Neural network analysis of water pollution for a main river, Tamagawa, in Tokyo metropolis

Yan Yuan, Junko Kambe *, T. Aoyama, U. Nagashima **

The Faculty of Engineering, Miyazaki University, Japan
Gakuen Kibanadai-Nishi 1-1, Miyazaki 889-2192, Japan
Fax: +81-0985-58-7411; *E-mail: tgb341u@student.miyazaki-u.ac
* Faculty of Foreign Language, Daito Bunka University
1-9-1 Takashimadaira, Itabashi, Tokyo 175-8571, Japan
** Grid Technology Research Center,AIST
1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan

Abstract: We proposed a method to compensate incomplete observations and made a study of environmental problem, water quality of Tama-River in Tokyo.The method is based on interpolations of the multi-layer neural networks. We call the approach as CQSAR method .which can compensate the defect data.The water quality data include defects which will give wrong effect to other normal data. The CQSAR method suppresses the wrong effect .Thus, we believe that the proposed CQSAR method has practical usability for environment examinations.

Keywords: neural networks, inverse optimization problems, interpolations,environments, water quality, Tama River

1. INTRODUCTION

Today, the environmental problem, especially the problem of the big city which the population concentrating catch more attention. Without soluting the problem, it is difficult not only to guarantee residents' health but also to maintenance of a big city. Extraordinarily is the water-resources problem. Now we are observing the big city at the water quality of the flowing river. By investigating transformation of water quality, it shows the level of the technology which maintains residence and natural environment of a big city. The technology is so significant that it is one of the technology which can change nature.

The water quality of the Tama River, the valley lies in the capital Tokyo in Japan, was investigated in 1994-2002 and the result have been publicationed. [1,2] Tokyo (capital = prefecture) consists of about 20 cities. The Tama River valley was fully developed from the 1970s. The lower stream of the river has been flowing from the middle class region in Tama River to the city area with much people since 1980 . In the upper region, as it has a long distance from the urban areas of Tokyo, nature remained. However development of an upper region was prospered from the 1990s. It is interested to investigate the influence of this development situation water quality.

It was the long-term economical stagnation period said to be "lost ten years" during 1990 to 2004.
We think that economical stagnation had influence good for an environmental problem. We are interested in the influence which it had on the water quality of economical stagnation.

1.2 Environmental measurement and deficit

The quantity (plurality) showing water quality is measured using latest technology like ISFET (ion sensitive field effect transistor). There are about ten kinds. The water quality data released is the 3-dimensional matrix of a water quality index, a measurement point, and time. Fluctuation and an error always follow on measurement. Still known as the extreme situation that can't be determined to present. There is the case of measured value could not be issued because of the political reason. Therefore, there is a deficit of some elements in the 3-dimensional matrix.

Some numerical processing technology is required in order to analyze a super-matrix and to draw a significant conclusion.

2. DEFINITION OF THE PROBLEM

If there are defect parts in data, they give wrong effects on the analysis. Specially, in case of environmental data, there is a doubt that is; the defects might be made by artificial and political requirements. So, we have considered a method for the compensations. The problem is a kind of the inverse optimization problems. There are many approaches. Now, we believe multi-neural networks is a useful approach in points of easy-programming and the robustness.

The multi-layer neural networks are vector-vector 'group' transformers, where these elements are different each other. We define the transformer as,
tr=NN(**xr**), (1)
where "**xr**" is a set of input vectors, and "**tr**" is a set of output vectors. Using the character, we wish to compensate the defects of data. Thus, "**xr**" is a structure indexes, and "**tr**" is for the environmental data. The structure indexes relate to sampling spots in case of water quality observations. We approximated the relations to the order of observed spots along a river. We expressed the order as an arithmetic series. The approximation is useful to interpolate unpublished data by the order only.

Equation (1) shoukd have following restrictions.
Restriction 1: Never defect part in learning data,
Restriction 2: Must not fit in prohibition, xr=xs & tr !=ts.

