傾向スコア・マッチング法を用いた新幹線が沿線地域の人口変動に及ぼす因果効果の推定

オウ　セイゲン[1]・寺部　慎太郎[2]・柳沼　秀樹[3]・海野　遥香[4]・鈴木　雄[5]

[1]非会員　東京理科大学大学院　理工学研究科土木工学専攻　修士課程（〒278-8510 千葉県野田市山﨑 2641）
E-mail: 7622505@ed.tus.ac.jp (Corresponding Author)

[2]正会員　東京理科大学教授　理工学部土木工学科（〒278-8510 千葉県野田市山﨑 2641）
E-mail: terabe@rs.tus.ac.jp

[3]正会員　東京理科大学准教授　理工学部土木工学科（〒278-8510 千葉県野田市山﨑 2641）
E-mail: yaginuma@rs.tus.ac.jp

[4]正会員　東京理科大学助教　理工学部土木工学科（〒278-8510 千葉県野田市山﨑 2641）
E-mail: unoharuka@rs.tus.ac.jp

[5]正会員　東京理科大学助教　理工学部土木工学科（〒278-8510 千葉県野田市山﨑 2641）
E-mail: yusuzuki@rs.tus.ac.jp

日本の新幹線は 1964 年に開業して以来，約 60 年にわたり整備が進められてきた．新幹線の整備は人の移動を促進し，沿線地域の人口動態の変化に大きな影響を与えた．本研究では，国勢調査の 1km グリッドデータを用いて，新幹線沿線地域の人口変動を分析することを目的とする．具体的には，1995 年から 2015 年までの間に新幹線駅が周辺地域の人口変動に及ぼす因果効果を，傾向スコア法を用いて推定する．分析においては，傾向スコア法の重み付け，層別解析とマッチングそれぞれの手法を使用して，ATT（処置群における平均処置効果）を計算する．特に，傾向スコア・マッチング法では，区間マッチング，最近傍マッチングおよびキャリパー・マッチングをそれぞれ用いて ATT を算出し，その結果を比較分析する．最後に，IPW （逆確率重み付け）法を用いて ATE（平均処置効果）を算出し，重回帰分析で得られた結果と比較する．

***Key words**: Causal inference, High speed railway, Propensity Score, IPW, Population change*

1. はじめに

日本の新幹線は 1964 年に東海道新幹線が開業して以来，山陽新幹線や東北新幹線，上越新幹線が続いて開業して約 60 年にわたり整備が進められてきた．2016 年に北海道新幹線の新青森駅から新函館北斗駅間が開業した時点で，全国に新幹線駅はすでに 92 駅存在する．新幹線は，遠距離移動に便利な交通手段として，全国の都市間の移動や経済活動に大きな影響を及ぼしている．

交通インフラへの投資は，経済全体の成長のための重要な手段である．現在，世界では 20 カ国以上で高速鉄道が運行または建設されている．現在，先進国，途上国を問わず，多くの国が高速鉄道への投資の意志と計画を持っている．日本は世界で最も古くから高速鉄道を有しており，高速鉄道の影響を研究するのに適したケースである．そして，限られた資源を有効に活用し，社会資本ストック効果を最大化するために，事後評価は欠かせない．

交通基盤整備は地域のアクセス性向上に伴う生産力の拡大，人口や雇用の増加，資産価値の向上など多様な影響を地域社会に及ぼすものの，その因果関係を明らかにすることは容易ではない．事後評価では，ストック効果の発現状況を多面的に捉え，統計データを有効に活用しながら，可能な限り定量的，客観的に効果を把握することが求められる．インフラ整備によってもたらされるストック効果を適切に評価するためには，実務で一般的に行われている単純な前後比較では十分とは言えない[1)]．

さて，統計的因果推論は近年，様々な分野で広く応用されている．特に，観察データを用いる際に交絡の影響を低減・除去する傾向スコアの手法に関心が集まってい

る．したがって，因果推論手法を交通の分野に適用する試みは，非常に重要なトピックである．

本稿の構成は，以下の通りである．2．では，既往研究を紹介した上で，本研究の着眼点について述べる．3．では，分析に使用したデータと本研究で使用する方法を示す．4．では使用した手法による推定結果とバランスチェックを説明する．また，様々な手法で得られた結果を集計した上で比較・分析する．最後に，5．で結論を述べる．

2. 本研究の基本的な考え方と位置付け

(1) 既往研究の概要

傾向スコアとは，観察された共変量に基づいて処置が割り当てられる確率のことである．傾向スコアによって，ランダム化比較試験の特徴のいくつかを模倣するように，観察（非ランダム化）研究を設計し分析することができる．傾向スコアはバランシング・スコアであり，それに基づいて，観察された共変量の分布が処置群と対照群の間で類似することになる．非ランダム化研究や観察研究において，処置や介入の因果関係を調べる目的で，観察された変数に関する処置前の不均衡をコントロールするために傾向スコアを用いることが，過去 10 年間に広く行われるようになった．これまでに教育 [2]，経済 [3]，心理 [4]，経営 [5]の分野でこの手法を応用した研究がある．

新幹線の整備効果に関する実証研究は数多くみられている．松永ら [6]は東北新幹線や九州新幹線の開業前後で利用者数や観光客の人数を比較し，その増加分を各新幹線による開業効果であるとして事後評価を行っている．しかし，これらの研究では開業直後の短期的な影響評価を行うことはできるものの，他の要因による影響を排除することが困難になるため，長期的な影響評価を行うことはできない．

在来線に関する研究として，中川ら [7]は 1981 年から 1990 年に廃止された駅を対象に，ローカル鉄道の廃止が駅周辺人口に及ぼす影響を定量的に推計した．傾向スコア・マッチング差分法（PSM-DID）を用いて，ローカル鉄道を廃止した場合，存続させた場合と比較して駅周辺人口が最大で 8.3%減少することを統計的に明らかにした．また，ローカル線廃止の影響は，駅周辺の特性によって異なることも明らかにした．特に，廃止前の人口が多い駅や公共交通機関のモーダルシェアが高い駅ほど，ローカル線の廃止による人口減少が大きいことを明らかにした．

Talebian et al. [8]は傾向スコア・マッチングによりカリフォルニア州が行ったアムトラック駅への経済的支援が地域の人口と雇用に与える影響を評価した．この研究では，群レベルと市レベルで経済的支援を受けた群と受けなかった群を傾向スコア・マッチングにより共変量のバランスを調整し，重回帰分析を行った，その結果，カリフォルニア州による経済的支援により整備が行われたアムトラック駅がある市(群)は，人々にとって魅力的な治線となり人口には正の影響がみられたが，地元の雇用に対する影響は限定的であった．

新幹線に関する研究では，室ら [9] は新幹線駅の利便性が，現在新幹線駅が立地する地方自治体の人口や就業者人口をはじめとした指標に正の影響を与えているという仮説を設定し，両者の経年変化の関係を実証的に分析した．その結果，利便性の高い新幹線駅がある地方自治体は，利便性の低い新幹線駅がある地方自治体と比較すると，それらの指標の減少幅が小さいことが明らかとなった．また，工業製品出荷額が増加しているとそれらの指標の減少幅が同様に小さいことが明らかとなった．

片岡ら [10]は新幹線の新規整備が我が国の実質 GDP の向上に寄与し，一定のマクロ経済改善効果があることを確認した．地方別の生産額及び人口の変化に着目した分析では，現状整備との比較においてリニア中央新幹線の整備や新幹線の全国整備を進めた場合，関東地方の人口が最大 4.2%，GRP は最大 5.3%の水準で少ない．一方，各地方においては人口等が多く，「分散化」効果があることがわかった．これらの結果は，新幹線の新規整備が日本全体の成長力向上に寄与し，また人口と経済力の偏在状況を改善する効果を持つことを示すと言える．

Jia et al. [11]は差の差分析と傾向スコア・マッチング差の差分析(PSM-DID)を用いて中国の高速鉄道が地域経済の発展に与える影響は路線を比較した場合で異なるかということを検証した．その結果，路線間で影響は異なり，中国の高速鉄道建設が経済に正の影響を与えたことがわかった．

