Discussion Paper Series A No.595

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――

栗原由紀子
（弘前大学）

2013年10月

Institute of Economic Research
Hitotsubashi University
Kunitachi, Tokyo, 186-8603 Japan

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 ―

栗原由紀子
(弘前大学)

2013 年 10 月

要　旨

本稿は，法人企業統計調査（財務省）の調査票情報を対象として，統計的マッチングによるパネルデータの作成可能性とその有効性を検証するために，統計的マッチング手法の比較とともに，精度の高いマッチング推定量（相関係数）を得るためのキー変数選択の条件を吟味し，その効果を捕捉した。結果は以下の三点に整理できる。まず，条件付き独立性（CIA）が成立するようなマッチングに適した条件において，マハラノビス法と回帰補定法（RIEPS および NIBAS）を比較したところ，回帰補定法がよりバイアスの小さい推定値を与えている。次に，回帰補定法で適切な推定量を得るための条件としては，キー変数を単純に増やしていくよりも，CIA がほぼ成立し，可能な限り目標変数 X および Y，それぞれとの相関がともに強いキー変数を用意する方が効果的である。最後に，ある程度相関の強いキー変数のもとでは，回帰補定法の多重代入法による 95%信頼区間において，その信頼区間に含まれる真値の割合を示すカバレッジ指標が約 90%以上という高いパフォーマンスを示しており，マッチングによる不確実性が多重代入法によりかなりの程度捉えられていることを確認した。

Key words: RIEPS, NIBAS, Mahalanobis distance, Multiple imputation, Sampling experiment

* 連絡先　栗原由紀子　E-mail: yukuri@cc.hirosaki-u.ac.jp

本研究は，「一橋大学経済研究所 共同利用共同研究拠点事業プロジェクト研究;景気変動を踏まえた就業行動と企業の生産性および賃金構造の動態変化に関する計量分析」（研究代表者：中央大学　坂田幸繁，平成 25 年度）の成果の一部である。また，本研究は，財務省から「法人企業統計調査 1983 年 4-7 月期-2011 年 10-12 月期」の調査票情報の提供を受け，個票データに基づいて分析を行っている。記して関係諸機関への謝辞とします。

1. はじめに

統計的マッチングは，異なるデータソースを個体ベースで接合することで，情報資源の統合的活用を可能にするとともに，新たな分析枠組みを提供するものである。しかしながら，異なる標本から構成される２つのデータセットに対して，両者にたまたま共通して存在する変数を接着剤代わりに接合するため，マッチング・データに基づく統計量に関してはマッチング誤差が極めて大きな問題となる。

特に，統計的マッチングの精度は，マッチング手法やデータがおかれる条件の違いによって影響を受けるにもかかわらず，実際にマッチングの適用が必要な場面においては，当然，真値や完全データによる推定値は不明であるから，マッチング・データから算出された推定値が利用可能な精度を保持しているか否かの判断は難しい。そのため，統計的マッチングの利用可能性については，対象となるデータセットがおかれる条件を考慮しながら，シミュレーションを含めて完全データによる推定値（あるいは真値）が入手できるような特殊な状況をうまく利用して，推定値の分布や特性を詳細に吟味・検討し，その成果を現実の場に敷衍するというやり方が有力な方法のひとつとなる。本稿のアプローチもそのような方向に沿っている。

統計的マッチングの先行研究として,主要な研究成果のひとつにRässler(2002) が挙げられる。パラメトリック・モデルによる統計的マッチングの手法を比較したものであり，消費支出に関するデータとテレビの視聴時間データとの接合を行うことで，マッチング手法の精度比較を行っている。日本においては，荒木・美添（2007）が，家計調査と貯蓄動向調査（総務省統計局）に関して統計的マッチングによる接合を行い，ノンパラメトリック手法である最近隣距離法の制約付きモデルと制約無しモデルによる結果の相違が検討されている。また，栗原（2012b）では，ノンパラメトリック手法のマハラノビス距離関数を用いて中小企業景況調査（中小企業整備基盤機構）の疑似パネルデータを作成し，景況調査のパネル分析を試みている。これに対して，坂田・栗原（2013）では，ノンパラメトリック手法およびパラメトリック手法を，法人企業統計調査（財務省）の調査票情報に適用し，マッチング・データから得られる統計量のバイアスや MSE を比較することで，有効な推定量を得るためのマッチング手法を検証している。

本研究は，法人企業統計調査（財務省）の調査票情報（以下では，法企データとも呼称する）に対して，統計的マッチングの手法を適用し，キー変数の選択条件の良し悪しの吟味を主要課題として，パネルデータの作製という観点からマッチング・データによる推定量の精度比較を試みるものである。法企データに関しては，資本金規模 10 億円以上の大企業に限定すれば全数調査が行われており，原理的には識別子によりパネル化できる。他方で，中小・中堅企業についてみると，それらは確率的に抽出されたサンプルであることから，識別子が利用できたとしても年度をまたがる（１年を超える）パネル化は困難である[1]。見かけ上，同一調査の接合なのですべてが共通変数と思われがちであるが，標本も異なり観測時点も異なるのではキー変数の役割を果たさない。そのため，時間的に一定，もしくは変動が少ないと想定される調査項目の異時点データをキー変数に用いるしかないが，作成されたデータセットの有効性という点では疑義が残る。

[1] 法人企業統計調査四半期調査では，1 年間は固定標本であるから，識別子（あるいは企業名，住所などの照合）により年度内については完全照合による接合は可能である。

しかしながら，法企データの一部項目については，当期の実績値に加え前期実績値も同時に記入されており，統計的マッチングにおいて問題となるキー変数の時点間のズレに関しては，これらの調査項目を利用すれば理論的には解消できる。いわば，統計的マッチングには比較的有利なデータセットの条件を法企データは有している。そこで，このような特性を逆手に取って，本稿では，法人企業統計調査から作製した疑似パネルデータ分析の有効性を高めるために，真値が把握可能な範囲において，異なる標本間のパネル接合による推定値の特性を精査し，統計的マッチング手法の選択と推定バイアスの関係，およびマッチングに使用するキー変数の選択条件を明らかにすることにしたい。

2. 検証内容

まずは，統計的マッチングの基本概念を整理しておこう。分析目標は変数 X と変数 Y（X,Y を目標変数と呼ぶ）の相関係数の推定に限定する。しかし X と Y は同時に観察されておらず，２つのデータセット A および B に分離されて観察されているものとする。A および B にはマッチングのために利用可能なキー変数 Z が含まれていることにして，A,B それぞれのデータセットの内容を Data A:[X,Z]，Data B:[Y,Z]と表すことにする。統計的マッチングでは，このようなデータセット A および B から共通のキー変数 Z を利用して，拡張データセット[X,Y,Z]を作成しようとするものである[2]。なお，マッチングのベースとするデータを recipient，変数情報を提供し接合される側のデータを donor と呼ぶことから，以下ではデータセット A に recipient ファイル，B に donor ファイルの役割を割り当てている。統計的マッチングの精度は，採用するマッチング手法，条件付き独立性の仮定の成否，目標変数とキー変数との相関特性に規定される[3]。以下に，それらの理論的要点を整理しておく。

2.1 マッチング手法

統計的マッチング手法は，ノンパラメトリック法とパラメトリック法の 2 つに大別できる。前者は，距離関数を定義して，キー変数に関して最も距離が近い個体同士を接合するものである。これに対して，後者は，キー変数と目標変数の間に統計モデルを想定し，その推定値や予測値を利用して理論分布のパラメータを求め，その分布形から確率的に発生させた値を補定値とする。

（a）マハラノビス法

ノンパラメトリック手法の一つであるマハラノビス法は，マハラノビス距離関数（Mahalanobis Distance；以下 MHL と略称）にキー変数を適用して，各要素の距離を測定し，最も距離が最小となる要素同士を接合するものである。

データ A に属するi番目の要素のキー変数ベクトルを$\boldsymbol{z}_i^A$，データ B に属するj番目の要素のキ

---

2 データ A と B に同一の標本が含まれ，かつキー変数 Z として個体識別子（ID）が付与されている場合には完全マッチングが可能となる。

3 統計的マッチングの手法やデータセットの条件などに関する詳細は,Rässler(2002),pp.15-43 および D’Orazio et. al.(2006),pp.13-64 を参照。

一変数ベクトルをz_j^B，またAとBをマージしたキー変数の分散共分散行列をΣ_{ZZ}としたとき，これら任意の要素間の距離は以下のように定義でき，マッチングの際には，この距離が最小となるような要素同士を接合することになる。

$$d_{AB} = \left(\boldsymbol{z}_i^A - \boldsymbol{z}_j^B\right)^T \Sigma_{zz}^{-1} \left(\boldsymbol{z}_i^A - \boldsymbol{z}_j^B\right) \quad (1)$$

(b) 回帰補定法

回帰補定法は欠損値処理のために開発されたものであり，データセットに多変量正規分布を仮定して，そのパラメータ（$\mu_{X|ZY}$，$\mu_{X|ZY}$，$\Sigma_{X|ZY}$，$\Sigma_{Y|ZX}$）を回帰モデルなどにより求めたうえで，目標変数への補定値を確率的に発生させるものである。

$$X|y,\beta,\Sigma \sim N\left(\mu_{X|ZY}\,;\,\Sigma_{X|ZY}\right) \quad (2)$$

$$Y|x,\beta,\Sigma \sim N\left(\mu_{Y|ZX}\,;\,\Sigma_{Y|ZX}\right) \quad (3)$$

多変量正規分布のパラメータの求め方は 2 つあり，一つは頻度論ベースで展開した RIEPS (Regression Imputation with Random Residuals)，もう一つはベイジアンベースによる NIBAS (Non-iterative Bayesian-based Imputation) である[4]。

