

離散形河川水質モデルと不完全データ問題

青山 智夫^{a*}, 神部 順子^b, 長嶋 雲兵^c

^a 宮崎大学工学部, 〒 889-2192 宮崎市学園木花台西 1-1

^b 大東文化大学外国語学部, 〒 175-8571 板橋区高島平 1-9-1

^c 産業技術総合研究所計算科学研究部門, 〒 305-8568 つくば市梅園 1-1-1

*e-mail: t0b217u@cc.miyazaki-u.ac.jp

(Received: November 21, 2005; Accepted for publication: February 6, 2006; Published on Web: June 5, 2006)

都市を流域とする河川水質シミュレーションのための離散形式の四モデル, 単純離散, 堰, 地下浸透, 双未定係数モデルを提示した. それらモデルで記述された人工河川の水質を神経回路網により解析した. 目的は水中物質質量変化を離散表示する妥当性を, モデルの支配方程式逆算可能性の観点から検討することである. 逆算可能ならば河川作用をデータのみから推定できる.

神経回路網は非線形現象解析に有用である. 解析には離散形式のデータ集合が必要である. データ集合は観測値と現象要因の離散形データを含む. 神経回路網はデータを学習することにより現象記述機能を獲得する. 獲得には観測, 要因データに不足があってはならない. 不足とはデータ集合要素の存在自明, 値不明の場合と, 測定要因データの一部種の全欠損, 要因存在不確実の三種類がある. 最初の場合を欠測という. 欠測を含むデータ処理は近年研究されているが後二者は検討例が少ない. 我々は環境問題解析に後者の検討は必要と考える. 同時にそれは神経回路網の出力値検定に関わる重要問題である.

本稿では神経回路網の非線形多変数解析機能を基にデータ不足状況における挙動を調査する. 主目的は同回路網の機能限界を明確にすることである. そのためデータ誤差の考察を除外した. 一般に観測誤差の統計的性質は確かでない. 不確実データを基に適用限界の議論は困難である. そのためにモデルを考案し人工河川を定義した. 河川データは一様乱数を基に作成した. そしてあえて解析に必要な現象要因データを省き, データ不足状況を作った.

神経回路網解析から以下の事実が判明した. 1. 水中物質質量変化を離散表示することは妥当である. データが完備していれば離散表示から精度良く現象要因を計算できる. 2. データ不足状態でも神経回路網は精度良く河川水中の物質質量をシミュレーションする. 一方, 同回路網の出力値の偏微分係数は正確な要因値を示さない. 3. 原因は欠損要因を他が補完するためである. 4. データ要素の偏微分値変化の大きさがその補完要因種を示す. 5. 偏微分値変化から欠損要因の性質が推定できる. 6. それが可能であるのはモデル支配方程式の概要が推測できる場合である. これら事実は水中物質質量変化が離散表示可能なこと, モデル支配方程式がデータのみから逆算できることを示し, 総じて河川作用を観測データのみから推定できる可能性を示唆する.

キーワード: 河川モデル, 浄化作用, 神経回路網, 偏微分係数, 欠測

1 はじめに

昔, 川は上流から下流にかけて連続的に変化していく水の流れであったが, 人口増加により用水を必要と

し取水が行われるようになり, 近年では下水道の整備により処理水が河川に流入するようになった. その量は支流量を超えることもある [1]. 下水処理水の水質

は炭素に関する生物学的酸素要求量 (C-BOD, Carbon-Biochemical Oxygen Demand) については次第に改善され、近年では支流とほぼ同じレベルである。しかし化学的酸素要求量 (COD, Chemical Oxygen Demand), 全窒素 (T-N, Total-Nitrogen), 全リン (T-P, Total-Phosphorus), N-BOD (Nitrogen-Biochemical Oxygen Demand) については 1 オーダ大きいこともある [2]。このような環境に対する小さくない負荷が断続的に生じているのが現代の都市圏を流れる河川である。

河川の水質問題についてさまざまな研究が成されてきた。それらの研究の多くは膨大な観測を必要としたり、現象記述方程式 (支配方程式) の知識を必要とする [3, 4]。環境に対する大負荷であっても詳細に測定でき、支配方程式を決定できれば詳細な検討が可能である。しかし一般的に公開されている数 km 間隔の観測点間距離の河川データを用いて河川水質を論じるには既存の方法では困難である。都市部を流域とする河川では環境負荷が大きく水質状況を定量的に把握するには従来とは別の、観測数をそれ程要求せず、明示的な支配方程式を要求しない、新しい処理が求められている。我々は点と点の間の水質指標の離散的变化として河川水質を定式化し、河川水質の断続的变化を的確に表したい。

現象を離散的に定式化する方法の一つに階層型神経回路網 (以下神経回路網という) がある。神経回路網が河川水量に適用された例 [5] はある。水質変化について適用された例はないように思われる。

神経回路網は非線形関数適合 (fitting という)、補完 (補間と補外の総称)、分類機能を有し、目的現象の支配方程式が不明であっても、説明変数と現象の観測値から同方程式の模擬関数 (emulation function という) を自動的に生成する。この技術は薬理学では確立されている。自動生成された関数は離散データが基になっていても、連続かつ微分可能な関数である。その性質を利用して予測が可能である。これらの機能は河川水質問題を研究する上で有用ではないかと考えられる。ただし、その諸機能が発現するためには、説明変数と観測値に不足がなくてはならない。

不足とは観測データが本来あるべきことが自明であるのに何らかの理由で不明な場合と、現象を説明する観測自体がされていないか、そういう観測自体の存在不確定の三種類がある。前者を欠測という。欠測を含んだデータの情報処理は近年かなり研究されている [6]。一方、後二者は前者以上に重要であると思われるが、ほとんど検討されていない。当然のことで、少な

い観測から非線形現象を再現すること自体容易でないのに、さらにその再現の背景に不足データの存在を指摘するのは極めて難しい。しかし、我々は環境問題の真の理解のために後者は回避できない問題と考える。同時にそれは神経回路網出力の検定に関わる問題である。我々は神経回路網の検定を出力値の信頼性という観点から調査したが、そこには後者の概念が検討されていない。本論文の目的はこの観測データ不足問題を考察することである。

ほとんど全ての環境データに欠測が見出されるが、水質問題に関する諸数値はデータ不足の状況である。それにも関わらず、環境保護団体や行政の議論にそれを考慮した例は少ない。定量性に疑問ある議論を一步前進させるため、不足データを基に現象を解析する方法を検討したい。本論文では神経回路網を使用した非線形多変量解析のデータ不足状況における挙動を調査し、同回路網解析の適用限界を明らかにすることである。

観測データには測定誤差が含まれる。その統計的性質は確かでない。それを基に神経回路網の適用限界の議論は困難である。この測定誤差問題を回避するため河川モデルを考案し、それと乱数を基に人工河川を定義し、そのデータを基に神経回路網の適用限界を明らかにする方針である。

2 河川水質の離散形式

2.1 定義

我々は都市河川の中下流域の河川浄化機能に焦点を絞り下記のモデルを考える。

【1】河川の性質は離散的な観測点列の連鎖で表せる。観測点間の水質変化を一指標の変化 (ベクトル形式) として表す。

【2】観測点への流入は本流と「支流」の二要素とする。ここで支流とは本流への全流入を加算したものである。

我々は人工河川を取り扱うので、観測点間距離、支流の流入量は区間指定した一様乱数で作成する。支流水に含まれる物質も同乱数で作成する。水面からの蒸散、地下浸透は単純化のため、ここでは考慮しない。河川に流入する物質をこのように単純化すると、本流流量と本流水に含まれる物質が計算できる。

【3】河川は浄化機能を有する。その浄化機能を観測点間距離に依存する。この機能は一次関数で表現できるとする。実河川は上流と下流で河川構造が異なり、水

中の物質も違ふと思われる。そこで我々は多摩川の1/4000縮尺の航空写真を羽村堰から河口まで調べた。その結果、日野橋から下流では河川構造の相異は堰の存在以外の顕著な相異は見られなかった。そのような状況ならば河川浄化機能を観測点間距離に比例するとしても良いように思われた。

以上【1】～【3】の仮定を用いて人工河川水質の変化を観測点で離散的に表現した。それは最も簡単なモデルである。以下、単純離散モデルという。さらに複雑なモデルは第3、4節で研究する。以下にアルゴリズムの詳細を示す。

- (1) モデルの各量は無次元である。
- (2) 河川は $N-1$ の観測点 $\{2,3,4,\dots,N\}=\{i\}$ で水質値が定義される。観測点1を本流の始点とする。始点を加え $\{1,2,3,4,\dots,N\}=\{Xi\}$ で河川の各点を番号づける。1側を上流側とする。
- (3) 観測点間距離を $[0.5, 2]$ 区間の一様乱数(以下、乱数)とする。 $\{di\}$ と表記する。 di は $Xi-1$ と Xi の距離である。本区間は多摩川の支流の流量 [7] から決定した。
- (4) 支流は $\{2,3,4,\dots,N\}=\{Bi\}$, $1 < i$ と番号づけされる。支流は川だけでなく下水道などを含む本流に流入する水の総称である。 Bi は Xi に流入する。
- (5) 支流の流量 (= 本流流入量) を $[0, 1]$ 区間の乱数とする。これを $\{fi\}$ と表現する。 $1 < i$ である。本流の流量は

$$g_i = g_{i-1} + f_i, g_1 = 1, \quad (1)$$

である。

- (6) 支流 Bi の水質を $[1, 1+ai]$ 区間の乱数とする。 $\{ai\}=\{0, 1/N, 2/N, \dots, 1\}$ である。これを $\{pi\}$ と表示する。 $1 < i$ である。このようにすると支流の水質は下流側で悪化することになる。それは多くの河川で見出される状況である。シミュレーションで支流効果を明らかにするため上流側の支流水質の最低値を1とした。
- (7) 本流 Xi 地点の水に含まれる物質量は、

$$s_i = s_{i-1} + f_i p_i + k_i d_i, \quad (2)$$

である。 $1 < i$ である。 $s_1=0$ とする。ここで k_i は観測点 $Xi-1$ と Xi の間の河川水浄化機能である。ここで仮説【3】「浄化機能は水に含まれる汚染物質質量によらない。距離に依存する」が採用されている。 k_i は普通負数である。またさらに近似して $k_1=k_2=\dots=K$ (定数) とすることもできる。観測データ数が少ない(約20