(2) 本研究の位置と意義

本稿は 3 種類の傾向スコア法で分析を行う．それらは，傾向スコアを用いた処置の逆確率重み付け，傾向スコアによる層別化，そして傾向スコアによるマッチングである．そして，これらの各手法を用いて，新幹線が沿線人口変化に与える影響を定量化し，各手法で得られた結果を比較，分析し，各手法の共変量バランスを図示し，検証し，調整する．本研究の独自の貢献は，様々な手法の中で最もバランスが良く，最も精度の高い手法を明らかにする点と，交通分野の因果分析への傾向スコア法の適用に関する今後の研究への参考となる点にある．

3. データと研究方法

(1) 分析対象と使用データ

新幹線路線は，東海道新幹線，山陽新幹線，九州新幹線，上越新幹線，北陸新幹線，東北新幹線，山形新幹線，秋田新幹線である．これらをすべて合計すると 2765 ㎞，駅数は 105 である．本研究では，そのうち，1995 年から 2015 年までに供用開始された新幹線駅の周辺地域を分析対象とする．

これらの路線の新幹線駅を中心とする半径 15km 圏内に存在する 1km グリッド(3 次メッシュ)を処置群に設定する．距離は，各グリッドの中心と，最寄りの新幹線駅までの直線距離である．直線距離が最も短い駅を選ぶため，半島の一部のグリッドでは，海を挟んだ対岸の駅を最寄り駅として選ぶ場合がある．その場合は，個別に確認して，陸路でアクセスできる駅を選ぶようにした．

一方，新幹線駅から 15km 以上離れているグリッドは対照群の候補とする．対照群の選定に当たっては，処置の割り当て（新幹線の整備）以外の特徴が処置群とできるだけ類似したグリッドを選ぶことが重要である．そこで，処置群の地理的，環境的および経済的な特徴を勘案し，以下の 5 つの条件のいずれかを満たす地点を対照群の候補から除外する．それらは，１）北海道と四国内のグリッド，２）離島（佐渡，沖縄県など）内のグリッド，３）標高が 1000m 以上であるグリッド，４）森林地域，自然公園地域内のグリッド，５）1995 年から 2015 年まで無人のグリッド，である．以上の条件のいずれかを満たすものを除外し，残ったグリッドは対照群の候補とする．

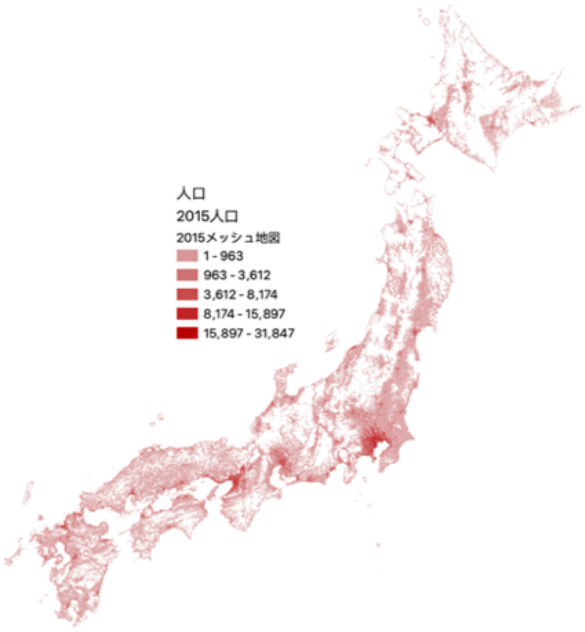

図-1 2015 年における人口分布（３次メッシュ）

アウトカムとして用いるデータは，e-Stat 政府統計の総合窓口の国勢調査（1995 年，2000 年，2005 年，2010 年，2015 年）の「人口数」に関する３次メッシュ（１㎞グリッド）データである．傾向スコアを算出する際の共変量データは，国勢調査の市町村データ（1995 年），土地利用メッシュデータ（1997 年，2006 年，2009 年，2016 年）および国土交通省の地形に関する標高・傾斜度３次メッシュ（2009 年）を用いる．なお，各種データを３次メッシュ単位で集計する作業は QGIS を用いて行った．

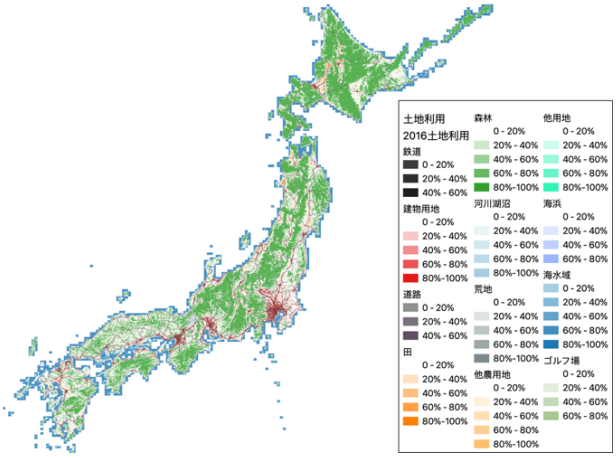

図-2 2016 年における土地利用分布（３次メッシュ）

(2) 傾向スコア法の紹介と推定 [12), 24)]

傾向スコアは，観察された共変量 X＝（X_1，X_2，…，X_k）で条件つけた，処置される（D＝１）確率

$$e_i(X_i)=P_r(D_i=1|X_i) \quad (1)$$

として定義される．傾向スコア e(X)で条件つけた場合，共変量の分布と処置の割り当てが独立となる．

$$\{Y(0),\ Y(1)\}\perp D|X \quad (2)$$

このとき，傾向スコアに関する条件付き独立性

$$\{Y(0),\ Y(1)\}\perp D|e(X) \text{かつ} D\perp X|e(X) \quad (3)$$

が成立し，傾向スコアに基づく共変量によって因果効果の推定が可能となる．本研究では，以下のロジットモデルを用いて，傾向スコアを算出する．

$$D_i \sim Bernoulli(\theta_i) \quad (4)$$

$$Logit(\theta_i)=a_0+a_1X_{1i}+\ldots+a_kX_{ik} \quad (5)$$

処置の割り当てを新幹線が整備された(D=1)か否(D=0)かとし，高速道路 IC までの距離，土地利用や地理的環境などを共変量 X_i として推定を行う．共変量は表-1 に示す．

表-1 共変量の選定

変量	平均値	最小値	最大値	中央値
ICまでの距離(km)	9.442	0.005	85.253	7.075
田(%)	17.69	0	100	90
他農用地(%)	8.929	0	100	3
森林(%)	47.34	0	100	51
荒地(%)	0.9213	0	94	0
建物用地(%)	15.27	0	100	60
道路(%)	0.6202	0	31	0
鉄道(%)	0.3952	0	50	0
他用地(%)	2.422	0	98	0
河川湖沼(%)	3.471	0	100	0
海浜(%)	0.0729	0	47	0
海水域(%)	2.351	0	100	0
ゴルフ場(%)	0.5165	0	76	0
平均標高(m)	207.6	-2.1	2894	128.2
最大傾斜角(°)	9.082	0	46.3	7.8
人口(人)	809.4	0	31847	108

(3) 傾向スコアによる重み付け [12)]

a) ATT（処置群における平均処置効果）の推定

ATT を推定するために，傾向スコア $e_i(X)$ を用いて次の重みを計算する．

$$w_i^{ATT}=D_i+(1-D_i\frac{e_i(X)}{1-e_i(X)} \quad (6)$$

この重みを使うと，処置群（$D_i=1$）のグリッドの重みは $w_i^{ATT}=1$ になる．つまり，処置群のグリッドには，傾向スコアに関係なく等しい重みが与えられる．