Environmental data include defect parts; therefore, the restriction 1 limits application fields strictly. Even if the solver were not based on neural networks but statistical ones, the restriction 1 remained.

We researched solvers for the problem [3,5,7]. The solver's outlines are followings.

approach-1: Remove all data including the defects from data.

approach-2: Replace the defects parts to averages that are calculated from other data set.

approach-3: Using prediction methods, compensate the defects, and calculate neural networks as non-defect case.

approach 4: Assuming a statistical model for the data set, evaluate the definitive parameters based on a likelihood around the defects.

The approach-1 is a simple approach, but it is not so wrong. It is handy and effective in case of small defect part. We often find the method equipped in commercial statistical program products.

However, if the defect part increases more, the efficiency becomes less; at the limit, the method cannot be applied.

On simple considerations, the approach-2 is superior to approach-1. However, recent researches show that there are wrong cases by comparison with approach-1 [3].

We are also sure that there is no ground for replacement of averages.

The approach-3 is practical and has wide application fields; however, prediction methods are required. Efficiency of this method depends on ability of the prediction methods.

The approach-4 would be the most accurate approach; however, the rule equation or distribution for observations must be known in advance. On environmental observations, the requirement is a fatal restriction.

Then, we believe that approach-1 or 3 is practicable.

Our objective is, by using multi-layer neural networks, to show a kind of approach-3 for environmental data, and examine it on practical water quality data.

3. NOMENCLATURE

We assume that an index “i” corresponds to a chemical compound, and write the physical and chemical properties as;

{Xi0,Xi1,Xi2,......Xin}. (2)

Hereafter, we rewrite index “i” as digit series, (0,1,2,.....m).

We write the physiological activities of “i” compound as followings,

{T0,T1,T2,....Tn}. (3)

The activities are sorted,

T0<T1<T2<…..<Tn. (4)

In this paper, we consider a kind of activities. The elements of {Xij} and {Tj} correspond each other. Where, the generality is kept.

Next, we introduce a matrix {Mij} whose elements are 0 or 1. Similarly, a vector {Nj} is defined.

The {Mij} and {Nj} correspond to {Xij} and {Tj}, respectively.

When Mij=0 or Nj=0, the corresponding data Xij or Tj are lost.

4. TRADITIONAL NEURAL NETWORK APPROACHS

Using neural networks, to solve inverse optimization problems, some trials have been published [5,7].

When we solve the problems by using neural networks, a difficult task is that there is no means to take account of “non datum calculation”. By considering the means simply, we may get that the learning process is to be locked as for the connections related to non-datum part. However, as tracing of the back-propagation algorithm, the idea is equivalent to a calculation in case of “defect-data=0.” On the back-propagation algorithm, any means is not defined for non data.

If we leave out of the learning algorithm and adopt the reconstruction learning, the non data is replaced by uniform random numbers in interval [0,1]. This would be only one way to take account of non-data. However, the replaced input data are random numbers. That is, we cannot expect the convergence of the reconstruction learning. In fact, the learning is never converged and the back-propagation error are swaying for ever. But, the learning of other data except the non-data is converged. When the learning becomes such situations, the endless learning is forced to stop.

Next, uniform random numbers are inputted in the network. The differences between the network output and teaching data are compared each other. Then, the minimum difference is selected and the input random number is a solution of the inverse problem. We call these scheme “swaying reconstruction-learning method.”

The method has some problems, which are too many CPU-times and to require teaching-data. Therefore, restricted and few type of inverse problems is solved.

5. INTER/EXTRAPOLATION THEORY

We discuss a new inter/extrapolation theory in this section.

We consider a case, Mik=0 and 0<=k<=n. Where, “k” represents plural cases.

Then, observations, **Xi**={Xi0,Xi2,.....,Xin}, include plural defects, whose locations are pointed by Mik=0. Now, we introduce a vector, **Y**={Y0,Y1,.....,Yn}. The element number is equal to **Xi**.