X，Y，Z は，それぞれq個，p個，h個の変数から構成されるベクトルとする。$\hat{\beta}_{XZ}$および$\hat{\beta}_{YZ}$は，XをZ_Bに回帰したときのパラメータと，YをZ_Aに回帰したときのパラメータをそれぞれ示す。また，$\hat{\Sigma}_{XX|Z}$および $\hat{\Sigma}_{YY|Z}$は，Z の条件つきXの分散共分散行列，およびZ の条件つき Y の分散共分散行列をそれぞれ示す。なお，Z の条件付きXと Yの相関係数と，Z の条件付きXの分散およびZ の条件付きYの分散から計算される$\hat{\Sigma}_{YX|Z}$（または$\hat{\Sigma}_{XY|Z}$）も利用する。ただし，Z の条件付きXと Yの相関係数は，完全データでのみ観測可能な値であり，マッチングの際には不明である。したがって，ここには補助情報から計測した数値を適用することになるが，補助情報がない（無情報の）場合にはゼロを投入せざるをえないことになる。本稿では，補助情報がないケースを想定している。

このような条件のもとで，RIEPS では，回帰により求めた$\hat{\beta}_{XZ}$，$\hat{\beta}_{YZ}$，各データセットから計算される$\hat{\Sigma}_{XX|Z}$，$\hat{\Sigma}_{YY|Z}$，および補助情報から得られる$\hat{\Sigma}_{YX|Z}$を用いて，回帰補定による理論値 $\mu_{X|ZY}$ および $\mu_{Y|ZX}$ を算出し，これを正規分布の平均に関するパラメータとする。さらに，この理論値と実現値の誤差を $\hat{\Sigma}_{X|ZY}$ と $\hat{\Sigma}_{Y|ZX}$ で計測し，これを正規分布の分散に関するパラメータに利用する。なお，$\otimes$ はクロネッカー積を示し，I_{na}，I_{nb}はA，Bそれぞれのサンプルサイズn_a，n_bを次数とする単位行列である。

$$\mu_{X|ZY} = Z_A\hat{\beta}_{XZ} + \left(Y - Z_A\hat{\beta}_{YZ}\right)\hat{\Sigma}_{YY|Z}^{-1}\hat{\Sigma}_{YX|Z} \quad (4)$$

$$\mu_{Y|ZX} = Z_B\hat{\beta}_{YZ} + \left(X - Z_B\hat{\beta}_{XZ}\right)\hat{\Sigma}_{XX|Z}^{-1}\hat{\Sigma}_{XY|Z} \quad (5)$$

$$\Sigma_{X|ZY} = \left(X - \mu_{X|ZY}\right)'\left(X - \mu_{X|ZY}\right)/\upsilon_A \otimes I_{na}\,, \qquad \upsilon_A = n_a - (k+p) \quad (6)$$

$$\Sigma_{Y|ZX} = \left(Y - \mu_{Y|ZX}\right)'\left(Y - \mu_{Y|ZX}\right)/\upsilon_B \otimes I_{nb}\,, \qquad \upsilon_B = n_b - (k+q) \quad (7)$$

[4] RIEPS および NIBAS の理論的詳細は Rässler (2002) p.96-107 を参照のこと。

これに対して NIBAS は，パラメータを構成するβ_{XZ}とβ_{YZ}，および分散共分散行列$\Sigma_{XX|Z}$と$\Sigma_{YY|Z}$を確率分布から発生させて適用する。

$$\mu_{X|ZY} = Z_A\beta_{XZ} + (Y - Z_A\beta_{YZ})\Sigma_{YY|Z}^{-1}\hat{\Sigma}_{YX|Z} \tag{8}$$

$$\mu_{Y|ZX} = Z_B\beta_{YZ} + (X - Z_B\beta_{XZ})\Sigma_{XX|Z}^{-1}\hat{\Sigma}_{XY|Z} \tag{9}$$

$$\Sigma_{X|ZY} = \left(\Sigma_{XX|Z} - \hat{\Sigma}_{XY|Z}\Sigma_{YY|Z}^{-1}\hat{\Sigma}_{YX|Z}\right)\otimes I_{na} \tag{10}$$

$$\Sigma_{Y|ZX} = \left(\Sigma_{YY|Z} - \hat{\Sigma}_{YX|Z}\Sigma_{XX|Z}^{-1}\hat{\Sigma}_{XY|Z}\right)\otimes I_{nb} \tag{11}$$

実際には，$\hat{\Sigma}_{XX|Z}$ および $\hat{\Sigma}_{YY|Z}$ の代わりの分散共分散行列パラメータとして，逆ウィシャート分布から $\Sigma_{XX|Z}$ および $\Sigma_{YY|Z}$ を発生させ，平均パラメータを$\hat{\beta}_{XZ}$および$\hat{\beta}_{YZ}$とする正規分布を仮定して，各マッチング回毎にβ_{XZ}およびβ_{YZ}を発生させる。

$$\Sigma_{XX|Z}\,|x \sim W^{-1}\left(v_A + 1\,;\,\left(X - Z_A\hat{\beta}_{XZ}\right)'\left(X - Z_A\hat{\beta}_{XZ}\right)\right) \tag{12}$$

$$\Sigma_{YY|Z}\,|y \sim W^{-1}\left(v_B + 1\,;\,\left(Y - Z_B\hat{\beta}_{YZ}\right)'\left(Y - Z_B\hat{\beta}_{YZ}\right)\right) \tag{13}$$

$$\beta_{XZ}|\Sigma_{XX|Z}, x \sim N\left(\hat{\beta}_{XZ}\,;\,\Sigma_{XX|Z}\otimes(Z_B'Z_B)^{-1}\right) \tag{14}$$

$$\beta_{YZ}|\Sigma_{YY|Z}, y \sim N\left(\hat{\beta}_{YZ}\,;\,\Sigma_{YY|Z}\otimes(Z_A'Z_A)^{-1}\right) \tag{15}$$

なお，回帰補定法では，多変量正規分布を仮定してランダムに補定値を発生させることになる。そのため，実際には現実のデータがとり得る範囲を超える可能性があり，本稿では法企データの最大値と最小値を閾値とした切断された多変量正規分布から発生させている。

(c) Multiple Imputation 法

ある特定の分布から確率的にパラメータや補定値を発生させる RIEPS や NIBAS を適用したとき，その補定値は変動し，同時に補定後のデータから得られる統計量も変動する。Multiple Imputation 法では，このような確率分布に基づいて発生させた変動を，統計的マッチングによりデータを作成することの不確実性を表すものと捉え，この不確実性まで含めて推定値の評価を行う。そのために，統計的マッチングを複数回実行し，各マッチング回毎に推定値を算出し，その推定値集合の平均値を統計的マッチングの推定値とする[5]。以下では，Multiple Imputation により得られた推定値を MI 値と略称する。

まず，統計的マッチングをM 回繰り返すものとする。そのうちの任意の試行回をm $(m = 1, \ldots, M)$ としたとき，得られる推定量（例えば平均，相関係数，回帰係数など）は$\hat{\theta}_m$とする。このとき，MI 値は $\hat{\theta}_1, \ldots, \hat{\theta}_M$の平均値として計測される。

$$\hat{\theta}^{MI} = \frac{\sum_{m=1}^{M}\hat{\theta}_m}{M} \tag{16}$$

5 これに対して 1 回限りの補定を Single Imputation と呼ぶ。

次に，MI 値の分散は，1 回の推定値に対する分散 W（Within Variance）と，推定値間のばらつき B（Between Variance）を複合的に考慮した値 T（Total Variance）で与えられる。Wは，M 回のマッチングから得られる推定値の分散 $\hat{V}(\theta_m)$ の平均値を，B はM 回分の推定値 $\hat{\theta}_m$の分散を意味している。

$$W = \frac{\sum_{m=1}^{M} \hat{V}(\hat{\theta}_m)}{M} \tag{17}$$

$$B = \frac{\sum_{m=1}^{M} (\hat{\theta}_m - \hat{\theta}^{MI})^2}{M-1} \tag{18}$$

$$T = \left(1 + \frac{1}{M}\right) B + W \tag{19}$$

MI 値については，その分散を Total Variance として，自由度 ν の t 分布に従うことが知られている。MI 値による推定や検定は，この性質を利用して行うことができる。

$$\frac{\hat{\theta}^{MI} - \theta}{\sqrt{T}} \sim t(\nu) \tag{20}$$

$$\nu = (M-1)\left[1 + \frac{W}{\left(1 + \frac{1}{M}\right) B}\right]^2 \tag{21}$$

なお，相関係数の MI 値は，相関係数の z 変換値（M回分）の期待値により算出している。また，この z 変換値が正規分布に近似する性質に基づいて推定量の分散 $\hat{V}(\hat{\theta}_m)$ を算出し，これを用いて Total Variance とともに信頼区間を計算している。そのため，本稿では，相関係数の z 変換値に関する MI 値や信頼区間を，さらに逆変換した値により分析結果を示している[6]。

2.2 条件付き独立性

Z をキー変数としてマッチングする場合，X と Y に関する Z の条件付き分布の独立性（CIA; Conditional Independence Assumption）が成立していることが前提となる。

$$f(X,Y|Z) = f(X|Z) f(Y|Z) \tag{22}$$

この条件の成否を捉えるには完全データが必要であるため，実際に統計的マッチングが必要とされる状況では観測不可能であるが，検証に際してはその成否の程度を確認しておかねばならない。そのために，目標変数に対してキー変数を説明変数として回帰した残差 ε_X および ε_Y の相関係数を，条件付き従属性（CID; Conditional Independence and Dependence Index）として定義し，これを基に CIA の成否を評価する[7]。CID がゼロに近ければ，CIA が成立していると判断する。