以下)では定数が妥当である。

$$s_i = s_{i-1} + f_i p_i + K d_i, \quad (3)$$

制約条件は $0 < s_i$ である。

〈8〉 本流の水質基準値 $\{qi\}$ は

$$q_i = s_i / g_i, 1 < i, \quad (4)$$

である。

このアルゴリズムの特徴について議論する。式(3)はベクトル $\{s_i\}, \{f_i p_i\}, \{d_i\}$, $2 \leq i$ が与えられれば K について解くことができる。すなわち本流の水質指標値、支流水量、支流水質指標値、観測点間距離の観測があれば、河川浄化機能がまだ働いているか否かを検討できる。最低観測点数は式(3)が与えられていれば2である。しかし式(3)は与えられていないので観測数は多いほど良い。神経回路網は不明の(現象)支配方程式、

$$s_i = F(s_{i-1}, f_i p_i, K d_i), \quad (5)$$

をデータから生成し、その偏微分、

$$\partial F(s_{i-1}, f_i p_i, K d_i) / \partial d_i = K, \quad (6)$$

から河川浄化係数 K を求めるのが本アルゴリズムの目的である。

「データから生成」できる可能性は以下のように考えられる。神経回路網に入力する一つの説明変数を x_i , W_{xij} , W_{yjk} を第1,2層間のニューロン間の結合荷重, 第2,3層間の結合荷重, 添字 i,j,k を第1,2,3層のニューロンを示すとし, 第1,2層の最後のニューロンそれぞれ1個を bias neuron, 値を1とする。添字 ijk は一対一対応しているとする。第2層のニューロンに入力する値を h_j とすると、

$$h_j = \sum_i W_{xij} x_i, \quad (7)$$

第2層の出力 q_j は、第2層のニューロン・エミュレーション関数、シグモイド関数 ff として、

$$q_j = ff(h_j) = 1 / \{1 + \exp(-h_j)\}, 0 \leq q_j \leq 1, \quad (8)$$

第3層からの出力 y_k は、

$$y_k = \sum_j W_{yjk} q_j, \quad (9)$$

となるから、式(7)で式(3)の項間の+演算子が生成され、 ff 関数部ではバイアスにより $average(h_j)$ 0と

なれば ff の直線部を使い式 (3) の線形結合が再現される可能性がある。式 (9) はその直線部の拡大/縮小作用なので再現を助ける。以上一変数で説明したが、多変数でも全く式 (7) の作用は同じで+,+,... という結合が生成される。

$\partial F(si-1, fipi, Kdi)/\partial di$ が定数でなく関数 $G(di)=\{Gi\}$ となる場合はアルゴリズムの一貫性 (self consistency) が満足されていないことになる。定数でない度合を距離 D ,

$$D = \sum_i \{G_i - \text{average}(G_i)\}^2 / N, \quad (10)$$

とする。 N はベクトル $\{Gi\}$ の要素数である。

2.2 神経回路網のための単純離散モデルデータ

第 2.1 節のアルゴリズムで作成した神経回路網用単純離散モデルデータの一例 ($K=-0.2$ の場合) を Table 1 に示す。

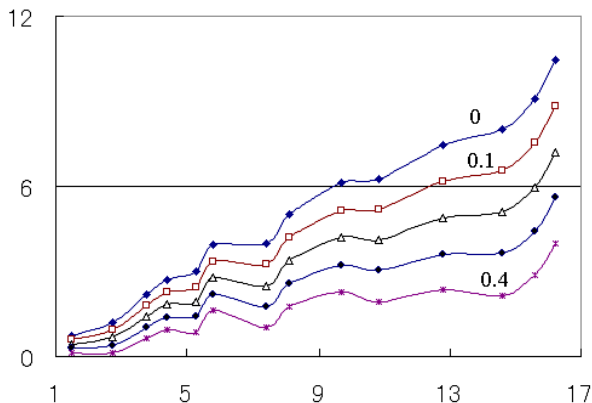


Figure 1. Substance amounts in the water of a virtual river. The vertical axis is the substance amounts that have a dimension of [mass]. The [mass] is a generic-dimension of [kg] etc. The horizontal axis is sampling points, which are numbered. The points are located by the accumulated distances. The left side is the upper stream. Whole data are generated by uniform random numbers. In the figure, digits are the river purification coefficient, K , which are multiplied by (-).

本流水中の $K=0,-0.1,-0.2,-0.3,-0.4$ の場合の各観測点の物質質量 (教師データ) を Figure 1 に示す。河川浄化係数 $-K(0-0.4)$ による変化を Figure 1 中に示した。各河川で観測された水質指標 (たとえば多摩川の BOD データなど) と類似している。

単位はデータを乱数で作成しているため具体的に [kg] のように書けないので、質量の意味があることを示すために [mass] とした。以下体積を [volume] などとする。

支流から流入する物質質量を Figure 2 に示す。

我々は Figure 2 において、様々な水が流入するように、かつ上流ほど清浄な水が流入するようにした。Table 1 のデータからデータを作成したモデル式 (2) を神経回路網に与えないで河川浄化係数 K を推測するのが目的である。一種の逆問題である。

Table 1. Neural network learning data for simple discrete river model. The d1,2,3 are descriptors. The d1 is $fipi$. The term $fipi$ is substance amounts from branches into the main stream. Where the branches are a generic name of branch rivers' water and sewage and rainwater. The d2 is distances between observation points, i.e., di . The d3 is substance amounts at Xi point in main stream, i.e., $si-1$. The T is teaching data, si .

$K=-0.2$	d1	d2	d3	T
1	0.7399	1.4803	0.0000	0.4438
2	0.4795	1.2653	0.4438	0.6703
3	0.9468	1.0436	0.6703	1.4083
4	0.5422	0.6244	1.4083	1.8256
5	0.2720	0.8682	1.8256	1.9240
6	0.9507	0.5127	1.9240	2.7721
7	0.0566	1.6193	2.7721	2.5048
8	1.0308	0.7241	2.5048	3.3908
9	1.1093	1.5378	3.3908	4.1926
10	0.1438	1.1489	4.1926	4.1066
11	1.1838	1.9709	4.1066	4.8962
12	0.5434	1.7954	4.8962	5.0806
13	1.0854	0.9882	5.0806	5.9684
14	1.3816	0.6495	5.9684	7.2201

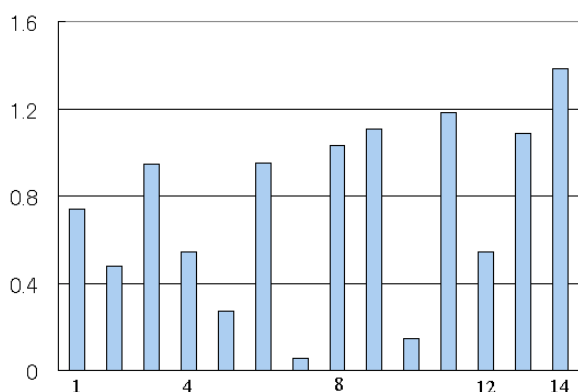


Figure 2. Inflow substance amounts from the branch in the virtual river system. The vertical axis is the amount whose dimension is [mass]. The branch is one at an 'observation' point in the main stream. The branch is a generic name of all inflow. Therefore, we avoid no branch stream. The horizontal axis is the number of the branch, $1 \leq B_i \leq 14$. We set water quality of the branch is to be clean at upper stream; and we set many kinds of branches flow into the main stream.

Table 2. Learning parameters for vector data in Table 1. The number of neurons on the second layer is only 3 that are decreased from initial value 8. One of the three neurons is a bias. The bias neuron is on the first layer, too.

# of data(descriptor., teach.)	(14×3, 14)
Neurons on 1,2,3-layers	4, 3, 1
Emulation on 2nd-layer	sigmoid
Emulation on 3rd-layer	linear
# of learning	3K
Learning Constant	0.2, 0.2
Reconstruction	60 times
Erasing factor	0.04
Square of BP error	0.025

2.3 河川浄化係数 K の計算

第 2.1 節のアルゴリズムで河川水質のデータを生成し、観測点 X_i への支流からの流入物質質量 f_{ipi} , X_{i-1} 観測点との距離 d_i , 本流 X_{i-1} 点の物質質量 s_{i-1} を説明変数(descriptor)とした。本流 X_i 点の物質質量 s_i を教師データとした。神経回路網にとって説明変数は入力データであるのでそのように記す。Table 2 に 3 説明変数, 1

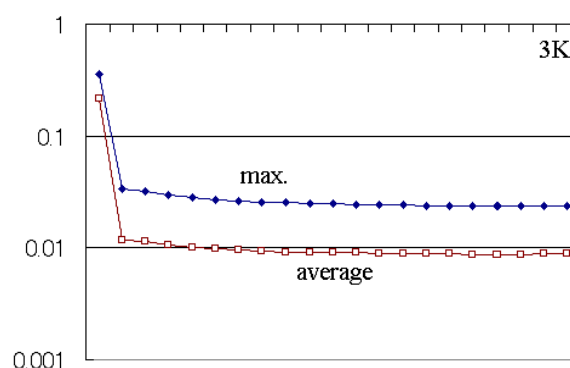


Figure 3. Maximum and average BP error in learning iterations. The vertical axis is the logarithm of the error. The horizontal axis is the learning iterations' number. The learning data are generated by the coefficient, $K=-0.2$.

