企業データの統計的マッチング及びその精度改善

高部　勲[†]
山下　智志[††]

Statistical Matching of Corporate Data and Improvement of Its Accuracy

TAKABE Isao
YAMASHITA Satoshi

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、Takabe and Yamashita(2020)及び高部・山下(2019)で提案された、多項ロジットモデルに基づく統計的マッチングの手法をさらに発展させ、通常のモデル及び Recipient と Donor の転置処理を行ったモデルによるマッチング確率の加重平均を距離として、ウエイト付き距離に基づく統計的最適マッチングを行った。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、従来の手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

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 based on Takabe and Yamashita (2019) and Takabe and Yamashita (2019). We did statistical matching using new distance measure which is the weighted mean of probabilities estimated by using previous multinomial logit model and transposed model that exchange the rolls of donor and recipient data. 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 previous statistical matching methods in terms of true match rate.

Key Words　Statistical matching, Multinomial logit model, Weighted distance function

[†]　総務省統計局統計データ利活用センター、統計数理研究所　Email：takabe.isao@ism.ac.jp
[††] 統計数理研究所　Email：yamasita@ism.ac.jp

1. はじめに

近年、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、情報量の多い有用なデータを構築することができる。こうした中、複数のデータを結合するためのデータリンケージの手法が、様々な分野で注目を集めている（Herzog et al.(2007)、Christen(2012)、Harron et al.(2015)）。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、それらを利用してレコードを結合する完全照合（Exact matching）を行うことができる（企業データの完全照合については村田・伊藤(2016)を参照。世帯データの完全照合については山口(2014)を参照）。しかし、このような照合キーの情報が利用できない場合には、各データの共通変数を基に算した距離が近い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（美添(2005)）。これらの方法の関係を整理したものが、以下の図１である。

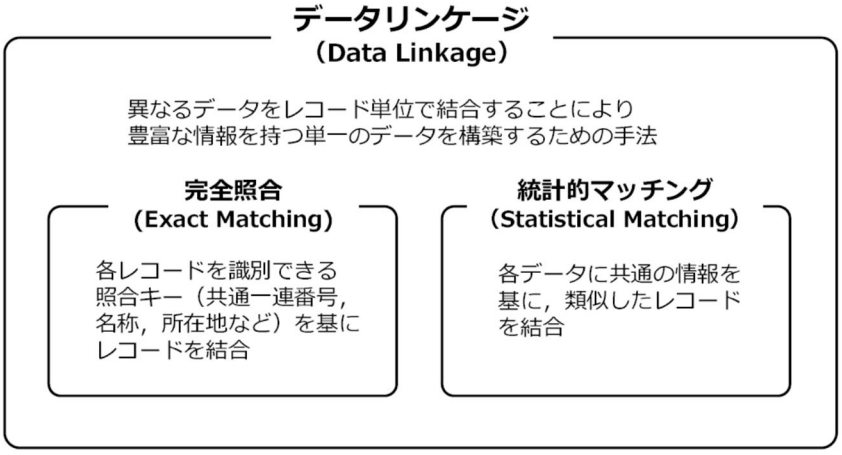

図１　データリンケージと統計的マッチング等との関係

統計的マッチングのイメージを示したものが、以下の図２である。図２は、照合キー（W_i及びW_j）が利用できない中で、データ A の i 番目のレコードに対して、共通変数 X_i 及び X_j を基に算出した距離が最も近いデータ B の j 番目のレコードをマッチングした結果が、新たなデータ（マッチングデータ）の l 番目のレコードになる様子を示している。

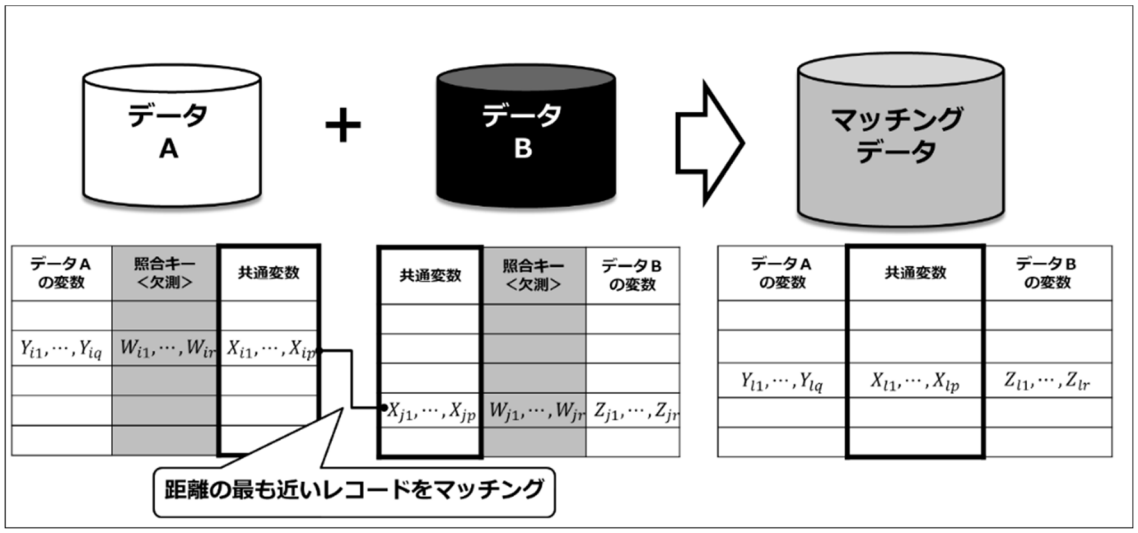

図２　統計的マッチングのイメージ

我が国では以前から、ミクロデータの利活用拡充の必要性とともに、既存のデータのリンケージによる新たな統計の作成の重要性についても指摘されていた（森(2008a)、森(2008b)及び山口(2014)）。伊藤(2018)は、今般の統計法の改正に関連して、公的統計ミクロデータと行政記録情報などの他のデータとの結合についても詳しく述べている。このような状況を踏まえると、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統計的マッチングに関する研究は、既存のデータを有効に活用した有用なデータの構築の進展につながるものであり、今後、重要な研究テーマになると考えられる。

統計的マッチングに関する研究は、諸外国では1960年代から行われてきており、これまでに、様々な手法が研究・開発されてきている。初期には、データリンケージの自動化を目的として、名称・所在地などを基に、異なるレコードを同一と判定する確率と、同一の対象を表すレコードが正しく同一であると判定される確率の比率を基にマッチングの適否を判定する方法（Newcombe(1959)、Fellegi and Sunter(1969)）が開発され、その後も長く研究されてきたが、この方法では、名称・所在地などの詳細な文字情報が利用できる状況を想定しており、一方で、そのような情報が使用できず、売上高や資本金額、従業者数などの限られた情報のみが利用できる状況では、このような方法は適していない。

共通変数以外の変数を欠測値とみなして、重回帰モデルやベイズ統計学の枠組みに基づき推測を行う方法（D’ Orazio et al.(2006)、Rässler(2002)、栗原(2015)）も研究されているが、企業データには売上高や従業者数など、はずれ値を含む歪んだ分布を持つ変数が含まれており、また、連続変数とカテゴリ変数（産業、地域等）が混在するなど、多変量分布などの特定の分布の仮定が当てはまらない状況が想定されることから、この方法も適していない。

各レコードがどちらのデータに属するかという確率（傾向スコア(Propensity Score)）の値が近いレコード同士をマッチングする方法（Rubin(1986)及びStuart(2010)）などの様々な手法が、研究・開発されている。ただしこの方法では、マッチングの正しさを定量的な形で評価することができず、また、複数のデータや手法間のマッチングの精度の比較を行うことができないという課題がある。

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きた方法である。これは、各データに共通の変数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方法である（D’ Orazio et al.(2006)）。この方法では、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一般的な基準が無く、各変数のウエイトの決定方法が恣意的になるおそれがある。この問題に対応するため、Takabe and Yamashita(2020)及び高部・山下(2019)では、多項ロジットモデルを用いた統計的マッチングの手法を提案している。この方法では、レコード間の距離を説明変数として多項ロジットモデルを構築することにより、レコード間の距離を推定し、マッチングを行う手法である。この手法は、前述の先行研究における課題を、以下のように克服している点が特長である。

・距離のウエイトをデータから推定することが可能
・名称・所在地などの詳細な文字情報が利用できない場合でも、適用可能
・連続変数とカテゴリ変数が混在する場合でも適用可能
・レコードの一致確率（マッチング確率）を推定し、マッチングの精度を定量的に評価することが可能
・データの構造として特定の分布（多変量正規分布等）を仮定する必要がない

