財務諸表データに対する
欠損値補完及び外れ値処理について

高橋　淳一

博士（統計科学）論文

総合研究大学院大学

複合科学研究科

統計科学専攻

目次

第 1 章：序論

1.1　財務諸表データとその特徴

事業を営む経済主体は，いずれも自己の経済活動の結果としての財務諸表を有する.この財務諸表は,事業主体の経営状況のバロメータともいえるもので，様々な情報が含まれている．その情報量は非常に多く，株式市場の参加者は上場企業の経営状況を確認し，投資対象を検討するために，必ずと言ってもいいほど財務諸表の内容を吟味するであろう．実際に，非常に多くのデフォルト予測モデルでは，財務諸表情報を用いた試みがなされている．代表的な論文として，企業の信用リスク評価モデルとして最も初期の論文に Altman（1968）があるが,その論文では製造業の財務諸表[1]から算出される財務指標を判別分析の説明変数として利用している．この他にも，Martin（1977）では，銀行の財務諸表を用いてロジット回帰分析を行い，信用リスク評価モデルが推計されている．また，Lane et al.（1986）でも銀行の財務諸表を用いて Cox 比例ハザードモデルが推計されている．これらの代表的な論文のように，財務諸表データを信用リスク評価モデルに活用した例は枚挙にいとまがない.

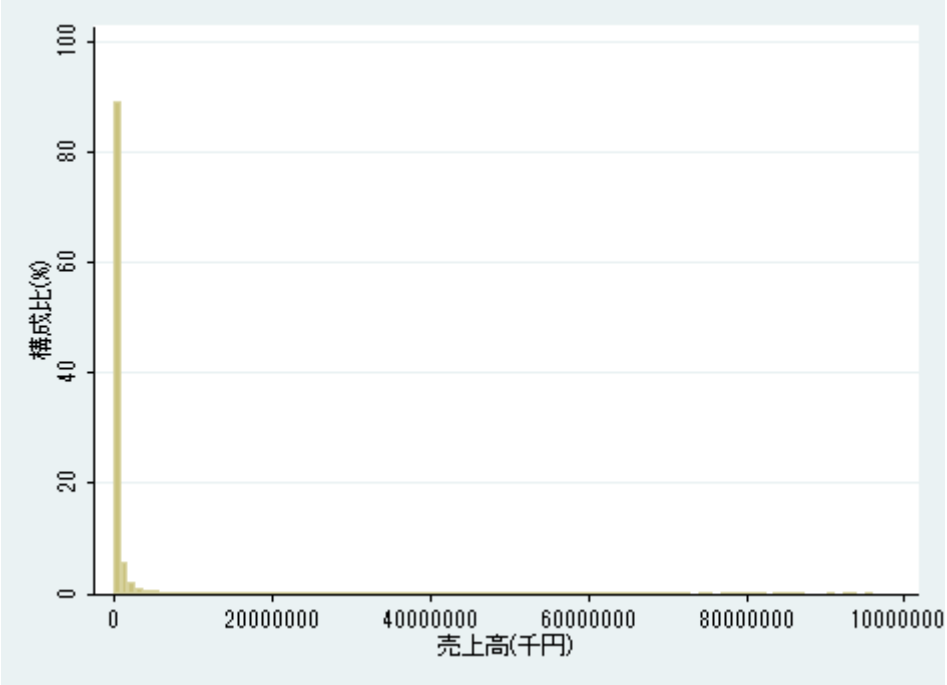
このように，事業主体の経営状況を判断するために非常に有用な財務諸表データであるが，上記の論文を初め，その他の信用リスク評価モデルに関する論文では，欠損値補完や外れ値処理等のデータの事前処理についてはあまり明示されていない．これは，Altman（1968）が規模の類似した製造業のデータのみを抽出したように，その他の論文でも分析対象を絞り込んでいることが主要因と考えられる．すなわち，ある程度均質で質の良いデータが分析に利用されているため，欠損値補完や外れ値処理が不要であったと考えられる．上場企業の財務諸表などは，この条件が満たされるデータの典型であろう.

ところが，中小企業の財務諸表については，上場企業ほど会計基準が統一されているわけではなく，また，規模もばらつきが非常に大きい．一般的に零細

[1] Altman（1968）の論文で利用されているデータは,『Moody's Industrial Manuals』及び『Annual Reports』から引用していると書かれている.

企業としてイメージされる企業と中堅企業としてイメージされる企業を比較すると，非常に大きな差が存在する．例えば，中小企業基本法で定義される中小企業には,零細企業から一般的にイメージされている中堅企業まで含まれるが,その中小企業のデータを大量に集積した CRD[2]のデータベースで売上高[3]の分布を図 1 で確認すると，非常に裾の長い分布となっている．

図 1 売上高の分布



(注)本研究の分析用データから 100 万レコードをランダムサンプリングして抽出したデータを使用した．使用データの分布詳細は，付録 A 参照．

[2] CRD（Credit Risk Database）は，データから中小企業の経営状況を判断することを通じて，中小企業金融に係る信用リスクの測定を行うことにより，中小企業金融の円滑化や業務の効率化を実現することを目指し，中小企業の経営データ（財務・非財務データ及びデフォルト情報）を集積する機関として，全国の信用保証協会を中心に任意団体 CRD 運営協議会として 2001 年 3 月にスタートした．現在の運営組織の名称は，一般社団法人 CRD 協会となっている（http://www.crd-office.net/CRD/index2.htm）．2015 年 3 月末時点で，過去 20 年分の法人及び個人事業主の財務諸表データが，2000 万件以上蓄積されている

[3] CRD のデータ項目の定義では「売上高・営業収益」である．

また，高橋・山下（2002）でも述べられているように，中小企業の財務諸表には，ある程度の欠損値が存在する．今回の分析に用いた CRD の財務諸表データでも，全く欠損が無い財務項目も多いが，40%前後の欠損率となっている財務項目も存在する[4]．財務項目が欠損値となる要因は様々である．例えば，CRD の財務項目については，売上高などの項目は必須項目となっており，その項目に欠損値は存在しない．しかし，必須項目になっていない項目については，データ提供元である保証協会や金融機関のシステム的な制約により取得できない項目も存在する．また，財務項目の一つである建設仮勘定などは建設業を営んでいる経済主体の財務諸表でしか使われないなど，業種によって使う項目と使わない項目が存在する．その他にも様々な要因が考えられるが，いずれにしても，財務諸表データに欠損値は付き物である．

以上のように，経済活動を行っている経済主体全体の財務諸表を集積することを考えた場合，規模面でかなりのばらつきの存在や欠損値の存在を考慮する必要がある．この点は，統計分析を行う上で，外れ値処理や欠損値処理の方法によって，結果が大きく変わることを意味しており，非常に重要な点である．

1.2 本研究の目的

本研究では，1.1 で説明したような外れ値や欠損値の存在する財務諸表データに関し，偏った特性のある一部のデータを抽出して使うのではなく，全体のデータ特性を利用する必要がある場合について，合理的で効率的な処理方法について考察を行っている．これにより，CRD のように，全国的に様々な規模の財務諸表データが集積している大量データに関して，偏った一部データの傾向ではなく，データ全体の特性を踏まえた統計分析が可能となる．

例えば，第 2 章でも言及されるが，CRD データにおいて，デフォルト債務者の財務項目の欠損率は非デフォルト債務者の財務項目の欠損率が高いという傾

[4] 付録 A に欠損率を示した表を掲載している．

向が確認されており,欠損値を含むレコードを削除するような処理をした場合,非デフォルト債務者が多く抽出されてしまうということを意味する.その結果,抽出されたデータを用いた統計分析の結果というのは，非デフォルト債務者の特性を強く反映した，偏った分析結果となってしまう.

しかし，実際の統計分析において，データ全体の特性を利用することが必要なケースは多い．例えば，規模の類似した製造業のデータを抽出した Altman（1968）だが，より多様な規模の企業が含まれているデータや他の業種が含まれたデータでも当該論文で示された結果が当てはまることが示されると，より一般的な結論として説得力のある結論になるであろう．また，物理的なハードディスクの容量拡大やクラウドのようなネットワーク面での革新により，情報を蓄積しておく容量にボトルネックがほぼ無くなりつつある昨今，利用可能なデータの偏った一部だけを用いた分析よりも，データ全体の特性を踏まえた統計分析の必要性が高くなっている．ただし，ここでは“データの全てを使って分析する”ことが必要ということを言っているのではない．データ全体の特性を保持するよう統計分析前のデータ整備を行うことによって，偏った情報に基づかない合理的な分析が可能になるということが重要である.

本研究では，まず，CRD の財務諸表データベースを利用し，財務諸表データ一般に関する標準的な欠損値補完の方法を確立し，扱いやすい標準的なデータベースの構築を行うことを目的とする．これにより，初期データ整備が主眼ではない中小企業研究等が，これまで以上に幅広く行われることが期待される．さらに，欠損値の多い中小企業財務諸表データに対して，欠損値を適切に補完することで，企業の財務状況のより正確な把握や，企業のより正確な信用リスク評価を行えるようになる.

次に，本研究では，信用リスク評価という面について，欠損値処理からさらに分析を深め，外れ値処理について考察する．財務諸表を用いた信用リスク評価モデルでは，最尤法によるロジットモデル推計が行われることが一般的であるが，最尤法による推定量は外れ値の影響を受けやすいという性質を有する．財務諸表データのように，分布に偏りが大きいデータの場合は，最尤法による

推定を行う前に，事前に外れ値の処理を行うことで，推計精度を高めることができる．そこで，欠損値処理で用いた処理手法を応用することを含め，いくつかの外れ値処理の方法論を試行することで，信用リスク評価のモデリング精度向上を目指す．これにより，前述の通り，企業のより正確な信用リスク評価を行えるようになる．

1.3 本論文の構成

第 2 章では，中小企業の経営データを大量に集積したデータベース（CRD）の財務諸表データを用い，財務諸表データの特性である分布の偏りや，時系列方向の自己相関性の強さ[5]を考慮した，欠損値補完方法を提案する．具体的には，欠損項目を含む財務諸表に対して，欠損していない項目に関して類似した財務諸表の値を補完する，k-NN（K-Nearest Neighbor）法という方法である．本研究では，CRD データベースから欠損値の存在しない完全データを抽出した上で欠損値を一定の法則に従って発生させ，今回提案する k-NN 法により欠損値を補完し，その補完値と真値との誤差を計測して，他の欠損値補完方法より誤差が小さくなることを示す．この際，大規模財務諸表データに対する効率的な計算方法として，売上高によるセグメント分割による効率的な距離計算を導入することで，計算効率を大幅に向上させた k-NN 法による欠損値補完が実現したことを示す．

第 3 章では，第 2 章の分析を発展させ，業種区分情報を利用して，補完値と真値の誤差がさらに小さい欠損値補完方法の開発に向けた考察を行っている．

[5] 時系列方向の自己相関性の強さとは，同じ債務者の異なる期の財務諸表は，比較的類似した数値になっているということである．すなわち，A という債務者の 2010 年 3 月決算の財務諸表は，2009 年 3 月決算や 2011 年 3 月決算の数値と類似する傾向があるということである．CRD データベースに対するデータ提供元である信用保証協会や金融機関は，保証先や融資先の経営状況を確認するために，財務諸表を定期的に保証先や融資先から徴求してチェックしている．したがって，CRD データベースには同一債務者で毎年連続して財務諸表が存在するデータは非常に多い．

一般的に，業種によって粗利率や自己資本比率の平均的な水準感は大きく異るように，BS/PL における各財務項目の総資産対比や売上高対比の平均的な構成比は業種による特性が出やすいと考えられる．この点は，k-NN 法においては，財務指標で計測する各レコード（財務諸表）間の距離に影響を及ぼすと考えられるが，本章ではその情報を利用することを考える．すなわち，財務諸表に表われている数値を効率よく距離計算に反映させる情報として業種区分情報を考慮して，より正確な欠損値補完を目指すものである．

第 4 章では，CRD の財務諸表データを用い，集積された財務諸表データを分析する際に常に課題となる裾の長い分布の処理と外れ値処理に焦点を当て，ロジスティック回帰モデルの予測精度を向上させる方法についての研究内容を説明する．この際，第 3 章までの研究成果である有効な欠損値補完方法である k-NN 法を活用することを考える．具体的には，各財務指標を neglog 変換した後，外れ値処理として財務指標毎に平均から 3 標準偏差もしくは 4 標準偏差以上離れている財務指標値を欠損値とする．その欠損値を k-NN 法で補完した後，ロジスティック回帰分析を行って AUC（Area Under Curve）を計測する．この計算手法を，従来の一般的な方法論である，外れ値を折り返し処理した後にロジスティック回帰分析を行った結果と，AUC で比較して精度が向上するかどうかを確認した．結果的に，本章の方法論では，AUC を明確に向上させることができなかった．

そこで，第 5 章では，第 4 章の結果を踏まえて，異なるアプローチによる外れ値処理をロジスティック回帰分析に応用し，AUC 向上を目指した．具体的には，外れ値処理として財務指標毎に分布の両端の数％の範囲に属する数値を両端数％点の値に置き換え（折り返し）処理した後，一般化 neglog 変換を施したものを説明変数としてロジスティック回帰モデルを推計し，その AUC が他の方法論によるロジスティック回帰モデル推計の AUC を上回ることを示す．一般化 neglog 変換を取り入れている点が，本研究における新しい取り組みであり，その有効性が示された．

最後に，第 6 章はまとめである．

第 2 章：大規模財務諸表データに対する k-NN 法による欠損値補完

2.1　はじめに

2.1.1.　研究目的

CRD（Credit Risk Database） には，過去 20 年以上の中小企業及び個人事業主の財務諸表データが 2000 万件以上蓄積しているが，中小企業や個人事業主の財務諸表データに関しては,外れ値や異常値,欠損値などが存在するため，分析を行う前の初期データ整備の負担が大きい．そこで，本研究では　CRD の財務諸表データベースを利用し，財務諸表データ一般に関する標準的な欠損値補完の方法を確立し，扱いやすい標準的なデータベースの構築を行う．これにより，初期データ整備が主眼ではない中小企業研究等が，これまで以上に幅広く行われることが期待される．さらに，欠損値の多い中小企業財務諸表データに対して，欠損値を適切に補完することで，企業の財務状況のより正確な把握や，企業のより正確な信用リスク評価を行えるようになる．

2.1.2.　欠損値補完に関する既存研究

標本に基づいて母集団を推測する統計的推測の際，実データの標本には欠損値や外れ値が存在するケースは少なくない．このうち外れ値については，外れ値が存在したデータに対してもロバスト性を確保できる統計的推測について，M 推定などの手法が知られている（藤澤（2006）, Maronna et al.（2006））．

欠損値処理についても，これまでに様々な処理方法が提案されている．欠損値処理については，特に遺伝子データなどの大量データを扱う分野における既存研究が多く, Aittokallio（2010）, Liew et al.（2011）, Moorthy et al.（2014）では，様々な欠損値補完方法に関する既存研究についてサーベイしている．

Moorthy et al.（2014）の分類に従えば，欠損値補完に用いる情報によって，グローバル・アプローチ(Global Approach)，ローカル・アプローチ(Local

Approach), ハイブリッド・アプローチ(Hybrid Approach), ノレッジ・アプローチ(Knowledge Approach)に分けられる. グローバル・アプローチは, 全データセットの相関情報を用いて補完値を推計するものと定義されているが Troyanskaya et al.（2001）の SVDimpute（Singular Value Decomposition）や Oba et al.（2003）の BPCA（Bayesian principal component analysis）が該当する. Liew et al.（2011）によれば, これらの方法は, 全データのグローバルな共分散構造が存在した場合や, データに局所的な類似構造が存在した場合, あまり正確な補完にはならないとされている. ローカル・アプローチは, データセットの一部分の類似性を利用した補完方法と定義されるが, 後述する Troyanskaya et al.（2001）の k-NN 法は, これに分類される. 他にも, Ouyang et al.（2004）の GMCimpute（Gaussian Mixture Clustering）, Sehgal et al.（2005）の CMVE(Collateral Missing Value), Kim et al.(2005)の LLSimpute（Local Least Squares formulation）, Zhang et al.（2008）の SLLSimpute（Sequential Local Least Squares formulation）, Sehgal et al.（2008）の AMVI(Ameliorative Missing Value Imputation), Burgette and Reiter(2010)の MICE-CART（Multiple Imputations by Chained Equations and Classifications by Regression Trees）などがローカル・アプローチに分類されている. ハイブリッド・アプローチは, グローバル・アプローチとローカル・アプローチの両方の特徴を有しているもので, Jörnsten et al.（2005）の LinCmb 法が該当する. ノレッジ・アプローチは, データに基づいて補完値を推定するプロセスに, その分野の専門知識や外部情報を活用するアプローチと定義される. Gan et al.(2006)の POCSimpute(Projection Onto Convex Set), Tuikkala et al.（2006）の GOimpute（Gene Ontology）, Xiang et al.（2008）の HAIimpute（Histone Acetylation Information Aided）などが該当する.

欠損値補完の方法論は, どのような情報を用いて補完するかという視点に立った上記の分類の他に, Little and Rubin（2002）において次のような 8 分類が行われている.

(1) mean imputation

欠損値を標本中の対応する集合の平均値で置き換える方法.

(2) Hot deck imputation

補完値が各欠損値の推定分布から選択される方法.

(3) Substitution

標本に含まれないレコードで既知の値を代替して置き換える方法.

(4) Cold deck imputation

外部の情報源から一定値を補完する方法. 例えば, 同じ種類の調査における前回調査の結果など.

(5) Regression imputation

観測された値から欠損値を回帰による予測値で補完する方法.

(6) Stochastic regression imputation

回帰分析による予測値に不確実性を加味して欠損値を補完する方法.

(7) Composite methods

異なる手法を合わせて使う方法.

(8) Multiple imputation methods

欠損値に複数値を補完する方法.

上記(1)～(7)は Single imputation に分類されるが, Single imputation と Multiple imputation の差異については, Rubin(1987)が詳しい. Rubin(1987)に従えば, 単一値代入法(single imputation)と多重代入法(multiple imputation : MI)に大別される. Rubin (1987) では, それぞれの特徴を次のように説明している. 単一値代入法は各欠損セルに対して, それぞれ一つの値を補完するものであり, この方法には二つの特長が存在する. 一つは, 補完後のデータセットに対して, 標準的な完全データで利用可能な分析手法をそのまま適用できる点にある. 二つ目は, データ収集者の知識に合致させた補完が可能である, という点である. 一般に, データ収集者はデータ分析者よりも, 欠損値が発生するメカニズムについての情報量が多い. データ収集者が欠損値補完を行うことで, データ分析者はより信頼感のあるデータを用いてより良い統計処理が可能となる. これに対して多重代入法(MI)は, 各欠損セルに対して,

複数の値を補完するものであるが，三つの特長を有する．一つ目はデータの分布に関する性質を保持できるという点である．二つ目は，多重代入法（MI）は繰り返しランダムに値を引き出しているため，頑健な推測値が得られるという点が挙げられる．三つ目は，完全データの手法を単純に繰り返し使うだけで，欠損に対する様々なモデル推定の感応度分析が容易になることである．本研究は，CRD という実用的なデータベースに対し，データ収集者が欠損値補完を行い，データ分析の前の初期データ整備を容易にする方法論を研究することを目的としているため，単一値代入法に焦点を当てている．

単一値代入法として最も基本的な方法は，欠損値セルに対して，その欠損値セルを含むフィールドの平均値や中央値を補完する方法である[6]．既存研究の多くで，この方法は比較対象の一つとして採り上げられている．ただし，この方法論では，状態の大きく異なる複数のレコード（観測値）における欠損値セルに対して同じ値が補完されるため，適切ではない．そこで，レコードの状態に応じた欠損値補完方法を考える必要がある．一般的なレコードの状態に応じた欠損値補完方法は，回帰分析である．しかし，回帰分析は，欠損フィールドが一つのみであれば良いが，多数のフィールドで欠損がある場合は，説明変数に欠損が含まれることとなる．そのような場合，欠損値に対して繰り返し回帰分析を行うことで順次補完を進めていく，ICE（Imputation by Chained Equations）という方法がある（Royston（2005））．本稿でも，ICE を比較対象の一つとしている．

この他にも，単一値代入法の一種として，金子（2005）では，ランダム補完法という欠損値補完方法を提示している．この論文では，欠損値補完法の優劣の決め方も特徴的である．すなわち，補完値と真値との誤差ではなく，補完後のデータに基づいて推計されたロジットモデルの，推計に用いられていない外部データに対する予測精度で決めている．比較対象として，削除法，平均値補

[6] レコードの平均値を補完する方法も考えられるが，財務諸表データの場合，フィールド毎の平均的な値が一般的に大きく異なるため，レコード平均値を補完することは行わない．

完法，メジアン補完法，k-NN 法の 4 手法が挙げられている．この論文では，ロバスト性という点でランダム補完法が最も優れているという結論であった．一方，k-NN 法について，論文中では Acuna（2004）を引用して，“いくつかの利点があるものの,距離のとり方に様々なバリエーションがあること,更に,「k」の設定により，この方法の特性が変わることから，取り扱いが非常に難しい”と紹介している．なお，データは実際の医療データを用いているが，この論文で提示された方法論は，財務諸表データに適用することが想定されている点で，問題意識について本稿と比較的近い．

2.1.3. 財務諸表データに対する欠損値補完に関する既存研究

財務諸表データを扱う論文のうち，多くのケースで上記のような欠損値と外れ値処理に関する問題が発生しているはずである．しかし，その処理方法を主眼とした論文は少ない．

その中でも，今井（2013）は，財務諸表データの共同データベースにおいてロジットモデルによる推計を行う場合に，多重代入法（MI）による欠損値補完方法が，欠損値が多くなるほど有効であることを示している．

宮本他（2012）では，欠損値を含むレコードを削除（deletion）したケース，時系列的な相関関係を考慮した前後期平均値を補完したケース，多重代入法（MI）により補完したケースの 3 種類の方法で財務諸表データに対し欠損値処理を施した上で，財務諸表データからデフォルト確率をロジットモデルにより推計している．その中では，前後期平均値補完のケースで最もロジットモデルのデフォルト予測精度が高いという結果となっている．本稿でも，前後期平均値補完による欠損値補完方法を，k-NN 法との有力な比較対象としている．

財務諸表データを扱うこの他の論文の中において，外れ値処理や欠損値処理に言及していることは稀である（高橋・山下（2002），山下・川口（2003））．多くの論文では，今井（2013）や宮本（2012）でも欠損値処理の方法の一つとして言及されているように,欠損値セルを含むレコードもしくはフィールド(財

務諸表項目）を削除（deletion）した上で分析を行っていると推測される．削除は欠損値処理の一つの方法ではあるが，削除によって分析用データにバイアスが生じる可能性がある（Rubin（1976），Little（1988），Allison（2001），岩崎（2010））．このようなバイアスの生じたデータによって信用リスク評価を行うためのモデリングを行うと，完全に誤った結果を導出することも考えうる．そこで，本研究では，欠損値を削除することを考えるのではなく，合理的に補完する方法について考察する．

2.1.4. k-NN 法に関する既存研究

2.1.2 節におけるほとんどの既存研究では，k-NN 法を何らかの形でデータの特性に合わせて発展させた計算方法を提示するか，もしくは，k-NN 法を新しい欠損値補完方法の最も有効な比較対象として参照している．そこで，本研究でも，財務諸表データ以外の研究分野でベースの欠損値補完方法として利用されている k-NN 法を財務諸表データに適用し，他の欠損値補完方法と比較して財務諸表データベースに対して有効な欠損値補完方法であることを示す．

k-NN 法に関する既存研究として，上記の金子（2005）の他には，Acuna（2004）が挙げられる．Acuna（2004）では，植物学，医学，経済学など 12 分野の実験用データベースを用いて，削除法（Case Deletion），平均値補完法（Mean Imputation），中央値補完法（Median Imputation），k-NN 補完法（k-NN Imputation）の 4 手法で判別分析を行い，その判別誤差を比較している．その中で，k-NN 法は，欠損率が上昇した時にも最もロバストな欠損値補完方法であり，全体として他の手法より良いという結論となっている．

Troyanskaya et al.（2001）は，遺伝子マイクロアレイの時系列データを用いた単一値代入法に関する代表的な論文である．この論文は，当該分野における初期の研究であり，3 つの欠損値補完方法（k-NN 補完法，SVD（Singular Value Decomposition）補完法，平均値補完法）を提案し，k-NN 補完法が平均値補完法や SVD 補完法と比較して，精度の良い欠測値補完方法であることが

示されている．また，欠損値補完の精度を確認する方法として，完全データから欠損値を発生させ，真値と補完値を比較する，という方法を採用している．

Troyanskaya et al.（2001）で採用された k-NN 法及び精度確認手法をもとに，遺伝子データの分野では，いくつかの派生的な補完方法及び精度確認手法が提案されている．例えば，Kim et al.（2004）では，k-NN 法を応用して，欠損値の少ないレコードの値から順に補完値として用いる SKNN（Sequential K-Nearest Neighbor）法を提案している．また，前述した Kim et al.（2005）の LLS-impute（Local Least Squares formulation）や Hsu et al.（2011）の KNN-DTW（dynamic time warping）など，k-NN 法と組み合わせて欠損値を補完する方法も提案されている．

遺伝子データ以外の分野では，田村他（2009）で，ソフトウェア開発の工程数予測に関するクロスセクションデータを用い，k-NN 法を応用した CF 応用法を，他の補完方法と比較して精度の良い方法として提案している．CF 応用法は，距離計算の際，ユークリッド距離ではなく，フィールド中央値からの乖離度を用い，補完値については，規模と類似度で重み付けがされている．Jönsson and Wohlin（2004）では，順位データに k-NN 法を適用している．ここでは，距離計算はユークリッド距離を利用しているが，補完値の計算方法を複数用意し，欠損率や K を変化させた時のシミュレーションを行っている．

なお，上述の k-NN 法を応用した手法については，外れ値の影響を受ける可能性が大きいこと，時系列データのみ，もしくはクロスセクションデータのみに適用可能な考え方であることなどの制約が存在する．外れ値が存在し，クロスセクションと時系列データの特徴を併せ持つ財務諸表データに対しては，上記諸論文の方法論をそのまま適用することは難しいことから，本研究では，最もロバストな方法論である，Troyanskaya et al.（2001）の方法論をベースに，財務諸表データに適合した欠損値補完方法を提案している．

本章の構成は次のようになっている．2.2 節では利用データの内容を説明し，そのデータに対して適用した k-NN 法について説明する．2.3 節では，k-NN 法による欠損値の補完精度が，平均値補完など，その他の補完方法と比較して，

どの程度の誤差となるのかについて検証する．2.4 節はまとめである．

2.2　利用データと k-NN 法の適用

2.2.1.　評価プロセス

財務諸表データに対する欠損値補完方法として k-NN 法を評価するプロセスについて説明する．本稿では，Troyanskaya et al.（2001）やそれに続く既存論文で採用されている評価プロセスと同様とした．以下に，そのプロセスを説明する．

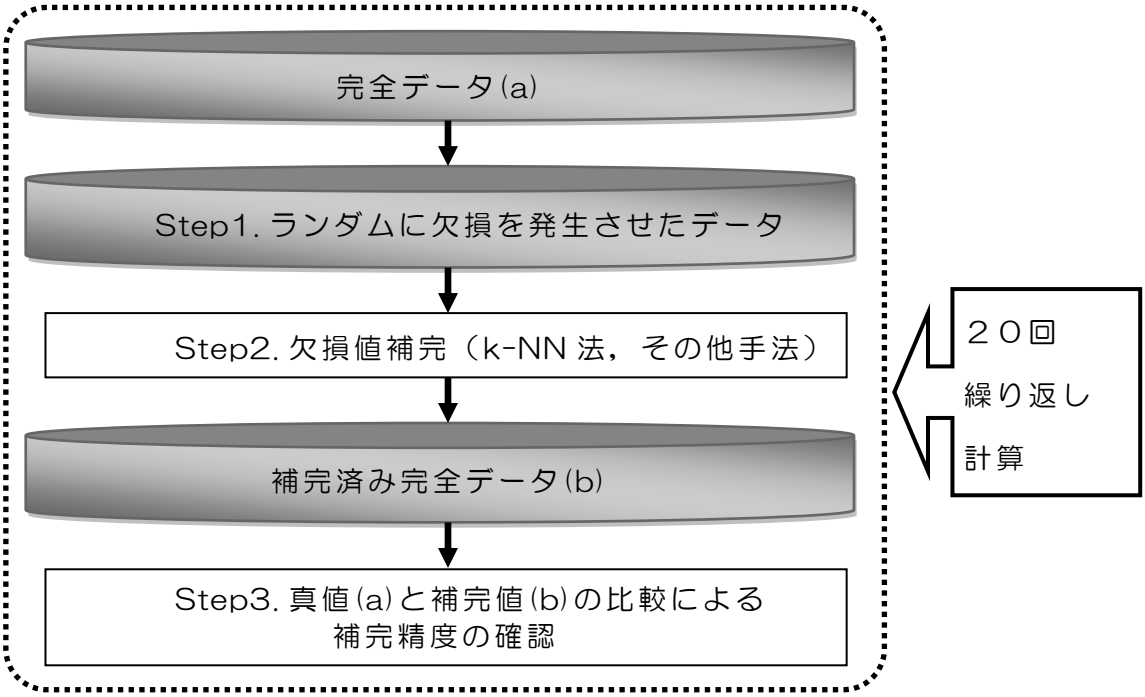
Step 1.　完全データから欠損値を発生させる．

Step 2.　k-NN 法により，欠損値を補完する．また，比較対象となる他の手法を用いた欠損値補完も行う．

Step 3.　Step2.において，k-NN 法や他の手法で補完された値と，Step1.で欠損値を発生させる前に実測された値を比較し，各手法間で欠損値と実測値の誤差の大きさを比較し，各手法の補完精度を評価する．