Except the case Mik=0, using Y and Xi as input and teaching data, we can make a multi-layer neural-network learn them.

After the learning completed, a relation, Xij=NN(Yj), j !=k, is organized in the network.

We can get Y-vector by sampling of an elementary function. So, a Yk can be evaluated by the function value corresponding to Mik=0. If there were plural Yks, the evaluation would be done.

As the elementary function, we adopt a linear function. So, Y is arithmetic progression.

In the function, prediction of Yk is easy.

If we substitute the Yk into the neural network, we get Xik=NN(Yk). Thus, we can know the defect datum Xik. Using the same processing, we can get Tk=NN(Yk).

Where, we can consider an advanced method. That is, the Y-vector is replaced by the compensation vector {Xij}. The method is iterative scheme. This is a kind of approach-4 in section 2. However, the neural network is a non-linear function converter; therefore, the revising points would be small.

As stated above, any defect in input and teaching data is compensated, and the compensation doesn’t depend on the number of defect data. The estimated data are complete; therefore, we can analyze them by using another neural network. This is normal neural network calculations.

Hereafter, we call the method CQSAR (compensated quantitaive structure activity relationships).

On the CQSAR, all observed data are used, where a lack of observations doesn't affect whole processing, and the wrong effect is suppressed. We wish to evaluate CQSAR numerically.

6.NUMERICAL EXAMINATIONS FOR ENVIRONMENTAL DATA

6.1 Generation of model data

We examine CQSAR for model functions. Because, they don't include noise, and it is convenience to check the true character. We selected 5 kinds of elementary functions; i.e., $x^{0.5}$, x, x^2, $(x-0.25)^2$, (x-0.1)(x-0.5)(x-0.9). These functions simulate physical and chemical characters of substituents of various compounds. The definition interval is [0,1]. Sampling them at 37 points regular intervals, and the sampled vectors are scaled within [0,1]. The 5 vectors were used as input data for neural networks. Next, we sample one function, $-(x-0.7)^2$. The function emulates physiological characters of the compound. We believe that such data should have a maximum point. Using same process, we got a teaching vector. We plotted them in figure 1.

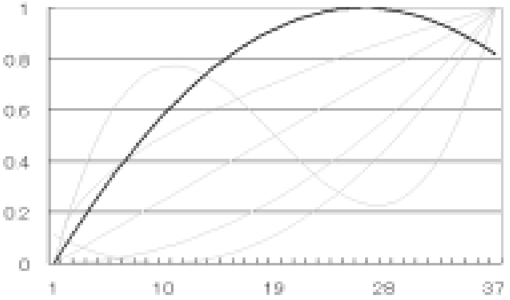

Fig. 1. Plotting of QSAR model data.
Horizontal axis indicates compounds' numbers.
Vertical axis is function values scaled in interval [0,1].
A bold curve is plotting of teaching data, and pale curves are 5-kinds input data.
The figure shows that there is a simple relation in learning data.

Those 6 kinds of vectors are all complete. We introduce following M-matix and N-vector, and under them, we process the 6 vectors as incomplete ones.
We considered 5 M-matrices and 5 N-vector as followings.

Case 1: Mii=0, 0<=i<=4, N5=0.
Case 2: Mii=0, 18<=i<=22, N23=0.
Case 3: Mii=0, 32<=i<=36, N37=0.
Case 4: Mii=0, 0<=i<=4, 6<=i<=10,..... N5=N11=....=0.
Case 5: Mij=0, i=even, j={0,2,4}, i=odd, j={1,3}, N(odd)=0.

An object of the selection is;
Case 1: Defects are found continuously in the first part of observation.
Case 2: They are found in central part.
Case 3: They are found in terminal part.
Case 4: They distribute over whole data.
Case 5: They distribute as a checkered pattern.

