機械学習による多産業を対象とした地域間交易額の同時推計

内田　瑞生[1]・杉本　達哉[2]・村上　大輔[3]・高山　雄貴[4]

[1]非会員　八千代エンジニヤリング株式会社（〒111-8648 東京都台東区浅草橋 5-20-8 CS タワー）
E-mail: mz-uchida@yachiyo-eng.co.jp (Corresponding Author)

[2]正会員　八千代エンジニヤリング株式会社（〒111-8648 東京都台東区浅草橋 5-20-8 CS タワー）
E-mail: tt-sugimoto@yachiyo-eng.co.jp

[3]正会員　統計数理研究所准教授　データ科学研究系（〒190-8562 東京都立川市緑町 10-3）
E-mail: dmuraka@ism.ac.jp

[4]正会員　東京工業大学教授　環境・社会理工学院（〒152-8552 東京都目黒区大岡山 2-12-1 M1-20）
E-mail: takayama.y.af@m.titech.ac.jp

本研究は，地域間産業連関表の作成に必要となる多産業の地域間交易額の精度向上に有効なモデルおよび特徴量を明らかにすることを目的とする．そこで，重力モデルおよび代表的な機械学習手法である GBDT および elastic net を対象とした精度検証を行う．さらに，地域間産業連関表の作成時に必要となる移出入額の制約条件に基づくバランス調整が精度に与える影響を把握する．最後に，これらの精度検証を踏まえて，地域間産業連関表の作成に適した機械学習手法や特徴量の選定方法およびバランス調整の方法について考察を行う．

Key Words: *interregional trade, interregional input-output table, machine learning, gravity model*

1. はじめに

近年，国の進める地方創生やエビデンスに基づく政策立案（EBPM）の推進に伴い，地域産業連関表の活用ニーズが高まっている．特に，地域間での財の取引関係をまとめた地域間産業連関表は，地域間での経済波及効果の分析や空間的応用一般均衡（SCGE）モデルによる社会資本整備評価などに活用されてきた．そのため，地域経済や社会資本整備の持続可能性を高めるうえで，地域間産業連関表の果たす役割は大きいと考えられる．一方で，高精度な地域間産業連関表の作成は，多大な調査・作成コストを要するため，現在，多くの都道府県や市町村は地域間産業連関表を作成していない．

通常，地域間産業連関表の作成では，既存の地域内産業連関表および産業別の地域間交易額の 2 種類のデータが必要となる．このうち，地域間交易額については，全国貨物純流動調査（物流センサス）などの実績値に基づいて設定することが望ましいとされている[1]が，市町村などの詳細な地域区分では充分なデータ量を確保できない可能性が高い．こうした場合には，多産業を対象とした地域間交易額の推計が必要となる．

地域間交易額の代表的な推計手法として，重力モデルが存在し，複数の先行研究[2,3,4]で用いられている．重力モデルには，以下に示す 2 つの課題が存在する．1 つ目は，対数線形モデルに属する重力モデルでは，特徴量間の非線形関係や相互作用などを柔軟に取り入れることが難しいことである．2 つ目は，説明変数が多い場合に過学習が生じうることである．

これらの課題を解決しうる機械学習手法として，勾配ブースティング決定木（GBDT）や elastic net が提案されている．これらの手法を用いることで，様々な特徴量の活用による精度向上や交差検証に基づく過学習の抑制が期待される．しかし，これまでこれらの手法による地域間交易額の精度向上効果については十分に検証されておらず，精度向上に寄与する特徴量についても十分な検討が行われてこなかった．

さらに，これらの手法によって推計された地域間交易額は，通常，既存の地域内産業連関表の移出入額と整合しない．そのため，地域間産業連関表の作成では，通常，事前にバランス調整を行う必要がある．しかし，これまでバランス調整が地域間交易額の精度に与える影響については，十分に検証されてこなかった．

これらを踏まえて，本研究では，地域間産業連関表の作成に必要となる多産業の地域間交易額の精度向上に有効なモデルおよび特徴量を明らかにするため，重力モデルおよび代表的な機械学習手法である GBDT，および elastic net を対象とした精度検証を行う．そこで，まず日本の 9 地方間交易額により学習したモデルを用いて都道府県間交易額を推計し，地域間交易額の実績値による精度検証を行う．さらに，バランス調整の前後での精度の変化について調べる．これらの精度検証を踏まえて，地域間産業連関表の作成に適した機械学習手法，特徴量の選定方法およびバランス調整の方法について考察する．