それに対し，対照群（$D_i=0$）のグリッドの重みは，$w_i^{ATT}=e_i(X)/[1-e_i(X)]$ になる．つまり，傾向スコアが高いほど，大きな重みが与えられる．これは，傾向スコアが高い，すなわち処置群に入る確率が高いにも関わらず対照群に入ったグリッドは，処置群の比較対象として重宝されるということを意味する．反対に，傾向スコアが低いグリッドは，処置群の比較対象としてあまり重要ではない（処置群にはそれによく似たグリッドが少ないのに，同種のグリッドが対照群にたくさんある）ので，割り引いて考えられるということである．

b) IPW: ATE（平均処置効果）の推定

ATE を推定するためには，IPW (Inverse Probability Weighting; 逆確率重み付け) という方法を用いる．IPW では，それぞれの群で珍しいグリッドほど重みが大きくなる． ATE =E[Y(1)]−E[Y(0)] のそれぞれの項を，傾向スコアを利用した重み付けによって推定する． E[Y(1)]の推定値は，重み $w_{1i}=\frac{D_i}{e_i(X)}$ を使って，

$$E[Y(1)]=\sum\frac{w_{1i}Y_i}{\sum w_{1i}}=\sum\frac{D_i/e_i(X)}{\sum[\frac{D_i}{e_i(X)}]}Y_i \quad (7)$$

である．E[Y(0)]の推定値は，重みを $w_{oi}=\frac{1-D_i}{1-e_i(X)}$ として，

$$E[Y(0)]=\sum\frac{w_{0i}Y_i}{\sum w_{0i}}=\sum\frac{(1-D_i)/(1-e_i(X))}{\sum[\frac{1-D_i}{1-e_i(X)}]}Y_i \quad (8)$$

である．これらの差で，ATE が推定できる．

$$ATE=E[Y(1)]-E[Y(0)]$$
$$=\sum\frac{D_i/e_i(X)}{\sum[\frac{D_i}{e_i(X)}]}Y_i-\sum\frac{(1-D_i)/(1-e_i(X))}{\sum[\frac{1-D_i}{1-e_i(X)}]}Y_i \quad (9)$$

(4) 傾向スコアによる層別解析

傾向スコアによる層別解析は，推定された傾向スコアに基づいて，グリッドを相互に排他的なサブセットに層別することである．グリッドは推定された傾向スコアに従って順位付けされる．次に，グリッドは事前に定義された傾向スコアの閾値に基づいてサブセットに層別化される．一般的なアプローチは，推定傾向スコアの 5 分位を使用してグリッドを 5 つの等しいサイズのグループに分ける方法である．Cochran[13)] は，連続交絡変数の 5 分位で層別化すると，その変数によるバイアスの約 90%が除去されることを実証した．Rosenbaum and Rubin[14)] は，この結果を傾向スコアでの層別化に拡張し，線形処置効果を推定するときに，傾向スコアの 5 分位で層別化すると測定された交絡因子によるバイアスの約 90%が除去されると述べている．

それぞれの傾向スコア層では，処置群と対照群はほぼ同じ傾向スコア値を持つことになる．したがって，傾向スコアが正しく指定された場合，測定された共変量の分布は，同じ層内の処置群と対照群の間でほぼ同様になる．各層では，処置したグリッドと処置していないグリッドとの間で直接アウトカムを比較することにより，アウトカムに対する処置の効果を推定することができる．そして、各層の処置効果推定値は，合計して集計され、全体的な処置効果を推定することができる．

(5) 傾向スコアによるマッチング [12) 15) 28)]

傾向スコア・マッチングは，傾向スコアの値が近い処置グリッドと対照（未処置）グリッドを対にしてマッチングすることを意味する．傾向スコア・マッチングによって，ATTを推定することができる．傾向スコア・マッチングの最も一般的な実装は 1 対 1 マッチングまたは k 対 1 マッチングで，処置グリッドと対照グリッドのペアを形成し，マッチングされたグリッドの傾向スコアの値が類似しているようにするものである．本研究で使用するデータ量（グリッド数）は十分にあり，処置群と対照群の数がかなり多いので，以下のマッチング方法はすべて非復元抽出，一対一のマッチングである．

a) 区間マッチング

区間マッチングの考え方は，分布している傾向スコアを複数の区間（層）に分割し，処置群と対照群の間の結

果の平均差を取ることによって，各区間内での影響を計算することである．この方法は，層別マッチングとも呼ばれる．

b) 最近傍マッチング

最近傍マッチングは，処置グリッドの傾向スコアに最も近い対照グリッドを処置グリッドにマッチングするために選択する方法である．複数の対照グリッドの傾向スコアが処置グリッドの傾向スコアに等しい場合，これらの対照グリッドのうち 1 つが無作為に選択される．なお，ペアになった 2 つのグリッドの傾向スコアの差の最大許容値には制限がない．

c) キャリパー・マッチング [22)]

キャリパー・マッチングは，近傍マッチングに似ているが，ペアになったグリッドの傾向スコアの絶対差が，事前に指定された閾値（キャリパー距離）以下でなければならないという制約が加えられている．したがって，ある処置グリッドについて，その傾向スコアが処置グリッドの傾向スコアと指定された距離内にあるすべての対照グリッドを限定することになる．この限定された複数の対照グリッドから，傾向スコアが処置グリッドのそれに最も近い対照グリッドが選ばれ，この処置グリッドとマッチングされることになる．もし，処置グリッドの傾向スコアの指定されたキャリパー距離内に位置する傾向スコアを持つ対照グリッドがない場合，その処置グリッドはどの対照グリッドともマッチングされない．そして，マッチングされなかった処置グリッドは，結果として分析対象から除外されることになる．

4. 推定結果と考察

傾向スコア分析の重要な要素は，傾向スコアモデルが適切に規定されているかどうかを検証することある．本章では，傾向スコアモデルが適切に規定されているかどうかを評価する方法 [21)] を示す．

傾向スコアモデルが適切に規定されているかどうかを評価するには，傾向スコアの分布を確認し，バランス調整などをすることである． IPW では，処置グリッドと対照グリッドを比較し，グリッドを処置の逆確率で重み付けして評価する．傾向スコアによる層別解析では，この評価は傾向スコアの層内での処置グリッドと対照グリッドの比較を必要とする．傾向スコア・マッチングでは，傾向スコアモデルが適切に規定されているかどうかを評価するために，傾向スコアをペアになった標本内の処置群と対照群を比較する．

標準化平均差 (standardized mean difference)は，プールされた標準偏差の単位で平均値の差を比較するものである．さらに，サンプル数に影響されず，異なる単位で測定された変数の相対的なバランスを比較することができる．均衡していないと判定される標準化平均差の閾値について，普遍的に合意された基準はないが，0.1 未満の標準差は，処置群間の共変量の平均値の差が無視できることを示すと考えられてきた． [13)] 標準化平均差は次式で表される．

$$d = \frac{\overline{X}_{D=1} - \overline{X}_{D=0}}{\sqrt{\frac{Var(X_{D=1}) + Var(X_{D=0})}{2}}} \tag{10}$$

厳しい基準: $|d| < 0.1$ (Austin[14])

緩い基準: $|d| < 0.25$ (Stuart[15])

標準化平均差は，傾向スコア・マッチングにおける処置群間の共変量の平均値を比較する．同様に，傾向スコア・マッチングにおける処置群間の連続的な共変量の分布を比較するために，サイドバイサイドボックスプロット，分位点プロット，累積分布関数，経験的ノンパラメトリック密度プロットなどのグラフ手法が使用されることがある． [15)]

(1) 推定された傾向スコアの分布の確認

共変量が多いので，各共変量の数値診断を精査するのは困難な場合がある．そこで，グラフを用いて共変量のバランスを素早く評価することができる．まず，元のグループとマッチングしたグループの傾向スコアの分布を調べる．これは，共有サポート [12)](common support)の評価に有効である．共有サポートとは，処置群と対照群で，傾向スコアが同じ範囲に分布していることを意味している．共有サポートがない相手の群に似ているグリッドがないということ比較できない，つまり、因果推論できないということである．したがって，傾向スコアの最初のステップは，共有サポートの検定である．