本稿では、Takabe and Yamashita(2020)及び高部・山下(2019)における多項ロジットモデルを用いた統計的マッチングの手法をさらに発展させ、通常のモデル及びマッチング元

（Donor）とマッチング先（Recipient）の転置処理を行ったモデルにより推定されたマッチング確率の加重平均を新たにレコード間の距離とみなして統計的マッチングを行う方法を提案する。提案手法を経済センサスの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、マッチングの正解率の観点から、従来の手法よりも優れていることが示された。

2. 多項ロジットモデルに基づく統計的マッチング

2.1 手法の概要

ここでは、Takabe and Yamashita (2020)及び高部・山下(2019)を基に、多項ロジットモデルに基づく統計的マッチングの手法について説明する。以下の２種類のデータ（データ A 及びデータ B）の統計的マッチングを行う場合を想定する。

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

このとき、データ A の i 番目のレコードと、データ B の j 番目のレコードが同一の対象である確率 P_{ij} を考える（以下、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。

$$P_{ij} = \exp(-D_{ij}) / \sum_{j=1}^{N} \exp(-D_{ij}) \tag{1}$$

距離 D_{ij} の形式については、以下のような、様々なものが考えられる。

絶対値距離（Manhattan 距離）：$D_{ij} = \sum_{k=1}^{p} \beta_k |X_{ik} - X_{jk}|$ (2)

Euclid 距離（２乗）：$D_{ij} = \sum_{k=1}^{p} \beta_k (X_{ik} - X_{jk})^2$ (3)

Mahalanobis 距離：$D_{ij} = (\boldsymbol{X}_i - \boldsymbol{X}_j)^T \Sigma_{XX}^{-1} (\boldsymbol{X}_i - \boldsymbol{X}_j)$ (4)

ここで、$\boldsymbol{X}_i = (X_{i1}, X_{i2}, \cdots, X_{ip})^T$、$\boldsymbol{X}_j = (X_{j1}, X_{j2}, \cdots, X_{jp})^T$ である。また、Σ_{XX}^{-1} は、共通変数の分散共分散行列を表している。

カテゴリ変数（離散変数）に対しては、以下の距離が用いられる。

$$D_{ij} = \sum_{k=1}^{p} \beta_k I(X_{ik} = X_{jk}) \tag{5}$$

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

共通変数に、連続変数とカテゴリ変数の両方が含まれる場合には、式(2)及び式(5)を組み合わせた Gower 距離が用いられる(Gower(1971))。

$$D_{ij} = \sum_{k=1}^{p} D_{ijk} / P \tag{6}$$

ここで D_{ijs} は、変数が連続変数の場合には $D_{ijs} = |X_{ik} - X_{jk}| / R_k$（$R_k$ は k 番目の変数のレンジ・範囲（最大値と最小値の差））、カテゴリ変数の場合には $D_{ijs} = I(X_{ik} = X_{jk})$ として定義される。これらは、距離のウエイト β_k をレンジの逆数あるいは 1 に固定したものとみることができる。

距離のウエイト β_k を推定することができれば、全てのレコードの組合せに対して距離 D_{ij} を計算することが可能となる。そして、距離 D_{ij} の値を基に、式(1)を用いてマッチング確率 P_{ij} を推定し、その値が最も大きいレコードと結合することにより、統計的マッチングを行うことができる。多項ロジットモデルに基づく統計的マッチングのイメージを示したものが、以下の図３である。

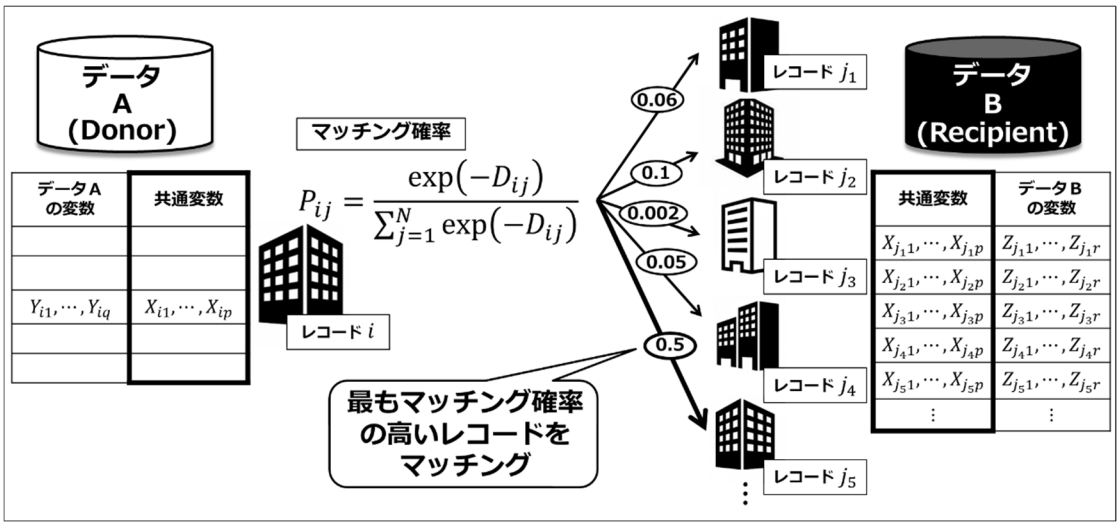

図３　多項ロジットモデルに基づく統計的マッチングのイメージ

次に、距離のウエイト β_k の推定方法について述べる。式(1)を基に、対数尤度関数 L を、以下のように構成することができる。

$$L(\boldsymbol{\beta}) = \log\left(\prod_{i=1}^{M}\prod_{j=1}^{N} P_{ij}(\boldsymbol{\beta})^{\delta_{ij}}\right) \qquad (7)$$
$$= \sum_{i=1}^{M}\sum_{j=1}^{N} \delta_{ij}\log\left(P_{ij}(\boldsymbol{\beta})\right)$$

ここで、$\boldsymbol{\beta}=(\beta_1,\beta_2,\cdots,\beta_p)$ は、距離に含まれるウエイトを表している。また、δ_{ij} は、データ A（Donner）のレコード i と、データ B（Recipient）のレコード j が表す対象が同一の場合に 1、それ以外の場合に 0 となる変数である。δ_{ij} に関する情報は、後述する学習用データから得られる。なお、上記の方法においては、マッチング元(Donner)及びマッチング先(Recipient)の両方のデータの中に、同一の企業を表すレコードが存在することが仮定されていることを注記しておく。

式(7)において、対数尤度関数 L はマッチング確率 P_{ij} に含まれる距離を通してウエイト $\boldsymbol{\beta}$ に依存していることから、このことを明示的に表すために、$L(\boldsymbol{\beta})$ 及び $P_{ij}(\boldsymbol{\beta})$ と表現している。式(7)の対数尤度関数 L をウエイト $\boldsymbol{\beta}$ に関して最大化することにより、ウエイトの最尤推定値 $\hat{\boldsymbol{\beta}}=(\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_p)$ が得られる（$\hat{\boldsymbol{\beta}}=\underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, L(\boldsymbol{\beta})$）。

式(7)の対数尤度関数 L の最大値は解析的に求めることができないため、ニュートン法に基づく逐次計算により数値的に求める。本研究では、R の最適化関数 optim を使用して数値的最適化を行っており、その際に BFGS 法に基づく準ニュートン法のオプション（method = BFGS）を使用して計算を行っている。準ニュートン法及び BFGS 法の詳細については、今野(1978)を参照。

多項ロジットモデルの枠組みで距離のウエイトの推定を行うことにより、 t 値や p 値などの統計量を計算することが可能となり、これらの統計量を用いてウエイトの推定精度を分析することが可能となる。また、Mcfadden の疑似決定係数（Hosmer et.al.(2013)及び山下(2005)）を用いることにより、異なるモデルのデータへの当てはまりの程度を比較することも可能となる。

2.2 マッチング元(Donner)及びマッチング先(Recipient)のデータ量に関する留意点

前節においては、マッチング先のデータに、同一の企業が必ず含まれる状況を想定していた。しかし、もし$M > N$であれば、式(1)で表されるマッチング確率に、N/Mを乗ずる必要がある。この点について、以下で説明する。

