, *7*(4), 1-4, (2017).

[15]. Okoye COB, *"Spectroscopic methods of analysis.Undergraduate Analytical Chemistry"*, Jolyn Publishers, Nsukka, pp. 98-119, (2005).

[16]. Soomro R, Jamahiddin MA, Menpou N, Khan H, *"A simple and selective spectrophotometric method for the determination of trace Gold on real Environmental"*, Biological, Geological and Soil samples using Bis, (Salicylaldehyde) Orthphenyldiamine. J. Anal. Chem. Insights 3, pp.75-90, (2009).

[17]. Zeiner, M., Rezic, I., & Steffan, I. *"Analytical methods for the determination of heavy metals in the textile industry"*. Kem. Ind, *56*(11), 587-59, (2007).

[18]. Ahmed, A., Singh, A., Padha, B., Sundramoorthy, A. K., Tomar, A., & Arya, S. *"UV–vis spectroscopic method for detection and removal of heavy metal ions in water using Ag doped ZnO nanoparticles"*. Chemosphere, *303*, 135208, (2022).

[19]. Echioda, S., Ogunieye, A. O., Salisu, S., Abdulrasheed, A. A., Chindo, I. Y., & Kolo, A. M. *"UV-Vis spectrophotometric determination of selected heavy metals (Pb, Cr, Cd and As) in environmental, water and biological samples with synthesized glutaraldehyde phenyl hydrazone as the chromogenic reagent"*. European Journal of Advanced Chemistry Research, 2(3), 1-5, (2021).

[20]. Nai-Liang H, Hong-wen G, Biao ZI, Guo-Qing Z. *"Simultaneous Determinations of Cobalt and Nickel in waste water with 2.- Hydroxyl – 5 – benzene azoformoamithiozone by spectral correction Technique"*. J. Chin. Chem. Soc. 52, pp.1145-1152, (2005).

[21]. Śliwińska, A., Smolinski, A., & Kucharski, P. *"Simultaneous analysis of heavy metal concentration in soil samples"*. Applied Sciences, *9*(21), 4705, (2019).

[22]. Ding, Y., Xia, G., Ji, H., & Xiong, X. *"Accurate quantitative determination of heavy metals in oily soil by laser induced breakdown spectroscopy (LIBS) combined with interval partial least squares (IPLS)"*. Analytical methods, 11(29), 3657-3664, (2019).

[23]. Ucun Ozel, H., Gemici, B. T., Gemici, E., Ozel, H. B., Cetin, M., & Sevik, H. *"Application of artificial neural networks to predict the heavy metal contamination in the Bartin River"*. Environmental Science and Pollution Research, 27, 42495-42512, (2020).

[24]. Alizamir, M., & Sobhanardakani, S. *"Forecasting of heavy metals concentration in groundwater resources of Asadabad plain using artificial neural network approach"*. Journal of Advances in Environmental Health Research, 4(2), 68-77, (2016).

[25]. Pyo, J., Hong, S. M., Kwon, Y. S., Kim, M. S., & Cho, K. H. *"Estimation of heavy metals using deep neural network with visible and infrared spectroscopy of soil"*. Science of the Total Environment, 741, 140162, (2020).

[26]. Liu, L., Huan, H., Zhang, L., Zhao, B., & Shao, X. *"Determination of heavy metal soil contaminants based on photoacoustic spectroscopy"*. International Journal of Thermophysics, 41, 1-10, (2020).

[27]. Emiko Ohyoshi, *"Relative stabilities of metal complexes of 4-(2-pyridylazo) resorcinol and 4-(2-thiazolylazo) resorcinol"*, Polyhedron, Vol. 5, No. 6, pp.1165-1170, (1985).

[28]. J. Ghasemi Et Al., *"Spectrophotometric studies on the protonation and nickel complexation equilibria of 4-(2-pyridylazo) resorcinol using global analysis in aqueous solution"*, J. Braz. Chem. Soc., Vol. 18, No. 2, pp.267-272, (2007).

TÓM TẮT

Xác định đồng thời các ion kim loại nặng trong mẫu nước bằng quang phổ kết hợp học máy

Trong nghiên cứu này, việc xác định đồng thời Co, Cd, Ni, Cu và Pb được thực hiện bằng cách tạo phức màu với resorcinol 4-(2-pyridylazo) trong dung dịch nước kết hợp với phương pháp học máy. Một mạng lưới nơ-ron nhân tạo và hồi quy tuyến tính đa biến bình phương nhỏ nhất một phần để phân tích hỗn hợp kim loại đã được phát triển. MATLAB là một phần mềm mạnh mẽ của chương trình học máy được sử dụng để hỗ trợ tính toán và hiển thị ma trận. Lợi ích của MATLAB trong việc xây dựng mô hình học máy cho phép phân tích nhanh chóng và hiệu quả đối với các mẫu có nhiều thành phần trong hỗn hợp mà không cần phân tách và làm giàu. Kết quả phân tích Ni2+, Pb2+ và Cd2+ bằng phương pháp phân tích kết hợp thuật toán bình phương tối thiểu một phần- PLS và mạng nơ ron thần kinh nhân tạo-ANN có độ nhạy và chính xác cao khi xác định đồng thời nồng độ các ion này trong hỗn hợp tổng hợp với hệ số hồi quy cao lần lượt là 0,993, 0,997, 0,997. Đối với Cu2+ và Co2+, độ chính xác cao hơn khi chỉ sử dụng phương pháp ANN.

Từ khoá: Mạng nơ ron nhân tạo ANN; Xác định đồng thời các kim loại nặng; Resorcinol 4-(2-pyridylazo) .