上記プロセスは，図 2-1 のように表される．その詳細について，2.2.2 節以降で説明する．

図 2-1 分析プロセス



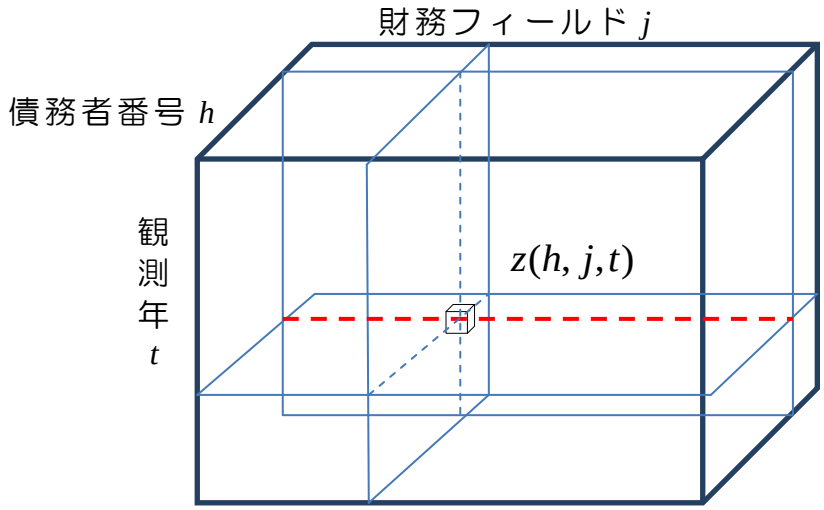
2.2.2 利用データと欠損値の発生

本研究では，まず CRD データから完全データを抽出した．この際，6 年連続$(t=1,2,3,4,5)$でレコード（財務諸表）が存在する債務者 h $(h=1,2,\cdots,H)$のデータで，主要 38 フィールド（財務諸表項目等）j $(j=1,2,\cdots,38)$が欠損でない完全データを抽出した．主要 38 フィールドは，表 2-1 のように，BS 項目 25 フィールド，PL 項目 9 フィールド，その他（脚注項目及び期末従業員数）4 フィールドである．これにより，時系列方向 t とフィールド方向 j の二次元で完全データとしている．したがって，完全データのあるセル $z(h,j,t)$は，図 2-2 のように h,j,tで定義される．

表 2-1 財務諸表主要 38 項目一覧

BS項目			PL項目	その他項目
流動資産合計	資産合計	負債合計	売上高･営業収益	受取手形割引高
現金･預金	流動負債合計	資本合計	売上原価･営業原価	受取手形裏書譲渡高
受取手形	支払手形	資本金	売上総利益	減価償却実施額
売掛金	買掛金	その他の資本	販売費および一般管理費	期末従業員数人
棚卸資産合計	短期借入金	負債･資本合計	営業利益	
その他流動資産合計	その他流動負債合計		受取利息･割引料･配当金	
固定資産合計	固定負債合計		支払利息･利子割引料	
有形固定資産合計	社債･長期借入金		経常利益	
土地	その他固定負債		当期利益	
繰延資産	長短借入金合計			

図 2-2 セル $z(h, j, t)$ の概念図



上記のように抽出された分析用完全データサンプルから，債務者数で 2 万件 $(H = 20000)$ を抽出した．同一債務者で 6 年連続でレコード（財務諸表）$i(h \times t \to i : i = 1, 2, \cdots, N)$ が存在するので，レコード数では 12 万件 $(N = 120000)$ となる．この時，デフォルト[7]債務者の財務諸表と非デフォルト債務者の財務諸表を，それぞれ同数の 6 万レコードずつランダム抽出している．このデフォルト債務者のレコード数は，主要 38 項目が完全データとなっている利用可能なデ

[7] なお，ここでのデフォルトは，CRD の財務諸表データベースの定義に従い，各レコードの決算年月から一年以内に 3 ヵ月以上延滞，実質破綻，破綻，代位弁済に該当した先としている．また，デフォルト債務者とは，6 年連続の財務諸表を有する債務者のうち，いずれかの年にデフォルトに該当した債務者としている．

フォルト債務者の財務諸表データとしては,ほぼ全数に近い数字である.なお,完全データの非デフォルト債務者については，デフォルト債務者の約 10 倍の件数を利用可能であるが，非デフォルト債務者の計算結果とデフォルト債務者の計算結果を比較する際の利便性を考慮して，デフォルト債務者と非デフォルト債務者のレコード数を同数としている．2.3.1 節では，上記のようなランダム抽出を 20 回繰り返して作成した，20 個のデータセットを使った結果を示している.

次に，この完全データから欠損値を発生させる必要がある．この際，今回利用した CRD データではデフォルト債務者ほど欠損セルが含まれる割合（欠損率）が高い，という特徴があったため，その特徴を取り入れて欠損値を発生させた[8]．すなわち，非デフォルト債務者のフィールド毎の欠損率は約 3%，デフォルト債務者のフィールド毎の欠損率は約 10%となるよう，乱数を用いて欠損値をランダムに発生させた．ここで，実際の財務諸表項目（フィールド）のうち，売上高は欠損値にならないことから，売上高については欠損値を発生させない．これにより，後に説明する，売上高を用いた効率的な計算方法を実現した[9]．本稿では，以上のような大規模な抽出データを用いて，様々な欠損値補完方法による補完精度がどのように異なるのかを観察する.

2.2.3 k-NN 法を用いた財務諸表補完方法

本節では，欠損値補完の分野で過去に多くの研究が存在し，本研究のベースである k-NN 法による欠損値補完方法について説明する.

k-NN 法では，まず，各レコード間の距離を，何らかの距離定義を用いて計算する必要がある．本研究では，k-NN 法に関する既存研究の多くで利用され

[8] 欠損率はフィールド毎に異なるが，本稿で利用した決算年が 2001 年から 2006 年のデータでは,デフォルト債務者の平均欠損フィールド率が非デフォルト債務者の平均欠損フィールド率を約 5%上回った．なお，この数字は，完全データ抽出前のデータベースにおける値である.
[9] 売上高の CRD データにおける項目名は売上高営業収益である.

ており，距離の定義として一般的なユークリッド距離を用いる．ただし，距離計算を行う前に，フィールドの値を標準化している．これは，財務諸表については，フィールド毎の値が大きく異なる[10]という特徴があるためで，これらの数値をそのまま距離計算に用いると，値の大きなフィールドほど距離に与える影響が大きくなる，という問題が発生するためである．また，k-NN 法の計算では，債務者単位の距離計算ではなく，財務諸表（レコード）単位で計算を行う．

これらを定式化すると，次のようになる．まず，レコード（財務諸表）$i(i \in I)$，フィールド（財務諸表項目）j のセルの数値を z_{ij}（i=1,2,…,N, j=1,2,…,38）とした時，標準化されたセル値を $x_{ij} = f(z_{ij})$ で表わす．ここで，N は分析で用いたレコード数，f は標準化の関数であり，（1）式に示す．

$$x_{ij} = f(z_{ij}) = \frac{z_{ij} - \overline{z_j}}{\sigma_j} \qquad (i=1,2,\cdots,N,\ j=1,2,\cdots,38) \qquad (1)$$

ただし，$\overline{z}_j$，σ_j は，それぞれセル z_{ij} の各レコード $i(i=1,2,\cdots,N$ に関するフィールド j の平均値及び標準偏差である．

次に，集合 I の中から任意の 2 つのレコード p と $q(p \neq q, p, q = 1,2,\cdots,N \in I)$ の間の距離として，標準化された値を用いて，ユークリッド距離 L_{pq} を計算する．通常は全 38 フィールドを用いて距離を計算するが，p もしくは q のレコードのどちらかの 38 フィールドのうちに欠損値が存在する場合には，両方のレコードで充足しているフィールドのみを用いて距離を計算している．この時，利用するフィールド数が多くなれば距離も大きくなる状況を避けるため，通常のユークリッド距離を，両方のレコードで欠損とならずに距離計算に利用したフィールド数 J によって修正される平均距離を用いる．

[10] 例えば，完全サンプルデータの売掛金の中央値は 61,537 千円であるが，資産合計は約 7 倍の 432,565 千円となっている．

$$L_{pq}=\left\{\frac{\sum_{j'=1}^{J_{pq}}\left(x_{pj'}-x_{qj'}\right)^2}{J_{pq}}\right\}^{\frac{1}{2}} \qquad (p\neq q, p,q=1,2,\cdots,N\in I, j'=1,2,\cdots,J_{pq}) \quad (2)$$

k-NN 法では次に，選ばれた距離の近い K 個のレコードを用いて欠損値を補完する．補完方法には様々な方法があるが，ここでは，相対ユークリッド距離の逆数で加重平均した値を用いる．この方法は，金子（2005）や Crookston and Finley（2008）で利用されている．（3）式で定義されるレコード i のフィールド j の補完値 $\hat{x}_{ij}$ には，i との距離が近いレコード K 個の標準値の加重平均値を用いる．この加重値は，相対ユークリッド距離の逆数である．K の値は，理論的な根拠は存在しないので，分析者で与えることが一般的であるが，本研究では 1～8 を試行している．なお，K 個に含まれるレコードのフィールド j が欠損である場合には，計算には用いていない[11]．

$$\hat{x}_{ij}=\sum_{k=1}^{K}x_{kj}\left(\frac{1/L_{ik}}{\sum_{k=1}^{K}\left(1/L_{ik}\right)}\right), \quad \left(k\neq i, i=1,2,\cdots,N\right) \qquad (3)$$

この後，標準化されていた値 $\hat{x}_{ij}$ を，（4）式のように原フィールドの分布に引き戻し，原フィールドの欠損値 $\hat{z}_{ij}$ を補完する．

$$\hat{z}_{ij}=f^{-1}(\hat{x}_{ij})=\hat{x}_{ij}\cdot\sigma_j+\overline{z}_j \qquad (i=1,2,\cdots,N,\ j=1,2,\cdots,38) \qquad (4)$$

最後に，試行した補完方法の評価指標として，補完値と削除された真値の誤差を計算する．この際に用いる指標は，Troyanskaya et al.（2001）を始め，

[11] ここで，K=1 の時，（2）式で定義されるレコード間のユークリッド距離が 0 となる場合，補完に利用するレコードの加重値は 1 とする．また，K=2 以上でユークリッド距離が 0 となるレコードが 1 つ存在した場合，0 のレコードに対して加重値は 1 として，その値のみ利用する．さらに，K=2 以上でユークリッド距離が 0 となるレコードが複数存在した場合，距離 0 以外のレコードは利用せず，距離 0 のレコードの加重値を 1 として，それらを単純平均する．

完全データから欠損値を発生させる方法での実験を行ったほとんどの論文で採用されていた，真値と補完値との標準平均平方誤差（normalised root mean squared error：NRMSE）を採用する．これを定式化すると，次のようになる．まず，欠損セルの集合を M とすると，欠損セル(i, j)は，$(i, j) \in \mathrm{M}$で表わされる．また，完全データにおける欠損セルに対応する真の値をz_{ij}^{*}とすると，補完値$\hat{z}_{ij}$，フィールドz_{ij}の標準偏差σ_j，欠損セル数 n を用いて，NRMSE は（5）式のように計算される．

$$\mathrm{NRMSE} = \sqrt{\frac{\sum_{(i,j)\in \mathrm{M}} \left(\frac{\hat{z}_{ij} - z_{ij}^{*}}{\sigma_j} \right)^2}{n}} \qquad (5)$$

2.2.4　完全フィールドを用いた効率的計算方法

k-NN 法では，あるレコードに対して，他の全てのレコードとの距離を計算して，そのうち距離の近い K 個のレコードを，欠損値補完に利用するレコードとして選ぶことになる．しかし，レコード数が多くなると，総当たりで距離を計算することは，計算コストの面から現実的ではない．（2）式の距離計算を行う際の組み合わせ計算量 X は，レコード数を x とした時，$X = (x/2) \times (x-1)$ で表わされる．従って，12 万レコード全件の組み合わせを計算する場合には，7.2×10^9 回の距離計算が必要とされる．

そこで，今回の分析では，売上高を基準にして，売上高が類似しているレコードのみを距離計算対象とすることで，計算効率を大幅に向上させた．CRD データにおいて，売上高が欠損値となることはない．現実的にも，売上高は財務諸表の基礎になる数字であることから融資を受ける際に必ず必要となる情報であり，欠損となることは考えにくく，売上高という完全フィールドを基準にしてランク分けすることが必ず可能となる．また，一般的に，企業規模によって

財務諸表の内容は類似しており，規模の一つの指標が売上高である．

売上高を基準にした計算方法は，次のようなものである．まず，売上高順にR個のランクを設ける．次に，あるランク$r(r=1,2,\cdots,R)$に属するレコードに対しては，同じランク内及びその隣接する上下のランク（$r-1$及び$r+1$）のレコードのみを計算対象とする．ただし，$r=1$に属するレコードの場合は，$r=1$及び$r=2$が計算対象となり，$r=R$に属するレコードは，$r=R$及び$r=R-1$を計算対象とする．ここで，1ランクに含まれるレコード数は，300レコードとした．これは，サンプル数3000レコード及び6000レコードで実験をしたケースにおいて，1ランクに300レコード以上のレコード数が確保されている場合，全件を対象に距離計算を行ったケースとほぼ同じ真値と補完値の誤差精度（NRMSE）であり，それよりも1ランク内のレコードが減少すると，誤差が大きくなることが確認されたからである．確認結果は表2-2及び図2-3に示している．

表 2-2 ランク分け法による NRMSE

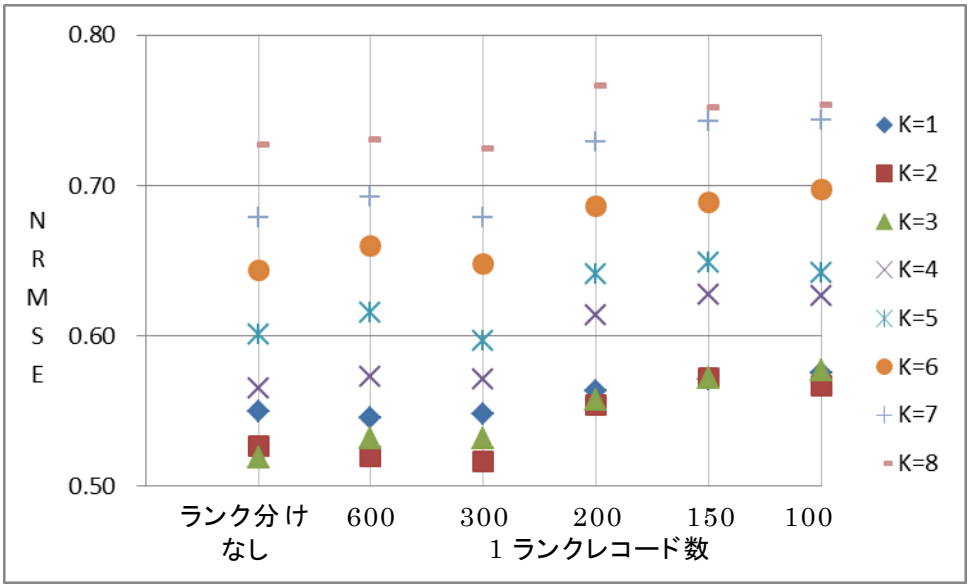
（上段：3000 レコードのケース，下段：6000 レコードのケース）

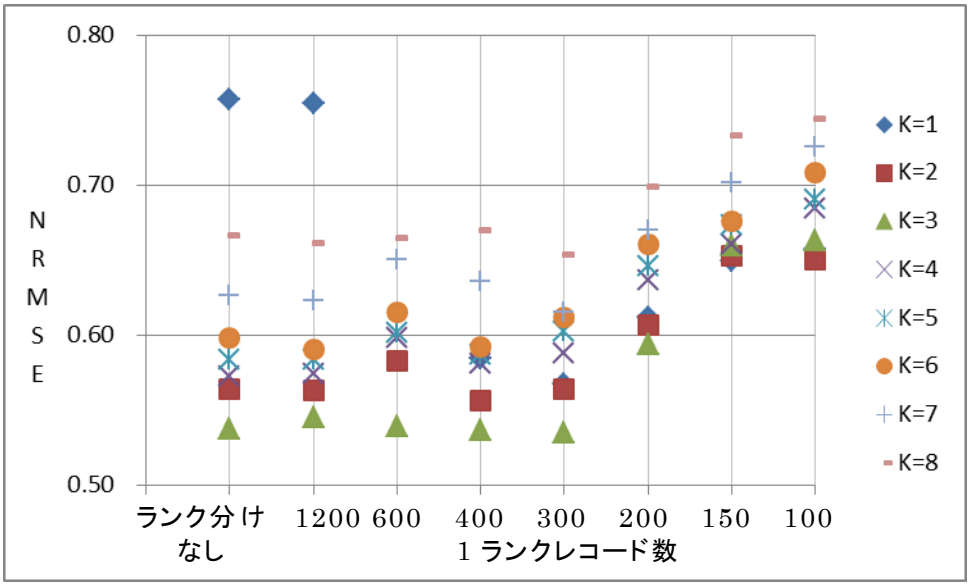
	ランク分けなし	1ランクレコード数				
		600	300	200	150	100
K=1	0.5501	0.5456	0.5482	0.5630	0.5713	0.5753
K=2	0.5263	0.5201	0.5166	0.5537	0.5719	0.5671
K=3	0.5186	0.5315	0.5321	0.5577	0.5719	0.5772
K=4	0.5654	0.5732	0.5711	0.6134	0.6278	0.6265
K=5	0.6013	0.6152	0.5964	0.6411	0.6491	0.6417
K=6	0.6438	0.6597	0.6476	0.6865	0.6887	0.6976
K=7	0.6786	0.6925	0.6785	0.7291	0.7424	0.7434
K=8	0.7271	0.7306	0.7244	0.7661	0.7518	0.7537

	ランク分けなし	1ランクレコード数						
		1200	600	400	300	200	150	100
K=1	0.7569	0.7542	0.8123	0.5850	0.5675	0.6122	0.6493	0.6569
K=2	0.5644	0.5631	0.5833	0.5569	0.5638	0.6071	0.6529	0.6504
K=3	0.5380	0.5453	0.5394	0.5366	0.5351	0.5940	0.6601	0.6630
K=4	0.5727	0.5742	0.5983	0.5815	0.5882	0.6364	0.6606	0.6847
K=5	0.5834	0.5840	0.6015	0.5873	0.6025	0.6464	0.6738	0.6903
K=6	0.5983	0.5906	0.6150	0.5920	0.6122	0.6604	0.6761	0.7086
K=7	0.6262	0.6228	0.6501	0.6357	0.6150	0.6702	0.7013	0.7254
K=8	0.6664	0.6613	0.6650	0.6700	0.6542	0.6991	0.7335	0.7440

図 2-3　ランク分け法による NRMSE のグラフ

（上段：3000 レコードのケース，下段：6000 レコードのケース）



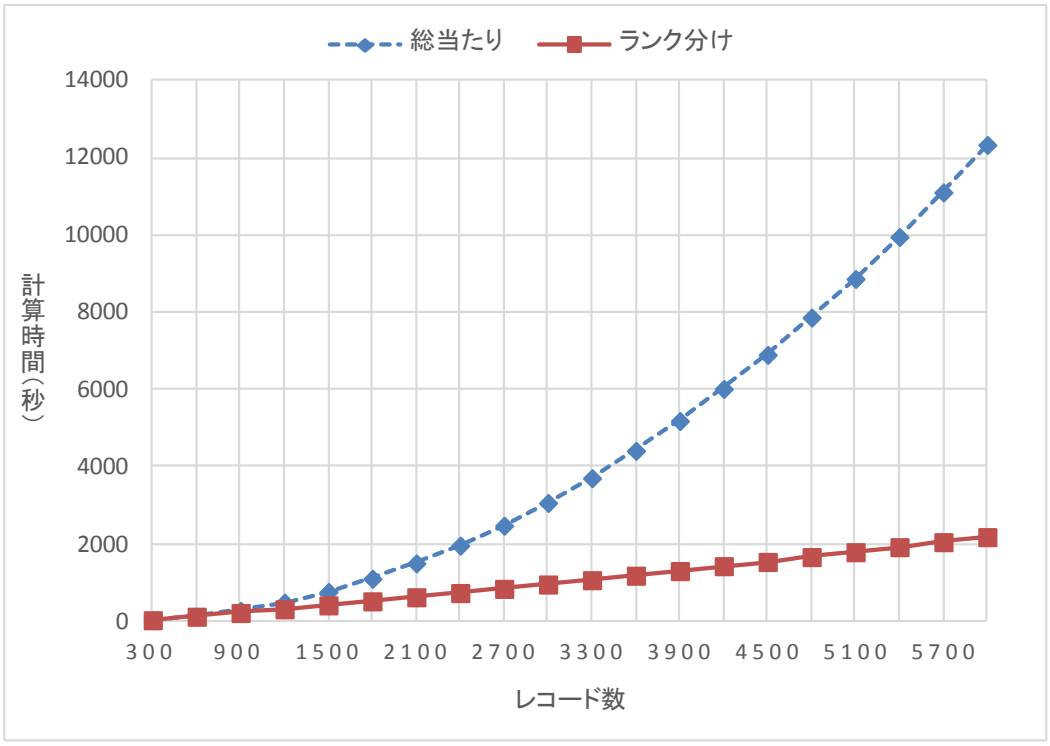


この計算方法であれば，総当たりで距離計算を行うケースのような組み合わせ数が比例級数的に増加するという問題は解消される．本稿の 12 万件のケースでは 400 ランク ($r = 400$に分けられるが，この売上ランクによる効率的な計算方法を用いれば，5.4×10^7 回の計算量で済む．総当たりの計算量である 7.2×10^9 回と比較すると，実に約 99.3%の計算量を削減することができた．この

計算ロジックは，レコード数が大きくなればなるほど計算量の削減効果が大きい方法論であり，他の分野においても，大規模データに対して k-NN 法を適用する際には，参考となる方法論ではないかと考えられる．

実際に売上ランクによる効率的な計算方法（ランク分け法）の計算時間を，総当たりでの距離計算方法（総当たり法）の計算時間と比較したものが，図 2-4 である[12]．図 2-4 では，レコード数 300 から 6000 までのデータセットに対して，各方法の計算時間を計測した結果をプロットしている．図 2-4 からは，レコード数の比例的な増加に伴って，総当たり法の計算時間がほぼその自乗に比例して増加する一方，ランク分け法の計算時間は比例的にしか増加していないことが確認された．この計算時間の増加傾向を 120000 レコードに適用すると，総当たり法の計算時間は 4.8×10^{6} 秒となるが，ランク分け法の計算時間は 4.3×10^{4} 秒となり，計算時間を大幅に短縮できることが確認された．

図 2-4．総当たり法とランク分け法の計算時間比較



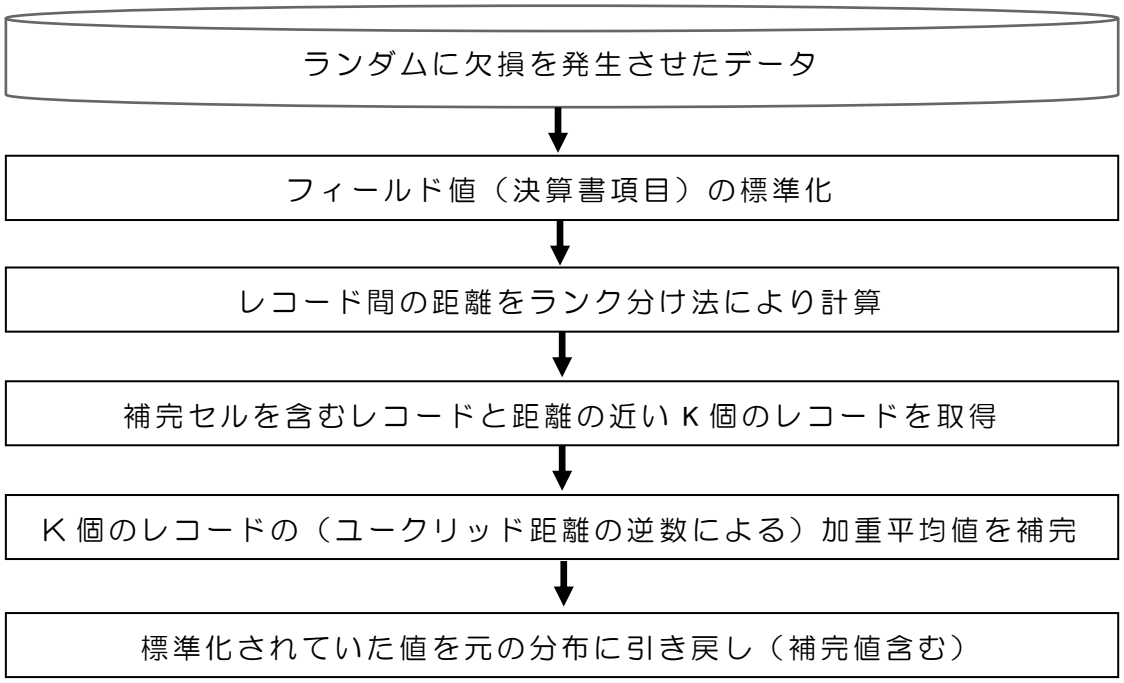
[12] 計算には，Intel® Core™ i7 CPU 2.80GHz，RAM24.0GB のスペックの PC を用いている．

（単位:秒）

レコード数	総当たり		ランク分け	
	計算時間	差分	計算時間	差分
300	31	-	31	-
600	121	90	122	91
900	274	153	218	96
1200	487	213	315	97
1500	764	277	416	101
1800	1105	341	520	104
2100	1503	398	626	106
2400	1964	461	735	109
2700	2479	515	843	108
3000	3068	589	956	113
3300	3710	642	1068	112
3600	4418	708	1184	116
3900	5190	772	1304	120
4200	6021	831	1431	127
4500	6904	883	1541	110
4800	7869	965	1667	126
5100	8870	1001	1793	126
5400	9957	1087	1923	130
5700	11111	1154	2053	130
6000	12341	1230	2183	130

以上のような効率的な計算方法を採用した k-NN 法による欠損値補完手順をまとめると，図 2-5 のようになる．

図 2-5. k-NN 法による欠損値補完の手順



2.2.5 比較対象となる欠損値補完方法

k-NN 法による CRD データに対する欠損値補完方法が，他の欠損値補完方法と比較して，相対的にどの程度誤差が小さくなるかを確認するため，フィールド平均値の補完法，ICE（Imputation by Chained Equations）を用いた回帰分析による補完法，時系列補完法（同一債務者の異なる決算期レコードによる補完法）の，3 通りの補完方法と比較した．

フィールド平均値の補完法としては，原数字の算術平均値 $\hat{z}_{ij}^{\,avg}$ と原数字の neglog 変換後の算術平均値 $\hat{y}_{ij}^{\,ngav}$ を用いている[13]．各フィールドにおいて欠損していないセルの平均値を，当該フィールドにおいて欠損している全てのセルに補完する．すなわち，フィールド j における欠損セルの数を $m_j\left(m_j \le N\right)$ とすると，$\hat{z}_{ij}^{\,avg}$ は （6）式のように表わされる．

$$\hat{z}_{ij}^{\,avg} = \frac{\sum_{i=1}^{N-m_j} z_{ij}}{N-m_j}, \quad \left(m_j \le N\right) \tag{6}$$

