〔共同研究：水インフラ整備の課題と展望〕

多変量外れ値に対する パラメトリックな生産フロンティアの感度分析*

――STATA を用いた多変量外れ値検出の水道事業分析への適用――

矢　根　真　二

要約

本稿の目的は，多変量外れ値（Multivariate Outliers）の除去に対する SFA（確率的フロンティア分析）の１ステップモデルによる生産フロンティアと技術効率性の推定値の感度（Sensitivity）を分析する点にある。具体的には，Wang（2002）のモデルを日本の上水道事業に適用した矢根・矢根（2018）が用いた12変数に，STATA で利用可能な３種のコマンド **hadimvo**・**bacon**・**mcd** を適用する。その主要な結論は，（1）デフォルトでの外れ値の検出数は **bacon**・**hadimvo**・**mcd** の順に次第に多くなるが，**bacon** の外れ値はいずれも **hadimvo** の外れ値であり **hadimvo** の外れ値はすべて **mcd** の外れ値になるという意味で整合的である，（2）1243事業者のうち明らかな異常値を示す２事業者を取り除くと，技術非効率性に対して有意でなかった環境変数である顧客密度 cusden が0.1％水準で有意になる，（3）この環境変数の影響を含めた生産フロンティアと技術効率性の **bacon** および **hadimvo** による外れ値に対する感度分析の結果はほぼ頑健である，（4）ただしサンプル数の３割弱を外れ値として検出する **mcd** は，非効率性を示す片側誤差項の分散を過度に収縮させるため，このフロンティア分析にはデフォルト値での適用はできないことの４点である。

目次

*本稿は，2014-16年度桃山学院大学総合研究所地域連携プロジェクト番号238「水インフラ整備の課題と展望」による研究成果の一部であり，桃山学院大学およびプロジェクト・メンバーに，ここに記して感謝します。

キーワード：多変量外れ値（Multivariate Outliers），感度分析（Sensitivity Analysis），SFA（Stochastic Frontier Analysis：確率的フロンティア分析），１ステップモデル（1-Step Model），頑健性（Robustness）

１　外れ値検出と生産フロンティア分析

パラメトリックな生産フロンティア分析の嚆矢は，Aigner et al.（1977）およびMeeusen and van den Broeck（1977）である。それ以来約40年，**SFA**（**Stochastic Frontier Analysis**：確率的フロンティア分析）におけるクロスセクション分析の主要課題の１つは，技術効率性に及ぼす環境変数の影響を，生産フロンティアの推定と同時に１ステップで的確に把握することである[1)]。そのために多様な１ステップモデルが開発・提案されてきたことは，Wang（2002）やKumbhakar and Wang（2015）による**KGMHLBC**型や**CFCFGHI**型といった１ステップモデルの分類や命名から伺える。

矢根・矢根（2018）は，国際的にも散度の高い日本の水道事業では，**KGMHLBC**型や**CFCFGHI**型ではなく，双方の要素を取り入れたWang（2002）のモデルが尤度比検定の結果として選択されることを示した。すなわち，Kumbhakar and Lovell（2000, p. 122）およびKumbhakar and Wang（2015, p. 35）の予測どおり，影響力の大きな環境変数を無視したSFAモデルでは，生産フロンティアと技術効率性の双方の推定値にバイアスが生じることを例証したのである。

しかし散度の高いデータにもかかわらず，1243事業者から成る生産フロンティアの推定にも，８つの環境変数で説明される個々の事業者の非効率性の推定にも，**外れ値**（**outlier**）への配慮は一切なされていない。Greene（2017, p. 105）によれば外れ値とは，当該モデルの適用外にあることが明白な観測値であり，その主因はおそらく異なったデータ生成過程に由来する。換言すれば，矢根・矢根（2018）による上記の結論は，総計10000を超える全変数の観測値の中に，誤植をも含めて分析結果を左右するような外れ値は存在しないという「**暗黙の想定**」に立脚しているのである。

そこで本稿の目的は，この暗黙の想定がどの程度妥当なのか，すなわち外れ値に対する環境変数を含めたSFAの１ステップモデルの**感度分析**（**Sensibility Analysis**）を行うことである。その特色は，矢根・矢根（2018）で用いられた環境変数を含む12変数をベースに，日本の水道事業の効率性分析では初めてパラメトリックな**多変量外れ値**（**Multivariate Outliers**）の検出を試みる点にある。多変量外れ値の手法を試みるのは，**単変量外れ値**（**Univariate Outliers**）手法では検出できない外れ値が存在するからである。しかも，フロンティア分析のような生産経済学や産業組織論といった応用経済学分野の対象となるデータの多くは高い散度を示すのが常であるのに，多変量外れ値検出の適用例は未だ少ないように思われるからである。

ただし，応用分野の実証研究で多変量外れ値の検出がさほど試みられていないという認識

1）ノンパラメトリックな２ステップ法を「根拠薄弱（invalid）」と批判したSimar and Wilson（2007, 注8）は，もちろんパラメトリックな２ステップ法の統計的な問題も指摘している。このバイアスの程度は，Wang and Schumidt（2002, p. 130）によれば，第１ステップで驚くほど下方に拡散するので，第２ステップでもバイアスが生じるため，２ステップ法は控えるべきだという結論に至るほどである。

が正しければ，それ相応の理由があるはずである。もちろん筆者のようなデータ化される現場の実情やその処理に関する統計学知識が不十分だという個人的責任は免れえないとしても，より一般的で広範な要因があるように思われる[2)]。

第１に，近年も進展中であるコンピューターの計算処理能力やソフトウェアの使い勝手の向上に，大学の教育・研究体制が取り残されてしまったのではないだろうか。実際，半世紀近く前にはOLS（最小二乗法）を機械に任せられる研究者は希少であったかもしれないが，今日では応用分野でも多用される基本モデルなら多くの学部生でも容易に実行できる環境が整えられている。一昔前にはOLSでさえ大型計算機がない環境では容易な作業ではなかったのに，今ではフロンティア分析のような基本モデルや多変量外れ値検出のような負荷の高かった作業もOLSと変わらぬ容易な作業になっているのだから，その使用を妨げているボトルネックは環境変化に対する教育・研究体制の対応の遅れが主因としか考えられない。実際，AI・IoT・RPAにおける技術進歩が既存の職種を半減させると話題になっているにもかかわらず，大学全体や個別学部の教育・研究体制にはこうした科学的進歩の恩恵・対策を先取るような大胆な変革はみられない。

もちろん，その責めの一部は，上述したような自ら分析対象と手法の熟知を怠る応用研究者や学生個人が負うべきだろうが，計量経済学のテキストでも，たとえば外れ値の検出や対応に多くの頁や時間を割く余裕はないのが現実である[3)]，ゆえに，たとえ大学で勤勉に学んだ学生でも，実際に研究されている手法とのギャップは拡大するばかりというのが第２の要因である。これはハードと，とりわけソフトの進展に対する統計学・計量経済学の対応の遅れともみなせようが，それだけ経済学における当該分野の理論と実践の重要性が急速に高まっているのではないだろうか。事実，多変量外れ値の検出でも，様々な基本モデルの応用の場合と同様，STATA・R・MATLABといったソフトウェアのモジュール・パッケージ・キットといった形の追加更新さえすれば，基本的な作業量はOLSの推定とほぼ変わらない時代である。学生を含めた多くの応用経済学者が今後も分析対象の拡大と使い勝手の向上を続けるソフトウェアに依存することを前提にすれば，実証の分野の理論と実践にもっと多くの時間を割くことが，経済学教育のみならず応用経済学の発展にも不可欠であると思われる。

第３に外れ値に焦点を戻すと，外れ値の検出とその対応はあまりに重要すぎて判断しかねるがゆえ，試行し難いのかもしれない。検出した外れ値の存在が分析結果に影響しないなら検出作業はさして重要でない一方，結果を左右する場合には外れ値を処理する十分な理論的な説明が必要になろう。たとえば，外れ値の存在がフロンティアの形状に直接影響し，それ