まず、$M > N$の場合には、マッチング先に同一の企業が含まれないレコードが、マッチング元のデータの中に、確実に存在する。そこで、マッチング確率を、以下のように分解することを考える。

$$P_{ij} = P_{ij}(u|t,\boldsymbol{\beta})P(t) \qquad (8)$$

ここで、$P(t)$は、データ A のあるレコードに対して、データ B の中に、対応する同一の企業のレコードが存在する確率を表す（存在する場合を$t = 1$、存在しない場合を$t = 0$とする）。また、$P_{ij}(u|t)$は、tを条件付けた場合に、レコードiがレコードjと一致する確率を表しており、$t = 1$の場合には、$P_{ij}(u|t = 1)$は式(1)で表される確率となり、$t = 0$の場合には、$P_{ij}(u|t = 0) = 0$となる。M個のレコードのうち、N個のレコードに、対応する同一企業が存在する場合には、$P(t) = N/M$となる。

以上を踏まえると、式(7)の対数尤度関数は、以下のように変形される。

$$\begin{aligned} L(\boldsymbol{\beta}) &= \sum_{i=1}^{M}\sum_{j=1}^{N}\delta_{ij}\log\left(P_{ij}(u|t,\boldsymbol{\beta})P(t)\right) \qquad (9)\\ &= \sum_{i=1}^{M}\sum_{j=1}^{N}\delta_{ij}\log\left(P_{ij}(u|t,\boldsymbol{\beta})\,N/M\right)\\ &= \sum_{i=1}^{M}\sum_{j=1}^{N}\delta_{ij}\log\left(P_{ij}(u|t,\boldsymbol{\beta})\right) - M(\log N - \log M) \end{aligned}$$

ここで、$\sum_{j=1}^{N}\delta_{ij} = 1$という事実を用いている。$N/M$に対応する最後の項（$M(\log N - \log M)$）は定数のため、最尤法によりパラメータの推定を行う際には、これらの項は影響しないが、後述するように、マッチング元(Donner)及びマッチング先(Recipient)の転置処理を行ったデータでは、マッチング確率の加重平均を行う際に、その計算結果に影響を与えることとなる。こうした結果については、次節及び 7 節で述べることとする。

3. 提案手法：Recipient と Donor の転置処理に基づく方法

ここでは、前節で紹介した、多項ロジットモデルに基づく統計的マッチング手法のマッチング精度を改善するために提案する手法について説明する。前節の方法では、マッチング確率を用いることにより、データ A のあるレコードの側から見た、データ B の最適なレコードの候補を見つけることはできる。しかし、そのデータ A のレコードが、逆にデータ B の側から見て、最適な候補であるとは限らず、より適切なマッチングの候補がデータ A の中に存在する可能性もある（次ページの図 4 を参照）。

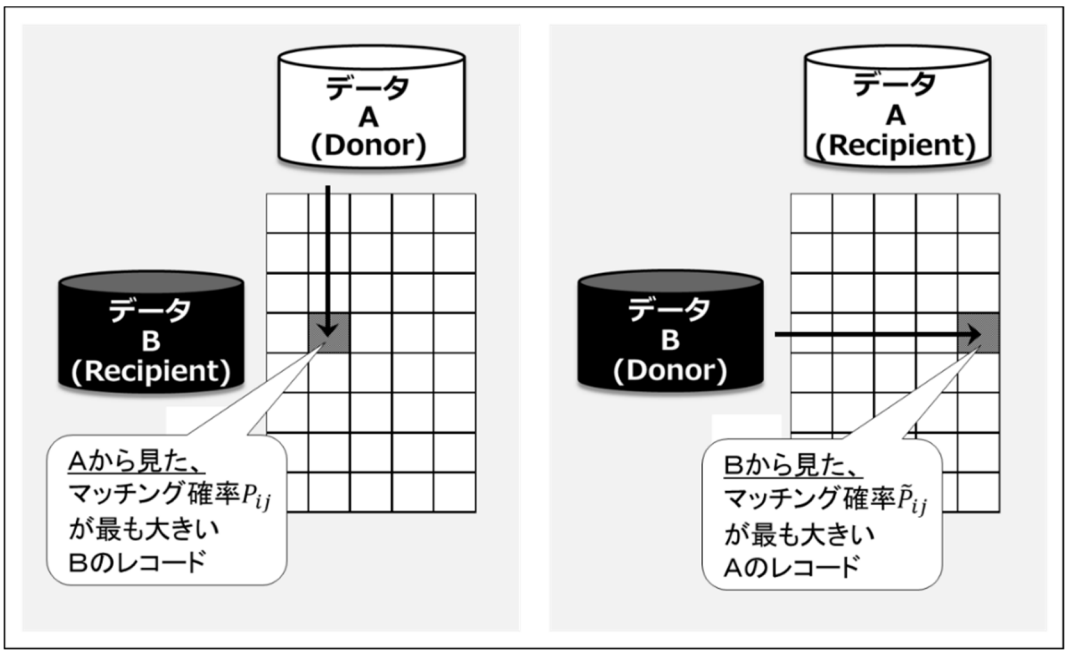

図４　各データから見てマッチング確率最大のレコードが一致しない場合

こうした状況を踏まえつつ、データ B の側から推定を行ったマッチング確率の情報も付加することによって、より適切なマッチングの結果が得られると考えられる。そこで、Donor と Recipient の役割を交換した（転置処理した）形で推定した多項ロジットモデルを基に、データ B から見た場合のマッチング確率 $\tilde{P}_{ij}$ を算出し、これに対して、適当なウエイト ω を用いて、元の（データ A から見た場合の）マッチング確率 P_{ij} との加重平均（Q_{ij}）を算出し、これを新たにレコード間の距離とみて、データ A から見て Q_{ij} が最大となるレコードを探索することにより、統計的マッチングを行うことを考える。

ここで、加重平均を計算する際に、単純な平均（$(1-\omega)P_{ij}+\omega\tilde{P}_{ij}$）と、幾何平均（${P_{ij}}^{1-\omega}{\tilde{P}_{ij}}^{\omega}$）の２種類の計算方法が考えられる。本稿では、以下の２つの式により Q_{ij} を算出し、それらを用いた場合の精度の比較も行う。

$$Q_{ij}=(1-\omega)P_{ij}+\omega\tilde{P}_{ij} \qquad (10)$$
$$Q_{ij}=(1-\omega)\log[P_{ij}]+\omega\log[\tilde{P}_{ij}] \qquad (11)$$

式(11)の幾何平均については、計算を行いやすいように対数変換を行った形で定義を行っている。式(10)及び式(11)のいずれの式においても、加重平均のウエイト ω が 0 の場合には、通常の多項ロジットモデルに基づくマッチング確率のみを用いること対応している。また、ω の値が大きいほど、Donor と Recipient の転置処理を行ったモデルに基づくマッチング確率 $\tilde{P}_{ij}$ を取り入れる割合が大きくなる。

ここで、Donor と Recipient を転置処理したデータについては、Donor のデータ量が Recipient のデータ量を上回っていることから（$M>N$）、2.2 節で述べた状況が成立しているため、データ量に応じた事前確率を考慮する必要がある（転置処理を行っているため、M と N が逆になっていることに注意）。よって、2.2 節で述べた事実を踏まえると、転置処理した後のマッチング確率は、$\tilde{P}_{ij}$ を式(7)で推定したものとすると、正確には $\tilde{P}_{ij}\,N/M$ となる。このような状況を考慮して式(10)及び式(11)を修正した式は以下のようになる。

$$Q_{ij}=(1-\omega)P_{ij}+\omega\tilde{P}_{ij}\,N/M \qquad (12)$$

$$Q_{ij} = (1-\omega)\log\left(P_{ij}\right) + \omega\left[\log\left(\tilde{P}_{ij}\right) N/M\right] \qquad (13)$$
$$= (1-\omega)\log\left(P_{ij}\right) + \omega\left[\log\left(\tilde{P}_{ij}\right) + \log N - \log M\right]$$

これらの式を距離として用いて、マッチングを行うこととする。以上の内容を踏まえた、本稿で提案する手法に関する手順のイメージを示したものが、次ページの図５である。

加重平均のウエイト ω の簡易な設定方法としては、$\omega = 0.5$として、単純な平均を用いる方法が考えられるが、データの量も考慮する必要があり、単純平均以外のウエイトの方が、マッチングの精度が向上する可能性もある。データに基づく最適なウエイトの見積もりについては、７節の提案手法と従来の手法との比較において分析する。

