氏　　名	髙部 勲
学位(専攻分野)	博士(統計科学)
学位記番号	総研大甲第 2073 号
学位授与の日付	平成 31年 3 月22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	企業データの統計的マッチング及び変数選択に関する研究
論文審査委員	主　査　教授　川﨑 能典 准教授　南 和宏 教授　山下 智志 特任教授　川﨑 茂　日本大学　経済学部 教授　伊藤 伸介　中央大学　経済学部

（様式３）

博士論文の要旨

氏　　名　高部　勲

論文題目　企業データの統計的マッチング及び変数選択に関する研究

近年，インターネット上の情報，公的統計ミクロデータ，民間企業のデータなど，様々なデータが利用可能となっており，これらを統合できれば，新たにデータ収集を行うことなく，情報量の多い有用なデータが得られると期待される．こうした中，複数のデータを統合するデータリンケージ（Data Linkage）の手法が様々な分野で注目を集めている．ところで，各レコードを識別できる照合キー（共通一連番号，名称，所在地など）が存在する場合には，それらを利用してレコードをリンケージする完全照合（Exact Matching）を行うことが可能である．しかし，例えば異なる機関が整備する企業データに関しては，秘匿性の観点から名称や所在地などの個体を特定できる情報を利用することができず，資本金や売上高などの限られた情報のみが利用可能である場合が多いと想定される．このような場合には，各データに共通の情報を基に，類似したレコードをリンケージする方法が用いられる．これを統計的マッチング（Statistical Matching）という．

このような状況を踏まえつつ，第１章ではデータリンケージ及び統計的マッチングについて，国内外の研究事例などを紹介しつつ，その課題について述べている．我が国ではこれまでに，特に公的統計ミクロデータに関して，多くのデータリンケージに関する研究が行われている．これについては，昨今，公的統計ミクロデータ研究コンソーシアムの設立や公的統計のオーダーメード集計の利用条件等の緩和の実施など，その利活用に向けた機運が急速に高まってきている．また各種の政府決定では今後，企業の保有するビッグデータの公的統計への活用について，検討を進めることとされている．政府統計を取り巻くこうした状況を鑑みれば，公的統計ミクロデータと企業の様々なデータとのリンケージは，今後重要な研究課題になると考えられる．

第２章では，複数のデータにおけるレコード間の類似性を計測する際によく利用されるウエイト付き距離を用いた，多項ロジットモデルに基づく新たな統計的マッチングの手法を提案している．提案手法により，名称・所在地のような詳細な情報がない企業データに対しても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，ウエイト付き距離のウエイトの合理的な決定方法については，これまで研究が行われていなかったが，提案手法により，ウエイトを最尤法の枠組みで合理的に推定することが可能となり，さらにマッチングが正しい確率を算出することができる．提案手法を経済センサスミクロデータ及び帝国データバンク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では，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よりも優れていることが示された．

以上の結果から，提案手法が優れた性能を発揮することが示されたものの，距離や尤度の計算量の問題は依然として残っている．統計的マッチングの対象となるデータのサイズ

が大きくなり，レコード数が増加するような場合には，距離や尤度を計算する対象となるレコードの組合せの数がデータのサイズの２乗に比例して増加するため，計算にはかなりの時間がかかると考えられる．これに対して第３章では，主成分分析によりデータを層化し，近隣の層のレコードのみを距離・尤度計算の対象とすることで計算の効率化を図り，マッチングの精度を大きく低下させない形で計算速度を向上させる方法について検討している．提案手法を経済センサスミクロデータ及び帝国データバンクデータに適用した結果，層の数を適切に設定することで，正解率の低下を最小限としつつ，計算時間を大幅に減少させることが可能なことが示された．

これまでに述べた統計的マッチングの手法を単純に適用した場合，レコードの使用回数に関する制約を設けていないため，一つのレコードに複数のレコードがマッチングされる可能性がある．このような場合，正しいマッチングが実現できず，マッチング精度が低下するおそれがある．そこで第４章では，多項ロジットモデルにより得られたマッチング確率を用いて，１対１の制約付き統計的マッチングの問題を重み付き２部グラフの最適マッチングの問題として定式化した上で，実装しやすく，計算速度が速い効率的なアルゴリズムであるハンガリー法を適用することにより，マッチング精度の向上を図っている．この方法を複数の地域のデータに適用したところ，多項ロジットモデルに基づく統計的マッチングの方法を単純に適用した場合と比較して，全ての地域において統計的マッチングの正解率が向上した．

