ノ ー ト

二 進 木 を 用 い る 河 川 水 質 デ ー タ の 特 徴 抽 出

貝原巳樹雄^{®1}, 高橋 信幸¹, 佐藤 成幸¹, 樋口 陽子¹, 遊佐 里美¹

The feature extraction of water-quality data by using classification and regression trees

Mikio KAIHARA¹, Nobuyuki TAKAHASHI¹, Seikou SATO¹,
Yohko HIGUCHI¹ and Satomi YUSA¹

¹ Department of Chemical Engineering, Ichinoseki National College of Technology, Hagisho, Ichinoseki-shi, Iwate 021-8511

(Received 2 September 2002, Accepted 28 February 2003)

Water-quality data of rivers, for example, COD and BOD are important indices for environmental conservation. Whether we can distinguish between the sampling positions or the fiscal year, the sampling time of the water are depend on the way of characterizing the sampling positions or time of the water. We focused our attentions to one of the decision trees, the classification and the regressions trees (CART), because we should have clear splitting rules between the sampling positions or the sampling fiscal years. As a result of using CART, we found that classification between the upper, middle and lower area of Tama river could be completely done based on the conductivity and the chlorine ion. Furthermore, from the splitting rules between the water-quality data of fiscal year '97, '98 and '99, the important point concerning the purification of Tama river's water-quality, would be considerable reduction of the nitrogen-related chemical species (Total nitrogen, NH₄, NO₂, NO₃). CART should be a useful method for finding the meanings of classifying or the splitting rules of the complicated analytical data or environment data, which have recently increased in number.

Keywords : quality of the river water ; feature extraction ; classification ; classification and regression trees.

1 緒 言

近年、各種測定機器や分析手法の開発に伴って多彩な環境、分析データなどが自動的かつ大量に入手できる状況が実現し、その結果として、大量の情報から種々の特徴を上手に抽出する必要性が生じている。例えば、一昨年、昨年、今年の河川の各採水位置における水質データから、各年ごとの特徴や傾向を知りたい、あるいは各採水位置における水質の特徴を把握したいといった空間的、時間的な変化を伴う情報の特徴抽出は典型的なケースと考えられる。

多種類かつ大量のデータからまとまった知見を得る方法として広く知られているのは、各種の記述統計によってデ

ータの代表値を比較する方法や大量データから情報の損失をなるべく少なくして情報を集約し、構造を単純化したり、視覚化する主成分分析、またデータ間の距離を求め、類似した集団を見つけるクラスター分析などである。

本報告では、クラスター分析と同様、分類の一手法ではあるが、理解しやすい分割条件を求めることができ、しかもその階層構造を含めて新たな知見の発見も可能とされる決定木^{1)~4)}を用いた河川水質の解析事例について報告する。

2 決定木について

決定木は分類や予測を行う際に使われる手法で、決定木を用いる利点としてはだれでも理解できるルール、分割条件を作成できる点が挙げられる。Fig. 1 (Fig. 1, 3, 4 は、

¹ 一関工業高等専門学校：021-8511 岩手県一関市萩荘字高梨

Cl	100.00	
COND	100.00	
T_P	75.68	
COD	50.00	
N03_N	50.00	
T_N	50.00	
BOD	50.00	
P04_P	44.71	
DO	19.10	
PH	0.00	
N02_N	0.00	
NH4_N	0.00	

Varibale Importance

Fig. 1 Classification tree on the water quality of upper, middle and lower area of Tama river and the variable importance

いずれも CART Ver.4.0³⁾にて出力)に見られるように、決定木は、最後にたどり着いたターミナルを葉と見なして上下方向に反転すると、樹木に見えることから分類樹(回帰の場合は回帰樹)とも呼ばれている。最上部の節(ノード)、これは樹木の根に相当していることから、ルートノード(根っこ)とも呼ばれており、最下部に葉(ターミナル)があるように描かれる。最上段のルートノードには多次元情報を持つすべてのサンプルが入っており、ターゲットを最もよく分割するという目標に基づいたテストが行われ、各々のサンプルが左右どちらの節(ノード)へ進むかが決められる。最終的に同じターミナルに属するサンプルは同じ方法で分類され、根っこからターミナルまでの道筋が分類するときに使われる階層的なルールに対応している。