---

6 RIEPS および NIBAS による推定値の算出には，Rässler（2002）pp.214-221 の SPLUS コードを参考に，独自に作成した統計ソフト R のためのプログラムを用いている。

7 CIA に関する計測方法は，荒木・美添（2007）に提示されており，栗原（2012a）では相関係数と CID の理論的関係とともにモンテカルロ・シミュレーションによりその特性を検証している。

$$X = Z'\beta + \varepsilon_X,\ Y = Z'\beta + \varepsilon_Y \qquad (23)$$

2.3 目標変数とキー変数との相関

マッチング精度を高める条件のひとつとしてrecipient側の目標変数Xとキー変数Zとの相関，またはdonor側の目標変数Yとキー変数Zとの相関が強いことが求められる。当然，XとZおよびYとZの両方の相関が極めて強いことが理想的であるが，入手したデータセットがそのような都合のよい条件を満たすとは限らない。そこで，より現実的な場面を想定して，許容できる範囲の精度で推定量を得るには，XとZの相関またはYとZの相関のうち一方だけでも強ければよいのか[8]，あるいはやはり両方の相関がある程度強い必要があるのか，そのときその相関の強さはどの程度あればよいのか，といった実際的な問題への指針となるべく検証作業がセットされる必要がある。

3. 検証方法

3.1 検証用データセット

本稿では，法人企業統計調査（四半期調査）の2001年第1四半期と2000年第4四半期に関する調査票情報を用いて検証を進める。検証対象は，製造業・大企業で識別子によりパネル化が可能である$n = 622$社を利用して，2001年第1四半期の収益性指標の総資本経常利益率とその二期（半年）前の安全性指標である2000年第3四半期の自己資本比率との相関係数の算出を目標とする[9]。

マッチング検証用のデータセットは，表1に示すように，目標変数としてRecipientには総資本経常利益率（Y），Donorには前期（2000年Q3の）自己資本比率（X）を設定し，キー変数はそ

表1　検証用データセット

[Recipient Data A: 2001年 Q1]		[Donor Data B: 2000年 Q4]
X　missing		X　前期自己資本比率
Y　総資本経常利益率		Y　missing
Z1　前期流動比率	⟷	Z1　当期流動比率
Z2　前期自己資本比率	⟷	Z2　当期自己資本比率
Z3　人件費_従業員数（対数変換）		Z3　人件費_従業員数（対数変換）
Z4　前期資本_資本金（対数変換）	⟷	Z4　当期資本_資本金（対数変換）
Z5　損益_売上高（対数変換）		Z5　損益_売上高（対数変換）
Z6　損益_経常利益		Z6　損益_経常利益
Z7　前期総資本（対数変換）	⟷	Z7　当期総資本（対数変換）
Z8　人件費_従業員給与（対数変換）		Z8　人件費_従業員給与（対数変換）

8 栗原（2012a）では，シミュレーション結果から，XとYの少なくとも一方がキー変数と相関が強ければ，統計的マッチングは利用可能であることを示している。

9 法企データの場合，1ファイル内に前期と当期の値が与えられていることから，統計的マッチングによりパネル化をせずとも，一期前の値との相関係数は容易に求められる。

れぞれ Z1～Z8 とする[10]。ここで，Z1,Z2,Z4,Z7 については，同時点の情報をキー変数として利用することができる。ただし，標本が重複していれば，それらの同時点の情報はほぼ識別子の役割を果たす可能性があるが，本研究では重複標本がないケースを検討するために，同時点であっても Recipient と Donor で異なる要素を割り当てており，これら同時点の変数が識別子と同等の役割を果たすものではないことに注意が必要である。

3.2 データセットの特徴

表 2 および表 3 には，検証に使用するデータの基本統計量を示している[11]。基本統計量に関しては，その多くが，右に裾野が長い分布形状を示していることが想定される。パラメトリック手法を適用する際には，各変数の正規性の成立が不可欠であることから，これを Q-Q プロットにより確認すると，図 1a からは X,Y,Z1 を除いて，正規性を満たしていないことが分かる。対数変換により正規化を図ることは可能であるが，負の値を含む変数については対数変換ができないため，本稿では Z3,Z4,Z5,Z7,Z8 のみ対数変換を行い，正規化を図った（図 1b）。

相関行列の特徴としては（表 4），キー変数 Z は X（または Y）との相関が強いほどマッチング精度の改善が見込めるので，単純に比較すると，Z1，Z2，Z6 はよいキー変数であり，そのほかのキー変数はマッチングに有効な情報をあまり含んでいないようにみえる[12]。

表 2　基本統計量（recipient: Data A）

	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Y
下位 3%平均	33.0	-8.1	13.6	267.2	178.8	-218.5	931.6	17.2	-2.5
中央値	114.4	30.2	218.0	494.0	1693.0	15.0	7327.0	272.0	0.3
上位 3%平均	294.2	75.6	849.9	971.9	6499.6	382.3	25221.7	1012.2	3.9
平均値	123.5	32.3	258.0	554.8	2058.2	29.1	8291.5	312.9	0.4
標準偏差	53.7	18.8	185.2	190.7	1468.5	111.5	5429.5	223.9	1.3

（出所）著者により作成.

表 3　基本統計量（donor: Data B）

	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	X
下位 3%平均	33.1	-8.1	13.5	267.2	195.9	-233.9	933.6	15.3	-7.2
中央値	114.9	30.1	215.0	494.0	1838.0	39.5	7319.0	260.5	30.0
上位 3%平均	291.5	75.6	843.9	972.0	7400.4	512.3	25468.4	974.4	75.0
平均値	124.1	32.3	256.6	554.8	2284.3	62.8	8327.0	308.4	32.1
標準偏差	53.4	18.8	184.0	190.7	1670.2	138.4	5478.1	222.3	19.0

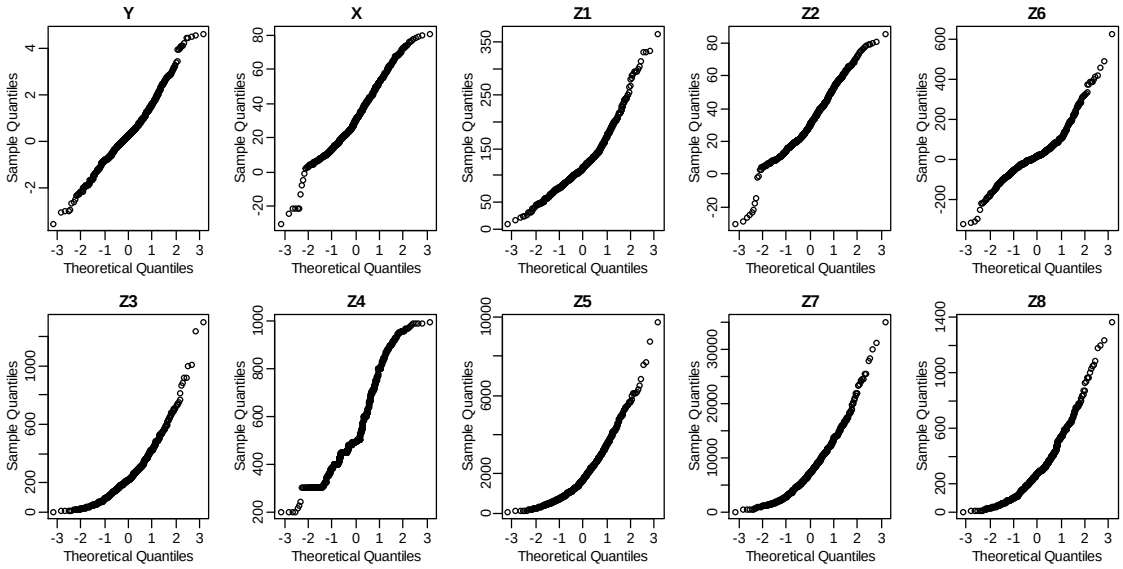
（出所）著者により作成.

10 統計的マッチングの基本は同時分布を捉えることにあるため，実際の分析に利用する変数が比率や合成値などの場合には，原データをマッチングした後に比率や合成値に変換するのではなく，変換後の値に対してマッチングを適用し，推定量を求めたほうが精度がよい。

11 検証用データセット（622 サンプル）からは，R パッケージ mvoutlier の Multivariate Outlier Method により外れ値となる要素を特定し，除外している。

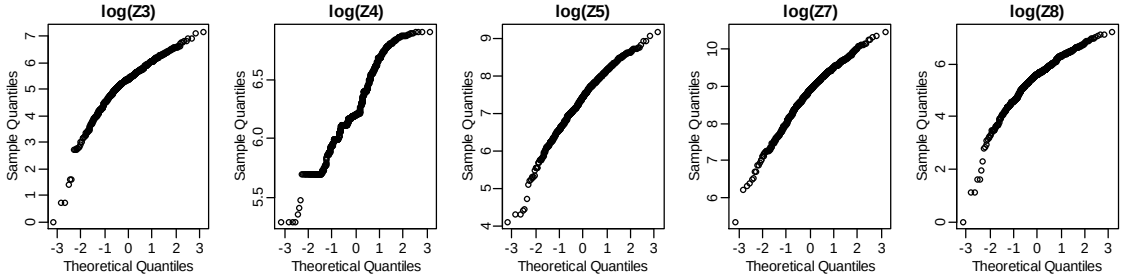
12 なお，完全データ（622 サンプル）による X と Y の相関係数は 0.21 であった。このことから，大企業・製造業（外れ値除外）サンプルに限れば，総資本経常利益率（Y）に対する相関は，1 期前の自己資本比率（Z2）であっても 2 期前の自己資本比率（X）であっても 0.21 と不変である。

図 1(a)　Q-Q プロット



(注) キー変数の Z1 から Z8 は, Data A の変数について分析したものであるが, Data B についても同様の傾向を示している。
(出所) 著者により作成.

