氏　　名	髙橋 淳一
学位(専攻分野)	博士(統計科学)
学 位 記 番 号	総研大甲第 1792 号
学位授与の日付	平成27年9月28日
学位授与の要件	複合科学研究科　統計科学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	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
論 文 審 査 委 員	主　査　教授 川崎 能典 教授 山下 智志 准教授 逸見 昌之 教授 津田 博史 同志社大学

(別紙様式 2)
(Separate Form 2)

論文内容の要旨
Summary of thesis contents

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，

(別紙様式 2)
(Separate Form 2)

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．

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，今後の研究課題となる．

(別紙様式 3)
(Separate Form 3)

博士論文の審査結果の要旨
Summary of the results of the doctoral thesis screening

高橋淳一氏の博士論文審査を，2015年8月6日午後2時から約2時間にわたって，本人と4名の委員全員の出席のもとに行い，論文発表会および審査のための会議を行った．

論文は全 6 章 95 ページからなり，日本語で執筆されている．目的は，企業財務データに含まれる欠損値の補完と外れ値の補正に関する方法を提案し，大規模データの効率的処理の観点からその有用性を実証することである．

第 1 章では，財務データの特徴を説明した上で，欠損値や外れ値を補正する意義を述べつつ本論文で発案する手法を紹介し，他の方法と比較する際の評価基準を定めている．

第2章では，欠測値補完に関する既存研究を整理した後，大規模財務データにk-最近傍法(以下k-NN法)を適用する際の組合せ爆発的な計算量の問題に対し,売上高によるランク分けに基づき類似レコードのみを距離計算対象とする効率的な計算法を提案し，その補完精度を数値実験で示している．実験によると，1ランクあたり300レコードあれば補完精度は十分保証され，最近傍の構成はk=3が平均的に誤差最小となることが示されており，連鎖回帰型補完や時系列補間より精度が良い．

第 3 章では，k-NN 法による欠損値補完法に業種要素を加味した分析を行っている．レコード間の距離を定義する際に，同一業種のレコードに対しては距離を定率で縮小することで，同一業種が最近接レコードとして採用される確率を増加させる仕組みを導入している．この方法により第 2 章の結果よりも若干良好な補完精度を得ることができたが，改善幅は小さかった．

第4章では，従来財務データの外れ値に適用されていた様々な経験的補正法について整理した後，データに対するneglog変換の有無，変数毎の上下限値での折り返しの有無，上下限値超を欠測として扱いk-NN法で補完するか否かで全5パターンを考慮し，2項ロジットモデルに基づく模擬デフォルト予測をAUCで評価している． k-NN法以上にneglog変換の影響が大きいことが示唆されている．

そこで第5章では，変数分布の歪みをより柔軟に行うためにべき母数を導入した一般化neglog変換を新規に提案している．個々の説明変数で一変量2項ロジットモデルに基づき最適変換を決め，その変換を所与として多変量で2項ロジットモデルを推定し，4章と同じ要領で模擬予測を行っている．一般化neglog変換に基づく方法は若干ではあるがAUCを改善する．第6章はまとめである．

技術的には，大規模データでの欠測値補完に対する k-NN 法の実行可能性を高める方法を提案し，縮小した探索空間のサイズと精度の関係や近傍数の選択に一定の指標を与えたことと，外れ値の影響軽減には分布の歪み補正が本質的であることを示し，一般化 neglog 変換を提案しその有効性を示した点が，貢献として高く評価できる．クレンジング次第で基礎データが変わるという意味で常に問題であった欠損値と外れ値の処理に対して，統計科学的知見に基づいて一定の実務的ガイドラインを与えるものであり，応用面でも高く評価できる．なお，本論文の第 2 章と第 3 章に相当する内容は，査読付和文誌「ジャフィー・ジャーナル」(2015 年, p.143-165)に掲載されている．

総合研究大学院大学複合科学研究科における課程博士の学位授与に係る論文審査等の手続き等に関する規程第10条に基づいて，口述による試験を実施した．この結果，出願者はその博士論文を中心としてそれに関連がある専門分野及びその基礎となる分野について博士（統計科学）の学位の授与に十分な学識を有するものと判断し，合格と判定した．