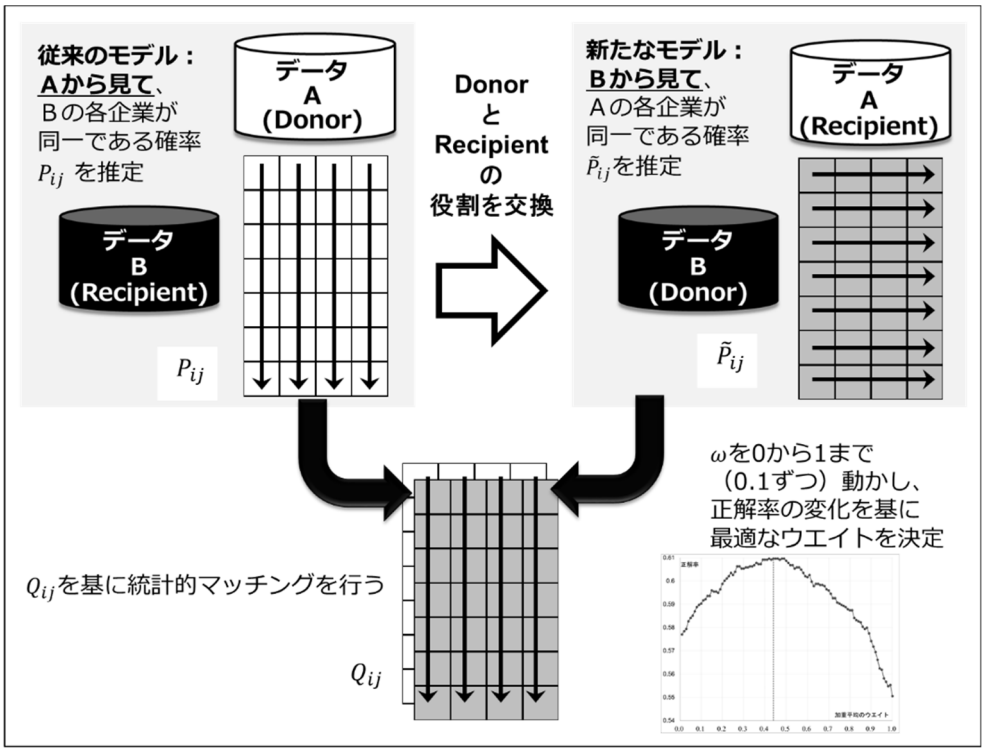

図５　Donor と Recipient の転置処理に基づく統計的マッチング

4. データ

4.1 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の概要

本稿では、平成 24 年経済センサス - 活動調査のミクロデータ（経済センサスミクロデータ）及び帝国データバンクの企業データ（帝国データバンクデータ）を対象として統計的マッチングを行う。また、今回の分析に当たり、推定の対象地域については、Takabe and Yamashita(2020)及び高部・山下(2019)よりも多い、３つの県のデータを対象に推定を行う。データの概要と、分析に当たって調整等を行った点については、以下のとおり。

【経済センサスミクロデータ】

・分析には、平成 24 年調査の結果を使用（調査の期日は平成 24 年２月１日現在）
・調査票情報については、統計法第 33 条２号に基づく二次的利用の制度により提供を受けたものである。
・データには一部の変数に関して、欠測値が含まれていることから、MICE：Multivariate Imputation by Chained Equations（Buuren(2012)）の手法に基づき、欠測

値の補完を行う。MICE の計算には R のパッケージ mice（Buuren and Groothuis-Oudshoorn(2010)）を使用する。その際に、連続変数の欠測値については Predictive Mean Matching により、カテゴリ変数の欠測値については多項ロジットモデルにより、それぞれ補完を行う。

【帝国データバンクデータ】

・「COSMOSII」企業概要ファイル・レイアウトＣを使用
・データの時点については、平成 24 年経済センサス - 活動調査の実施時期と合わせるために、平成 24 年２月時点とした。
・本稿では、利用可能な情報が少ない中小企業を対象とし、帝国データバンクのデータについては、資本金 300 万円以上 5,000 万円未満の企業を対象としている。
・日本標準産業分類と類似した TDB 産業分類コードが付与されている。
・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。データには欠測値は含まれていない。

以上の処理により、帝国データバンクデータのレコードが経済センサスミクロデータの中に必ず存在するという状況となっている。本研究においては、このように、δ_{ij} に関する情報、すなわち完全照合の有無に関する情報を持つ学習用データが存在することが仮定されていることを注記しておく。

4.2 分析用データの作成

次に、上記の完全照合後のデータについて、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の各データから 2/3 のレコードを無作為抽出して学習用データとした。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のテストデータとした。

なお、本稿では、多項ロジットモデルのパラメータを推定する際の、完全照合が完了しているレコードに関するデータを学習用データと呼び、完全照合ができていないレコードからなるマッチングの精度検証を行うためのデータをテストデータと呼んでいるが、これらは、機械学習におけるハイパーパラメータの推定など、汎化誤差を考慮した分析の際に、一般的に利用されている用語とは意味が異なる点を、注記しておく。
地域ごとの学習用データ及びテストデータのレコード数について示したものが、以下の表１である。

表１　学習用データ及びテストデータのレコード数

		地域Ａ	地域Ｂ	地域Ｃ
(1)	学習用データ	13,267	14,735	18,137
	経済センサスミクロデータ	9,105	9,649	12,583
	帝国データバンクデータ	4,162	5,086	5,554
(2)	テストデータ	6,668	7,297	9,051
	経済センサスミクロデータ	4,552	4,825	6,292
	帝国データバンクデータ	2,116	2,472	2,759
合計（(1)＋(2)）		19,935	22,032	27,188

経済センサスミクロデータおよび帝国データバンクデータの両方のデータに共通に含まれる変数（共通変数）は、以下の表２に示した７種類である。

表２　分析に使用する変数

変数	種類	単位・区分
従業者数（従業員数）	連続変数	人
売上高	連続変数	百万円
資本金額	連続変数	万円
産業分類（大分類）	カテゴリ変数	18 区分
開設年	カテゴリ変数	4区分
地域	カテゴリ変数	市又は郡に応じた区分
経営組織	カテゴリ変数	3区分（株式・有限・不明）

なお、経済センサスミクロデータと帝国データバンクデータでは、変数の定義などに違いがあり、そのままでは分析に用いることができないことから、以下に示すように、各種の調整を行っている。

【従業者数及び従業員数】
- データを事前に比較・分析した結果、帝国データバンクデータの従業員数には、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想定されるデータが見受けられた。
- これに対応する経済センサスミクロデータの従業者数については、パート・アルバイトを含む場合と含まない場合のどちらの情報も得られる。
- そこで、上記の２つの場合に関して距離を計算し、このうち小さい方を、従業者数・従業員数に関する距離とする。

【産業】
- 各データには産業大分類の情報を付与する。その際に、帝国データバンクデータで用いられている TDB 産業分類の大分類を、平成 24 年経済センサス - 活動調査で用いら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する。
- 産業大分類のうち、「S：公務(他に分類されないものを除く)」及び「T：分類不能の産業」については、本研究における分析の対象外とする。
- 日本標準産業分類については、平成 25 年 10 月改定(第 13 回改定)を用いる。

【開設年】
- 帝国データバンクデータには、企業の開設年の情報が年単位で記録されている。
- 一方で、平成 24 年経済センサス - 活動調査では、開設時期について、いくつかのカテゴリから選択する形になっている。
- 本稿では、両データの開設年の粒度を合わせるために、経済センサスの調査事項を参考に、開設年を以下の４つの時期に区分して、カテゴリ変数として使用する。
  (1) 1984 年以前
  (2) 1985 年 ～ 1994 年
  (3) 1995 年 ～ 2004 年
  (4) 2005 年以降

なお、完全照合を行うための照合キーが利用不可能な場合に、本稿では、距離に利用するデータが共通であり、分布が同一であることを仮定しており、両者の分布（平均値

など）に大きな違いがないことを前提としている。上記の従業者数と従業員数のように定義が異なる場合など、一方のデータのある変数が他方のデータの対応する変数の代理変数となっており、それらの分布が大きく異なると想定されるような場合には、こうした前提が成立しているかという点に留意する必要がある。

5. 多項ロジットモデルの推定

データには連続変数とカテゴリ変数が含まれていることから、距離として、式(2)及び式(5)を組み合わせた形の距離を用いる。さらに本研究では、式(2)の絶対値距離の対数変換値（以下の式(14)）を用いた多項ロジットモデルについても推定を行う。その際に、距離が０となり、対数を計算できない可能性があることから、１を加算した上で対数変換を行うこととする。

$$\text{絶対値距離（Manhattan 距離）の対数：} D_{ij} = \sum_{k=1}^{p}\left(\beta_k \left|X_{ik} - X_{jk}\right| + 1\right) \qquad (14)$$