2）Hadi（1992）やBillor et al.（2000）のような多変量外れ値検出法の提案者が異口同音に必ず指摘する理由として，計算費用・速度の問題がある。しかし以下では，数十万の観測値を抱える分析に限らず，むしろ今日ではさほど計算時間を要しない経済分析一般について焦点を当てる。

3）たとえばGreene（2017, pp. 104-7）では，1000頁を超えるテキスト中の約３頁を割き，回帰分析へのスチューデント化残差の機械的な適用法とその問題点を説明している。しかし少なくとも日本では，この分量の知識だけを教えるのも困難だろうから，実際にソフトウェアを使った実践的な教育と組み合わせるには課題をさらに大幅に絞るしかないはずである。

ゆえ個々の効率性水準の評価をも歪めるノンパラメトリックなフロンティア分析では，外れ値の検知や除去が **DEA**（Data Envelopment Analysis：包絡分析法）の施行前の不可欠な吟味だと強調されるのが常である。実際，R や MATLAB で作動するノンパラメトリックなフロンティア分析の外れ値の検出法は，それぞれ Wilson（1993）や Simar（2003）によって公開されている。しかし Simar（2003, pp. 402-3）が強調するように，閾値の設定等に定まった規則などはなく，外れ値を自動で検出することなど決して容易ではない。むしろ，Simar（2003, pp. 404）が指摘するように，外れ値の数の上限に何の理論的な規則もないならば，外れ値の除去が結果を左右するとしても，その外れ値の閾値や個数の設定を正当化するのは容易ではない。

そのうえ第４に，STATA を用いてモンテカルロ法による環境変数の効果に関するシミュレーションを行った Wang and Schumidt（2002, p. 135）が指摘するように，尤度を最大にする数値計算には外れ値に伴う技術的な問題が起こりうる。特に，切断正規分布するパラメーターの推定から安定した平均値を得るため，上下0.3％の極端な推定値を外れ値として除外している。なぜ0.3％なのかという上記の閾値の説明はさておいたとしても，データの生成過程が既知であるシミュレーションにおいても外れ値を無視できないならば，最尤法が多用される今日ではもっと外れ値の検出が普及してもよいはずである。しかし，日常的にも少なくないと思われる大局的に凹であることが疑わしい最大化問題に対して，多変量変数外れ値の検出を試みるよりも，むしろ変数を減らしたり計算アルゴリズムを変更したりする対応の方が多いように思われる。そこで本稿では，変数や計算アルゴリズムを固定したうえで，外れ値に対する SFA の１ステップモデルの感度を分析する。

本稿の構成は，以下のとおりである。第２節では，本稿で使用する STATA で利用可能な３種の多変量外れ値検出コマンド **hadimvo**・**bacon**・**mcd** について説明する。これらの検出法を矢根・矢根（2018）の SFA の１ステップモデルに用いられた12変数に適用するのが，第３節である。そこでは，デフォルトによる外れ値の検出数は大きく異なるが，検出数の少なかった検出法の外れ値は必ず検出数の多かった検出法の外れ値であるという包含関係が成立するという意味で整合的であることが示される。第４節では，**bacon** および **hadimvo** による外れ値検出に対する矢根・矢根（2018）の１ステップ SFA モデルにおける生産フロンティアと技術効率性の感度を分析し，双方ともほぼ頑健であることを示す。それどころか，明らかな２個の異常値を取り除くだけで，環境変数である顧客密度 **cusden** の効率性に及ぼす影響が0.1％水準で有意になることが確認される。以上の分析結果は第５節で要約され，Appendix では除外した外れ値の属性が要約される。

2　多変量外れ値の検出法と使用変数

外れ値は，文字どおり，ある基準から測って非常に離れた距離にある観測値だから，その基準と距離の取り方によって様々な判定をなしうる。たとえば和田（2010, 表１）は，単変

量か多変量か，ロバストかロバストでないかという視点から 4 種の手法に分類している[4)]。標本平均値や標準偏差がロバストでない理由は，平均や標準偏差の算出自体に含まれる外れ値の影響が大きくなる場合があるからである。他方，ボックス・プロット（箱ヒゲ図）が単量外れ値検出法としてロバストな理由は，中央値や四分位差が標本平均値や標準偏差より外れ値の影響を受けにくいからである[5)]。

しかし，単変量外れ値検出を繰り返したとしても，極端な値を取る外れ値を検出できても，他の変数との関係における外れ値は検出できない点に注意すべきである。すなわち，当該変数だけを見た場合には極端な値ではなくとも，他の変数との関係において観測値の大半とは異なる傾向を持つ外れ値を検出できないのである。

そこで多変量データの距離指標として伝統的に用いられてきたのが，マハラノビス（平方）距離である。m 個の変数を持つサンプル数 n の n 行 m 列の行列 X の第 i 行ベクトルを x_i で表すと，マハラノビス距離 D_i は次のように定義できる。

$$D_i=\sqrt{(x_i-x)\Sigma^{-1}(x_i-x)'} \qquad (1)$$

ここで，x は平均値の行ベクトル，Σ は m 行 m 列の分散共分散行列である。ゆえに，諸変数が独立であれば Σ は対角行列となり，（正規化）ユークリッド距離になる。換言すれば，マハラノビス距離は，変数間の相関も考慮した方向によって調整された尺度なのである。さらに X が多変量正規分布からのランダムなサンプルであれば，マハラノビス平方距離の検定統計量は自由度 m および $n-m$ の F 分布に従うことが知られている。

しかし，単変量の標本平均値や標準偏差と同じように，（1）式の平均や共分散が外れ値を含む標本全体から算出される限り，マハラノビス距離 D_i もロバストではない。Hadi（1992, p. 762）も指摘するように，マスキング問題やスワンピング問題を回避できないからである。前者は，ある外れ値の存在が，平均値や距離を歪めることにより，他の外れ値の検出をマスクしてしまう効果である。後者は，多数のパターンに従う観測値であっても，その反対方向に平均や距離を歪める小群の外れ値があれば，距離が大きくなってしまうために外れ値とみなされかねない問題である。こうした問題を避けるには，外れ値に対してロバストな基準を定める必要がある。

そこで本稿では，STATA で利用可能な 3 種のロバストな多変量外れ値検出法を使用する。それぞれコマンド名は，**hadimvo**・**bacon**・**mcd** である。

hadimvo は，Hadi（1992, pp. 762-3）によって **MVE**（Minimum Volume Ellipsoid）の近似手続きとして提案されたアルゴリズムである。基本的なアイデアは，最初に外れ値を含みそうにない少数の基本サブグループとそうでないサブグループに分け，基本サブグループに含

4）さらに Ben-Gal（2005, p. 2）は，パラメトリックな手法とノンパラメトリックな手法に分類している。こうした分類に従うなら，本稿の焦点はロバストな多変量外れ値のパラメトリックな検出法にある。

5）和田（2010, p. 94）は，ロバストでない多変量外れ値検出法として後述するマハラノビス平方距離を，ロバストな多変量外れ値検出法として MSD（Modified Stahel-Donoho）法を挙げている。

まれる観測値を少しずつ増加させ，一定の基準に達した時点で停止し，その終了時点で基本サブグループに含まれなかった観測値を外れ値とみなす点にある。

STATA での **hadimvo** の外れ値検出数に影響するオプションは外れ値検出に関する有意性水準のみであり[6]，デフォルトは0.05である。フロンティア分析への適用例としては，Wang（2002, p. 246）があり，明らかに異常値と思われる２つの外れ値を検出している。

bacon の特徴は，Billor et al.（2000, p. 284）によれば，Hadi（1992）のアイデアをベースにしながらも，計算速度を引き上げるために，個々の観測値を吟味して基本サブグループを拡大するのではなく，ブロックに束ねた観測値を吟味するよう改善したアルゴリズムにある。この名称は，まさにそのアルゴリズム名 blocked adaptive computationally efficient outlier nominators の略称である。それゆえ，検出結果は上記の **hadimvo** や **MVE**，後述の **mcd** と整合的であると報告されている。