本稿の構成は次のとおりである．まず，第 2 章では，実績値やモデルに基づく地域間交易額の推計方法およびバランス調整方法について整理し，推計上の課題および本研究の位置づけを示す．次に，第 3 章で，精度検証の方法，検証データ・学習データ（特徴量）の作成方法およびモデルの学習方法についてまとめる．さらに，第 4 章で，各モデルの精度比較およびバランス調整が精度に与える影響に関する分析を行い，第 5 章に，本研究のまとめと今後の課題を述べる．

2. 地域間交易額の推計方法と本研究の位置づけ

(1) 地域間産業連関表と地域間交易額の関係

はじめに，地域間産業連関表の作成における地域間交易額の役割を整理する．

表-1 は，一般的な地域間産業連関表を表したものである．地域間産業連関表では，本来，財の供給地・需要地別に中間需要額x_{ij}^{rs}および最終需要額f_i^{rs}の詳細な取引額を設定する必要があるが，こうしたデータ入手は困難である [5]．そのため，実際の推計では，既存の地域内産業連関表と地域間交易額から地域間産業連関表を作成可能な Chenery-Moses モデル [5]が用いられることが多い．

Chenery-Moses モデルでは，式(1)に示す交易係数$\hat{t}_i^{rs}$が用いられる．ここで，$\hat{t}_i^{rr}$は，地域rにおける産業iの自給率に一致し，$\sum_r \hat{t}_i^{rs} = 1$が成立する．さらに，交易係数$\hat{t}_i^{rs}$を用いて，地域間産業連関表の中間需要額$x_{ij}^{rs}$および最終需要額$f_i^{rs}$が，それぞれ式(2)および式(3)により推計される．式(1)の分母の域内需要額d_i^rは，あらかじめ地域内産業連関表で与えられるため，Chenery-Moses モデルでは，産業別の地域間交易額t_i^{rs}を推計することにより地域間産業連関表の作成が可能となる．

$$\hat{t}_i^{rs} = \frac{t_i^{rs}}{d_i^s} \tag{1}$$

t_i^{rs}: 地域rから地域sに移出される産業iの交易額

d_i^s: 地域sにおける産業iの域内需要額

$$x_{ij}^{rs} = \hat{t}_i^{rs} x_{ij}^{s} \tag{2}$$

$$f_i^{rs} = \hat{t}_i^{rs} f_i^{s} \tag{3}$$

(2) 実績値に基づく地域間交易額の推計方法

地域間交易額の設定は，地域間交易に関連する何らかの実績値に基づいて行うことが望ましいとされており，宮城ら [1]では，物流センサス，通話トラフィックデータ，居住地・従業地別 15 歳以上従業者数や旅客純流動調査などを用いて地域間交易額が設定されている．

地域間交易額の設定では，上記の地域間交易関連データが交易額に比例するという仮定のもとで推計を行うのが一般的であるが，こうした仮定に加えて，既存の地域内産業連関の移出入額と整合させる必要がある．そのため，地域間産業連関表の作成では，地域間交易額の実績値が直接使用されることは少なく，通常，何らかの実績値を初期値としたバランス調整が行われる．

また，これらの地域間交易関連データには，市町村間などの詳細な地域のデータが入手できないものや十分なデータ量を確保できないものが含まれる．たとえば，全国幹線旅客純流動調査は，都道府県をまたぐ旅客流動を対象としているため，市町村間データを取得することができない．また，物流センサスは，3 日間調査であるため，市町村などの詳細な地域区分では，十分なデータ量を確保することは難しいと考えられる．

こうした課題への対処として，重力モデルにより地域間交易額の初期値を推計する方法が存在する．

(3) 重力モデルによる地域間交易額の推計方法

重力モデルは，これまで貿易額や人口などの様々な地域間流動データの推計に活用されてきた．

重力モデルは，地域rから地域sへの地域間交易額（バランス調整前の初期値）を$t_{rs}^{(0)}$と表すと，通常，式(4)で表される．ただし以下では，産業iの添え字を省略する．何らかの学習データにより，式(4)のパラメータを推定すれば，移出入額や輸送距離から地域間交易額の初期値$t_{rs}^{(0)}$を推計することが可能となる．

$$E\left[t_{rs}^{(0)} | u_r, n_s, d_{rs}\right] = \beta_0 u_r^{\beta_1} n_s^{\beta_2} d_{rs}^{\beta_3} \tag{4}$$