Since the defect ratio is small in cases 1-3 (~6/37=16.2%), by using approach-1, they can be processed. However, in case 4, the ratio is 100%, and moreover in case 5, 50% of whole data is lost.
(In case 4, 1/6=16.7% of whole data is lost.)
On such cases, the approach-1 cannot be applied.
When such a large percentage data are lost, normally usual methods would not process them.
On other hand, CQSAR can be done for all cases; and it can take account of the unused data in column including defect datum.

6.2 Defination of neural networks

The defect parts are predicted by using 3-layer neural networks whose structure is;

(1) Input, hidden, output-layer's neurons are 2, 4, 1, respectively.
(2) The neuron's emulation functions are a sigmoid-function for hidden and output-layers.

The non-linear function fitting ability of neural networks is caused by the sigmoid-function and the number. Therefore, it should be limited in a small number. The 4-sigmoid functions can simulate any function having two peaks. The fitting ability is not so much. We would prevent the neural networks from excessive learning. The design is important at first predictions for defect data.

(3) Initial guess of connection weights among neurons is uniform random numbers in [-0.5,0.5].
(4) Bak propagation learning, learning coefficients are 0.2 and 0.15 for hidden and output-layers.
(5) The learning iterations are 20K.

If the back propagation error is not converged to 0, the learning is stoped by force.
It is also a treatment to stop excessive learning. In case of model data, the error were under O(-5).

(6) Whole calculations are done by using 64-bits IEEE floating point format.
(7) The compiler is digital Fortran version 6 (producted by DEC).

After the predictions, normal QSAR was calculated by using secondary 3-layer neural network.
The network structure is below.

[1] Input, hidden, output-layer's neurons are 6, 8, 1, respectively.

The second network has many kinds of input data; therefore, we used large number of hidden-layer's neurons.

[2] Other conditions are same as the first networks.

6.3 Examination of CQSAR method

We examined the effects of CQSAR based on standard deviations.

Table 1. Standard deviations for defect parts

Table 1 shows prediction ability of two methods.
The ability is represented by the standard deviations; therefore, "0" signifies complete predictions.
"0.1" means that error of 10% order is found.

	CQSAR	approach-1
Case-1	0.0152	0.1301
Case-2	0.0046	0.0063
Case-3	0.0343	0.1107
Case-4	0.0132	impossible
Case-5	0.0195	impossible

The base data are sampled from model functions to simulate QSAR, which don't include noises (ideal data). Then, we can know the ideal standard deviation when complete data are given. It was 0.0062, which signifies calculation-precision of second network. Even if a lower value might be got, it was calculated by an accident. We believe the value under 0.0062 is non-significant.

CQSAR method is developed in this paper, and approach-1 is traditional, which is column or row including a defect is left out of the calculations. On whole cases, CQSAR is superior to traditional.

On cases-3,4,5, CQSAR's predictions are 8.61, 1.37, 3.23 times accurate. Especially, on the cases-4 and 5, CQSAR gave reasonable standard deviations. These results are to be noticed.

Cases-1 and -3 show the network's precisions near two terminal points. They are less precisions compared with that of the case-2, which indicates the defects in central parts. It is well known that extrapolation ability of neural networks is less than interpolation. The simulations indicated the same results.

Comparing CQSAR and approach-1, CQSAR gives high precision standard deviations at both terminals of observations. This shows that CQSAR method has extrapolation, which is a useful character for QSAR.

We checked the compensation ability of CQSAR in case-4 and 5. To check it, we plotted the differences between the outputs and real values in figure 2.

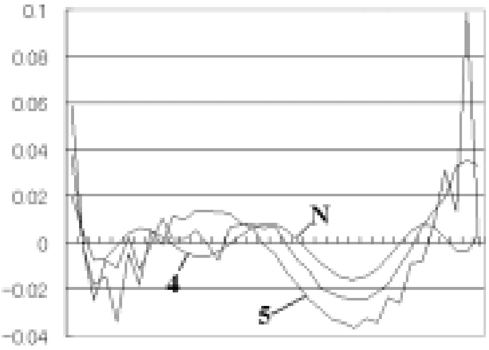

Fig. 2 Differences between outputs and teaching data
Horizontal axis indicates compounds' numbers.
Vertical axis is amplitude of the differences.
Curve N is the case of non-defect data, which shows fitting-ability of the second neural network.
Curve 4 is case-4, and "5" is case-5. They show fitting-abilities of this CQSAR method.

