様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月 １４ 日現在

機関番号：６２６０３
研究種目：基盤研究(B)（一般）
研究期間：2015～2018
課題番号：１５Ｈ０３３９０

研究課題名（和文）企業の信用力評価のための大規模財務データベースの欠損値補完・異常値処理方法の開発

研究課題名（英文）Development of missing value complement / outlier treatment method for large scale financial database for corporate credit risk evaluation

研究代表者
山下　智志（Yamashita, Satoshi）

統計数理研究所・データ科学研究系・教授

研究者番号：５０２４４１０８

交付決定額（研究期間全体）：（直接経費）　　8,000,000円

研究成果の概要（和文）：統計学や生物・医療分野で発達した欠損値異常値処理を財務データへの適用を検討し、財務データ特有の性質をもとにした新たな手法を開発した。CRD協会データや地銀５行の財務・与信データ、政府調査のミクロデータ、不動産賃貸業データに適用する。これらの正常化されたデータベースのうち、秘匿性の低いデータベースについては研究用に公開した。法人統計・事業所統計などの政府データと上記の企業データとの結合に関する方法論を研究し、高い精度のデータ結合を実現した。これによって企業の全数データである政府統計データと、サンプル標本ではあるが高質である信用データを下にした、企業プロファイリングが可能となった。

研究成果の学術的意義や社会的意義
本研究では、欠損値補間、異常値補正、データ結合などのデータ構造化手法を開発した。その結果、信用リスク評価において多種多様なデータベースを用いて予測精度の向上を実現した。この結果は、構造化データベースの相互利用や手法の公開などによって、研究者で去有されている。具体的には以下の成果を得た。
１．欠損値異常値に関する既存研究のサーベイ。２．データクレンジング手法の開発。３．経営・経済系データに対するクレンジング手法の適用。４．統合化信用リスクデータベースの作成。５．統計モデルによる期待損失モデルの構築。６．賃貸不動産収益データベースの作成。７．政府データと信用データの結合。

研究成果の概要（英文）：We examined the application to the financial data of the missing value outlier processing developed in the statistics and bio / medical fields, and developed a new method based on the characteristics of the financial data. We applied it to CRD Association data, credit database of 5 regional banks, government survey microdat and real estate leasing data. Of these structured databases, those with less secrecy were released for research. We researched the methodology for combining government data such as corporate statistics and business location statistics with the above-mentioned enterprise data, and realized highly accurate data combining. This made it possible to conduct corporate profiling based on government statistical data, which is census data of companies, and credit data of banks, which is sample data but high quality.

研究分野：ファイナンス統計学

キーワード：データ構造化　信用リスク　財務データ　欠損値補間　データ結合　異常値補正　アパートローン　公的ミクロデータ

１．研究開始当初の背景

本研究に関連する国内・国外の研究動向及び位置づけ

欠損値や異常値が存在する不完全なデータベースに対する統計学的処理方法（以下、データクレンジング手法）については 1980 年代以降活発に開発され、EMA や ICE などの多くの研究成果がある。しかし、それらの成果は、一定の数学的仮説のもとに成り立つ方法論であり、実際のデータに対して適用可能であるとは限らない。そこで 2000 年以降は、現実のデータの特性を踏まえた、特定分野を前提したデータクレンジング手法の提案がなされている。特に生物・医療分野において顕著な研究成果がある。一方、リーマンショックなどの金融危機の経験から、複数の金融機関から経営財務データを統合・ビッグデータ化を行い、信用リスクモデルを作成することが重要な課題になっている。
我々の研究班のこれまでの研究成果は大きく分けて３つの流れがある。

信用データベースの作成と判別モデル

企業の財務データと倒産データを統合した CRD 協会データを活用したデフォルト確率推計モデルを 2003 年に作成したが、担保や保証などの与信データと毀損データがないため、回収不能額を評価した信用リスク計量化をすることができなかった。担保・保証・毀損の情報が含まれるデータベースが存在しないため、複数の銀行の全数データを得て、総合的な信用リスクデータベースの統合作業を行っている(高度信用リスク統合データベースコンソーシアム:CDSC)。このプロジェクトの問題点は、大企業と違い、中小企業の財務データには多くの欠損値・ 異常値があり、モデルの精度を極端に低下されていることである。