図 1(b)　対数変換した変数の Q-Q プロット



(注) Data A の変数について分析したものであるが, Data B についても同様の傾向を示している。
(出所) 著者により作成.

表 4　相関行列

	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	X	Data B
Z1		0.65	0.01	0.04	-0.09	0.13	-0.03	0.03	0.65	Z1
Z2	0.65		-0.14	0.06	-0.25	0.06	-0.18	-0.12	0.98	Z2
Z3	0.01	-0.13		0.17	0.73	0.23	0.70	0.94	-0.12	Z3
Z4	0.04	0.06	0.17		0.21	0.13	0.28	0.17	0.04	Z4
Z5	-0.10	-0.23	0.71	0.18		0.33	0.83	0.77	-0.25	Z5
Z6	0.16	0.20	0.07	-0.02	0.22		0.31	0.27	0.04	Z6
Z7	-0.03	-0.18	0.70	0.29	0.81	0.15		0.72	-0.17	Z7
Z8	0.02	-0.11	0.95	0.17	0.74	0.12	0.73		-0.11	Z8
Y	0.17	0.21	-0.06	0.00	0.09	0.81	-0.01	-0.02		X
Data A	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8		

(注) 行列の下三角部分は DataA の相関行列, 上三角部分は DataB の相関行列をそれぞれ示す。また, Z3, Z4, Z5, Z7, Z8 は対数変換した値を用いている。
(出所) 著者により作成.

3.3 検証のプロセス

本稿では，以下のプロセスにより検証を進める（図 2)。

(1) まず，母集団として，識別子により完全マッチングが可能な検証用のデータセット A，B（各データのサンプルサイズはそれぞれ$n = 622$である）を用意し，ここから真値θ_aを算出する。

(2) 母集団からサンプルサイズ n_1（100 または 300）でランダムにサンプリングを行う。ただし，データ A と B からはそれぞれ異なる要素を抽出する。データ A のサンプリングデータには，X が含まれないためデータ A^{Xmis} とし，同様に，B からは Y が得られないためデータ B^{Ymis} とする。

(3) この二つのデータ A^{Xmis} および B^{Ymis} を統計的マッチングにより接合することで，[X,Y,Z]が揃ったデータセットを作成する。

(4) マッチングにより X が補定されたデータ A^{Xmis} から必要な統計量（相関係数）を算出する。この一回限りのマッチングから得られた推定結果は Single Imputation による推定値 $\hat{\theta}_a^{SI}$ となる。

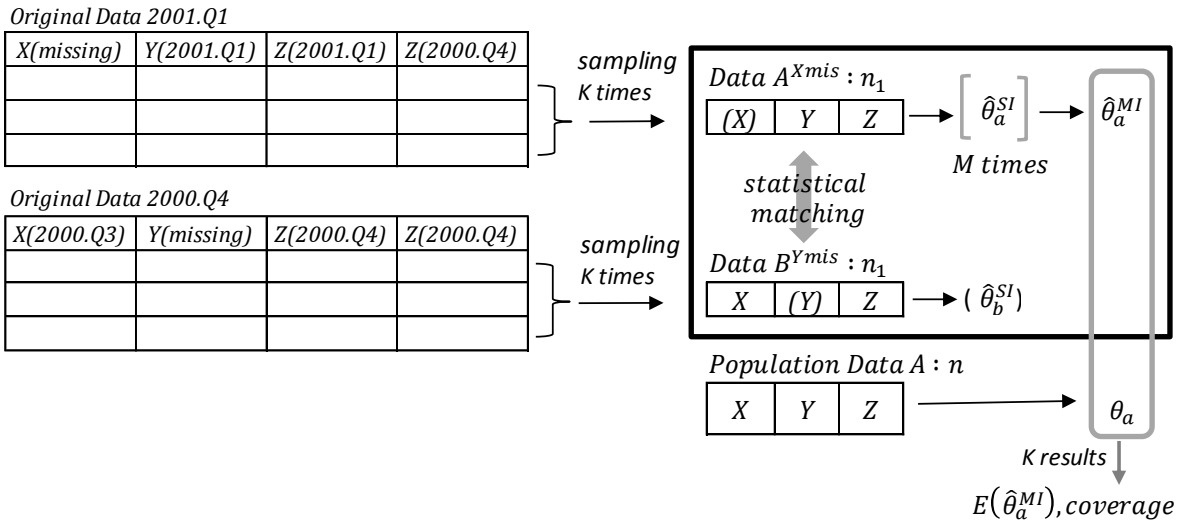
(5) (3) と (4) を$M = 30$回繰り返して得られる推定値の集合から，Multiple Imputation による推定値 $\hat{\theta}_a^{MI}$ およびその 95%信頼区間$[\underline{\theta}_{a,k}^{MI}, \overline{\theta}_{a,k}^{MI}]$を算出する。

(6) 標本の違いによる影響を考慮するために，(2) から (5) の作業を$K = 50$回繰り返し，$\hat{\theta}_a^{MI}$の期待値 $E(\hat{\theta}_a^{MI})$ およびカバレッジを算出する。

$$E(\hat{\theta}_a^{MI}) = \frac{\sum_{k=1}^{K} \hat{\theta}_{a,k}^{MI}}{k} \tag{24}$$

なお，カバレッジは$K = 50$回の試行のうち，95%信頼区間$[\underline{\theta}_{a,k}^{MI}, \overline{\theta}_{a,k}^{MI}]$に真値が含まれる割合を示す。

図 2　統計的マッチングの検証プロセス



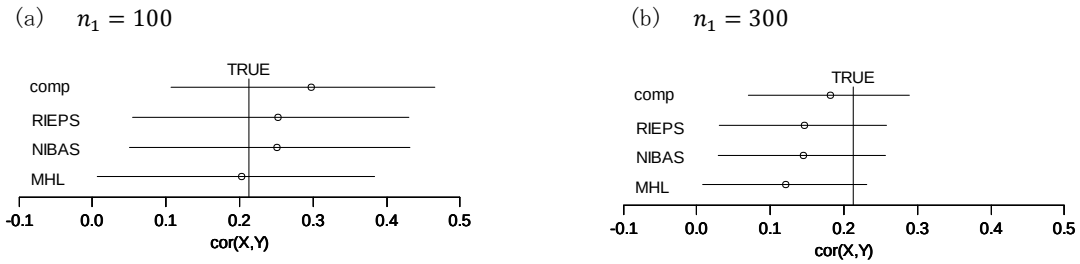
（出所）著者により作成.

4. 検証結果

4.1 統計的マッチング手法の比較

まずはマッチング手法による結果の違いを評価するために，Z1～Z8 の 8 個全てのキー変数を適用したケースから始めよう。図 3 には，完全データによる推定値と統計的マッチングの推定値が示されている（母集団からのサンプリングは 1 回のみである)。母集団要素をすべて使った相関係数の推定値 TRUE(以下では真値と呼ぶ)を基準としたとき,統計的マッチングによる推定値(RIEPS, NIBAS, MHL）は，完全データによる推定値（comp）と比べて大きな差異はなく，それぞれ真値の近傍に位置している。次に，95%信頼区間については，統計的マッチングによる不確実性がその標準誤差に含まれるため，完全データの区間幅よりも若干広めに示されている。また，完全データであれば，サンプルサイズが大きいとき信頼区間は狭まるが，(a）と（b）の比較を通して，統計的マッチングによる推定においても同様の特性を確認できる。

図 3　統計的マッチングによる相関係数と信頼区間



(注）(a）および（b）ともに，1 回限りのサンプリング（k=1）に基づいた相関係数の推定値およびその信頼区間を，○印と横線でそれぞれ示している。完全データのサンプリングデータ（サイズ 100 または 300）による推定値は comp として示している。なお，RIEPS および NIBAS については，Multiple Imputation によりその推定値を算出している。
(出所）著者により作成。

標本の相違による推定値の変動を考慮するため，母集団からのサンプリングを 50 回繰り返した場合の相関係数の期待値と，50 回のうちで 95%信頼区間に真値を含む比率を示したカバレッジを表 5 に整理している。まず，このデータセットでは CID が 0.03 前後と低い値を示していることから，マッチングに適したデータセットであることが分かる。相関係数の期待値については，RIEPS および NIBAS が若干低めではあるが，ほぼ真値に近い値を示しており，MHL はこれらパラメトリック・モデルより精度が悪く，下方にバイアスをもつ。また，カバレッジについては，RIEPS, NIBAS ともに完全データの値に近い数値を示している[13]。なお，完全データではあるがサンプルサイズが 50 の場合のカバレッジを算出したところ 94%であったことから，CID が低く統計的マッチングに適したデータセットの場合には，このようなサンプルサイズの小さい完全データから相関係数を算出するよりも，サンプルサイズは大きいが目標変数が 2 つのファイルに分かれたデータセットから，統計的マッチングにより推定値を算出した方が，より真値を捉えた結果が得られるものと