傾向スコアによる共有サポートの確認は 2015 年のデータを例示する（以下に示すグラフはすべて 2015 年のデータの例である）．

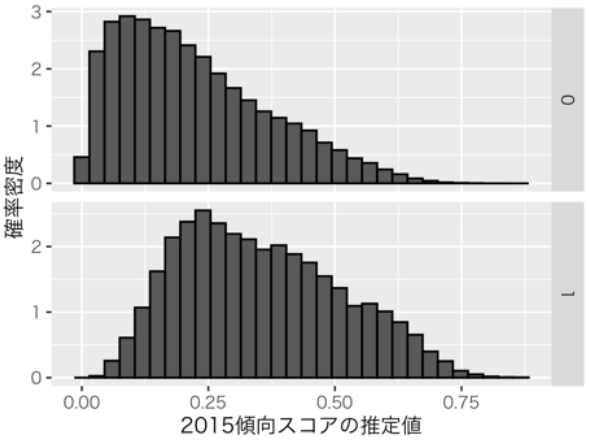

図-3 共有サポートの検定のための傾向スコアの比較
対照群（0; 上段）と処置群（1; 下段）

図-3 より，対照群（0; 上段）と処置群（1; 下段）で傾向スコアの分布の傾向が少し異なることがわかる（上段と下段で縦軸のスケールが異なる．）ただし，傾向スコアは同じ範囲に分布していることから，共有サポートはありそうだといえる．

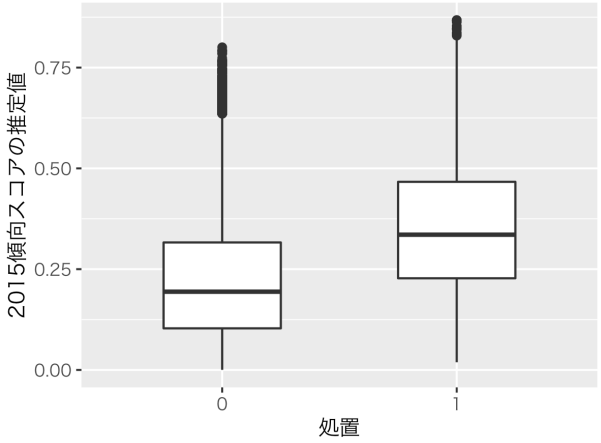

図-4　傾向スコア分布のボックスプロット
対照群（0; 左）と処置群（1; 右）

さらに，図-4 に示すボックスプロットで見ても，分布が少し異なることがわかる．ヒストグラムでは，対照群に傾向スコアが 0.75 以上のグリッドがあるかどうかがよくわからなかったが，この図から，分布の範囲は0.75 以上の部分もあることがわかる．すなわち，共有サポートがあると判定した．

(2)　傾向スコアによる重み付けのバランス調整

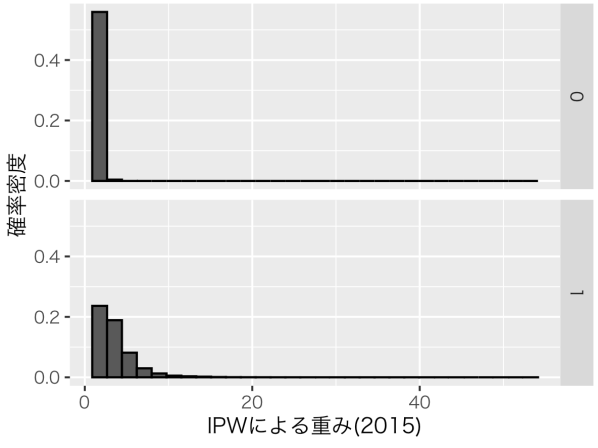

図-5　IPW による重みの分布
対照群（0; 上段）と処置群（1; 下段）

図-5 に調整前の IPW による重みの分布を示す．これをみると，処置群で傾向スコアが大きいグリッドの重みが割り引かれている一方で，対照群に傾向スコアが大きいグリッドがあまりないため，重みが割増されているグリッドはあまりないことがわかる．

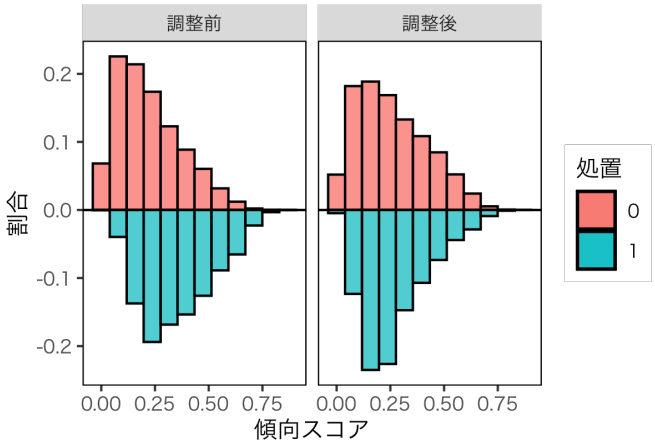

図-6　調整前後の傾向スコアの分布の比較

図-6 で調整前後の傾向スコアの分布を確認する．これをみると，調整後に処置群の分布が対照群の分布に近づいていることがわかる．

(3)　共変量のバランスチェック

a)　重み付けのバランスチェック

標準化平均値の差をプロットすると，個々の共変量についてバランスが改善されたかどうかを簡単に概観することができる．図-7に示すように，各共変量の平均の標準化された差が減少していることがわかる．この図ではないが，状況によっては，いくつかの共変量の平均の標準化された差が増加することがある．これは，マッチング前の差が小さい共変量に当てはまることが多く，それらは傾向スコアモデルに大きく影響しない（それらは処置割り付けを予測しないため）．このような場合，これらの共変量でのバイアスの増加が問題であるかどうかを検討すべきである．この例では，すべての共変量についてバランスの改善は見られるものの，標準化平均の絶対値が 0.1 より大きい共変量はまだ 2 つある．そのため，このまま処置効果を推定してもうまくいかない．

b)　層別解析

図-8 には，グリッド全体での共変量のバランスと，各層での共変量のバランスが表示されている．ほとんどの共変量について，全体よりも層別した方が標準化平均差の絶対値が小さくなっている．しかし，基準である 0.1 または 0.25 よりも大きい差が，一つの共変量（distance_ic）に残っており，この共変量については，層別によるバランスはあまり良くないことがわかる．