欠損値補間方法の検討

上記の信用リスクデータベース関係のプロジェクトの推進時に、データのクレンジング方法の差異によって、判別モデルの予測値や精度が大きく異なることを認識することができた。そのため、経営財務のデータに対して欠損値補間方法の検討を行っている。モデルのフィットを上げる目的に行うシングル補完(single imputation)と分散を保存するマルチ補完（multiple imputation）の２つの方法を試している。しかし、データ特有の問題点の多さや、数学的な方法論の脆弱性から目標の精度を確保できていない。

医学データなどの欠損値補間方法の理論的検討

分担者の野間は医療・医薬品データに対する欠損値補間方法や効果の判別問題の数学的評価を行い、欠損値補間問題の効率的解決方法について数学的検証を行ってきた。

２．研究の目的

これまでこの分野におけるデータクレンジング手法が詳しく研究されたことがない。とくに財務データは、会計関係諸規則にしたがってデータが作成されているため、変数間に独特の性質があり、医療・生物などの自然科学データとは構造が異なる。財務データベースの精緻化は、これまでの企業財務分析や企業倒産の判別問題に多大な発展をもたらすにもかかわらず、その具体的な方法論が存在しないため対応が遅れており、人為的な手法に頼っていた。そこで本研究では、統計学や生物・医療分野で発達を遂げた欠損値異常値処理方法の財務データベースへの適用可能性を検討するとともに、財務データ特有の性質を元にした新たなデータクレンジング手法を開発する。具体的には、CRD 協会の 1500 万件（150 万社×10 年）のデータや地銀５行から入手した財務･与信データベースに適用する。これらの正常化されたデータベースのうち、秘匿性の低いデータベースについては研究用に公開することによって財務分析学の進展に寄与することを最終成果とする。