d_{rs}: 地域rから地域sまでの輸送距離

$\beta_0, \beta_1, \beta_2, \beta_3$: パラメータ

表-1　地域間産業連関表（2 地域r, s・2 産業）

供給＼需要			中間需要				域内最終需要		輸出	輸入	生産額
			地域r		地域s		地域	地域			
			産業 1	産業 2	産業 1	産業 2	r	s			
中間投入	地域r	産業 1	x_{11}^{rr}	x_{12}^{rr}	x_{11}^{rs}	x_{12}^{rs}	f_1^{rr}	f_1^{rs}	e_1^r	$-m_1^r$	x_1^r
		産業 2	x_{21}^{rr}	x_{22}^{rr}	x_{21}^{rs}	x_{22}^{rs}	f_2^{rr}	f_2^{rs}	e_2^r	$-m_2^r$	x_2^r
	地域s	産業 1	x_{11}^{sr}	x_{12}^{sr}	x_{11}^{ss}	x_{12}^{ss}	f_1^{sr}	f_1^{ss}	e_1^s	$-m_1^s$	x_1^s
		産業 2	x_{21}^{sr}	x_{22}^{sr}	x_{21}^{ss}	x_{22}^{ss}	f_2^{sr}	f_2^{ss}	e_2^s	$-m_2^s$	x_2^s
粗付加価値			v_1^r	v_2^r	v_1^s	v_2^s					
生産額			x_1^r	x_2^r	x_1^s	x_2^s					

重力モデルのパラメータ推定では，伝統的に，式(4)に対数変換を行った式(5)に対して最小二乗法（OLS）が適用されてきた．しかし，こうしたパラメータ推定では，Jensen の不等式，すなわち$E[\ln y] \neq \ln E[y]$となることに起因した推定バイアスが生じることが知られている．さらに，特定の地域間において交易額が生じないことは珍しくないが，式(5)では，直接，ゼロ値のデータを目的変数として用いることができない．そのため，データの除外や微小量を付加するといった対処がなされてきたが，こうした処理も推定バイアスの原因となりうる．

$$\ln t_{rs}^{(0)} = \ln \beta_0 + \beta_1 \ln u_r + \beta_2 \ln n_s + \beta_3 \ln d_{rs} + \varepsilon_{rs} \quad (5)$$
$$\varepsilon_{rs}: \text{誤差項}$$

Silva ら[6]は，こうした課題を解決する方法として，式(5)を対象としたポアソン疑似最尤推定法（PPML）によるパラメータ推定を推奨している．PPML では，条件付き期待値と目的変数の条件付き分散が比例関係をもつという仮定のもとでパラメータ推定を行う．PPML を用いることで，OLS 推定の課題である対数変換に伴う推定バイアスの解消およびゼロ値を含む目的変数に対するパラメータ推定が可能となる．さらに，通常のポアソン回帰と異なり，PPML では，目的変数として交易額などの連続変数を用いることができるという利点が存在する．

一方で，対数線形モデルに属する重力モデルには，特徴量間の非線形関係や相互作用などを柔軟に取り入れることが難しいという課題や，説明変数が多い場合に過学習が生じるという課題があることが知られている．

(4) 機械学習による地域間交易額推計の可能性と課題

近年，こうした課題を緩和しうる様々な機械学習手法が用いられるようになっている．これらの手法の多くは学習に関するハイパーパラメータをもつが，交差検証によって，これらのハイパーパラメータをチューニングすることで過学習の抑制や精度向上を可能とする．

代表的な機械学習手法である勾配ブースティング決定木（GBDT）は，複数の決定木のアンサンブルを用いることで，線形モデルなどでは扱いづらい非線形関係や相互作用を柔軟に扱うことが可能となる．そのため，多くのタスクにおいて，他の手法よりも高精度となることが知られている[7]．また，GBDT には，欠損値に対する補完が必要ではないという利点もある．一方で，GBDT は外挿に向かないことが知られており，外挿では線形モデルのほうが精度が高くなる場合がある．

重力モデルと同様，一般化線形モデルに属する Lasso・Ridge・elastic net は，偏回帰係数の大きさを抑制する正則化項を用いることで，説明変数が多い際に問題となる過学習を防止することができるモデルである[8]．正則化項として L1 ノルムを用いる Lasso では係数がゼロとなりやすいため変数選択に適している一方で，L2 ノルムを用いる Ridge では共線性をもつ変数の係数が一定となりパラメータ推定が安定的であるという利点がある．さらに，L1 ノルムと L2 ノルムを組み合わせた elastic net は，Lasso と Ridge の利点を兼ね備えたモデルであるといえる．