Thus, we believe that CQSAR is effective for a model QSAR calculation.

We evaluated other model calculations that include many kinds of bend-functions and step-functions, which are not differential and continuous. Even if there are such un-fit points, where the influences are limited at local regions, CQSAR gave reasonable results.

7. NUMERICAL CALUATION FOR WATER QUALITY DATA FOR TAMA RIVER

7.1 Characters of water qualities

The effects of CQSAR should be examined for water quality data of Tama River, which are published in references [5]. Observation spots are distributed uniformly along the riverside and the number is 17. Where, we numbered the spots as 1,2,....17 from the upper to lower regions.

The qualities are observed as pH, DO (), BOD (biological), COD (), T-N (total nytrogen), T-P (total phosphrus), Cl- (cloride ion), NH4N (ammonium ion), NO2-N (sub-?nitrate ion), NO3-N (nitrate ions), PO4-P (phosphrus ions), COND (electric conductivity).

Ratios of the defect parts are in table 2. The parts give wrong effects to other data. The effects are calculated as analysis means for horizontal/vertical directions, and they are listed in table 2.

Where the horizontal direction analysis is as for the observation spots, and the vertical is for the water qualities' indexes. If you want to analyze environment for spots and water qualities, the ratio of defects is the maximum value of both ratios. On reference of the table, a small partial defects make the analysis be inenable for 2001/2002.

Table 2 Ratios of defect parts and scopes of two methods

years	defects	horizontal/vertical	approach-1	CQSAR
1994	0.000	0.000/0.000	practicable	practi cable
1995	0.000	0.000/0.000		
1996	0.000	0.000/0.000		
1997	0.000	0.000/0.000		
1998	0.000	0.000/0.000		
1999	0.000	0.000/0.000		
2000	0.0245	0.2941/0.0833	impossible	
2001	0.1471	0.7059/1.000		
2002	0.2010	0.7059/1.000		

Using procedures of the sections 3-5 (i.e. CQSAR method), we can compensate the defects and make the ratio be zero. Therefore, the CQAR method makes the scope of environmental analyses expand.

7.2 Neural network calculations

The defect parts are predicted by using 3-layer neural networks whose structures are same in section 6. The back-propagation error converged to O(-1.5) for compensations of input data defects, and O(-3) for teaching data. The values are not sufficient. But, considering the characters of learning data, they should be accepted. If we used more neurons on a hidden-layer, the error might be less.

However, the direction doesn't equal to increase the prediction ability of the neural network. By using many parameters on neural networks, it is impossible to get accurate predictions, as well as statistical methods. Here, it is true that less information gives uncertain predictions. In this section, the objective is to research influences of the uncertain compensations to water qualities of Tama River. The results are super-matrix whose element number is 9*17*12=1836; so, we don't list it here.

The structure of second neural network is followings;

(1) Input, hidden, output-layer's neurons are 13, 6, 1, respectively.
(2) Other structures are same as section 6.

When the structures were used, the back-propagation-error of 20K iterations were 0.0047 and 0.0046 for CQSAR and approach-1.

They are O(-2.5) and are same level to that of model data in section 6.

Some one may feel it is not sufficient. We believe that forced convergence means excess-learning.

So, we stoped the learning on this level. To check content back-propagation-error for full set of learning data, we calculated the non-defect case on 100K iterations. The error value was 0.005; that is O(-2.5). So, we believe that the CQSAR compensates water qualities data of Tama River.

7.3 Signigicance of water quality indexes

We evaluate environmental protections on use of many indexes. On such evaluations, what is significance index? We often use a word, clean environments; in that case, what is the judgement standard?

It is believed that the water quality of a river is clean at source, and it may be not at the sink.