また，neglog 変換後のセル値 y_{ij} は（7）式のように表されることから，neglog 変換後の算術平均値 $\hat{y}_{ij}^{\,ngavg}$ は（8）式のように表される．NRMSE を計算する際の補完値 $\hat{z}_{ij}^{\,ngav}$ には，(8)式の neglog 逆変換した(9)式の値を使う．

$$y_{ij} \equiv \mathrm{neglog}\left(z_{ij}\right) = \begin{cases} \log_{10}\left(1+z_{ij}\right) & \text{if } z_{ij} \ge 0 \\ -\log_{10}\left(1-z_{ij}\right) & \text{otherwise} \end{cases} \tag{7}$$

$$\hat{y}_{ij}^{\,ngavg} = \frac{\sum_{i=1}^{N-m_j} y_{ij}}{N-m_j}, \quad \left(m_j \le N\right) \tag{8}$$

$$\hat{z}_{ij}^{\,ngavg} = \mathrm{neglog}^{-1}(\hat{y}_{ij}^{\,ngavg}) \tag{9}$$

ICE による補完法では，欠損セルを含む複数のフィールドが存在する場合に，

[13] neglog 変換については，森平（2009）参照．

各フィールドに関する回帰方程式を連鎖的に生成し，欠損セルをその回帰式に基づいて補完する[14]．ICE による補完値を $\hat{z}_{ij}^{ice}$，レコード i の j 以外のフィールド値ベクトルを y_i，連鎖回帰式による係数ベクトルを $\hat{\beta}$ とすると，（10）式のような推計式を連鎖的に計算して，補完値が作成される．

$$\hat{z}_{ij}^{ice} = y_i \bullet \hat{\beta}, \qquad (i=1,2,\cdots,N,\ j=1,2,\cdots,38) \qquad (10)$$

時系列補完法は，同一債務者であれば，財務諸表データは時系列で大きな変動をすることが少ない，という点に着目した補完方法である．例えば，資産合計が 1 億円から 10 億円に急増するケースは稀であり，前後期の情報を用いることで，ある程度合理的な補完値を推定することが可能となる．この場合，同一債務者で観測年月が近接しているレコードは，決算項目の数値が類似しているということを根拠にしている．今回は，当該レコードから過去に 1 決算年分遡る前期値補完と，当該レコードの前後の決算年レコードを単純平均した値を利用する前後期平均値補完の 2 パターンで計算を行っている[15,16]．なお，財務諸表データには，同一債務者であれば，時系列で見た時に，同じフィールドが欠損となる傾向があるため，現実的にはこの方法を採用することは難しい．すなわち，前期値補完や前後期平均値補完は，本研究で用いる人工欠損データのように，あるフィールドでランダムに欠損を発生させた場合にのみ，利用可能な方法と言える．

---

[14] MICE については，Van Buuren. et al.（1999）参照．

[15] 1 決算年遡ったセル値も欠損である場合，更に過去に遡った直近のセル値を用いる．過去に非欠損セルが無い場合は，当該決算年後の直近セル値を用いる．

[16] 前期セルもしくは後期セルが欠損セルであった場合，ないしはその両方が欠損セルであった場合，それぞれ直近の欠損でない 2 期分のセル値の平均値を補完値とする．1 債務者の中で 2 期分の非欠損セル値が揃わなかった場合は，前期値補完と同様の補完方法を採用する．

2.3. 実データを用いた補完精度の検証結果

2.3.1 k-NN 法と他の方法の精度比較

計算は，完全データから抽出した 120,000 レコードを含む 20 個のデータセットに対して，2.2.2 節で説明した方法で欠損セルをランダムに発生させて行った．各データセットに対して，2.2.2 節及び 2.2.3 節で説明した k-NN 法について，K =1 から K =8 まで変えた時に，NRMSE がどのように変化するかを確認したものが図 2-6 である．比較対象として，算術平均値補完，neglog 化平均値補完，ICE，前期値補完，前後期平均値補完も示している．計算結果は，20 個のデータセットの結果を図 2-6 において分布で示している．また，表 2-3 では，図 2-6 で示した数値と合わせて，平均値も示している．

図 2-6　補完方法別 NRMSE（12 万レコード 20 データセット結果分布）

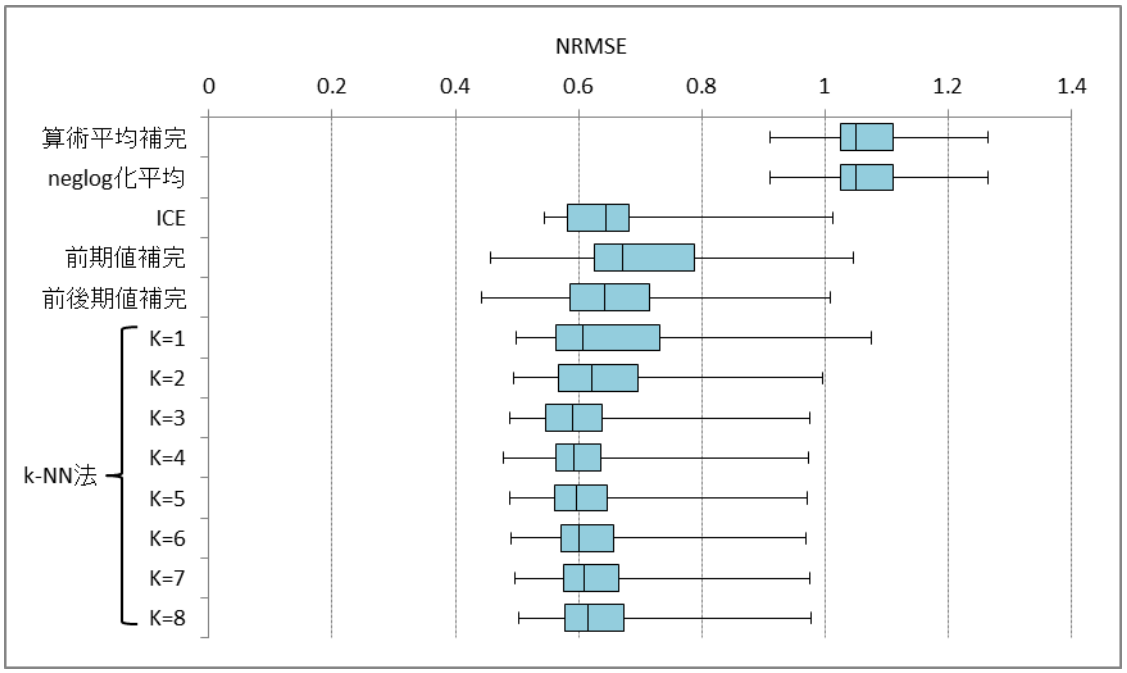


表 2-3　補完方法別 NRMSE（12 万レコード 20 データセット結果分布）

	算術平均補完	neglog化平均	ICE	前期値補完	前後期値補完	k-NN法							
						K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8
最大値	1.263	1.263	1.013	1.045	1.009	1.074	0.996	0.975	0.974	0.971	0.969	0.975	0.978
75%点位	1.111	1.111	0.682	0.788	0.714	0.731	0.696	0.637	0.637	0.646	0.657	0.666	0.672
中央値	1.050	1.050	0.644	0.671	0.643	0.607	0.622	0.590	0.592	0.596	0.601	0.608	0.615
25%点位	1.025	1.025	0.582	0.626	0.587	0.563	0.566	0.547	0.564	0.562	0.572	0.576	0.577
最小値	0.911	0.911	0.544	0.457	0.443	0.500	0.494	0.488	0.479	0.489	0.491	0.497	0.503
平均値	1.076	1.076	0.684	0.721	0.673	0.666	0.654	0.638	0.641	0.643	0.648	0.654	0.660

図 2-6 及び表 2-3 の結果を見ると，k-NN 法の K =3 の時の NRMSE の水準が 0.638 となっており，他の欠損値補完方法と比較して最も低くなっていることが分かる．次に NRMSE が小さいケースは，K =4, K =5 と続き，比較対象となる補完方法の中では，前後期平均値補完が k-NN 法の次に NRMSE の小さい方法となっている．時系列補完は，k-NN 法と比較して相対的に NRMSE が大きい欠損値補完方法という結果となった．このことは，前後期平均値補完のように，同一債務者の時系列情報を利用するだけでなく，他の類似した財務諸表情報を持つレコードを参照することで，時系列方向に変動が大きい場合でも，相対的に誤差の小さい欠損値補完が可能となることを示している．

なお，表 2-4 では，20 個のデータセットの平均 NRMSE が最も小さい K=3 の時の k-NN 法と，k-NN 法以外の補完方法の中で最も平均 NRMSE が小さい前後期値補完とを比較して，両者の NRMSE の差が有意な差であることを，t 検定で示している．各データセット d に対して，k-NN 法 K=3 の NRMSE を $\mathrm{NRMSE(k\text{-}NN}_{K=3})_d$，前後期値補完の NRMSE を $\mathrm{NRMSE(前後期値補完)}_d$ とした時，その差分を $\mathrm{diff}_d = \mathrm{NRMSE(前後期値補完)}_d - \mathrm{NRMSE(k\text{-}NN}_{K=3})_d$ で表わす．帰無仮説は $\mathrm{diff}_d > 0$ であり，t 値は(11)式のように計算され，95%の信頼区間で帰無仮説は棄却された．

$$t = \frac{\mathrm{E(diff}_d)}{\mathrm{std(diff_d)}/\sqrt{20}} = \frac{\left(\sum_{d=1}^{20} \mathrm{diff}_d / 20\right)}{\mathrm{std(diff_d)}/\sqrt{20}} \tag{11}$$

ここで，$\mathrm{std(diff_d)}$ は 20 個のデータセットに関する NRMSE の差分の標準偏差を表わす．

表 2-4　k-NN 法 *K*=3 と前後期値補完の NRMSE の有意差確認

	20データセット		t値
	平均 $E(diff_d)$	標準偏差 $std(diff_d)$	
NRMSE差分 （前後期値補完 － $k\text{-}NN_{K=3}$）	0.034	0.073	2.1157

その他の結果を確認すると，前期値補完よりも前後期平均値補完の NRMSE が小さくなっている．これは，同一債務者の情報だけを使う場合，1 期分よりも 2 期分の情報を使った方が，相対的に正確な値を補完できることを示している.ICE は前期値補完よりも NRMSE は小さいが,前後期値補完には及ばない．算術平均補完や neglog 化平均値補完の NRMSE は，その他の補完方法と比較して，明らかに大きな誤差（NRMSE）となっていることが分かる．これは，算術平均補完及び neglog 化平均値補完は,各レコードに含まれる情報を全く用いずに，単一の値を入れていることが原因と思われる．

2.4. まとめと課題整理

本章の結果から，遺伝子データや工程数予測などの欠損値補完に用いられてきた k-NN 法は，大規模な財務諸表データについても，十分な有効性が確認された．欠損値を含む財務諸表データに k-NN 法を適用することにより，欠損値補完の一般的な方法である平均値補完や，連鎖的な回帰方程式による欠損値補完方法である ICE，同一債務者の時系列データによる補完よりも，真値と補完値の誤差を小さくすることが確認された．欠損値を含む現実のデータに対して補完する際も，同一債務者の情報は使えないケースが多いと考えられることから，k-NN 法や ICE に次ぐ安定的な欠損値補完方法である時系列補完は現実のデータに対する活用可能性は低く，本章で提示した k-NN 法による補完方法の有効性は高いものと考えられる．

本章の特徴として，特に大規模なデータの k-NN 法を計算する時に有効となる,売上高ランクを導入した効率的な計算方法について提示した.この方法は，売上高のように完全フィールドを想定できるフィールドが存在する場合には，財務諸表以外の他のデータでも応用可能であると考えられる.

本章の研究内容に関する課題として，欠損値が存在する財務諸表情報を用いた信用リスク評価に対して，この欠損値補完方法がどのように影響を及ぼすかを確認する点が挙げられる．その際，k-NN 法により欠損値を補完した場合とその他の欠損値処理方法を比較し，二項ロジットモデルの AUC などの予測精度がどの程度異なるのかを確認する必要がある．欠損値を含む財務諸表のリスク評価をどのように行うかは，非常に重要な問題である．仮に，本章で提示した k-NN 法により欠損データを補完したことにより，信用リスクモデルの予測精度が上昇すれば，信用リスク計量化の前進に大きな貢献となる．この取り組みについては，本研究の第 4 章で考察する.

この他に，本章の結果が一般的に成立するかどうかについては，CRD データ以外の外部データに対する有効性についても確認する必要がある．また，遺伝子研究の分野では，k-NN 法を異常値修正に応用しているケースがある．この点についても,財務諸表データでの有効性を確認する必要がある.第 4 章では，この点についても応用した分析を行う.

第 3 章：業種情報を利用した k-NN 欠損値補完法の精度向上の試み

3.1 はじめに

本章では，第 2 章の分析を発展させ，業種区分情報を利用して，補完値と真値の誤差がさらに小さい欠損値補完方法の開発に向けた考察を行っている．業種分類については，CRD 業種コードをベースとして，各セグメントにある程度のレコード数が存在するよう，表 3-1 のような 9 区分のセグメント $S(S=1,2,\ldots,8,9)$ を新たに作成し，各レコード i を 9 区分のいずれかに分類した．この業種分類は，3.2 節，3.3 節で共通である．

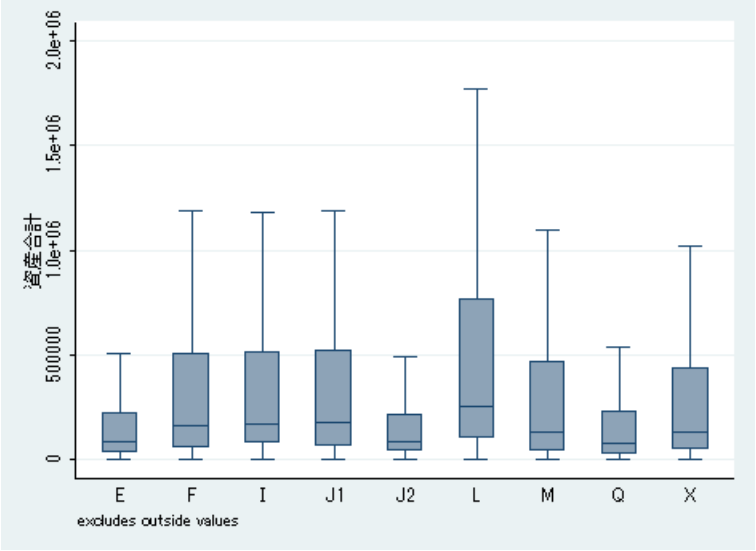
表 3-1 業種分類とレコード数

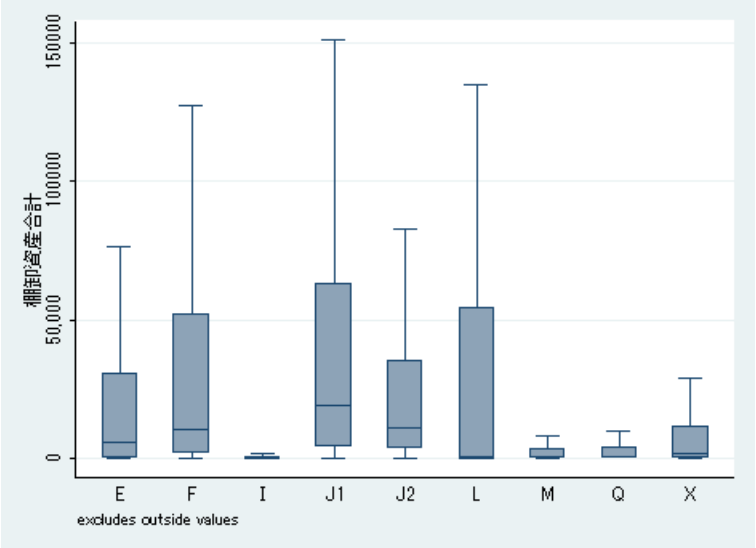
No.	Code	業種分類	レコード数	構成比(%)
1	E	建設業	30,337	25.28
2	F	製造業	26,655	22.21
3	J2	小売業	15,982	13.32
4	J1	卸売業	15,764	13.14
5	Q	サービス業	11,988	9.99
6	I	運輸業	5,512	4.59
7	M	飲食店，宿泊業	5,476	4.56
8	L	不動産業	5,025	4.19
9	X	その他	3,261	2.72
計			120,000	100

一般的に，業種によって平均的な売上高や自己資本比率の水準感は大きく異なるように，BS/PL における各財務項目の財務諸表項目の数値的構造は業種による特性が出やすい．図 3-1 には，本章の分析で用いたデータに関し，業種別にみた主要財務項目（資産合計，棚卸資産合計，売上高営業収益）の箱ひげ図を，表 3-2 にはその分位点の数値を示している[17]．

[17] その他の財務項目については，付録 B 参照．

図 3-1 業種別主要財務項目の箱ひげ図





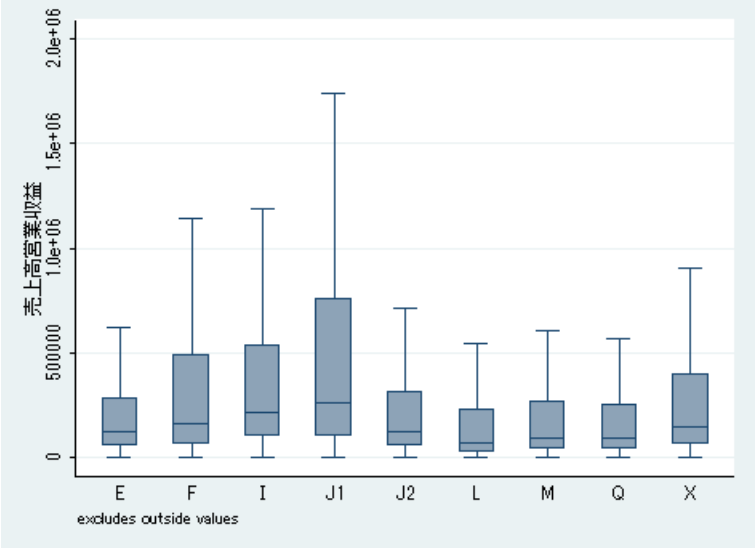


表 3-2 業種別主要財務項目の分位点

資産合計

No.	Code	業種分類	5%点位	25%点位	中央値	75%点位	95%点位
1	E	建設業	11,400	37250	89300	227000	1050000
2	F	製造業	13,900	56300	165000	511000	2590000
3	J2	小売業	21,200	77700	174000	519000	2570000
4	J1	卸売業	17,300	66700	181000	520000	2660000
5	Q	サービス業	11,700	38300	86500	221000	1120000
6	I	運輸業	17,800	103000	258000	772000	3590000
7	M	飲食店，宿泊業	10,400	40300	136000	467000	2020000
8	L	不動産業	8,530	29100	76900	235000	1560000
9	X	その他	12,500	49550	134000	441500	2700000
全体			12,300	45500	120000	354000	1940000

棚卸資産合計

No.	Code	業種分類	5%点位	25%点位	中央値	75%点位	95%点位
1	E	建設業	0	500	5920	30900	208000
2	F	製造業	0	1830	10700	52200	342000
3	J2	小売業	0	0	0	700	9870
4	J1	卸売業	0	4120	19200	63200	304000
5	Q	サービス業	200	3440	11400	35300	177000
6	I	運輸業	0	0	0	54300	853000
7	M	飲食店，宿泊業	0	300	1000	3450	13000
8	L	不動産業	0	0	700	4040	39300
9	X	その他	0	200	1865	11750	124500
全体			0	500	5470	30800	229000

売上高営業収益

No.	Code	業種分類	5%点位	25%点位	中央値	75%点位	95%点位
1	E	建設業	21,300	60600	126000	285000	1110000
2	F	製造業	17,700	61900	161000	495000	2470000
3	J2	小売業	35,600	102000	217000	540000	2190000
4	J1	卸売業	26,100	102000	263000	760000	4190000
5	Q	サービス業	17,600	55000	127000	318000	1690000
6	I	運輸業	6,160	26000	73400	233000	1360000
7	M	飲食店，宿泊業	15,800	43900	98800	268000	845000
8	L	不動産業	12,400	40800	93500	254000	1270000
9	X	その他	18,850	63750	145000	401500	1900000
全体			17,300	58750	141000	375000	1950000

図 3-1 及び表 3-2 から，運輸業(I)，飲食店・宿泊業(M)，サービス業(Q)は，資産規模（資産合計）に対する棚卸資産の規模（棚卸資産合計）が，他の業種と比較して相対的に小さいことが確認できる．また，不動産業(L)及び飲食店・宿泊業(M)は，資産規模（資産合計）に対する売上規模（売上高営業収益）が，他の業種と比較して相対的に小さくなっている．これらの二つの財務指標だけを見ても，平均的な傾向が業種によって大きく異なることが確認された．

このような業種による平均的な財務指標の違いは，k-NN 法においては，財務指標で計測する各レコード（財務諸表）間の距離に影響を及ぼすと考えられる．そこで，本章ではその業種情報を利用することを考える．すなわち，財務諸表に表われている数値を効率よく距離計算に反映させる情報として業種区分情報を考慮して，より正確な欠損値補完を目指すものである．

本研究では，上記のような業種情報の利用方法について，2 通りの方法で考察した．1 つ目が，業種をセグメント化し，同一セグメントからの情報のみを用いて k-NN 法による補完を行う方法である．この方法については，3.2 節で説明する．2 つ目が，レコード間の距離計算の際，同一業種の距離を短縮化する方法である．この方法については，3.3 節で説明する．3.2 節の方法と 3.3 節の方法の相違については，3.2 節では他業種セグメントのレコードを補完用情報として一切利用しない一方，3.3 節では他業種セグメントのレコードを補完用情報として利用する可能性は低くなるのみ，という整理になる．

3.2 業種をセグメント化した補完精度向上の試み

3.2.1 分析手法に関する第 2 章との相違点

一般的に，製造業やサービス業などの業種により，財務諸表項目の数値的構造は異なる．例えば，製造業では棚卸資産が多く，サービス業では小さい．この点を考慮すると，k-NN 法において，同一業種の財務諸表を重視して補完する方が，単純に全ての業種を均等に距離計算して補完するよりも，適切である

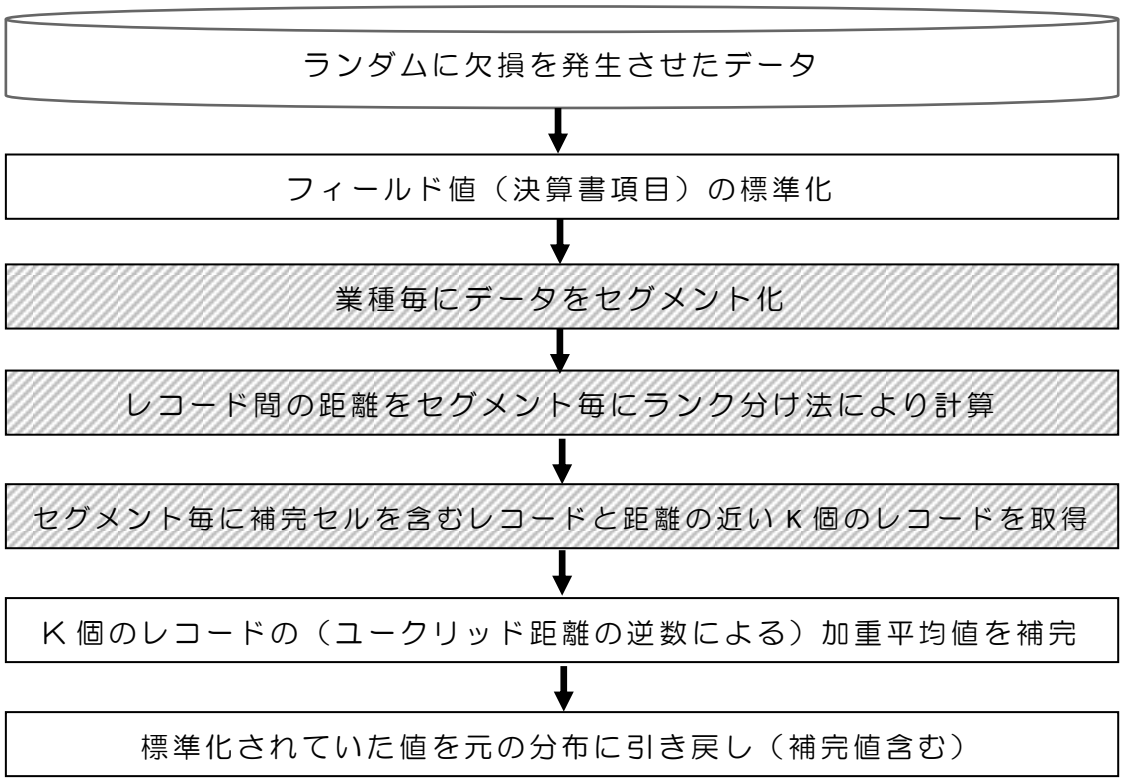
と推測される．そこで，本節では，業種分類を用いてセグメント化を行い，同一セグメントからの情報のみを用いて k-NN 法による補完を行い，補完精度が改善するかどうかを確認した[18].

分析手順及び売上高区分を用いた効率的計算方法については第 2 章と同様であるが，業種をセグメント化し，セグメント内でのみ距離計算を行った点が第 2 章と異なる．すなわち，第 2 章の（3）式で表わされる補完値は，（3’）式のように，同一セグメント内のレコードから作成されることになる．

$$\hat{x}_{ij} = \sum_{k=1}^{K} x_{kj} \left(\frac{1/L_{ik}}{\sum_{k=1}^{K} (1/L_{ik})} \right), \quad (k \neq i, i = 1,2,\cdots,N \in S, S = 1,2,\ldots,8,9) \qquad (3')$$

その手順を図 2-4 に準じて表すと，図 3-2 のようになる．

図 3-2．k-NN 法による欠損値補完の手順（斜線部分が図 2-4 からの変更点）



[18] 本研究とは異なる問題意識であるが，山下・川口（2003）では，CRD データを用いて業種分類でセグメント化した場合，信用リスク計測モデルの AR（Accuracy Ratio）で計測した精度が向上することを示している．

3.2.2　分析結果

表 3-3 では，業種セグメントを用いた k-NN 法による補完値と真値との NRMSE を,業種セグメントを用いずに計算した第 2 章の結果と比較している．この結果を見ると，業種セグメントを用いた k-NN 法の方が，NRMSE で見た誤差が明らかに大きくなっている．

表 3-3　業種セグメント導入前後の k-NN 法の NRMSE

		オリジナル	業種セグメント
NRMSE	K=1	0.588	0.739
	K=2	0.573	1.036
	K=3	0.569	0.868
	K=4	0.580	0.796
	K=5	0.585	0.766
	K=6	0.592	0.771
	K=7	0.600	0.805
	K=8	0.609	0.803

表3-3の結果が導かれた理由について考察するため,業種セグメントを導入する前のオリジナルのデータセットを使って，次のような確認を行った．

① 同一債務者で，前年の決算書から業種が変化している割合は，7.8%と少数である．

② K=1で欠損値補完をする場合，補完のために取得されたレコードが同一債務者である割合は20.8%である．すなわち，約8割のレコードが，他の債務者のデータを最も距離の近いレコードとして取得している．

③ 最も距離の近いレコードとして取得されたレコードのうち，補完されるレコードと同一業種である割合は45.2%である．

上記①～③の確認結果から，まず，最も距離の近いレコードを取得するという点に関して，必ずしも同一業種である必要はないということが分かる．しかし，補完に用いるレコードとして，多くが同一債務者でないレコードを取得しているにもかかわらず，結果的に半数近くは同一業種となっていることから，

同一業種であることには，ある程度の類似性が存在するものと考えられる．そこで，3.3節では，業種をセグメント化するのではなく，業種情報に重み付けを行い，他業種でも距離の近いレコードの情報を利用できるようにした．

3.3　同一業種の距離を短縮する方法による補完精度向上の試み

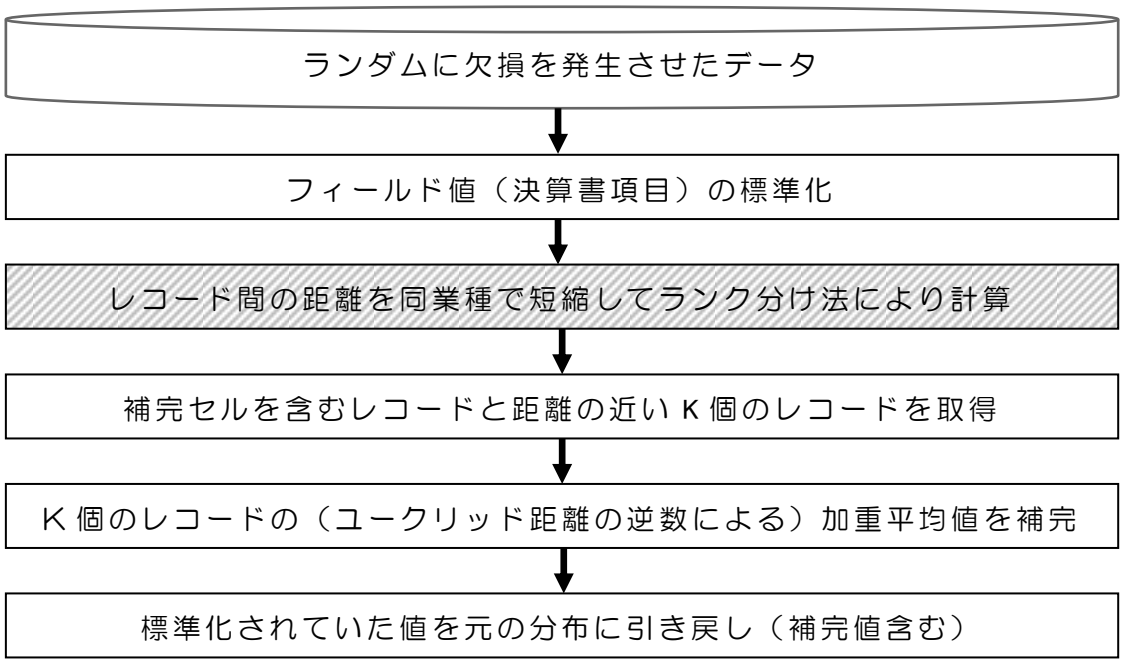
3.3.1　分析手法に関する 3.3 節との相違点

3.3 節では，業種情報をセグメント化することで k-NN 法の計算に利用したが，同一業種以外の財務諸表情報を完全に排除した結果，NRMSE は大きくなった．そこで，異なる年度で業種が変化する債務者の，異時点間の情報を完全に排除せず,一方で同一業種であるという情報を有効に利用することを試みた．すなわち，ある欠損セルを補完する値として，そのレコードが属する業種と同じ業種のレコードとの距離を計算する際，その距離を G%に短縮する処理を施した．したがって，同一業種の距離計算に限って，第 2 章の（2）式を（12）式のように変更することとなる．

$$L_{pq}=\left\{\frac{\sum_{j=1}^{J_{pq}}\left(x_{pj}-x_{qj}\right)^2}{J_{pq}}\right\}^{\frac{1}{2}}\times(G/100) \qquad (12)$$

その手順を図 2-4 と同様に示すと，図 3-3 のようになる．

図 3-3. k-NN 法による欠損値補完の手順（斜線部分が図 2-4 からの変更点）



3.3.2 分析結果

(12)式を用いて計算した結果が表 3-4 である. 表 3-4 では, レコード数 120000 のケースについて, 全ての距離を（2）で計算した通常のケース（業種間距離短縮率：100%）と, 同一業種内レコードの距離のみ G%（G=96,92,88,84,80）にした場合の NRMSE を示している.