実際，Weber（2010, p. 334）は，STATA における **hadimvo** と **bacon** をデフォルト値で使った検出結果がほぼ同様であることを確認すると同時に，**hadimvo** の計算速度の遅さを強調している。**bacon** の外れ値検出数に影響するオプションも外れ値を決める閾値だけであるが，カイ二乗分布のパーセンタイルを与える形になっており，デフォルトは0.15である[7]。Weber（2010）によるデフォルト値を使った検出例では，**hadimvo** も **bacon** も74個中２個（2.7%），そして28101個中それぞれ20個（0.07%）と29個（0.1%）であり，整合的な結果を得る速度の違いが強調されている。

mcd は，ロバストな統計量として知られる minimum covariance determinant の略であり，Verardi and Dehon（2010）によって STATA で使える速い検出法のアルゴリズムとして提案されたものである。mcd の基本的なイデアは，最小の一般化分散，つまり最小の分散の行列式を有するグループが標本の半数を占めるようにして，ロバストな平均値と共分散を求めることにある。Verardi and Dehon（2010, p. 260-1）によれば，実際には観測値の半数から成る膨大な組み合わせをすべて計算する必要がない点にアルゴリズムの特徴があり，**hadimvo** の最初の基本サブグループの選択よりロバストである点が利点だという。

事実，Verardi and Dehon（2010）による STATA における **hadimvo** と **mcd** の比較でも，速度はもちろん外れ値検出力も **mcd** に軍配を上げている。ただし比較の対象は，１つのシミュレーション・データに関してのみである。**mcd** のオプションは，外れ値の最大期待パーセント等を含めて７個あり，非常に柔軟である。

以上の３種のパラメトリックな多変量外れ値検出コマンドの対象とする変数は，矢根・矢根（2018）の SFA の１ステップモデルの推定に使用された12変数である。半切断-正規モデ

6）外れ値なら１を取るダミー変数と算定された距離を示す変数を作成する機能もあるが，これらは他の２つのコマンドも共有する機能であり，検出数とは無関係である。

7）Billor et al.（2000）の中央値を使うバージョンも可能であり，最初の基本サブグループの数の設定も可能であるが，本稿では使用しない。さらに Weber（2010）は，何度も計算し直せる replace 機能が便利だと強調している。

ルと呼ばれる基本的なSFAモデルは，次のように定義できる。

$$\ln(Y_i)=\beta_0+\beta_1\ln(K_i)+\beta_2\ln(L_i)+\beta_3\ln(O_i)+v_i-u_i \quad (2)$$

$$v_i \sim iid \quad N(0,\ \sigma_v^2) \quad (3)$$

$$u_i \sim iid \quad N^+(\mu,\ \sigma_u^2) \quad (4)$$

すなわち，(4)式の非効率性を表す半切断の片側誤差項がなければ，(2)式は生産量Yの対数値を資本K・労働L・その他の生産要素Oの対数値での回帰にすぎない。実際，(4)式の平均値も分散も0なら，SFAは成り立たない。換言すれば，Coelli (1995) による非効率性の要素がないという帰無仮説を棄却できてはじめて，(2)式は意味を持ち，非効率性を推定できるのである。この非効率性は，Jondrow et al. (1982) による次式に従い効率性に変換されるのが常である。

$$\mathrm{E}_i=\exp(-E(u_i|\varepsilon_i)),\quad where\ \varepsilon_i=v_i-u_i \quad (5)$$

ゆえに，コブダグラス型生産関数の推定に使用する変数は産出量Yと生産要素$K,\ L,\ O$の4変数である。これらの変数を含め，矢根・矢根（2018）で用いられた12変数のすべては2007年度『地方公営企業年鑑』から作成されており，定義は以下のとおりである。

Y（産出量）：年間総有収水量（千m^3）

K（資本）：減価償却累計額を含む（建設仮勘定を除く）有形固定資産額（千円）

L（労働）：損益勘定所属職員数と資本勘定所属職員数の合計人数

O（その他）：動力費，光熱水費，通信運搬費，修繕費，材料費，薬品費，路面復旧費，委託費，資本相当分を除いた受水費の合計金額（千円）

いわゆる2ステップ法とはこの第1ステップで得た効率性を環境変数で回帰するアプローチであるのに対し，矢根・矢根（2018）が用いた1ステップモデルは(4)式の平均と分散を(6)式のように不均質性や不均一分散を許容できる形に拡張し，(7)-(8)式のように環境変数で説明する方法である。定数項を含めた環境変数ベクトルをZiとすれば，1ステップモデルでは(4)式の平均と分散は次のように書き直せる。

$$u_i \sim iid \quad N^+(\mu_i,\ \sigma_{ui}^2) \quad (6)$$

$$\mu_i=Z_i'\delta \quad (7)$$

$$\sigma_{ui}^2=\exp(Z_i'\gamma) \quad (8)$$

次に，矢根・矢根（2018）で用いられた8個の環境変数の定義は，以下のとおりである。

lwc：取水規模を表す指標としての取水能力の対数値

cusden：給水人口を水道管総延長距離で除した**顧客**（給水人口）**密度**

Load：1日平均給水量を1日最大給水量で除した**負荷率**

dtypews2：取水能力に占める**受水比率**が50％ を超えれば1をとるダミー変数

dtypewe2：自己水源に占める**地下水比率**が90％を超えれば1をとるダミー変数

rraw：**低水質**を示す指標としての取水能力当たりの薬品使用額の比率

rsubp：**補助金比率**としての給水収益に占める他会計繰入金等の損益勘定５項目の総額の比率

Uprice：供給単価としての給水収益を有収水量で除した**料金水準**

２ステップモデルでは，最初にフロンティア推定のために４変数を用い，次にその効率性の推定値をこれらの８個の環境変数で回帰する。他方，１ステップモデルは，一度に12変数すべてを使うことになる。次節では，この点を念頭に置いたうえで，３種の多変量外れ値検出法を適用する。

3　SFA の１ステップモデルの外れ値の検出結果と検出法の選択

第３節では，前節で説明した STATA で利用可能な３種の多変量外れ値検出コマンド**hadimvo**・**bacon**・**mcd**（のデフォルト値）を矢根・矢根（2018）の１ステップモデルの推定に用いられた12変数に適用する。前節で説明したように，仮に２ステップ法を取る場合には，まず生産フロンティアに関する４変数が使われ，次に効率性を回帰する８個の環境変数が使われることになるので，参考のためにそれぞれケースＡ・Ｂとして別個に検出し，１ステップモデルの一度に12変数を使ったケースＣと比べたのが**表１**である。

表１　多変量外れ値検出法による外れ値の検知数

	bacon	hadimvo	mcd
A 生産フロンティア	0（0%）	5（0.40%）	131（10.54%）
B 環境変数	9（0.72%）	76（6.11%）	490（39.42%）
C 全変数（A＋B）	2（0.16%）	81（6.52%）	361（29.00%）

＊()内は，全サンプル数1243事業者に対する比率

表１の３種のコマンドのデフォルト値での外れ値検出数を比較すると，**bacon**・**hadimvo**・**mcd** の順に次第に大きくなり，その格差は前節の文献の示唆に比べればかなり大きいように思われる。しかし，Ａ・Ｂ・Ｃのいずれのケースにおいても，**bacon** による外れ値は必ず **hadimvo** の外れ値であり，その **hadimvo** の外れ値はすべて **mcd** の外れ値であるという包含関係が成立するという意味では，整合的である。というのも，互いに閾値を調整すれば，格差は縮小すると予想されるからである。ただし，デフォルト値であっても，これほどの格差が現れるという事実は，前節で言及した多くのシミュレーション結果から考えれば，予想し難い結果かもしれない。

次に，各コマンド別にＡ・Ｂ・Ｃでの検出数を比較すると，Ａの生産フロンティアでの外れ値が最も少ないことがわかる。これは，矢根・矢根（2018，表１）が指摘するように，産出量や生産要素の散度は高くても，その対数値を使っているからである。他方，**bacon** と **mcd** ではＢの環境変数の検出数が最大であるが，**hadimvo** ではＣの全変数での検出数が最