[13] 95%信頼区間のカバレッジが非常に高いのは，母集団サイズ 622 に対してサンプルサイズが 100（300）と非常に抽出率が高いためである。

考えられる。

MHL のカバレッジは，サンプルサイズの増大とともに減少している。本稿で適用したマハラノビス法による統計的マッチングでは，通常の標準誤差を用いており，当然，これには統計的マッチングの不確実性が反映されていない。そのため，サンプルサイズの増大とともに信頼区間が狭まり真値をカバーする比率が落ちていく結果となった。目標統計量が相関係数であり Z1～Z8 の全てのキー変数を使用する場合，バイアスの観点からも，また統計的マッチングの精度を適切に評価するうえでも，MHL より RIEPS または NIBAS が適していると言える。

表 5　統計的マッチングによる推定量の期待値およびカバレッジ

	n_1 = 100 CID = 0.029		n_1 = 300 CID = 0.038	
	E[cor(X,Y)]	Coverage	E[cor(X,Y)]	Coverage
COMP	0.213	98 %	0.211	100%
RIEPS	0.195	98 %	0.194	100%
NIBAS	0.192	98 %	0.194	100%
MHL	0.160	(92 %)	0.166	(90%)

（注） 母集団の完全データによる真の相関係数は$cor(X,Y) = 0.213$である。また，
$n_1 = 50$の COMP による値は，$E[cor(X,Y)] = 0.203, Coverage = 94\%$である。
（出所）著者により作成。

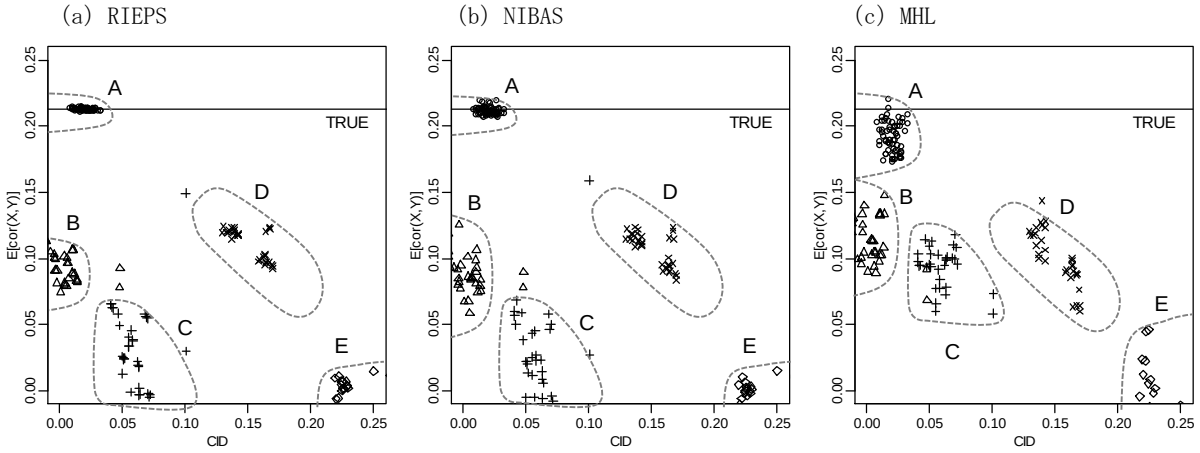
4.2　キー変数の選択とバイアス

統計的マッチングの精度を規定する条件付き独立性やキー変数と目標変数との相関は，キー変数に左右されることから，キー変数の数やその組み合わせがマッチング精度に与える影響を明らかにしたうえで，利用可能な精度でマッチング推定量が得られるキー変数の条件を特定しておく必要がある。そこで，キー変数 Z1～Z8 について，1 個のみをキー変数として利用した場合から，8 個全てを利用した場合まで，全ての組み合わせ（全 255 通り）についてマッチング実験を行った。

その結果を，マッチングにより得られた相関係数の MI 値の期待値（50 回試行）を縦軸，条件付き従属性 CID を横軸として，マッチング手法別に図 4 に示している。まず RIEPS と NIBAS については，NIBAS が若干推定値のブレが大きいが，両者とも極めて類似した傾向を示している。これらは，CID がゼロ付近であるときバイアスが小さく，CID の値が高い場合にはバイアスが大きくなる傾向がみてとれる。

しかしながら，A 群と B 群のように CID がゼロ付近（0.03 以下）にあっても，バイアスが小さい場合と大きい場合の 2 群に分かれるケースがある。さらに，CID が低い C 群よりも CID が高い D 群が，バイアスが若干小さいケースもある。すなわち，キー変数の組み合わせによって CID は異なるが，CID とバイアスは直線的な関係で捉えることはできず，統計的マッチングの精度と CIA の関係に関する理論的条件が示すように「CID がゼロ付近=バイアスが小さい」という関係が必ず

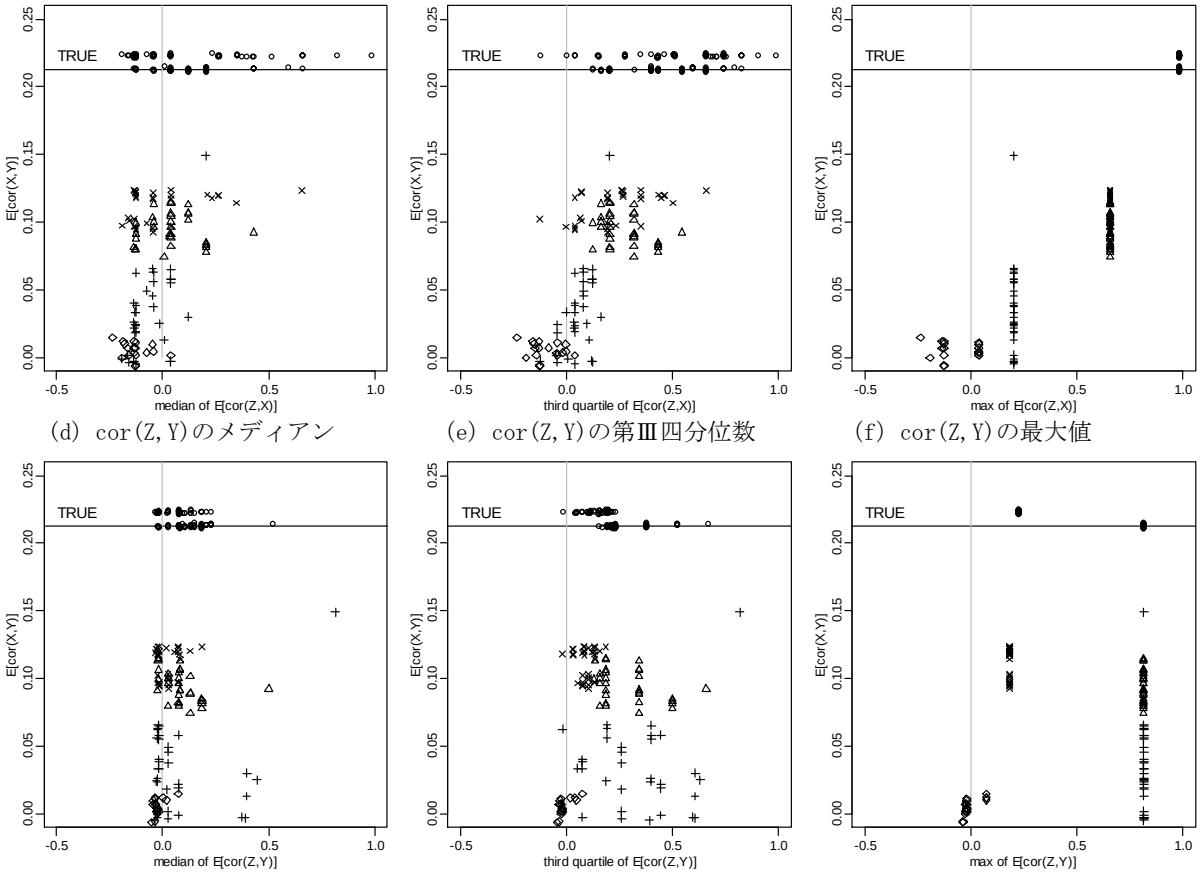
図 4　キー変数の組み合わせ別，CID の期待値と相関係数の期待値



（注）推定値の分布状況を基に 5 つに分類し，それぞれマークを割り当てている。目安として，「○」は CID<0.03，「△」は「○」の中でも E[cor(X,Y)]が 0.15 以下，「＋」は 0.03<CID<0.1，「×」は 0.1<CID<0.2，「◇」は CID>0.2 であるケースに相当する。

（出所）著者により作成。

図 5　キー変数の組み合わせ別，目標変数とキー変数の相関と相関係数の期待値（RIEPS）



（注）　複数のキー変数を利用している場合，複数の E[cor(Z,X)]が算出されるため，それらの特徴を捕捉するため，メディアン，第Ⅲ四分位数，最大値を用いた。キー変数が 1 つのみの場合には，すべて同じ値を利用している。なお，複数の E[cor(Z,X)]の算術平均は，メディアンと類似した傾向が示されている。

（出所）著者により作成。

しも成立していないことが分かる。

ただし，パラメトリック手法と比較して，マハラノビス法（MHL）は，A 群についてはバイアスが大きく，B 群，C 群についてはバイアスが小さい傾向が見られる。利用可能なキー変数セットが B 群や C 群に属するような場合には,バイアスの観点からは MHL の利用が適切である場合もある。

マッチングによる推定量のバイアスを，目標変数とキー変数の相関の強さとの関連から捉え直してみよう。なお，RIEPS と NIBAS は類似の傾向を示すことから，以下では主に RIPES について検討を加える。図 5 には，縦軸にマッチングによる X と Y の相関係数の期待値，横軸にキー変数と目標変数（X または Y）との相関係数の期待値（複数のキー変数を利用している場合には,その中での中央値，第Ⅲ四分位数，最大値）を示している。目標変数とキー変数の相関に関する中央値や第Ⅲ四分位数からは，特定の傾向は見られないが，最大値からは，明らかにキー変数と目標変数 X（または Y）との間に強い相関を持つ変数があれば,バイアスが小さくなることが示されている。すなわち，複数のキー変数の平均的な相関の強さよりも，目標変数との相関が極めて強いただ一つの変数の存在が，マッチング推定量の精度に作用していることが示唆される。