これまでに述べた内容を踏まえ、本項では、以下の３種類の距離を用いた多項ロジットモデルの推定結果を比較する。

(1) ウエイト付き Euclid 距離（２乗）
(2) ウエイト付き絶対値距離
(3) ウエイト付き絶対値距離の対数変換

なお、Euclid 距離や絶対値距離を用いた場合、距離の大きさ（数値の桁数）が変数間で大きく異なることがあり、その場合、最尤法の数値計算が安定せず、パラメータの推定に失敗する可能性がある。そこで、変数の大きさに応じて適当な整数で割って標準化を行う方法（スケーリング）が用いられる（今野(1978)）。本研究においても、連続変数に関して、以下の表３に示した定数を乗じることにより、各変数のスケーリングを行う。

表３　各変数のスケーリング

	Euclid 距離（２乗）	絶対値距離
従業者数・従業員数	1/1,000	1/10
売上高	1/1,000,000	1/100
資本金額	1/1,000,000	1/100

これらの距離に基づく多項ロジットモデルの推定結果を示したものが、表４～表６である。各表には、多項ロジットモデルの回帰係数（ウエイト）とともに、それらの t 値も示している。また、各モデルのデータへの当てはまりをみるために、McFadden の疑似決定係数（自由度調整済みを含む）についても、合わせて示している。

推定結果を見ると、全ての地域で、どのモデルにおいても回帰係数の標準誤差は十分に小さくなっており、ほぼ全ての変数について 0.1 パーセントの有意水準で有意となっている。また、McFadden の疑似決定係数を比較すると、対数変換し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となっている。

以上の結果から、マッチングの正解率の観点からみた場合についても、対数変換を行ったウエイト付き絶対値距離（式(14)）に基づくモデルが、全ての地域を通じて最も優れていることが示された。よって、次節以降のマッチング確率の推定の際には、対数変換を行ったウエイト付き絶対値距離を用いることとする。

表４　多項ロジットモデルの推定結果（地域Ａ）

	[1]Euclid距離（二乗）	[2]絶対値距離	[3]絶対値距離（log）
従業員数	0.9227 ***	2.0864 ***	1.0630 ***
	(11.710)	(29.672)	(34.021)
資本金額	1.3526 ***	0.5561 ***	0.8024 ***
	(28.841)	(40.871)	(54.745)
売上高	0.1617 **	0.7826 ***	0.9288 ***
	(5.758)	(33.472)	(64.071)
産業	3.7348 ***	3.6093 ***	3.5074 ***
	(71.751)	(65.336)	(62.438)
開設年	1.5765 ***	1.4919 ***	1.4813 ***
	(43.637)	(38.059)	(34.944)
地域（市・郡）	9.9063 ***	16.4215 ***	9.4954 ***
	(17.137)	(10.519)	(18.940)
経営組織（株式・有限会社）	4.7835 ***	4.6189 ***	4.4628 ***
	(23.016)	(19.894)	(20.164)
初期対数尤度	-37943	-37943	-37943
対数尤度	-16419	-12259	-9350
疑似決定係数	0.5673	0.6769	0.7536
自由度調整済疑似決定係数	0.5671	0.6767	0.7534

*** p < 0.001, ** p < 0.01, * p < 0.05, - p < 0.01　( )は t 値

表５　多項ロジットモデルの推定結果（地域Ｂ）

	[1]Euclid距離（二乗）	[2]絶対値距離	[3]絶対値距離（log）
従業員数	0.2789 ***	1.6242 ***	1.1215 ***
	(13.712)	(35.424)	(46.898)
資本金額	0.0350 ***	0.3331 ***	0.8064 ***
	(9.452)	(49.826)	(72.644)
売上高	0.0003 ***	0.1970 ***	0.7894 ***
	(39.819)	(20.020)	(61.920)
産業	3.6448 ***	3.4903 ***	3.3997 ***
	(82.588)	(78.082)	(73.641)
開設年	1.4869 ***	1.3786 ***	1.3427 ***
	(47.377)	(42.281)	(38.697)
地域（市・郡）	9.0754 ***	9.4280 ***	9.0213 ***
	(18.313)	(19.331)	(22.027)
経営組織（株式・有限会社）	4.1955 ***	3.7069 ***	3.8687 ***
	(33.968)	(29.377)	(29.407)
初期対数尤度	-46662	-46662	-46662
対数尤度	-26261	-20869	-15574
疑似決定係数	0.4372	0.5528	0.6662
自由度調整済疑似決定係数	0.4371	0.5526	0.6661

*** p < 0.001, ** p < 0.01, * p < 0.05, - p < 0.01　( )は t 値

表６　多項ロジットモデルの推定結果（地域Ｃ）

	[1]Euclid距離（二乗）	[2]絶対値距離	[3]絶対値距離（log）
従業員数	1.0390 ***	1.8428 ***	1.0744 ***
	(16.742)	(36.109)	(47.554)
資本金額	0.1825 ***	0.4565 ***	0.7507 ***
	(17.890)	(48.271)	(68.587)
売上高	0.1004 **	0.3780 ***	0.7321 ***
	(8.185)	(34.082)	(60.766)
産業	3.7187 ***	3.5794 ***	3.4925 ***
	(86.053)	(80.652)	(78.185)
開設年	1.6310 ***	1.5154 ***	1.4784 ***
	(51.603)	(45.715)	(43.628)
地域（市・郡）	8.1823 ***	8.7183 ***	8.1163 ***
	(38.129)	(32.609)	(37.931)
経営組織（株式・有限会社）	4.3065 ***	3.7643 ***	3.7897 ***
	(34.379)	(28.066)	(27.880)
初期対数尤度	-52430	-52430	-52430
対数尤度	-27701	-22358	-19070
疑似決定係数	0.4717	0.5736	0.6363
自由度調整済疑似決定係数	0.4715	0.5734	0.6361

*** p < 0.001, ** p < 0.01, * p < 0.05, - p < 0.01　( )は t 値

6. マッチングの正解率の比較

マッチングの正解率の観点から、提案手法と、従来の研究でよく用いられている最近隣法(Nearest Neighbor Method)との比較を行う。具体的には、以下の３つの統計的マッチングの手法について比較を行う。

(1) ウエイト付き絶対値距離の対数変換
(2) 最近隣法(Mahalanobis 距離)
(3) 最近隣法(Gower 距離)

Gower 距離及び Mahalanobis 距離の計算には R のパッケージ StatMatch を用いる(D'Orazio (2006))。なお、Mahalanobis 距離の計算に当たり、StatMatch の仕様により連続変数しか使用できないため、離散変数は用いずに距離を計算した。

マッチング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。この方法は、マッチング元(Donor)の各レコードから見て、マッチング確率の高い上位 R 件のマッチング先（Recipient）レコードの中に、正しい（同一の対象を表す）レコードが含まれる割合を算出するものである。以下では、その算出方法について示す。

帝国データバンクデータ(Donor)のテストデータの各レコード i $(i = 1,2,\cdots,M_{test})$に対して、経済センサスミクロデータ(Recipient)のテストデータで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、帝国データバンクデータのテストデータのレコード i に対して、経済センサスミクロデータ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、正しいマッチング先のレコードが上位 R 件の候補レコードに含まれているものの割合(マッチングの正解率)を表す$P(R)$は、以下の式(15)の形で表現できる。

$$P(R) = \frac{1}{M_{test}} \sum_{i=1}^{M_{test}} I\big(t \in C(i,R)\big) \qquad (15)$$

ここで $I\big(t \in C(i,R)\big)$ は、$t \in C(i,R)$ の場合に 1、それ以外の場合に 0 となる関数である。

マッチングの正解率 $P(R)$ を地域ごと比較したものが、図６～図８である。これらの結果を見ると、いずれの地域においても、多項ロジットモデルを用いた統計的マッチングの手法は、Gower 距離や Mahalanobis 距離に基づく最近隣法に基づく方法と比較して、大幅に正解率が高くなっている。特に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方法が、最も正解率が高くなっている。

なお、前述のとおり、Mahalanobis 距離の算出に当たっては、統計解析ソフトウェア R のパッケージ StatMatch で利用できる変数に関する制約から、連続変数(従業者数、売上高及び資本金額)のみを用いて距離を計算している。また、Gower 距離の算出に当たっては、そこで用いられるウエイトをテストデータのみから算出している。よって、結果の比較を行う際には、統計的マッチング手法の違いのほか、上記のように距離によって使用している情報量(変数)が異なる点についても留意する必要がある。