表 3-4 同一業種内の距離を短縮させた場合の NRMSE

	業種間距離短縮率(G)					
	100%	96%	92%	88%	84%	80%
K=1	0.588	0.587	0.587	0.586	0.590	0.590
K=2	0.573	0.573	0.572	0.573	0.576	0.578
K=3	0.569	0.569	0.569	0.568	0.574	0.574
K=4	0.580	0.579	0.579	0.580	0.582	0.584
K=5	0.585	0.585	0.586	0.586	0.586	0.586
K=6	0.592	0.592	0.593	0.593	0.593	0.594
K=7	0.600	0.601	0.597	0.596	0.597	0.600
K=8	0.609	0.611	0.599	0.598	0.601	0.605

この結果を見ると，*K*=3 において，同一業種内レコードの距離を 88%（G=88）に短縮した時の NRMSE が全体の最小で 0.568 となっている．K=1,7,8 でも，G=88 において NRMSE が最小となっている．一方，K=2,4 では，G=92 で NRMSE が最小となっており，K=5,6 では，G=100，すなわち通常のケースにおいて，最も NRMSE が小さくなっている．

このように，k-NN 法において，同一業種の財務諸表を重視して補完することで，多くの K のケースでは，単純に全ての業種を均等に距離計算して補完するよりも，適切に補完できることが確認できたが，その NRMSE の減少幅は微少である．G を小さくしても NRMSE が減少しないケースも確認された．

また，K=3 のケースで，同一業種内レコードの距離を 84%以下に短縮した場合，NRMSE は大きくなっている．G を小さくすることで NRMSE が小さくなるその他の K においても，G を一定程度小さくすると，必ず NRMSE が増加する方向に転じている．これは，欠損値補完の際に，業種という 1 種類の情報に対するウェイトを大きくし過ぎたことによるものと考えられる．以上の結果を勘案すると，全業種で均等に距離を計算する（2）式に基づく計算方法は，業種情報も比較的内包しており，類似したレコードを補完値として採用する，比較的頑健な手法であると考えられる．

3.4. まとめと課題整理

本章では，第 2 章の分析を発展させ，業種区分情報を利用することで，補完値と真値の誤差をさらに小さくできるのではないか，という問題意識で分析を行った．その結果，k-NN 法については，業種情報を使ったとしても，ほとんど精度が改善しないことが示された．この結果は，他の業種情報を全く使わないケースでも，他の業種情報を利用するウェイトを可変的にしたとしても，同様の結果となった．これは，k-NN 法で計算される距離の中に，業種による財務諸表情報の差異も含まれた形で計算されていることを示しており，距離情報

に加えて，改めて業種情報を利用する必要性は無いことを意味する．

ただし，財務諸表の項目の数値については，業種による差異が大きく，信用リスク計測の際には業種セグメントを設けてモデリングを行うケースも多い．したがって，k-NN 法による欠損値補完では業種情報は有効ではなかったが，他の欠損値補完方法において業種情報を利用した場合，有効になる可能性は残っている．

第 4 章：財務諸表データに対する k-NN 法を利用した外れ値処理と信用リスク評価モデリング

4.1　はじめに

4.1.1.　研究目的

金融機関が融資先の信用リスクを測定するための情報として財務諸表データを用いた信用リスクモデリングは，日本や欧米でかなり定着している．上場企業については株価等のマーケット情報から信用リスク判定を行うモデリングも行われているが，マーケット情報のない中小企業の信用リスクモデリングについては,財務諸表情報及びデフォルト情報によるモデリングを行う必要がある．その際，個別金融機関で蓄積されているデータ量が少ない場合には，CRD などの財務諸表データを大量に集積した外部データベース機関の情報を活用したモデリングが盛んに行われている.しかし,中小企業財務諸表データに関しては，外れ値や欠損値などが存在するため,適切な信用リスクモデリングのためには，モデリングの前に，適切な外れ値処理や欠損値処理が不可欠となる．欠損値処理については，既に高橋・山下（2015）で精度の高い欠損値補完方法として k-NN（k-Nearest Neighbor）法を提示した．そこで，本研究では外れ値処理に焦点を当てる．すなわち，外れ値を含む CRD の財務諸表データベースを利用し，信用リスクモデルの予測精度を向上させる外れ値処理の新しい方法を提示する．この際，欠損値補完で利用した k-NN 法を外れ値処理に応用することを考える．k-NN 法による外れ値処理が有効であることを示すことができれば，従来以上に中小企業の財務状況を正確に把握することが可能となり，金融機関にとっては，経営支援や信用リスク管理に資することができる．

4.1.2. 先行研究

Takayasu and Okuyama（1998）で指摘されているように，財務指標データの分布に正規分布を仮定することが難しいケースは多い．また，高橋・山下（2002）でも述べられているように，財務データには欠損値が存在する．このように，実データを扱う場合，何らかの外れ値処理や欠損値処理が必要となるケースが多い．また，財務諸表データのように集積すると分布の裾が長くなるデータについては，正規分布を前提とした統計理論は適用できないため，分析の事前準備として正規化処理を行うか，もしくは正規分布を仮定しない手法を採用することが必要となる．

統計的推測の際，外れ値や裾の長い分布に対してロバストな推計を行う方法については，古くから研究されてきた．Dixon and Yuen（1974）では，外れ値や裾の長い分布を有するデータに対して，trimmed mean（刈り込み平均）や Winsorized mean（ウィンザード平均，折り込み平均）による統計的推測に関するサーベイを行っている．これらの手法は，ファイナンス分野の直近の論文においても頻繁に採用されている．例えば，Kyaw and Zhang（2014）や Pyo and Chung（2015）では，企業統計データや株式市場データを用いて実証分析を行う際に，連続変数の分布の 1%点において Winsorized（折り込み）している．Meinl and Sun（2015）では，大量データを分析する際，いかにノイズを取り除くか，という視点で多くの手法が紹介されている．その中で，Peltonen et al.（2001）の分類に従い，Nonlinear Filters の一つとして Trimmed Mean Filter や Winsorized mean filter などが紹介されている．このように，分布の端の方を折り込む処理というものは，非常に多くのデータ分析で利用されており，本研究でも，外れ値処理の一つの方法として，Winsorized（折り込み）を採用している．

上記のような方法論の他に，Whitecker et al.（2005）では，銀行のクレジットカードデータベースを用いて信用リスク評価のスコアリングモデルを推計する際，Quantile regression を行っているが，その説明変数には，裾の長い分

布を正規分布に近づけるため，neglog 変換が施されている．森平・岡崎（2009）では，財務データによる信用リスク評価のモデリングの際，マクロ指標を取り込むために多期間ロジットモデル推計を行っているが，その際の説明変数として，財務指標の neglog 変換を用いている．

4.1.3 利用データ

本研究では，まず CRD データから企業の財務分析で利用される代表的な財務指標 j を表 4-1 のように 10 種類 $\left(j=1,2;\cdots,10 \in J\right)$ 作成した．

表 4-1 主要財務指標 10 指標一覧

j	財務指標名
1	総資本当期利益率(ROA)
2	総資本回転率
3	棚卸資産回転日数
4	支払準備率
5	現預金比率
6	自己資本比率
7	デットキャパシティレイシオ
8	預借率
9	売上高支払利息割引料率
10	流動資産その他流動資産比率

次に，上記 10 種類の財務指標全てが欠損でない完全データを抽出した．この完全データの中から，更に 12000 レコードの財務諸表 $i\left(i=1,2,\cdots,12000 \in I\right)$ をランダムに抽出して，分析用データセットとした．なお，12000 レコードのうち，300 レコードがデフォルトデータ[19]である．

[19] 本研究のデフォルト定義は，CRD の財務諸表データベースの定義に従い，各レコードの決算年月から一年以内に 3 ヵ月以上延滞，実質破綻，破綻，代位弁済に該当した先としている．

4.1.4　本章の構成

4.2 節では，外れ値処理に欠損値処理の方法論である k-NN 法を応用したアプローチの分析手法について説明する．4.3 節では，その分析結果について示す．4.4 節はまとめである．

4.2　分析手法

本節では，第 2 章や高橋・山下（2015）で示した，欠損値に対する精度の高い補完方法である k-NN 法を活用することを考える．第 2 章及び高橋・山下（2015）では，本研究と同様の財務データを利用し，k-NN 法が他の欠損値補完方法と比較して最も精度の高い補完方法であることを示した．具体的には，まず，完全データにランダムに欠損セルを発生させ，そこに時系列補完，ICE（Imputation by Chained Equations）や k-NN 法など，複数の補完方法を試行した．次に，欠損化前のセルの値と，欠損セルに補完した値の標準平均平方誤差を計算し，k-NN 法が他の補完方法と比較して最もその標準平均平方誤差を小さくすることを示した．

上述の方法を応用し，外れ値を欠損値化した後に精度の高い補完を行うことで，2 項 Logit モデルの推計精度が向上するのではないか，という仮説に基づき，本アプローチの分析を進める．この際，折り返し（Winsorized）処理を，k-NN 法と比較対象となる手法として考えている．k-NN 法の計算負荷は，折り返し処理と比較すると非常に大きくなるが，k-NN 法による外れ値処理が折り返し処理と比較して最終的に 2 項 Logit モデルの推計精度を大きく向上させるのであれば，利用価値はあると考えられる．

k-NN 法を外れ値処理に応用する分析方法の手順は，次のとおりである．

i)　12000 レコードの完全データを用い，各財務指標 x_j に対して neglog 変換を行う．neglog 変換後の財務指標 $\nu(x_j)$ は，(13)式のように表される．

$$\nu(x_j) = \mathrm{sgn}(x_j) \times \ln(|x_j| + 1) \tag{13}$$

ここで $\mathrm{sgn}(x_j)$ は，(14)式のように定義される符号関数である．

$$\mathrm{sgn}(x_j) = \begin{cases} 1 & if \quad x_j \geq 0 \\ -1 & if \quad x_j < 0 \end{cases} \tag{14}$$

ii) i）で決定されたラムダを用いた neglog 変換後の各財務指標 $\nu(x_j)$ の分布に対し，上限値 $\bar{\nu}(x_j)$ 及び下限値 $\underline{\nu}(x_j)$ を設定し，$\bar{\nu}(x_j)$ を上回る値及び $\underline{\nu}(x_j)$ を下回る値を欠損値化する．上下限値の設定は，$\nu(x_j)$ の平均を μ_j，標準偏差を σ_j とした時，以下のように 2 パターン計算した．

① $\bar{\nu}(x_j) = \mu_j + 3\sigma_j$，$\underline{\nu}(x_j) = \mu_j - 3\sigma_j$

② $\bar{\nu}(x_j) = \mu_j + 4\sigma_j$，$\underline{\nu}(x_j) = \mu_j - 4\sigma_j$

iii) ii）で欠損となった $\nu(x_{ij})$ を，k-NN 法で補完する．k-NN 法による欠損値補完は，高橋・山下（2015）と同じ方法論を採用しており，その表記に従うと，以下のような 2 ステップの計算となる．

Step 1. 集合 I の中から任意の 2 つのレコード p と q $(p \neq q, p, q = 1, 2, \cdots, 12000 \in I)$ の間の距離として，標準化された値を用いて，ユークリッド距離 L_{pq} を(15)式のように計算する．通常は全 10 フィールドを用いて距離を計算するが，p もしくは q のレコードのどちらかの 10 フィールドのうちに欠損値が存在する場合には，両方のレコードで充足しているフィールドのみを用いて距離を計算している．この時，利用するフィールド数が多くなれば距離も大きくなる状況を避けるため，(15)式では，通常のユークリッド距離を，両方のレコードで欠損とならずに距離計算に利用したフィールド数 J によって修正される平均距離を用いる．

$$L_{pq}=\left\{\frac{\sum_{j'=1}^{J_{pq}}\left(x_{pj'}-x_{qj'}\right)^2}{J_{pq}}\right\}^{\frac{1}{2}} \quad (p\neq q, p, q=1,2,\cdots,12000\in I, j'=1,2,\cdots,J_{pq}) \qquad (15)$$

Step 2. 次に，Step1 で定義された距離の近い K 個のレコードを用いて，欠損値を補完する．補完方法には様々な方法があるが，ここでは，相対ユークリッド距離の逆数で加重平均した値を用いる．この方法は，金子（2005）や Crookston and Finley（2008）で利用されている．(16)式で定義されるレコード i のフィールド j の補完値 $\nu(\hat{x}_{ij})$ には，i との距離が近いレコード K 個の標準値の加重平均値を用いる．この加重値は，相対ユークリッド距離の逆数である．K の値は，理論的な根拠は存在しないので，分析者で与えることが一般的であるが，本研究では高橋・山下（2015）に従って，1～8 を試行している．なお，K 個に含まれるレコードのフィールド j が欠損である場合には，計算には用いていない[20]．

$$\nu(\hat{x}_{ij})=\sum_{k=1}^{K}\nu(x_{kj})\left(\frac{1/L_{ik}}{\sum_{k=1}^{K}(1/L_{ik})}\right), \quad (k\neq i, i=1,2,\cdots,N) \qquad (16)$$

iv) 変数変換及び欠損値補完した数値 $\nu(\hat{x}_j)$ を用い，デフォルトイベント発生有無を目的変数とした多変量 2 項ロジットモデルを最尤推定する．これを定式化すると，次のようになる．まず，企業 i の倒産確率 p_i は，ロジスティック関数を用いて，(16)式のように表される．

[20] ここで，K=1 の時，(15)式で定義されるレコード間のユークリッド距離が 0 となる場合，補完に利用するレコードの加重値は 1 とする．また，K=2 以上でユークリッド距離が 0 となるレコードが 1 つ存在した場合，0 のレコードに対して加重値は 1 として，その値のみ利用する．さらに，K=2 以上でユークリッド距離が 0 となるレコードが複数存在した場合，距離 0 以外のレコードは利用せず，距離 0 のレコードの加重値を 1 として，それらを単純平均する．

$$p_i = \frac{1}{1+\exp(\boldsymbol{\beta}\bullet\boldsymbol{\nu}(\hat{x}_i))} \tag{16}$$

ここで，β_0 を定数項とした $\boldsymbol{\beta}=(\beta_0,\beta_1,\beta_2,\cdots,\beta_{10})$ はパラメータベクトル，$\boldsymbol{\nu}(\hat{x}_i)=(1,\nu(\hat{x}_{1i}),\nu(\hat{x}_{2i}),\cdots,\nu(\hat{x}_{10i}))^T$ は説明変数ベクトルである．パラメータ $\boldsymbol{\beta}$ を最尤法で推定するための尤度関数 $L(\boldsymbol{\beta})$ は，(17)式で表される．

$$L(\boldsymbol{\beta}) = \prod_i p_i^{\delta_i}\bullet(1-p_i^{1-\delta_i}) \tag{17}$$

ここで，

$$\delta_i = \begin{cases} 1 \quad if \quad \text{企業}i\text{がデフォルトした場合} \\ 0 \quad if \quad \text{企業}i\text{がデフォルトしなかった場合} \end{cases}$$

である．(17)式を最大化することで最尤推定量 $\hat{\boldsymbol{\beta}}$ が得られる．これはすなわち，(18)式のように表される．

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) \tag{18}$$

v) ⅳ)の推定結果を AUC(Area Under Curve)で評価する．最尤推定量 $\hat{\boldsymbol{\beta}}$ から得られる倒産確率 $\hat{p}_i$ を，

$$\hat{p}_i = \frac{1}{1+\exp(\hat{\boldsymbol{\beta}}\bullet\boldsymbol{\nu}(\hat{x}))}$$

とした時，山下・三浦（2011）の表記を用いれば，AUC の計算式は(19)式のように表される．

$$\mathrm{AUC} = \frac{1}{n_D n_{ND}}\sum_{d=1}^{n_D}\sum_{m=1}^{n_{ND}} I(\hat{p}_d^D - \hat{p}_m^{ND}) \tag{19}$$

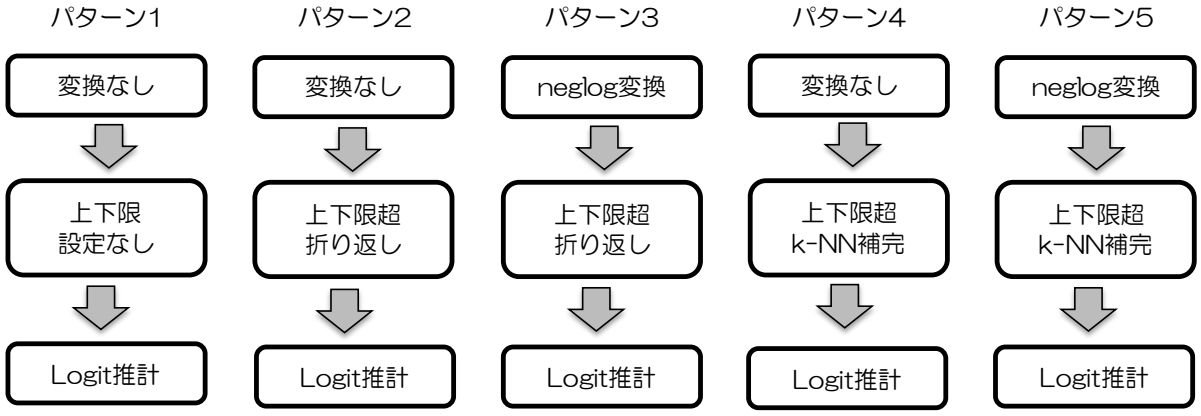
ここで，n_D, n_{ND} はそれぞれデフォルト企業数，非デフォルト企業数を表し，$I(\bullet)$ は(20)式のようなヘビサイド関数である．

$$I(\hat{p}_d^D - \hat{p}_m^{ND}) = \begin{cases} 1 \quad \text{if} \quad \hat{p}_d^D \geq \hat{p}_m^{ND} \\ 0 \quad \text{if} \quad \hat{p}_d^D < \hat{p}_m^{ND} \end{cases} \tag{20}$$

上述のｉ)～ｖ)のアプローチをパターン 5 とし，パターン 5 の比較対象となる方法論は次の 4 種類である．まず，パターン 1 では，変数変換も外れ値処理もせず，原指標 x_j のまま多変量 2 項ロジットモデル推計を行う．パターン 2 では，変数変換せず，各財務指標の上下限値 $\bar{x}_j, \underline{x}_j$ 超の値を，それぞれ $\bar{x}_j, \underline{x}_j$ に折り返した後，多変量 2 項ロジットモデル推計を行う．パターン 3 では，(13)式で定義される neglog 変換後の各財務指標の上下限値 $\bar{\nu}(x_j), \underline{\nu}(x_j)$ 超の値を，それぞれ $\bar{\nu}(x_j), \underline{\nu}(x_j)$ に折り返した後，多変量 2 項ロジットモデル推計を行う．パターン 4 では，変数変換なしの各財務指標 x_j の上下限値 $\bar{x}_j, \underline{x}_j$ 超の値を欠損値化し，欠損値を k-NN 補完（K=1～8）した後，多変量 2 項ロジットモデル推計を行う．

これらの全パターンの計算手順を図でまとめると，図 4-1 のようになる．

図 4-1 分析パターン毎の計算手順



4.3　分析結果

4.2 節で説明した分析パターン毎に計算し，AUC を計算した結果は，表 4-2 のとおり．

表 4-2　パターン別 AUC 一覧

		AUC	
パターン	knn	3σバージョン	4σバージョン
1	−	0.7771	左記に同じ
2	−	0.7872	0.7872
3	−	0.8065	0.8058
4	k＝1	0.7846	0.7845
	k＝2	0.7840	0.7848
	k＝3	0.7840	0.7848
	k＝4	0.7842	0.7850
	k＝5	0.7843	0.7849
	k＝6	0.7842	0.7846
	k＝7	0.7843	0.7844
	k＝8	0.7843	0.7847
5	k＝1	0.8082	0.8063
	k＝2	0.8073	0.8069
	k＝3	0.8066	0.8069
	k＝4	0.8074	0.8068
	k＝5	0.8072	0.8069
	k＝6	0.8079	0.8070
	k＝7	0.8080	0.8070
	k＝8	0.8079	0.8071

表 4-2 の結果を見ると，パターン 1 と比較して，パターン 2 及び 4 は，AUC がやや上昇しているが，3 及び 5 と比較するとそれほど上昇していない．また，パターン 2 と 4，パターン 3 と 5 をそれぞれ比較すると，それほど大きな差は見られない．さらに，3σバージョンと 4σバージョンを比較すると，ほとんど差が見られない．

このことから，まず，何らかの外れ値処理を行うことは，AUC を上昇させることに有効であると言える．次に，上下限設定処理を行うよりも，neglog 変換変換を行うことの方が，AUC 上昇に寄与することが確認された．一方で，上下

限値超の値を折り返すか k-NN 補完するかという点や，上下限設定を 3σ にするか 4σ にするかという点は，AUC 向上にあまり関係ないことが確認された．すなわち，AUC 向上には，変数変換を行なって財務指標の分布の歪みを補正することが重要であることが示された．次節では，この変数変換の重要性に注目し，neglog 変換を一般化した変換を行うことで，AUC 向上を目指す．

なお，本節の結果では，僅かではあるが，パターン 5 の方が AUC は高くなったが，次節ではパターン 3 を一般化することを考えている．これは，2.1 節でも述べたように，k-NN 法は計算負荷が折り返し処理と比較して非常に大きいことによる．実際に，k-NN 法を利用した本研究のパターン 4 もしくはパターン 5 の計算には約 8000 秒の時間がかかる一方，パターン 3 は数秒で計算を終えている[21]．本分析は 12000 レコードの計算であるが，レコード数が多くなるに従い，本分析で用いた k-NN 法は比例的に計算量が増加する（高橋・山下（2015)）．折り返し計算も同様に比例的に計算量は増加するが，計算負荷は比較にならないほど小さい．したがって，レコード数が多くなった場合に，k-NN 法を利用することは現実的ではなく，次節では，より現実的な利用が見込まれる折り返し処理の手法を発展されることを考えている．なお，次節では，neglog 変換を一般化する最適な変換率 λ^* と，最適な折り返し点 θ^* を求めるが，仮にこの計算に k-NN 法を適用すると，さらに最適な近傍数 k^* も求める必要があり，計算負荷はさらに大きくなる．以上のような理由から，本章の僅かにパターン 5 の AUC がパターン 3 を上回っているにもかかわらず，次章ではパターン 3 を一般化することを考える．

[21] 計算には，Intel® Core™ i7 CPU 2.80GHz，RAM24.0GB のスペックの PC を用いている．

4.4　まとめと課題整理

本章では，外れ値処理に対して，欠損値に対する精度の高い補完方法である k-NN 法を応用し，外れ値を欠損値化した後に精度の高い補完を行うことで，2 項 Logit モデルの AUC で見た推計精度が向上するのではないか，という仮説の下，検証を進めた．結果的に，k-NN 法を応用した外れ値処理方法は，AUC 向上にそれほど有効ではなく，変数変換の方が比較的重要であることが確認された．

本章の結果を受けて，第 5 章では，第 4 章で利用した neglog 変換をより一般化することで，AUC 向上が見込まれることから，その手法について考察する．

第 5 章：財務諸表データに対する一般化 neglog 変換を利用した外れ値処理と信用リスク評価モデリング

5.1　はじめに

第 4 章では，k-NN 法のように精度の高い欠損値補完方法を応用して外れ値処理を行ったとしても，必ずしも Logit モデル推計の AUC で見た精度向上にはつながらず，変数変換の方が相対的に重要であることが分かった．これを踏まえて，本章では，変数変換の手法を一般化することを考える．具体的には，既存の手法である neglog 変換を一般化した一般化 neglog 変換（generalized neglog transformation）を提案し，一般化 neglog 変換を用いる方が，Logit モデル推計の際に AUC 向上という面で有効であることを示す．

本章の構成は，次のとおり．5.2 節では，neglog 変換を一般化した一般化 neglog 変換について詳細に説明し，それを踏まえた Logit モデル推計の分析手順を示す．5.3 節では，分析結果を示す．5.4 節はまとめである．

5.2　分析手法

5.2.1　neglog 変換と一般化 neglog 変換の関係

第 4 章において，折り返し処理と neglog 変換により，ロジットモデル推計の AUC が向上することを確認できたが，このうち，neglog 変換については一般化の余地がある．本節では，この点に焦点を当てて，AUC 向上を目指す．

neglog 変換については，正値の対数変換を負値にも適用できるような関数形となっており，両側に裾の非常に長い分布をモデリングする際に有効な変換手法である．しかし，正値の対数変換をより一般化したものとして Box-Cox 変換（Box and Cox（1964））が存在するように，neglog 変換についても一般化が可能である．本研究では，neglog 変換をより一般化した変換（一般化 neglog

変換）を提示し，その最適な曲率についても考慮した分析を行う[22].