A 群から D 群のキー変数と目標変数との相関とマッチング精度の関係について,横軸に cor(Z,X)の期待値の最大値，縦軸に cor(Z,Y)の期待値の最大値を示した図 6 からみておこう。キー変数と目標変数の相関の特徴として，A 群は 2 つに分けられ，双方とも X については cor(Z2,X)=0.98 が最大となるようなキー変数の組み合わせであるが，Y については，一つは cor(Z2,Y)=0.21 が最大となり，他の一つは cor(Z6,Y)=0.81 が最大となるようなキー変数の組み合わせである。これらのうち cor(Z6,Y)の相関が強いことから，一つはキー変数セットに Z2 を含むが Z6 を含まないケース，もう一つは Z2 と Z6 を含むケースと整理できる。このうち，X，Y ともに最も強い相関を与えるキー変数[Z2,Z6]を含むマッチングがより精度のよい結果を与えている。

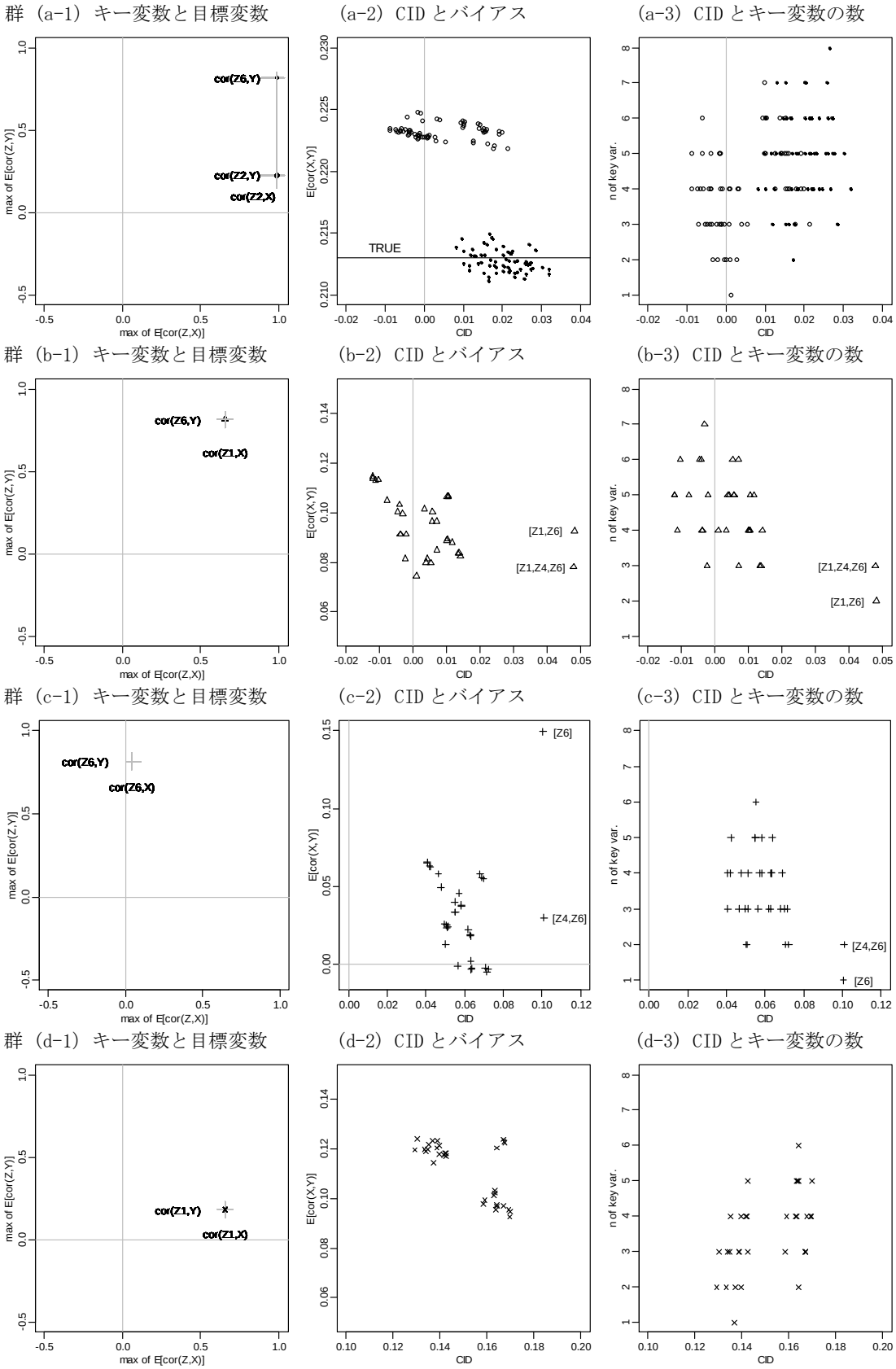
ここから，キー変数と目標変数との相関が，少なくとも 0.8 以上あれば，CID が低く比較的バイアスの小さいマッチング推定量が得られる。なお,CID とキー変数の数の関係を示した図 6 (a-3)を踏まえれば,CID の大きさはキー変数の数に左右されるものではなく，目標変数との相関が可能な限り強いキー変数を利用できるかどうかにより決まり，同時に目標とする統計量 cor(X,Y)の精度も高まることがわかる。

B 群は，cor(Z6,Y)=0.81 と cor(Z1,X)=0.65 が最大であるケースであることから，Z2 は含まないが，Z1 と Z6 を含むようなキー変数の組み合わせによるマッチングの結果である。B 群の中でも Z1,Z6（または Z4）のみをキー変数とする場合には，CID の値は若干大きくなる傾向にあるが，バイアスの大きさは B 群の傾向とそれほど変わらない。

C 群は Z1 と Z2 を含まないが Z6 を含むケースであり，D 群は Z2 と Z6 を含まないが Z1 を含むケースである。CID は C 群が低いにもかかわらず，バイアスは D 群が小さい理由は，キー変数 Z と X の相関の最大値が，D 群で大きいためと考えられる。すなわち，Z と Y の相関および Z と X の相関の両方が強いことが望ましいが，X を donor として補定する場合には，Y よりも X との相関が強いキー変数 Z を適用する方が，より精度の高い推定値の算出が期待できる。

以上の結果から，法企データの統計的マッチングにおけるパネル化により，バイアスの小さい相関係数を得るためには,CID がゼロ付近（0.03 以下）であり，キー変数と目標変数 X（または Y）

図 6　キー変数の組み合わせ別，目標変数とキー変数の相関とマッチング精度（RIEPS）



（注）図中の[ ]内には，マッチングに使用したキー変数を示している。
（出所）著者により作成。

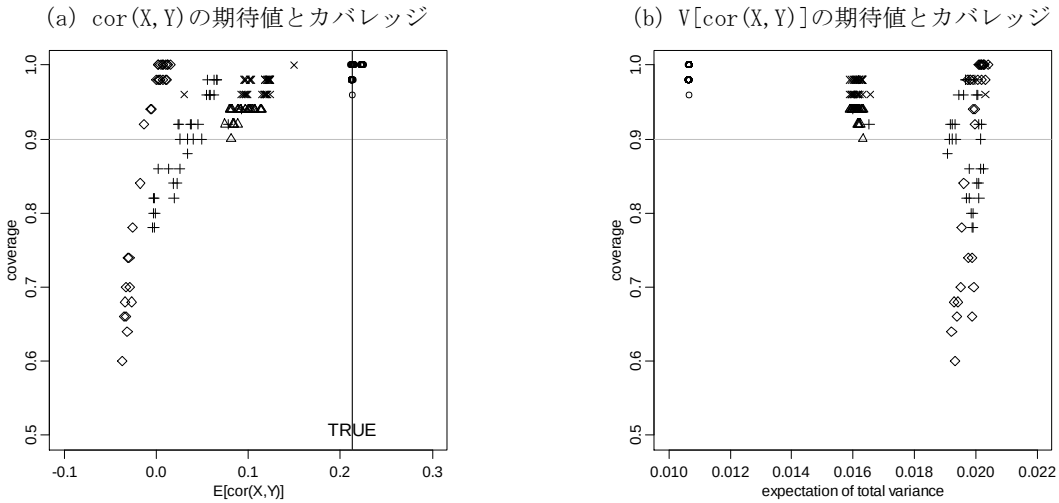
との相関が極めて強い（少なくとも 0.8 以上）という条件を満たすキー変数の組み合わせが必要である。統計的マッチングの実際では，キー変数が多ければ多いほどバイアスを低下させるというわけではないため，キー変数の数よりも可能なかぎり目標変数 X または Y との相関が強いキー変数を用意することが効果的と考えられる。

4.3　キー変数の選択とカバレッジ

まず，点推定量の特性を踏まえて，統計的マッチングにおける区間推定量の特性，とくに統計的マッチングによる MI 値の 95%信頼区間が有用であるのか，キー変数の組み合わせとの関連も含めて確認しておこう。

