アンサンブル学習法による河川満足度調査データの評価

下川　敏雄[1]・武藤由香里[2]・御園生拓[3]・北村　眞一[4]

[1]会員　博士(工学)　山梨大学大学院医学工学総合研究部（〒400-8511　山梨県甲府市武田4-3-11 ,E-mail; shimokawa@yamanashi.ac.jp）
[2]学生会員　山梨大学大学院医学工学総合教育部持続社会形成専攻（〒400-8511　山梨県甲府市武田4-3-11 ,E-mail;g07mf014@yamanashi.ac.jp）
[3]非会員　Ph.D in Biology　山梨大学大学院医学工学総合研究部（〒400-8511　山梨県甲府市武田4-3-11 , E-mail; mist@yamanashi.ac.jp）
[4]正会員　工博　山梨大学大学院医学工学総合研究部（〒400-8511　山梨県甲府市武田4-3-11 ,E-mail;skita@yamanashi.ac.jp）

近年，河川開発と景観に関する多くの報告が行われている．その調査の多くがアンケートに基づいており，その結果は重回帰分析のような線形モデルによって処理される．ただし，真の構造が単純な線形関係で得られることは殆どなく，非線形構造，および交互作用を伴うことが多い．本論文では，これらの複雑な構造を適切に捉えるだけでなく，その結果をグラフィカルに解釈できるアンサンブル型学習法の適用について述べた．そして，その有用性はワード研究所(2004)によって実施された，河川満足度に関するアンケート調査への適用により提示した．その結果，アンサンブル型学習法を適用することで，重回帰分析，あるいは樹木構造接近法に比べて良好な予測確度をもつだけでなく，より有用な知見を見出すことに成功した．

キーワード: *多重加法型回帰樹木，　交互作用，　変数重要度，河川満足度調査*

１．序

近年，河川開発と景観に関する多くの報告が行われている．例えば，和田他（2003）による大正川に関する調査によると，河川の改善事業として望むことは，水質の改善、豊かな自然、親水空間の演出であることが報告されている．また，小路他（2005）は，緑，人工構造物，景観障害物が，景観評価に与える影響が大きいことが提示されている．さらに，市民満足学会・（株）ワード研究所(2006)では，河川満足度に対する調査をインターネットによって大規模に調査を行っている．そこでは，重相関分析により，河川満足度に与える影響を調査し，その結果，「自然，うるおい，町並み・田園との調和，統一感，水質，堰や橋，安全であることを報告している．

アンケート調査だけでなく，多くの予測問題で頻繁に用いられている重回帰分析(線形モデル)は，個々の質問項目(影響要因)がどれぐらい河川満足度(結果)にどのように，そしてどれほど寄与しているかを解釈できない．さらに，説明変数間の交互作用関係(相乗関係)は予めモデルに組み込まなければならないだけでなく，外れ値や多重共線性などの問題も含んでいる．

一般に，回帰分析の目標は，(1)応答の予測，および(2)応答と説明変数の関連性(推定モデルの解釈)の二つに分けて捉えられるが，重回帰分析は二つの目標を十分に満たしているとは考え難い．

近年，目標(2)に重点をおいた，重回帰分析の代替手法として，樹木構造接近法(あるいは決定木)が注目されている．その利点は，モデルの非線形性あるいは交互作用関係を解釈が平易なプロダクション・ルールによって与えることができることにある．その有用性は，多くのデータ・マイニングの成書で指摘されている．

ただし，樹木構造接近法による近似はステップ関数に基づくため，真のモデルが線形構造を持つ場合には不適切な結果を導くだけでなく，一般的に樹木構造接近法の予測確度は低いことが指摘されている．

重回帰分析および樹木構造接近法の予測確度の低さは，結果を誤った解釈に導く惧れがある．予測確度を向上させることは，アンケート調査の結果に内在する複雑な構造を適切に捉えるだけでなく，そこに新たな知見を見出すことに強く寄与すると考えられる．そのための一つの戦略は，目標(2)を満たす樹木構造接近法の予測確度を向上させる(目標(1)を満たす)ことである．本論文では，このような統計的学習法として，アンサンブル学習法を

