【研究論文】（『統計学』第115号 2018年9月）

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

高部　勲*・山下智志**

要旨

統計的マッチングは，異なるデータを組み合わせて有用なデータを構築するための手法である。統計的マッチングにより，追加の調査やデータの収集を行うことなく，有益なデータを作成することが可能となり，近年，様々な分野で利用が進んでいる。本研究では，多項ロジットモデルを用いた新たな統計的マッチング方法を提案する。提案手法では，ウエイト付き距離を基にマッチング確率を推定するモデルを構築する。このとき，分析対象となるデータの規模が拡大するとともに，距離を計算するための時間が急激に増加することとなるが，この問題を解決するために主成分分析を基にデータを層化し，レコードの検索対象を縮小することで，より効率的に検索を行っている。提案手法を商用データと経済センサスのミクロデータに適用した結果，マッチングの正解率の観点から，最近隣法よりも優れていることが示された。

キーワード

統計的マッチング，多項ロジットモデル，ウエイト付き距離関数，主成分分析

はじめに

データリンケージは，異なるデータをレコード単位で結合して豊富な情報を持つ単一のデータを構築するための手法であり，レコードリンケージと呼ばれることもある（Herzog et al.（2007），Christen（2012），Harron et al.（2015））。データリンケージにより，新たな統計調査やデータ収集などを行うことなく有用なデータの作成が可能となることから，近年，様々な分野で利用されるようになってきている。

我が国ではこれまでに，特に公的統計の分野において，以下の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に関する研究が行われてきている。

- 家計調査と貯蓄動向調査（荒木・美添（2007））
- 生産動態統計調査と工業統計調査（小西（2012））
- 異時点間の中小企業景況調査の結果（坂田・栗原（2011））
- 異時点間の法人企業統計調査の結果（栗原（2015），坂田・栗原（2013））
- 賃金構造基本統計調査と経済産業省企業活動基本調査（村田・伊藤（2016））

* 正会員，総務省統計局
東京都新宿区若松町19-1
総合研究大学院大学
東京都立川市緑町10-3
e-mail：i.takabe@soumu.go.jp
** 非会員，統計数理研究所
東京都立川市緑町10-3
e-mail：yamasita@ism.ac.jp

ところで上記のような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（植松（2016a）及び植松（2016b））。また昨年決定された「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，先般の統計委員会において答申が行われた「第Ⅲ期公的統計基本計画」（平成29年12月19日統計委員会）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保有する様々なデータとのリンケージは，既存のデータを有効に活用した有用なデータベースの構築につながるものであり，今後重要な研究課題になると考えられる。

ところで，データリンケージを行う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1対1で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，何らかの意味で類似したレコード同士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（美添（2005））。これらの関係を整理したものが，以下の図1である。

統計的マッチングのイメージを示したものが図2である。図2において，$X_{i1}, X_{i2}, \cdots, X_{iK}$（又は$X_{j1}, X_{j2}, \cdots, X_{jK}$）は，データ$A$とデータ$B$に共通に含まれている変数（共通変数）である。また$Y_{i1}, Y_{i2}, \cdots, Y_{iq}$はデー$A$のみに含まれる変数であり，$Z_{j1}, Z_{j2}, \cdots, Z_{jr}$はデータ$B$のみに含まれる変数である。図2では，共通変数を用いてデータAのi番目のレコードとデータBのj番目のレコードをマッチングした結果が，融合データのl番目のレコードになる様子を示している。

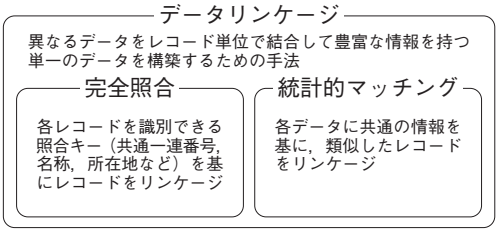

図1　データリンケージと統計的マッチング等との関係

ここで，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，レコードの組合せは膨大になり，これら全てについての類似度を計算することは計算時間の面からみて現実的ではないと考えられる。

これまでに述べてきた背景・論点を踏まえると，今後，公的統計ミクロデータと企業データとのリンケージや統計的マッチングを効果的に進めていくためには，以下に示した3点の課題に対応していく必要がある。

(1) 公的統計ミクロデータと民間の保有するデータの更なる活用を見据えた統計的マッチングの実証分析
(2) 企業データの秘匿性に配慮した上で，限られた情報を基に統計的マッチングを行うモデル・技術の開発
(3) データベースの容量拡大にも対応可能な統計的マッチングの計算効率化・高速化の実現

本稿では計量経済学等の分野で広く利用されている多項ロジットモデル（McCullagh and Nelder（1989），Hosmer et al.（2013））及び主成

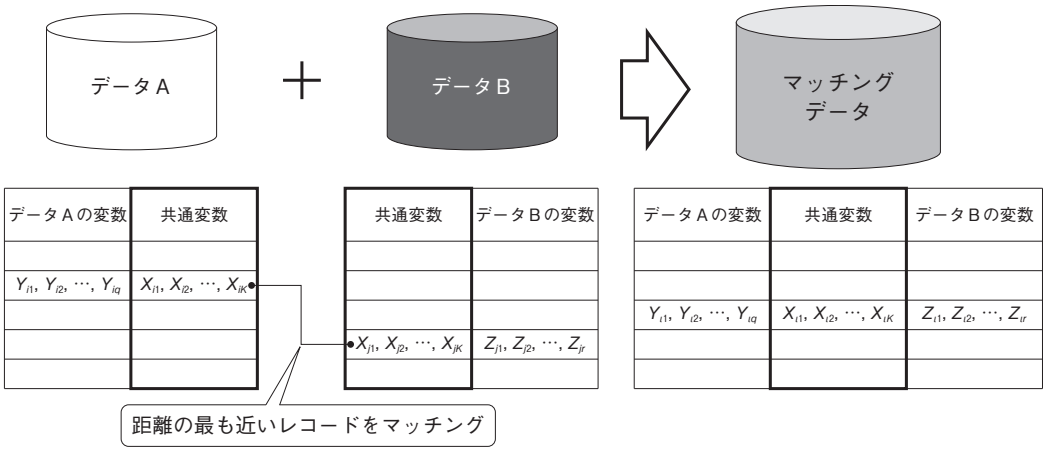

図2　統計的マッチングのイメージ

分分析を基に，上記の課題に対応した新たな統計的マッチング手法を提案する。提案手法を経済センサス－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，正解率などの点で，従前のマッチング手法よりも優れていることが示された。

1．統計的マッチングに関する先行研究と課題

統計的マッチングに関する研究は1960年代から行われてきており，これまでに様々な手法が開発されている。初期の段階から研究が行われているのは，異なる2つのレコードが偶然マッチングする確率（U-statitics）と，同一のレコードが正しくマッチングする確率（M-statistics）を用いて，しきい値によりマッチングの適否を判定する方法である（Newcombe（1959），Fellegi and Sunter（1969））。この方法では，名称・所在地のような詳細な共通情報が利用できることを仮定しているが，今回の分析では企業の名称，所在地などの詳細な情報が利用できず，限られた共通情報のみを活用できる状況を想定していることから，この方法は適していない。

また，データの構造に多変量正規分布を仮定した上で，共通情報以外の変数を欠測値とみなし，これらを重回帰モデルやベイズ統計学の枠組みに基づき推測する方法も存在する（D'Orazio et al.（2006），Rässler（2002），栗原（2015），Kurihara（2015））。本稿で対象とする企業のデータについては，売上高や従業者数などかなり歪んだ分布を持つ変数が存在しており，また連続変数とカテゴリ変数（業種や開設年など）が混在するなど，多変量正規分布の仮定が当てはまらないケースが想定されるため，この方法も適していない。