(5) 地域間交易額のバランス調整方法と課題

これまでに紹介した地域間交易関連の実績値，重力モデルおよび機械学習手法で推計される地域間交易額$t_{rs}^{(0)}$は，いずれも既存の地域内産業連関表の移出入額と整合しないことがほとんどである．

そのため，$t_{rs}^{(0)}$に基づく何らかのバランス調整が行われることが多い．以下では，代表的なバランス調整方法とその課題をまとめたうえで，本研究における検証方針について述べる．

地域間産業連関表のもととなる既存の地域内産業連関表では，域内需要額のうち自地域で賄う金額を把握可能である．具体的には，これらの金額は，式(6)に示す地域rにおける内々の交易額t_{rr}として算出される．

$$t_{rr} = (x_r - e_r) - u_r = (d_r - m_r) - n_r \quad (6)$$
$$u_r: \text{地域}r\text{における移出額}$$
$$n_r: \text{地域}r\text{における移入額}$$

そのため，地域間産業連関表の作成では，財の供給地rと需要地sの異なる内外の交易額t_{rs}のみを推計すればよい．ここで，移出額$u_r = \sum_r t_{rs}$および移入額$n_s = \sum_r t_{rs}$が成り立つため，移出入額と地域間交易額の関係は，**表-2** の分割表のように表される．したがって，地域間産業連関表の作成では，財の供給地rと需要地sの異なる内外の交易額t_{rs}のみを推計すればよい．ここで，移出額$u_r = \sum_r t_{rs}$および移入額$n_s = \sum_r t_{rs}$が成り立つため，移出入額と地域間交易額の関係は，**表-2** の分割表のように表される．

地域間交易額の推計は，**表-2** の分割表の周辺度数にあたる移出額u_rおよび移入額n_sのもとで，地域間交易額t_{rs}に対する非負の解を求める問題として捉えることができる．

表-2　地域間交易額と移出入額の関係（3 地域）

需要 / 供給	地域 1	地域 2	地域 3	移出額
地域 1		t_{12}	t_{13}	u_1
地域 2	t_{21}		t_{23}	u_2
地域 3	t_{31}	t_{32}		u_3
移入額	n_1	n_2	n_3	T

ただし，移出額u_rおよび移入額n_sによっては，**表-2** の分割表は非負の解をもちえない．そのため通常，事前に移出額u_rおよび移入額n_sを修正する必要がある．**表-2** の分割表が非負の解をもつためには，移出額u_rおよび移入額n_sが非負であること，および$\sum_r u_r = \sum_s n_s = T$が成り立つ必要がある．また，Erdös ら[9]より，これらの条件および式(7)を満たすことが，解をもつための必要十分条件であることが明らかとなっている[注1]．式(7)は，いずれの地域の移出額および移入額も，それぞれ他の地域の移入額および移出額の総額を上回ってはならないことを表している．

$$\max_r(u_r, n_r) \le T \tag{7}$$

これらの条件を満たす移出額u_rおよび移入額n_sのもとで分割表の度数を推定する手法は複数提案されているが，特に精度の高い手法として Deming ら[10]による IPF（iterative proportional fitting）法が存在する[11]．IPF 法は，産業連関分析で用いられる RAS 法と同等の手法であり実装も容易であるため，地域間交易額の推計に適した手法であると考えられる[注2]．IPF 法は，地域間交易額の初期値$t_{rs}^{(0)}$に基づき，式(8)に示す推計値t_{rs}に関する制約付き最適化問題を解く手法である[11]．この最適化問題は，式(9)に示す繰り返し計算により解くことができる．

$$\begin{aligned} \text{minimize,} \quad & \sum_r \sum_s t_{rs} \ln \frac{t_{rs}}{t_{rs}^{(0)}} \\ \text{subject to} \quad & \sum_s t_{rs} = u_r \\ & \sum_r t_{rs} = n_s \end{aligned} \tag{8}$$

$$\begin{aligned} t'_{rs} &= t_{rs}^{(n-1)} \frac{u_r}{\sum_s t_{rs}^{(n-1)}} \\ t_{rs}^{(n)} &= t'_{rs} \frac{n_s}{\sum_r t'_{rs}} \end{aligned} \tag{9}$$