とり上げる．アンサンブル学習法とは，複数個の樹木モデルの加法型でモデルを構築することで，樹木構造接近法の利点を保持しながら予測性能を向上することができる方法である．また，変数重要度や部分従属度といった統計量を用いることで，影響要因と結果の間の関係を明らかにすることもできる．これにより，これまでのアンケート調査の分析では得られなかった，非線形関係，あるいは要因の影響の大きさが高い予測性能のもとで解釈できる．

ここでは，ワード研究所の河川快適性に関するアンケート調査をもとに，影響要因の評価を行う．そこでは，アンサンブル型学習法のなかでも，とくに多重加法型回帰樹木(MART法, Friedman, 2001)を俎上にあげる．そして，重回帰分析，および樹木構造接近法との性能を比較するだけでなく，MART樹木によるデータ分析が，これまでの方法よりも，より多くの知見を与えることができることを示す．

２．樹木構造接近法

樹木構造接近法は，モデルへの適用結果が「樹木」によってグラフィカルに表現される．樹木表現は，複雑な非線形効果や交互作用効果に対する鋭い洞察を与えるだけでなく，さらに，変数の重要度を表現することもできる．Breiman, *et al.*(1984)によって提案された分類回帰樹木(CART:Classification And Regression Tree)は，樹木構造接近法を飛躍的に進歩させ，社会科学・工学・医学などの諸種の応用分野に広く関心を与えている(杉本他, 2005)．

ここでは，回帰モデルの構成に関心があることから，CART法のなかでも，とくに回帰樹木法の構成方法(因みに，分類樹木法は分岐基準が異なるだけで，アルゴリズムは同一である)について述べる．さらに，CART法により得られたモデルから，応答に寄与する要因(変数)を解釈するための変数重要度にも触れる．

(1)CART法

CART法は，データを説明変数空間に沿って2分岐させることで，モデルをあてはめる．このとき，分岐させた部分集合(ふし)内の応答の予測値には，平均値あるいは中央値が用いられる．したがって，CART法では，ステップ関数によってモデルが近似される．そのモデル推定の過程は，(1)分岐過程，(2)刈り込み過程，(3)最適樹木の選定過程，から構成される．

分岐の評価基準には，2分岐されたそれぞれの部分集合の応答に対する不均一性の測度が用いられる．
次いで，分岐過程により得られた「過剰適合」な樹木を，刈り込み基準(複雑度コスト)に基づいて，根幹ふし(分岐のない状態)まで，分岐点を逐次に削除する．この過程が刈り込み過程である．これにより，大きな樹木から根幹ふしまでの巣籠もり状の樹木系列が得られる．そのなかから，最適な部分樹木を選定するのに， CART法では応答の予測の観点からテスト標本法あるいは交差確認法が頻用される．

このような過程により得られたモデルは

$$h(\boldsymbol{x};\{t_m\}_{m=1}^{M}) = \sum_{m=1}^{M} \hat{\beta}_m \mathrm{I}(\boldsymbol{x}_n \in t_m) \qquad (1)$$

で与えられる．ここに，$\hat{\beta}_m = \bar{y}(t_m)$ である．

(2) 変数重要度

CART法の利点の一つは，ある応答に対して各説明変数がどの程度の重要度をもつかを統計量により明らかにできる点にある．ただし，得られたCART樹木のみに基づく解釈だけでは，分岐に用いられていない説明変数の効果は隠される．例えば，ある説明変数は最初の分岐で2番目に良い分岐ルールを持ったが，最終的には樹木分岐として採用されなかったとする．この場合に，その変数の重要度を0とすることが必ずしも適切であるとは限らない．そのため，変数重要度の算出方法には若干の工夫が行われている．すなわち，CART樹木の変数重要度は，得られたCART樹木の分岐変数とは異なる変数(代理変数)で分岐を行ったときの残差の減少量(改善度)を用い，この改善度を(分岐過程で得られた)樹木の全ての分岐点で計算する．そして，それぞれの変数の重要度は，これらの改善度の総和により定義される．