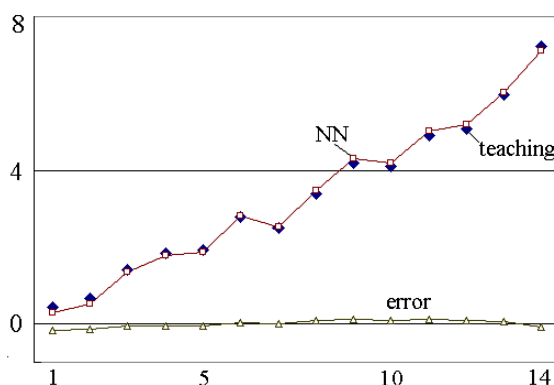


Figure 4. Neural network outputs and the teaching data after learning iterations of 3K. The vertical axis is substance amounts that are calculated by $K=-0.2$. The dimension is [mass]. The horizontal axis is the observation points, X_i . The left side is the upper stream. The black points are the teaching data, and the bend lines are the outputs. The error line is the difference.

教師データ, 各データ数 14 の場合の神経回路網学習パラメータを示す。学習方法は back propagation + reconstruction learning を用いた。本計算の目的は、学習が停留状態になった神経回路網から河川浄化係数 K を求める具体的方法を示すことである。

Figure 3 に学習時の教師データとの差の変化を示す。Figure 3 において左端が 1 回, 右端が 3K 回, 150 回

とに点描した。河川浄化係数 K は 0 ~ -0.4 まで 5 段階に変化させたが K の値によらず Figure 3 と同様な変化が得られた。

Figures 3, 4 から判断すると、観測誤差の無い場合、神経回路網の計算精度は $O(-2)$ 、2 桁である。河川浄化係数 K は 0 ~ -0.4 まで 5 段階に変化させたが K の値によらず Figure 4 と同様な変化が得られた。単純離散モデルと神経回路網法は環境解析に十分な精度を備えている。

河川浄化係数 K は 0 ~ 0.4 まで 5 段階に変化させたが K の値によらず Figure 5 と同様な変化が得られた。折線 d1, d3 は X_1-14 で期待値 1 である。d2 はそのまま河川浄化係数値である (河川浄化項が線形でその偏微分を取ったのであるから、偏微分値 = K である)。負の値は観測点間距離が水中の物質量を減じる方向 (浄化方向) に作用したことを示す。折線 B は神経回路網の非線形関数機能の変化の指標である。各説明変数の self consistency を Table 3 に示す。期待値は共に 0 である。

各説明変数の偏微分係数値の変化は少なく神経回路網が正しく式 (3) を再現したことを示す。特に $O(-4)$ の河川浄化係数 K の値は信頼できる。Table 4 に $K=0 \sim -0.4$ の各 d_j の偏微分係数値を示す。d2 が本計算で求める河川浄化係数である。expectation はその期待値

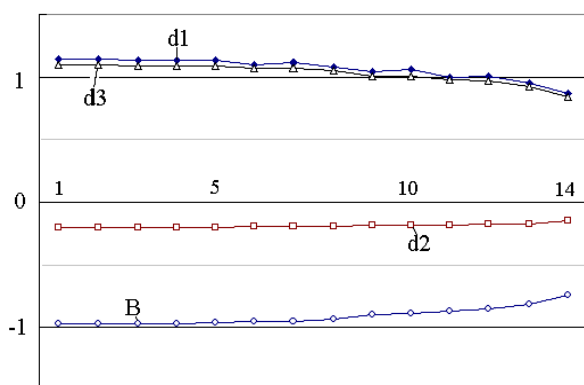


Figure 5. Partial derivatives of three descriptors and a bias. The vertical axis is index for changes of substance amounts. The changes are calculated by partial derivatives of a neural network on $K=-0.2$. The dimension is [mass]. The horizontal axis is the observation points, X_i . The left side is the upper stream. The curves d1-3 are partial derivatives of $\{s_{i-1}\}$, $\{d_i\}$, $\{f_{ipi}\}$. The B shows a partial derivative of a bias.

である。

Table 4 は河川浄化係数 K が高精度で計算されたことを示す。神経回路網に学習させたデータに誤差が無く想定モデルに合致したデータで、かつ不足が無いならば $\pm 2\%$ の精度で計算可能である。

3 堰効果を導入した離散モデル

3.1 堰モデル定義

多摩川の航空写真調査から無視できないと考えられる堰の効果を単純離散モデルに導入する。元来、堰は農業用水等の取水目的に作られたが堰は小さなダムであるから、そこに水が滞留し化学反応により水質指標値が変化する。その変化は考慮する必要がある。

堰の性質として 1. 水の滞留時間, 2. 水温, 3. 水深, 4. pH などのデータが公表されていることが望ましいが、そういうデータは入手困難である。その場合、航空写真 [8] で堰の規模と水面の状況を調べる。多摩川、上中流域の 8 堰で調査したところ規模はほぼ同じと判断した。水面の状況から滞留時間、水温についてもおおむね同じであろう。外国の大川のように著しい相異がある場合は以下の方法は適用が困難であるが「同

Table 3. Self consistency of the simple discrete model as for descriptors and a bias. The d1-3 are same as Figure 5.

d1	d2 (K)	d3	B
0.0063	0.0002	0.0057	0.0046

Table 4. Partial derivative coefficients for descriptors d1-3 and a bias.

expectation	0.00	-0.10	-0.20	-0.30	-0.40
d1	1.07	1.07	1.06	1.06	1.05
d2 (K)	0.02	-0.08	-0.19	-0.29	-0.40
d3	1.02	1.02	1.03	1.03	1.04
B	-1.63	-1.28	-0.91	-0.43	-0.13

じ」場合には「各堰の特徴を無視し全堰で同じ」と近似する。そうすると神経回路網のために堰を表現する形式として scaling[9] により堰の存在非存在のみが有意となる。したがって 0/1 要素のベクトルとなる。このベクトルの存在は構造活性相関研究分野 [10] では、ダミー変数として用いられてきた。構造活性相関では構造を表現する適切な物理化学的測定手段が不明の場合、あるいは従来指標では表せない効果を想定して使用されている。それらは主要な原因を説明する変数として導入されず、また少数の大きな外れ値 (outlier) を説明する変数としては用いられていない。

環境問題ではダミー変数は構造活性相関的にも外れ値、突発的事故などによる環境負荷の一時的増大の解析にも使用できる可能性がある。本論文は前者の場合を扱う。後者は今後の課題である。我々の考えている外れ値を付録に示した。

ダミー変数は多用すると危険性である [11] が、注意して使用すれば多摩川、ドナウ川の水質問題については有効であった [12, 13]。注意深く使用することが前提であるが、我々はダミー変数は常時有効データ不足である環境科学の研究にもっと採用されて良い方法と信じる。

我々はダミー変数形式の堰データを導入した堰モデルを提案する。アルゴリズムは以下である。

- (1)-(6) まで第 2.1 節と同じである。
- (7) 堰を観測点 X_i , $i=2,6,10,14,\dots$ に設定する。
- (8) 本流 X_i 地点の水に含まれる物質は

$$s_i = s_{i-1} + f_i p_i + K d_i + L h_i \quad (11)$$

である。 $1 < i$ である。 $s_1=0$ とする。ここで K は観測点 X_{i-1} と X_i の間の河川水浄化係数 (負が浄化方向, 正が汚染方向) である。 L は堰の浄化係数, h_i は堰の存在 / 非存在を示すダミー変数である。制約条件は $0 < s_i$ である。

- (9) 本流の水質基準値 $\{q_i\}$ は,

$$q_i = s_i / g_i, \quad 1 < i, \quad (12)$$

である。

3.2 神経回路網用堰モデルデータ

第 3.1 節のアルゴリズムで作成した神経回路網用堰モデルデータの一例 ($K=-0.2$ の場合) を Table 5 に示す。

d1-4 の四種類の説明変数 (ベクトル) から河川と堰の浄化係数 K, L を神経回路網を使用して数値的に計算することが目的である。これは第 1.3 節の逆問題よりも説明変数が多く困難な問題である。堰モデルで各 X_i 点で計算した河川水中の物質量を Figure 6 に示す。

Figure 6 において、実際の河川 (多摩川) で観測されたデータと類似している $K=-0.1/-0.2$ がシミュレーション用として適当に思われる。 $K=-0.4$ では物質量が 2 番目の観測点で -0.013 となり、わずかに式 (5) の制約条件に触れる。神経回路網シミュレーションは可能であるが、実際の河川で採用するのは適当でない。

Table 6 に $K=0 \sim -0.4$ の各 d_j の偏微分係数値を示す。 $d_{2,4}$ が本計算で求める河川と堰の浄化係数である。expectation はその期待値である。

Table 6 において、 $d_{1,3}$ は支流からの流入物質、本流の上流側から流入する物質である。河川と堰以外に物質を増減する要因は無いので、それらの期待値は 1 である。B 欄は神経回路網の非線形関数機能の変化を示す。本表は河川と堰の浄化係数 K, L が高精度で