Rivers through metropolises are polluted generally; however, the rivers would be clean at the source, and the causes of the pollutions are not distributed uniformly but be localized.

The localization gives stepwise changes for the water quality indexes.

Under the assumptions, we made reconstruction learning (BP-learning with forgetfulness) of a neural network, and selected significance indexes, wich were DO, T-P, and PO4-P. The ratio of PO4-P was very small. The DO values are listed in figure 3.

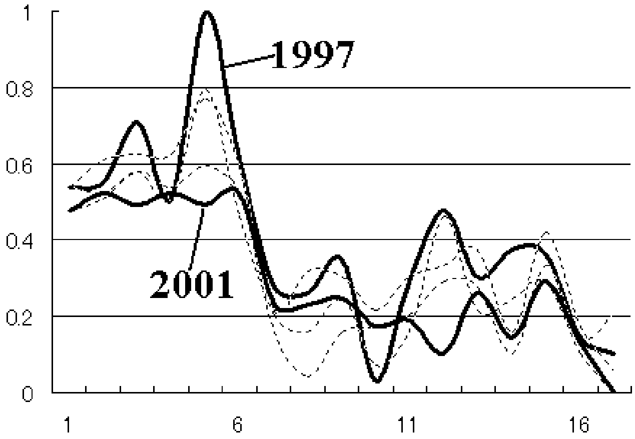

Fig. 3 The changes of DO-index at 17 spots along Tama River between 1997 and 2001.

The vertical axis is normalized, the maximum is 1.0 and the minimum is 0.0.

The number of horizontal axis means the spots.

The bold curves are the values of 1997 and 2001. The dotted curves are that of 1998-2000.

The high values for the DO-index indicate more clean environments.

So, the graph shows that the water quality becomes wrong gradually.

Especially, the quality becomes wrong rapidly at the upper region.

However, it becomes slightly clean at the middle and lower regions.

The T-P values are listed in figure 4.

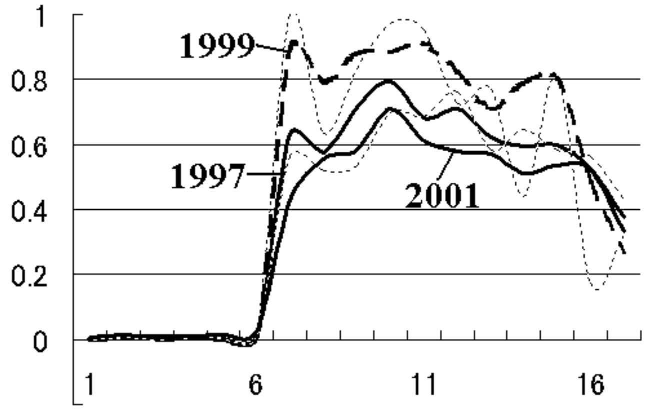

Fig. 4 The changes of T-P-index at 17 spots along Tama River between 1997 and 2001.

The vertical axis is normalized, the maximum is 1.0 and the minimum is 0.0.

The number of horizontal axis means the spots.

The low values for the T-P-index indicate more clean environments.

So, the graph shows that the water quality in middle region was wrong at 1999, and it returns to that of 1997's level now.

The T-P index gives no information for the upper region.

7.4 Discussion of Asahi press report

Asahi press makes a comment on Tama River; that is, the water quality is becoming clean now. Hamura Dam (a control dam in upper region) has been discharging from 1993s. This is a method to purify the water of rivers by using dilution and powers of microorganisms.

Asahi is a newspaper for the general public, where the meanings of the clean are not discussed in many water quality indexes. As we calculated the water quality of Tama River by CQSAR method; then, we wish to compare them with the development of Tama River valley and Japan's economic stagnation in 1990s. The development gives wrong effects, and the stagnation and discharging give the reverse. The both effects are continue in all time of 1990s. The three actions are not examined qualitatively. We believe our calculations may be a indicater at least.