大になる。必ずしも変数が多くなるほど外れ値が増えるわけではない点に留意すべきである。

第３に，本稿で使用する全変数を用いたＣの **bacon** での外れ値は新潟県の聖籠町と佐賀県の鳥栖市の２事業者のみであるが，いずれも吉川・他（2012, p. 86）が水道管総延長距離が（前後の期間と比べて）異常に低いと指摘した事業者に他ならない点である。すなわち，前節で言及した Wang（2002, p. 246）による外れ値除去と同じように，除去する確固たる客観的な理由が存在するのである。

事実，矢根・矢根（2018，表１）によれば，顧客密度 **cusden** の平均値は157だが，その最大値は聖籠町の11464，次に大きいのが1410の鳥栖市であり，異常に短い水道管総延長距離によってあまりに高い顧客密度になったと考えられる。この２事業者を除けば，顧客密度の標準偏差は三分の一以下になり，平均値も147まで低下する。

本稿の目的は外れ値に対する矢根・矢根（2018）の１ステップモデルにおける生産フロンティアと技術効率性の感度を分析する点にあるので，Ｃの全変数の外れ値検出の結果を利用するのが妥当である。しかし，特に **mcd** のように多数の外れ値が除外された場合には，対数尤度の最大値をうまく探し出せず，STATA の具体的な警告でいえば「not concave」や「Backed Up」で終わってしまう場合が起こりうる。もちろん STATA でも，計算アルゴリズム等を変更することにより解決できる場合もあるが，本稿では外れ値に対する感度分析が目的なのでアルゴリズム等は変更しないことを原則とする[8)]。

その際，問題になるのが，やはり **mcd** によって３割近くを除外されたサンプルである。（２）-（４）式から成る基本的な SFA モデルでも，（２）・（６）-（８）式から成る環境変数を含めた１ステップモデルでも，Backed Up の警告行で終わるからである。さらに本質的な問題は，基本的な SFA モデルの推定はアルゴリズムの変更で可能だとしても，前節で説明した非効率性が存在しないという帰無仮説を10％水準でも棄却できないことである[9)]。すなわち，２ステップ法で言えば，そもそも最初のステップで説明すべき非効率性が除去されてしまうのだから，SFA ではなく OLS の適用が妥当になってしまうのである。ケース A のフロンティア変数のみを mcd で検出し１割近い外れ値を除去した場合にも，基本的な SFA モデルでの非効率性が存在しないという帰無仮説を棄却できず，第２ステップに移る余地などないのである。

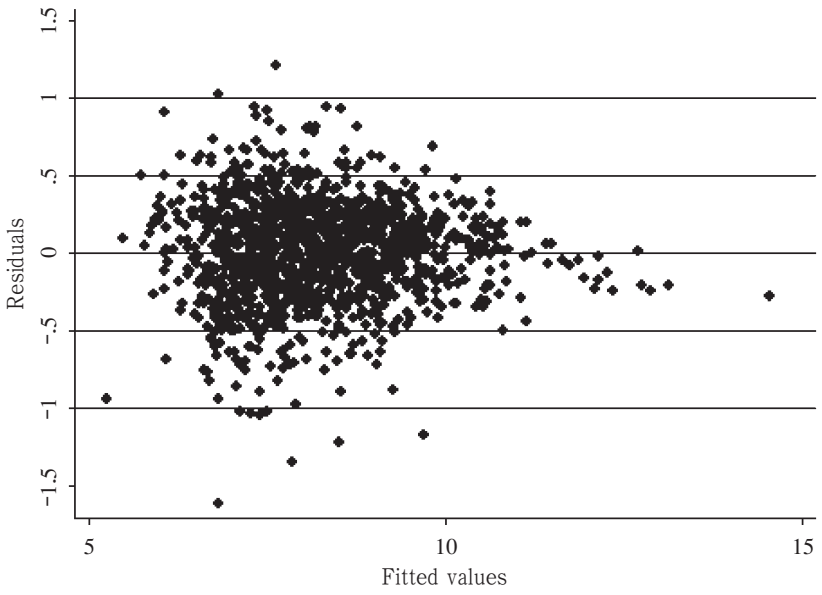
実際，OLS によって生産関数を推定すると，**mcd** による外れ値除外によって，自由度調整済み決定係数は0.93から0.95へと上昇する。この理由は，生産関数の予測値に対する OLS の残差をプロットした**図１**からも伺える。

8）STATA の technique では，nr（Newton-Raphson），bhhh（Berndt-Hall-Hall-Hausman），dfp（Davidon-Fletcher-Powell），bfgs（Broyden-Fletcher-Goldfarb-Shanno）の４種のアルゴリズムを選択・混合できる。本稿では一貫して，technique（dfp 5 bhhh 2 nr 5）という回数での混合を用いる。

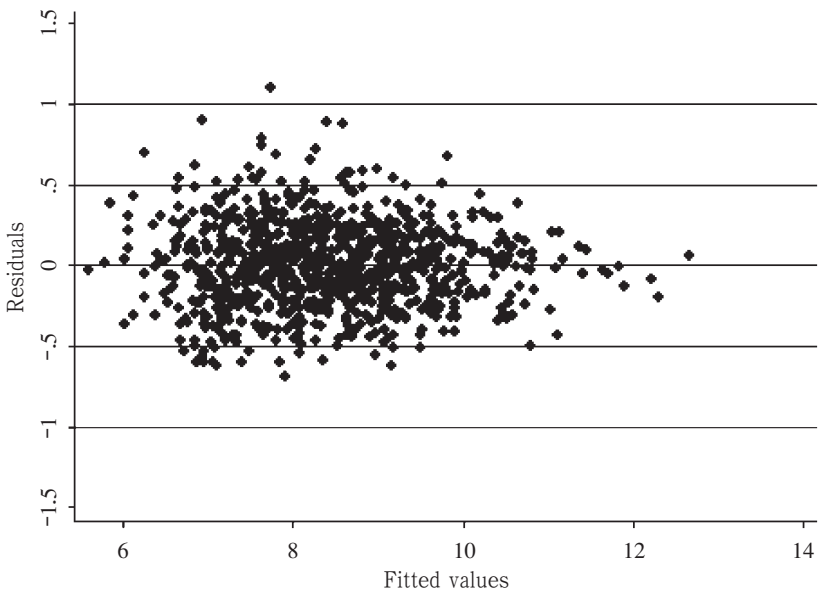
9）たとえば technique（bhhh 2 nr 15）に変更すれば，対数尤度-70で収束する。しかし，片側誤差項の標準偏差が片側および両側誤差項の標準偏差の和に占める比率は0.14になり，帰無仮説を棄却できない。

図１　生産関数推定における予測値に対する残差

全サンプル1243事業の生産関数の最小二乗推定の rvf プロット



mcd による外れ値除去後の881事業の生産関数の最小二乗推定の rvf プロット



外れ値を除外する前の上図での残差はおよそ－1 から 1 までの広範な分布だが，**mcd** によって３割近くの外れ値を除去した下図の残差のかなりの分布は-0.5より大きい範囲内にほぼ収まるように変化しているからである。すなわち，デフォルト値での **mcd** の適用は，非効率

性を表す片側誤差項の分散を無視できるまで縮小させるほど強力だといえよう。換言すれば，デフォルト値を緩めない限り，生産関数の推定ならともかく，本稿のようなフロンティア分析には適用できないので，**mcd** による外れ値の更なる検討は Appendix にて言及する。

ゆえに，デフォルト値で１ステップモデルに適用できる外れ値検出法は，**bacon** と **hadimvo** の２種になるが，両者の外れ値検出の百分比は0.2％と6.5％という開きがある。そこで，**bacon** の閾値を緩めたり，**hadimvo** の閾値を厳しくしたりすることで，その中間にある外れ値も検出して吟味することも有益かもしれない。本稿では，確度の高い最少の異常値を検出した **bacon** の閾値を緩めるアプローチを採用しよう[10]。