距離に基づく統計的マッチングも比較的初期の段階から研究が行われている方法である。この方法では，各データに共通の情報を用いてレコード間の距離を計算し，最も近いレコード同士のマッチングを行う。このとき連続変数に対する距離関数としては，例えば以下のものが用いられる（D'Orazio et al.（2006））。

・絶対値距離（Manhattan距離）：

$$D_{ij} = \sum_{k=1}^{K} \beta_k \left| X_{ik} - X_{jk} \right| \tag{1}$$

・Euclid距離：

$$D_{ij} = \sqrt{\sum_{k=1}^{K} \beta_k (X_{ik} - X_{jk})^2} \tag{2}$$

・Mahalanobis距離：

$$D_{ij} = (X_i - X_j)^T \sum_{XX}^{-1} (X_i - X_j) \tag{3}$$

D_{ij}：レコードiとレコードjの距離
X_{ik}, X_{jk}：レコードi及びレコードjに含まれるk番目の共通変数
$X_i = (X_{i1}, X_{i2}, \cdots, X_{iK})^T$
$X_j = (X_{j1}, X_{j2}, \cdots, X_{jK})^T$
Σ_{XX}：共通変数の分散共分散行列

ここでβ_kはk番目の共通変数にかかるウエイトを表す。またカテゴリ変数に対しては，以下の距離が用いられる。

$$D_{ij} = \sum_{k=1}^{K} \beta_k I(X_{ik} - X_{jk}) \quad (4)$$

$I(X_{ik}=X_{jk})$は，$X_{ik}=X_{jk}$のときに1，$X_{ik}\neq X_{jk}$の場合に0となる関数である。

Lie（2001）では企業の財務データに絶対値距離を適用して統計的マッチングを行っている。Yoshizoe and Araki（1999）では，絶対値距離及びEuclid距離の2乗を基に，家計調査及び貯蓄動向調査の統計的マッチングを行っている。坂田・栗原（2011・2013）及び栗原（2015）ではマハラノビス距離を用いて企業データを接続し，パネルデータの作成やマッチングのバイアスの分析等を行っている。

連続変数とカテゴリ変数の両方が含まれる場合には，上記の式(1)及び式(4)を組み合わせた以下のGower距離[1)]が用いられる（Gower（1971））。

$$D_{ij} = (\sum_{k=1}^{K} \delta_{ijk} D_{ijk}) / \sum_{k=1}^{K} \delta_{ijk} \quad (5)$$

ここでD_{ijk}は以下のように定義される変数である。

・X_{ik}, X_{jk}が連続変数の場合
$D_{ijk} = \left| X_{ik} - X_{jk} \right| / R_k$
R_k：k番目の共通変数のレンジ
・X_{ik}, X_{jk}がカテゴリ変数の場合

$$D_{ijk} = \begin{cases} 0 \, (X_{ik} = X_{jk}) \\ 1 \, (X_{ik} \neq X_{jk}) \end{cases}$$

また，δ_{ijk}は変数が欠測している場合に0，それ以外の場合に1となる変数である。今回の分析では事前に欠測値を補完したデータを扱うためδ_{ijk}は常に1となり，$D_{ij} = (\sum_{k=1}^{K} D_{ijk}) / K$となる。

式(1)(2)(4)におけるウエイトβ_kは，変数間の重要度の違いの反映やスケールの調整のために用いられる。このようなウエイトとしては各変数の標準偏差の逆数やレンジ（最大値－最小値）の逆数が用いられる（D'Orazio et al.(2006)）。しかし各変数の重要度やスケール調整の方法をどのように決定するかについては一般的な基準が無く，上記のようなウエイトを使用する理論的な根拠は特段ないため，より最適なウエイトを検討する余地が残されている。本稿で提案する統計的マッチング手法ではウエイトをデータから推定することが可能であり，連続変数とカテゴリ変数が混在する場合でも問題なく扱うことが可能である。

2．多項ロジットモデルを用いた統計的マッチング手法

ここでは，本稿で提案する多項ロジットモデルを用いた統計的マッチング手法におけるモデルの詳細について説明する。以下の2種類のデータを考える。

・データA（マッチング元）：レコード数M
・データB（マッチング先）：レコード数N

このとき，マッチング元のデータAから取り出したある企業iに対して，データBのある企業jが正しいマッチング先である確率P_{ij}を考える（以下ではこれを「マッチング確率」という。）。ここでP_{ij}は，レコード間の距離D_{ij}を用いて次のように表現できるものとする。

$$P_{ij} = \frac{\exp(-D_{ij})}{\sum_{j=1}^{N} \exp(-D_{ij})} \quad (6)$$

距離が小さいほど正しいマッチング先である

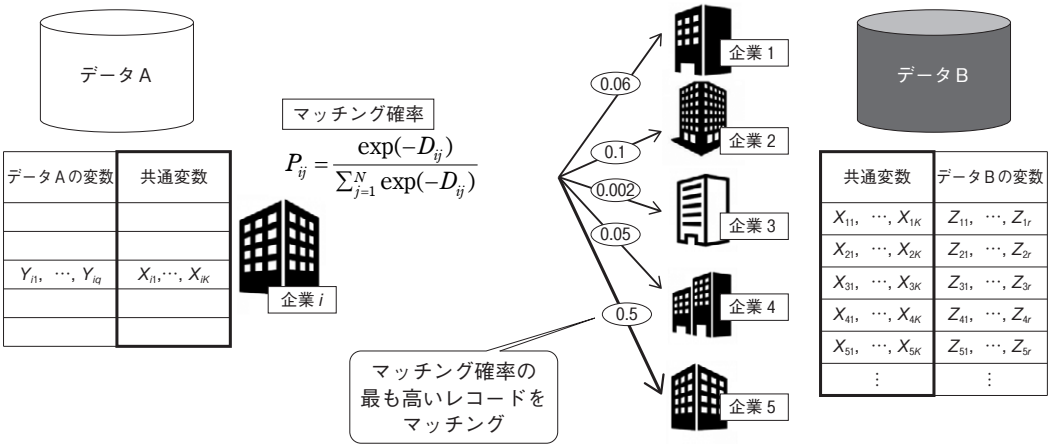

図3　多項ロジットモデルに基づく統計的マッチングモデルのイメージ

可能性が高くなるという状況を想定していることから，式(6)では距離D_{ij}の－1倍を用いている。多項ロジットモデルに基づく統計的マッチングモデルのイメージを示したものが，図3である。図3では，データAの企業iとデータBの全ての企業に対して共通変数を基にP_{ij}を計算し，この値が最も高い5番目の企業とマッチングを行う様子を示している。なお，データAの各企業に対してマッチング先を決定する際に，データBのある企業が複数回選ばれる可能性がある。

モデルの中の距離D_{ij}としては，どのようなものを用いてもよく，複数の距離を組み合わせてもよい。本稿の分析では連続変数とカテゴリ変数を扱うので，式(1)及び式(4)を組み合わせたものを距離として用いる（具体的な距離の形状は3.3節で示す。）。