ところで，統計的マッチングにより分析に利用できる変数が増えるというメリットはあるものの，それらの変数の中から分析に適したものを絞り込むためには，データによっては膨大なコストがかかる可能性がある．また，変数間に複雑な非線形の関係がある場合には，それらの関係を考慮した変数選択は一層困難なものとなる．そこで第５章では，銀行データに基づく企業のデフォルト確率推定モデルの構築の事例を取り上げ，非線形性と変数選択という２つの課題を同時に解決することを目的として，B-スプラインに基づく非線形・ノンパラメトリック回帰モデル及び Adaptive Group LASSO に基づく効率的な変数選択という２つの手法を組み合わせた形でのデフォルト確率予測モデルの構築を試みた．複数の銀行のデータを統合した独自のデータベースを用いてデフォルト確率予測モデルの構築を行った結果，提案手法は，t 値・p 値に基づく変数選択や単純な LASSO と比較して，最も説明変数の数が少なくなり，効率的な変数選択を行うことができた．また AR 値などの指標の観点から，推定精度が向上していることが確認された．

第６章では，本研究の成果について総括するとともに，今後の展望について述べている．今回のデータを用いて構築したモデルを，マッチングの正解が不明な他の企業データに適用することが考えられる．また，各レコードに対してマッチング確率という新たな変数が付与されることとなり，このマッチング確率の有効な利用，例えば回帰分析などにおいて，マッチング確率を説明変数に加えることや，マッチング確率を用いて複数のレコードの変数を加重平均した値を用いることなどが考えられる．マッチング後のデータを用いた様々な分析を行うことにより，情報量の増えたデータベースを用いることの有用性を示すこともまた，重要な課題である．例えば，これまで財務指標がメインであった企業のデフォルト予測モデルの中に，労働量や生産性，付加価値などに関する変数を加えることで，モデルの予測精度が向上する可能性がある．

今後，公的統計のミクロデータや企業の保有するビッグデータの利活用が進められていく中で，様々なデータの特性に応じた統計的マッチング手法の開発は，一層重要なテーマになっていくものと考えられる．将来的に，より多くの多様なデータが利用可能になることを念頭に，本研究で提案した手法も含め，より効果的な統計的マッチングの手法の開発・改善を続けていく必要があると考える．

Form8・Separate Sheet
（様式８・別紙）

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏　　名　高部　勲

Title
論文題目　企業データの統計的マッチング及び変数選択に関する研究

提出された論文は全 6 章 126 ページからなり，日本語で執筆されており，多項ロジットモデルに基づく統計的マッチング（名寄せ・プロファイリング）のための新たな手法を提案している．提案手法により，名称・所在地のような詳細な情報がない企業データに対しても，効果的な統計的マッチングを行うことが可能となったことが示されている．

第 1 章では，データマッチングに関する用語の定義を行った上で，企業データマッチングの重要性を政府統計に関する社会的変化をふまえ解説している．

第 2 章では，統計的マッチングに関するこれまでの研究について概観するとともに，それらの課題について整理している．本研究において設定した課題への対応として，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を提案した．提案手法を経済センサス活動調査のミクロデータと帝国データバンクの企業データに対してマッチング実験を行い，その結果について考察を行った．

第 3 章では，マッチングを行うデータのサイズが拡大した場合を念頭に，より効率的・高速な統計的マッチングを行うための手法について検討している．具体的には，主成分分析の結果に基づいてあらかじめデータを層化しておき，近隣の層のレコードのみを距離・尤度計算の対象とすることにより，計算量を減少させることに成功している．

第 4 章では，第 2 章の方法においてマッチングの正答率を下げている原因として，マッチング元のデータベースの 1 つの企業が，マッチング先のデータベースの複数の企業とマッチングすることがある（この分析では帝国データバンクの 1 レコードが経済センサスの複数のレコードと紐付けられることがある）ことに着目した．そこで，統計的マッチングの精度改善を図るため，統計的マッチングモデルにより得られたマッチング確率を用いて，統計的マッチングの問題を，重み付き 2 部グラフの最適マッチングの問題として定式化し，ハンガリー法を適用することによって正答率の向上を達成した．

第 5 章では，統計的マッチングにより変数が増加した企業データの分析を念頭に置いて，データに含まれる変数の非線形な関係を考慮しつつ，効率的に変数選択を行う方法について検討した．具体的には，銀行データに基づく企業のデフォルト確率推定モデルの構築を例として， B-スプラインに基づくノンパラメトリック回帰及び Adaptive Group LASSO を組み合わせた方法を適用した結果について分析した．

第 6 章では，それまでの分析結果を基に，全体的な考察を行うとともに，将来の展望についても示している．なお，本論文第 2 章の内容は経済統計学会の機関誌『統計学』第 115 号に，第 5 章の内容は『統計数理』第 66 巻 2 号に掲載されている．

本論文の意義は，これまで名前や生年月日などの特定フィールドの完全一致や手動によって行われていたデータマッチングを，多項ロジットモデルやハンガリー法などの統計科

学的な手法を用いることにより，精度の高いマッチングツールを提供できたことにある．今後，研究に利用できるミクロデータベースが増加し，複数のデータベースを利用した分析が一般的になるにつれて，本研究の意義が大きくなると予想される．

以上から，博士論文審査委員会は，本論文が博士（統計科学）の授与に値すると全員一致で判断した．