すなわち，財務指標 x_j 毎に折り返し処理を行った後，一般化された neglog 変換を単変量2項ロジットモデルのAUCが最大となるように変換率を設定[23]して，その後に多変量 2 項ロジットモデル推計を行えば，AUC が上昇すると考えられる．第 4 章の(13)式で表される neglog 変換を一般化した変換式は(21)式のように表される．これを一般化 neglog 変換とよぶ．

$$\psi(\lambda_j, x_j) = \begin{cases} \mathrm{sgn}(x_j) \times \dfrac{(|x_j|+1)^{\lambda_j} - 1}{\lambda_j} & (\lambda_j \neq 0) \\ \mathrm{sgn}(x_j) \times \ln(|x_j|+1) & (\lambda_j = 0) \end{cases} \tag{21}$$

ここで，λ_j は一般化 neglog 変換の変換率を決定するパラメータである．(21)式から明らかなように，$\lambda_j = 0$ の時は neglog 変換に対応し，$\lambda_j = 1$ の時は無変換のケースに対応する．これを図で表すと，図 5-1 のようになる．

[22] なお，一般化 neglog 変換と類似した考え方として，Yeo and Johnson（2000）で提示されている Yeo-Johnson 変換がある．Yeo-Johnson 変換は，x-y 軸の図で表すと y=－x で線対称となる変換式であるが，一般化 neglog 変換は，原点で点対称となる変換である．

[23] 各財務指標の分布を正規分布に近づけるように一般化 neglog 変換の曲率を決定し，その後 2 項ロジットモデル推計を行う方法も考えられる．しかし，各財務指標の分布を正規分布に近づけたとしても AUC が高くなるとは限らないため，本研究では採用していない．各財務指標の分布を正規分布に近づけた結果については，付録 C を参照．

図 5-1 neglog 変換と一般化 neglog 変換

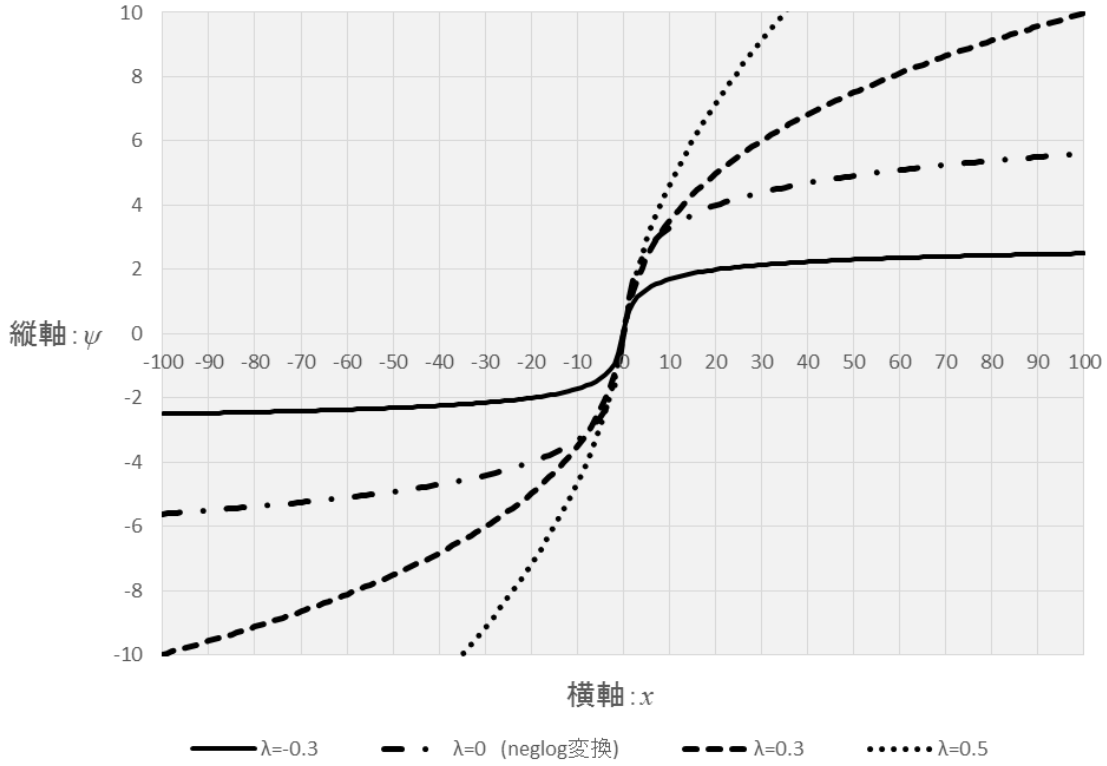
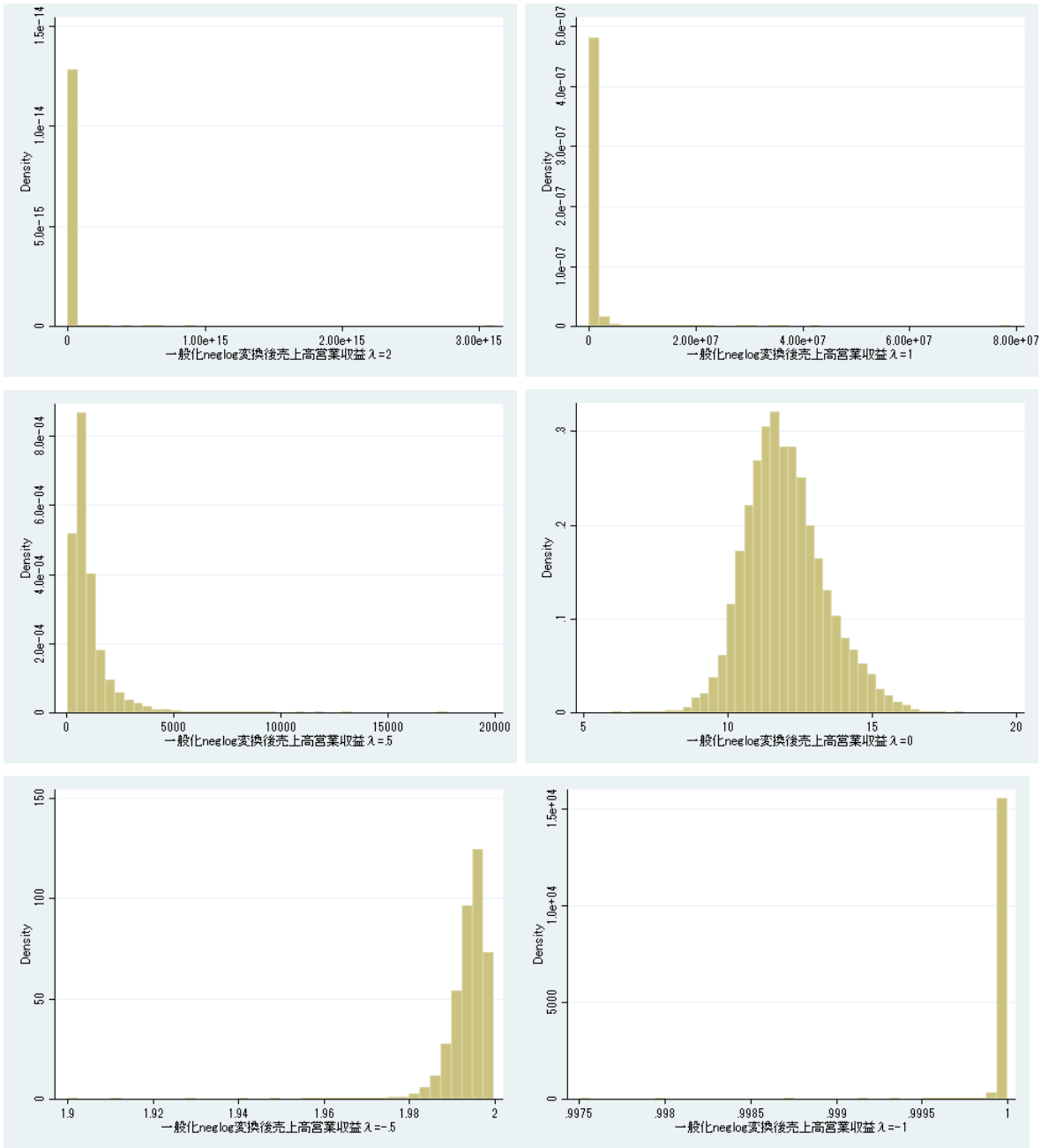


図 5-1 から分かるように，$0<\lambda<1$の場合は，neglog 変換と無変換の中間の変換率であり，neglog 変換では縮約しすぎるような分布に対して，より緩やかな縮約をすることができる．$\lambda<0$の場合は，neglog 変換では縮約が十分でないような分布に対して，より大きな縮約をすることができる．$\lambda>1$の場合は，分布を縮約するのではなく，拡張させるような変換となる．例えば，売上高営業収益という財務項目の分布は，裾の非常に長い分布となっているが，様々なλの値で一般化 neglog 変換すると，図 5-2 のようになる．

図 5-2 売上高営業収益の一般化 neglog 変換



5.2.2 分析手順

本節における分析手順は，以下のとおりである．

i) 12000 レコードの完全データを用い，各財務指標 x_j に対して分布の上下 θ％点の値 $\overline{x}_j^{\theta}, \underline{x}_j^{\theta}$ を設定し，上下限値 $\overline{x}_j^{\theta}, \underline{x}_j^{\theta}$ 超の値を $\overline{x}_j^{\theta}, \underline{x}_j^{\theta}$ に折り返し処理する．折り返し処理後の財務指標分布を x_j^{θ} で表す．折り返し処理後の各財務指標 x_j^{θ} に対し，(21)式で表される一般化 neglog 変換を施す．すなわち，(21)式は(22)式のように変更される．

$$\psi(\lambda_j, x_j^{\theta}) = \begin{cases} \mathrm{sgn}(x_j^{\theta}) \times \dfrac{(|x_j^{\theta}|+1)^{\lambda_j} - 1}{\lambda_j} & (\lambda_j \neq 0) \\ \mathrm{sgn}(x_j^{\theta}) \times \ln(|x_j^{\theta}|+1) & (\lambda_j = 0) \end{cases} \tag{22}$$

折り返し処理及び一般化 neglog 変換を施した各財務指標 $\psi(\lambda_j, x_j^{\theta})$ を説明変数とし，デフォルトイベント発生有無を目的変数とした単変量 2 項ロジットモデルを推計する．この際，グリッド計算により単変量で AUC が最大となる最適な折り返し点 θ^* と一般化 neglog 変換の最適な λ^* を決定する．単変量 2 項ロジットモデルは，財務指標 j に関する企業 i の倒産確率 p_i を用いて，(23)式のように表される．

$$p_{ij}(a_j, b_j, \lambda_j, \theta_j) = \frac{1}{1+\exp(a_j + b_j \bullet \psi(\lambda_{ij}, x_{ij}^{\theta}))} \tag{23}$$

ここで，a_j, b_j はパラメータである．(23)式から計算される AUC は(24)式のようになる．

$$\mathrm{AUC}(a_j, b_j, \lambda_j, \theta_j) = \frac{1}{n_D n_{ND}} \sum_{d=1}^{n_D} \sum_{m=1}^{n_{ND}} I(p_d^D(a_j, b_j, \lambda_j, \theta_j) - p_m^{ND}(a_j, b_j, \lambda_j, \theta_j)) \tag{24}$$

(24)式を最大にする $(a_j^*, b_j^*, \lambda_j^*, \theta_j^*)$ を求める．すなわち，(25)式のように定式化される．

$$(a_j^*, b_j^*, \lambda_j^*, \theta_j^*) = \arg \max_{a_j, b_j, \lambda_j, \theta_j} \mathrm{AUC}(a_j, b_j, \lambda_j, \theta_j) \tag{25}$$

ii) 各変数の最適な $(\lambda_j^*, \theta_j^*)$ を用いて財務指標 x_j を変換し，多変量 2 項ロジットモデルを最尤法で推計する．なお，iv)においても，iii)と同様に，AUC を最大化するようなグリッド計算を考えたが，多変量では計算負荷が大きすぎることから，微分可能な尤度関数を用いた最尤法で推計を行った．すなわち，企業 i の倒産確率 p_i は，$\boldsymbol{\lambda}^*(=\lambda_1^*, \lambda_2^*, \cdots, \lambda_{10}^*)$, $\boldsymbol{\theta}^*(=\theta_1^*, \theta_2^*, \cdots, \theta_{10}^*)$ を所与の下，(26)式のように表される．

$$p_i = \frac{1}{1+\exp(\boldsymbol{b} \bullet \boldsymbol{\psi}(x_i \mid \boldsymbol{\lambda}^*, \boldsymbol{\theta}^*))} \tag{26}$$

ここで，b_0 を定数項とした $\boldsymbol{b} = (b_0, b_1, b_2; \cdots, b_{10})$ はパラメータベクトル，$\boldsymbol{\psi}(x_i \mid \boldsymbol{\lambda}^*, \boldsymbol{\theta}^*) = (1, \psi(x_{1i} \mid \boldsymbol{\lambda}^*, \boldsymbol{\theta}^*), \psi(x_{2i} \mid \boldsymbol{\lambda}^*, \boldsymbol{\theta}^*), \cdots, \psi(x_{10i} \mid \boldsymbol{\lambda}^*, \boldsymbol{\theta}^*))^T$ は説明変数ベクトルである．パラメータ $\boldsymbol{b}$ を最尤法で推定するための尤度関数 $L(\boldsymbol{b} \mid \boldsymbol{\lambda}^*, \boldsymbol{\theta}^*)$ は，(27)式で表される．

$$L(\boldsymbol{b} \mid \boldsymbol{\lambda}^*, \boldsymbol{\theta}^*) = \prod_i p_i^{\delta_i} \bullet (1 - p_i^{1-\delta_i}) \tag{27}$$

したがって，このアプローチは(28)式のように定式化される．

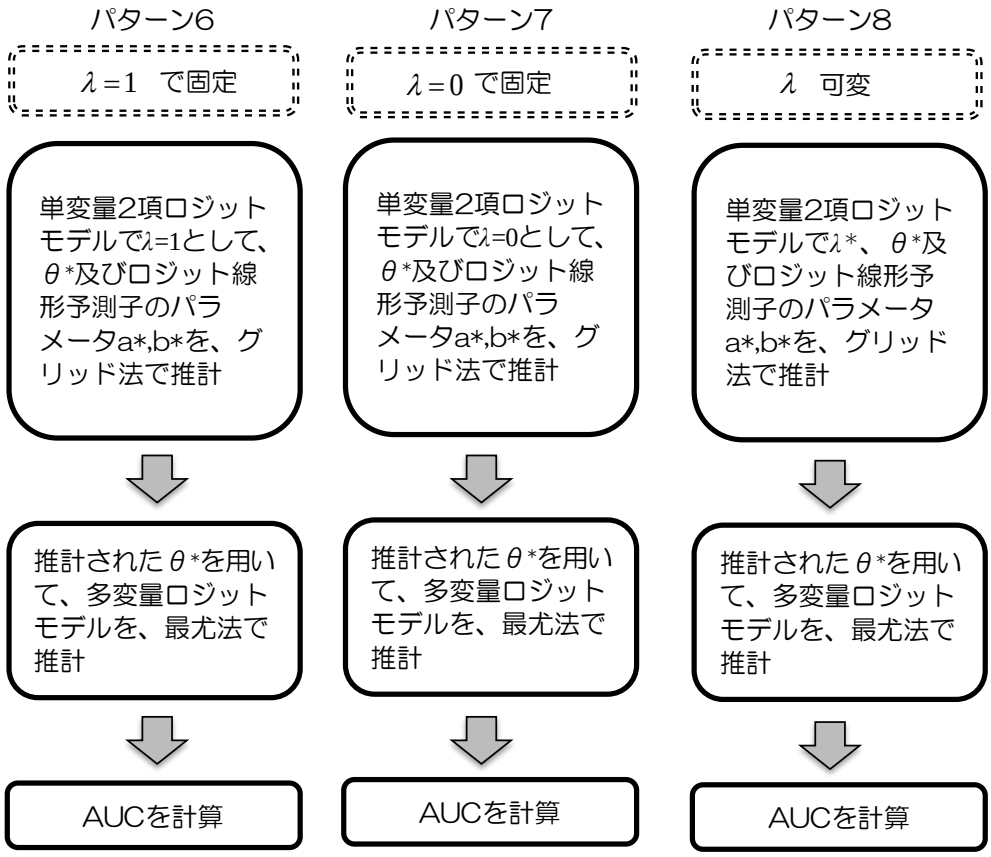
$$(\hat{\boldsymbol{b}} \mid \boldsymbol{\lambda}^*, \boldsymbol{\theta}^*) = \arg\max_{\boldsymbol{b}} L(\boldsymbol{b} \mid \boldsymbol{\lambda}^*, \boldsymbol{\theta}^*) \tag{28}$$

iii) 最後に，(29)式のように AUC を計算する．

$$\mathrm{AUC}(\hat{\boldsymbol{b}} \mid \boldsymbol{\lambda}^*, \boldsymbol{\theta}^*) = \frac{1}{n_D n_{ND}} \sum_{d=1}^{n_D} \sum_{m=1}^{n_{ND}} I(\hat{p}_d^D - \hat{p}_m^{ND}) \tag{29}$$

上述のⅰ)～vii)の計算手順を踏むアプローチをパターン 8 としたとき，上述のⅱ)で変数変換をしないケース，すなわち $\lambda = 1$ のケース（パターン 6），及び，neglog 変換のケース，すなわち $\lambda = 0$ のケース（パターン 7）が比較対象となる．パターン 6,7 はいずれも，λ を固定の上，最適な θ^* のみ決定した後，最尤法による多変量 2 項ロジットモデル推計を行うことと同義である．以上の手順を，図 4-1 と同様にフローチャートにまとめると，図 5-3 のようになる．

図 5-3 分析パターン毎の計算手順



5.3　分析結果

5.2 節で説明したアプローチにおけるパターン 8 の最適な (λ^*,θ^*) は，表 5-1 のようになった．

表 5-1　パターン 8 の最適な (λ^*,θ^*)

j	財務指標名	λ *	θ *
1	総資本当期利益率(ROA)	0.32	–※
2	総資本回転率	−2.45	5.9
3	棚卸資産回転日数	0.62	0.3
4	支払準備率	−0.52	3.1
5	現預金比率	−0.69	2.5
6	自己資本比率	0.26	−
7	デットキャパシティレイシオ	−0.31	−
8	預借率	−0.27	1.8
9	売上高支払利息割引料率	−0.12	6.2
10	流動資産その他流動資産比率	0.59	3.9

※θ *の“－”は、最適計算の結果、折り返しなし(θ ＝0)となったことを意味する。

表 5-1 の数値を財務指標毎 x_j に適用し，折り返しのみのケース（パターン 6），折り返しかつ neglog 変換のケース（パターン 7），折り返し一般化 neglog 変換のケース（パターン 8）でそれぞれ多変量 2 項ロジットモデル推計を行い，AUC を比較した結果が表 5-2 である[24]．

表 5－2　計算パターン別 AUC

パターン	AUC
6　変換なし	0.7865
7　neglog変換	0.8078
8　一般化neglog変換	0.8121

[24] なお，5.2 節で説明したとおり，パターン 6,7 は最適な θ^* として，パターン 8 の数値，すなわち，表 3 と同じ数値を用いている．

表 5-2 からは，一般化 neglog 変換が，通常の neglog 変換よりも AUC 向上という面で有効な変数変換の手法であることが確認できた．

5.4 まとめと課題整理

第 5 章では，既存の変数変換手法である neglog 変換よりも，新しい変数変換手法である一般化 neglog 変換を用いた．この際，最適な変換率 λ^* 及び折り返し点 θ^* を同時推計する新しい方法を採用した．この結果，一般化 neglog 変換は，2 項 Logit モデル推計の際に AUC 向上という面で有効であることが確認された．

最後に，今後の展望について議論する．5.3 節の分析では，各指標の両端 θ% を折り返し処理したが，指標によっては片側だけ折り返した方が AUC 向上に資するケースも考えられる．同様に，各指標を正値と負値で同じ変換率ラムダを適用したが，それぞれ別の変換率を最適に設定した方が AUC 向上に資するはずである．これらの計算手法開発については，次の課題である．また，一般化 neglog 変換は，単変量での最適な (θ^*, λ^*) を求めたが，多変量での最適点の決定は課題として残っている．試行段階では最尤法による探索は不安定になることが分かっており，他の最適解を計算する方法論を適用する必要がある．

第 6 章：結語

本研究では，CRD 協会の財務諸表データを用い，財務諸表データに対して有効な欠損値処理及び外れ値処理とはどのようなものか，という点について主に考察を行った．

第 2 章では，遺伝子データや工程数予測などの欠損値補完に用いられてきた k-NN 法は，大規模な財務諸表データについても，十分な有効性が確認された．欠損値を含む財務諸表データに k-NN 法を適用することにより，欠損値補完の一般的な方法である平均値補完や，連鎖的な回帰方程式による欠損値補完方法である ICE，同一債務者の時系列データによる補完よりも，真値と補完値の誤差を小さくすることが確認された．欠損値を含む現実のデータに対して補完する際も，同一債務者の情報は使えないケースが多いと考えられることから，k-NN 法や ICE に次ぐ安定的な欠損値補完方法である時系列補完は現実のデータに対する活用可能性は低く，本章で提示した k-NN 法による補完方法の有効性は高いものと考えられる．

第 2 章の特徴として，特に大規模なデータの k-NN 法を計算する時に有効となる，売上高ランクを導入した効率的な計算方法について提示した．この方法は，売上高のように完全フィールドを想定できるフィールドが存在する場合には，財務諸表以外の他のデータでも応用可能であると考えられる．

第 2 章の研究内容に関する課題として，欠損値が存在する財務諸表情報を用いた信用リスク評価に対して，この欠損値補完方法がどのように影響を及ぼすかを確認する点が挙げられる．その際，k-NN 法により欠損値を補完した場合とその他の欠損値処理方法を比較し，二項ロジットモデルの AUC などの予測精度がどの程度異なるのかを確認する必要がある．欠損値を含む財務諸表のリスク評価をどのように行うかは，非常に重要な問題である．仮に，第 2 章で提示した k-NN 法により欠損データを補完したことにより，信用リスクモデルの予測精度が上昇すれば，信用リスク計量化の前進に大きな貢献となる．この取り組みについては，本研究の第 4 章で考察した．

この他に，第 2 章の結果が一般的に成立するかどうかについては，CRD データ以外の外部データに対する有効性についても確認する必要がある．また，遺伝子研究の分野では，k-NN 法を異常値修正に応用しているケースがある．この点についても，財務諸表データでの有効性を確認する必要がある．

第 3 章では，第 2 章の分析を発展させ，業種区分情報を利用することで，補完値と真値の誤差をさらに小さくできるのではないか，という問題意識で分析を行った．その結果，k-NN 法については，業種情報を使ったとしても，ほとんど精度が改善しないことが示された．この結果は，他の業種情報を全く使わないケースでも，他の業種情報を利用するウェイトを可変的にしたとしても，同様の結果となった．これは，k-NN 法で計算される距離の中に，業種による財務諸表情報の差異も含まれた形で計算されていることを示しており，距離情報に加えて，改めて業種情報を利用する必要性は無いことを意味する．

ただし，財務諸表の項目の数値については，業種による差異が大きく，信用リスク計測の際には業種セグメントを設けてモデリングを行うケースも多い．したがって，k-NN 法による欠損値補完では業種情報は有効ではなかったが，他の欠損値補完方法において業種情報を利用した場合，有効になる可能性は残っている．

第 4 章では，外れ値処理に対して，欠損値に対する精度の高い補完方法である k-NN 法を応用し，外れ値を欠損値化した後に精度の高い補完を行うことで，2 項 Logit モデルの AUC で見た推計精度が向上するのではないか，という仮説の下，検証を進めた．結果的に，k-NN 法を応用した外れ値処理方法は，AUC 向上にそれほど有効ではなく，変数変換の方が比較的重要であることが確認された．第 4 章の結果を受けて，第 5 章では，第 4 章で利用した neglog 変換をより一般化することで，AUC 向上が見込まれることから，その手法について考察する．

第 5 章では，既存の変数変換手法である neglog 変換よりも，新しい変数変換手法である一般化 neglog 変換を用いた．この際，最適な変換率 及び折り返し点 を同時推計する新しい方法を採用した．この結果，一般化 neglog 変換は，

2 項 Logit モデル推計の際に AUC 向上という面で有効であることが確認された.

最後に，今後の展望について議論する．5.3 節の分析では，各指標の両端 % を折り返し処理したが，指標によっては片側だけ折り返した方が AUC 向上に資するケースも考えられる．同様に，各指標を正値と負値で同じ変換率ラムダを適用したが，それぞれ別の変換率を最適に設定した方が AUC 向上に資するはずである．これらの計算手法開発については，次の課題である．また，一般化 neglog 変換は，単変量での最適な (θ^*, λ^*) を求めたが，多変量での最適点の決定は課題として残っている．試行段階では最尤法による探索は不安定になることが分かっており，他の最適解を計算する方法論を適用する必要がある．

全体として，財務諸表データに対して有効な欠損値処理及び外れ値処理について，それぞれ示すことができた．しかし，欠損値処理と外れ値処理を同時に解決する方法論の提示という点については，第 5 章のような折り返し処理と AUC 最大化グリッド法による一般化 neglog 変換という組み合わせの提示に留まっている．一回の処理もしくは同一の方法論で欠損値処理と外れ値処理を行うことができれば，従来よりもデータ整備段階での処理内容及び処理負担が大きく軽減されるはずであり，今後の研究課題としては，そのような方法論について研究を続けたい.