このとき，尤度関数$L(\beta_1, \beta_2, \cdots, \beta_K)$及び対数尤度関数$l(\beta_1, \beta_2, \cdots, \beta_K)$については，以下のように表現できる。

$$\begin{aligned} L(\beta_1,\beta_2,\cdots,\beta_K) &= \prod_{i,j} P_{ij}^{\delta ij} \\ l(\beta_1,\beta_2,\cdots,\beta_K) &= \log[L(\beta_1,\beta_2,\cdots,\beta_K)] \\ &= \log\left[\prod_{i,j} P_{ij}^{\delta ij}\right] \\ &= \sum_{i,j} \delta_{ij} \log[P_{ij}] \end{aligned} \tag{7}$$

ここでδ_{ij}は，企業iと企業jが等しい場合に1となり，異なる場合に0となる変数である。距離のウエイト$\beta_1, \beta_2, \cdots, \beta_K$は対数尤度関数の式(7)に組み込まれているので，式(7)を最大化することにより，以下のような形でウエイトの最尤推定値$\beta_1, \beta_2, \cdots, \beta_K$が得られる[2)]。

$$\begin{aligned} \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_K &= \underset{\beta_1,\beta_2,\cdots,\beta_K}{\arg\max}\, l(\beta_1,\beta_2,\cdots,\beta_K) \\ &= \underset{\beta_1,\beta_2,\cdots,\beta_K}{\arg\max} \sum_{i,j} \delta_{ij} \log[P_{ij}] \end{aligned} \tag{8}$$

このような形で距離のウエイトの最適化を行っている研究事例については現在のところ確認されていない。

モデルのパラメータ（ウエイト）が推定されれば，ウエイト付き距離及びマッチング確率を全てのレコードの組合せに対して計算することができる。このようにして得られたマッチング確率を用いて，データAの各企業に対して，最も高いマッチング確率を有する企業をデータBから検索する。

モデルの構築方法やパラメータの推定方法からわかるとおり，提案手法には，以下のような利点がある。

【提案手法の利点】
・距離関数のウエイトを統計的に推定することが可能。

・連続変数とカテゴリ変数が混在する場合でも，問題なく扱うことが可能。
・レコードが一致する程度をマッチング確率の形で推定できるので，マッチングの精度の確率的な評価が可能。
・p値，t値などにより，ウエイトの推定精度を分析することが可能[3)]。
・変数の背後に特定の分布を仮定する必要がない。

提案した手法が実際に機能するか，また従前の手法と比較してどの程度のパフォーマンスを発揮するかについて，この後の節で，企業に関する実際のデータを用いた検証を行う。

このように，マッチングの正解のわかっているデータを用いてモデルの構築を行うことができれば，マッチングの正解が不明な他のデータに当該モデルを適用してマッチングを行うことも可能となる。このような方向性については，今後の課題で触れる。

3．実際のデータに基づく分析

3.1　分析に用いるデータの作成

本稿では，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データに対して，提案手法に基づく統計的マッチングを行う[4)]。なお，分析に要する時間と費用の関係から，滋賀県のデータについて分析を行っている。どちらのデータも分析対象は株式会社又は有限会社の企業データとしている。また，帝国データバンクのデータでは，資本金300万円以上5,000万円未満の会社を対象としている。

これらのデータについて，名称及び所在地を照合キーとして，事前に完全照合を行っておく。こうして完全照合されたデータについて，名称及び所在地の情報を削除し，分析用データとする。各データに対する処理等の詳細については，以下を参照。

(1)「帝国データバンク」データ
・平成24年2月分の滋賀県のデータ7,720レコードのうち，名称・所在地を用いて完全照合できた6,278レコードを分析に用いる。
・完全照合できなかったレコードについては，分析対象から除外する。

(2)「平成24年経済センサス－活動調査」ミクロデータ
・滋賀県の企業13,657レコードを分析に用いる。
・データには一部，欠測値が含まれていることから，ICE：Imputation by Chained Equations（Buuren（2012））により補完を行う[5)]。

以上の処理により，帝国データバンクのレコードに対応するレコードが，経済センサス－活動調査のデータの中に必ず存在するという状況となっている。

次に，上記の完全照合後のデータセットについて，帝国データバンクデータと経済センサス－活動調査のミクロデータの各レコードから2/3を無作為抽出してモデル構築用（学習用）のデータとした。また両データにおける残りの1/3のレコードをモデルの性能の検証用（テスト用）のデータとした。学習用とテスト用のそれぞれのデータのレコード数は，以下のとおりである。

・学習用データ
　帝国データバンク：4,240レコード
　経済センサス－活動調査：9,105レコード
・テスト用データ
　帝国データバンク：2,038レコード
　経済センサス－活動調査：4,552レコード

3.2　使用する変数

今回の分析で用いた帝国データバンクの

データと経済センサス－活動調査ミクロデータの両方に共通して存在する変数は，以下の6種類である（2つのデータで単位等をそろえている。）。

⑴　従業者数（従業員数）：(X)【人】
⑵　資本金額（Y)【万円】
⑶　売上高（Z)【百万円】
⑷　産業（S)
⑸　開設年（O)
⑹　地域（市・郡）(R)

⑴⑵⑶が連続変数であり，⑷⑸⑹がカテゴリ変数である。

データを事前に比較・分析した結果，帝国データバンクのデータの「従業員数」($X1$) には，パート・アルバイトを含む場合とそうでない場合が混在していると考えられるケースが存在することが確認できた。経済センサス－活動調査の「従業者数」については，パート等を含む ($X2$) と含まない場合 ($X3$) のどちらの場合の情報も得られる。そこで，2つの場合 (($X1, X2$) 又は ($X1, X3$)) に関して距離を計算し，小さい方を従業者数・従業員数に関する距離とした。

産業については，帝国データバンクのデータで用いられている産業分類（TDB産業分類）を，経済センサスで採用されている日本標準産業分類の大分類に合うように組み替えて使用した。開設年については，⑴1984年以前，⑵1985年～1994年，⑶1995年～2004年，⑷2005年以降という形でカテゴリ化して使用した[6]。地域情報については，詳細な情報は利用できないが，県内の市・郡レベルの情報のみ利用できるという状況を想定している。

学習用データに関して各変数の記述統計量を示したものが表1である。テストデータに関する同様の表は，論文の最後に付表として示している。また連続変数の距離に関する記述統計量及びカテゴリ内での一致率を学習用データ及びテスト用データについて示したものが表2である。

3.3　分析に用いるモデル

分析に用いる多項ロジットモデル[7]を改めて次のとおり示す。

$$
\begin{aligned}
P_{ij} &= \frac{\exp(-D_{ij})}{\sum_{j=1}^{N}\exp(-D_{ij})} \\
D_{ij} &= \beta_1\left|X_i - X_j\right| + \beta_2\left|Y_i - Y_j\right| + \beta_3\left|Z_i - Z_j\right| \\
&\quad + \beta_4 I(S_i = S_j) + \beta_5 I(O_i = O_j) + \beta_6 I(R_i = R_j)
\end{aligned}
\tag{9}
$$

X_i, X_j：従業者数・従業員数（人）
Y_i, Y_j：資本金額（万円）
Z_i, Z_j：売上高（百万円）
S_i, S_j：産業（大分類）
O_i, O_j：開設年
R_i, R_i：地域（市・郡）

ここで距離が0に近い部分の差を強調するために，以下のように絶対値距離を対数変換した量を用いたモデルについても別途，推定を行う（0の値を含む変数があるため，従業者数・従業員数であれば，以下のように1を足した上で対数変換を行う。）。

$$\log(\left|X_i - X_j\right| + 1) \tag{10}$$

このように変換された量については，距離の定義（三角不等式など）を満たしていないものの，レコード間の類似度を測る指標として活用することはできる。Gower距離及びMahalanobis距離[8]に基づく最近隣法（Nearest Neighbor Method）[9]を提案手法に対する比較対象として，結果の比較・分析を行う。