前節で説明したように，**bacon** において外れ値を決めるカイ二乗分布のパーセンタイルのデフォルト値は0.15であり，この値を引き上げれば外れ値の検出数も増加するはずである。そこで，0.15から0.25・0.35・0.45へと0.1ずつ増やしてみると，外れ値の検出数も２から３・６・12へと増加することを確認できる。ちなみに，ケースＢの環境変数で検出された９個の外れ値はすべて，この最も緩い閾値で検出された12個の外れ値に含まれている。

そこで次節では，この **bacon** の閾値の緩和に伴う外れ値数の変化に対する SFA の１ステップモデルにおける生産フロンティアと技術効率性の感度を分析する。その際，**hadimvo** による81個の外れ値をその上限値として参照・利用する。

4　生産フロンティアと技術効率性への影響

第４節では，前節で検出した外れ値の除去に対する SFA の１ステップモデルによるフロンティアと技術効率性の推定値の感度を吟味する。具体的には，bacon の外れ値検出の閾値であるカイ二乗分布のパーセンタイルをデフォルトの0.15から0.25・0.35・0.45へ増加させた場合に，矢根・矢根（2018）による尤度比検定の結果選択された Wang（2002）モデルの推定値がどのように変化するかを検討する。その際，第１・３節で説明したように，外れ値の有無が引き起こす影響に焦点を当てるため，STATA の最大化アルゴリズムを固定して使用する。

表２は，これらの SFA モデルの推定結果，すなわち（２）-（４）式の基本モデル **S**，（２）・（６）-（８）式で拡張した１ステップモデル **SMU**，その１ステップモデルに **bacon** の閾値0.15と0.45で除去したサンプルを用いた **SMU-B0.15**・**SMU-B0.45**，**hadimov** のデフォルト値で除去したサンプルを使った **SMU-Hadi** の係数推定値のみを要約したものである。**bacon** の閾値0.25と0.35の推定結果は省略しているが，同一のアルゴリズムで収束し，推定結果もほ

10）81個の外れ値を検出した **hadimvo** を基準にしないもう１つの理由は，１ステップモデルでは収束するものの，その他での点では **mcd** の場合と同じ問題に直面するからである。すなわち，アルゴリズムを変更して収束させた基本的な SFA モデルにおける非効率性が存在しないという帰無仮説を10％の有意水準でも棄却できないからである。**hadimvo** によるケース A のフロンティアでの検出数が５であることからも推測されるように，本稿のフロンティア分析での6.5％の外れ値は過多であるように思われる。

表２　SFA の１ステップモデルの係数推定値

モデル	S	SMU	SMU-B0.15	SMU-B0.45	SMU-Hadi
（２）式の生産関数の係数値					
lk	0.280***	0.217***	0.288***	0.250***	0.239***
ll	0.288***	0.128***	0.119***	0.119***	0.116***
lo	0.438***	0.307***	0.327***	0.288***	0.308***
_定数項	−2.128***	3.070	0.731*	2.537	1.856***
（７）式の平均への係数値					
lwc		−0.348***	−0.259***	−0.333***	−0.331***
cusden		−0.000	−0.000***	−0.000***	−0.000***
Load		−0.005***	−0.004***	−0.005***	−0.004***
dtypews2		−0.071***	−0.058***	−0.068***	−0.057***
dtypewe2		−0.026**	−0.023**	−0.020*	−0.016*
rraw		0.019***	0.023***	0.019**	0.027***
rsubp		0.005***	0.005***	0.004***	0.006***
UPrice		0.003***	0.004***	0.003***	0.003***
_定数項	−0.639*	5.662***	3.673***	5.288***	4.592***
（８）式の分散への係数値					
lwc		−0.382***	−0.250***	−0.475	−0.268**
cusden		−0.004*	−0.001	−0.058***	−0.002
Load		−0.026**	−0.020***	0.006	−0.015*
dtypews2		0.534*	0.351**	1.683**	0.053
dtypewe2		−0.614**	−0.365***	−0.899*	−0.430**
rraw		−0.173	−0.136*	0.094	−0.273*
rsubp		0.005	0.005	−0.018	−0.005
UPrice		−0.003	−0.002*	−0.012*	−0.001
定数項	−1.525***	2.482	0.556	5.917	0.356
サンプル数	1243	1243	1241	1231	1162
対数尤度	−344.261	787.654	778.946	788.138	808.899

係数の右肩のアスタリスクは，* $p<0.05$，** $p<0.01$，*** $p<0.001$ を示す

ぼ同様である。

まず，全1243事業者をサンプルとする基本モデル S と拡張した１ステップモデル SMU の推定結果を比べれば，矢根・矢根（2018）が強調したように，環境変数の考慮により（２）式のフロンティアの係数値が大きく変化していることを確認できる。事実，基本モデルでは収穫一定の帰無仮説を10％の有意水準でも棄却できないが，拡張モデルでは0.1％水準で棄却されるという意味で，適切な方法で十分な環境変数を考慮しないパラメーターの推定値には無視できないバイアスを伴うと結論づけられている。

次に，前節で指摘した明らかな異常値を示す２事業のみを除外した１ステップモデル **SMU-B0.15** を SMU の推定結果と比べると，（７）式の平均値に及ぼす環境変数の影響にお

いて，10%でも有意でなかった顧客密度 **cusden** が0.1%水準で有意になることがわかる[11]。わずか0.2%足らずの外れ値の除去によって結果が変わりうるという意味では，少なくとも明らかな異常値の検出は非常に有益であることを示唆している。

第３に，この環境変数が(７)式の平均値に及ぼす影響と(２)式の生産フロンティアのパラメーターは，仮に外れ値の上限として **hadimov** による81事業者の除外まで想定したとしても，定数項への影響を除けば，ほぼ頑健なことである。**表１**の上段の生産フロンティアでは，資本やその他の生産要素の係数推定値に多少の変動はあるものの，一次同次の帰無仮説の棄却とすべての有意水準が変化しないという意味で，推定結果は安定しているからである。同様に，中段の平均値に及ぼす環境変数の影響も，地下水 **dtypewe2** が有意水準を５%まで低下させるものの，全体として符号も有意性も安定しているといえよう。

他方，同じ環境変数でも下段の分散に対する影響は，そもそも全サンプルの SMU でも補助金比率 **rsubp** や料金水準 **UPrice** は有意でないうえ，有意であった変数も有意でなくなる場合がある。矢根・矢根（2018）が指摘するように，仮にこの分散への影響が上記の平均への影響と密接に関係しているならば，両者の関係を吟味したうえでの更なるモデルの精緻化が必要かもしれない。

このような１ステップモデルによる環境変数の考慮は，矢根・矢根（2018）によれば，生産フロンティアのパラメータだけでなく，技術効率性の推定値にも大きな影響を与える。そこで，**表２**のモデルを使って(５)式の技術効率性 $\boldsymbol{E}$ を求め，それらの相関・順位相関係数を要約したのが**表３**である。技術効率性 $\boldsymbol{E}$ の後の()の中は，表２のモデル名である。

表３　技術効率性推定値 E の相関係数・順位相関係数

	E(S)	E(SMU)	E(B0.15)	E(B0.45)	E(Hadi)
E(S)	1	0.435	0.572	0.734	0.465
E(SMU)	0.580	1	0.970	0.812	0.995
E(B0.15)	0.680	0.986	1	0.8897	0.989
E(B0.45)	0.761	0.986	0.92077	1	0.840
E(Hadi)	0.571	0.997	0.993	0.883	1