通常，変数重要度は，最大の変数重要度をもつ変数の値を100としたときの相対指標として解釈される．

３．アンサンブル型学習法

線形回帰法は，説明変数と応答の非線形構造を適切に捉えることができないだけでなく，説明変数間の交互作用効果を事前にモデルにとり込まなければならない．他方，CART法では非線形構造および交互作用効果を解釈が平易な樹木によって提示できるものの，その近似はステップ関数に基づくため，真のモデルが線形構造を持つ場合には不適切な結果を導く(Breiman, *et al.*, 1984)．近年，諸種のアンサンブル学習法，例えばBoosting法(Schapire, 1990)やBagging法(Breiman, 1996)が提案されている．

アンサンブル学習法とは，線形回帰モデルあるいは，CARTモデルといった学習器を反復してあてはめることで，より強力な予測力をもつモデルを生成する統計的学習法の一つである．

とくに，CART樹木をアンサンブル結合するBoosting 樹木法は，CARTの長所を保持しながら，その欠点の予測性能を向上することに成功している．これは，MART法

(Multiple Additive Regression Trees: 加法型重回帰樹木)と名づけられている(杉本・下川・後藤, 2005).

(1) MART法

MART法は，CART法の予測確度を向上させる目的で，CART樹木にBoostingを加味する方法として考案された．Boostingとは，関数空間における最適化アルゴリズムとして提案された方法である(Breiman, 1999)．また，ブースティング法の非常に興味深い側面は，データの背後にある潜在モデルに含まれる主効果項あるいは交互作用項のような構造的な推定問題にも魅力的な効用を発揮することである．このことは，アンケート調査のような複雑な構造を適切に捉えることに寄与すると考えられる．

観測値$\{y, \boldsymbol{x}\}$が与えられたとき，MART法によるブースティング法の目標は，(微分可能な)損失関数

$L(y, f(\boldsymbol{x}))$の期待値を最小にするモデル

$$f^*(\boldsymbol{x}) = \arg\min_{f(\boldsymbol{x})} E[L(y, f(\boldsymbol{x}))] \qquad (2)$$

を推定することである．本論文では，線形回帰法あるいはCART法と同様に，損失関数には，最小2乗基準損失関数

$$L(y, f(\boldsymbol{x})) = (y - f(\boldsymbol{x}))^2 \qquad (3)$$

を用いる．

B個のCART樹木$h(\boldsymbol{x}; \{t_m^{(b)}\}_{m=1}^{M_b})$より得られる，MART法の推定モデル$\hat{f}_{\text{boost}}^{(B)}(\boldsymbol{x})$は，加法的な展開のもとで逐次的に式(2)を近似することで，

$$\begin{aligned}\hat{f}_{\text{boost}}^{(B)}(\boldsymbol{x}) &= \sum_{b=1}^{B} \nu \hat{\omega}_b h(\boldsymbol{x}; \{t_m^{(b)}\}_{m=1}^{M_b}) \\ &= \sum_{b=1}^{B} \nu \sum_{m=1}^{M_b} \hat{\omega}_b \hat{\beta}_m^{(b)} \mathrm{I}(\boldsymbol{x}_n \in t_m^{(b)})\end{aligned}$$

により得られる．ここに，$\hat{\omega}_b$は，b番目のCART樹木での重みパラメータの推定値であり，$\nu(0 < \nu \leq 1)$は，学習効率を制御するための任意の縮小パラメータである．

このとき，重みパラメータおよび樹木の推定を同時に行うことは困難なため，MART法でのブースティングでは，重みパラメータω_bを推定する過程と，樹木$h(\boldsymbol{x}; \{t_m^{(b)}\}_{m=1}^{M_b})$を推定する過程を交互に繰り返す反復アルゴリズム(ステイジワイズ戦略)を採用している．