The conclusions are followings:

1. In upper region, the development is main influence for environments. The water quality is becoming to wrong certainly, which is shown by DO-index.

2. In middle and lower regions, the influence of developments has not been so large. Under the economic stagnation, activities of existing factories have been suppressed. The effect would be operated. So, the discharging of a dam gives an influence for environments correspondingly. This is shown by time-changing of T-P index.

8. CONCLUSIONS

We discussed a neural network approach for incomplete data set. We proposed a new solver that used 2 kinds of multi-layer neural networks. One is to compensate the defect data, and another is to analyze the compensated data set. We call the approach as CQSAR method, which has following facilities.

1) It can solve problems that a traditional method cannot process.

2) It has prediction-ability to compensate 50% defects of observations.

3) It revises standard deviations of neural networks about 2-5

times.
The solver can completely predict the defects in model data. By using them, we get very high precision neural network calculations. It gives 5-10 times accurate standard deviations in comparison with a traditional method. We tried to other 5 models, and got same effective results.

Next, we adopted the CQSAR to an environmental problem, the water quality of Tama River. The river flows to Tokyo Bay through many big cities in Tokyo prefecture. Under the economic developments, to maintain the river and to keep the environments, it is a difficult problem, and it requires high technologies. We wished to check our abilities for the environments, and as a practical examinations, we analyzed the water quality indexes.

The water quality data had defects whose ratio is 2.5-20%, which is not so large. However, the defects give wrong effect to other data, as the effect, the 29-70% of data is not included in neural network calculations. The CQSAR method suppresses the wrong effect.
Using the method, we got the following results:

1. In upper region of Tama River, the development is main influence for environments. The water quality is becoming to wrong certainly, which is shown by DO-index.
2. In middle and lower regions, the influence of developments has not been so large. Under the economic stagnation, activities of existing factories have been suppressed. The effect would be operated. So, the discharging of a dam gives an influence for environments correspondingly. This is shown by time-changing of T-P index.

Those conclusions are roughly same to that of Asahi press report.
Thus, we believe that the proposed CQSAR method has practical usability for environment examinations.

We have published the Fortran program-codes and their data on Internet; the URL is,
http://www.miyazaki-u.ac.jp/aoyama_t/index.html.

REFERENCES

[1] Bureau of Environment, Tokyo Metropolitan Government,
http://www.kankyo.metro.tokyo.jp/

[2] Social Survey Research Information Co., Ltd.,
http://www.ssri.com/index.ja.html

[3] M. Watanabe, K. Yamaguchi,
"EM Algorithm and the problems for incomplete data set (in Japanese)",
Taga Publishing Co. Ltd. (2000, Tokyo), ISBN4-8115-5701-8.

[4] J. Kambe, T.Fukuda, U.Nagashima, T.Aoyama, "Extraction of chemical parameter charcterizing the upper stream, middle, stream, and lower stream by principal component analysis and neural network -The case of Tamagawa river, Tokyo-", J.Chem Softwere, Vol.8, No.1, pp27-36, (2002.3).

[5] K. Bitou, Y.Yuan, T. Aoyama, U. Nagashima, "A new learning algorithm for incomplete data sets and multi-layer neural networks", ICCAS'03, CD-ROM(TA06-04), (2003.10).

[6] T.Aoyama, Y.Suzuki, H.Ichikawa,
Neural networks applied to quantitative structure activity relationship,
J. Med. Chem., Vol.33, pp.2583-2590, 1990.9.

[7] Qianyi WANG, Tomoo AOYAMA, Umpei NAGASHIMA, Eui-Sung KANG,
Inverse optimization problem solver on use of multi-layer neural networks,
Proc. of International Conference on Control Automation and Systems'01,
CD-ROM (949.pdf), 2001.10.17-21.

[8] T.Aoyama, H.Ichikawa, "Reconstruction of weight matrices in neural network, a method correlating output with input", Chem. Pharm. Bull. Vol.39, pp.1222-1228. (1991.9).