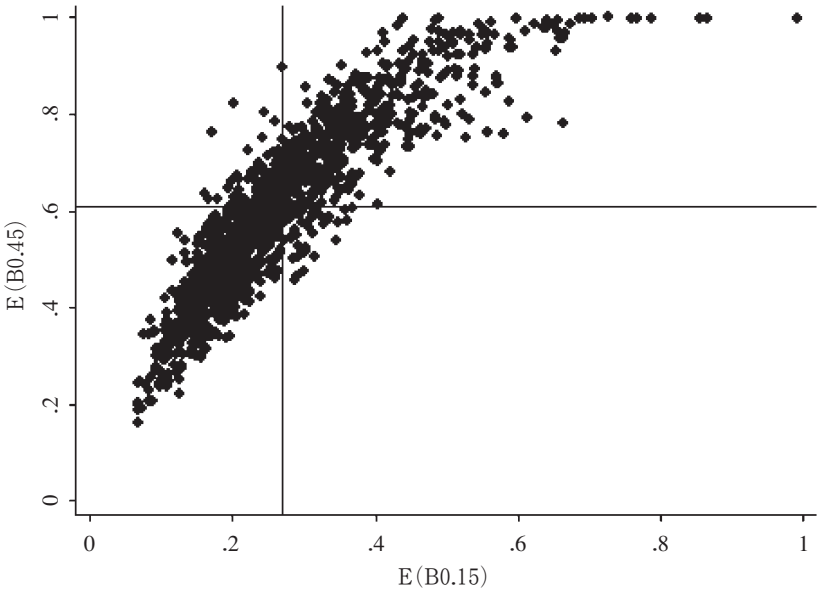
対角線の右上は相関係数，左下は順位相関係数
ただし，それぞれの対象サンプル数は少ない方に調整して算出

表３から，矢根・矢根（2018, 図３）が指摘したように，環境変数を考慮した１ステップモデルと比べれば，環境変数を考慮しない基本モデル **S** による効率性の推定値 E(S) のバイアスは明らかであろう。他方，全サンプルの１ステップモデルの効率性 E(SMU) は，順位相関でみる限り，外れ値を除去した３種のモデルと0.99程度の高い相関を維持している。ゆ

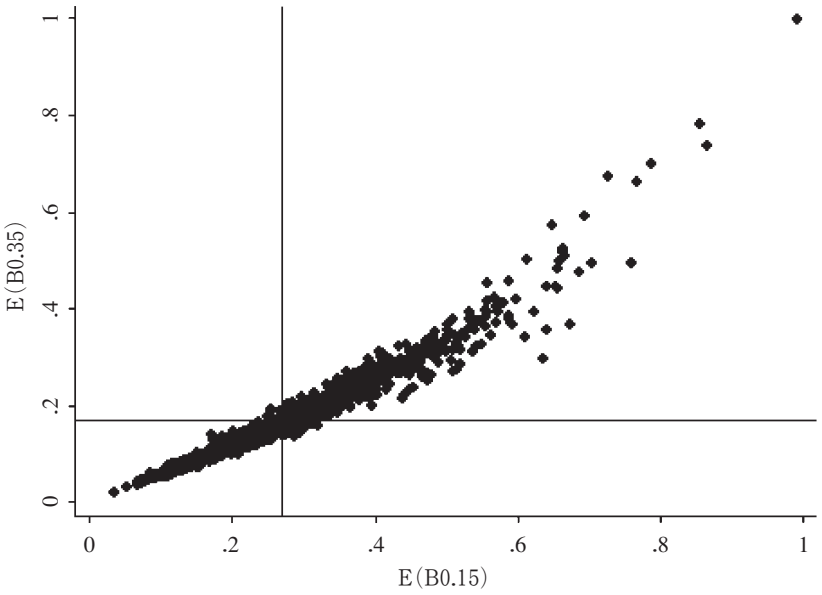
11）**表１**のモデル SMU の平均値に及ぼす顧客密度 **cusden** の p 値は0.93である。ただし矢根・矢根（2018, 表２）では，顧客密度を含む８環境変数のすべてが0.1%水準で有意であると報告されている。これは，デフォルトの最大化アルゴリズム（nr）を使っているからであり，その場合には対数尤度は791まで上昇する。

図２　**E**（**B0.15**）に対する技術効率性の散布図

E（**B0.45**）の **E**（**B0.15**）に対する散布図（1231事業者）



E（**B0.35**）の **E**（**B0.15**）に対する散布図（1237事業者）



えに，技術効率性の推定値も，外れ値の除去に対して頑健であると結論づけられよう。

ただし，**表３**に示した外れ値を除外した３種のモデルによる効率性を比較した場合，**bacon** による２事業者を除去した場合の **E**（**B0.15**）が，同じ手法を用いて12事業者を除去

した**E**(**B0.45**) よりも，上限値の参考として81事業者を除去した**E**(**Hadi**) と高い相関・順位相関係数を有する点に留意すべきかもしれない。というのも，**bacon** の閾値を0.25・0.35とした場合の効率性**E**(**B0.25**)・**E**(**B0.35**) と **E**(**B0.15**) との順位相関係数は（表３には記載していないが）0.986・0.987と十分に高いからである。

図２は，それぞれ**E**(**B0.45**) と **E**(**B0.35**) の **E**(**B0.15**) に対する散布図であり，各軸の垂直線はその平均値を示している。たとえば，横軸の **E**(**B0.15**) の平均値は0.27である。下段の縦軸の **E**(**B0.35**) の平均値は0.17とやや低下しているのに対し，上段の縦軸の **E**(**B0.45**) の平均値は0.61に大きく上昇していることがわかる。**E**(**B0.25**) と **E**(**Hadi**) の平均値がそれぞれ0.16と0.26であることからも，**E**(**B0.45**) の技術効率性はフロンティアの係数変化から予想される以上に大きく変化しているのである。

ゆえに，技術効率性の推定値が外れ値の除去に対して頑健であるのは事実だとしても，図２が示唆するように，**E**(**B0.35**) から **E**(**B0.45**) へと外れ値を６個増加させただけでも効率性水準や相関関係をかなり変化させる場合もありうることに留意すべきである。この **bacon** の外れ値除去の影響を念頭に置けば，**E**(**Hadi**) の **E**(**B0.15**) とのきわめて高い相関関係には偶然の要素も加味する必要があるかもしれないからである。

以上から，全12変数を対象にして **hadimov** による81外れ値を上限参考値とした **bacon** の閾値緩和による外れ値の増加に対して，SFA の１ステップモデルのフロンティアのパラメーターもその推定フロンティアから算出される技術効率性の推定値も頑健であると結論づけることができる。パラメーター推定値はいずれも一次同次性を棄却し，技術効率性の順位相関係数は0.9を超えるという意味で，整合的な推定結果を得るからである。

5　分析結果の要約と帰結

本稿では，日本の水道事業の効率性分析に初めて多変量外れ値の除去を適用し，SFA の１ステップモデルのフロンティアのパラメーターおよび技術効率性の推定値の感度を検討した。具体的には，Wang（2002）のモデルを上水道事業に適用した矢根・矢根（2018）が用いた12変数に，STATA で利用可能な３種のコマンド **hadimvo**・**bacon**・**mcd** を適用したわけである。その結果，少なくともデフォルト値では **mcd** は361個（29％）に及ぶ外れ値を非効率性要素とともに除去してしまうため本稿の分析には妥当ではないことを確認する一方，**bacon** による閾値の緩和に対しては矢根・矢根（2018）の推定結果が頑健であることを例証したのである。すなわち，本稿の主要な結論は，以下の４点に集約できよう。

第１は，デフォルトでの外れ値の検出数は **bacon**・**hadimvo**・**mcd** の順に次第に多くなるが，**bacon** の外れ値はいずれも **hadimvo** の外れ値であり，**hadimvo** の外れ値はすべて **mcd** の外れ値になるという意味で，整合性を有する点である。**表１**に要約されているように，たとえばＣ欄の外れ値検出数は２（0.2％）から361（29％）と100倍以上という予想以上に大きな格差はあるものの，おそらくデフォルト値の変更によって整合性を高められる可

能性が高いからである。

第２は，1243事業者のうち明らかな異常値を示す２事業者を取り除くと，技術非効率性に対して有意でなかった環境変数である顧客密度 **cusden** が0.1％水準で有意になることである。そもそも矢根・矢根（2018）では８個のすべての環境変数が有意だったのに本稿では顧客密度のみが有意ではなくなった理由は，最大化のアルゴリズムを原則固定したために，対数尤度が若干低下したからである。しかし，デフォルト値での **bacon** が検出した２事業者は，吉川・他（2012, p. 86）が異常に高い顧客密度だと指摘していた事業者であるという意味で，外れ値検出の重要性を裏づけている。さらに，わずか0.2％に満たない異常値の除去によって有意水準が変わりうる事実は，外れ値への対応の重要性を例証している。