通常，回帰モデルの推定において，「ステップワイズ」戦略が頻用されている．ステップワイズ戦略では，変数(あるいは基底関数)が追加される毎に，モデルに含まれる全てのパラメータが調整されるのに対して，ステイジワイズ戦略では，パラメータの調整は必要としない．さらに，ステイジワイズ過程では，これらの推定を最急降下法に類似する更新手続きとして捉え，損失関数(4.2)に対してCART樹木を当てはめるのではなく，その偏微分したもの(これを疑似残差と呼ぶ)に対して当てはめる．

(2) 変数重要度の拡張

MARTモデルが，複数のCART樹木の加法形であることから，変数重要度は，CART樹木の場合と同様に定義できる．ただし，CART樹木における代理変数による変数重要度の定義は採用しない．もともと，CART法で代理変数を用いた理由は，単一の樹木の分岐でマスキングされた変数の影響をできる限り公平に要約することにある．これに対して，MART法では多数の樹木を扱うが，このことは単一の樹木におけるマスキング変数の影響を考慮することにも繋がる．したがって，MART法での変数重要度は，推定された複数のCART樹木での改善度(それぞれのCART樹木の構成に出現しなかった変数の改善度は0である)の算術平均値によって定義される．

4．河川満足度調査に対するアンサンブル型学習の適用

河川景観の満足度の現況を調査する目的で，平成17年6月30日から7月24日にかけて，㈱ワード研究所によってアンケートが実施されている．このアンケートは，全国を対象にしたインターネット調査の方法で行われており，12,189名の回答が得られた．

本アンケートの内容のなかから，ここでは，河川満足度に影響をあたえる項目(説明変数)として，

- 水量が富に流れている(水量).
- 自然がしっかり保全され，残っている(自然度).
- 水質がしっかり保全されている(水質).
- ゴミがなくてきれい(ゴミ).
- 時の変化(季節・気象・朝夕)がすばらしい(四季).
- 水辺の遊びなど，人々の活動がたのしい(親水性).
- 安全で安心する(安全性).
- 鳥や魚や蛍など生物に親しめる(生態系).
- 堰や橋などが美しい(橋・堰).
- コンクリートが少なくてよい(コンクリート).