3.4　モデルの推定結果及び正解率の評価

ウエイト付き距離及びその対数変換を用いたモデルについて，パラメータの推定結果を示したものが表3である[10]。推定された係数とともに，それらの標準誤差も併せて示している。またモデルのデータへの当てはまり具

表1　学習用データにおける記述統計量（要約統計量及びカテゴリ別企業数）

	帝国データバンク			平成24年経済センサス活動調査			
	従業員数（人）	資本金（万円）	売上高（百万円）	従業者数（パート等含む）（人）	従業者数（パート等除く）（人）	資本金（万円）	売上高（百万円）
第1四分位	2.0	500.0	41.0	3.0	0.0	300.0	23.0
中央値	4.0	1000.0	100.0	6.0	2.0	1000.0	63.1
平均値	10.5	1163.0	311.3	18.6	9.7	2782.0	376.9
第3四分位	10.0	1112.0	234.2	14.0	6.0	1000.0	177.7
標準偏差	26.6	961.4	1063.0	71.6	49.1	52659.1	3641.9

産業大分類	帝国データバンク	平成24年経済センサス－活動調査
A 農業，林業	0	81
B 漁業	0	4
C 鉱業，採石業，砂利採取業	0	12
D 建設業	1411	1739
E 製造業	841	1771
F 電気・ガス・熱供給・水道業	0	－
G 情報通信業	39	135
H 運輸業，郵便業	0	281
I 卸売業，小売業	1181	2297
J 金融業，保険業	0	105
K 不動産業，物品賃貸業	249	803
L 学術研究，専門・技術サービス業	145	399
M 宿泊業，飲食サービス業	72	463
N 生活関連サービス業，娯楽業	73	332
O 教育，学習支援業	12	72
P 医療，福祉	27	129
Q 複合サービス事業	0	－
R サービス業（他に分類されないもの）	190	477

地域（市・郡）	帝国データバンク	平成24年経済センサス－活動調査
大津市	807	1930
彦根市	341	705
長浜市	459	960
近江八幡市	258	492
草津市	370	847
守山市	197	437
栗東市	212	498
甲賀市	331	663
野洲市	149	331
湖南市	160	345
高島市	231	422
東近江市	360	706
米原市	107	234
蒲生郡	96	202
愛知郡	90	180
犬上郡	72	153

※少数のデータに関しては秘匿を行っている。

期間	帝国データバンク	平成24年経済センサス－活動調査
1984年以前	1825	3624
1984年～1994年	1303	1939
1995年～2004年	850	2125
2005年以降	262	1417

表２　学習用データ及びテスト用データにおける距離に関する記述統計量
（要約統計量及びカテゴリ別一致率）

	学習用データ			テスト用データ		
	従業者数・従業員数 $\lvert X_i - X_j \rvert$	資本金額 $\lvert Y_i - Y_j \rvert$	売上高 $\lvert Z_i - Z_j \rvert$	従業者数・従業員数 $\lvert X_i - X_j \rvert$	資本金額 $\lvert Y_i - Y_j \rvert$	売上高 $\lvert Z_i - Z_j \rvert$
第1四分位	1.0	200.0	39.5	1.0	200.0	39.2
中央値	3.0	700.0	111.0	3.0	700.0	109.8
平均値	13.8	2701.0	544.5	14.4	2230.0	556.0
第3四分位	10.0	1500.0	316.5	10.0	1500.0	304.0
標準偏差	53.6	52620.6	3755.0	83.5	31771.8	5058.2

	学習用データ			テスト用データ		
	産業（大分類）	開設年	地域（市・郡）	産業（大分類）	開設年	地域（市・郡）
一致率	0.89	0.68	1.00	0.88	0.69	1.00

表３　多項ロジットモデルの推定結果

	ウエイト付き絶対値距離	ウエイト付き絶対値距離（対数変換）
従業員数	0.20447 ***	1.05785 ***
	（0.00702）	（0.03037）
資本金額	0.00458 ***	0.82099 ***
	（0.00009）	（0.01374）
売上高	0.00965 ***	0.95353 ***
	（0.00026）	（0.01404）
産業	3.62859 ***	3.50038 ***
	（0.05449）	（0.05540）
開設年	1.55769 ***	1.50823 ***
	（0.03797）	（0.04106）
地域（市・郡）	13.98638 ***	9.27463 ***
	（1.43508）	（0.44812）
初期対数尤度	− 38654	− 38654
対数尤度	− 13918	− 10302
疑似決定係数 ρ^2	0.6399	0.7335
自由度調整済疑似決定係数	0.6398	0.7333
サンプルサイズ	4552	4552

※*** は，0.1％で有意であることを示している。
※（　）は各回帰係数の標準誤差である。

合をみるために，McFaddenの疑似決定係数（Pseudo-R-square）ρ^2も併せて示している。

どのモデルにおいても各ウエイト（回帰係数）の標準誤差は十分に小さく，有意な結果となっている。特に産業と地域のウエイトが大きな値となっており，マッチング確率に強い影響を与えていることがわかる[11)]。さらに疑似決定係数を比較すると，対数変換した距

離を用いたウエイト付き絶対値距離に基づくモデルの方が，データへの当てはまりが良いという結果になっている。

次に正解率の観点から，提案手法をGower距離及びMahalanobis距離に基づく最近隣法と比較する。ここでマッチング手法間の正解率の比較を定量的に行うために，Yoshikawa et al.(2015) で示されている評価方法を用いる。これはマッチング元の各レコードについて，マッチング確率の高い上位 R 件の候補レコードに正しいマッチング先が含まれる割合を用いるものである。ここでマッチング確率は表2に示した回帰係数の推定値に基づき，テストデータを用いて算出する。具体的には以下の方法により計算する。

帝国データバンクのテストデータの各レコード i ($i=1, 2, \cdots, M_{test}$) に対して，経済センサス－活動調査のテストデータにおける対応する正しいレコードのインデックスを $t(i)$ とする。次に帝国データバンクのテストデータの各レコード i に対して，経済センサス－活動調査のテストデータのレコードの中でマッチング確率の高かった順に上位 R 件のレコードを取り出し，その集合を $C(i, R)$ とする。このとき，正しいレコードが上位 R 件の候補レコードに含まれる割合（正解率）$P(R)$ は，以下の式(11)の形で表現できる。

$$P(R)=\frac{1}{M_{test}}\sum_{i=1}^{M_{test}} I(t(i)\in C(i,R)) \qquad (11)$$

ここでI(・) は，(・)内の命題が真の場合に1，それ以外の場合に0となる関数である。

マッチング確率の上位の件数 R を横軸にとり，対応する正解率 $P(R)$ をマッチング手法ごとにプロットしたものが図4である。

多項ロジットモデルによる方法は，Gower距離やMahalanobis距離に基づく最近隣法と比較して，大幅に正解率が高くなっていることがわかる。また絶対値距離を対数変換したモデルが最も正解率が高くなっている。なお，