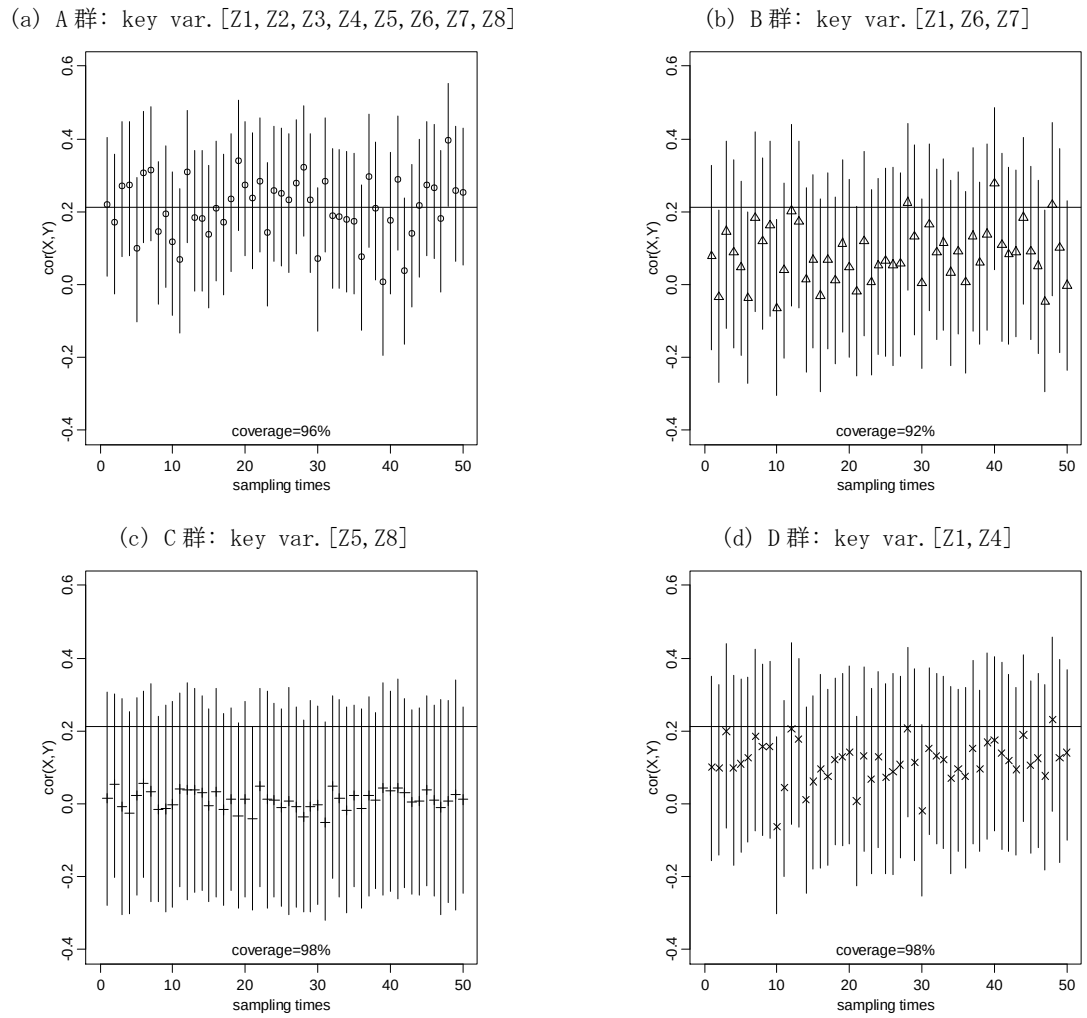
図 7 には，RIEPS についてキー変数の組み合わせ別に，横軸を X と Y の相関係数の期待値 (a)，もしくは X と Y の相関係数の分散の期待値 (b) としたときのカバレッジ (縦軸) が示されている。図 7(a)によれば，真値の近傍にある A 群（○）は 90%以上のカバレッジを示しているが，若干バイアスのある B，D 群（△，×）の 95%信頼区間についても，同様に 90%以上の比率で真値をカバーしている。図 7(b)から推察できるように，バイアスが大きい B～E 群については，推定量の分散（Total variance）が大きくなることでカバレッジが高く保たれているようである。これは，各群の一例として抽出標本別の 95%信頼区間の捕捉範囲を示した図 8 からも，抽出回数 50 回に対してカバレッジは 90%以上であり，バイアスが大きい場合にはその信頼区間が広がることで真値を捕捉していることが確認できる。ただし，95%信頼区間に対してカバレッジが 95%に満たないケースも存在しており，また，バイアスが大きい C,E 群（＋，◇）については，カバレッジが 90%以下となるケースもあることから，統計的マッチングの適用による MI 値の信頼区間については若干の補正が必要と考えられる。

図 7　信頼区間のカバレッジ（RIEPS）



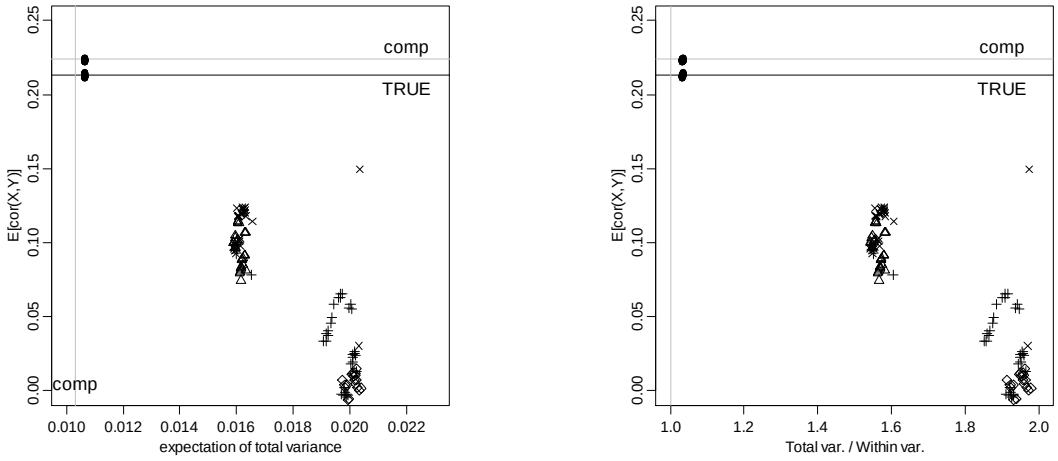
（注）マークの種別は図 4 と同様である。また Total Variance は，相関係数の z 変換値に対する分散である。
（出所）著者により作成。

図 8　抽出標本別，信頼区間とカバレッジ（RIEPS）



（注）図中の cor(X,Y)=0.213 に位置する横線は真値を示している。
（出所）著者により作成。

図 9　cor(X,Y)の期待値と total variance の期待値の関係（RIEPS）



（注）Total Variance，Within Variance は，相関係数の z 変換値に対する分散である。
（出所）著者により作成。

これまでの結果を整理すると，推定量の期待値と Total Variance（T）[14]の期待値との関係は図 9(a)のように示される。明らかに，バイアスが小さい場合には分散Tも小さく，バイアスが大きい場合には分散Tも大きい傾向が見てとれる。そこで,Total Variance（T）と Within Variance（W）との関係がバイアスに与える影響を捉えるために，Within Variance と Total Variance の比率（T/W比）を算出し，これを横軸とした図 9(b)を作成した。2.1 節の（19）式から，T/W比は Between Variance がゼロのとき 1 となり，Multiple Imputation の回数Mが一定であれば，Between Variance が大きな値をとるにつれて 1 から乖離していくことになる[15]。実際に図 9(b)からは，T/W比が 1 付近にある A 群はバイアスが小さく，T/W比が 1 から乖離するにつれて，バイアスも大きくなる様子が見て取れる。

さらに，統計的マッチングを適用する際には，1 つのデータセットが利用できるだけであるから，推定量の期待値ではなく，各標本についての標本相関係数とT/W比との関係を確認しておこう（図 10)。バイアスの小さい A 群からは，Z1 から Z8 まで全ての変数を使ったケースを一例として取り上げた。50 回分の抽出実験におけるマッチング推定値の分布は，完全データによる推定値（マーク・）の分布とほぼ同様の傾向を示しており，T/W比についても「1」の近傍に位置している。これに対して，バイアスの大きい B,C,D の各群から，適当なキー変数の組み合わせに対応する推定値の分布とT/W比の関係をみると,いずれのケースも完全データの分布とは大きく異なり，またT/W比も 1 から離れた値を示している。

これらの結果は,T/W比の数値が 1 の近傍にあるか否かがキー変数選択の良し悪しの目安を与えてくれる可能性を示唆している。ただし，バイアスの程度とT/W比の関係は，データセットの条件によって変動するため，その理論的背景を踏まえつつ,CID やキー変数と目標変数との相関の強さなど，諸条件をコントロールしながらさらに検討を加える必要がある。

なお，MHL から得られた推定量の期待値とカバレッジとの関係からは（図 11a)，推定量のバイアスが大きくなるにつれカバレッジは低下しており，また C 群（一例）における各標本の信頼区間のパフォ-マンス（図 11b）をみても，95%信頼区間とは名ばかりの結果である。とくに,本稿で適用したマハラノビス距離関数に基づく信頼区間に関しては，マッチングによる不確実性をその評価方法に反映させることができないため,これを推定量としてそのまま分析に利用するのは問題である。マハラノビス法に関しては，マッチング誤差の評価方法を含めてさらなる検討が必要である。