をとりあげ，全体の満足度を表す項目(応答変数)として，「全体としての雰囲気が素晴らしい」をとり上げる．その他の3項目「水辺の町並みや田園と調和して美しい」

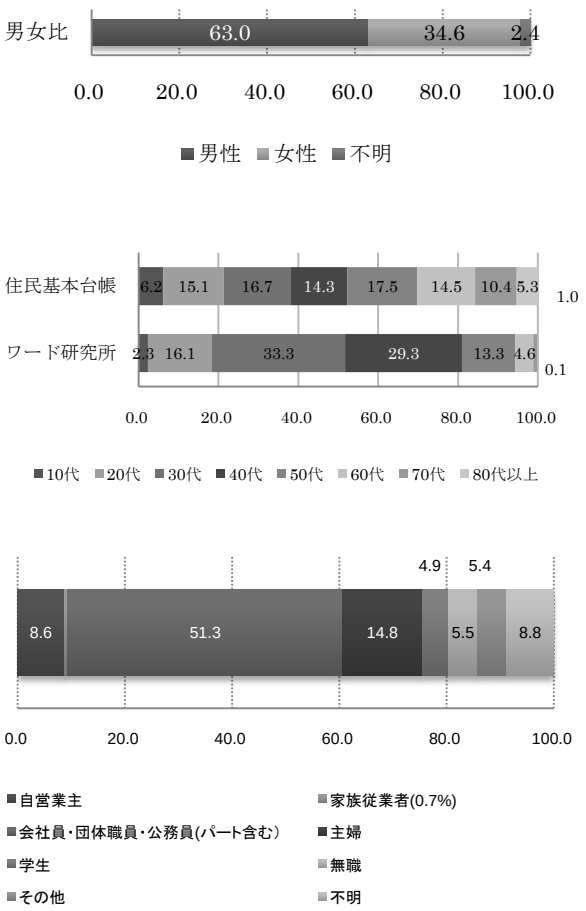

図-1 河川満足度アンケートの要約

「日々の生活の中で潤いを与えてくれる」「整然として統一感がある」に関しては，応答変数と同様の意味であり，またその内容が曖昧であることから除外した．

回答者の背景を図1に提示する．男女比は，男性のほうが女性に比して約30%程度，比率が高かった (n=12189)．年代では，30代から40代にかけての比率が住民基本台帳に比して高く，他方，60代以上の高齢者の比率が低かった(n=11509)．さらに職業別では，会社員・団体職員・公務員が半数を占めており，また，主婦の比率も14.8%であり，2番目に高かった (n=12189)．これは，本アンケートがインターネットで実施されていることから，インターネット利用率の比較的多い層が，そのままアンケートの回答者背景に反映されていると推察できる．

(1) 重回帰分析の適用

既存のアンケート結果の分析に倣い，重回帰分析を実施した．このときの回帰係数の一覧を図2に示す．その結果，自然度の回帰係数が最も高く，次いで橋・堰が高かった．他方，水量，水質，生態系といった項目の回帰係数が低かった．因に，全ての回帰係数に対するt検定のp値は，有意水準0.0001のもとで有意であり，係数の適切性が示唆されている．さらに，自由度調整済み寄与率

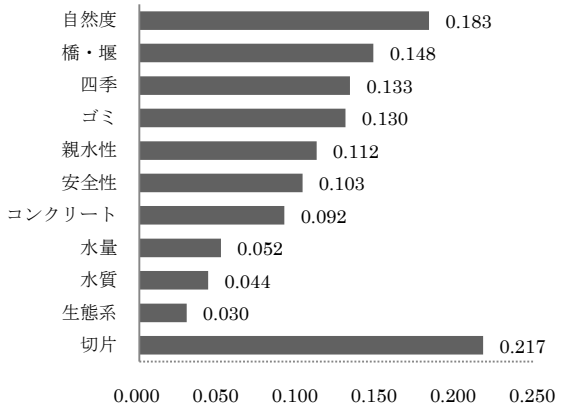

図-2 重回帰分析の回帰係数

$R^{2*}=0.580$ であることから，比較的良くあてはまっている．ただし，橋・堰といった項目が河川満足度に2番に強く寄与していることは考えにくい．さらに，自然度や四季といった景観のポジティブな要因が，河川満足度に強く寄与している一方で，ゴミや水質といったネガティブな要因の効果がそれほど高くなかった．そこで，交互作用効果あるいは非線形効果を探索するために，次項では，CART法を適用することで，重回帰モデル(線形モデル)との結果の違いを省察した．

(2) CART法の適用

図3にCART法での結果を示す．ここで，灰色の四角の上側の数字がふし内の応答の平均値を表しており，下側の数字がふしに含まれる標本サイズを表している．全観測値は，先ず，自然度によって分割され，その点数が2以下の観測値は次いで橋・堰によって分割され，それ以外の観測値はゴミによって分割され，最終的に9個の終結ふしに分割された．このとき，分割変数として自然度，橋・堰，水質，ゴミ，四季，および親水性が選択された．他方，生態系や水量，安全性，そしてコンクリー