(6) 本研究の位置づけ

本研究では，重力モデルの課題解決に有効な手段となり得る複数の機械学習手法を対象に，多産業の地域間交易額の精度検証を行った．こうした検証は，従来研究ではほとんど行われておらず，本研究の最大の特長である．さらに，バランス調整が地域間交易額の精度に与える影響や地域間交易額の精度向上に寄与する特徴量を明らかにした点も特長と言える．

本検証結果は，機械学習による多産業の地域間交易額の推計に関する新たな知見となり得る．本知見の活用やさらなる蓄積により，将来的に，これまで作成が困難であった詳細な地域区分での高精度な地域間産業連関表の作成が期待できる．

3. 地域間交易額の精度検証とモデル学習の方法

本章では，まず，多産業の地域間交易額を対象とした精度検証の内容と評価指標を示す．次に，検証データおよび学習データの作成方法について示す．さらに，精度検証に使用したモデルの概要と学習方法を示す．

(1) 精度検証の内容と使用する評価指標

重力モデルなどによる地域間交易額の初期値$t_{rs}^{(0)}$の推計が必要となるのは，通常，交易関連データが存在しない状況においてであるため，学習データの取得先地域とモデル適用先地域が異なることが想定される．

そこで本研究では，重力モデル，GBDT の代表的な学習アルゴリズムである XGBoost[12]および elastic net を対象に，地域細分化におけるパラメータ移転可能性の有無を検証する．精度検証では，**表-2** の分割表の移出額と移入額をそれぞれ独立とした場合の期待度数によるヌルモデルを基準として，バランス調整の前後およびモデル間での精度比較を行う．さらに，従来手法で用いられることの少ない産業構造や輸送経路に関する特徴量でモデルを学習することにより，精度向上に寄与する特徴量を明らかにする．

本研究では，まず，実績値に基づく都道府県間交易額を日本の 9 地方間に集計した交易額によりモデルを学習した．さらに，これらのモデルで産業別の都道府県間交易額を推計し，モデル間の精度比較を行った．ただし，9 地方の地域区分については，2005 年の日本の 9 地域間産業連関表と同様に，**表-3** に示す地域区分を用いた．

また，地域間交易額の評価指標として，平均絶対誤差（MAE）・二乗平均平方根誤差（RMSE）・加重絶対パーセント誤差（WAPE）の 3 種類の評価指標を用いた．

このうち，RMSE は，外れ値の影響を受けやすいという特徴がある．地域間交易額では，地域や産業によって交易額が大きく異なるため，RMSE に加えて MAE を使用することとした．また，割合で誤差を評価可能な評価指標として，平均絶対パーセント誤差（MAPE）や WAPE が知られているが，MAPE では，真値がゼロ値を含む場合に評価できない．そのため，ゼロ値を含むデータの精度評価が可能な WAPE を使用した．

表-3 本研究で用いた 9 地方の地域区分

地方	都道府県
北海道	北海道
東北	青森，岩手，宮城，秋田，山形，福島
関東	茨城，栃木，群馬，埼玉，千葉，東京，神奈川，新潟，山梨，長野，静岡
中部	富山，石川，岐阜，愛知，三重
近畿	福井，滋賀，京都，大阪，兵庫，奈良，和歌山
中国	鳥取，島根，岡山，広島，山口
四国	徳島，香川，愛媛，高知
九州	福岡，佐賀，長崎，熊本，大分，宮崎，鹿児島
沖縄	沖縄

(2) 検証データの作成方法

本研究では，検証データとして，2015 年産業連関表の統合大分類別の都道府県間交易額を用いた．

地域間産業連関表の作成では，地域間交易額を既存の地域内産業連関表の移出入額に整合させる必要がある．そこで，まず，47 都道府県の 2015 年の産業連関表を収集した．ただし，奈良県については 2011 年産業連関表を用いた．また，各都道府県の産業連関表の部門分類はそれぞれ異なっており，統合大分類の部門分類と必ずしも一対一対応していない．そのため，各都道府県の公表する最も詳細な部門分類（いずれも約 100 部門程度以上）の産業連関表に対する按分・集計により，国の統合大分類（37 部門）に一致させた．特に，移輸出または移輸入が分離されていない県については国の移輸入率・移輸出率を用いて分離させる処理を行った．

また，地域間交易額の推計では，都道府県産業連関表の産業別の移出入額が非負条件，合計額の一致条件および式(7)の条件を満たす必要があるため，適宜，移出入額の一部を輸出入額に振り分ける処理を行った．