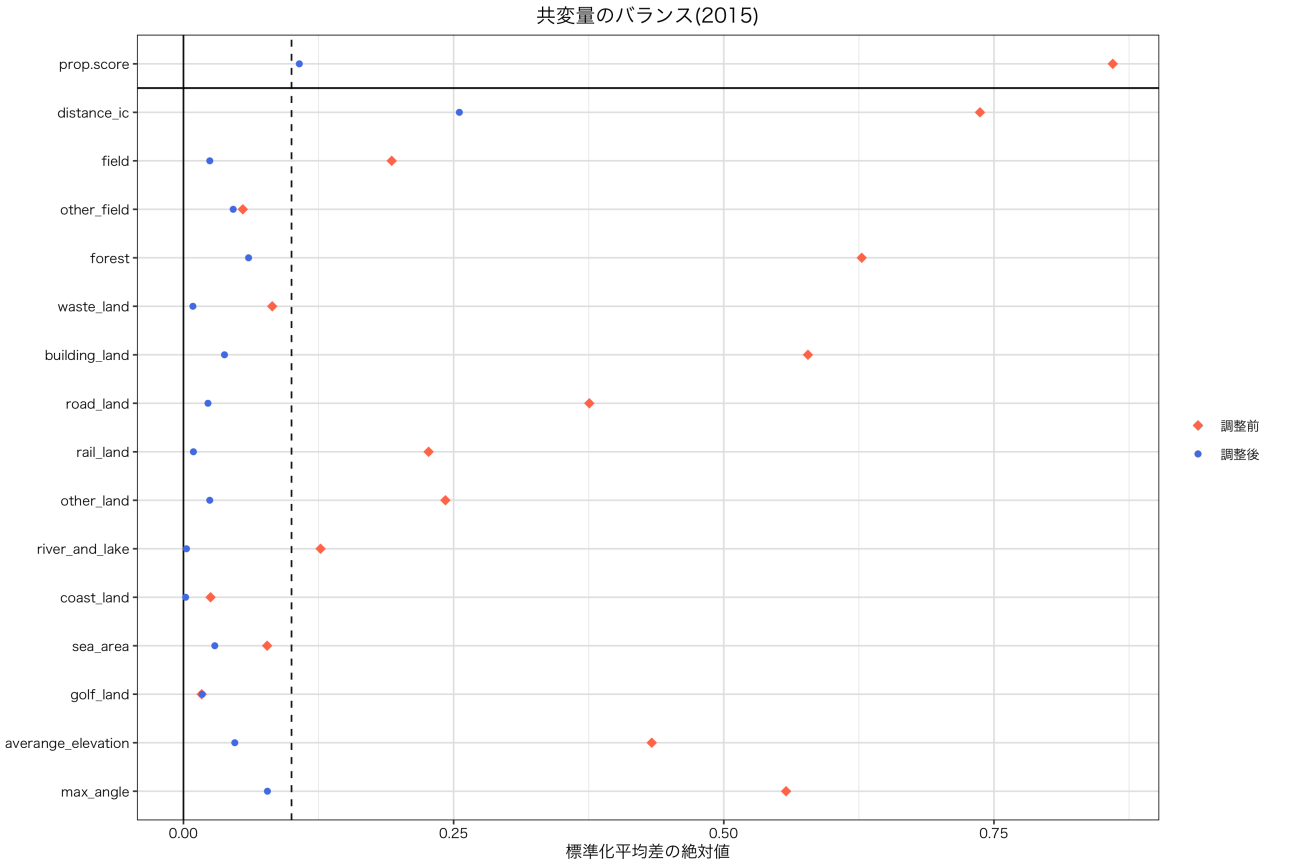

図-7　IPW による ATE 推定のバランスチェック

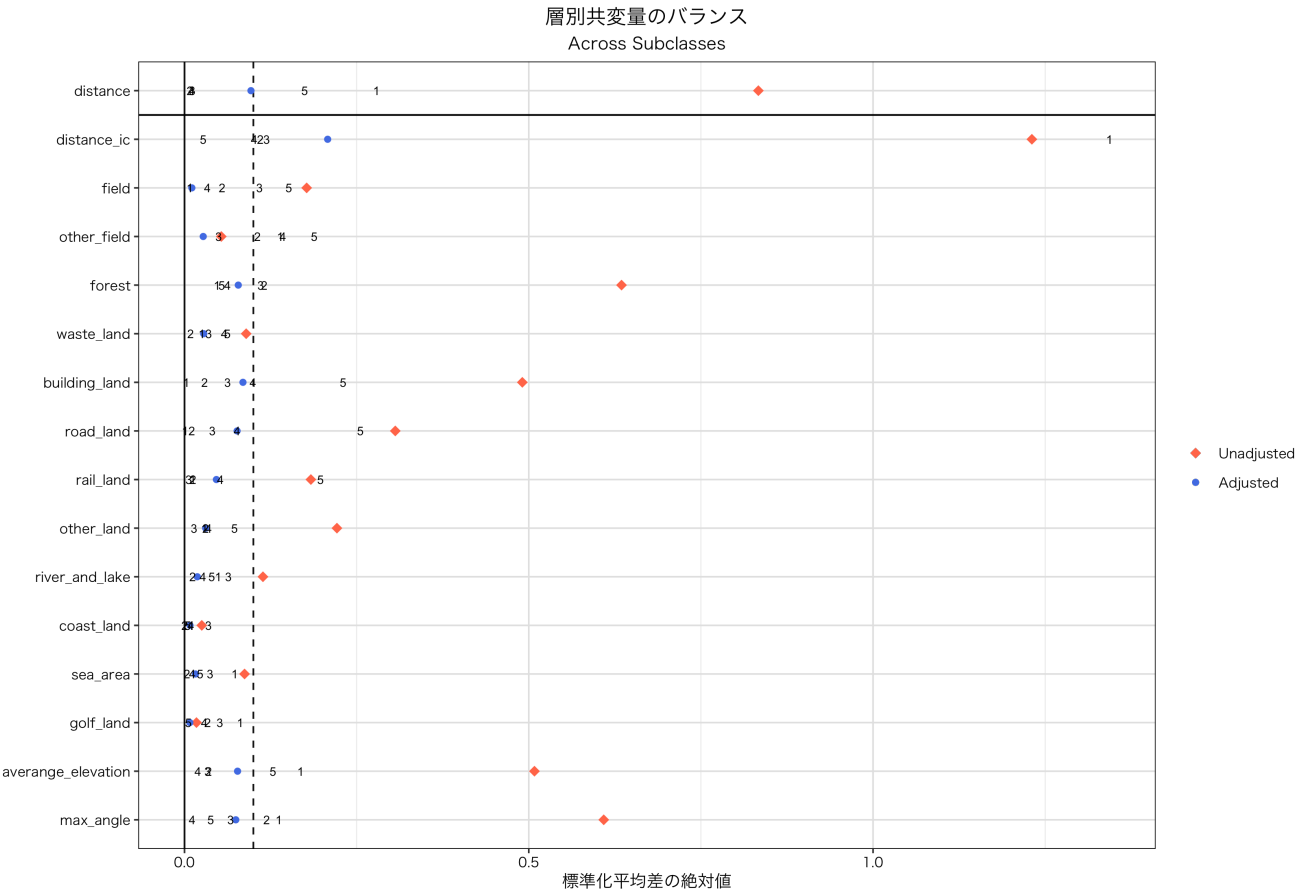

図-8　層別のバランスチェック

c) マッチング法

マッチング法を使用する上で最も重要なステップは，結果として得られるペアになったグリッドのバランスを診断することである． すべてのマッチングは，マッチングされたグループの共変量バランスについて評価されるべきであり，非常にバランスの悪いグリッドとなるマッチング方法は不適切である．[22)]

区間マッチング

第 3 章で説明したように，全体を 5 層に分けると誤差を大きく減らすことができる．今回使用した処置群グリッド数は 27390 個と，非常に多い． そこで，処置群を 5 つに等分し，それぞれ 5478 個のグリッドとすることでマッチングを行った．対応する対照群のグリッド数は，46720，14354，12024，9404，4434 であった．

図-9 には，傾向スコアの分布について，処置群（左上），対照群（左下），区間マッチング後の処置群（右上），区間マッチング後の対照群（右下）を比較したものである．区間マッチング後のグラフは，グリッドの数によって均等に分配されるため，層別は破線で表示されている．マッチング後のヒストグラムを見ると，処置群の層によって対照群のグリッド数が調整されていることがわかる．これらのヒストグラムによると，処置群と対照群で，層別後の分布の範囲は共有されている（共有サポートがある）ようである．