付録 A　分析用データの特徴

ここでは，本研究で用いたデータの特徴について示す．

P.4 図 1-1 の売上高分布のヒストグラムに関する基礎統計量は，表 A-1 の No.72 のとおり．なお，表 A-1 には，ヒストグラム作成には，分析用データから 2001 年から 2006 年までの財務諸表から 100 万件をランダムサンプリングしたデータを用いている．表 A-1.No.72 から分かるとおり，売上高分布の平均値は中央値の約 3.5 倍，標準偏差は平均の約 3 倍となっており，大きく右に偏っていることが分かる．その点は歪度でも確認できる．

表 A-1. 財務項目の基礎統計量

番号	項目名	件数	1%点位	中央値	99%点位	平均	標準偏差	歪度	尖度
No.	items	N	p1	p50	p99	mean	sd	skewness	kurtosis
11	流動資産合計	1000000	8970	147000	6500000	521132	1686429	16	444
12	現金・預金	999999	70	11700	794000	60566	262243	36	2645
13	受取手形	999767	0	0	343000	17199	113480	31	2224
14	売掛金	999969	0	12800	857000	61713	244985	24	1478
15	有価証券	622979	0	0	78900	4301	82232	145	31834
16	棚卸資産合計	999959	0	4710	704000	51777	516247	103	16627
17	商品・製品	526921	0	0	572000	38372	502614	115	19660
18	半製品・仕掛品	526846	0	0	280000	18643	362420	112	18338
19	原材料・貯蔵品	526887	0	0	146000	7732	75799	90	15990
20	その他棚卸資産	399727	0	0	70700	4720	138694	147	27423
21	その他流動資産合計	999965	−1820	2450	352000	25955	917514	465	226499
22	前渡金	561301	0	0	26300	1576	31766	125	25983
23	前払費用	561497	0	100	22100	1677	14300	71	8505
24	未収入金	561427	0	10	136000	7729	68422	95	19140
25	未収収益	289561	0	0	17100	1185	60322	263	75710
26	短期貸付金	561379	0	0	186000	11400	154935	93	15788
27	その他流動資産	559126	−3810	50	127000	11350	1206774	365	135370
28	(▲)貸倒引当金流動資産	561311	−23000	0	0	−1413	27016	−242	83812
29	固定資産合計	1000000	500	40100	2760000	214172	1200025	96	17306
30	有形固定資産合計	999999	80	28200	2210000	168388	891634	83	14375
31	建物・構築物	613106	0	10500	1150000	85133	474269	65	8385
32	機械・装置	561565	0	4110	333000	24456	158378	61	7472
33	工具・器具・備品	559927	0	700	129000	10255	238768	130	21590
34	土地	999797	0	0	1090000	76244	516565	151	40830
35	建設仮勘定	612851	0	0	29400	2350	100809	298	119984
36	その他固定資産	948345	0	5240	539000	39510	439698	152	35418
37	無形固定資産	623201	0	200	52300	3416	64944	182	50487
38	投資等合計	674841	0	6580	792000	56751	519735	126	24928
39	投資有価証券	565094	0	100	305000	20675	452990	177	40623
40	長期貸付金	564820	0	0	110000	6836	146808	123	24002
41	その他投資	565079	0	3440	336000	24971	168115	99	16810
42	(▲)貸倒引当金固定資産	563159	−500	0	0	−478	25399	−194	53955
43	繰延資産	999748	0	0	19400	1087	23583	320	144016
44	資産合計	1000000	5300	105000	5380000	435143	2036909	86	16060
45	流動負債合計	1000000	700	38300	2350000	173877	1148874	250	95663
46	支払手形	999787	0	0	575000	30640	166032	23	1214
47	買掛金	999935	0	4590	508000	34200	188994	117	46084
48	短期借入金	1000000	0	10000	1070000	72997	430409	61	9004
49	その他流動負債合計	1000000	30	7050	432000	36050	923655	468	228415
50	未払金	557587	0	2600	230000	16387	78145	28	1467

表 A-1:続き 財務項目の基礎統計量

番号	項目名	件数	1%点位	中央値	99%点位	平均	標準偏差	歪度	尖度
No.	items	N	p1	p50	p99	mean	sd	skewness	kurtosis
51	設備未払金・設備関係支払手形	319298	0	0	13000	936	20591	94	14400
52	未払費用	557463	0	0	89000	5494	56262	383	220670
53	前受金	557402	0	0	125000	7323	130402	106	18180
54	前受収益	528694	0	0	400	217	13987	182	42477
55	従業員預り金	234040	0	0	800	67	9648	462	219029
56	短期引当金	549537	0	0	17400	942	16489	99	16924
57	その他流動負債	551335	0	1140	171000	15334	1224097	364	134322
58	固定負債合計	1000000	0	42900	2120000	175463	950249	69	8918
59	社債・長期借入金	1000000	0	40800	1930000	160709	794582	64	8042
60	その他固定負債	999718	0	0	226000	14760	278099	102	15529
61	長短借入金合計	1000000	0	61900	2820000	233700	1085506	66	9562
62	引当金	601208	0	0	0	48	6774	235	63942
63	負債合計	1000000	5580	92200	4280000	349369	1722209	106	23115
64	資本合計	1000000	-120000	10800	1390000	85772	507559	42	3911
65	資本金	1000000	3000	10000	100000	16868	83463	58	4810
66	その他の資本	999997	-146000	2090	1310000	69723	484282	43	4136
67	資本準備金	560868	0	0	29400	3316	134173	181	44962
68	利益準備金	540966	0	0	30000	2710	61230	254	71096
69	任意積立金	560944	0	0	1190000	67588	452621	46	4209
70	当期未処分利益	565118	-195000	900	490000	17278	215156	17	2594
71	負債・資本合計	1000000	5300	105000	5380000	435145	2036911	86	16060
72	売上高・営業収益	1000000	8970	147000	6500000	521132	1686429	16	444
73	売上原価・営業原価	1000000	0	89300	5370000	396756	1449092	17	521
74	うち外注加工費	581486	0	0	1100000	75898	2415002	270	82968
75	うち労務費	581474	0	0	676000	45497	200606	67	11242
76	うち賃借料原価	570532	0	0	86000	5159	47866	82	11596
77	うち租税公課原価	518939	0	0	14800	728	7947	116	23739
78	売上総利益	1000000	1790	44500	1310000	124378	365314	30	2953
79	販売費および一般管理費	1000000	4290	43300	1150000	113003	315808	34	4293
80	うち人件費	582874	0	27900	690000	72399	169915	15	530
81	うち賃借料販管費	519188	0	2400	126000	9668	44022	46	4864
82	うち租税公課販管費	519202	0	1290	46400	4233	13907	26	1416
83	営業利益	1000000	-47500	1420	222000	11375	96348	73	15348
84	営業外収益合計	623275	0	1290	102000	7569	54333	192	58123
85	受取利息・割引料・配当金	999976	0	10	10000	650	13438	254	106923
86	その他営業外収益	565100	0	1240	95500	7239	49311	190	55346
87	営業外費用合計	623276	0	1900	105000	8505	53767	178	54314
88	支払利息・利子割引料	1000000	0	1510	67400	5766	55191	739	662924
89	その他営業外費用	564939	0	0	39700	2370	45251	308	117922
90	経常利益	1000000	-43100	900	211000	10334	90477	72	15659
91	特別利益	564901	0	0	73000	4335	69944	123	27049
92	特別損失	564979	0	0	134000	7466	83865	76	9542
93	税引前当期利益	616732	-66900	1130	248000	12049	119447	49	9592
94	法人税等充当額	616596	0	80	106000	6070	44601	77	15009
95	当期利益	1000000	-53800	500	110000	3728	69536	30	12211
96	株主配当金	555078	0	0	11700	705	13253	123	23291
97	役員賞与	551677	0	0	8500	299	3115	68	10606
98	受取手形割引高	978866	0	0	250000	12503	77622	24	1237
99	受取手形裏書譲渡高	973782	0	0	55000	2677	75184	553	400341
100	有形固定資産減価償却累計額	506141	0	0	1250000	66789	1298729	400	216457
101	減価償却実施額	999520	0	2000	142000	10881	78088	97	15219
102	保証債務合計	521675	0	0	0	2324	556352	407	169907
103	期末従業員数人	1000000	0	7	181	20	281	218	54452

※1%点位、中央値、99%点位、平均、標準偏差の単位は全て千円。
※2001年から2006年までの決算書から100万件をランダムサンプリングした数字。

上記と同様にランダムサンプリングした 100 万件のデータのうち，各財務項目（フィールド）でどの程度の欠損数があるのかを示したものが，表 A-2 である．欠損数を 100 万件で割った欠損率も表示している．表 A-2 から，“資産合計”，“売上高営業収益”，“経常利益”などの重要な項目は欠損数が 0 となって

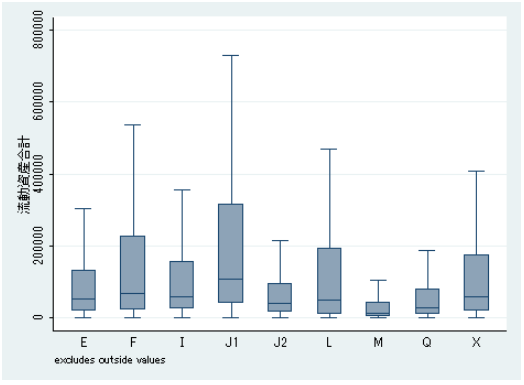
いることが分かる．一方，“商品・製品”などの“棚卸資産”の内訳項目や，“人件費”などの“販売費および一般管理費”の内訳項目など，内訳項目となっている財務項目は，約 40%程度の欠損率となっていることが多い．他にも，“未収収益”など，約 70%程度の欠損率となっている項目が存在することも確認できる．

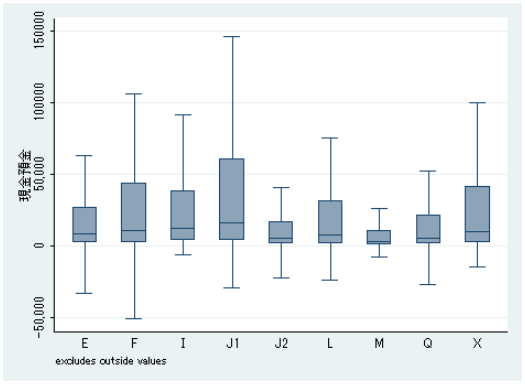
表 A-2. 分析用データ(サンプリング後)の欠損数と欠損率

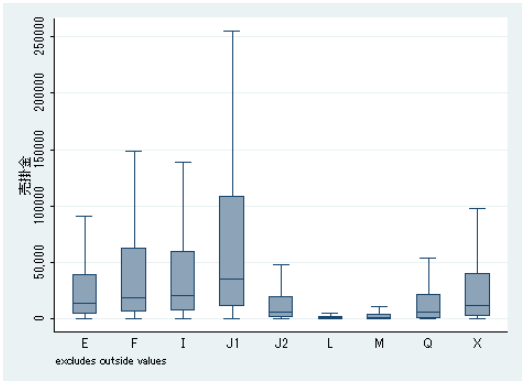
No.	項目名	欠損数	欠損率
11	流動資産合計	0	0%
12	現金・預金	1	0%
13	受取手形	233	0%
14	売掛金	31	0%
15	有価証券	377,021	38%
16	棚卸資産合計	41	0%
17	商品・製品	473,079	47%
18	半製品・仕掛品	473,154	47%
19	原材料・貯蔵品	473,113	47%
20	その他棚卸資産	600,273	60%
21	その他流動資産合計	35	0%
22	前渡金	438,699	44%
23	前払費用	438,503	44%
24	未収入金	438,573	44%
25	未収収益	710,439	71%
26	短期貸付金	438,621	44%
27	その他流動資産	440,874	44%
28	(▲)貸倒引当金流動資産	438,689	44%
29	固定資産合計	0	0%
30	有形固定資産合計	1	0%
31	建物・構築物	386,894	39%
32	機械・装置	438,435	44%
33	工具・器具・備品	440,073	44%
34	土地	203	0%
35	建設仮勘定	387,149	39%
36	その他固定資産	51,655	5%
37	無形固定資産	376,799	38%
38	投資等合計	325,159	33%
39	投資有価証券	434,906	43%
40	長期貸付金	435,180	44%
41	その他投資	434,921	43%
42	(▲)貸倒引当金固定資産	436,841	44%
43	繰延資産	252	0%
44	資産合計	0	0%
45	流動負債合計	0	0%
46	支払手形	213	0%
47	買掛金	65	0%
48	短期借入金	0	0%
49	その他流動負債合計	0	0%
50	未払金	442,413	44%
51	設備未払金・設備関係支払手形	680,702	68%
52	未払費用	442,537	44%
53	前受金	442,598	44%
54	前受収益	471,306	47%
55	従業員預り金	765,960	77%
56	短期引当金	450,463	45%
57	その他流動負債	448,665	45%
58	固定負債合計	0	0%
59	社債・長期借入金	0	0%

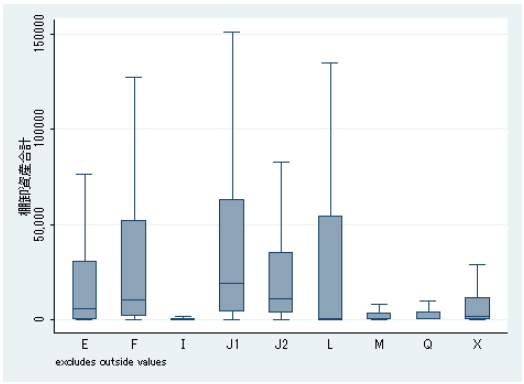
No.	項目名	欠損数	欠損率
60	その他固定負債	282	0%
61	長短借入金合計	0	0%
62	引当金	398,792	40%
63	負債合計	0	0%
64	資本合計	0	0%
65	資本金	0	0%
66	その他の資本	3	0%
67	資本準備金	439,132	44%
68	利益準備金	459,034	46%
69	任意積立金	439,056	44%
70	当期未処分利益	434,882	43%
71	負債・資本合計	0	0%
72	売上高・営業収益	0	0%
73	売上原価・営業原価	0	0%
74	うち外注加工費	418,514	42%
75	うち労務費	418,526	42%
76	うち賃借料原価	429,468	43%
77	うち租税公課原価	481,061	48%
78	売上総利益	0	0%
79	販売費および一般管理費	0	0%
80	うち人件費	417,126	42%
81	うち賃借料販管費	480,812	48%
82	うち租税公課販管費	480,798	48%
83	営業利益	0	0%
84	営業外収益合計	376,725	38%
85	受取利息・割引料・配当金	24	0%
86	その他営業外収益	434,900	43%
87	営業外費用合計	376,724	38%
88	支払利息・利子割引料	0	0%
89	その他営業外費用	435,061	44%
90	経常利益	0	0%
91	特別利益	435,099	44%
92	特別損失	435,021	44%
93	税引前当期利益	383,268	38%
94	法人税等充当額	383,404	38%
95	当期利益	0	0%
96	株主配当金	444,922	44%
97	役員賞与	448,323	45%
98	受取手形割引高	21,134	2%
99	受取手形裏書譲渡高	26,218	3%
100	有形固定資産減価償却累計額	493,859	49%
101	減価償却実施額	480	0%
102	保証債務合計	478,325	48%
103	期末従業員数人	0	0%

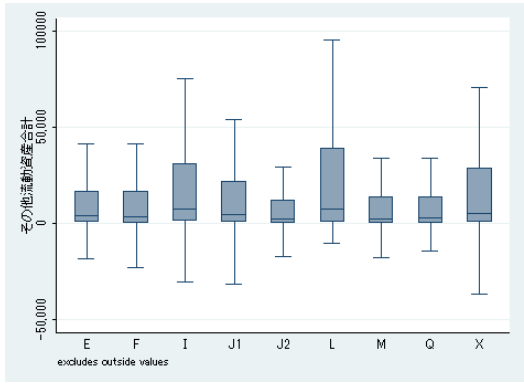
付録 B　業種別財務指標分布の箱ひげ図

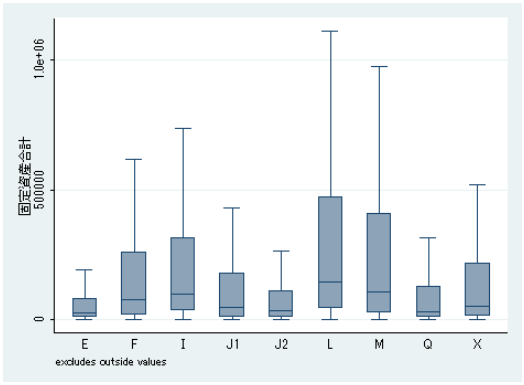


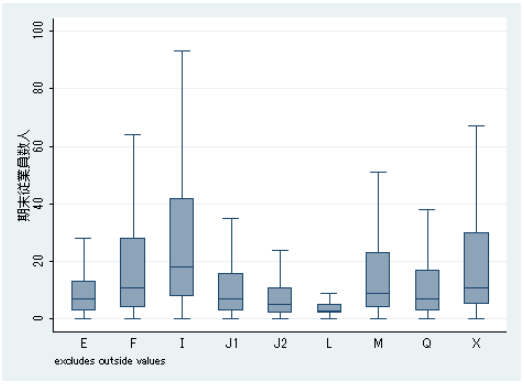


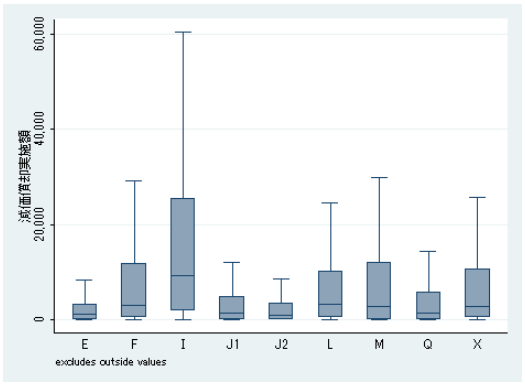




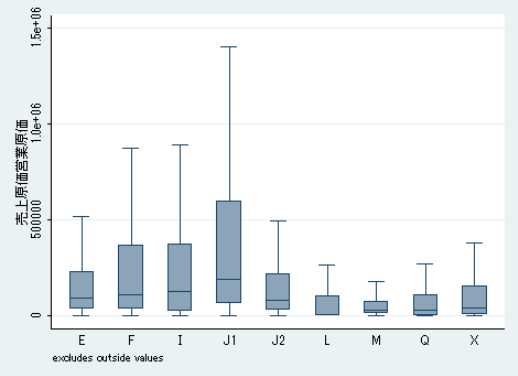


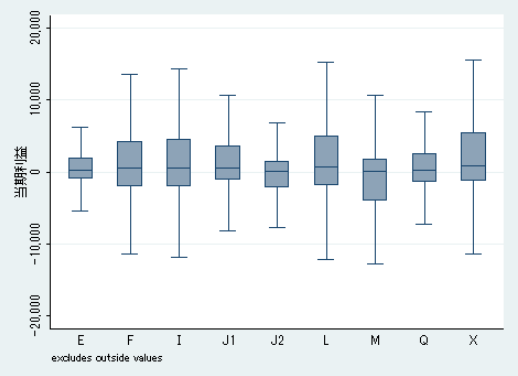


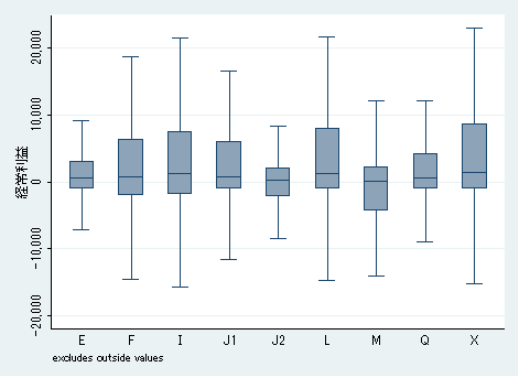


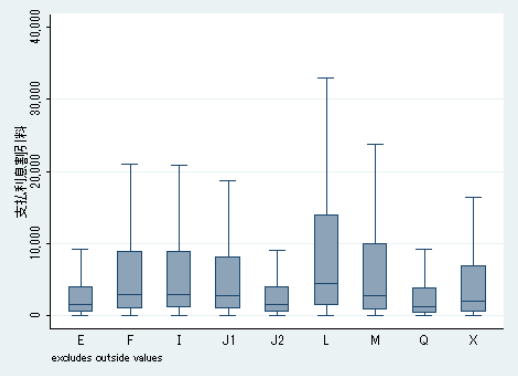


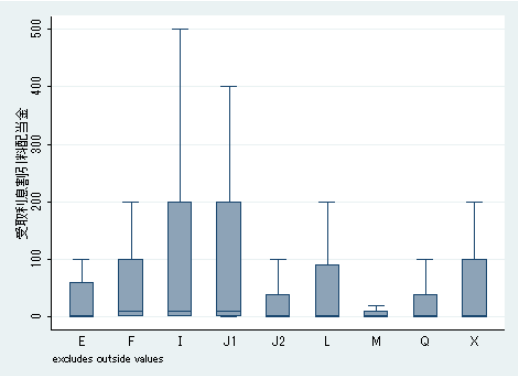
業種別財務指標分布の箱ひげ図（続き）

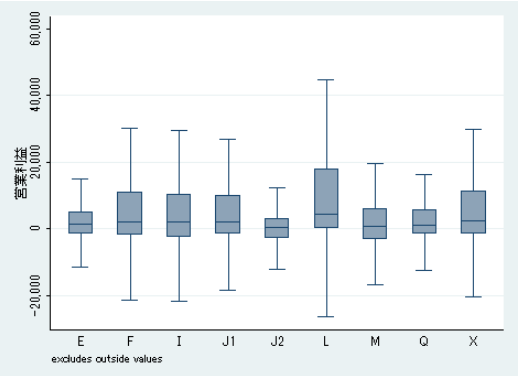


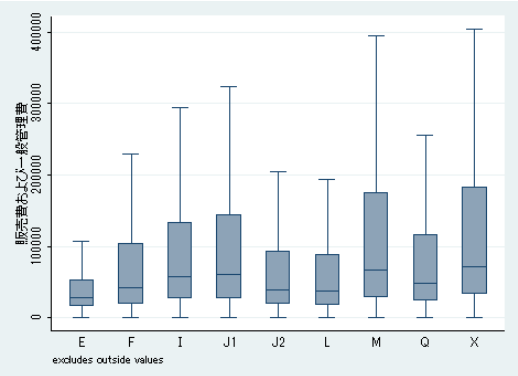


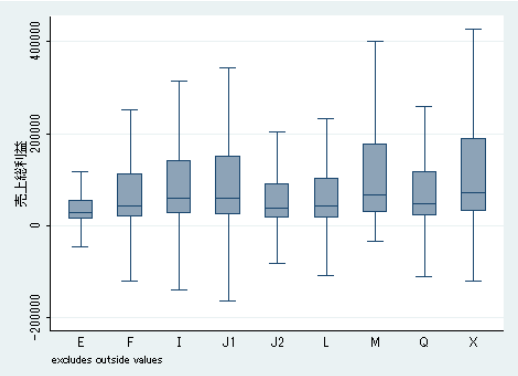




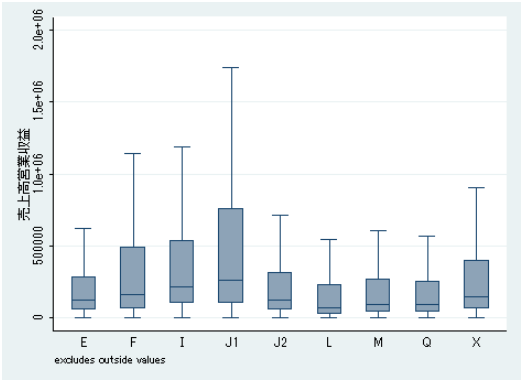


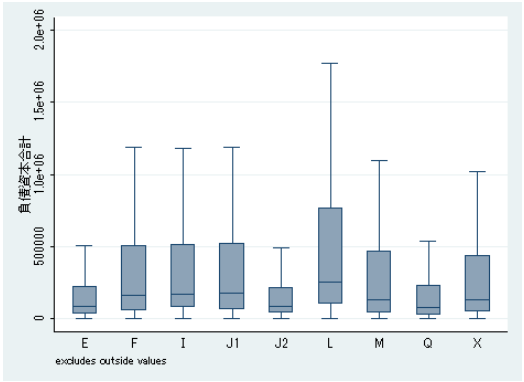


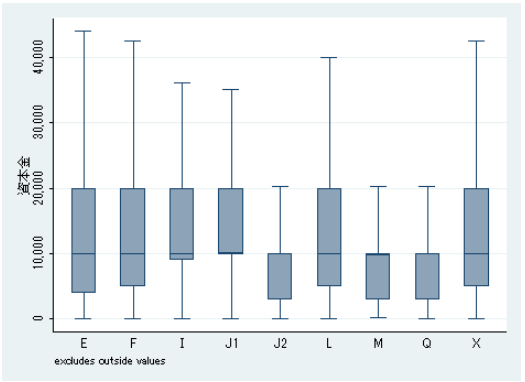


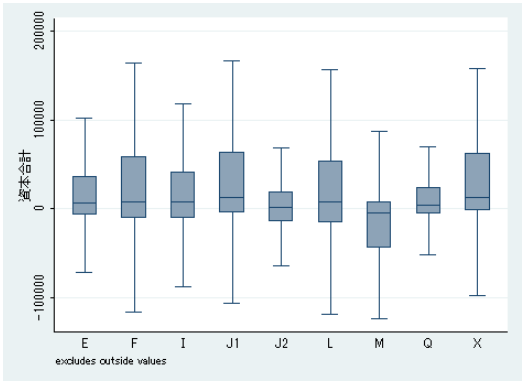


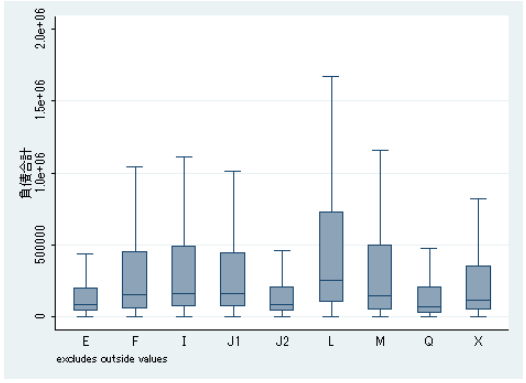
業種別財務指標分布の箱ひげ図（続き）

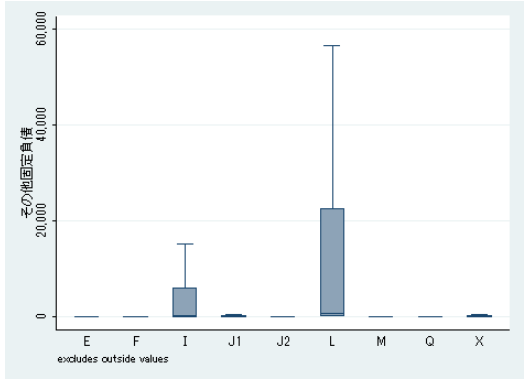


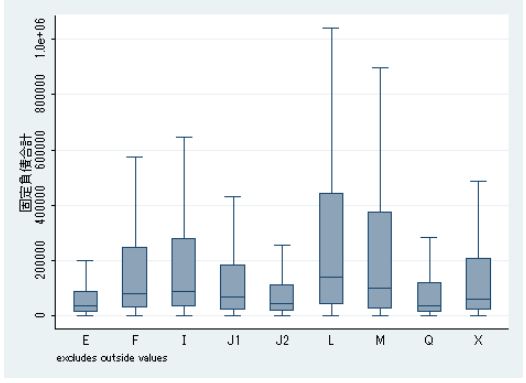


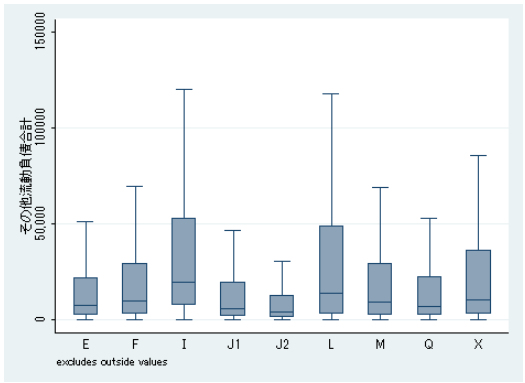




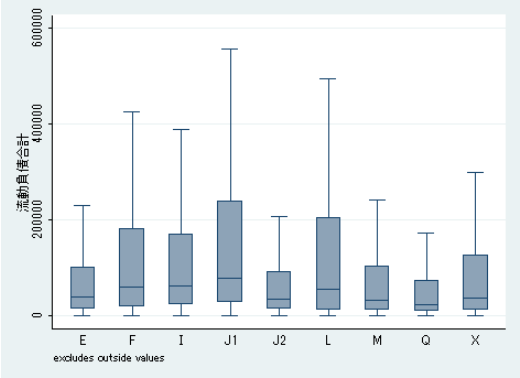


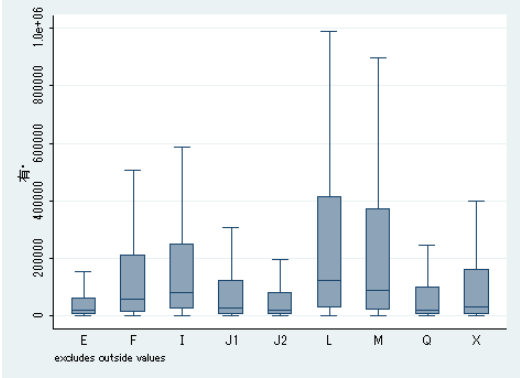


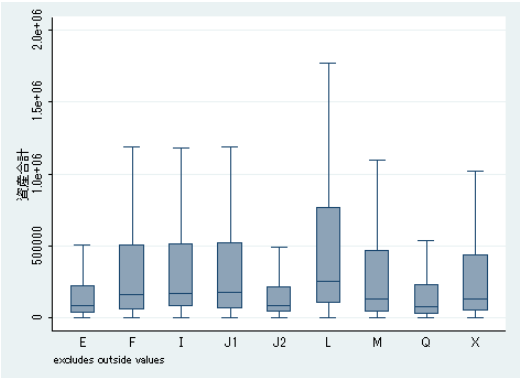


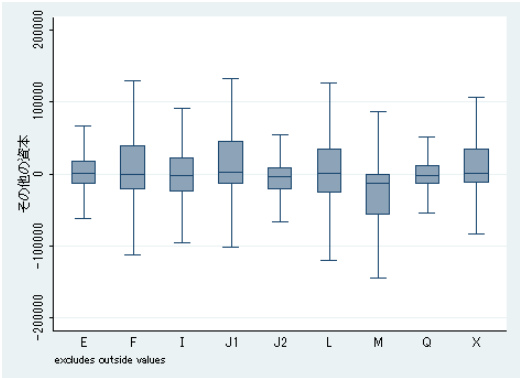


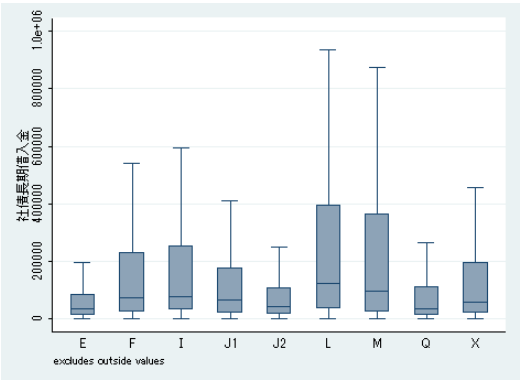
業種別財務指標分布の箱ひげ図（続き）

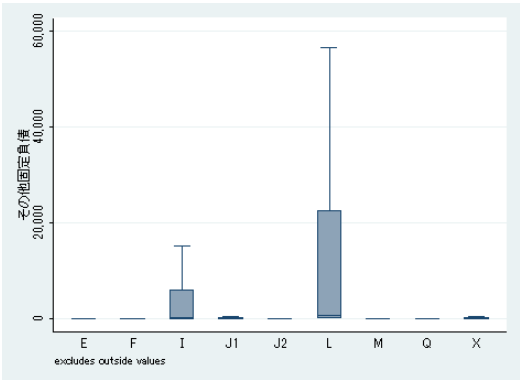


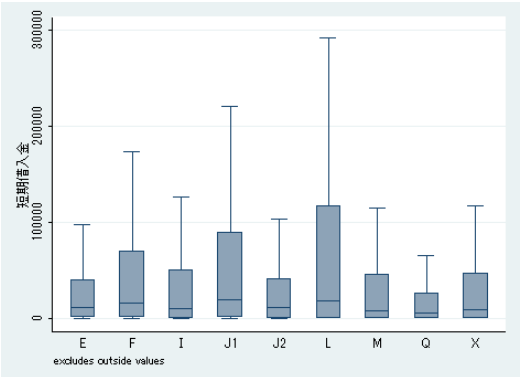


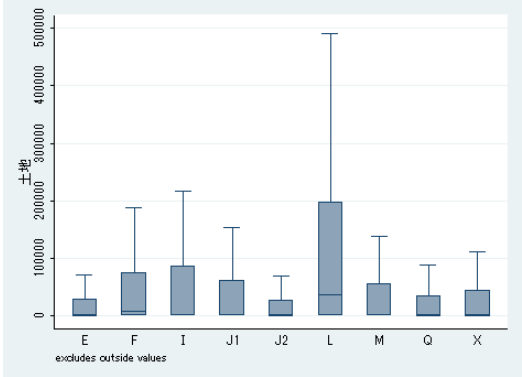




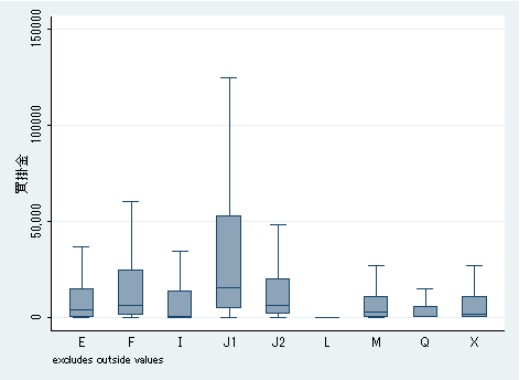


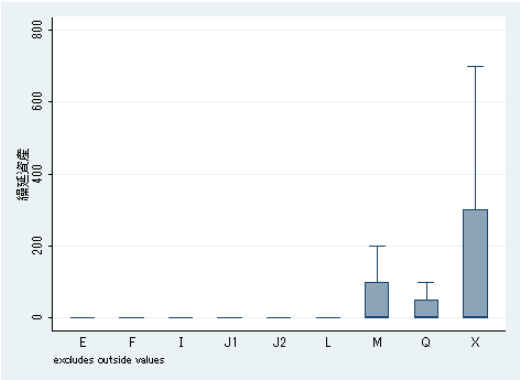






業種別財務指標分布の箱ひげ図（続き）





付録 C　一般化 neglog 変換による分布の正規化

一般化 neglog 変換を用いることで,歪度の大きい分布を正規分布に近づけることが可能となる. 一般化 neglog 変換は歪度を柔軟に調整できるため, 財務指標の正規化処理として広範に使われているものの歪度の調整ができない neglog 変換よりも有用であると考えられる. そこで, 本節では, 一般化 neglog 変換の λ を調整することで, 財務指標が最も正規化されることを示す.