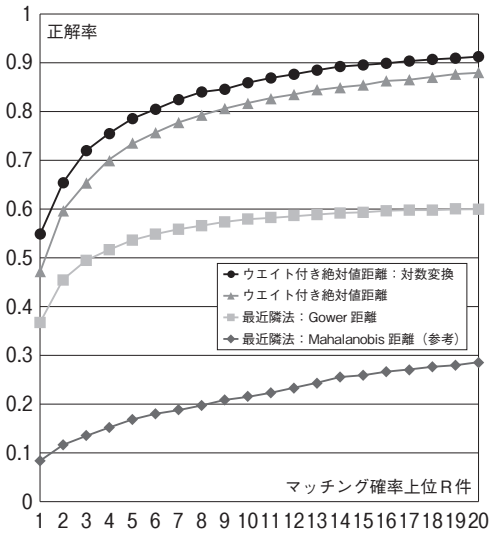

図4　マッチング正解率の比較

Mahalanobis距離の算出に当たっては，R のパッケージの制約から連続量の変数のみを用いている。また，Gower距離の算出に当たっては，そのウエイトをテストデータのみから算出している。手法の違いのほか，上記のように距離によって使用している情報量が異なる点が正解率の違いに影響を与えている可能性がある。

4．主成分分析に基づく計算の効率化

距離に基づく統計的マッチングでは，マッチング先とマッチング元の各データのレコード数が増加した場合，距離計算の対象となるレコードの組合せが急激に増大し，現実的な時間での計算が困難となる問題が生じる。そこで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により計算の効率化を図る方法を検討する。

ところで，3.1節で述べたとおり，帝国データバンクの従業員数（$X1$）に対応する経済センサス－活動調査の従業者数は2つの場合があり（$X2$ 及び $X3$），このままでは2つのデータで変数の数が合わず，主成分分析を行うこ

とができない。そこで変数の数をそろえるために，従業者に関する変数として，従業者1（$W1$）及び従業者2（$W2$）を導入した。

・帝国データバンク：
$W1=X1, W2=X1$
・経済センサス－活動調査：
$W1=X2, W2=X3$

なお，主成分分析の計算[12]の際には，学習用データに関して，帝国データバンクのデータ4,552レコードに対して，経済センサス－活動調査のミクロデータ9,105レコードを縦に追加する形で結合した13,657レコードのデータを用いている。また，連続変数については（1を足した上で）対数変換を適用している。

マッチング元及びマッチング先のデータを合わせて主成分分析を行った結果について示したものが，表4である。主成分分析の結果をみると，第1主成分は，企業のサイズを表す成分となっていることがわかる。また第2主成分は産業及び開設年を強調した成分となっており，第3主成分は地域（市・郡）を強調した成分となっている。

今回の分析では，第1主成分の情報を用いて層化を行う。具体的には，マッチング元及びマッチング先のレコードを合わせたデータについて，各レコードの第1主成分得点の大きさを基に，層内のレコード数が等しくなるように，n 個の層に分割する（例えば層の数が4つの場合には25％点・50％点・75％点で分割する。）。そしてマッチング元のレコードが属する層が第 h 層の場合，マッチング先のレコードに関しては，第 h 層に隣接する層を加えた，第 $h-1$ 層・第 h 層・第 $h+1$ 層に属するレコードのみを対象として，距離及び尤度の計算を行う。ここでマッチング元のレコードが属する層が第1層（第 n 層）の場合には，第2層（第 $n-1$ 層）のみを隣接する層として扱うことにする。このようにして一部の計算対象を省くことで，計算速度の向上が期待されるが，一方で省かれた対象の中に正解のレコードが含まれる場合には，正解率は低下することになる。

層の数を変化させた際の，$R=1$ とした場合の正解率（$P(1)$）及び計算速度との関係について示したものが，以下の図5及び図6である（層の数が6個に対応する部分に点線を引いている。）。図5をみると，層の数が6個ま

表4　主成分分析の結果（上段：寄与率等，下段：因子負荷量）

	第1主成分	第2主成分	第3主成分	第4主成分	第5主成分	第6主成分	第7主成分
標準偏差	1.7080	1.0819	1.0022	0.9339	0.7602	0.5484	0.3964
寄与率	0.4168	0.1672	0.1435	0.1246	0.0826	0.0430	0.0224
累積寄与率	0.4168	0.5840	0.7275	0.8520	0.9346	0.9776	1.0000

	第1主成分	第2主成分	第3主成分	第4主成分	第5主成分	第6主成分	第7主成分
従業者数1	－0.534	0.139			0.211	－0.298	0.746
従業者数2	－0.51	0.27			0.229	－0.429	－0.654
資本金額	－0.401	－0.254	－0.122	0.141	－0.853	－0.114	
売上高	－0.511			－0.115		0.839	
産業		0.653		0.727	－0.114	0.104	
開設年	0.158	0.614	0.299	－0.592	－0.396		
地域		－0.182	0.939	0.291			

では正解率がそれほど大きく低下しておらず，層の数が7個を過ぎたあたりから右下がりで低下していることがわかる。図6で層の数が6個の場合をみると，層化を行わない場合（約1,600秒）と比較して，計算時間（テストデータに基づく回帰係数の推定にかかる時間）が半分以下（約600秒）にまで減少していることがわかる。

マッチング元及びマッチング先のデータの各レコードについて，第1主成分に対応する主成分得点をプロットしたものが図7及び図8である。これらの図には，層の数が6個の場合の各層の境界を併せて示している。主成分得点の大きさに応じて適当な数の層に分割されている様子がわかる。

なお，今回の主成分分析の目的は情報の縮約ではなく，マッチングの精度を維持しつつ計算効率を向上させるための層化を行うことにある。第1主成分は精度を落とさない層化に寄与しており，層化後のマッチング確率の推定では，地域や産業の情報がモデルの当てはまりに寄与していると考えられる。

