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

高部　勲

博士（統計科学）

総合研究大学院大学
複合科学研究科
統計科学専攻

平成 30 年度（2018 年度）

要約

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の様々なデータが利用可能になっており，これらのデータを何らかの形で結合することができれば，新たに統計調査やデータの収集等を行うことなく，情報量（変数）の多い有用なデータを構築することが可能となる．このような状況の中，複数のデータをレコード単位で結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．例えば，企業の過去のデータを基にデフォルト確率予測モデルを構築し，信用力の評価を行う場合には，企業のデフォルトに関する大規模なデータが必要となるが，その際に多様な性質を持つ複数のデータを結合することにより，様々な財務指標や企業の属性情報などの分析に利用可能な情報を効率的に増加させることが可能となり，データを収集する際のコストの削減が期待される．

ところで，複数のデータを結合する際に，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結合する完全照合（Exact Matching）を行うことが可能である．しかし，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相互に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．そのような場合には，複数のデータに共通に含まれる変数を基に，何らかの意味で類似したレコード同士を結合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．

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，我が国及び海外の研究事例などを紹介しつつ，その課題について述べる．我が国ではこれまでに，特に公的統計の分野において，公的統計ミクロデータに関する事例を含む多くのデータリンケージ及び統計的マッチングに関する研究が行われている．ところで，公的統計ミクロデータに関しては，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など，その利活用に向けた機運が急速に高まっている．また，平成 29 年に決定された「統計改革推進会議最終取りまとめ」（平成 29 年 5 月 19 日統計改革推進会議決定）や，統計委員会の答申を受けて平成 30 年に閣議決定された第Ⅲ期「公的統計の整備に関する基本的な計画」（平成 30 年 3 月 6 日閣議決定）では，今後，企業の保有するビッグデータなど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．これらの決定等を踏まえ，政府は平成 30 年 3 月 6 日に，公的統計ミクロデータの更なる利活用を含む 「統計法及び独立行政法人統計センター法の一部を改正する法律案」（閣法第 34 号）を国会に提出し，平成 30 年 5 月 25 日に可決・成立した（平成 30 年法律第 34 号）．政府統計を取り巻くこのような状況を鑑みれば，公的統計ミクロデータと企業の保有する様々なデータとのデータリンケージや統計的マッチングは，既存のデータを有効に活用した有用なデータの構築につながるものであり，今後一層，重要な研究課題になると考えられる．

第 2 章では，ウエイト付き距離を用いた多項ロジットモデル（Multinomial Logit Model）に基づく新たな統計的マッチングの手法を提案する．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．

以上の結果から，従来の研究で用いられている統計的マッチングの手法と比較して，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，レコード間の距離や対数尤度の計算に伴う計算量の問題は依然として残っている．例えば，経済センサスのようにサイズの大きなデータを扱う場合には，距離や対数尤度の計算の対象となるレコードの組合せの数も非常に多くなることから，多項ロジットモデルの推定やレコードのマッチングに相当な時間がかかるものと考えられる．このような問題に対して，第 3 章では，主成分分析の結果（第 1 主成分得点）に基づいてデータを層化し，同一又は近隣の層のレコードのみを距離や対数尤度の計算の対象とすることによって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する．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により，正解率の低下を最小限に抑えつつ，計算時間を大幅に削減できることが示された．

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，レコードが複数回使用されることに関する制約を設けていないため, 1 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り，そのような場合，正しいマッチングが実現できず，マッチングの精度の低下につながるおそれがある．そこで第 4 章では，多項ロジットモデルにより推定さ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（Weighted Bipartite Graph）の最適マッチングの問題として定式化した上で，この問題に対する効率的なアルゴリズムであるハンガリー法（Hungarian Method）を適用することにより，1 対 1 の制約付きマッチング（Constrained Matching）を実現しつつ，更なるマッチング精度の向上を図っている．ハンガリー法のアルゴリズムは実装しやすく，その計算速度は速い．提案手法を複数の地域のデータに対して適用した結果，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することが確認できた．

ところで，統計的マッチングによりデータに含まれる変数が増加した場合，その後の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，それらの全てが必ずしも分析に役立つというわけではなく，多くの変数の中から分析に適したものを選択する必要があり，その作業に膨大なコストがかかる可能性がある．また，企業データでは変数間に非線形な関係が存在する場合もあり，そのような関係を考慮した変数選択は一層困難なものになると考えられる．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予測モデル構築の事例を取り上げ，変数間の非線形な関係も考慮した，効率的かつ効果的な変数選択の方法について検討する．具体的には，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（B-spline）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（Least Absolute Shrinkage and Selection Operator）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ることにより，効率的かつ効果的なデフォルト確率予測モデルの構築を行う．複数の銀行のデータを統合した独自のデータ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，いずれの期間のデータにおいても最も説明変数の数が少なくなっており，また，B--スプラインにより，デフォルト確率と財務指標の非線形な構造をある程度捉えられることが確認できた．さらに，AR 値（Accuracy Ratio）や McFadden の疑似決定係数などの複数の指標に基づく比較の結果，提案手法は推定精度の面で，他の手法よりも優れていることが確認された．

以上の結果を踏まえ，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る．今回のデータを用いて構築したモデルを，全く別の企業データ，特にマッチングの正解が不明なデータに適用することが考えられる．このようにして構築されたデータは，成長産業の要因に関する分析や企業・事業所の開廃業に関する分析，信用リスクモデルの精度向上など，様々な分野における分析等に役立つと考えられる．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，これらを用いた複数のレコードの加重平均や，回帰分析における説明変数としての利用など，様々な活用方法が考えられる．統計的マッチングを行った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（変数）の増えたデータ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，企業の労働生産性や付加価値などに関する変数を加えることにより，モデルの予測精度が向上する可能性がある．このような課題については，本研究で提案したマッチング手法により結合したデータを用いて更なる分析を重ねていく必要があると考える．

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効果的な統計的マッチング手法の開発は，今後一層，重要な研究テーマになっていくものと考えられる．今後は，本研究で提案した手法も含め，分野横断的に様々な統計的マッチング及びデータリンケージの手法を組み合わせて，より効果的な手法の提案や，複数のデータに基づく実証を重ねていく必要がある．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，更なる効果的な統計的マッチングの手法の開発及び改善を続けていく必要があると考える．