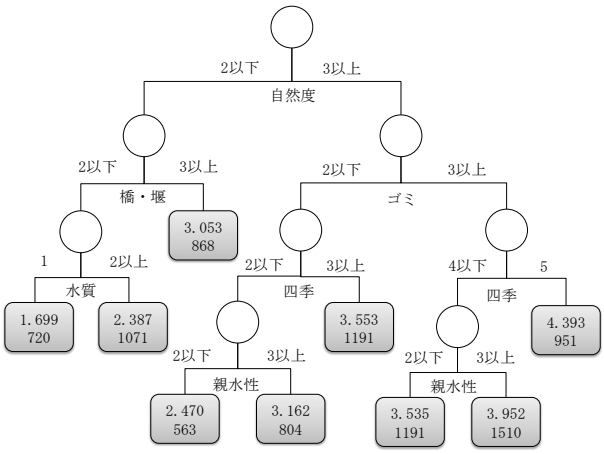

図-3 CART 樹木の結果

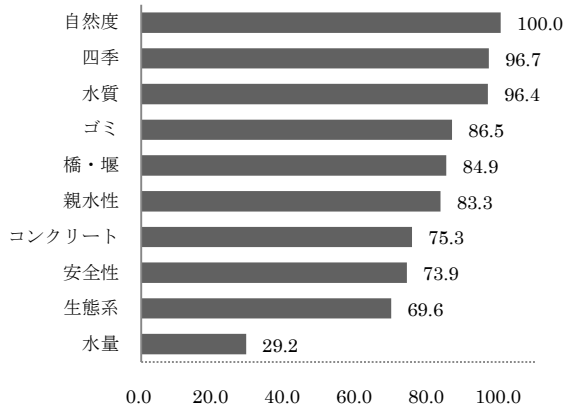

図-4 CART 法の変数重要度

トといった項目は出現しなかった．したがって，重回帰分析でも最も重要な要因として示唆された，自然度による影響が最も顕著であることが示唆された．また，重回帰係数の低かった水質によって分岐されるふしも存在した．このときの変数重要度を図4に示す．自然度での重要度が最も高く，次いで四季，水質，ゴミの順で高かった．すなわち，河川に対してポジティブな質問項目に対する重要度が最も高かった．次いで，ネガティブな質問項目が続いた．他方，水量や生物など，川辺からは視覚的に捉えにくい要因に対する重要度は高くなかった．多くの傾向が重回帰解析での結果に類似したものの，水質の重要度が顕著に高かった．これは，水質が主効果として影響を及ぼすのではなく，自然度や橋・堰との交互作用としての効果が顕著であることが推察できる．このときの寄与率を残差平方和の10重クロスバリデーション推定値に基づいて推定した．その結果，$R^2_{cv} = 0.479$ であった．本データでは，線形構造の傾向が強く，これにより，ステップ関数で近似するCART樹木でのあてはまりが悪かったと考えられる．また，CART樹木は，重回帰分析の適合性能を劇的に上昇させることはないことは広く知られており(線形構造が顕著な場合には，むしろ悪化する)，本データではそれを裏付ける結果を示した(杉本ほか,2005).

(3) MART法の適用

MART法では，損失関数(4.2)として，最小2乗損失，最小絶対損失，Huberの損失関数などが用いられるが，ここでは，他の方法との比較のため最小2乗損失を用いた．MART法は，ステイジワイズ戦略のなかで，複数の樹木を構成し(ブースティングさせていき)，その加法形によってモデルを推定する．図5にブースティング回数と最小2乗損失のプロット示す．ここでは，損失関数の交差確認推定量と学習標本での損失関数の推定値を示す．学