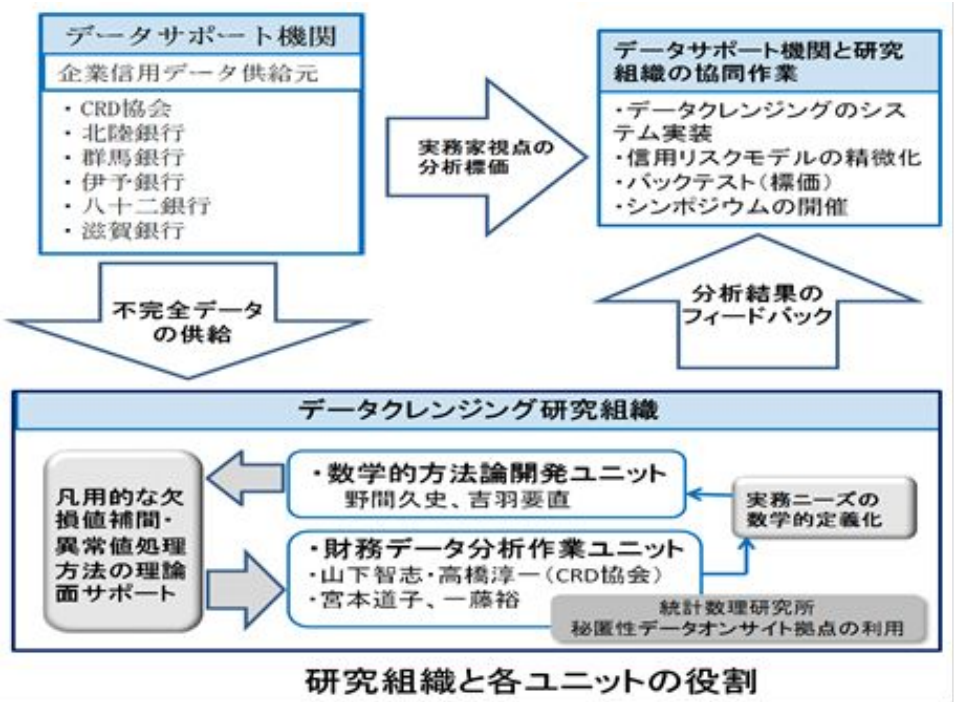
３．研究の方法

【第１作業ユニット】数学的方法論開発ユニット

第 1 作業群は欠損値補間･異常値処理の数学的方法論を検討する。特に技術の発展の著しい 医療・医薬品・生物科学分野の研究成果を詳しくサーベイし、経営財務データのようなバイアスの大きいデータベースに対して有効なクレンジング方法を発見する。さらに、会計・財務の特徴を把握することにより、既存の統計的方法に改良を加えることによって、実務的な有効性を確保したクレンジング方法を開発する。特に k-NN 法などの非パラメトリック補完および、ICE、MICE などの連鎖的回帰分析方法に着目し、財務データへの適用可能性について精査する。

【第２作業ユニット】財務データ分析作業ユニット

第２作業群は経営財務データの統合データベースの構築作業とメンテナンス、およびデフォルト・倒産に関する判別モデルを構築する。CRD 協会と地銀５行から供給されるデータを、分析可能な形に加工する（定形 加工：名寄せなど）。また、第１作業群から提供されるクレンジング手法を適用して計算作業を行う。さらに、クレンジングされたデータを元に判別問題のモデルを作成し精度の向上を確認する。



研究組織と各ユニットの役割

作業は秘匿性データを扱うため、統計数理研究所内の秘匿性データオンサイト拠点（統計法に基づく秘匿性データセキュリティ基準を満たすデータ管理室）を利用する。

４．研究成果

債務回収実績データベース作成事業に関する研究成果

LGD(Loss Given Default)は PD(Probability of Default)とともに信用リスクの構成要素であり、正確な推定を必要とされている。これまで、市場データやリスク・プレミアムから LGD を推計するモデルについては提案がなされているが、回収実績データから作成される統計モデルについてはほとんど存在しない。とくに我が国においては、銀行の回収実績データが未公開なため回収率の決定要因や推計モデルが提案されていない。本研究では、地方銀行における事業法人向け融資の回収実績データをもとに LGD の要因分析と LGD・EL(Expected Loss)の推計モデルの開発を行った。その結果、LGD については担保、保証、貸出額（エクスポージャー）が重要であることがわかった。また、LGD のレベルは FIRB の回収率関数（LGD は 0.35～0.45 となる）に比較して、低いことがわかった。EL 推計には多段階モデルを用いたが、その結果、担保や保証などの貸出要件が PD に影響していることなどが判明し、EL 推計精度の向上に対する知見が得られた。

欠損値補間方法に関する研究成果

「k 最近傍法(以下 k-NN 法)」を財務データに適用する際の計算ステップについて検討した。また、大規模データに対して k-NN 法を適用した際に、計算時間が膨大になる問題を売上高フィールドを利用したインデックスセグメントを用いることにより解決した。連鎖方程式型代入法や前後期線形補完よりも、k-NN 法は真値に近い値を補完することがわかった。

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た。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縮小することによって、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である。この方法では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。

k-NN 法を用いて外れ値の補正を行った。財務データの外れ値については経験的に様々な手法がとられてきた。まず、それらの既存手法についてまとめた後、k-NN 法を外れ値処理に適用する方法について検討した。実際の財務データとデフォルトデータ用いた 2 項ロジットモデルに対して適用し、デフォルト予測精度が向上することを確認した。

さらに変数に対して事前変換を行うことを提案した。具体的には Box-Cox 変換を負の領域にまで拡大した一般化 neglog 変換を用いる。2 項ロジットモデルによるデフォルト予測精度について実証分析を行い、予測精度の向上に成功した。