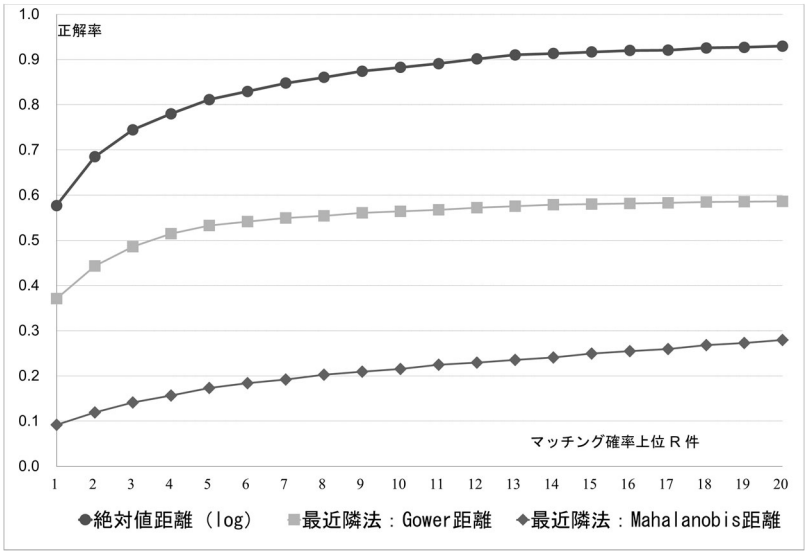

図６　正解率の比較（地域Ａ）

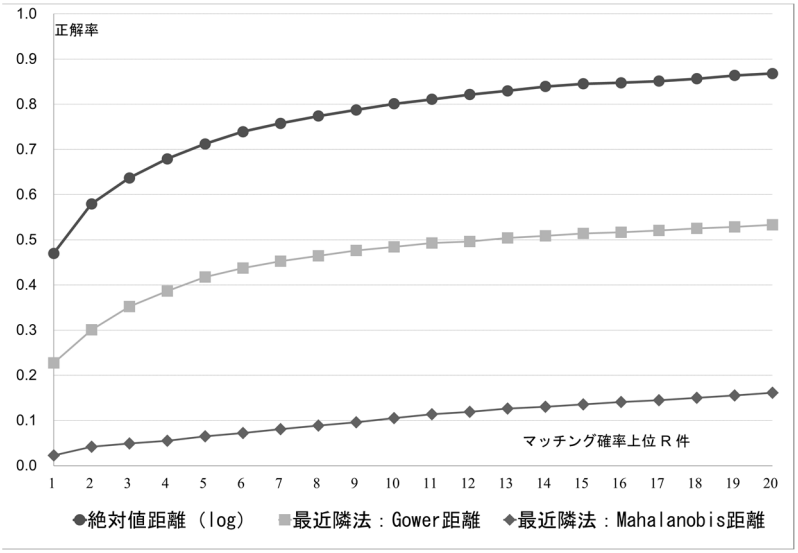

図７　正解率の比較（地域Ｂ）

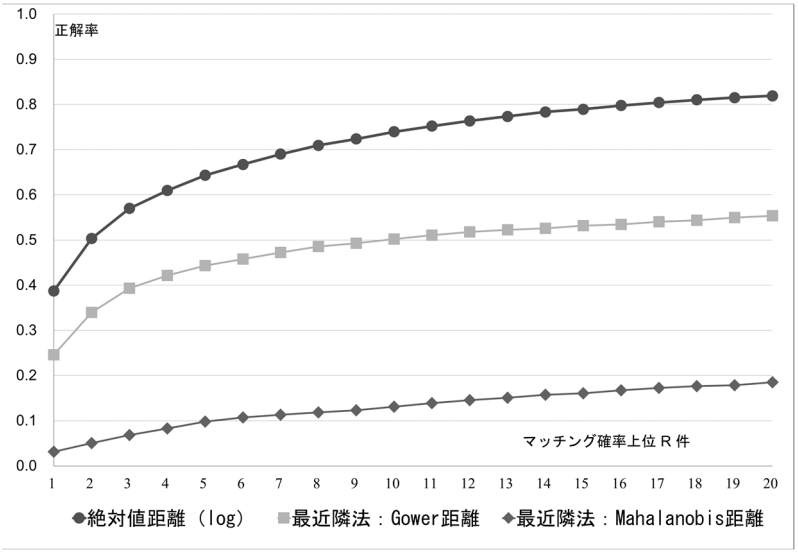

図８　正解率の比較（地域Ｃ）

7. 照合キーのランダムな欠測に対応する場合の試算

ここまでの分析では、照合キーが利用できない（テストデータとなる）企業がランダムに発生する場合を想定していた。しかし、特定の属性を持つ企業が照合キーを使うことができない場合も想定され、こうした状況が結果に影響を与える可能性も考えられる。そこで、こうした状況に対応する場合の試算も行う。

具体的には、小規模企業ほど情報が少なく、照合キーが欠測する（テストデータとなる）ことを想定して、経済センサスミクロデータにおいて資本金額が 1,000 万円以下かつ従業者数が 10 人以下の企業から、全体の 1/3 に当たるレコードをテストデータとして抽出し、それ以外の企業を学習用データとし、当該データに基づき、多項ロジットモデル（絶対値距離の対数を使用）を推定するとともに、それを用いて推定したマッチング確率を基に統計的マッチングを行い、マッチングの正解率を、前節の結果と比較する。新たに抽出した各データのサイズを示したものが、表７である。また、多項ロジットモデルの推定結果を示したものが、表８である。表８を見ると、多項ロジットモデルは適切に推定され、係数も前述の結果と大きくは変わっていないことがわかる。

表７　学習用データ及びテストデータのレコード数（再抽出）

		地域Ａ	地域Ｂ	地域Ｃ
(1)	学習用データ	13,417	14,819	18,314
	経済センサスミクロデータ	9,104	9,649	12,583
	帝国データバンクデータ	4,313	5,170	5,731
(2)	テストデータ	6,518	7,213	8,874
	経済センサスミクロデータ	4,553	4,825	6,292
	帝国データバンクデータ	1,965	2,388	2,582
合計（(1)＋(2)）		19,935	22,032	27,188

表８　多項ロジットモデルの推定結果（再計算）

	地域Ａ	地域Ｂ	地域Ｃ
従業員数	1.0856 ***	1.1279 ***	1.0800 ***
	(36.859)	(50.837)	(52.279)
資本金額	0.8288 ***	0.7980 ***	0.7755 ***
	(58.221)	(80.029)	(73.482)
売上高	0.9397 **	0.7767 ***	0.7148 ***
	(64.589)	(61.316)	(60.260)
産業	3.4873 ***	3.4491 ***	3.4857 ***
	(60.738)	(72.492)	(78.349)
開設年	1.3408 ***	1.2582 ***	1.3738 ***
	(31.035)	(35.388)	(40.390)
地域（市・郡）	9.1061 ***	9.2070 ***	8.0449 ***
	(22.210)	(20.528)	(38.609)
経営組織（株式・有限会社）	4.3003 ***	3.9420 ***	3.8320 ***
	(19.733)	(28.337)	(28.232)
初期対数尤度	-39319	-47433	-54101
対数尤度	-8570	-14398	-18381
疑似決定係数	0.7820	0.6965	0.6602
自由度調整済疑似決定係数	0.7819	0.6963	0.6601

マッチングの正解率について、前節までの結果と比較を行う形で示したものが、以下の図９～11 である。照合キーが欠測する企業と、全体の企業との間に差がある場合には、モデルの推定結果やマッチングの精度に影響があることがわかる。