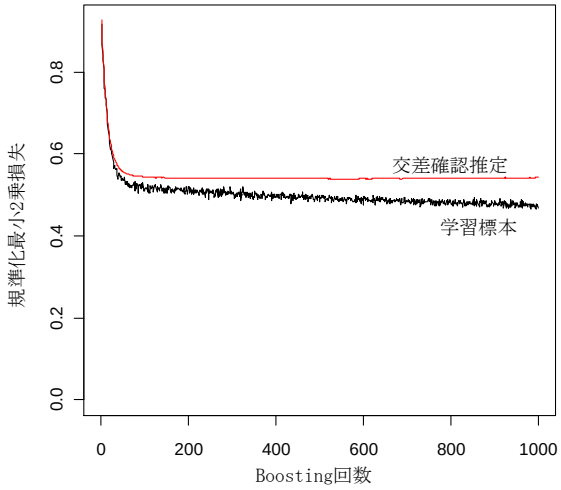

図-5 ブースティング回数と規準化２乗損失のプロット

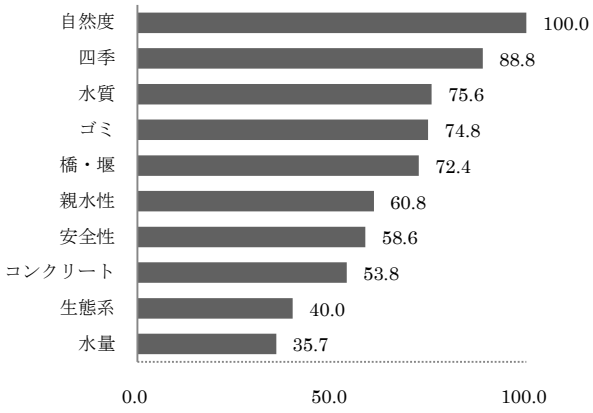

図-6 MART 法の変数重要度

習標本では，ブースティング回数を増加させるほど損失が減少する傾向にある．これに対して交差確認推定量は，あるブースティング回数からは飽和傾向を示した．交差確認推定量の結果，本データにおける差異的ブースティング回数は549回であることが示唆された．このときの変数重要度のプロットを図6に示す．CART樹木と同様に自然度の重要度が最も高く，次いで四季，水質，ゴミの順で変数重要度が高かった．すなわち，ポジティブ・イメージの要因の影響が最も高く，ネガティブ・イメージの要因が続く傾向には変化がなかった．これは，MART樹木のステイジワイズ戦略では，最初の樹木(1回目のブースティング樹木)の変数重要度に対する影響が最も強いためであると推察される．しかしながら，CART法の変数重要度は，水量以外で大きな差異が認められなかったものの，MART法では，自然度，および四季の変数重要度が他に比して大きな値を示した．このときの寄与率をCART法と同様の流儀で計算した．その結

果，$R_{cv}^2 = 0.708$ であり，最も高い値を示した．すなわち，寄与率の最も低かったCART樹木をアンサンブルさせることで，大幅に寄与率を上昇させた．

５．結び

本研究では，河川の満足度に影響を与える要因をアンケート調査の結果に基づいて探索した．アンケート調査の分析には，通常，重回帰分析が用いらるが，これらの要因が単純な線形結合によって結びついていることはなく，要因間の交互作用構造，あるいは応答と要因のあいだの非線形構造を含むことは少なくない．ただし，これらの構造を捉えることは困難である．

本論文では，予測確度と結果の解釈の両方を満たすことができる，重加法型回帰樹木法をアンケート調査の分析に応用し，その有用性を既存の回帰分析手法と比較した．その結果，河川満足度にポジティブな要因(自然度，四季)のほうがネガティブな要因(ゴミ，水質)よりも影響度が高かった．

謝辞：本研究の資料調査において(株)ワード研究所の大島章嘉氏には多大なご協力を頂いた．厚く謝意を表する．

参考文献

1) 杉本知之，下川敏雄，後藤昌司，樹木構造接近法と最近の発展，計算機統計学, No.18, pp.123-164, 2005.
2) Hastie, T., Tibshirani, R., Friedman. J.H. *The Elements of Statistical Learning: Data mining, inference and prediction*. Springer, NewYork, 2001.
3) Friedman, J.H. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, Vol.29, pp.1189-1232, 2001.
4) Breiman, L., Friedman, J.H. Olshen, R.A., Stone. C.J. *Classification and Regression Trees*. CRC Press, Florida ,1984.
5) 市民満足学会・（株）ワード研究所，河川景観満足度調査中間報告書，2006