Table 5. Dam model data for a neural network. These data are generated by uniform random data in algorithm (1-9). Descriptor d1: inflow substance amounts at observation points, $\{f_i p_i\}$. Descriptor d2: distances among observations, $\{d_i\}$. Descriptor d3: substance amounts at upper observations, $\{s_{i-1}\}$. Descriptor d4: dummy variable for existence of a dam. Teaching data: $\{s_i\}$. $K=-0.2$, $L=-0.15$;

	d1	d2	d3	d4	T
1	0.7399	1.4803	0.0000	0	0.4438
2	0.4795	1.2653	0.4438	1	0.5203
3	0.9468	1.0436	0.5203	0	1.2583
4	0.5422	0.6244	1.2583	0	1.6756
5	0.2720	0.8682	1.6756	0	1.7740
6	0.9507	0.5127	1.7740	1	2.4721
7	0.0566	1.6193	2.4721	0	2.2048
8	1.0308	0.7241	2.2048	0	3.0908
9	1.1093	1.5378	3.0908	0	3.8926
10	0.1438	1.1489	3.8926	1	3.6566
11	1.1838	1.9709	3.6566	0	4.4462
12	0.5434	1.7954	4.4462	0	4.6306
13	1.0854	0.9882	4.6306	0	5.5184
14	1.3816	0.6495	5.5184	1	6.6201

Table 6. Partial derivative coefficients for descriptors d1-4 and a bias in the dam model. Descriptor d1: inflow substance amounts at observation points, $\{fipi\}$. Descriptor d2(K): distances among observations, $\{di\}$. Descriptor d3: substance amounts at upper observations, $\{si-1\}$. Descriptor d4(L): dummy variable for existence of a dam. "B" is that of a bias neuron.

expectation	0/-0.15	-0.1/-0.15	-0.2/-0.15	-0.3/-0.15	-0.4/-0.15
d1	1.00	1.01	1.01	1.01	1.01
d2 (K)	-0.04	-0.12	-0.22	-0.31	-0.40
d3	1.00	1.00	1.00	1.01	1.01
d4 (L)	-0.13	-0.13	-0.13	-0.13	-0.14
B	-4.25	-3.52	-2.73	-1.96	-1.30

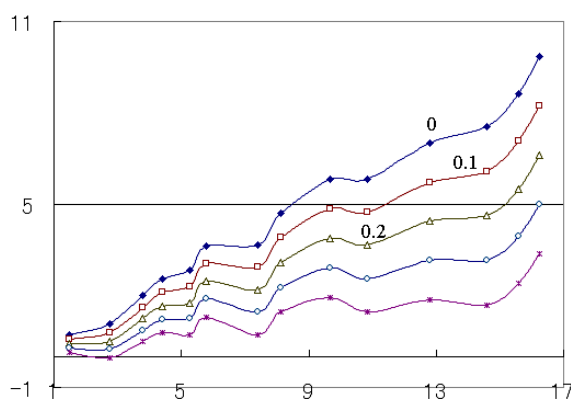


Figure 6. Substance amounts calculated by the dam model. The vertical axis is the substance amounts that have a dimension of [mass]. The horizontal axis is sampling points, which are numbered. The points are located by the accumulated distances. The left side is the upper stream. Whole data are generated by uniform random numbers. In the figure, digits are the river purification coefficient, K , which are multiplied by (-). The dam coefficient, L , is fixed as -0.15 .

算出されたことを示す。神経回路網に学習させたデータに誤差が無く、想定モデルに合致しデータで、かつ不足が無いならば $\pm 4\%$ の精度で河川浄化係数が、 $\pm 2\%$ の精度で堰の浄化係数が同時に予測可能である。Table 7 に各説明変数の self consistency を示す。d2(K), d4(L) とともに $O(-4)$ でモデルの一貫性に問題はない。以上の調査から河川水質データの離散形は環境解析に充分適用できると考えられる。

Table 7. Self consistency of d1-d4 and bias, in case of $K=-0.20, L=-0.15$.

d1	d2 (K)	d3	d4 (L)	B
0.0030	0.0001	0.0027	0.0001	0.0247

4 地下浸透を導入した離散モデル

前節までのモデルの検証では説明変数に不足は無かった。モデルに多くの要因を説明変数として導入すると現象記述機能は高くなる。一方データを用意することが困難になる。説明変数データが不足した状態のモデル精密化の効果が神経回路網出力にどう影響するかを調査する。

4.1 地下浸透モデル定義

堰モデルに地下浸透効果を導入する。地下浸透データの実測はほとんど無い。本節では地下浸透データを説明変数に加えずに神経回路網を学習させる。そのようなデータ不足の状態でも学習しても神経回路網は自動関数適合機能により学習を完了する。それは「不十分な説明変数で現象を再現する」という状態である。その状態の現象再現性の程度を調べることは、我々が不十分な情報から結論を導く時の「結果の信頼性」を調査することに通じる。環境問題解析はこのような状況にあるのではないと思われる。地下浸透モデルのアルゴリズムは以下である。

<1>-<7> まで第 3.1 節と同じである。

<8> 地下浸透量を $[0, 0.5]$ 区間の乱数とする。理科年表環境編 [14] の日本の河川データを見る限り妥当な

オーダである．地下浸透は流出する支流として扱う． $\{y_i\}$ と表記する．

(9) 本流 X_i 地点の水に含まれる物質は，

$$s_i = s_{i-1} + f_i p_i + K d_i + L h_i - q_{i-1} y_i \quad (13)$$

である． $1 < i$ である． $s_1=0$ とする．ここで K は観測点 X_{i-1} と X_i の間の河川水浄化係数である． L は堰の浄化係数， h_i は堰の存在 / 非存在を示すダミー変数である． q_i は (10) で定義される本流の水質基準値である (第 2.1 節，第 3.1 節の式 (4) と同じ)．制約条件は $0 < s_i$ である．

(10) 本流の水質基準値 $\{q_i\}$ は $q_i = s_i / g_i$ ， $1 < i$ である．

4.2 神経回路網用地下浸透モデルデータ

第 4.1 節のアルゴリズムで作成した神経回路網用地下浸透モデルデータの一部 ($K=+0.1, L=-0.15$ の場合) を Table 8 に示す．

Table 8. Learning data for the underground-penetration model. These data are generated by uniform random data in algorithm (1-10). Descriptor d1: inflow substance amounts at observation points, $\{f_i p_i\}$. Descriptor d2(K): distances among observations, $\{d_i\}$. Descriptor d3: substance amounts at upper observations, $\{s_{i-1}\}$. Descriptor d4(L): dummy variable for existence of a dam. Teaching data is T, $\{s_i\}$. The descriptor for $\{q_{i-1} y_i\}$ is not given.

	d1	d2 (K)	d3	d4 (L)	T
1	0.7399	1.4803	0	0	0.8879
2	0.7949	1.1492	0.8879	1	1.4106
3	0.0919	0.6848	1.4106	0	1.5015
4	0.2582	0.8682	1.5015	0	1.8440
5	0.4866	1.7244	1.8440	0	2.4889
6	0.2062	1.4851	2.4889	1	2.5818
7	1.0336	1.5378	2.5818	0	3.6001
8	0.0202	0.7137	3.6001	0	3.3069
9	1.0437	1.1885	3.3069	0	4.1506
10	1.0353	0.9882	4.1506	1	5.0930
11	0.4466	1.9617	5.0930	0	5.3559
12	0.1531	0.6278	5.3559	0	5.5334
13	0.5765	1.8577	5.5334	0	6.1387
14	0.0724	1.4772	6.1387	1	6.0167

4.3 説明変数不足の場合の地下浸透モデルの解析

Table 8 のデータは地下浸透に関する説明変数 (式 (13) の $q_{i-1} y_i$ 項) が不足している．それと Table 9 の学習パラメータを用いて河川の神経回路網解析を行い Figure 7 の結果を得た．

Table 9. Learning parameters for the underground-penetration model. The number of neurons on the second layer is 3 that are decreased from initial value 8. One of the three neurons is a bias. The bias neuron is on the first layer, too.

# of data(descriptor., teach.)	(14×4, 14)
Neurons on 1,2,3-layers	5, 3, 1
Emulation on 2nd-layer	sigmoid
Emulation on 3rd-layer	linear
# of learning	3K
Learning Const.	0.2, 0.2
Reconstruction	60 times
Erasing factor	0.04
Square of BP error	0.034*

* in case of $K=+0.1, L=-0.15$.

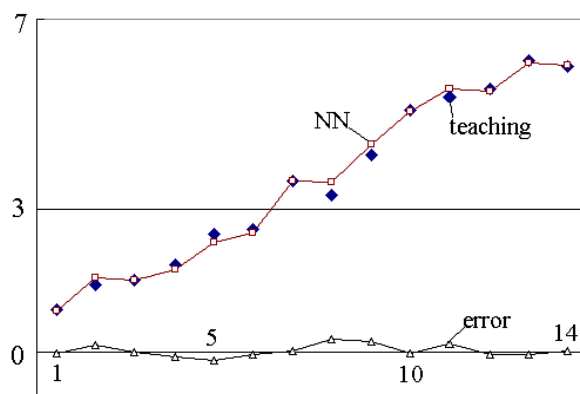


Figure 7. Neural network outputs and the teaching data after learning iterations of 3K. The vertical axis is substance amounts that are calculated by $K=+0.1, L=-0.15$. The dimension is [mass]. The horizontal axis is the observation points, X_i . The left side is the upper stream. The black points are the teaching data, and the bend lines are the outputs. The error line is the difference.