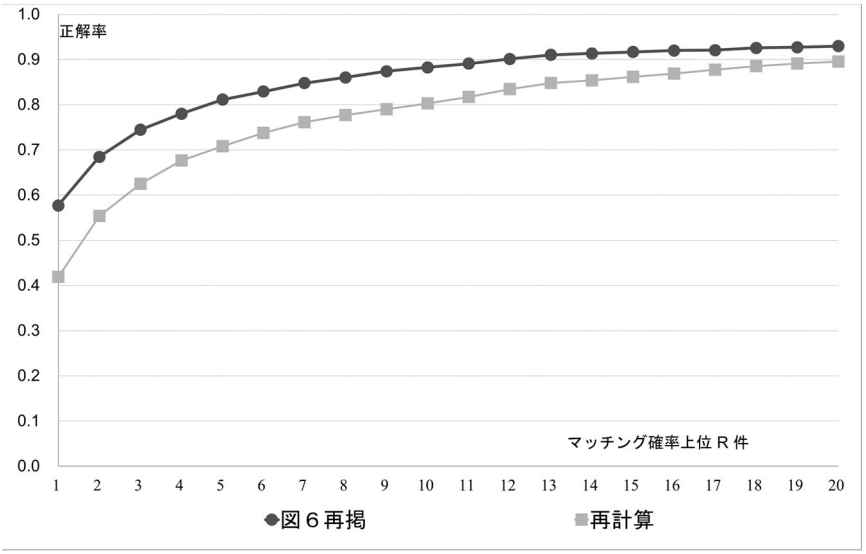

図９　正解率の比較（地域Ａ）再計算

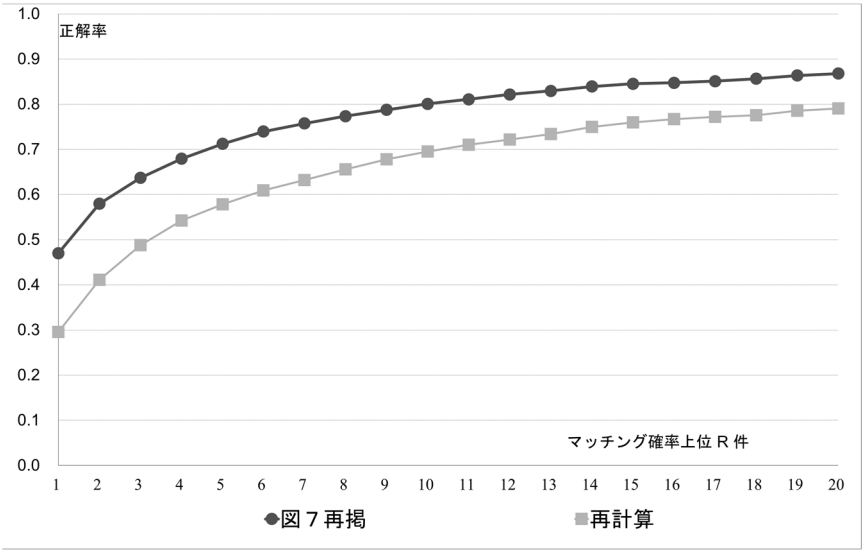

図 10　正解率の比較（地域Ｂ）再計算

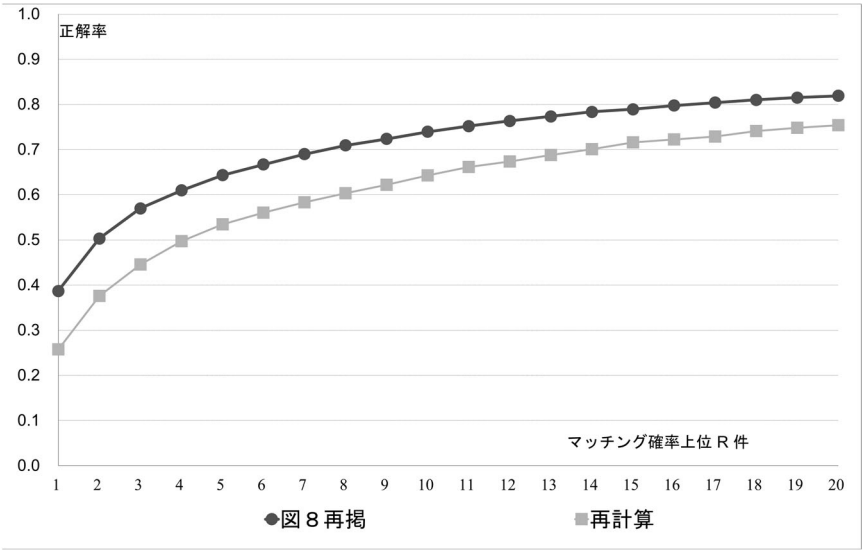

図 11　正解率の比較（地域Ｃ）再計算

8. 提案手法と従来の手法との比較

加重平均の式(12)（単純平均）及び式(13)（幾何平均）において、加重平均のウエイト ω を 0 から 1 まで 0.01 ずつ動かし、対応する Q_{ij} を用いて統計的マッチングを行い、式(15)で $R=1$ とした正確率 $P(1)$ を算出する。ω と $P(1)$ との関係を地域ごとに示したものが、以下の図 12～図 14 である。正解率が最大のウエイトに縦線を描いている。

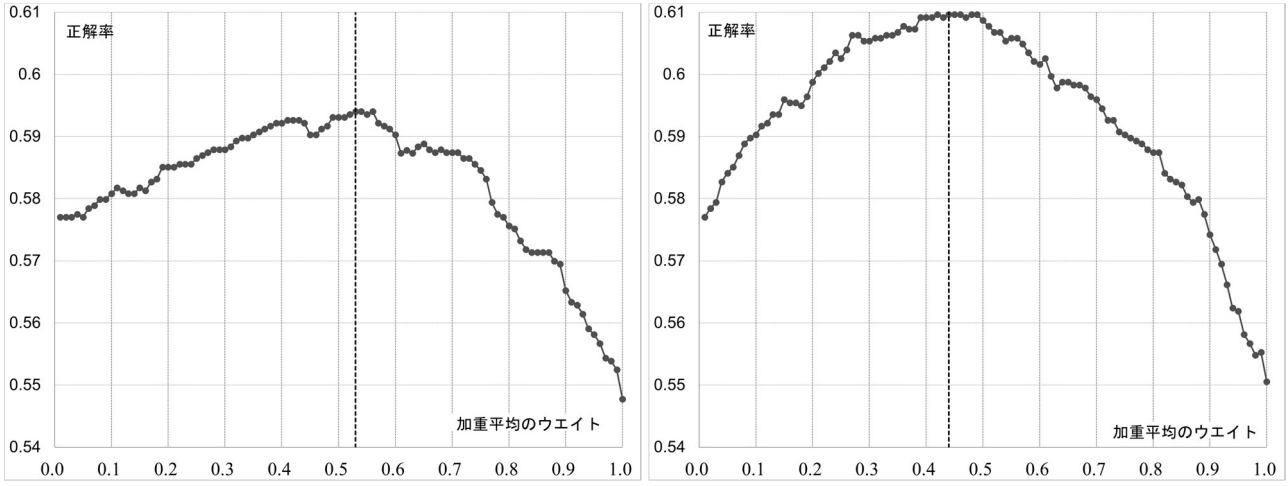

図 12　加重平均のウエイトと正解率の関係（地域Ａ）左：式(12)、右：式(13)

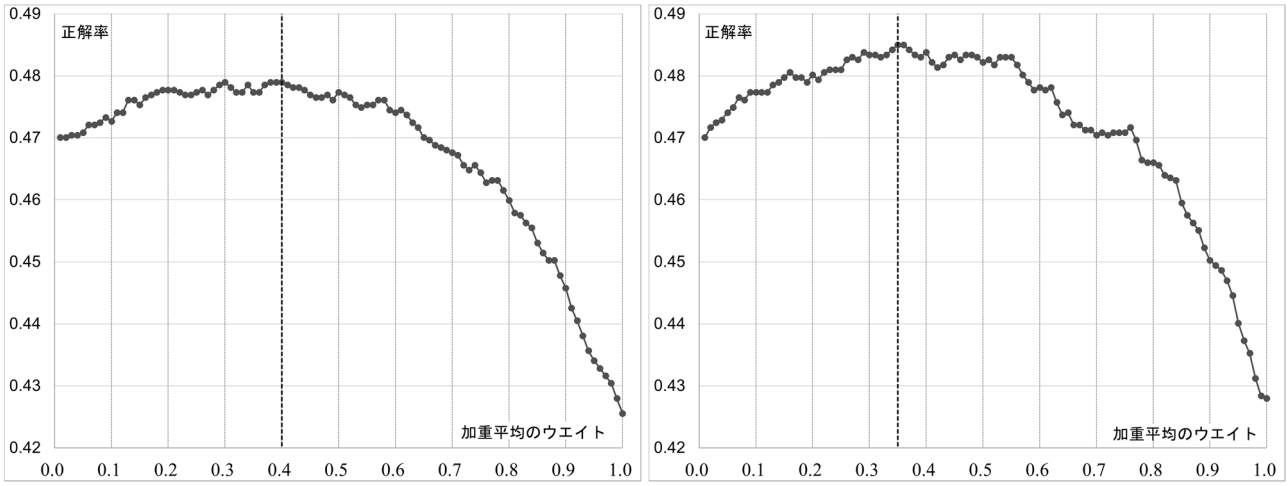

図 13　加重平均のウエイトと正解率の関係（地域Ｂ）左：式(12)、右：式(13)