5．おわりに

本稿では，多項ロジットモデルを用いた新たな統計的マッチングの方法を提案した。提案手法により，距離のウエイトを統計的に推

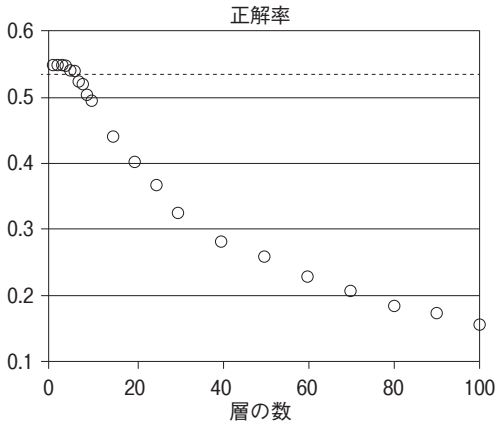

図5　マッチング正解率の比較

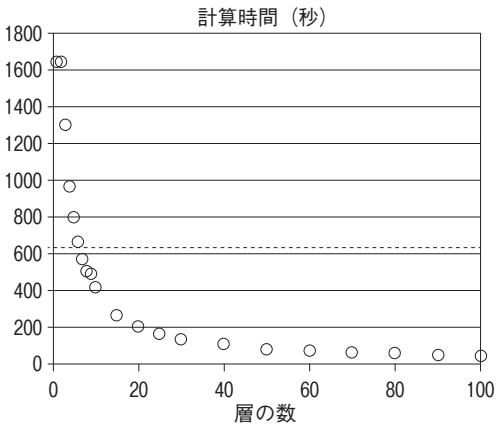

図6　層の数と計算時間との関係

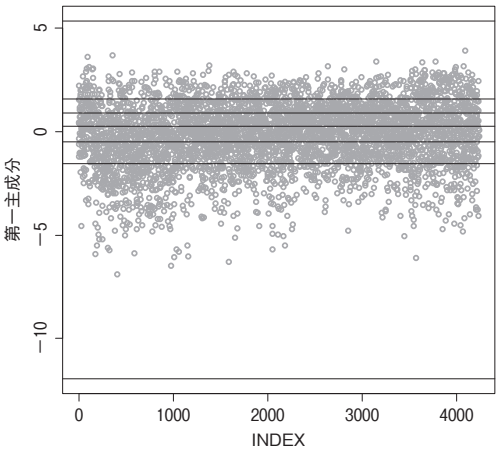

図7　マッチング元（帝国データバンク）第1主成分の主成分得点

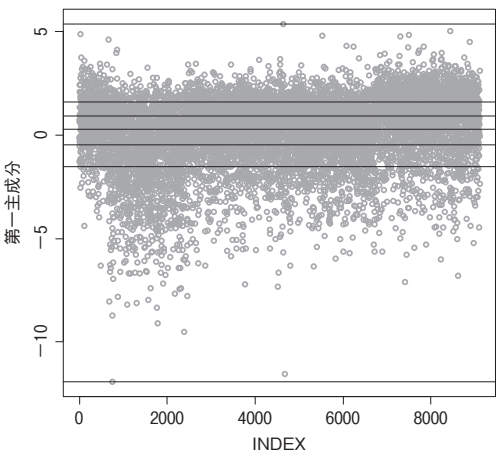

図8　マッチング先（経済センサス−活動調査）第1主成分の主成分得点

定することが可能となり，マッチングの程度を確率の形で表現することが可能となる。また，提案手法はデータの分布についての仮定を必要とせず，p値・t値や疑似決定係数によるモデルの比較・評価が可能となる。

提案手法は正解率の観点で，従前の研究で用いられている最近隣法よりも優れており，共通に利用できる情報が少ない企業データ間のマッチングにおいて有力な手法であることが分かった。また，絶対値距離を単純に用いるのではなく，対数変換することにより，距離が0に近い部分を強調し，距離が大きい部分の影響を抑えることで，モデルの当てはまりや正解率が向上することが分かった。

さらに，データの拡大に伴う計算時間の増加を考慮して，主成分分析による層化を基に計算を効率化する方法についても検討を行った。その結果，本稿で扱ったデータでは，層の数を6にすることで，正解率の低下を抑えつつ，計算時間を半分以下にまで減少できることがわかった。

本稿で提案した手法は，完全照合が可能なデータが得られる場合にのみ実行可能であるという点に注意を要する。完全照合が可能なデータが部分的にでも得られる場合には，それらの有効活用という観点から提案手法は有用であるものの，完全照合を行うための情報が存在しない場合には，モデルの推定を行うことができない。

今後の課題として，以下の3点が挙げられる。まず，主成分分析に基づく計算速度の向上に関して，今回は第1主成分のみを用いて層化を行ったが，第2主成分以降も有効に活用することで計算の範囲をさらに縮小し，計算速度を一層向上できる可能性がある。次に，推定されたマッチング確率を用いた，より効率的なマッチングのアルゴリズムの構築が考えられる。例えば企業は複数回選ばれることはないという条件の下で，マッチング確率の合計が最も小さくなるような割り当てを見出すことなどが考えられる。最後に，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。このようにして構成されたデータは，様々な分析に利用できる有用なものとなる可能性がある。

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。本稿で提案した手法も含め，継続的な手法の開発・改善を続けていく必要があると考える。

付表　テストデータにおける記述統計量（要約統計量及びカテゴリ別企業数）

	帝国データバンク			平成24年経済センサス活動調査			
	従業員数（人）	資本金（万円）	売上高（百万円）	従業者数（パート等含む）（人）	従業者数（パート等除く）（人）	資本金（万円）	売上高（百万円）
第1四分位	2.0	500.0	43.0	3.0	0.0	300.0	23.0
中央値	4.0	1000.0	95.0	6.0	2.0	975.0	62.0
平均値	9.6	1162.0	264.5	23.6	11.0	2309.0	431.6
第3四分位	9.0	1100.0	235.0	14.0	6.0	1000.0	175.3
標準偏差	25.4	950.9	765.9	334.7	81.1	31818.6	5028.5

産業大分類	帝国データバンク	平成24年経済センサス－活動調査
A 農業，林業	0	44
B 漁業	0	－
C 鉱業，採石業，砂利採取業	0	6
D 建設業	726	873
E 製造業	402	889
F 電気・ガス・熱供給・水道業	0	－
G 情報通信業	12	62
H 運輸業，郵便業	0	153
I 卸売業，小売業	520	1049
J 金融業，保険業	0	67
K 不動産業，物品賃貸業	136	406
L 学術研究，専門・技術サービス業	66	222
M 宿泊業，飲食サービス業	37	254
N 生活関連サービス業，娯楽業	39	187
O 教育，学習支援業	11	42
P 医療，福祉	9	67
Q 複合サービス事業	0	0
R サービス業（他に分類されないもの）	80	226

地域（市・郡）	帝国データバンク	平成24年経済センサス－活動調査
大津市	416	959
彦根市	166	384
長浜市	227	457
近江八幡市	105	241
草津市	175	417
守山市	92	221
栗東市	102	222
甲賀市	139	322
野洲市	56	160
湖南市	70	180
高島市	135	262
東近江市	178	360
米原市	60	126
蒲生郡	48	95
愛知郡	36	77
犬上郡	33	69

※少数のデータに関しては秘匿を行っている。

期間	帝国データバンク	平成24年経済センサス－活動調査
1984年以前	884	1785
1984年～1994年	618	981
1995年～2004年	399	1079
2005年以降	137	707

謝辞

本稿の内容の一部について，経済統計学会東北・関東支部例会及び全国研究大会において報告を行った際に，多くの方々から貴重なコメントをいただいた。ここに記して感謝の意を表したい。また，本稿について有益なコメントをしていただいた匿名の2名の査読者にも，感謝申し上げたい。なお，本稿の意見は筆者個人のものであり，所属する組織を代表するものではない。

本研究は科研費（16H02013 及び15H03390）の助成を受けている。

注

1）Gower距離は，統計的マッチングを扱うRのパッケージStatMatchでも採用されている（D'Orazio（2016））。

2）今回の分析では，統計解析ソフトウェアRと，その最適化関数optim（準ニュートン法）を使用して最尤推定の際の数値最適化の計算を行った。

3）p値，t値は，変数の選択に利用することができる。p値が大きく（t値の絶対値が小さく）回帰係数が有意でないと判断される場合には，モデルの当てはまりを向上させるために，当該変数をモデルから落とすことが考えられる。

4）これらのデータの中に大規模な親子企業やグループ企業などが存在する場合には，同一グループ内の企業を誤ってマッチングする可能性はあるものの，今回の分析では資本金額が300万円～5000万円とそれほど規模の大きくない企業を扱っており，影響はそれほどないと考えられる。また単一事業所企業と複数事業所企業では，企業の規模も異なってくるが，それらの違いについては規模に関する変数（売上高，従業員数等）により，ある程度捉えられると考えられる。

5）ICEの計算にはRのパッケージmiceを使用した。その際に，カテゴリ変数の欠測値は多項ロジットモデルにより，連続変数の欠測値はPredictive Mean Matchingにより，それぞれ補完を行った。

6）平成24年経済センサス－活動調査の産業共通調査票では，開設時期について，まず①昭和59年以前，②昭和60年～平成6年，③平成7年～平成16年，④平成17年以降の4つの区分について回答し，平成17年以降に開設した場合には年月を回答する形式となっている。今回の分析では，上記の区分に沿った形で開設年を4つに区分し，各カテゴリ内の企業数をある程度そろえるため，平成17年以降を1つの区分として用いている。