アパートローンデータ構造化に関する研究成果

本研究ではこれまでリスク計量化モデルが考案されてこなかったアパートローン与信リスクについて、Web データから賃貸住宅の入居化要因を分析しながら、不動産鑑定士による賃貸住宅の定性状況の実地調査を行った。それぞれのデータを統合することにより、より正確な賃貸住宅の収益予測を行うことを目指しているが、主にサーベイデータから得られた情報をもとに、空室が占室になる空占確率モデルと占室が空室になる占空状態推移を示す確率行列モデルを示した。それを利用した時系列シミュレーションと空室状態モデルとの差異について言及した。

今後の課題としては
① 空→占モデル、占→空モデルの精度向上を図る
② 対象地域を増やすことによって、場所転移性を確認する
③ 賃貸物件キャッシュフロー以外の原因でおこるアパートローンのデフォルトについて
モデル化を試みる
などが考えられる。①については今後パネル調査データの増加が見込まれるため、データの蓄積とともに精度の向上が達成されると予想している。

信用リスクモデルデータの構造化手法に関する研究成果
提案手法により、利用可能な変数が少なく、名称、所在地などの詳細な文字情報がない企業データに対しても効率的かつ効果的な統計的マッチングを行うことが可能となる。また、本研究で提案する統計的マッチングのモデルにより､距離のウエイトを最尤法の枠組みで統計的に(最尤法により）推定することが可能となり、これまで過去の経験や専門的な知識に基づいて設定されることが多かった距離のウエイトについてデータに基づき最適な値を推定することができる。さらに、マッチングの正しさに関する確率（マッチング確率）を推定することが可能となり、マッチングの精度の定量的な比較を行うことができる。
なお、名称、所在地などの詳細な文字情報に基づく統計的マッチングでは、同一の対象に対する複数の表現（漢字、平仮名、片仮名、アルファベット等）が存在する標記ゆれの問題があり、これがマッチングを困難なものとしているが、距離に基づく統計的マッチングではそれらの文字情報を用いないため、そのような表記ゆれの問題は生じない。
また、詳細な文字情報によるマッチングは個別のレコードの特定につながるおそれがあるが、提案手法ではマッチング確率を算出するのみであり、直接的な対象の特定を行っているわけではない。提案手法を実際のデータ（平成 24 年経済センサス--活動調査のミクロデータ及び帝国データバンクのデータ）に適用した結果、多項ロジットモデルは適切に推定されており、最も当てはまりの良いウエイト付き絶対値距離の対数変換を用いたモデルに基づく統計的マッチングは､マッチングの正解率の観点から従来の研究で用いられている最近隣法(Nearest Neighbor Method）よりも優れていることが示された。

５．主な発表論文等
〔雑誌論文〕（計１３件）

(1) 右京芳文，野間久史，欠測を伴う経時測定データにおける MMRM（Mixed-Effects Model for Repeated Measures）の並べ替え法に基づく推測手法，計量生物学，査読有，2019，掲載決定
(2) Ukyo，Y.，Noma，H.，Maruo，K.，Gosho，M.，Improved small sample inference methods for a mixed-effects model for repeated measures approach in incomplete longitudinal data analysis，Stats 2，査読有，2019，174-188
(3) 高部勲，山下智志，多項ロジットモデルを用いた新たな統計的マッチング手法の提案，統計学，査読有，115，2018，1-17
(4) 高部勲，山下智志，B-スプライン及び Adaptive Group LASSO に基づく正則化非線形ロジットモデルによるデフォルト確率の推定，統計数理，査読有，66-2，2018，295-317
(5) 高部勲，山下智志，多項ロジットモデルに基づく企業データの統計的マッチング（企業分析），JAFEE proceedings，査読有，2018 年度夏，2018，1-8
(6) Tanoue，Y. and Yamashita，S.，When banks venture beyond home turf:consequences for loan performance，Journal of Credit Risk，査読有，13-3，2017，1-19，DOI:10.21314/JCR.2017.225.
(7) 山下智志，藤山秋佐夫，吉野諒三，越前功，北本朝展，データサイエンスによる大学との連携・協働、そして発展へ②　オープンサイエンスと協働が支える社会・人文学研究の新展開，文部科学教育通信，査読有，422，2017，22-23
(8) Tanoue，Y.，Kawada，A. and Yamashita，S.，Forecasting loss given default of bank loans with multi-stage model，International Journal of Forecasting，査読有，33，2017，513-522，DOI:10.1016/j.ijforecast.2016.11.005
(9) 野間久史，連鎖方程式による多重代入法，応用統計学，査読有，46，2017，1-99
(10) 一藤裕，岡本基，山下智志，曽根原登，ソーシャル･ビッグデータ駆動の観光政策決定支援システム，月刊統計，査読有，9，2015，20-25
(11) 山下智志，一藤裕，鈴木雅人，大島容大，Web ビッグデータとサーベイデータの統合による賃貸住宅価値評価システムの構築，土木計画学研究，査読有，52，2015，219-227
(12) Yamashita，S. and Yoshiba，T.，Analytical solutions for expected loss and standard deviation of loss with an additional loan，Asia-Pacific Financial Markets，査読有，22-2，2015，113-132，DOI:10.1007/s10690-014-9196-5
(13) 野間久史，メタアナリシスのエビデンスを正しく読み解くために～アカデミアの生物統計家の立場から～，薬理と治療，査読有，43，2015，615-620