決定木の作成においては、ルートノードにおいて、まず、どの変数が最もよく分類を行うか、つまり、最良の分岐変数は何かを定める。最良の分岐とは、データの多様性を最も減少させるものであり、Giniの多様性指標(次式)や情報量利得比などの基準を用いて分岐の良さが決定される。

$$i(t) = 1 - \sum_{j=1}^C [P(j|t)]^2$$

$i(t)$: ノード t における多様性, 分散指数, C : 分類の

個数, $P(j|t)$: ノード t における第 j 番目の分類に属する試料の存在確立。

CART (Classification and regression trees) では、分岐変数と同じように分岐に使用できる変数を競合変数とし、競合変数を用いて変数重要度 (Variable importance, I_v) を計算する。本報告では、枝葉が複雑になりすぎて、あてはめすぎ(過学習)が起こらないように枝刈り {剪(せん)定} を行う CART を用いて二進木解析を行った。

3 河川水質データの二進木解析

解析に用いた河川の水質の観測データは、東京都環境局発行の「公共用水域および地下水の水質測定結果」⁵⁾(1997～1999年度)のうちから、源流から河口までのデータが比較的均質にそろっている多摩川の多摩湖から多摩川本流を通じて東京湾に注ぐ17のサンプル採取地点 {和田橋, 調布橋, 羽村堰(ぜき), 永田橋, 拝島原水補給点, 拝島橋, 日野橋, 関戸橋, 是政橋, 多摩川原橋, 多摩水道橋, 砧取水点, 第三京浜多摩川橋, 調布取水点, 田園調布堰上, 六郷橋, 大師橋} での水質分析データ3年分を用いた(毎年2月のある1日における2回測定結果の平均値⁵⁾⁶⁾)。このデータ集では、和田橋から拝島橋までを上流域, 日野橋から田園調布堰までを中流域, 六郷橋と大師橋を下流域としてあるため、それを上流, 中流, 下流の指標としてそのまま用いた。また、「公共用水域および地下水の水質測定結果」には、水質を示す化学的パラメーターとして55種類のパラメーターが掲載されている。しかし、本報告では17地点のすべてに測定結果があり、かつ17地点ですべてが同じ値ではない12種類の化学的パラメーター {pH, 溶存酸素(DO), 生物化学的酸素要求量(BOD), 化学的酸素要求量(COD), 全窒素(T_N), 全リン(T_P), 塩化物イオン(Cl⁻), アンモニア性窒素(NH₄-N), 亜硝酸性窒素(NO₂-N), 硝酸性窒素(NO₃-N), リン酸性リン(PO₄-P), 電気伝導度} について解析を行った。電気伝導率(COND)の単位は, mS/m/10で, pHを除く残りの10種類のパラメーターの単位はいずれもmg/lである。また、12種のパラメーターのうち、BODのみ、定量下限値, 0.5 mg/l未満のデータが全51サンプル中14個あり、これらはすべて0.5 mg/lとして解析した。17か所の採水位置は、上流から下流へと順次, aからqのアルファベットに対応させた。Fig. 2に解析に用いた多摩川の採水位置と水質データの年度別プロットを示す。

まず最初に、CARTを用いて河川の上流, 中流, 下流の区別の基準及び各変数の重要度について検討し、上流, 中流, 下流の特徴を考察する。解析方法は、以下、すべてGini法とした。Fig. 1には、各種水質指標から多摩川を上流, 中流, 下流水域と3種にクラス分けする二進木(分類樹)を示す。この二進木の見方を簡単に説明する。

Fig. 2 Water quality data of 17 sampling positions of Tama river ('97, '98, '99 fiscal year)
 The alphabet codes mean the sampling positions (see Table 2).
 ◇: '97 fiscal year; □: '98 fiscal year; △: '99 fiscal year

Fig. 3 Classification tree on the water quality of 6 points of the upper area of Tama river