Table 10. River and dam purification coefficients calculated by the underground-penetration model.

expectation	0.1/-0.15	0.0/-0.15	-0.1/-0.15	-0.2/-0.15	-0.3/-0.15
d1	0.99	0.99	0.99	0.99	1.00
d2 (K)	0.01	-0.08	-0.16	-0.25	-0.33
d2(K+q _{i-1} y _i)	0.07	-0.03	-0.13	-0.23	-0.33
d3	0.96	0.96	0.96	0.96	0.97
d4 (L)	-0.13	-0.14	-0.15	-0.15	-0.16
B	-4.08	-3.31	-2.62	-1.90	-0.97

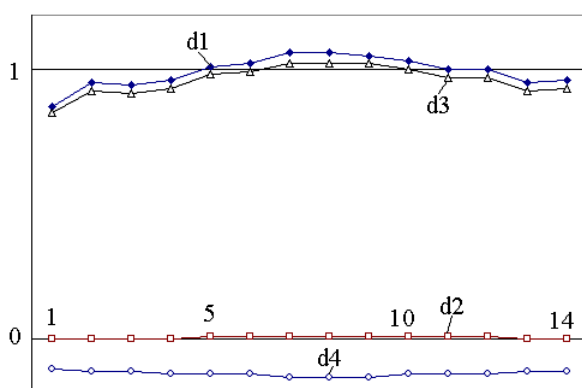


Figure 8. Changes of partial derivatives of four descriptors. The vertical axis is an index for changes of substance amounts. The changes are calculated by partial derivatives of a neural network on $K=+0.1$, $L=-0.15$. The dimension is [mass]. The horizontal axis is the observation points, X_i . The left side is the upper stream. The curves d1-4 are partial derivatives of $\{s_{i-1}\}$, $\{d_i\}$, $\{f_{ipi}\}$, and dummy variable for existence of a dam.

河川浄化係数 K は $-0.1 \sim -0.3$ まで 5 段階に変化させたが K の値によらず Figure 7 と同様な変化が得られた。説明変数不足にもかかわらず神経回路網の出力値は期待値を精度良く再現する。神経回路網に自動関数適合機能が存在するためである。これは環境問題解析に重要な問題で、注意して解析しないと誤った結論を導き易い。Figure 8 に同神経回路網から得られた各 X_i 点の偏微分係数を示す。

河川浄化係数 K は $0.1 \sim 0.3$ まで 5 段階に変化させたが K の値によらず Figure 8 と同様な変化が得られた。説明変数 $d1, d3$ は $X1-14$ で期待値 1 である。Table 10 に各 K, L 値の期待値と神経回路網の計算した値を示す。

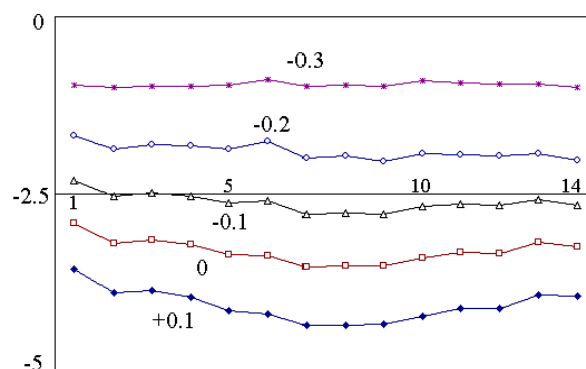


Figure 9. Changes of partial derivatives for a bias neuron. The vertical axis is amplitude of partial derivatives calculated by $K=+0.1$, $L=-0.15$. The horizontal axis is the observation points, X_i . The left side is the upper stream.

Table 10 において、河川浄化係数値 ($d2, K$) が $K > -0.30$ で正しくない。しかし堰の係数 ($d4, L$) は正しい。河川浄化係数値は地下浸透項の平均値 (-0.0287) を加算した値 ($d2, K+q_{i-1}y_i$) にほぼ等しい。シミュレーションではなく実測定値で解析する場合、データ不足の場合は正しくない結果となる恐れがある。それを self consistency 値で判定できるか否かを調べる。同値を Table 11 に示す。

Self consistency の値では $K=+0.1$ の場合のバイアス・ニューロンに関する偏微分係数値が異常をしめしている。これは神経回路網で不足する説明変数を補う何らかの関数機能変化が起こっていることを示す。しかし、他の偏微分係数値は Table 7 と比較して特に異常ではない。 $K=+0.1$ の場合は不足が著しいため検出できたと思われる。ゆえに self consistency 値だけではデータ不足状態を検出するには不十分である。この結果は神経回路網解析の限界と思う。バイアス・ニューロンに関する偏微分係数を Figure 9 に示す。

Table 11. Self consistency of the underground-penetration model.

K	L	d1	d2(K)	d3	d4(L)	B
+0.1	-0.15	0.0029	0.0000	0.0026	0.0001	0.0482
0.0	-0.15	0.0024	0.0000	0.0024	0.0001	0.0268
-0.1	-0.15	0.0022	0.0001	0.0022	0.0001	0.0159
-0.2	-0.15	0.0025	0.0001	0.0026	0.0001	0.0102

Table 12. Correlations between river/dam purifications and underground-penetration terms.

K	+0.10	0.00	-0.10	-0.20	-0.30
d2 vs $q_{i-1}y_i$	-0.16	-0.17	-0.18	-0.20	-0.23
d4 vs $q_{i-1}y_i$	0.03	0.03	-0.00	-0.04	-0.13

地下浸透効果は各 X_i 点ごとに違い、河川浄化係数にパターンの類似している。そのため d2 が本来神経回路網に入力されるべき地下浸透量を包含するように神経回路網の中で作用したと考える。その証拠が Table 10 の $d2(K + q_{i-1}y_i)$ 欄と Figure 9 の $K=0.1$ の曲線である。確かに $K=0.1$ のとき d2 の値は最大誤差を示している。一方、堰の効果は Table 12 に示したように相関係数が小さく影響が少ないと判断する。

4.4 観測点で異なる河川浄化係数 k_i の計算

説明変数不足の場合の神経回路網の出力の特徴をさらに調査するため、単純離散モデルを再検討する。第 2.1 節では式 (3) により河川水質のデータを生成し、観測点 X_i への支流からの流入物質質量 f_{ipi} 、 X_{i-1} 観測点との距離 d_i 、本流 X_i 点の物質質量 s_{i-1} を説明変数とし、本流 X_i 点の物質質量 s_i を教師データとし、それらから河川浄化係数 K を求めた。未定係数 K 以外は全てベクトルの要素で与えられている。従って、それらの各要素を満足する平均的な値として K が求まる。

本来の式 (2) では観測点の河川浄化係数 k_i は $kidi$ という両項未定形である。故に 3 説明変数 $\{s_{i-1}, f_{ipi}, kidi\}$ の学習データでは原則として k_i を求めることはできない。それでもその 3 変数ベクトルにより神経回路網を学習し、その偏微分係数から k_i を求めようとすると、

$$s_i = s_{i-1} + f_{ipi} + average(k_i)d_i, \quad (14)$$

となる。神経回路網は非線形ではあるが複数のデータから尤もな未定定数を求める方法の一つであるので、

平均化された k_i が求まる。しかし式 (7) は個別の k_i を求められる可能性を示唆する。ただし可能性であるので数値計算で検証する。神経回路網の k_i 算出機能について追跡容易にするため、 $\{k_i\}$ を特徴ある数列、等差数列とする。河川浄化係数 K を使用して、

$$k_i = K(2i/N), \quad i = \{0, 1, 2, \dots, N\}, \quad (15)$$

とする。このとき式 (7) の $average(k_i)$ 項は項数が $N+1$ なので、

$$average(k_i) = \{K/(N+1)\} \left[\sum_{i=0, N} (2i/N) \right], \quad (16)$$

$$= \{K/(N+1)\} \{N(N+1)/2\} \{2/N\} = K, \quad (17)$$

ここで改めて

$$average(k_i) \equiv average(k_i)_0, \quad (18)$$

と書く。

式 (16) より神経回路網の学習をデータ数 $N+1$ 個ではなく要素 m を除去した N 個で行うと、 $average(k_i)_m$ は $N+1$ の場合より除去した km 分だけ変化する。これを δ_m とすると、

$$\begin{aligned} \delta_m &= K - (K/N) \left[\sum_{i=0, N} (2i/N) - 2m/N \right] \\ &= -(K/N) \{1 - 2m/N\}. \end{aligned} \quad (19)$$

δ_m が高精度に計算できるのなら、

$$\begin{aligned} km &\equiv 2mK/N \\ &= -\{N/(1 - 2m/N)\} \delta_m \\ &\sim -N\delta_m \quad (\because 2m/N \ll 1). \end{aligned} \quad (20)$$

である。負号は基準の取り方による。

式 (20) は小さい値 δ_m を N 倍するので、実測の誤差を含むデータでは実行困難である。無誤差のシミュレーションならば可能性がある。このような観測集合の部

分集合でモデルの性質を調べる方法は、モデルの正当性をその出力から判断する一方式として leave-one-out 法 [15] と呼ばれて来た。式 (20) の方法は神経回路網の出力値ではなく偏微分係数から算出される項の性質に関する点が本来の leave-one-out 法とは異なる。

式 (20) は小値 δ_m を N 倍しているので除去要素 m の単独計算だけからは精度良く km が求まらない。ここで $\{km\}$ の等差数列の性質を利用し、除去要素 m の集合 $\{m; m=0,1,\dots,N\}$ を考え各々の km を求め、平均変化 (勾配という) を求めた。式 (20) の *average* 項を使い、

$$\begin{aligned} average(ki)_m - average(ki)_{m+1} \\ = -N(\delta_m - \delta_{m+1}) = 2K/N, \end{aligned} \quad (21)$$

となる。ゆえに $\{km\}$ の勾配 $2K/N$ が求められる。3 変数の尤度方程式が複数あることを利用して、4 説明変数の情報を得ることになるので、一種の平均量が求められることを限度と考える。

ここで δ_m をどの説明変数について調べれば良いかを議論する。常識的には式 (2) の *kidi* 項であるが、神経回路網に入力するのは値だけであり、その値が ki と di から計算されたのものであることを提示していない。従って、神経回路網は他の説明変数と区別することが出来ず、どの説明変数の偏微分係数に表れるかは不明である。ここで「神経回路網は複数データの最尤定数求解法」であることを利用する。従って偏微分係数は本来ならば定数になるので、そうならない係数がその候補である。

環境問題解析では有効観測数は多くないので観測データの部分集合化は計算精度低下をもたらす。leave-one-out 法の拡張 leave-n-out 法は不可能な場合が多い

Table 13. Leave-n-out method expression for the slope of arithmetic progression.

leave-n-out	expression
1	$2K/N^*$
2	$6K/(N-1)$
3	$12K/(N-2)$
4	$20K/(N-3)$
5	$30K/(N-4)$
6	$42K/(N-5)$
7	$56K/(N-6)$

* N is the number of data.