〔学会発表〕（計３８件）

(1) 野間久史，欠測データの統計解析，昭和大学実践臨床統計学専門セミナー，2019
(2) 山下智志，AI と機械学習の直感的理解と金融への応用，日本銀行金融機構局金融高度化センターWS，2018
(3) 山下智志，ビッグデータ時代におけるデータベース結合の目的・方法・効果，統計関連学会連合大会，2018
(4) 山下智志，ビッグデータ時代における企業データの統計的名寄せ手法，統計数理研究所第 6 回金融シンポジウム，2018
(5) 高部勲，山下智志，多項ロジットモデルに基づく企業データの統計的マッチング（理論的側面），日本分類学会第 37 回大会，2018
(6) 山下智志，医療・健康科学における統計リテラシー：情報・システム研究機構統計数理研究所の取り組み，横幹連合フォーラム，2018
(7) 右京芳文，野間久史，MMRM におけるブートストラップ法を用いた高次漸近理論に基づく近似推測手法，統計関連学会連合大会，2018
(8) 野間久史，Precision Medicine，Comparative Effectiveness Research とデータサイエンス，大阪大学データ科学特別セミナー，2018
(9) 野間久史，先端医学研究の発展を担うデータサイエンス，日本統計学会春季集会，2018
(10) 園田桂子，山下智志，銀行－企業間貸出マッチデータを用いた取引関係の変化の要因分析，統計関連学会連合大会，2017
(11) 岡本基，山下智志，国際ミクロ統計データベースの整備と利用，統計関連学会連合大会，2017
(12) 高部勲，山下智志，多項ロジットモデル及び主成分分析を用いた統計的マッチング手法の提案，統計関連学会連合大会，2017
(13) 宮本道子，安藤雅和，山下智志，欠測値を含む大規模財務データを用いたコピュラによる企業の信用リスク評価について，統計関連学会連合大会，2017
(14) 山下智志，金融機関のリスク管理における人工知能・機械学習(1)，ＣＲＤ信用リスク管理セミナー，2017
(15) 山下智志，金融機関のリスク管理における人工知能・機械学習(2)，ＣＲＤ信用リスク管理セミナー，2017
(16) 高部勲，山下智志，多項ロジットモデル及び主成分分析を用いた新たな統計的マッチング手法の提案，経済統計学会全国研究大会，2017
(17) Takabe，I. and Yamashita，S.，A new statistical matching methodology using multinomial logistic regression and multivariate analysis，International Federation of Classification Societies(IFCS)，2017
(18) 高部勲，山下智志，非線形・正則化ロジットモデルに基づく企業のデフォルト確率予測，JAFEE 夏季大会，2017
(19) 野間久史，Precision Medicine とビッグデータ、統計科学，第 56 回大分統計談話会大会，2017
(20) 野間久史，臨床研究における欠測データの取り扱いと解析の方法：最近の JAMA の事例から，昭和大学実践臨床統計学専門セミナー，2017
(21) 野間久史，臨床研究における欠測データの統計解析：最新の動向と実践的な方法論について，国立がん研究センター生物統計セミナー，2017
(22) 山下智志，企業の成長要因の構造分析と成長率予測の同時推計，中小企業等の事業性評価に向けたモデル構築調査事業，2017
(23) 渡邊隼史，一藤裕，鈴木雅人，山下智志，Web データとサーベイデータの融合：地方圏における住宅投資リスク評価の実験，社会データ構造化センターシンポジウム，2017
(24) 山下智志，データ構造化とは何か？，社会データ構造化センターシンポジウム，2017
(25) 山下智志，賃貸住宅の空室率要因分析とアパートローンのリスク計量化モデルの開発(1)，ＣＲＤ信用リスク管理セミナー，2016
(26) 山下智志，賃貸住宅の空室率要因分析とアパートローンのリスク計量化モデルの開発(2)，ＣＲＤ信用リスク管理セミナー，2016
(27) Noma，H. and Gosho，M.，A generalized Akaike ’s information criterion for multiple imputation，XXVIIIth International Biometric Conference，2016
(28) 岡本基，山下智志，「国際ミクロ統計データベース」の利活用について，統計関連学会連合大会，2016
(29) 山下智志，岡本基，公的統計ミクロデータ研究コンソーシアムによる高等教育研究支援，統計関連学会連合大会，2016
(30) 山下智志，リスク科学と目的・データ・統計的方法論，統計関連学会連合大会，2016
(31) 野間久史，多重代入法によるロバストな推測方法，統計関連学会連合大会，2016
(32) 山下智志，金融機関データに関する人工知能と機械学習のこれまでとこれから，日本アクチュアリー会年次大会， 2016
(33) Yamashita，S.，A new approach of micro-data analysis through international cooperation，The 8th International Workshop on Analysis of Micro Data of Official