5. おわりに

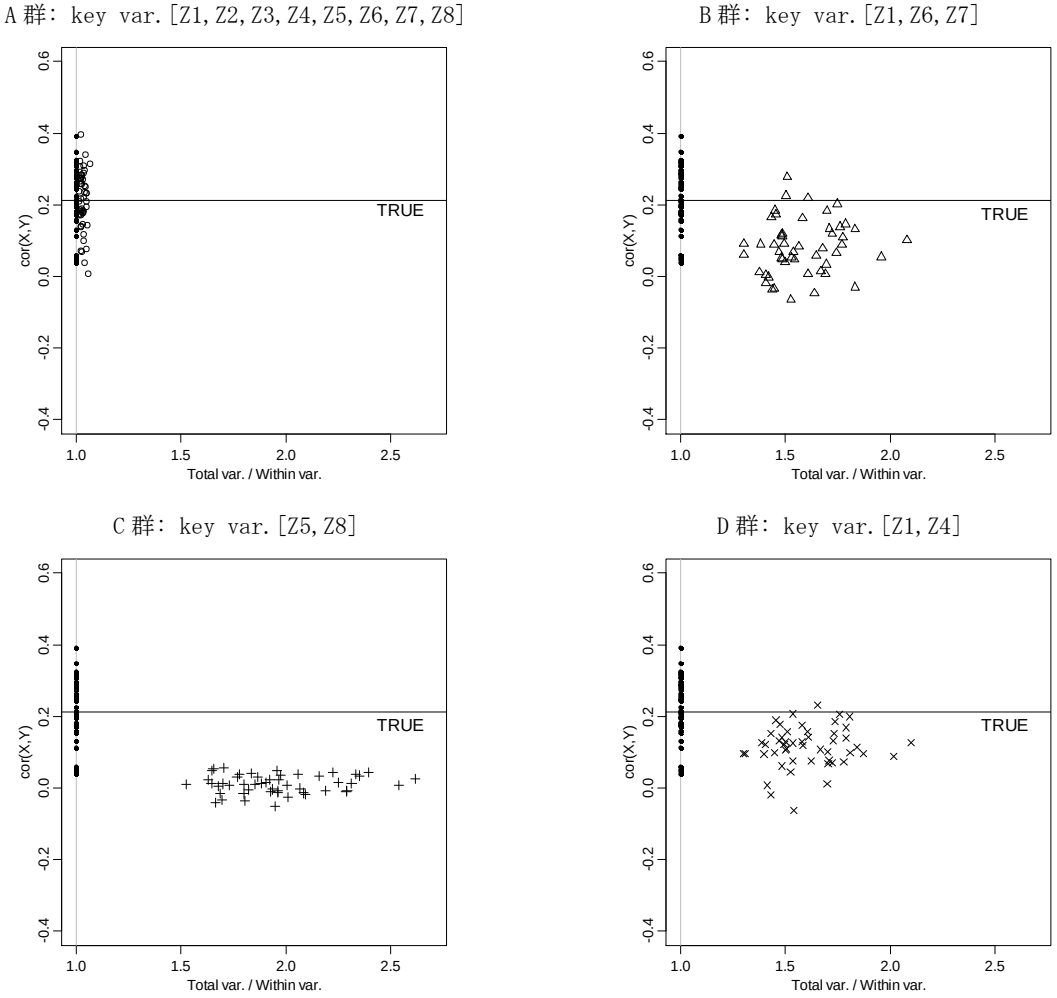
本稿では，法人企業統計調査の調査票情報を対象に，統計的マッチングによるパネルデータの作成可能性とその有効性を吟味するために，マッチング法も含めて異なる条件の下で作成したマ

---

[14] Total Variance, Within Variance, Between Variance は，z 変換値についての値を用いている。

[15] 相関係数の z 変換値が正規分布に近似する特性によれば,z 変換値の分散は$1/(n-3)$で与えられることから，サンプルサイズが一定であれば,完全データに基づく z 変換値の分散と,統計的マッチングによる z 変換値の Within Variance は一致する。すなわち,相関係数の z 変換値によるT/W比は,完全データによる推定量の分散に対する,マッチング推定量の分散を表しており,これは統計的マッチングによる不確実性として解釈することもできる。

図 10　抽出標本別，相関係数と T/W 比（RIEPS）

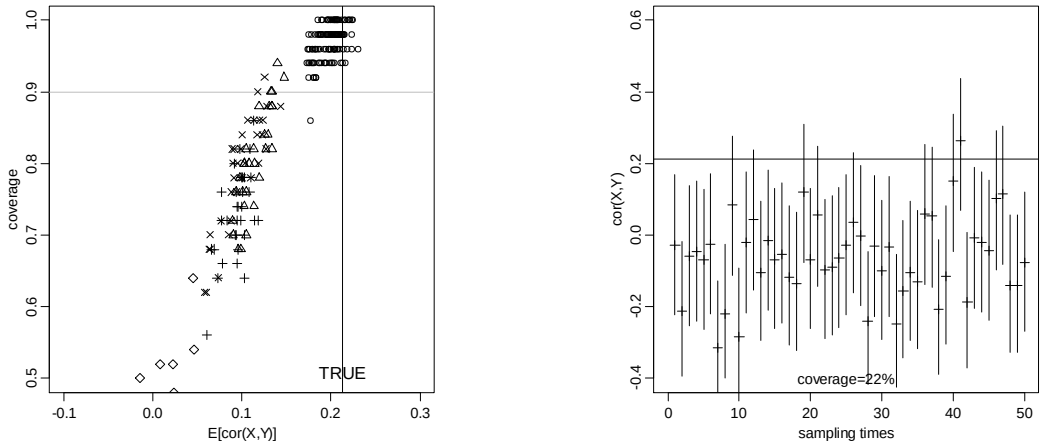


（注）Total Variance，Within Variance は，相関係数の z 変換値に対する分散である。
（出所）著者により作成。

図 11　信頼区間のカバレッジ（MHL）

(a) 推定値の期待値とカバレッジ

(b) 抽出標本別，信頼区間とカバレッジ
C 群：[Z5,Z8]



（注）図(b)中の cor(X,Y)=0.213 に位置する横線は真値を示している。
（出所）著者により作成。

ッチング・データからの推定量（相関係数）の精度検証を行った。法企データの一部の調査変数に関しては，調査票情報として前期と当期のデータが与えられているため，パネルデータを作成する際の障壁となるキー変数の時点間のズレに関する問題を，ある程度回避できる。そのため法企データは統計的マッチングによるパネル化という点では，他統計に比して有利な条件が揃っている。このような条件を活用しながら，とりわけ精度の良いマッチング推定量（相関係数）を得るためのキー変数の選択条件をマッチング実験により明らかにすることを試みた。

まず，法企データに対して，ノンパラメトリック手法であるマハラノビス法，パラメトリック手法である RIEPS または NIBAS を適用したところ，バイアスの観点において，CID が低く統計的マッチングに適したデータセットの下では RIEPS または NIBAS が適していた。ただし，CID が比較的高い数値を示すデータセット条件の基ではマハラノビス法の利用が適切である可能性も考えられる。

また，RIEPS または NIBAS に関して，統計的マッチングから適切な推定量を得るための条件としては，CID がゼロ付近（0.03 以下）であり，かつキー変数と目標変数 X（または Y）との相関が極めて強い（少なくとも 0.8 以上）ことが必要条件となる。同時に，キー変数の数の多寡はマッチングの精度に強い作用を及ぼすものではないため，キー変数を増やすことよりも，可能なかぎり目標変数 X および Y との相関が両者ともに強いキー変数を用意する方が効果的といえる。

さらに，95%信頼区間に含まれる真値の割合を示すカバレッジ指標（RIEPS または NIBAS）については，目標変数との相関が強いキー変数の組み合わせにおいて，約 90%以上という高いパフォーマンスが観測され，マッチング推定による不確実性が，ある程度，多重代入法により捉えられていることを確認した。ただし，その捕捉率は 95%に満たないケースもあるため，いかなるケースでもカバレッジ 95%以上を保持し，通常の信頼区間と同じ程度の確率的意味合いで利用するためには，若干の補正が必要と考えられる。なお，本稿で適用したマハラノビス距離関数に基づく 95%信頼区間に関しては，とくにバイアスが大きい場合にはそのカバレッジは極めて低い値を示している。マハラノビス法に関しては，マッチング誤差の評価方法を含めて未だ多くの問題が残されている。

なお，キー変数の選択に着目したとき， Total Variance と Within Variance の関係を示す数値（T/W 比）が「1」に近いキー変数セットであれば（すなわち Between Variance がゼロに近いほど），バイアスも小さくなる傾向が確認された。統計的マッチングの有効性を担保するキー変数選択の条件にも関わる特徴であることから，より一般的な条件下でのその妥当性については，さらに検討を重ねることにしたい。

【 参考文献 】

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

[2] Goel, P.K. & T. Ramalingam(1980), *The Matching Methodology: Some Statistical Properties*, Springer-Verlag.

[3] Haltiwanger,J.C.,etc(1999),*The Creation and Analysis of Employer-Employee Matched Data*, North-Holland.

[4] Little, R.J.A. & D.B. Rubin(2002), *Statistical Analysis with Missing Data*, Wiley Series in Probability and Statistics.

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

[6] 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.

[7] 奥野忠一, 山田文道 (1995),『情報化時代の経営分析』, 東京大学出版会.

[8] 栗原由紀子(2012a),「相関特性推定における統計的マッチングの有効性についてーモンテカルロ・シミュレーションによる精度検証ー」,『中央大学経済研究所年報』, 中央大学経済研究所, 第 43 号, pp.489-551.

[9] 栗原由紀子(2012b),『疑似景況パネルによる予測パフォーマンスの計測ーマハラノビス・マッチングの適用からー』, 法政大学日本統計研究所, オケージョナル・ペーパー, No.35, pp.1-38.

[10] 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」,『中央大学経済研究所年報』, 中央大学経済研究所, 第 44 号, pp.271-306.

[11] 間瀬茂 (2007),『R プログラミングマニュアル』, 数理工学社.