ルートノード (Node1: 各年度とも 17 か所で, 3 年分, 合計 51 サンプル) における最良分岐変数は, COND 及び Cl^- (最良の競合変数) で, これが 231.5 以下の場合, 左に進み, ターミナル (Terminal Node1: 18 サンプルを上流域と判定) にたどり着く. 一方, COND が 231.5 より大きい場合, 右に進み, ノード (Node2) にたどり着く. ここで, 更に COND が 4567 より小さい場合, 左に進み, ターミナル (Terminal Node2: 27 サンプルを中流域と判定) にたどり着く. COND が 4567 より大きい場合, 右に進み, ターミナル (Terminal Node3: 6 サンプルを下流域と判定) にたどり着く. すなわち, 多摩川の上, 中, 下流域の分類では, 最良分岐変数は COND であり, 競合変数としての Cl^- も, COND と同等の分岐変数で, Cl^- を用いる場合は, 判別条件は, 17.5 及び 1725 となった. また, その次に T_P が重要な変数であった. この二進木による分類の誤分類率は 0.00 であった. 但し, 元のデータに戻って, Fig. 2 の COND と Cl^- を観察してみると, 海水の影響が大きいのが下流域, 少ないのが上流域であり, したがって, Cl^- と COND が上, 中, 下流を区別する要素であることは明白である. しかし, 果たしてどちらの変数の影響が大きいのか, 等しいのか, また, どこに分割点があるのか, 更にこの二つ以外の変数の影響がどの程度な

のかについては, 必ずしも明確ではない.

次に, 上流の採水位置 6 か所について, その分類条件を求めた. Fig. 3 に変数重要度と得られた分類樹を示す. この分類では, 6 個のターミナルノードが形成されることが望ましいが, Fig. 3 に示すように, 5 個のターミナルノードが得られた. そのため, 上流 6 か所の分類樹では, 採水位置, 羽村堰 (c) がターミナルノード 1 ('98, '99 年度) と 4 ('97 年度) に誤分類されてしまうが, それ以外の 5 か所の採水位置, 和田橋 (a) はターミナルノード 1, 調布橋 (b) はターミナルノード 2, 永田橋 (d) はターミナルノード 4, 拝島原水補給点 (e) はターミナルノード 3, 拝島橋 (f) はターミナルノード 5 に分類され, 誤分類率は 0.17 であった (Table 1). この結果から, COND, COD, pH などが上流水域における採水位置 6 か所の分類における重要変数であることが示唆されている.

次に, 中流の採水位置 9 か所についてその分類条件を求めた. 中流の分類樹では, ターミナルノードが 4 個得られたが, この樹を用いた分類では誤分類率は 0.56 であり, 採水位置の個別の分類は全くできなかった. すなわち, 中流域における各採水位置の特徴を明確化することはできなかった.

次に, 97, 98, 99 年度の採水年ごとの水質分類について

Table 1 Water sampling positions with fiscal years of the upper area of Tama river and their classified terminal nodes in Fig. 3

Position name	Position code	Fiscal year	Terminal node	Fiscal year	Terminal node	Fiscal year	Terminal node
Wadabashi	a	1997	1	1998	1	1999	1
Tyohubashi	b	1997	2	1998	2	1999	2
Hamurazeki	c	1997	4	1998	1	1999	1
Nagatabashi	d	1997	4	1998	4	1999	4
Haijimagensuihokyuuten	e	1997	5	1998	5	1999	5
Haijimabashi	f	1997	3	1998	3	1999	3

Varibale Importance

Fig. 4 Classification tree on the water quality ('97, '98, '99 fiscal year) of Tama river

て検討した。この分類の分割条件が明確化されれば、各採水年の特徴を知ることができる。一般に、水質の変数は値が大きいほど汚染の程度は大きい。例えば、 T_N やBODは、大きいほど汚染の程度が進んでいる。したがって、分類樹では値の小さいほうを必ず左へ分類していくことから、左側のターミナルノードほど清浄であり、逆に右側のターミナルノードほど汚染が進んでいる傾向を持っている。Fig. 4の分類樹の誤分類率は0.06であった(51試料のうち、98, 99年度の拝島橋を97年度と誤判定した2試料と98年度の是政橋を99年度と誤判定した1試料の合計3ケースが誤りであった)。

この結果から、97年度の上流水域a~fはターミナルノード5に、また中、下流水域g~i, n, oはターミナルノード5, j~mは13に、また、98年度の上流水域a~dは

ターミナルノード2に、また中、下流水域h, j, k, m, oはターミナルノード10, l, n, pは11に集中し、99年度の上流水域a~dはターミナルノード1に、また中、下流水域h~pはターミナルノード12に集中していた(Table 2)。したがって、97~99年度の水質を区別する重要な指標はCONDと、窒素関連物質(T_N , NO_3_N , NO_2_N)にあり、上流水域については、これらの指標が年を経るごとに清浄化する傾向にあるが、中、下流水域では、97年度から98年度にかけては清浄化の傾向であったものの、99年度には逆戻りか、足踏みの状態にあること、したがって、中、下流域における窒素関連物質の低減が水質浄化のポイントであることなどが示唆されている⁷⁾。実際に個別の変数についてFig. 2より、生データの詳細を調べてみると、COND, NO_3_N , T_N では、97年度から