Statistics，2016
(34)山下智志，デフォルト・倒産予測モデルから進化した中小企業信用リスク計量化モデル，OLIS-慶應義塾大学保険フォーラム，2016
(35)山下智志，宮本道子，安藤雅和，欠測を考慮したロバストな一般化線形モデルを用いた信用リスクの予測について－中小企業大規模財務データベースにおける考察－，統計関連学会連合大会，2015
(36)山下智志，岡本基，「国際ミクロ統計データベース」の利用方法について，統計関連学会連合大会，2015
(37)山下智志，一藤裕，鈴木雅人，大島容大，Web ビッグデータとサーベイデータの統合による賃貸住宅価値評価システムの構築，土木計画学研究発表会，2015
(38)野間久史，Quantifying indirect evidence in network meta-analysis via composite likelihood methods: Evaluation of inconsistency and contribution rates of direct and indirect evidence，統計関連学会連合大会，2015

〔図書〕（計２件）
(1) 高井啓二，星野崇宏，野間久史，岩波書店，欠測データの統計科学：医学と社会科学への応用，2016，240
(2) Noma, H. and Matsui, S., Boca Raton: Chapman and Hall/CRC, Univariate analysis for gene screening: Beyond the multiple testing. In Design and Analysis of Clinical Trials for Predictive Medicine(Chapter13), 2015, 400(207-226)

６．研究組織

(1)研究分担者

研究分担者氏名：野間久史

ローマ字氏名：（NOMA, Hisashi）

所属研究機関名：統計数理研究所

部局名：データ科学研究系

職名：准教授

研究者番号（8 桁）：70633486

(2)研究協力者

研究協力者氏名：吉羽要直
ローマ字氏名：（YOSHIBA, Toshinao）

研究協力者氏名：高橋淳一
ローマ字氏名：（TAKAHASHI, Junichi）

研究協力者氏名：田上悠太
ローマ字氏名：（TANOUE, Yuta）

研究協力者氏名：宮本道子
ローマ字氏名：（MIYAMOTO, Michiko）

研究協力者氏名：一藤裕
ローマ字氏名：（ICHIFUJI, Yu）