7）ロジットモデルの係数が目的変数（今回の分析では企業）によって異なると想定した場合を「多項ロジットモデル」とし，（今回のように）共通であると想定した場合を「条件付きロジットモデル」と定義する場合がある（Greene（2002））。また，今回のモデルには定数項を含めていないが，基準となる企業を定めて定数項を追加した場合でも，結果は変わらない。

8）Gower距離及びMahalanobis距離の計算には，統計解析ソフトウェアRのパッケージStatMatchを用いた。また，Mahalanobis距離の計算に当たっては，StatMatchの仕様に基づき，連続変数のみを使用している。

9）最近隣法（Nearest Neighbor Method）は，距離に基づき，最も近いレコードをマッチング先として探索する方法である。

10）交互作用項を導入したモデルについても検討を行ったが，疑似決定係数が向上するものの，その上昇分はわずかであり，一方で有意でなくなる変数が多くなったため，本稿では交互作用を扱っていない。

11）産業及び地域に関するウエイトが大きな値となっていることから，産業・地域別にサンプルを分割してマッチングを行うことも考えられる。ただし2つのデータの定義の違いや調査の時期の違い等により，正しいマッチング先であるレコードが同一のカテゴリに入らないケースがまれに生じることから，今回はサブサンプルによるマッチングという手法を行っていない。

12）主成分分析の計算には，Rのprincomp関数を使用した。

参考文献

［1］荒木万寿夫，美添泰人（2007）「家計データを利用した完全照合と統計的照合」『青山経営論集』第42巻第1号，pp.175-210.

［2］植松良和（2016a）「『公的統計ミクロデータ研究コンソーシアム』への期待：公的統計ミクロデータ研究コンソーシアム設立記念シンポジウムの報告」『ESTRELA』No. 269，pp.2-8.

［3］植松良和（2016b）「オーダーメード集計の見直し：基本計画策定から省令改正までの2年の経緯」『ESTRELA』No. 269，pp.14-20.

［4］栗原由紀子（2015）「統計的マッチングにおける推定精度とキー変数選択の効果：法人企業統計調査ミクロデータを対象として」『統計学』第108号，pp.1-15.

［5］小西葉子（2012）「生産動態統計調査と工業統計調査の事業所マッチング法について（2005年-2009年）」『RIETIディスカッションペーパー』No. 12-P-020.

［6］坂田幸繁，栗原由紀子（2011）「統計的マッチングによる疑似パネルデータの作成と精度検証：中小企業景況調査ミクロデータを利用して」『法政大学日本統計研究所オケージョナル・ペーパー』No. 27.

［7］坂田幸繁，栗原由紀子（2013）「法人企業統計のデータ・リンケージとその有効性の検証：統計的マッチング手法の比較から」『中央大学経済研究所年報』第44号，pp.271-306.

［8］村田磨理子，伊藤伸介（2016）「事業所・企業系のミクロデータを用いたデータリンケージの可能性：賃金構造基本統計調査を例に」『統計学』第110号，pp.1-17.

［9］美添泰人（2005）「統計的照合手法の基礎理論と最近の適用例」『青山経済論集』第56巻，pp.43-71.

［10］Buuren, S.（2012）, *Flexible imputation of missing data*, CRC press.

［11］Christen, P.（2012）, *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*, Springer.

［12］D'Orazio, M., Di Zio M., and Scanu, M.（2006）, *Statistical Matching: Theory and Practice*, Wiley.

［13］D'Orazio M.（2016）StatMatch: Statistical Matching. R package version 1.2.5. http://CRAN.R-project.org/package=StatMatch

［14］Fellegi, I.P. and Sunter, A.B.（1969）, "A theory for record linkage", *Journal of the American Statistical Association*, 64, pp.1183-1210.

［15］Gower J.C.（1971）, "A general coefficient of similarity and some of its properties", *Biometrics*, 27, pp.623-637.

［16］Greene. W.H.（2002）, *Econometric Analysis: 5th edition*, Prentice Hall.

［17］Harron, K., Goldstein, H. and Dibben, C.（2015）, *Methodological developments in data linkage*, Wiley.

［18］Herzog, T.N., Scheuren, F.J. and Winkler, W.E.（2007）, *Data quality and record linkage techniques*, Springer.

［19］Hosmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X.（2013）, *Applied logistic regression: Third edition*, Wiley.

［20］Kurihara, Y.（2015）, "Estimation of Durability of Profit of Small and Medium Enterprises by Statistical Matching", *Journal of Mathematics and System Science*, 5, pp.173-182.

［21］Lie, E.（2001）"Detecting abnormal operating performance: Revisited", F*inancial Management*, 30, pp.77-91.

［22］McCullagh, P., and Nelder, J.A.（1989）, *Generalized linear models: Second edition*, CRC press.

［23］Newcombe, H.B., Kennedy, J.M., Axford, S.J. and James, A.P.（1959）"Automatic Linkage of Vital Records", *Science*, 130, pp.954-959.

［24］Rässler, S.（2002）, *Statistical Matching*, Springer.

［25］Yoshikawa, Y., Iwata, T., Sawada, H. and Yamada, T., E.（2015）"Cross-domain matching for bag-of-words data via kernel embeddings of latent distributions", *Advances in Neural Information Processing Systems*, pp.1405-1413.

［26］Yoshizoe, Y. and Araki, M.（1999）, "Statistical matching of household survey files", Working Paper, No. 10, Aoyama Gakuin University.

New Statistical Matching Method Using Multinomial Logit Model

Isao Takabe*, Satoshi Yamashita**

Summary

Statistical matching techniques aim to build a useful data by combining different data sources. These techniques make it possible to create informative data without conducting any survey or collecting additional data. In recent years, matching techniques have been employed in various fields. In this study, we proposed a new statistical matching methodology by employing multinomial logit model. The weighted distance were used to compute the probability of true match pairs through the model. It is worth noting that working with large datasets entails a considerable amount of time to calculate the distances of all the possible pairs. To address this problem, we resorted to the principal component analysis method by dividing the data sources and making strata to shrink the searching space of the record pairs and search the true matched pairs more efficiently. We applied these techniques to a commercial company data and the official economic census microdata. The results showed that our method performs better than the nearest neighbor method in terms of true match rate.

Key Words

Statistical matching, multinomial logit model, weighted distance function, principal component analysis

* Statistics Bureau of Japan, The Graduate University for Advanced Studies
** The Institute of Statistical Mathematics

機関誌『統計学』投稿規程

経済統計学会（以下，本会）会則第3条に定める事業として，『統計学』（電子媒体を含む。以下，本誌）は原則として年に2回（9月，3月）発行される。本誌の編集は「経済統計学会編集委員会規程」（以下，委員会規程）にもとづき，編集委員会が行う。投稿は一般投稿と編集委員会による執筆依頼によるものとし，いずれの場合も原則として，本投稿規程にしたがって処理される。