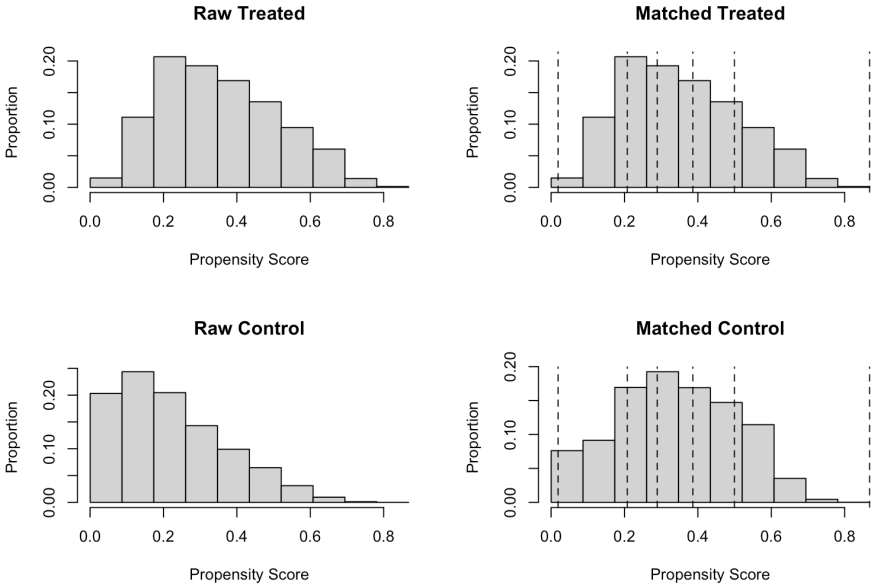

図-9 区間マッチング前後の傾向スコアの分布 処置群（左上），対照群（左下），区間マッチング後の処置群（右上），区間マッチング後の対照群（右下）

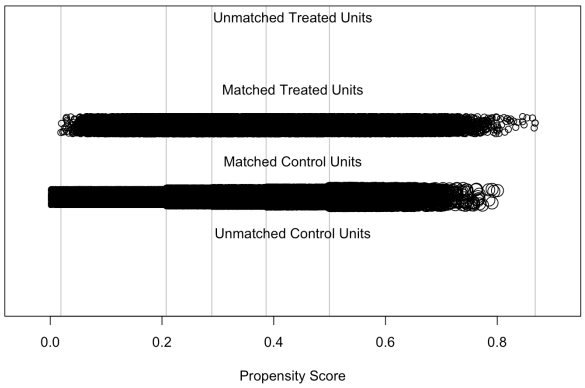

図-10 傾向スコアの分布を確認

図-10 に示すように，区間マッチング後の傾向スコアの分布を確認する．グラフの縦に入る薄い線が，層を区分している．処置群数に応じて均等に分布する 5 つの階層は，各階層間の分布確率が不均等であることがわかる．そして，傾向スコアが 0.2～0.5 の範囲にある処置群のグリッド数が多く，この層別化の確率密度は相対的に小さいことがわかる．逆に，傾向スコアが極端に小さいグリッドの数は少ないので，もしこれらを含めると層の範囲は大きくなる．傾向スコアが極端に小さいために層別から除外された対照群グリッドがある．それに伴い，処置群には，同様のスコアの対照群グリッドがない高スコアの層で，非常に高い傾向スコアを持つグリッドがあることがわかる．

図-11 に示したバランスチェックによると．全グリッドに比べてマッチング後の方で標準化平均差が小さくなっており，区間マッチングによってバランスが改善されていることがわかる．また，前述のように coast_land や golf_land などの共変量は標準化平均値の差を大きくしているものもあったが，この 2 つの共変量は傾向スコアへの影響が小さいため，ここで個別に議論する必要はない．しかし，共変量 diatance_ic（IC までの距離）は 0.25 という緩和基準を満たすものの，0.1 未満という厳密な基準はまだ満たしていないことがわかる．このため，この区

間マッチングで得られる結果は，依然として高いバイアスがかかっている．

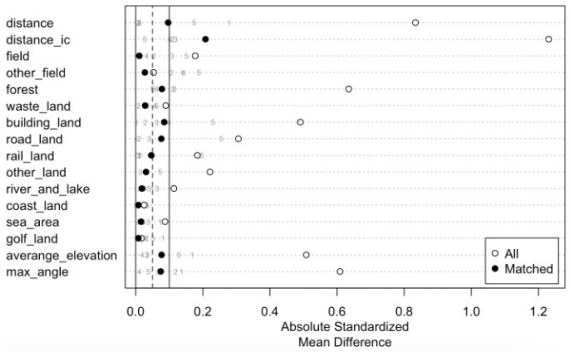

図-11　全グリッドと区間マッチング後の標準化平均差の比較

最近傍マッチング

続いて，最近傍マッチングのバランスチェック結果を**図-12** に示す．

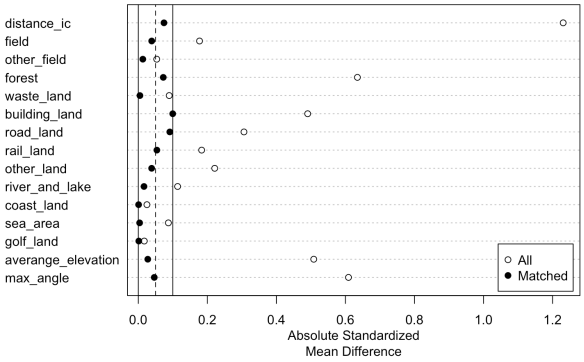

図-12　最近傍マッチング前後の標準化平均差の比較

ここに示すように，最近傍マッチングの結果，すべての共変量の標準化平均差は 0.1 以下という厳しい基準を満足していることがわかる．このことは，最近傍マッチングで得られる結果の精度がかなり高くなることを示している．

連続共変量については，処置群と対照群での各変数の経験分布を比較する QQ プロットも検討できる．QQ プロットは，処置群における変数の分位数と，対照群における対応する分位数を比較する．もし 2 つのグループが同じ経験的分布を持つなら，すべての点は 45 度線上にあることになる．

図-13 に示す QQ プロットから明らかなように，マッチング後は全点の分布が 45 度線に近づき，ほぼ 0.1 破線内に収まり，共変量のバランスも大きく改善されていることがわかる． ただし，傾向スコアが高い 0.1 破線の外側に分布している共変量（road_land，rail_land など）もある．これは**図-10** で分析した内容と一致している．これは，最近傍マッチングでは 2 つのペアの傾向スコアの最大許容差に制約がないためである．傾向スコアの高い処置群グリッドの中には，同じように傾向スコアの高い対照群グリッドがないため，傾向スコアが比較的低い対照群グリッドとのマッチングで共変量がアンバランスになった問題がある．

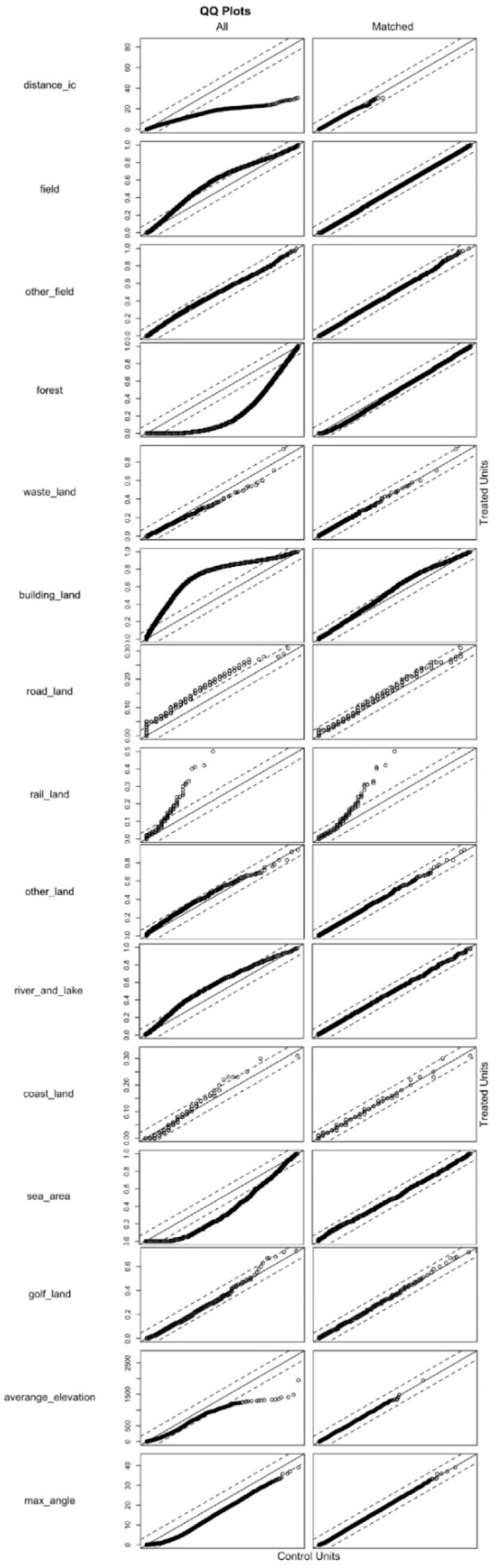

図-13　最近傍マッチング共変量の QQ プロット

キャリパー・マッチング

さらにマッチングのバランスを向上させるため，最近傍マッチングにキャリパーを追加してキャリパー・マッチングを行った．ここでは閾値を 0.1 に設定し，0.1 以上の距離を持つ低品質のマッチングをすべてフィルタリングして除外する．

表-2 キャリパー・マッチングによるスクリーニングの様子

項目	グリッド数
分析対象の全グリッド数	142980
処置群のグリッド数	36596
マッチングされた数	34640
重みづけされていないマッチング数	34640

表-2 から，元のマッチングでは処置群のグリッドが 36596 個，すなわちマッチング対象のペアが 36596 組あったことがわかる．キャリパー・マッチングの結果，マッチンググリッド数は 34640 グループとなり，1956 組の低品質（距離 0.1 以上）マッチンググリッドが除去された．