正規化の評価関数として, (C-1)式のような, Shapiro-Wilk 検定統計量 W （Shapiro and Wilk（1965））を用いる.

$$W = \frac{\left(\sum a_i y_i\right)^2}{\sum (y_i - \bar{y})^2} \tag{C-1}$$

ここで $y_1 < y_2 < \cdots < y_n$ は，正規性を検定する n 個の順序づけられた標本を表す. また， $\mathbf{a} = (a_1, \cdots, a_n)^{\mathbf{T}}$ は， $\left(n-1\right)^{-1/2} \sum a_i y_i$ が正規性を想定する y_i の標準誤差の最良線形不偏推定量（BLUE）となる値である． $\mathbf{a}$ の正確な値は，(C-2)式のようになる.

$$\mathbf{a} = (\mathbf{m}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{m})^{-\frac{1}{2}}\mathbf{m}^{\mathbf{T}}\mathbf{V}^{-1} \tag{C-2}$$

ここで $\mathbf{V}$ は，期待値ベクトル $\mathbf{m}$ となる n 個の標準正規乱数からの，順序づけられた標本統計量の共分散行列である.

Roysten(1992)でも述べられているように,(C-1)式で表される Shapiro-Wilk 検定統計量 W は正規性からの乖離に関する検定として確立しており，大きければ大きいほど正規性を排除できないという指標である. 本節では，その W を最大にするように，グリッド法により一般化 neglog 変換の変換率 $\hat{\lambda}$ を決定する. 計算結果は表 C-1 のようになる. また，変換前の分布形と変換後の分布形を比較すると，図 C-1 のようになる.

表 C-1 財務指標別最適 $\hat{\lambda}$ とその時の $\hat{W}$

j	*Code.*	財務指標名	$\hat{\lambda}$	$\hat{W}$
1	sihyo3	総資本当期利益率(ROA)	−0.02	0.964763
2	sihyo10	総資本回転率	−0.44	0.998344
3	sihyo12	棚卸資産回転日数	0.06	0.989753
4	sihyo17	支払準備率	0.01	0.987400
5	sihyo18	現預金比率	−0.02	0.996473
6	sihyo19	自己資本比率	0.29	0.875107
7	sihyo23	デットキャパシティレイシオ	−0.04	0.961583
8	sihyo24	預借率	−0.03	0.992684
9	sihyo26	売上高支払利息割引料率	−0.77	0.991077
10	sihyo31	流動資産その他流動資産比率	−0.02	0.970777

図 C-1 変換前の財務指標の分布(左)と最適 $\hat{\lambda}$ で変換した時の分布(右)

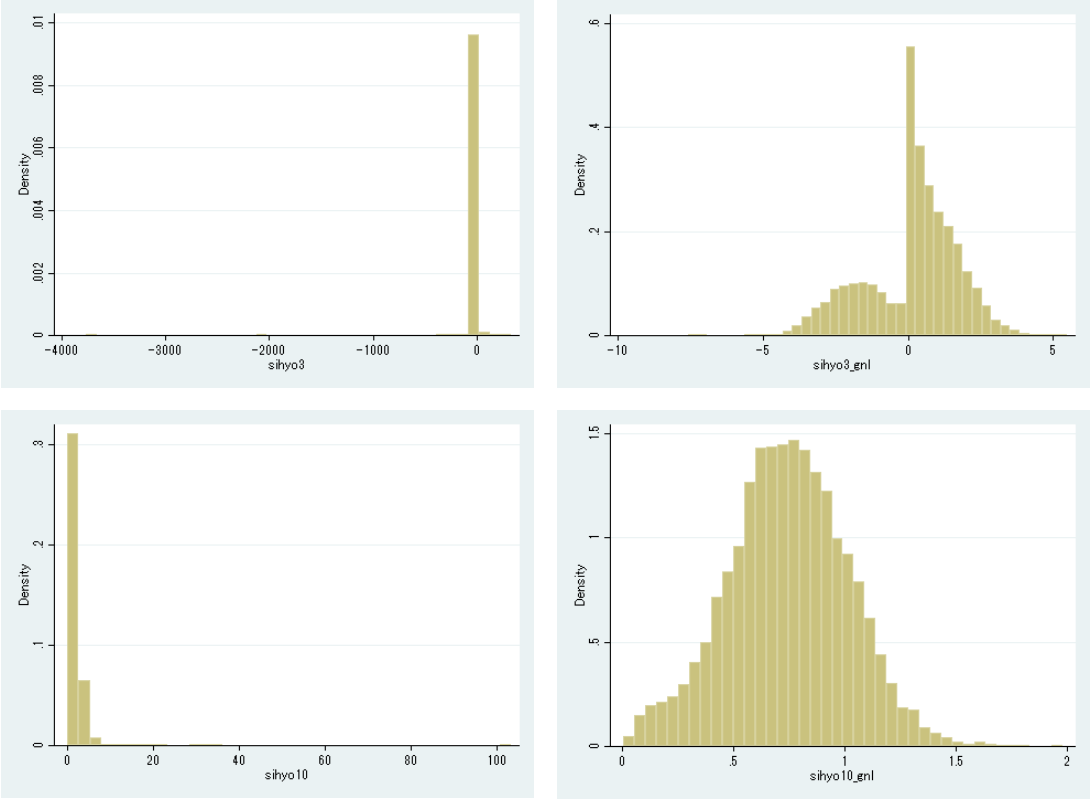
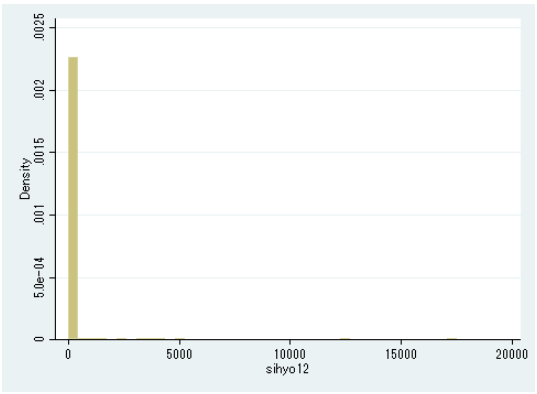
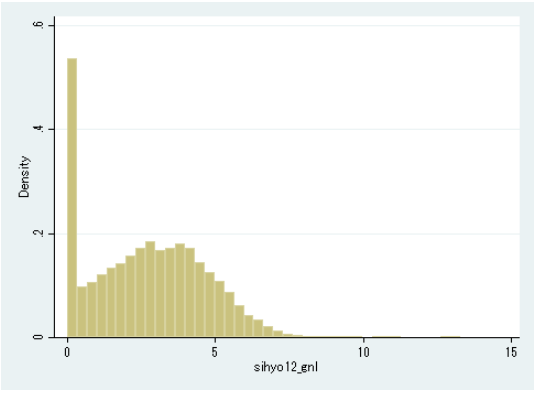
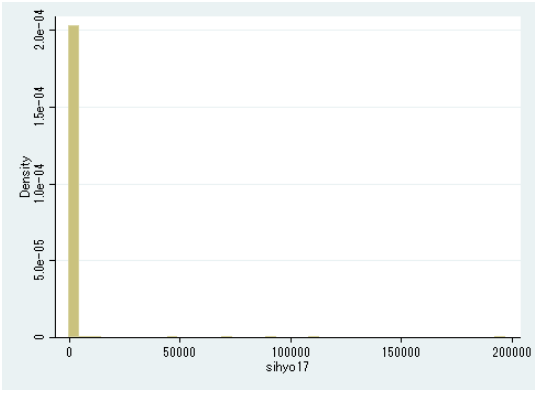
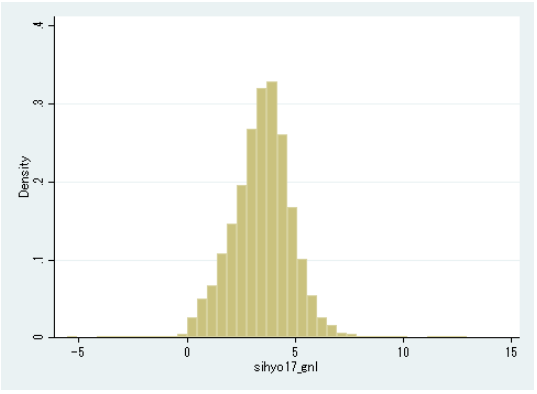


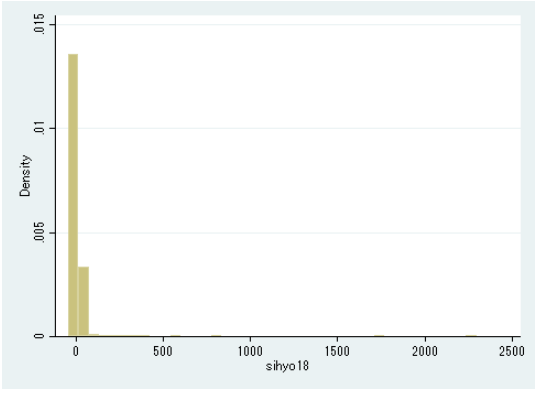
図 C-1　変換前の財務指標の分布（左）と最適 $\hat{\lambda}$ で変換した時の分布（右）（続き）

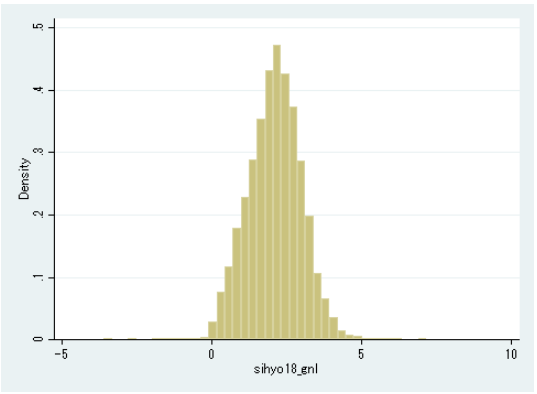


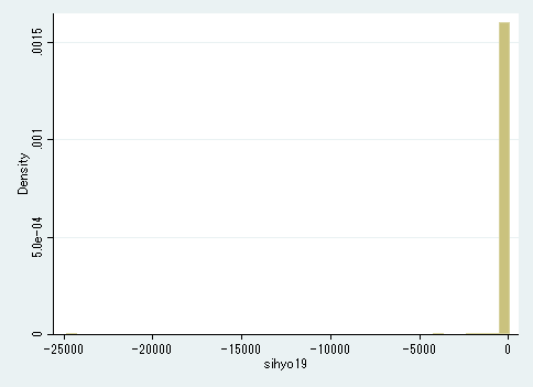












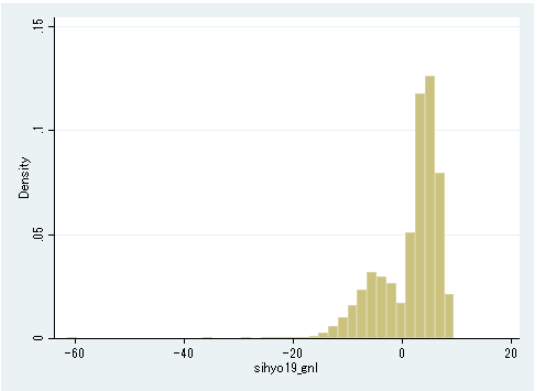
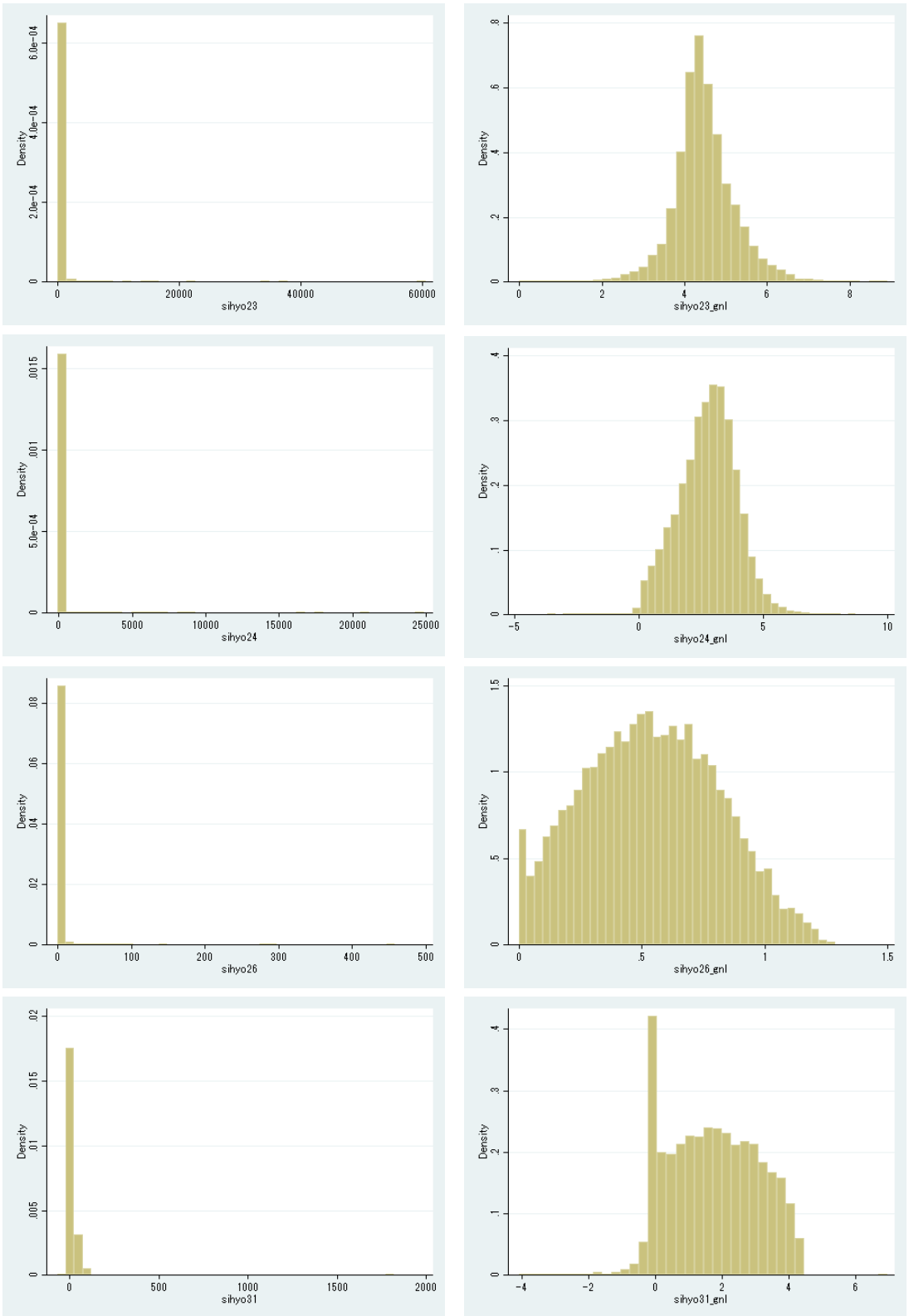


図 C-1　変換前の財務指標の分布（左）と最適 $\hat{\lambda}$ で変換した時の分布（右）（続き）



以上のように,財務指標は一般化 neglog 変換を用いることで正規化できることを示した．ただし，信用リスクモデルを推計する上で，正規化された財務指標を用いることが必ずしもモデルの予測精度を向上させるとは限らない．実際に，変換率 $\hat{\lambda}$ を用いて一般化 neglog 変換した財務指標に対し，5.2.2 節と同様のロジットモデル推計を行った場合，AUC= 0.8074 となった．この値は，5.3 節で示した neglog 変換の結果（AUC=0.8078）や単変量 2 項ロジットモデルで最適計算した場合の結果(AUC=0.8121)と比較して,低い AUC 水準となった．これらの情報を基に，5.2.2 節以降では，ロジットモデルの予測精度を最も向上させる一般化 neglog 変換の利用方法について論じた．

付録 D　財務指標分布の変換結果の比較

付録 D では，第 4 章及び第 5 章で利用した 10 財務指標に関し，以下のような変数変換後の分布を図示する．

- 変数変換なし（$\lambda=1,\theta^*$）
- neglog 変換（$\lambda=0,\theta^*$）
- 一般化 neglog 変換（λ^*,θ^*）

なお，図示する際の最適なθ^*はいずれも共通で，表 D-1（表 5-1 と同じ）のようになる．表 D-1 には，一般化 neglog 変換の最適なλ^*も示している．

表 D-1　パターン 8 の最適な(λ^*,θ^*)

j	財務指標名	λ *	θ *
1	総資本当期利益率(ROA)	0.32	–※
2	総資本回転率	−2.45	5.9
3	棚卸資産回転日数	0.62	0.3
4	支払準備率	−0.52	3.1
5	現預金比率	−0.69	2.5
6	自己資本比率	0.26	−
7	デットキャパシティレイシオ	−0.31	−
8	預借率	−0.27	1.8
9	売上高支払利息割引料率	−0.12	6.2
10	流動資産その他流動資産比率	0.59	3.9

※θ *の“－”は、最適計算の結果、折り返しなし(θ =0)となったことを意味する。

本章における図表の横軸は，コード表記となっている．図示している財務指標のコードは，表 D-2 のとおり．また，各財務指標コードの次に示されている“_nl”“_gnt”は，それぞれ neglog 変換，一般化 neglog 変換を表す．

表 D-2　指標コード一覧

Code.	財務指標名
sihyo3	総資本当期利益率(ROA)
sihyo10	総資本回転率
sihyo12	棚卸資産回転日数
sihyo17	支払準備率
sihyo18	現預金比率
sihyo19	自己資本比率
sihyo23	デットキャパシティレイシオ
sihyo24	預借率
sihyo26	売上高支払利息割引料率
sihyo31	流動資産その他流動資産比率

図 D-1　分布比較（左から無変換，neglog 変換，一般化 neglog 変換）

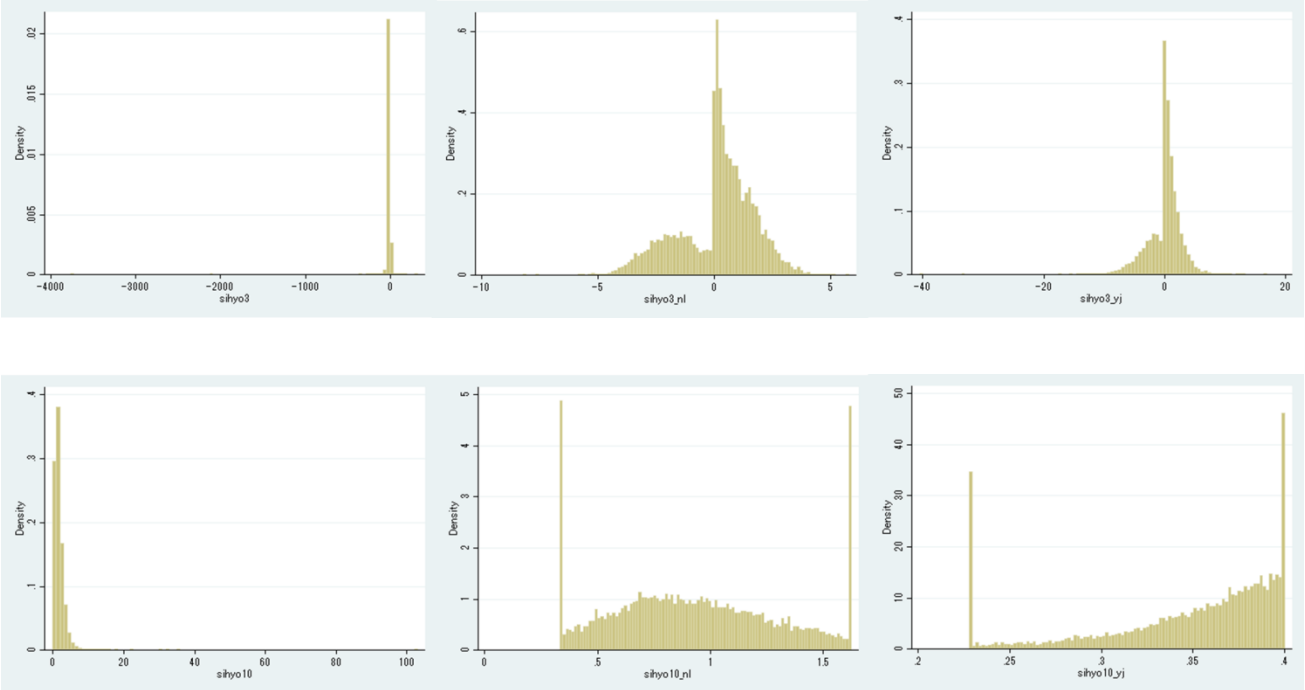


図 D-1 分布比較(左から無変換, neglog 変換, 一般化 neglog 変換)続き

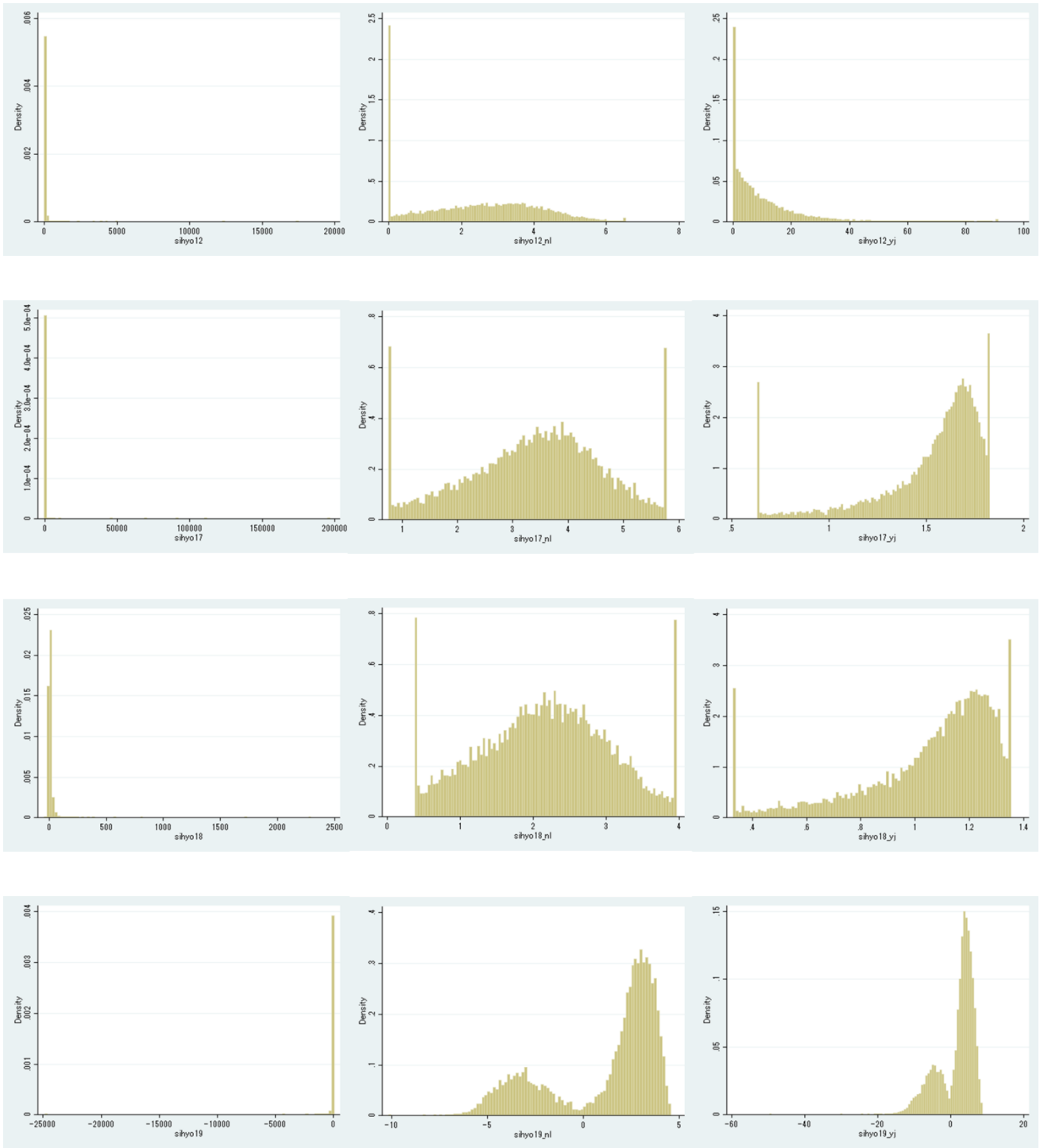


図 D-1 分布比較(左から無変換, neglog 変換, 一般化 neglog 変換)続き

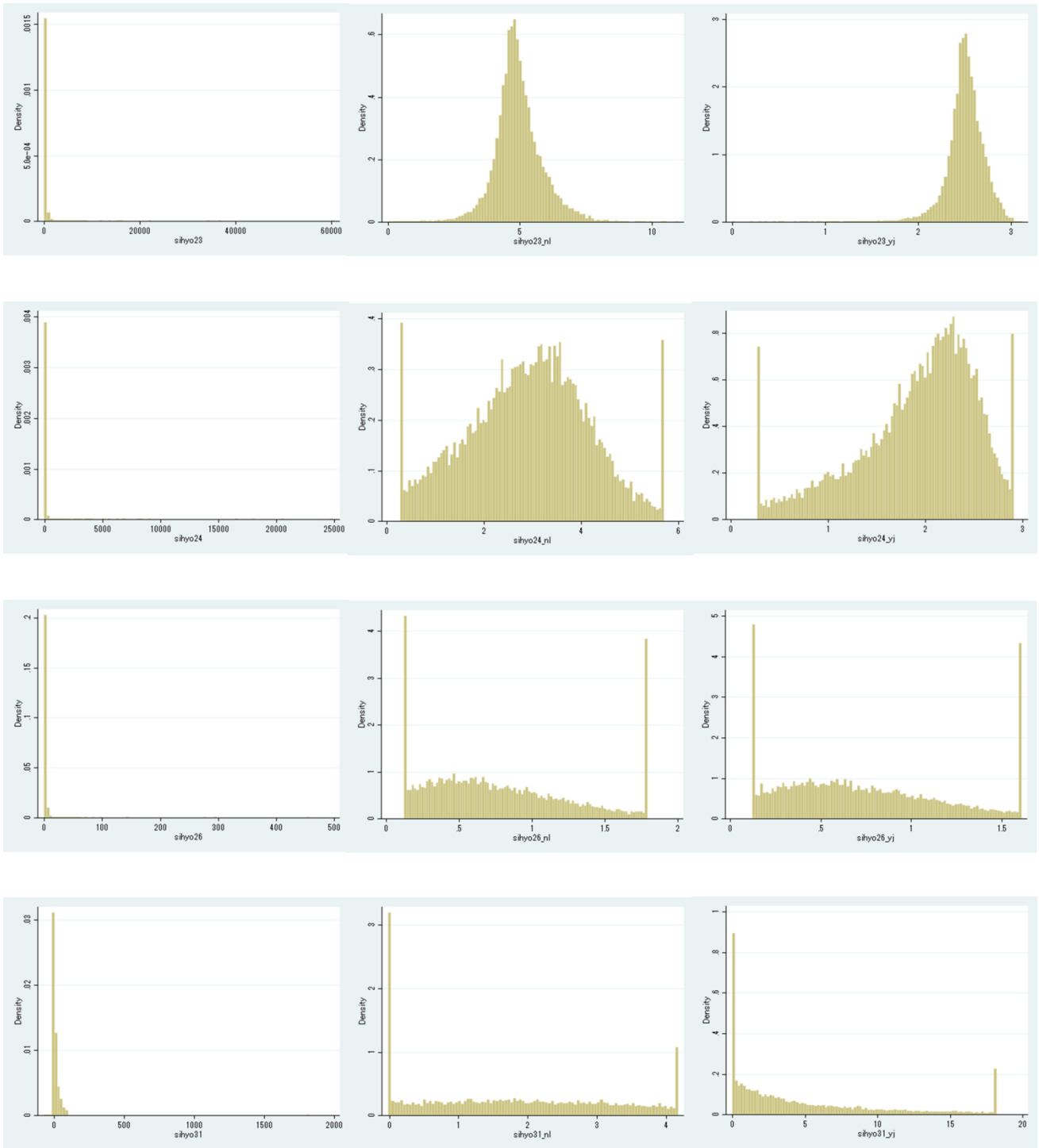


表 D-1 より，単変量ベースではあるが，全てのケースで一般化 neglog 変換のラムダは 1 を下回っており，尤度を最大にするためには，何らかの変換を行うことが有効であると言える．図表 D-1 の右側の図表は，尤度を最大にするラムダを適用した分布であるが，外れ値をかなり縮約した分布となっていることが見て取れる．