第３は，この環境変数の影響を含めた生産フロンティアと技術効率性の **bacon** による外れ値に対する感度分析の結果が，頑健とみなせることである。具体的には，**hadimvo** による81個（約７％）の外れ値を上限の参考値として，bacon の閾値であるカイ二乗分布のパーセンタイルをデフォルトの0.15から0.45まで0.1ずつ増加させても，環境変数の有意性やフロンティアの一次同次性の棄却は変わらず，効率性の順位相関係数も0.9を維持することを確認できたのである。

第４は，サンプル数の３割弱を外れ値として検出する **mcd** は，非効率性を示す片側誤差項の分散を過度に収縮させてしまうため，少なくともデフォルト値ではこのフロンティア分析には適用できないことである。これは，そもそも(２)式で表されるフロンティアモデルが(３)式の両側誤差項に収まりきらない(４)式の片側誤差項を要するのに，外れ値除去によって(４)式の必要性自体が喪失するパラメトリックなフロンティア分析固有の問題を示唆するのかもしれない。同じフロンティア分析でも，ノンパラメトリックな DEA 等においてフロンティアの形状や各事業者の技術効率性に重要な影響を与えるのは，主にフロンティア上にある外れ値だけだからである。しかしパラメトリックなフロンティア分析では，上記のような散度とその形状も重要になりうる。たとえば，Appendix でも例示されるように，**E(B0.45)** の平均効率性の上昇は，**E(B0.35)** からむしろ非常に非効率な６個の外れ値を除去したことによると考えられるからである。

以上の分析結果に基づけば，矢根・矢根（2018）の SFA の１ステップモデルによる生産フロンティアおよび技術効率性の推定は，外れ値の除去に対して頑健であるといえよう。ただし，それは STATA における **bacon** の閾値の緩和に対してであり，**mcd** は少なくともデフォルト値では適用できない。すなわち，検出手法の選択や閾値の設定に関する判断が結果に及ぼす影響は決して小さいとはいえず，判断力の向上にはむしろ今後も多様な外れ値検出の蓄積が不可欠だといえよう。

参　考　文　献

Aigner, D., C. A. K. Lovell and P. Schmidt (1977), “Formulation and Estimation of Stochastic Frontier

Production Function Models," *Journal of Econometrics* 6, no. 1: 21-37.

Ben-Gal, I. (2005), "Chapter 1: Outlier detection," In *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers,* edited by O. Maimon and L. Rockach, pp. 1-16. Dordrecht: Kluwer Academic Publishers.

Billor, N., A. S. Hadi, and P. F. Velleman (2000), "BACON: blocked adaptive computationally efficient outlier nominators," *Computational Statistics & Data Analysis* 34: 279-298.

Coelli, T. J. (1995), "Estimators and hypothesis tests for a stochastic frontier function: A Monte Carlo analysis," *Journal of Productivity Analysis* 6: 247-268.

Greene, W. H. (2017), *Econometric Analysis,* 8th edition. New York: Pearson Education, Inc.

Hadi, A. S. (1992), "Identifying multiple outliers in multivariate data," *Journal of the Royal Statistical Society* Series B 54, no. 3: 761-771.

Jondrow, J., K. Lovell, I. Materov, and P. Schmidt (1982), "On the estimation of technical inefficiency in the stochastic frontier production function model," *Journal of Econometrics* 19: 233-238.

Kumbhakar, S. C. and C. A. K. Lovell (2000), *Stochastic Frontier Analysis.* Cambridge: Cambridge University Press.

Kumbhakar, S. C. and H.-J. Wang (2015), "Estimation of technical inefficiency in production frontier models using cross-sectional data," In *Benchmarking for Performance Evaluation: A Production Frontier Approach,* edited by S. C. Ray, S. C. Kumbhakar and P. Dua, pp. 1-73. New Delhi: Springer India.

Meeusen, W. and J. van den Broeck (1977), *International Economic Review* 18, no. 2: 435-44.

Penny, K. I. and I. T. Jolliffe (2001), "A comparison of multivariate outlier detection methods for clinical laboratory safety data," *The Statistician* 50, no. 3: 295-308.

Rousseeuw, P. (1985), "Multivariate estimation with high breakdown point," In *Mathematical Statistics and Applications,* edited by W. Grossmann, G. Pflug, I. Vincze, and W. Wertz, pp. 283-297. Dordrecht: Reidel Publishing Company.

Simar, L. (2003), "Detecting outliers in frontier models: A simple approach," *Journal of Productivity Analysis* 20: 391-424.

Simar, L. and P. W. Wilson (2007), "Estimation and inference in two-stage, semi-parametric models of production processes," *Journal of Econometrics* 136, no. 1: 31-64.

Verardi, V. and C. Dehon (2010), "Multivariate outlier detection in Stata," *The Stata Journal* 10, no. 2: 259-266.

Wang, H.-J. (2002), "Heteroscedasticity and non-monotonic efficiency effects of a stochastic frontier model," *Journal of Productivity Analysis* 18, no. 3: 241-253.

Wang, H.-J. and Schmidt (2002), "One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels," *Journal of Productivity Analysis* 18, no. 2: 129-144.

Weber, S. (2010), "bacon: An effective way to detect outliers in multivariate data using Stata," *The Stata Journal* 10, no. 3: 331-338.

Wilson, P. W. (1993), "Detecting outliers in deterministic nonparametric frontier models with multiple outputs," *Journal of Business & Economic Statistics* 11, no. 3: 319-323.

Yane, S. and S. V. Berg. (2013), "Sensitivity Analysis of Efficiency Rankings to Distributional Assumptions: Applications to Japanese Water Utilities," *Applied Economics,* March (45): 2337-48.

中山徳良（2015),「日本の水道事業の技術効率性に影響を与える要因の分析」『オイコノミカ』52(1), 101-112.

矢根真二（2016),「水道料金格差の解消と道州制レベルの広域化：市町村原則の罪と政治的な価格決定」『桃山学院大学総合研究所紀要』第42巻　第２号, 23-40.

矢根遥佳・矢根真二（2018),「パラメトリックな確率的生産フロンティアへの環境要因の影響」『経済経営論集』第59巻第４号，3-26.
吉川丈・礒合良輔・矢根遥佳・矢根真二（2012),「確率的生産フロンティアと環境変数：技術効率性効果フロンティアモデルの上水道事業への適用」『経済経営論集』（桃山学院大学）第53巻第４号，155-179.
和田かず美（2010),「多変量外れ値の検出　～MSD 法とその改良手法について～」『統計研究彙報』第67号，89-157.