1．総則

1-1　投稿者

会員（資格停止会員を除く）は本誌に投稿することができる。

1-2　非会員の投稿

(1)　原稿が複数の執筆者による場合，筆頭執筆者は本会会員でなければならない。

(2)　常任理事会と協議の上，編集委員会は非会員に投稿を依頼することができる。

(3)　本誌に投稿する非会員は，本投稿規程に同意したものとみなす。

1-3　未発表

投稿は未発表ないし他に公表予定のない原稿に限る。

1-4　投稿の採否

投稿の採否は，審査の結果にもとづき，編集委員会が決定する。その際，編集委員会は原稿の訂正を求めることがある。

1-5　執筆要綱

原稿作成には本会執筆要綱にしたがう。

2．記事の分類

2-1　研究論文

以下のいずれかに該当するもの。

(a)　統計およびそれに関連した分野において，新知見を含む会員の独創的な研究成果をまとめたもの。

(b)　学術的な新規性を有し，今後の研究の発展可能性を期待できるもので，速やかな成果の公表を目的とするもの。

2-2　報告論文

研究論文に準じる内容で，研究成果の速やかな報告をとくに目的とする。

2-3　書評

統計関連図書や会員の著書などの紹介・批評。

2-4　資料

各種統計の紹介・解題や会員が行った調査や統計についての記録など。

2-5　フォーラム

本会の運営方法や統計，統計学の諸問題にたいする意見・批判・反論など。

2-6　海外統計事情

諸外国の統計や学会などについての報告。

2-7　その他

全国研究大会・会員総会記事，支部だより，その他本会の目的を達成するために有益と

思われる記事。

3．原稿の提出

3-1　投稿

原稿の投稿は常時受け付ける。

3-2　原稿の送付

原則として，原稿は執筆者情報を匿名化したPDFファイルを電子メールに添付して編集委員長へ送付する。なお，ファイルは『統計学』の印刷レイアウトに準じたPDFファイルであることが望ましい。

3-3　原稿の返却

投稿された原稿（電子媒体を含む）は，一切返却しない。

3-4　校正

著者校正は初校のみとし，大幅な変更は認めない。初校は速やかに校正し期限までに返送するものとする。

3-5　投稿などにかかわる費用

⑴　投稿料は徴収しない。

⑵　掲載原稿の全部もしくは一部について電子媒体が提出されない場合，編集委員会は製版にかかる経費を執筆者（複数の場合には筆頭執筆者）に請求することができる。

⑶　別刷は，研究論文，報告論文については30部までを無料とし，それ以外は実費を徴収する。

⑷　3-4項にもかかわらず，原稿に大幅な変更が加えられた場合，編集委員会は掲載の留保または実費の徴収などを行うことがある。

⑸　非会員を共同執筆者とする投稿原稿が掲載された場合，その投稿が編集委員会の依頼によるときを除いて，当該非会員は年会費の半額を掲載料として，本会に納入しなければならない。

3-6　掲載証明

掲載が決定した原稿の「受理証明書」は学会長が交付する。

4．著作権

4-1　本誌の著作権は本会に帰属する。

4-2　本誌に掲載された記事の発行時に会員であった執筆者もしくはその遺族がその単著記事を転載するときには，出所を明示するものとする。また，その共同執筆記事の転載を希望する場合には，他の執筆者もしくはその遺族の同意を得て，所定の書面によって本会に申し出なければならない。

4-3　前項の規定にもかかわらず，共同執筆者もしくはその遺族が所在不明のため，もしくは正当な理由によりその同意を得られない場合には，本会が承認するものとする。

4-4　執筆者もしくはその遺族以外の者が転載を希望する場合には，所定の書面によって本会に願い出て，承認を得なければならない。

4-5　4-4項にもとづく転載にあたって，本会は転載料を徴収することができる。

4-6　会員あるいは本誌に掲載された記事の発行時に会員であった執筆者が記事をウェブ転載するときには，所定の書類によって本会に申し出なければならない。なお，執筆者が所属する機関によるウェブ転載申請については，本人の転載同意書を添付するものとする。

4-7 会員以外の者，機関等によるウェブ転載申請については，前号を準用するものとする。

4-8 転載を希望する記事の発行時に，その執筆者が非会員の場合には，4-4，4-5項を準用する。

1997年7月27日制定（2001年9月18日，2004年9月12日，2006年9月16日，2007年9月15日，2009年9月5日，2012年9月13日，2016年9月12日一部改正）

編集委員会からのお知らせ
機関誌『統計学』の編集・発行について

編集委員会

2016年9月より，新しい規程にもとづいて，「研究論文」と「報告論文」が設定されました。皆様からの積極的な投稿をお待ちしております。

1．投稿は，常時，受け付けています。なお，書評，資料および海外統計事情等については，下記の［注記3］をご確認下さい。
2．次号以降の発行予定日は，
　第116号：2019年3月31日，第117号：2019年9月30日です。
3．投稿に際しては，新規程にもとづく「投稿規程」，「執筆要綱」，および「査読要領」などをご熟読願います。最新版は，学会の公式ウェブサイトをご参照下さい。
4．原稿は編集委員長（下記メールアドレス）宛にお送り下さい。
5．原稿はPDF形式のファイルとして提出して下さい。また，紙媒体での提出も旧規程に準拠して受け付けます。紙媒体の送付先は編集委員長宛にお願いします（住所は会員名簿をご参照下さい）。
6．原則として，すべての投稿原稿が査読の対象となります。
7．投稿から発刊までに要する期間は，通常3ヶ月以上を要します。投稿にあたっては十分に留意して下さい。

編集委員会，投稿応募についての問い合わせは，
下記編集委員長宛メールアドレス宛に連絡下さい。

editorial@jsest.jp

編集委員長　水野谷武志（北海学園大学）
副委員長　池田　伸（立命館大学）
編集委員　小林良行（総務省統計研究研修所）
　　松川太一郎（鹿児島大学）
　　山田　満（東北・関東支部）

［注記1］『統計学』の定期刊行に努めておりますので，できるかぎり早期のご投稿をお願いします。116号（2019年3月31日発行予定）への掲載を想定した場合，「研究論文」と「報告論文」の原稿は，2019年1月初旬を目途として，遅くともそれまでにご投稿下さい。
［注記2］「研究論文」と「報告論文」は，別個に査読し，区分を変更しません。投稿に当たっては自分で申告して投稿しますが，この点ご留意下さい。
［注記3］書評，資料および海外統計事情等について，執筆，推薦，および依頼等をお考えの会員がおられましたら，企画や思いつきの段階で結構ですので，できるだけ早い段階で，編集委員会にご一報下さい。

以上

編集後記

本誌に投稿していただきました執筆者の皆様，そして快く査読をお引き受けいただきました査読者の皆様に改めてお礼申し上げます。引き続き，会員の皆様からの積極的な投稿をお待ちしております。

（水野谷武志　記）

執筆者紹介

高部　勲　（総務省統計局）
大澤理沙　（釧路公立大学経済学部）
稲葉和夫　（立命館大学経済学部）
山下智志　（統計数理研究所）
橋本貴彦　（立命館大学経済学部）

支部名　事務局

支部名		事務局	
北海道 ……	062-8605	札幌市豊平区旭町 4-1-40 北海学園大学経済学部 （011-841-1161）	水野谷武志
東北・関東 ……	192-0393	八王子市東中野 742-1 中央大学経済学部 （042-674-3406）	伊藤伸介
関西 ……	640-8510	和歌山市栄谷 930 和歌山大学観光学部 （073-457-8557）	大井達雄
九州 ……	870-1192	大分市大字旦野原 700 大分大学経済学部 （097-554-7706）	西村善博

『統計学』編集委員

統　計　学　No.115

2018年9月30日　発行

発行所　経済統計学会
〒112-0013　東京都文京区音羽1-6-9
音羽リスマチック株式会社
TEL/FAX　03（3945）3227
E-mail：office@jsest.jp
http://www.jsest.jp/

発行人　代表者　西村善博

発売所　音羽リスマチック株式会社
〒112-0013　東京都文京区音羽1-6-9
TEL/FAX　03（3945）3227
E-mail：otorisu@jupiter.ocn.ne.jp
代表者　遠藤　誠

昭和情報プロセス㈱印刷