(4) 処置効果（因果効果）の推定

a) ATE(Average Treatment Effect)の結果

ATE（平均処置効果）を推定する方法は傾向スコア IPW の 1 つだけであり，傾向スコアモデルはロジスティック回帰モデルを用いて導き出され，適合度を判断する決定係数がないため，ここでは重回帰分析を行い，IPW から推測される結果の正確さをより検証するために，重回帰分析の結果を IPW の結果と比較する．この重回帰分析の決定係数は 0.67 であった．したがって，これは適合度が高く，正確な予測をする重回帰モデルなので，IPW の推定結果はそれと比較して良い．

表-3 ATE の推定結果（単位：人）

分析年	1995	2000	2005	2010	2015
重回帰	382	407	380	442	345
IPW	336	241	211	275	219
差	46	166	169	167	126
グリッドごとの人口の基本統計量					
最小値	1	1	1	0	0
第一四分位	88	82	68	61	25
中央値	232	224	204	191	108
平均値	992	1004	994	999	809
第三四分位	707	716	683	668	442
最大値	30862	31669	31123	32079	31847
差÷平均値	4.6%	16.5%	17.0%	16.7%	15.6%

表-3 に示すように，IPW と重回帰分析の結果を集計し，両者の結果を比較したところ，両者の結果の差は 46～169 人の範囲であった．さらに，各メッシュの人口の基本統計量と合わせて考察し，年別メッシュごとの人口の平均と比較すると，平均人口の 4.6%から 17%の範囲で差分が発生していることがわかる．メッシュでの人口の範囲では，IPW で推測される結果は重回帰分析の適合から得られる結果に近く，4-(2)のバランスチェックの結果と合わせても，IPW で得られる結果は悪くないと言える．

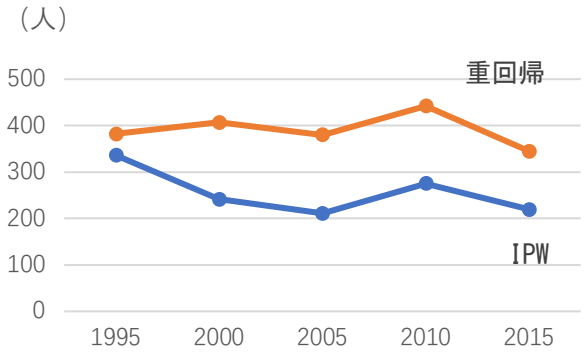

図-14 ATE(Average Treatment Effect; 平均処置効果)の比較

図-14 に示すように，IPW で得られた結果も，重回帰分析で得られた結果と傾向的に一致している．すなわち，新幹線が利用できる地域では，一貫して人口が増加していることがわかる．特に 1995 年は 336 人の増加となり，最大の人口増加を記録した．その後の人口増加は徐々に緩やかになり，2010 年に 275 人の増加で再びピークを迎えた．

b) ATT(Average Treatment Effect for the Treated)の結果

上記の傾向スコア法のうち，重み付け法，層別解析法，マッチング法はいずれも ATT（処置群における平均処置効果）を推定するための方法である．各手法で推定した ATT の結果を以下の**表-3** にまとめた．

表-3 ATT(Average Treatment Effect for the Treated; 処置群における平均処置効果)の推定結果

分析年	1995	2000	2005	2010	2015
重み付け	526	478	462	513	413
層別	685	622	566	613	541
区間マッチング	1579	1369	1402	1439	1233
最近傍マッチング	659	665	620	643	661
キャリパー・マッチング	291	258	225	321	292

（単位：人）

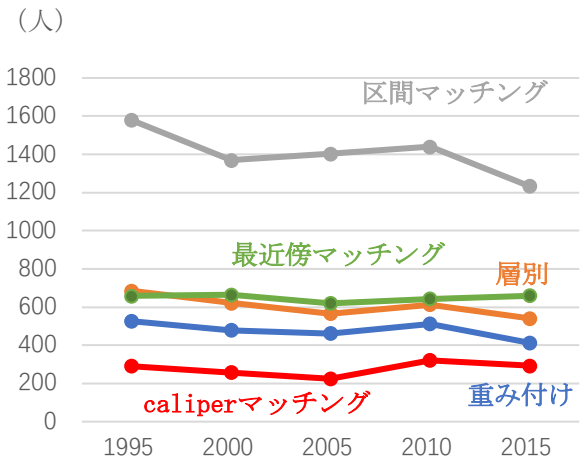

図-15 ATT(Average Treatment Effect for the Treated; 処置群における平均処置効果)推定結果の比較

先のバランスチェックでは，区間マッチングはバランスが悪く（共変量 distance_ic の標準偏差が 0.1 の厳密な範囲外），極端に偏った結果が出ることが分かっている．従って，グラフが示すように，区間マッチングで得られる ATT 結果の数は他の手法に比べて非常に多く，他の手法で得られる結果とは大きく異なることがわかる．

そして，図-11 で見たように，共変量 distance_ic のバランスが 0.1 より大きいため，層別マッチング法に理想的でないが，第 3 章で述べたように層別マッチング法では一般的に 5 層で 90%のバイアスを減らすことができる．そのため，層別マッチング法で得られた結果は，図-15 の緑とオレンジの線で示すように．バランスチェックの結果が非常に良い最近傍マッチング法で得られた結果に非常に近いものとなっている．

Lunceford and Davidian[17] は，層別化によって，様々な重み付け推定量よりもバイアスの大きい平均処置効果の推定値が得られることを実証した．重み付け法のバランスチェックも，すべての共変量が非常によくバランスしており，すべての標準化平均差が厳密な基準である 0.1 以下を満たしていた．したがって，この方法で得られた結果は，層別マッチング法や最近傍マッチング法よりも優れており，すべての方法の中でキャリパー・マッチング法の結果に最も近いものであった．

キャリパー・マッチング法は，最近傍マッチング法に閾値を加えることで低品質のマッチングを除去するため，すべての手法の中で最もバランスチェックの結果が良く，この手法も最も正確な結果を得ることができる．この方法で得られた結果から，新幹線のある地域のメッシュでは，平均 225 人から 321 人まで人口が増加することがわかる． また，人口増加のピークを迎えたのも 2010 年であった．

(5) 様々な方法の比較[26) 27)]

どの手法でもぶつかる問題のひとつが「共有サポート」である．両群の傾向スコアの分布はかなり重なっているはずだが，密度の違いがあるのかもしれない．しかし，場合によっては，分布が完全に重ならないこともある．例えば，対照群に含まれる多くのグリッドは，処置群のすべてのグリッドとは非常に異なっており，ATT を推定する際の比較対象としては不適当である可能性がある．キャリパー・マッチングでは，共有サポートエリア内（または近傍）にあるグリッドのみが自動的に使用される．そのため，これらの方法を使用する際には，共有サポート領域内のグリッドに明示的に限定して分析するよう注意が必要である．

重み付けアプローチの潜在的な欠点は，Horvitz-Thompson 推定と同様に，重みが極端な場合（すなわち，推定された傾向スコアが 0 または 1 に近い場合），分散が非常に大きくなる可能性があることである．

層別解析では，通常，各層で効果を推定し，その後，層全体で集計する． ATT は，各層の推定値を各層における処置グリッドの数で重み付けして推定されるが，特に処置群のグリッドが少ない場合には，各層でかなりのアンバランスが生じる可能性もある[13)]．

全部のデータを用いる重み付けと層別法とは対照的に，最近傍マッチングの明らかな欠点は，必ずしもすべてのデータを使うわけではないことで，対照群の一部のグリッド，さらには傾向スコアが処置群のスコアの範囲内にある一部のグリッドは捨てられ，分析に使われない．そのため，当然ながら推論の検出力の低下につながる．しかし，処置群の大きさはそのままで，対照群の大きさだけを小さくした場合，実際には全体の検出力はそれほど低下しない可能性がある[15)]． Smith[17)]は，1:1 マッチングからの推定値が線形回帰からの推定値よりも標準偏差が低いという例を示している．したがって，この方法を用いる場合は，処置群の大きさを一定に保つよう注意する必要がある．