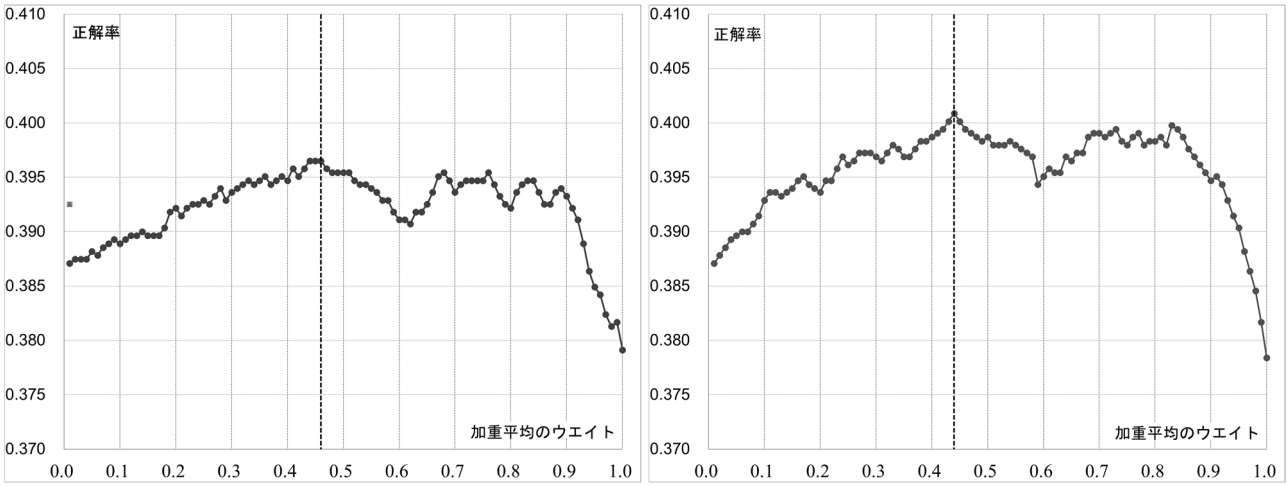

図 14　加重平均のウエイトと正解率の関係（地域Ｃ）左：式(12)、右：式(13)

各地域において、式(12)及び式(13)を用いた場合の正解率の最大値と、対応する加重平均のウエイトを示したものが、表９である。

表９　各変数の正解率向上の程度

		地域Ａ	地域Ｂ	地域Ｃ
通常モデル(絶対値距離(log))	正解率	0.5770	0.4701	0.3871
Donor-Recipient 転置処理（式(12)）	正解率	0.5940	0.4790	0.3965
	最適ウエイト	0.53	0.40	0.46
Donor-Recipient 転置処理（式(13)）	正解率	0.6096	0.4850	0.4009
	最適ウエイト	0.44	0.35	0.44

いずれの地域の結果においても、ω が 0 と 1 の中間の値においてピークがあり、マッチング確率の加重平均を用いることで、正解率が向上することが示されている。また、いずれの地域においても、式(13)の幾何平均を用いた方が、ピークの部分での正解率が高くなっている。以上の結果から、式(13)の幾何平均を用いて、ω を 0.4 から 0.5 程度に設定することにより、正解率が向上することが示された。

このように、単純な平均（$\omega = 0.5$）を用いるよりは、データから探索されたウエイトを用いる方が、マッチングの精度が向上する。ただし、最適なウエイトが不明な場合であっても、単純な平均（$\omega = 0.5$）を用いることにより、少なくとも、基の方法よりはマッチングの精度が向上していることがわかる。

9. おわりに

本研究では、Takabe and Yamashita(2020)及び高部・山下(2018)で提案された多項ロジットモデルに基づく統計的マッチングの手法をさらに発展させ、通常のモデル及び Recipient と Donor の転置処理を行ったモデルによるマッチング確率の加重平均を距離として、ウエイト付き距離に基づく統計的最適マッチングの手法を提案した。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、マッチングの正解率の観点から、従来の手法よりも優れていることが示された。

今後の課題として、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。

本研究では、統計的マッチング及び変数選択に関する手法の提案に焦点を当てているが、量（変数）が増加したデータを用いることの有用性を示していくこともまた、今後の重要な課題であると考える。その場合、正解が不明な中で、マッチングの有効性をどのように考えていくかが課題となる。これについては、例えば、マッチングを行う前のデータと、マッチングを行って変数が増加したデータで、回帰分析などの各種の計量分析を行い、分析結果の精度がどの程度改善するかを見ることにより、データの量を増加させたことに対する有効性の判断を行うことも、ひとつの方法であると考える。

また、本研究では、利用できる変数が少ない場合を想定して分析を行っているが、マッチングに用いる多項ロジットモデルにおいて、多くの変数が利用できる場合には、変数の種類や数が、その後のマッチングの精度に影響を与える可能性があることから、マッチングの精度を考慮した適切な変数群の選択方法を検討していくことも、今後の課題である。

本研究における手法は、企業データだけでなく、世帯・個人などが対象のデータにも適用することが可能であることから、企業データ以外の様々なデータに対しても本研究における手法を適用することにより、提案手法の有効性を確認していくことも必要であると考える。

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、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、今後も、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。

謝辞

本稿について丁寧な査読をしていただき、多くの改善点の指摘及び有益なコメントをしていただいた匿名の２名の査読者に対し、深く感謝を申し上げたい。本研究は科研費（16H02013 及び 15H03390）の助成を受けている。本研究で使用した平成 24 年経済センサス - 活動調査のミクロデータについては、統計法に基づく調査票情報の二次的利用の制度により提供を受け、総務省統計データ利活用センター（和歌山県）のオンサイト施設において独自集計を行ったものである。同データの提供に当たり、関係者の方々に多くの面で御支援いただいことに感謝を申し上げる。

参考文献

[1] 伊藤伸介(2018)，公的統計ミクロデータの利活用における匿名化措置のあり方について，日本統計学会誌 ，47, 77-101.

[2] 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 ，1-15.

[3] 今野浩(1978)，非線形計画法，日科技連.

[4] 高部勲 ，山下智志(2018)，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 ，115, 1-16.

[5] 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 ，110, 1－17.

[6] 森博美(2008a)，我が国における統計法制度の展開（21世紀の統計科学Ⅰ：社会・経済の統計科学，国友直人，山本拓 監修・編），121-145，東京大学出版会.

[7] 森博美(2008b)，情報資産としての統計と政府統計データアーカイブ，統計学，94, 15-25.

[8] 山口幸三(2014)，失われし20 年における世帯変動と就業異動：1991 年～2010 年のミクロ統計データの静態・動態リンケージにもとづく分析 ，日本統計協会.

[9] 山下智志(2005)，公共事業モデルのヴァリデーション（モデルヴァリデーション ，北川源四郎 ，岸野洋久 ，樋口知之 ，山下智志 ，川崎能典 著） ，155−180，共立出版.

[10] 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 ，56, 43-71.

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

[12] Buuren, S. V., and Groothuis-Oudshoorn, K. (2010), mice: Multivariate imputation by chained equations in R, Journal of statistical software, 45, 1-68.

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

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

[15] Fellegi, I. P. and Sunter, A. B. (1969), A theory for record linkage, Journal of the American Statistical Association, 64, 1183-1210.
[16] Gower, J. C. (1971), A general coefficient of similarity and some of its properties, Biometrics, 27, 623－637.
[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] Newcombe, H. B., Kennedy, J. M., Axford, S. J. and James, A. P. (1959), Automatic Linkage of Vital Records, Science, 130, 954-959.
[21] Rubin, D. B. (1986), Statistical matching using file concatenation with adjusted weights and multiple imputations, Journal of Business and Economic Statistics,4, 87-94.
[22] Rässler, S. (2002), Statistical Matching, Springer.
[23] Stuart, E. A. (2010), Matching methods for causal inference: A review and a look forward, Statistical science, 25, 1-21.
[24] Takabe, I. and Yamashita, S. (2020), New Statistical Matching Methods Using Multinomial Logistic Regression Model. (In Tadashi, I., Okada, A., Miyamoto, S., Sakaori, F., Yamamoto, Y. and Vichi, M. (Eds.), Advanced Studies in Classification and Data Science , 265-274, 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, 1405-1413.