が、それが可能ならば Table 13 を利用すれば勾配に関する情報は多くなる。この単純離散形式の拡張型を双未定係数モデルと呼ぶ。

神経回路網は隠れ層のニューロン数を多くすれば学習点に関し教師データと出力値の差は 0 に漸近する。そのため神経回路網を使用したデータ解析において外れ値 (outlier) を見出すのは難しくその議論は少ない。しかも本論文の神経回路網は隠れ層を第 2 層のみとし、そのニューロン・エミュレーション関数を sigmoid 関数とし、かつ第 3 層の同関数を線形関数とした (Tables 2, 9)。かつ reconstruction learning により第 2 層のニューロン数を 2 まで減少させ、神経回路網の非線形 fitting 能力を限定した (Figure 3 参照)。さらに leave-one-out 法を使用している。従って、外れ値の存在が疑われる「観測データ」の場合、それが与える影響を除去しなければならぬ場合がある。ただし外れ値の議論は本論文の趣旨とは少し離れるので付録 A に記した。

4.5 河川浄化係数 ki の計算例

第 4.4 節の方法を検証するため、第 2.1 節のアルゴリズムで河川水質のデータを生成し、観測点 Xi への支流からの流入物質質量 $fipi$ 、 $Xi-1$ 観測点との距離 di 、本流 $Xi-1$ 点の物質質量 $si-1$ を説明変数とした。

河川浄化係数 ki については第 4.4 節の式 (8) を用いた。このとき $K=-0.1, N=14$ であった。本流 Xi 点の物質質量 si を教師データとした。Table 14 に 3 説明変数 (*kidi* は 1 変数である)、1 教師データ (各データ数 14) の場合の神経回路網学習パラメータを示す。Figure 10 に $K=$ 定数の場合との相異を図示した。定数との相異は図示すると僅かである。神経回路網の学習方法は back propagation+reconstruction learning を用いた。本計算の目的は、学習が停留状態になった神経回路網から第 4.4 節の方法で河川浄化係数 $\{ki\}$ の勾配の平均値 $average(ki) = -0.1*2/(14-1) = -0.015$ を求めることである。

学習が停留状態になった神経回路網の出力値とその期待値、その差を Figure 11 に示す。図は不完全な学習データであっても神経回路網が高精度で現象を再現することを表している。それは神経回路網の非線形 fitting 機能であるが、これで現象が説明できたと考えると危険である。真に説明できたか否かは偏微分係数を調べる必要がある。Figure 12 に各説明変数とバイアスの偏微分係数を示す。

Table 14. Learning data to calculate variable river-purify coefficients $\{ki\}$. The d1,2,3 are descriptors. The d1 is *fipi*. The term *fipi* is substance amounts from branches into the main stream. Where the branches are a generic name of branch rivers' water and sewage and rainwater. The d2 is the multiplied values, *kidi*, which are calculated by distances and river-purify coefficients. They are unknown variables; however, they are processed as a descriptor here. The d3 is substance amounts at X_i point in main stream, i.e., *si-1*. The T is teaching data, *si*.

	d1	d2	d3	T
1	0.7399	1.4803	0.0000	0.7399
2	0.4795	1.2653	0.7399	1.1999
3	0.9468	1.0436	1.1999	2.1146
4	0.5422	0.6244	2.1146	2.6279
5	0.2720	0.8682	2.6279	2.8465
6	0.9507	0.5127	2.8465	3.7578
7	0.0566	1.6193	3.7578	3.6649
8	1.0308	0.7241	3.6649	4.6177
9	1.1093	1.5378	4.6177	5.5377
10	0.1438	1.1489	5.5377	5.5224
11	1.1838	1.9709	5.5224	6.4030
12	0.5434	1.7954	6.4030	6.6426
13	1.0854	0.9882	6.6426	7.5456
14	1.3816	0.6495	7.5456	8.7973

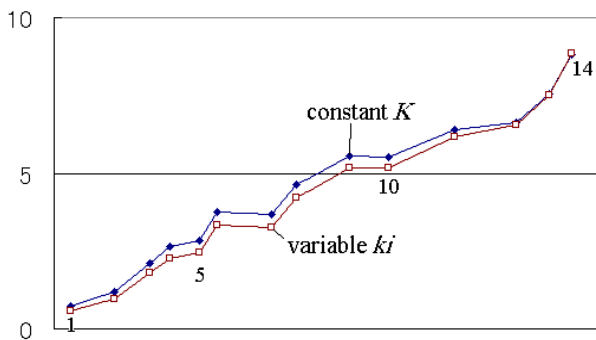


Figure 10. The difference between variables and a constant for the river purification. The vertical axis is the substance amount, whose dimension is [mass/volume]. The horizontal axis is ordered observation points, where the left side is upper stream. Total river purification power is same for the both conditions; i.e., $K=-0.1=\sum ki/N$.

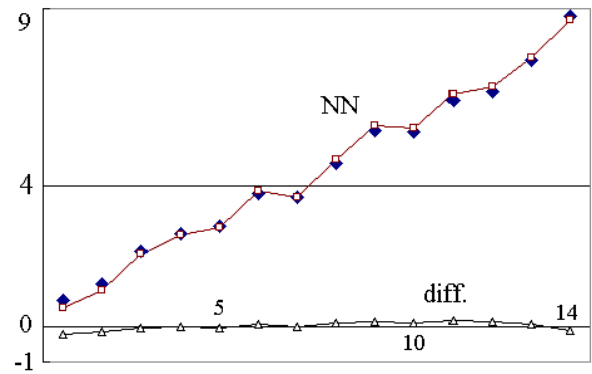


Figure 11. The difference between neural network's outputs and the expectations. The diamond-shaped black points are expectations, and the NN curve is outputs, which show the fitting-ability of the neural network. The figure shows fitting-ability even if the learning data are incomplete.

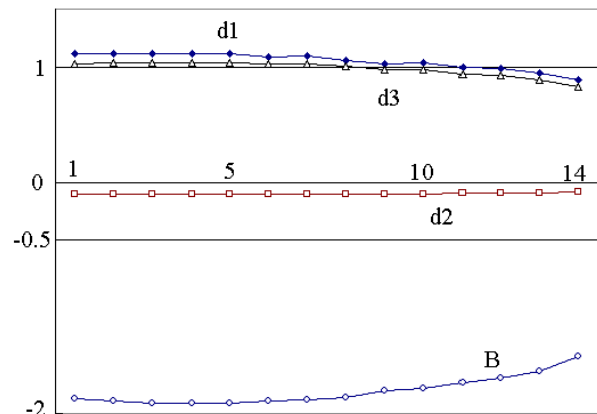


Figure 12. Partial derivatives for 3-descriptors and bias. The d1-3s are in Table 11. The B is partial derivative for a bias on the first layer. The vertical axis is amplitude of partial derivatives, whose unit is [mass/volume]; where the denominator is uniform scaled derivatives in [0.05, 0.95]. The bias is fixed values; however, it is same characters to descriptors for a neural network. Therefore, we calculate the bias distribution to estimate neural network's function.

Table 15. Self consistency of extended discrete river model.

d1	d2	d3	B
0.0045	0.0000	0.0041	0.0141

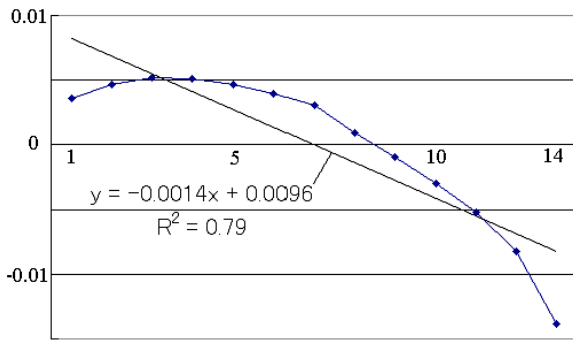


Figure 13. Change of $\{P_i\}$ that is derived from eq. (24). The change relates the slope of $\{k_i\}$; this is a hypothesis. Calculated slope is -0.0014 , the expectation is -0.015 .