最後に，ATE を推定する場合，一般的には IPW 法が良い選択となる．ATT が処置群より大きな対照群のグリッドなしで（あるいは多くのグリッドなしで）推定される場合，適切な選択は，一般に，層別と重み付けである．ATT が推定され，対照群が処置群より多い場合，置換なしの 1:1 キャリパー最近傍マッチングは，単純でよく機能するので，良い選択である[15)]．

傾向スコアには他にも様々なマッチング方法があり，応用研究者がそれらを選択するための指針はほとんどない．これまでの主な推奨は，最適なバランスを生み出す方法を選択することであった．しかし，最適なバランス

を定義するには，複数の共変量のバランスとトレードオフする必要があるため，複雑である．本稿では，最大数の共変量で標準化平均値の差が最小になる原則に着目して，様々な方法を分析した．これは，傾向スコア・マッチング法を適用する際の手法の選択について，後続の研究に情報を提供するだろう．

5. おわりに

本研究では，新幹線が沿線地域の人口変動に与える影響を実証的に示すことを目的に，傾向スコア・マッチング法を用いて因果効果を推定した．

IPW法を用いてATEを推定し，その結果を重回帰の結果と比較したところ，IPW 法が高い精度を持つことがわかった．IPW 推定量による新幹線駅の整備効果から見ると，新幹線沿線地域の人口に対してプラスの影響を与えている．

その上で，重み付け法，層別解析と様々なマッチング方法を用いて ATT の推定を行った． 推計された ATT の結果によると，1995 年から 2015 年にかけて，新幹線のある地域で人口が増加していることがわかる．様々な手法の中で，重み付け法とマッチング法のキャリパー・マッチング法が最も良い結果を出し，結果の精度も高かった．

本稿では，交通分野のインフラ整備の事後評価に，因果分析の統計手法である傾向スコア法を適用した．今後の課題として，傾向スコア・マッチング法と差の差分法を組み合わせ，本研究に関する因果推論をさらに洗練させることが期待される．

REFERENCES

1) 織田澤利守，大平悠季：交通インフラ整備効果の因果推論:論点整理と展望，土木学会論文集 D3 (土木計画学), Vol.75, No.5 (土木計画学研究・論文集第 36 巻), I_1-I_15, 2019.
2) 中澤渉：通塾が進路選択に及ぼす因果効果の異質性―傾向スコア・マッチングの応用，教育社会学研究第 92 集(2013)
3) Wang Huiling，Kong Rong：Does Formal Lending Promote Rural Households' Consumption? An Empirical Analysis based on PSM Method , *The Chinese Economy*,51(1): 97-114.
4) Cheng Gang, Du Sihui, Yu Qian：Are Your Educational Expectations Rational? Research on Effects of Parent-Child Discrepancies in Educational Expectations on Academic Performance, *JOURNAL OF EAST CHINA NORMAL UNIVERSITY Educational Sciences* No.1 2022
5) 村上暢子，山田雄二：傾向スコア・マッチング法を用いた 買収による生産性改善効果の検証，ジャフィー・ジャーナル 第 17 巻 67–75 2019 年 4 月
6) 松永卓也，山口修司:整備新幹線の開業効果に ついて，土木計画学研究・講演集，Vol.33,CD-ROM,No.357.
7) 中川大，西村嘉浩，波床正敏：鉄道整備が市町村人口の変遷に 及ぼしてきた影響に関する実証的研究，土木計画学研究・論文集 No.111993 年 12 月
8) Talebian, A., Zou, B., & Hansen, M. ：Assessing the impacts of state-supported rail services on local population and employment: A California case study, *Transport Policy*,63,108–121.
9) 室祥太朗，寺部慎太郎，栁沼秀樹，田中皓介，康楠：新幹線駅の利便性に着目した地方自治体における統計指標の経年変化，土木学会論文集 D3(土木計画学)， Vol.75,No.3,128-138,2019.
10) 片岡将，柳川篤志，田中皓介，川端祐一郎，藤井聡：全国新幹線整備が国土構造と国民経済にもたらす影響の計量分析，土木学会論文集 D3 (土木計画学)， Vol.75,No.5 (土木計画学研究・論文集第 36 巻), I_375-I_386,2019.
11) Shanming Jia, Chunyu Zhou, Chenglin Qin.: No difference in effect of high-speed rail on regional economic growth based on match effect perspective? *Transportation Research Part A* Volume 106,December 2017,Pages 144-157
12) 矢内勇生:計量経済学応用, 講義資料：5. 傾向スコア．https://yukiyanai.github.io/
13) Cochran, W. G.: The effectiveness of adjustment by subclassification in removing bias in observational studies, *Biometrics*,24,295–313.
14) Rosenbaum, P. R. & Rubin, D. B. (1984): Reducing bias in observational studies using subclassification on the propensity score, *Journal of the American Statistical Association*,79,516–524.
15) Peter C. Austin. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies, *Multivariate Behavioral Research*, 46:399–424,2011
16) Elizabeth A. Stuart.: Matching methods for causal inference: A review and a look forward, *StatSci*,2010 February 1; 25(1): 1–21.doi:10.1214/09-STS313.
17) Lunceford, J. K. & Davidian, M. (2004): Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study, *Statistics in Medicine*,23,2937– 2960.
18) Smith H.: Matching with multiple controls to estimate treatment effects in observational studies, *Sociological Methodology* 1997;27:325–353.
19) Ho DE, Imai K, King G, Stuart EA.: Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference, *Political Analysis* 2007;15(3):199–236.
20) Austin, P. C.: Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score， *Pharmacoepidemiology and Drug Safety*,17,1202–1217.doi:10.1002/pds.1673.2008
21) Austin, P. C.: Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples, *Statistics in*

Medicine,28,30833107. doi:10.1002/sim.3697.2009
22) Austin, P. C.: Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies, *Pharmaceutical Statistics*, DOI: 10.1002/pst.433.2011
23) Qin, Yu: 'No county left behind?' The distributional impact of high-speed rail upgrades in China, *Journal of Economic Geography*, Vol.17, pp.489-520,2017.
24) M. Alan Brookhart1, Sebastian Schneeweiss1, Kenneth J. Rothman1, 2, Robert J. Glynn1, 3, Jerry Avorn1, and TilStu¨rmer1.: Variable Selection for Propensity Score Models, *American Journal of Epidemiolgy*, Vol.163, No.12 Advance Access publication April 19, 2006
25) Alberto Abadie And Guido W. Imbens.: Matching on the Estimated Propensity Score, *Econometrica*, Vol.84, No.2(March,2016),781–807
26) B. B. Hansen.: The essential role of balance tests in propensity-matched observational studies: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, *Statistics in Medicine*.Statist.Med.2008; 27:2050–2054
27) Keisuke Hirano, Guido W. Imbens, And Geert Ridder. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score, Econometrica, Vol. 71, No. 4 (July 2003), 1161–1189
28) Marco Caliendo, Iza Bonn, Sabine Kopeinig.: Some Practical Guidance for The Implementation of Propensity Score Matching, *Journal of Economic Surveys* (2008) Vol.22, No.1, pp.31–72

The Causal Effect of the Shinkansen on Population Change in Areas along the Line: An Application of Propensity Score Method

Jingyuan WANG, Shintaro TERABE, Hideki YAGINUMA, Haruka UNO and Yu SUZUKI

Japan's Shinkansen train system has been under construction for about 60 years since it started service in 1964. The Shinkansen has promoted the movement of people and has had a significant impact on the demographic change of the areas along the Shinkansen lines. This study aims to analyze the population change along the Shinkansen line using the tertiary grid data of the National Population Census. Specifically, we estimate the causal effects of Shinkansen stations on population change in the surrounding areas between 1995 and 2015 using the propensity score method. In the analysis, we calculate the ATT (average treatment effect in the treatment group) using the weighting, stratification, and matching methods of the propensity score method, respectively. In particular, the propensity score matching method calculates the ATT using subclassification matching, nearest neighbor matching, and caliper matching, respectively, and the results are compared and analyzed. Finally, the IPW (inverse probability weighting) method is used to calculate the ATE (average treatment effect), which is compared with the results obtained by multiple regression analysis.