APPENDIX　検出された外れ値の属性

本稿では矢根・矢根（2018）による SFA の１ステップモデルの推定が全12変数を対象にした多変量外れ値を除いても頑健であることを例証したが，この Appendix はこれらの推定で除かれた外れ値を中心にそれらの属性を要約しておこう。というのも，第１節で述べたような除去理由の十分な説明は **bacon** のデフォルト値で検出された２事業者以外にはなされておらず，第５節で言及したようにノンパラメトリックなフロンティア分析のようなフロンティア上の外れ値のみが重要ではないからである。さらに，検出された外れ値を伝統的に多用されてきた規模や事業主体によるサンプル区分と比較することは，今後の研究にとっても有益と考えるからである。

表 A1　bacon によって検出された外れ値の事業者名とその属性

	E(SMU)	E(S)	lwc	事業主体
新潟県聖籠町	0.07	0.86	8.9	町村
佐賀県鳥栖市	0.11	0.82	10.6	市
千葉県山武市	0.02	0.45	8.2	市
群馬県長野原町	0.01	0.48	7.6	町村
沖縄県宮古島市	0.09	0.70	10.6	市
北海道浜中町	0.02	0.45	7.7	町村
千葉県市原市	0.05	0.41	11.0	市
青森県東通村	0.03	0.55	8.6	町村
鹿児島県与論町	0.03	0.60	8.2	町村
兵庫県播磨高原	0.02	0.35	9.0	企業団
福岡県立花町	0.01	0.29	7.5	町村
茨城県鉾田市	0.03	0.39	9.4	市

まず，本稿の SFA の１ステップモデルの推定に利用した bacon の閾値であるカイ二乗分布のパーセンタイルをデフォルトの0.15から0.45まで0.1ずつ増加させた場合に外れ値となった事業者リストが**表 A1** である。これら12事業者はデフォルトで計算された距離順に並べられており，聖籠町と鳥栖市は初期値0.15での検出事業者，0.25になると千葉県の山武市が加わって３事業者となり，さらに閾値を0.35・0.45と緩くすると上から６事業者・12事業者と検出事業者も増加する。その右の列には，環境変数を含めた１ステップモデルの効率性 E(SMU)，参考値としての環境変数を含まない基本モデルの効率性 E(S)，規模指標としての取水能力の対数値 **lwc**，事業主体を掲載している。全1243事業者の平均値は，E(SMU) が0.11，E(S)が0.81，**lwc** が9.9である。

第４節で言及したように，**bacon** の閾値を0.35から0.45に引き上げると，その技術効率性の平均値は E(B0.35) の0.17から E(B0.45) の0.61に急増する。この大きな変化は，**表 A1** の市原市から鉾田市までのわずか６事業者の除外によって生じるわけだが，これらの効率性はすべて平均以下である点に注目す

表 A2　bacon・hadimov・mcd によって検出された外れ値の事業者属性

		B0.45		Hadi		mcd		mcd-A	
		個数	%	個数	%	個数	%	個数	%
E(SMU)	最低効率	9	75	54	67	147	41	57	44
	低効率	1	8	10	12	82	23	16	12
	中効率	1	8	9	11	59	16	11	8
	高効率	1	8	3	4	30	8	7	5
	最高効率	0	0	5	6	43	12	40	31
E(S)	最低効率	10	83	54	67	149	41	48	37
	低効率	0	0	8	10	74	21	22	17
	中効率	1	8	12	15	51	14	20	15
	高効率	1	8	3	4	46	13	14	11
	最高効率	0	0	4	5	41	11	27	21
取水規模	最小規模	7	58	40	49	112	31	56	43
	小規模	2	17	18	22	87	24	20	15
	中規模	0	0	8	10	72	20	9	7
	大規模	2	17	8	10	45	12	7	5
	最大規模	1	8	7	9	45	12	39	30
事業主体	都道府県	0	0	2	2	2	1	3	2
	政令都市	0	0	2	2	6	2	14	11
	市	5	42	18	22	147	41	33	25
	町村	6	50	50	62	187	52	75	57
	企業団	1	8	9	11	19	5	6	5

べできである。実際，これらの 6 事業者の効率性 E(B0.35) はその平均値を大きく下回っている。すなわち，E(B0.45) の平均値の上昇は，E(B0.35) の非効率な 6 事業者を除外することによってもたらされたのである。

表 A2 は，上述の閾値を0.45にした場合の **bacon**，表 1 の全12変数を使ったケース C の **hadimov** と **mcd**，さらに参考としてフロンティア 4 変数のみを対象としたケース A に **mcd** を用いた **mcd-A** の外れ値の属性を要約したものである。効率性や取水規模は全1243事業者をそれぞれ 5 分割し，その分位に属する外れ値の個数と，その個数が外れ値全体に占める百分比を示している。

上述した **bacon** ほどではないにせよ，いずれの検出法も最も非効率な分位から20%を大きく超える外れ値を検出していることがわかる。これは，いずれの検出法も最小規模から最も多くの外れ値を検出している事実と符合するようにみえる。ただし，**mcd**，とりわけ 4 変数のみに適用した **mcd-A** では，その傾向が弱まる一方で，最高効率や最大規模からの外れ値の検出比が高まるようである。この特性は，第 3 節や**図 1** に示されたような(4)式の片側誤差項の消滅に関連しているように思われる。したがって，下段の主体別の区分でも町村や市の比重が高いことになるが，上記の効率性や規模のように 5 等分されたカテゴリーではないので，注意が必要である。

そこで，全1243事業者の主体別の事業者数と構成比の隣に，**表 A2** のデフォルトで大量の外れ値を検出した **mcd** の結果を並べたのが**表 A3** である。**mcd** の外れ値の構成比は，最大の標本である市からの構成比のみが低下し，その他のを主体の構成比はすべて上昇していることがわかる。とりわけ，フロンティアのみに適用された **mcd-A** では，都道府県と政令都市からの外れ値の検出が顕著である。

この事実は，伝統的な事業主体別のフロンティア分析に一定の根拠を与えるかもしれない。たとえば

表 A3　mcd によって検出された外れ値の事業主体属性

	全1243事業者		mcd		mcd-A	
都道府県	4	0.3%	2	0.6%	3	2.3%
政令都市	16	1.3%	6	1.7%	14	10.7%
市	646	52.0%	147	40.7%	34	26.0%
町村	531	42.7%	187	51.8%	74	56.5%
企業団	46	3.7%	19	5.3%	6	4.6%
合計	1243	100.0%	361	100.0%	131	100.0%

中山（2015, p. 106）は，政令指定都市は規模が大きいため，都道府県と企業団は複数の市町村にまたがるため，この３種の事業主体をサンプルから除いており，確かに**表 A3** の **mcd** による検出結果と符合する側面も有するからである。すでに第３節で言及したように，**mcd** による外れ値の除去が SFA の基本モデルの非効率性の存在自体も除去してしまうことを想起すれば，十分に合理的な伝統的説明の重要性はむしろ高まるのかもしれない。特に，サンプル次第で効率性水準が大きく変わりうるフロンティア分析では，合理的な説明が重要になろう。

しかし，少なくとも本稿の分析結果によれば，パラメトリックな **bacon** による多変量外れ値検出が異常値の発見に有効であり，矢根・矢根（2018）による全事業者のパラメトリックな推定がその外れ値の除去に対して頑健であることを例証している。さらに，たとえ数パーセントのサンプルの相違でも，推定結果が大きく変わりうる危険性も同時に物語っている。すなわち，一定の整合性は確認できたとはいえ，同じパラメトリックな方法といえども，採用する手法や閾値に明確な原則はないのが現状である。さらに，ノンパラメトリックな外れ値検出やフロンティア分析との整合性の検討も，今後の大きな課題となろう。

（2018年２月６日受理）

Sensitivity Analysis of Parametric Production Frontiers to Multivariate Outliers: Application of Multivariate Outlier Detection using Stata to Japanese Water Utility Analysis

YANE Shinji

The purpose of this paper is to conduct a sensitivity analysis of production frontiers and technical efficiencies estimated by a one-step model in stochastic frontier analysis (SFA) against the removal of multivariate outliers. Specifically, this study applies three commands available in STATA, **hadimvo**・**bacon**・**mcd**, on the 12 variables used by Yane and Yane (2018), which applied Wang (2002)'s model to Japanese water utilities. The main findings are as follows: 1) although by default the number of outliers detected is the highest with **mcd** followed by **hadimvo** and **bacon**, it is consistent in that **bacon**'s outliers are those of **hadimvo** and **hadimvo**'s outliers are those of **mcd**; 2) removing the only two observations with extreme values among the 1,243 water utilities, customer density variable *cusden*, which is an environmental variable and was not statistically significant to technical inefficiency, becomes statistically significant at the 0.1% level; 3) the results of sensitivity analysis against outliers using **bacon** and **hadimvo** on production frontier and technical efficiency including the forementioned environmental variable, are mostly robust; and 4) **mcd**, however, detects almost 30% of the sample as outliers, and extremely contracts the variance of the one-sided error term (which indicates technical inefficiency). Hence, it cannot be used for this frontier analysis as a default.