また，右側の一般化 neglog 変換のグラフを確認する限り，尤度を最大にする分布型は必ずしも正規分布に近い形状である必要は無いことが分かる．例えば，指標 17（支払準備率），指標 18（預金準備率），指標 23（デットキャパシティレシオ），指標 24（預借率）のように，neglog 変換（中央）の方が一般化 neglog 変換（右方）より正規分布に近いケースが見受けられる．

さらに，指標 3（総資本当期利益率（ROA）），指標 12（棚卸資産回転日数），指標 19（自己資本比率），指標 31（流動資産その他流動資産比率）のように，ラムダが 0 と 1 の間の値をとり，neglog 変換では変換率が強すぎる指標も見受けられる．

このように，既存の変数変換手法で最もよく利用される neglog 変換は，尤度を最大にする方法としては必ずしも適切でないケースが多いことは，本研究での発見である．

なお，表 D-2 に示している各財務指標の計算式は以下のとおり．

```
sihyo3 = (var95/var44)*100 // 総資本当期利益率(ROA)
sihyo10 = (var72/var44) // 総資本回転率
sihyo12 = (var16/var72)*365 // 棚卸資産回転日数
sihyo17 = (var12/var45)*100 // 支払準備率
sihyo18 = (var12/var72)*100 // 現預金比率
sihyo19 = (var64/var44)*100 // 自己資本比率
sihyo23 = ((var48+var59+var98)/(var12+var30))*100 // デットキャパシティレイシオ
replace sihyo23 = . if sihyo23 < 0 // 負数の場合は評価しない
sihyo24 = ((var12)/(var48+var59+var98))*100 // 預借率
sihyo26 = (var88)/(var72)*100 // 売上高支払利息割引料率
sihyo31 = (var21)/(var11)*100 // 流動資産その他流動資産比率
```

ここで，var＊の＊の数字は，表 A-2 の No.に対応している．

また，各財務指標の原数値，neglog 変換，一般化 neglog 変換の基礎統計量は，表 D-3 のとおり．

表 D-3 財務指標基礎統計量一覧

(1)変数変換なし

コード *Code.*	財務指標名 *name*	レコード数 *N*	最小値 *min*	中央値 *med*	最大値 *max*	平均値 *mean*	標準偏差 *sd*	尖度 *skewness*	歪度 *kurtosis*
sihyo3	総資本当期利益率(ROA)	12,000	−3766.7	0.4	334.0	−1.3	42.0	−70.7	5907.9
sihyo10	総資本回転率	12,000	0.0	1.5	103.2	1.8	1.8	17.4	827.2
sihyo12	棚卸資産回転日数	12,000	0.0	11.5	17526.3	41.8	227.8	53.6	3658.0
sihyo17	支払準備率	12,000	−218.2	32.1	197000.0	118.3	2372.1	64.1	4632.8
sihyo18	現預金比率	12,000	−41.1	7.8	2295.3	12.8	32.4	44.4	2767.9
sihyo19	自己資本比率	12,000	−24833.3	10.4	96.7	−2.7	240.4	−92.5	9489.2
sihyo23	デットキャパシティレイシオ	12,000	0.0	127.7	60596.0	253.7	908.2	37.6	2048.2
sihyo24	預借率	12,000	−47.9	18.5	24900.0	56.7	438.1	37.5	1719.8
sihyo26	売上高支払利息割引料率	12,000	0.0	1.0	458.0	1.9	6.6	42.7	2483.4
sihyo31	流動資産その他流動資産比率	12,000	−68.1	4.7	1820.0	13.4	25.6	30.1	2059.7

(2)neglog変換

コード *Code.*	財務指標名 *name*	レコード数 *N*	最小値 *min*	中央値 *med*	最大値 *max*	平均値 *mean*	標準偏差 *sd*	尖度 *skewness*	歪度 *kurtosis*
sihyo3	総資本当期利益率(ROA)	12,000	−8.2	0.3	5.8	0.1	1.6	−0.6	3.2
sihyo10	総資本回転率	12,000	0.3	0.9	1.6	0.9	0.4	0.2	2.2
sihyo12	棚卸資産回転日数	12,000	0.0	2.5	6.5	2.4	1.7	0.0	2.0
sihyo17	支払準備率	12,000	0.8	3.5	5.8	3.4	1.2	−0.2	2.6
sihyo18	現預金比率	12,000	0.4	2.2	4.0	2.2	0.9	0.0	2.4
sihyo19	自己資本比率	12,000	−10.1	2.4	4.6	1.1	2.9	−1.0	2.5
sihyo23	デットキャパシティレイシオ	12,000	0.0	4.9	11.0	4.9	0.9	0.3	5.8
sihyo24	預借率	12,000	0.3	3.0	5.7	2.9	1.2	−0.1	2.6
sihyo26	売上高支払利息割引料率	12,000	0.1	0.7	1.8	0.8	0.5	0.6	2.4
sihyo31	流動資産その他流動資産比率	12,000	0.0	1.7	4.2	1.8	1.3	0.2	1.8

(3)一般化neglog変換

コード *Code.*	財務指標名 *name*	レコード数 *N*	最小値 *min*	中央値 *med*	最大値 *max*	平均値 *mean*	標準偏差 *sd*	尖度 *skewness*	歪度 *kurtosis*
sihyo3	総資本当期利益率(ROA)	12,000	−40.4	0.3	17.0	0.1	2.7	−1.1	11.7
sihyo10	総資本回転率	12,000	0.2	0.4	0.4	0.3	0.0	−1.1	3.4
sihyo12	棚卸資産回転日数	12,000	0.0	6.1	91.3	10.3	13.0	2.5	11.7
sihyo17	支払準備率	12,000	0.6	1.6	1.8	1.5	0.3	−1.6	5.1
sihyo18	現預金比率	12,000	0.3	1.1	1.4	1.1	0.2	−1.2	4.0
sihyo19	自己資本比率	12,000	−49.6	3.4	8.8	1.6	5.1	−1.2	4.5
sihyo23	デットキャパシティレイシオ	12,000	0.0	2.5	3.1	2.5	0.2	−2.2	17.4
sihyo24	預借率	12,000	0.3	2.0	2.9	1.9	0.6	−0.8	3.2
sihyo26	売上高支払利息割引料率	12,000	0.1	0.7	1.6	0.7	0.4	0.5	2.3
sihyo31	流動資産その他流動資産比率	12,000	0.0	3.0	18.2	5.0	5.3	1.1	3.2

付録 E　主要計算ロジックに関する STATA の計算コード

本論文では，ほとんどの計算を統計ソフト STATA で行った．付録 E では，本論文中の主な計算ロジックに関して，STATA のコードをまとめておく．

E-1．k-NN 計算

E-1-1．レコード間の距離計算（売上高ランク分け法）

```
foreach rank of num #{
  foreach j of num # {
    forvalues b=#/#{
      qui : use `j'_`b', clear
      qui: egen rank=cut(var72), g(`rank')
      foreach m of numlist 11/14 16 21 29 30 34 43/49 58/61 63/66 71/73 78 79 83 85 88 90 95 98 99 101 103{
        qui: summarize var`m'_miss
        scalar mean`m'=r(mean)
        scalar sd`m'=r(sd)
        qui: ge var`m'_std = ( var`m'_miss - mean`m' ) / sd`m'
      }

      sort idy
      local k=#
      local n1 1
      local n=_N
      tempname disproc
      postfile `disproc' elem1 elem2 distan using distant`k'_`j'_`b', replace
      display "Calculating Euclidian Distance..."
      while `n1'<=`n' {
        local n2 1
        while `n2'<=(`n'-`n1') {
          local xx=`n1'+`n2'
          local dist 0
          local var 1
          local k12 0
          if (rank[`n1']==(rank[`xx']))|(rank[`n1']==(rank[`xx']+1))|(rank[`n1']==(rank[`xx']-1)){
            while `var'<=`k' {
              if x`var'_std[`n1']!=. & x`var'_std[`xx']!=. {
                local dist=`dist'+abs(x`var'_std[`n1']-x`var'_std[`xx'])^(2)
                local k12=`k12'+1
              }
            local var=`var'+1
            }
          }
          else{
            local dist=.
          }
          if `k12'==0 {
            local dist=.
          }
```

```
            else {
               local dist=(`dist'/`k12')^(1/2)
            }
            if `dist'!=.{
               post `disproc' (`n1') (`xx') (`dist')
            }
            local n2=`n2'+1
         }
         di "`b'_`n1'"
         local n1=`n1'+1
      }
      postclose `disproc'
   }
  }
}
```

E-1-2.　k-NN 法による補完

```
foreach a of numlist 1/8{
  qui:use complete_std_distan_`j'_`b', clear

  local K=`a'
  local n=_N
  forvalues i = 1/`n' {
    foreach m of numlist 11/14 16 21 29 30 34 43/49 58/61 63/66 71/73 78 79 83 85 88 90 95 98 99 101 103{
      forvalues k = 1/`K' {
        local temp`k' = elem2`k'[`i']
        local imp_dist_`k' = var`m'_std_0[`temp`k'']*(1/(distan`k'[`i']))
        local distan`k' = var`m'_nonmiss[`temp`k'']*(1/(distan`k'[`i']))
      }
      local distan0_sum = 0
      local k = 1
      while `k'<=`K'{
        local l=`k' - 1
        local distan`k'_sum = `distan`l'_sum' + `distan`k''
        local k = `k' + 1
      }
      local imp_dist_0_sum = 0
      local k = 1
      while `k'<=`K'{
        local l=`k' - 1
        local imp_dist_`k'_sum = `imp_dist_`l'_sum' + (`imp_dist_`k''/`distan`K'_sum')
        local k = `k' + 1
      }
      qui: replace var`m'_std_imp = `imp_dist_`K'_sum' if var`m'_std==.&elem1==`i'
    }
  }
}
```

E-1-3.　NRMSE の計算

```
foreach m of numlist 11/14 16 21 29 30 34 43/49 58/61 63/66 71/73 78 79 83 85 88 90 95 98 99 101 103{
   qui: ge var`m'_diff=.
   qui: replace var`m'_diff = ((var`m' - var`m'_imp)^2)/((sd`m')^2) if var`m'_std==.
   qui: replace var`m'_diff = 0 if (abs(var`m' - var`m'_imp)<1&var`m' - var`m'_imp!=.)&var`m'_std==.
}

qui:egen rowtotal_diff = rowtotal(var11_diff-var103_diff), missing
qui:egen rownonmiss_diff = rownonmiss(var11_diff-var103_diff)
qui:egen total_value_diff = total(rowtotal_diff)
qui:egen total_count_diff = total(rownonmiss_diff)

qui:egen total_value_diff_ndf = total(rowtotal_diff) if max_dfflg==0
qui:egen total_count_diff_ndf = total(rownonmiss_diff) if max_dfflg==0

qui:egen total_value_diff_df = total(rowtotal_diff) if max_dfflg==1
qui:egen total_count_diff_df = total(rownonmiss_diff) if max_dfflg==1

qui:ge nrmse_k`a'=sqrt(total_value_diff/total_count_diff)
qui:ge nrmse_k`a'_ndf=sqrt(total_value_diff_ndf/total_count_diff_ndf)
qui:ge nrmse_k`a'_df=sqrt(total_value_diff_df/total_count_diff_df)

scalar nrmse2_`a' = nrmse_k`a'_df[1]
scalar list
```

E-2.　グリッド法による最適折り返しポイント θ^* 及び一般化 neglog 変換の最適変換率 λ^* の決定

```
local count_cut = 0

forvalues cut = 0.1(0.1)10{
   local count_cut =`count_cut'+1
   di "cut`cut'%"
   use $filename, clear

   local locut_num = 12000*(`cut'/100)
   local hicut_num = 12000*(1-(`cut'/100))

   foreach var in sihyo3 sihyo10 sihyo12 sihyo17 sihyo18 sihyo19 sihyo23 sihyo24 sihyo26 sihyo31{
      sort `var'
      scalar locut_`var'=`var'[`locut_num']
      scalar hicut_`var'=`var'[`hicut_num']
      ge `var'_cut = `var'
      replace `var'_cut = locut_`var' if `var' <= locut_`var'
      replace `var'_cut = hicut_`var' if `var' >= hicut_`var'
   }
```

E-2.（続き）

```
qui: logit dfflg sihyo3_cut-sihyo31_cut
qui: lroc, nog
matrix auc_`count_cut'_p1 = r(area)

save yj_$filename , replace

*******************************************
global varname "sihyo3_cut sihyo10_cut sihyo12_cut sihyo17_cut sihyo18_cut sihyo19_cut /*
*/sihyo23_cut sihyo24_cut sihyo26_cut sihyo31_cut"
global depvar "dfflg"
global filename "outlier_crdit_12000_a1"
*******************************************
use yj_$filename, clear

foreach var in $varname{
   di "◆`var'"

   ge `var'_yj = .

   local i = 0
   forvalues k = -10(0.1)10{
      local i = `i' +1
      if abs(`k') > 1e-10 {
         qui: replace `var'_yj = (sign(`var')*(exp(`k'*log(abs(`var')+1))-1))/`k'
      }
      qui: else replace `var'_yj = sign(`var')*log(abs(`var')+1)

      qui: logit $depvar `var'_yj,  iterate(100)
      capture matrix drop ll_`i'
      matrix ll_`i' = (e(ll), `k')
   }

   capture matrix drop result_ll_`var'
   forvalues i = 1(1)201{
      matrix result_ll_`var' = (nullmat(result_ll_`var') ¥ ll_`i')
   }
}

foreach var in $varname{
   clear
   svmat result_ll_`var'
   gsort -result_ll_`var'1
   local max_ll_`var' = result_ll_`var'2[1]
}

*******************************************
```

E-2.（続き）

```
foreach var in $varname{
   di "◆`var'"

   use yj_$filename, clear
   local min_`var' = `max_ll_`var'' - 0.1
   local max_`var' = `max_ll_`var'' + 0.1

   ge `var'_yj = .
   local i = 0
   forvalues k = `min_`var''(0.01)`max_`var''{
      local i = `i' +1
      if abs(`k') > 1e-10 {
         qui: replace `var'_yj = (sign(`var')*(exp(`k'*log(abs(`var')+1))-1))/`k'
      }
      qui: else replace `var'_yj = sign(`var')*log(abs(`var')+1)

      qui: logit $depvar `var'_yj,  iterate(100)
      capture matrix drop ll2_`i'
      matrix ll2_`i' = (e(ll), `k')
      qui: lroc, nog
      matrix ll2_`i' = (ll2_`i', r(area))
   }

   capture matrix drop result_ll2_`var'
   forvalues i = 1(1)20{
      matrix result_ll2_`var' = (nullmat(result_ll2_`var') ¥ ll2_`i')
   }
}

foreach var in $varname{
   clear
   svmat result_ll2_`var'
   gsort -result_ll2_`var'1
   local max_ll2_`var' = result_ll2_`var'2[1]
   matrix max_ll2_`var' = `max_ll2_`var''

   gsort -result_ll2_`var'3
   local max_k_auc_`var' = result_ll2_`var'2[1]
   matrix max_k_auc_`var' = `max_k_auc_`var''
   local max_auc_`var' = result_ll2_`var'3[1]
   matrix max_auc_`var' = `max_auc_`var''
}
capture matrix drop result_auc_`count_cut'
foreach var in $varname{
   matrix result_auc_`count_cut' = (nullmat(result_auc_`count_cut') , max_k_auc_`var', max_auc_`var')
}

*****************************************
```

E-2.（続き）

```
    use yj_$filename, clear

    foreach var in $varname{
        if abs(`max_ll2_`var'') > 1e-10 {
            qui: ge `var'_yj = (sign(`var')*(exp(`max_ll2_`var''*log(abs(`var')+1))-1))/`max_ll2_`var''
        }
        qui: else ge `var'_yj = sign(`var')*log(abs(`var')+1)
    }

    qui: logit dfflg sihyo3_cut_yj-sihyo31_cut_yj,  iterate(100)
    qui: lroc, nog
    matrix auc_`count_cut'_p2 = r(area)

    capture matrix drop result_`count_cut'
    matrix result_`count_cut' = (auc_`count_cut'_p1, auc_`count_cut'_p2)
    foreach var in $varname{
        matrix result_`count_cut' = (result_`count_cut', max_ll2_`var')
    }
}

capture matrix drop result_all
forvalues count_cut = 0(1)100{
    matrix result_all = (nullmat(result_all) ¥ result_`count_cut')
}

capture matrix drop result_auc
forvalues count_cut = 0(1)100{
    matrix result_auc = (nullmat(result_auc) ¥ result_auc_`count_cut')
}

matrix list result_all
matrix list result_auc
```

謝辞

本論文の執筆に当たり，指導教官の山下智志教授からは，論文作成に必要な初歩的なことから親身にご相談に乗っていただき，丁寧かつ熱心なご指導をいただきました．副指導教官である川崎能典教授には，統計学と実際のデータを扱う際の考え方等について，多くの知見をご教授いただきました．同じく副指導教官である逸見昌之准教授には，統計的思考方法について丁寧にご指導をいただきました．また，本論文作成に必要不可欠なデータをご提供いただいた一般社団法人 CRD 協会を始め，ご協力いただいた皆様へ，心から感謝の気持ちと御礼を申し上げたく，謝辞にかえさせていただきます．

參考文獻

Acuna, E. and C. Rodriguez (2004), “The treatment of missing values and its effect in the classifier accuracy”.(http://academic.uprm.edu/~eacuna/IFCS04r.pdf)

Aittokallio, T. (2010), “Dealing with missing values in large-scale studies: microarray data imputation and beyond,” *Briefings in Bioinformatics*, **11**, No.2. 253-264.

Allison, P. D. (2001), *Missing Data.*, Sage University Papers Series on Quantitative Applications in the Social Sciences., Thousand Oaks, CA: Sage.

Altman, E. I. (1968), “Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy”, *The Journal of Finance*, **23**(4), 589-609.

Box, G. E. P. and D. R. Cox (1964), “An analysis of transformations”, *Journal of Royal Statistical Society*, Series B, **26**, 211–252.

Burgette LF and JP. Reiter (2010), “Multiple imputation for missing data via sequential regression trees,” *American Journal of Epidemiology*, **172**, 1070-76.

Crookston, N.L. and A.O. Finley (2008). “yaImpute: An R package for kNN imputation,” *Journal of Statistical Software*, **23**(10), 1-16.

Dixon, W. J. and K. K. Yuen (1974), “Trimming and winsorization: A review”, *Statistische Hefte*, **15**, 157-170.

Gan, X., AW. Liew and H. Yan (2006). “Microarray missing data imputation based on a set theoretic framework and biological knowledge,” *Nucleic Acids Research*. **34**, 1608-1619.

Hsu, Yang and Lu (2011). “KNN-DTW Based Missing Value Imputation for Microarray Time Series Data,” *Journal of Computers*, **6**, No.3, 418-425.

Jönsson, P. and C. Wohlin (2004), An evaluation of k-nearest neighbour imputation using Likert data, *Software Metrics, Proceedings. 10th International Symposium on*, *108 – 118.*

Jörnsten R, HY. Wang, WJ. Welsh and M. Ouyang (2005), “DNA microarray data imputation and significance analysis of differential expression,” *Bioinformatics*. **21**, 4155-4161.

Kim, H., G.H. Golub and H. Park (2004), "Imputation of missing values in DNA microarray gene expression data," *in: Proc. of the IEEE Computational Syst. Bioinformatics. Conf.*, 572-573.

Kim, H., G.H. Golub and H. Park (2005), "Missing value estimation for DNA microarray gene expression data: local least squares imputation," *Bioinformatics*, **21**, no.2, 187-198

Kyaw, K. and H. Zhang (2014), "OWNERSHIP STRUCTURE AND EARNINGS ANNOUNCEMENTS: Evidence from China," *Economics and Finance Review*, **4**, No.2, 1-9.

Lane, W. R., S. W. Looney and J. W. Wansley (1986), "An application of the cox proportional hazards model to bank failure", *Journal of Banking and Finance*, **10**(4), 511-531.

Little, R. (1988), "A Test of Missing Completely at Random for Multivariate Data With Missing Values," *Journal of the American Statistical Association*, **83**, No.404, 1198-1202.

Little, R. and D. B. Rubin (2002), *Statistical Analysis with Missing Data*, Second edition, John Wiley & Sons, Inc.

Liew, AW., NF. Law and H. Yan (2011), "Missing value imputation for gene expression data: computational techniques to recover missing data from available information," *Brief Bioinform*, **12**, 498-513.

Maronna, R., R. Martin and V. Yohai (2006), *Robust Statistics - Theory and Methods.*, John Wiley & Sons, Inc.

Martin, D. (1977), "Early warning of bank failure: A logit regression approach", *Journal of Banking and Finance*, **1**(3), 249-276.

Meinl, T. and E. W. Sun (2015), "Methods of Denoising Financial Data", *Handbook of Financial Econometrics and Statistics*, 519-538.

Moorthy, K., M.S. Mohamad and S. Deris, (2014), "A Review on Missing Value Imputation Algorithms for Microarray Gene Expression Data," *Current*

Bioinformatics, **9**, 18-22.

Oba S, MA. Sato, I. Takemasa, M. Monden, K. Matsubara and S. Ishii (2003). "A Bayesian missing value estimation method for gene expression profile data," *Bioinformatics*, **19,** 2088-2096.

Ouyang M, WJ. Welsh and P. Georgopoulos (2004). "Gaussian mixture clustering and imputation of microarray data," *Bioinformatics*, **20**, 917-923.

Peltonen, S., M. Gabbouj, and J. Astola (2001), "Nonlinear filter design: Methodologies and challenges", *In Proceedings of the 2nd International Symposium on Image and Signal Processing and Analysis*, ISPA 2001, 102–107.

Pyo, G. and Y. Chung (2015), "Inventory Changes With Information Asymmetry And Informed Trading Patterns In The Korean Stock Market," *The Journal of Applied Business Research*, **31**(2), 701-714.

Royston, P. (1992), "Approximationg the Shapiro-Wilk W-test for non-normality", *Statistics and Conputin*, **2**, 117–119.

Royston, P. (2005), "Multiple imputation of missing values: update," *Stata Journal*, **5**(2), 1–14.

Rubin, D. B. (1976), "Inference and missing data," *Biometrika*, **63**, 581-592.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, Inc.

Sehgal, MSB., I. Gondal and LS. Dooley (2005). "Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data," *Bioinformatics*. **21**, 2417-2423.

Sehgal, MSB., I. Gondal, LS. Dooley and R. Coppel (2008). "Ameliorative missing value imputation for robust biological knowledge inference," *Journal of Biomedical Informatics*. **41**, 499-514.

Shapiro, S. S. and M. B. Wilk (1965). "An analysis of variance test for normality (complete samples)", *Biometrika*. **52**(3,4), 591-611.

Takayasu, H. and K. Okuyama (1998), "Country Dependence on Company Size

Distributions and a Numerical Model Based on Competition and Cooperation", *Fractals*, **06**, 67-79.

Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R.B. Altman (2001), "Missing value estimation methods for DNA microarrays," *Bioinformatics* **17**(6), 520–525.

Tuikkala, J., L. Elo, OS. Nevalainen and T. Aittokallio (2006). "Improving missing value estimation in microarray data with gene ontology," *Bioinformatics*. **22**, 566-572.

van Buuren, S., H. C. Boshuizen and D. L. Knook (1999), "Multiple imputation of missing blood pressure covariates in survival analysis," *Statistics in Medicine* **18**: 681–694.

Whittaker, J., C. Whitehead, and M. Somers (2005), "The neglog transformation and quantile regression for the analysis of a large credit scoring database", *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**(5): 863–878.

Xiang, Q., X. Dai, Y. Deng, C. He, J. Wang, J. Feng and Z. Dai (2008). "Missing value imputation for microarray gene expression data using histone acetylation information," *BMC Bioinformatics*. **9**, 252-269.

Yeo, I-K. and R. A. Johnson (2000), "A new family of power transformations to improve normality or symmetry", *Biometrika*, **87**(4), 954–959.

Zhang, X., X. Song, H. Wang and H. Zhang (2008), "Sequential local least squares imputation estimating missing value of microarray data," *Computers in Biology and Medicine*. **38**, 1112-1120.

今井健太郎(2013)「共同 DB における欠損値解析法の利用」SAS ユーザー総会 2013.（http://www.riskdatabank.co.jp/rdb/library/files/20130718MissingDataAnalysis2.pdf）

岩崎学（2010）『不完全データの統計解析』エコノミスト社.

金子拓也（2005）「データマイニングにおける新しい欠損値補完方法の提案」『電子情報通信学会論文誌』，**J88-D-Ⅱ**，No.4，675-686.

高橋久尚・山下智志（2002）「大規模データによるデフォルト確率の推定－中小企

業信用リスクデータベースを用いてー」『統計数理』，**50**，2，241-258.

高橋淳一・山下智志（2015）「大規模財務諸表データに対する k-NN 法による欠損値補完」『ジャフィージャーナル（ファイナンスとデータ解析）』朝倉書店

田村晃一・柿元健・戸田航史・角田雅照・門田暁人・松本健一・大杉直樹（2009）「工数予測における類似性に基づく欠損値補完法の実験的評価」『コンピュータソフトウェア』，**26**，No.3，44-55.

藤澤洋徳（2006）『確率と統計』朝倉書店.

宮本道子・山下智志・安藤雅和・逸見昌之・高橋淳一（2012）「中小企業大規模財務データベースの欠測処理に対する問題点と対策について」『2012 年度統計関連学会連合大会報告集』.

森平爽一郎（2009）『信用リスクモデリング―測定と管理』朝倉書店.

森平爽一郎・岡崎貫治（2009）「マクロ経済効果を考慮したデフォルト確率の期間構造推定」『2009 年度日本ファイナンス学会第 17 回大会予稿集』，103-112.

山下智志・川口昇（2003）「大規模データベースを用いた信用リスク計測の問題点と対策（変数選択とデータ量の関係）」『金融研究センターディスカッションペーパー（2003 年 2 月 18 日）』.

山下智志・三浦翔（2011）『信用リスクモデルの予測精度－AR 値と評価指標－』朝倉書店.