Figure 12 において，各説明変数の値は「神経回路網は複数データの最尤定数求解法」であるためには一定である必要がある．Self consistency を Table 15 に示す．

Self consistency が不十分な係数はバイアス・ニューロンの係数である．この変化は神経回路網の関数機能がデータにより変化している割合を示す．そこでは不足する説明変数を神経回路網の内部で補っている可能性を示している．その内部補正量の大きさは，神経回路網の出力値を y_i (添字 i はデータの順序番号である) とするとき，

$$\partial y_i / \partial B = const., \quad (22)$$

である条件から求められる．定数値は不明であるので，

$$\left(\sum_i \partial y_i / \partial B \right) / N = K_B, \quad (23)$$

と仮定し，

$$P_i = K_B - \partial y_i / \partial B, \quad (24)$$

とする．ベクトル $\{P_i\}$ の要素値の傾きを最小二乗法で直線に fitting して調べる．それを Figure 13 に示す．計算された値は -0.0014 で期待値 -0.015 に符号は一致する．式 (23) の仮定は傾きには関係しないので，この結果は受け入れることができる．ただし最小二乗法の直線 fit 時の $R^2=0.79$ は理論値 1 に近くない．それはこの方法の限界を示すように思われる．

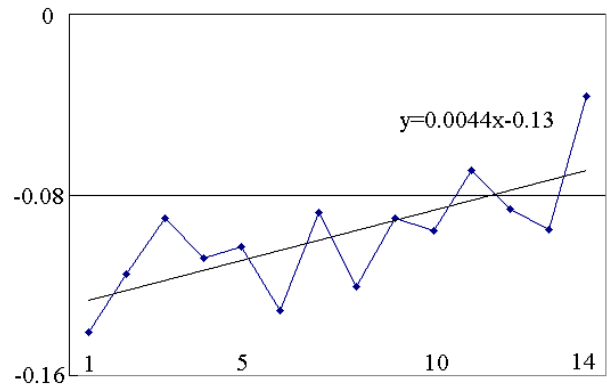


Figure 14. Coefficient $\{k_i\}$ estimated by descriptor-d2 from leave-one-out method and eq. (20). The vertical axis is the river purify coefficient, [mass/distance]. The horizontal axis is the observation points. The left side is upper stream.

以上 leave-one-out 法とバイアスの偏微分係数から説明変数不足の場合でも「データのみから不足する説明変数の性質を逆算できる」可能性を示した．それが可能であるのは「モデルが推測でき，かつそれが正しい場合」である．

説明変数 $d2(kidi$ 項に相当する) の leave-one-out 法の偏微分係数の変化と式 (20) からは km が直接求められる可能性がある．説明変数 $d2$ の偏微分係数は各観測点でほとんど変わっていない (Table 15 の $d2$) ため，その可能性がある．計算結果を Figure 14 に示す．河川浄化係数が各観測点ごとに違い，かつその変化がおおむね直線的であることは正しく算出されている．しかし傾きが 0.0044 で期待値 -0.015 とは違う．この結果は第 4.4 節の式 (20) 方法の限界を示している．

5 結論

現代の都市部を流域とする河川の水質について離散形の四モデルについて考察した．その離散モデルで記述された人工河川の水質を神経回路網により解析可能であるか否かを研究した．神経回路網は非線形現象の解析に有用であるが，その機能が学習により獲得するためには説明変数と観測値に不足があってはならない．不足とは観測データが本来あるべきことが自明であるのに何らかの理由で不明な場合と，現象を説明する観測自体がされていないか，そういう観測自体の存在不明の三種類がある．前者を欠測という．欠測を含

んだデータの情報処理は近年かなり研究されているが後二者はほとんど検討されていない。我々は環境問題の真の理解のために後者を回避できない問題と考える。同時にそれは神経回路網出力の検定に関わる問題である。本論文の目的はこの観測データ不足問題を考察することであった。

本論文では神経回路網を使用した非線形多変量解析のデータ不足状況における挙動を調査し同回路網解析の適用限界を明らかにするため、観測データに含まれる測定誤差を除外することにした。測定誤差の統計的性質は確かでない。それを基に神経回路網の適用限界の議論は困難である。この測定誤差問題を回避するため河川モデルを考案し、乱数を基に人工河川を定義し、そのデータを基に神経回路網の適用限界を明らかにする方針を採用した。

我々の提示した河川モデルは単純離散モデル、堰モデル、地下浸透モデル、双未定係数モデルである。単純離散モデル、堰モデルについてはデータが完全な場合を計算し河川と堰の水質浄化力を誤差 $\pm 2 \sim 4\%$ の精度でシミュレーションできた。このレベルが神経回路網の本来の能力と我々は考える。環境問題を議論するのに必要十分である。

地下浸透モデル、双未定係数モデルでは、あえて必要な説明変数データを省略してデータ不足な場合を作った。このデータ不足状態であっても神経回路網は精度良く河川水中の物質量をシミュレーションする。しかし偏微分値を計算すると、地下浸透モデルでは河川の浄化機能が正しく計算できていないことが明らかになった。その理由は説明変数にない地下浸透効果を河川浄化機能が補完したためである。両者の物理化学的性質はモデルでは類似している。双未定係数モデルでは河川浄化機能の平均的变化を神経回路網のバイアス部が補完した。その補完機能を利用し偏微分係数から河川浄化機能の平均的变化を逆算できることを示した。ただし理論上の可能性だけである。数値計算では符号のような大まかな特徴だけが逆算できた。以上の結果は与えられていない説明変数の効果がデータのみからある程度は推定できることを示す。それが可能であるのは「モデルが推測でき、かつそれが正しい場合」であるが、このような神経回路網の未知の使用法が明らかになったことは環境問題の議論を定量化する観点からは有用である。

参考文献

- [1] 大垣眞一郎監修, 河川環境管理財団編, 「河川と栄養塩類 管理に向けての提言」, *foundation of river and watershed environment management*, 技報堂出版 (2005), ISBN4-7655-3403-0.
- [2] 日本水環境学会編, 日本の水環境 3 関東・甲信越編, 技報堂出版 (2000), ISBN4-7655-3167-8.
- [3] 末次忠司, 河原能久, 賈仰文, 倪广恒, 都市河川流域における水・熱循環の統合解析モデルの開発, 土木研究所資料, 第 3713 号 (2000), 河川部都市河川研究室 ISSN 0386-5878.
- [4] 賈仰文, 倪广恒, 河原能久, 末次忠司, 都市河川流域の水循環解析と雨水浸透施設の効果の評価, 水工学論文集, 44, 151-156 (2000).
- [5] Goloka B. Sahoo, Chittaranjan Ray, Jack Z. Wang, Stephen A. Hubbs, Rengao Song, Jay Jasperse, Donald Seymour, Use of artificial neural networks to evaluate the effectiveness of riverbank filtration, *Water Research*, 39, 2505-2516 (2005).
www.elsevier.com/locate/watres
- [6] 渡辺美智子, 山口和範, *EM アルゴリズムと不完全データの諸問題*, 多賀出版, ISBN4-8115-5701-8.
東洋大, 渡辺教授の Web 資料:
[http://stat.eco.toyo.ac.jp/~michiko/em/emohp\(0\)/emohp.ppt](http://stat.eco.toyo.ac.jp/~michiko/em/emohp(0)/emohp.ppt)
- [7] 公共用水域水質測定結果データ集:
<http://www2.kankyo.metro.tokyo.jp/kansi/mizu/sokutei/sokuteikekka/suishitu.htm>. (CSV 形式)
東京都下水道局 HP:
<http://www2.kankyo.metro.tokyo.jp/kansi/mizu/sokutei/sokuteikekka/suishitu.htm>.
東京都下水道局事業概要 平成 16 年版
第 4 章「流域下水道主要事業の展開」
<http://www.gesui.metro.tokyo.jp/gijyutou/jg16/jigyougaiyou16/no4.pdf>. (PDF 形式)
- [8] 国土地理院 (Geographical Survey Institute)
<http://mapbrowse.gsi.go.jp/airphoto/indexmap200k/5339/533934.html>.
1/4000 ~ 1/20000 大縮尺航空写真 (無料) の一例
<http://www.ikutoko.com/>

- [9] 青山智夫, 神部順子, 長嶋雲兵, 欠測データ集合を扱う神経回路網法 CQSAR: Compensation Quantitative structure-Activity Relationships の開発, *J. Comput. Chem., Jpn.*, 投稿中 (2005.10.27).
- [10] 構造活性相関懇話会, 薬物の構造活性相関 ドラッグデザインと作用機作研究への指針 化学の領域増刊 122 号, 南江堂, 東京 (1979).
- [11] 青山智夫, 神部順子, 長嶋雲兵, 階層型神経回路網出力の検定法, *J. Comput. Chem., Jpn.*, 投稿中 (2005.10.13).
- [12] J. Kambe, Y. Yuan, T. Aoyama, U. Nagashima, Neural network analysis of water pollution for a main river, Tamagawa, in Tokyo metropolis, *Proceedings of International Conference on Control Automation and Systems 2004*, TP04-4 (2004).
- [13] J. Kambe, T. Aoyama, U. Nagashima, Comparison with Water Quality of main Rivers in the world, based on OECD reports, *Proceedings of International Conference on Control Automation and Systems 2005*, FE10-4 (2005).
- [14] 大島康行, 浅島誠, 高橋正征, 原沢英夫, 松本忠夫編, 理科年表環境編, 丸善, 東京 (2003), ISBN4-621-07335-4.
国立天文台編, 理科年表環境編 第 2 版, 丸善, 東京 (2006.1.182003), ISBN4-621-07641-8.
- [15] T. Sekiya, in T. Fujita, ed., *Structure-Activity Relationship and Drug Design*, Kagadojin, Kyoto (1986).
- [16] 市川新, 都市河川の環境科学, 培風館, 東京 (1980), 3043-4140-6955.
- [17] Environmental Performance and Information Division, OECD Environment Directorate, "OECD Environmental data, compendium 2002".