Table 2 Water sampling positions with fiscal years of the Tama river and their classified terminal nodes in Fig. 4

Position name	Position code	Fiscal year	Terminal node	Fiscal year	Terminal node	Fiscal year	Terminal node
Wadabashi	a	1997	5	1998	2	1999	1
Tyohubashi	b	1997	5	1998	2	1999	1
Hamurazeki	c	1997	5	1998	2	1999	1
Nagatabashi	d	1997	5	1998	2	1999	1
Haijimagensuihokyuuten	e	1997	5	1998	4	1999	3
Haijimabashi	f	1997	5	1998	5	1999	5
Hinobashi	g	1997	5	1998	7	1999	6
Sekidobashi	h	1997	5	1998	10	1999	12
Koremasabashi	i	1997	5	1998	12	1999	12
Tamagawagenbashi	j	1997	13	1998	10	1999	12
Tamasuidoubashi	k	1997	13	1998	10	1999	12
kinutasimotorisuiten	l	1997	13	1998	11	1999	12
Daisankeihintamagawabashi	m	1997	13	1998	10	1999	12
Tyohusyusuiten	n	1997	5	1998	11	1999	12
Denentyohuzekikami	o	1997	5	1998	10	1999	12
Rokugoubashi	p	1997	13	1998	11	1999	12
Daishibashi	q	1997	9	1998	7	1999	8

99 年度にかけて上流では浄化の傾向、中下流では汚染の程度が進行している。また、 $\text{NO}_2\text{-N}$ については、上流で汚染の程度が進行したものの大きく改善され、中下流では汚染の程度がいったん改善されたものの足踏み状態の傾向が伺える。

本報告の解析データ数は、51 試料でこの種の解析に用いられるデータとしては少数である。したがって、今回の解析結果からは、あくまでも示唆が得られたにすぎない。但し、たかだか 51 試料、12 パラメーターであっても、元の生データを観察することによって、各採水年の特徴的な変数とその分割条件、パラメーターの重要度を見いだすことは容易ではない。また、二進木は、各分類グループの特徴らしい特徴に着目する方法ではなく、各分類グループの境界線、分割条件、違いを明確化することにより、逆に各グループの特徴を浮彫りにする方法である。したがって、この点を念頭において解釈することが大切であろう。

4 結 言

多種多様な環境、分析データを総括する、あるいは、大量情報のデータ構造を把握したり、知識を得る方法の重要性は、ますます増大している。今回、河川の水質の採水位置や採水年の特徴を把握する目的で、分類樹を用いた分類条件の明確化により、データの階層構造や、それぞれの水質の特徴をとらえることができた。分類樹を用いることに

より、どの変数が幾ら以上か、以下なのかを見れば、目的変数の何がどの程度分かるのかといった詳細な議論が可能となる。大量の環境、分析データから、まとまった知見を得ようとする際、今後の分類樹の活用が期待される。

機械学習におけるニューラルネットワークを用いた神部、福田、長嶋、青山氏らによる、河川水質に関する研究論文⁶⁾が今回の検討の出発点になりました。また、日ごろから CART に関する助言、ご指導をいただいている明治大学大滝 厚教授に感謝します。更に今後の検討の方向付けについて貴重な示唆、助言をいただいた統計数理研究所シンポジウム「環境科学と統計科学の新たな融合」(2002 年 8 月 19・20 日)のオーガナイザー、広島大学大学院岡田光正教授、岩瀬見盛教授、統計数理研究所馬場康維教授に感謝します。

文 献

- 1) Brieman, L., J. H. Friedman, R. A. Olshen, C. J. Stone: "Classification and regression trees", (1984), (Wadsworth. Inc.).
- 2) J. R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA, 1993.
- 3) CART Ver.4.0: <http://www.salford-system.com>.
- 4) 大滝 厚, 堀江宥二, D. Steinberg: "CART による応用二進木解析法", (1998), (日科技連出版).
- 5) "公共用水域および地下水の水質測定結果", (1997-1999 年度), 東京都環境局.
- 6) 神部順子, 福田朋子, 長島雲兵, 青山智夫: *J. Chem. Software*, **8**, 27 (2002).
- 7) 東京都環境科学研究所: 研究所ニュース, No.21, (1999).