A 付録 外れ値

A.1 教師データの外れ値の検出

神経回路網計算において, 外れ値検出は関数適合機能と関係する. 第 2 層 sigmoid 関数, 第 3 層線形関数

の場合, 神経回路網の機能 F は,

$$F(x) = \sum_i ff(x; \alpha_i, \theta_i) + B, \quad (A1)$$

なる有限 sigmoid 関数展開である. ff は式 (8) に同じであるが, パラメータを明示すると,

$$ff(x; \alpha_i, \theta_i) = 1 / \{1 + \exp(-\alpha_i x + \theta_i)\}, \quad (A2)$$

である. パラメータ α_i は BP 学習により第 1,2 層間のニューロン間結合から自動的に算出される. θ_i は第 1 層のバイアス・ニューロンと第 1,2 層間のニューロン間結合から同じく自動的に算出される. 式 (1) の B 項は第 2 層のバイアス・ニューロンと第 2,3 層間のニューロン間結合から自動的に計算される. 式 (22) の F を使い任意の 2 ベクトル $X = \{x_1, x_2, \dots, x_N\}$, $Y = \{y_1, y_2, \dots, y_N\}$ を対応づけると,

$$Y - F(X) = \varepsilon \neq 0, \quad (A3)$$

である. $\varepsilon = \{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N\}$ において各要素は平均値 $\text{average}(\varepsilon) = \mu$ を中心とした正規分布,

$$\rho(x; \sigma, \mu) = \{1 / (\sigma(2\pi)^{0.5})\} \exp[-0.5\{(x - \mu) / \sigma\}^2], \quad (A4)$$

をなすと仮定する. 標本分散 σ を,

$$\sigma = \sum_i (\varepsilon_i - \mu)^2 / (N - 2), \quad (A5)$$

とする (回帰分析なので $N - 2$). ε_i の検定統計量 T_i を,

$$T_i = |\varepsilon_i - \mu| / \sigma^{0.5} > 0, \quad (A6)$$

とする.

$$\int_{[\mu, \mu + \nu]} \rho(x; 1, \mu) dx = P(\nu) \leq 1/2, \nu \geq 0, \quad (A7)$$

なる関数は, 中心 μ から ν 離れた ε_i が正規分布全要素に対比率 $1 - 2P(\nu) = P'(\nu)$ であることを示す. ゆえに $P'(T_i)$ の小さい ε_i が外れ値の候補である. 普通は $P'(T_i) < 0.01$ (or 0.05) なる ε_i 要素を外れ値とする. この基準で (観測値の) 外れ値を除外して神経回路網を使用すれば本論文の議論は外れ値が存在する場合も適用可能と考える. ただし本論文の乱数から作成したデータには外れ値は無いので本節の処理を施す必要はない. M 個の多変量解析の場合式 (A3) は,

$$Y - F(X_1, X_2, \dots, X_M) = \varepsilon \neq 0, \quad (A8)$$

となる。

式(A4)～(A7)の処理は同じである。正規分布でなくとも対称分布ならば上記の考え方はそのまま踏襲できる。

A.2 Leave-one-out法における神経回路網の関数機能の非同一性の検出

Leave-one-out法においては学習データ集合から1データを除外する。するとそのデータ集合に外れ値が無くとも、また同一の初期値をニューロン間の結合に採用し、同一の回数学習を行っても、神経回路網の関数機能が全て同じとはならない。それをどのように調べるのか検討する。

学習データ集合の $m(\leq N)$ 要素を除外したとする。神経回路網に入力する説明変数(複数を) $\{x\}$ とする。出力を $\{y\}$ とする。この学習データから生成された関数を $Fm(x)$ とする。

$$y_{m,j} = Fm(x_j), j = \{1, 2, \dots, N; \text{except } m\}, \quad (\text{A9})$$

こうして $N-1$ 要素のベクトル $\{y_{m,j}\}$ を得る。1データを除去しない場合の出力ベクトル(N 要素)を $\{Y0j\}$ と書く。神経回路網の関数機能に関する標本分散に相当する次の量を得る。

$$\sigma_m = \{1/(N-1)\} \sum_{j(\text{except } m)} \{y_{m,j} - Y0j\}^2, \quad (\text{A10})$$

σ_m もまた N 要素のベクトルである。従って、

$$\sigma_{\alpha\nu} = (1/N) \sum_m \sigma_m, \quad (\text{A11})$$

$$\tau = (1/(N-1)) \sum_m \{\sigma_m - \sigma_{\alpha\nu}\}^2, \quad (\text{A12})$$

$$T_m = |\sigma_m - \sigma_{\alpha\nu}|/\tau^{0.5}, \quad (\text{A13})$$

となるから、正規分布と仮定した場合のその値の正規分布と仮定した場合の T_m 値で示された関数機能の「位置」 $P'(T_m)$ を求めうる。

m 番目のデータを除外した場合、著しく関数機能が他と違った場合は、その値が「外れ値」である可能性が高い。

A.3 外れ値を除外しない場合

A.1-2の方法で検出した外れ値は、通常現象とは無関係な異常値として解析から除外する。

河川水質の予測方法はいくつか知られている。その中に多変量解析を使い測定データ(training data)を学習して、別の測定データ(checking data)の再現性を評価し、水質を決定している要因を取捨選択するGMDH(Group Modeling of Data Handling)法[16]がある。この方法は本論文のように少ない観測値から予測する目的で開発された方法である。このGMDH法でも外れ値の所は予測できないとしている。

我々は一般的な場合、外れ値除外は妥当と考えている。しかし、重金属イオン濃度の水質解析では除外できない場合がある(欧州第二の国際河川、ドナウ川の支流の1994-5年頃の Cd^{2+} , Cu^{2+} , Pb^{2+} イオン濃度値などがその場合である可能性がある[17])。そのとき外れ値をダミー変数により支配方程式の中に取り入れる方法がある。式(11)に取入れた場合、

$$s_i = s_{i-1} + f_i p_i + K d_i + L h_i + M m_i, \quad (\text{A14})$$

である。ここで K は観測点 X_{i-1} と X_i の間の河川水浄化係数(負が浄化方向、正が汚染方向)である。 L は堰の浄化係数、 h_i は堰の存在/非存在を示すダミー変数である。 M は未知の新たにダミー変数として取入れた仮定の「現象」 m_i の係数である。 $\{m_i\}$ は外れ値の所だけが1, 他が0のベクトルである。制約条件は $0 < s_i$ である。

式(A14)を適用して観測値と計算値の差の標準偏差値が急激に0に接近した場合は、ダミー変数に具体的な意味があると考え、変数に該当する事件が起こっていないか調査する必要がある。事件が無ければ、外れ値と考え除外するのが妥当であるが、もし「ある」のなら、それは外れ値を利用した観測データ不足問題の一解決法となる。

Discrete Expressions for the Water Purification in a River, Based on Neural Network Calculations under Incomplete Data Set

Tomoo AOYAMA^{a*}, Junko KAMBE^b and Umpei NAGASHIMA^c

^aFaculty of Engineering, University of Miyazaki
GakuenKihanadaiNishi, Miyazaki, 889-2192 Japan

^bFaculty of Foreign Language, Daito Bunka University
1-9-1 Takashimadaira, Itabashi, Tokyo 175-8571, Japan

^cResearch Institute of Computational Science, National Institute of Advanced Industrial Science and Technology
1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan

*e-mail: t0b217u@cc.miyazaki-u.ac.jp

We propose four models for water purification in a river that flows in a big city area. All models are discrete expressions. They are named a simple expression-, plus dam's effects-, plus underground penetration-, and pair unknown coefficient-models. Using a neural network, we analyze changes of the water quality in a virtual river defined by the models. The objective is to test the simulation-ability of the discrete expressions, and to discuss the possibility of the inverse prediction of the model. If we could predict the inverse operation, we estimate the water purification of a river on use of a data set only.

The neural network is a useful tool to analyze non-linear phenomena. Discrete data set is required in the analysis, which includes observations and descriptor data. The neural network has ability to emulate the phenomena through learning iterations. Defects in the set suspend the iteration. The defects are classified into three cases. The first is that the existence of defect elements is certain but the value is unknown. The second is loss of whole data for a descriptor. The third is uncertainty for the existence of a descriptor. The first case is called "defect", and it was studied recently. However, the latter two were not. It is necessary to discuss the latter two for researches of environmental problems. At the same time, they are important for significance-tests of outputs of the neural networks. We researched neural-network functions on the latter two cases, which are for multi-regression analysis. The main point is to evaluate limits of the functions. The statistical characters of the error are not clear; therefore, to simplify the research, we consider no-error cases. Thus, we define a virtual river whose data are constructed by the uniform random numbers. The defect part is made in the data set on purpose. The researches show the following; 1. A neural network outputs reasonable water quality in a river, even if there is a defect descriptor. 2. The partial derivatives don't indicate accurate descriptor characters, when the target descriptor is not defect one. 3. The cause is another descriptor makes up for the defect; i.e., there are interactions among descriptors. 4. The largest change of whole partial derivatives indicates the complement descriptor. 5. It is possible to calculate characters of the defect descriptor approximately. 6. The possibility is enabling when outlines of phenomena are known. Thus, the discrete expression of a river makes changes of the water quality calculable in general cases, and by the inverse calculations, we can predict the descriptive equation of phenomena by using observation data only.

Keywords: River model, Water purification, Neural network, Partial derivatives, Missing data