さらに，宮城ら[1]の示した地域間交易関連データと部門分類の対応関係を参考に，都道府県間交易関連データを収集した．物流センサスの品目と統合大分類の部門（37 部門中 20 部門）の対応関係を**表-4** に示す．ただし本研究では，積載重量データを使用した．また，物流センサスが対応していない残りの 17 部門の地域間交易関連データとの対応関係を**表-5** に示す．

最後に，収集した都道府県間交易関連データに対して，都道府県産業連関表の移出入額を制約条件とするバランス調整を行い産業別の地域間交易額を推計した．以降では，この推計結果を“実績値に基づく都道府県間交易額”と呼ぶ．

表-4　統合大分類の一部の部門（37 部門中 20 部門）と 2015 年物流センサスの品目の対応関係

部門分類	物流センサスの品目
01_農林漁業	麦，米，雑穀・豆，野菜・果物，羊毛，その他の畜産品，水産品，綿花，その他の農産品，原木，薪炭，樹脂類，その他の林産品
06_鉱業	石炭，鉄鉱石，その他の金属鉱，砂利・砂・石材，石灰石，原油・天然ガス，りん鉱石，その他の非金属鉱物
11 飲食料品	動植物性油脂，砂糖，その他の食料工業品，飲料
15 繊維製品	糸，織物，衣服・身の回り品
16 パルプ・紙・木製品	製材，パルプ，紙，家具・装備品，木製品
20 化学製品	原塩，化学薬品，化学肥料，染料・顔料・塗料，合成樹脂，その他の化学工業品
21_石油・石炭製品	重油，揮発油，その他の石油，ＬＮＧ・ＬＰＧ，その他の石油製品，コークス，その他の石炭製品
22 プラスチック・ゴム製品	ゴム製品
25 窯業・土石製品	セメント，生コンクリート，セメント製品，ガラス・ガラス製品，陶磁器，その他の窯業品
26 鉄鋼	鉄鋼
27 非鉄金属	非鉄金属
28 金属製品	金属製品
29 はん用機械	産業機械
30 生産用機械	産業機械
31 業務用機械	精密機械，その他の機械
32 電子部品	電気機械
33 電気機械	電気機械
34 情報通信機器	電気機械
35 輸送機械	自動車，自動車部品，その他の輸送機械
39 その他の製造工業製品	書籍・印刷物・記録物，がん具，文房具・運動娯楽用品，その他の日用品，その他の製造工業品

表-5　その他の統合大分類の部門（37 部門中 17 部門）と地域間交易関連データの対応関係

産業分類	物流	通話	従業	旅客	生産	無し
41 建設						●
46 電力・ガス・熱供給					●	
47 水道						●
48 廃棄物処理						●
51 商業	●					
53 金融・保険					●	
55 不動産			●			
57 運輸・郵便	●					
59 情報通信		●				
61 公務						●
63 教育・研究			●			
64 医療・福祉			●			
65 他に分類されない会員制団体		●				
66 対事業所サービス		●				
67 対個人サービス				●		
68 事務用品						●
69 分類不明						●

※物流：2015 年物流センサスの総積載重量，通話：2015 年通話トラフィック（テレコムデータブック 2017（TCA 編）），
従業：2015 年国勢調査の居住地・従業地別 15 歳以上従業者数，旅客：2015 年全国幹線旅客純流動調査の純流動データ，
生産：2015 年産業連関表の合計生産額，無し：交易がない（自給率 1）と仮定

ただし，**表-5** には，交易なしが仮定される 6 部門（建設，水道，廃棄物処理，公務，事務用品および分類不明）が含まれる．そのため，これらの部門については検証対象外とし，残りの 31 部門を対象とした．

また，IPF 法の初期値$t_{rs}^{(0)}$にゼロ値が多い産業では，一部の地域において地域間交易額と移出入額が整合せず，計算が収束しない場合があった．こうした場合には，収束していない行または列のゼロ値に微小量を足して，再度 IPF 法を適用する処理を行った．

(3) 学習データの作成方法

本研究では，従来の重力モデルと同様に，地域内産業連関表から得られる地域・産業別の移出入額，および輸送距離を説明変数として用いた．さらに，従来の重力モデルで用いられることの少ない産業構造および輸送経路に関する特徴量を学習データに使用することで，機械学習手法の有効性や精度向上に寄与する特徴量に関する検証を行った．

まず，輸送距離については，都道府県庁間における交通網上の最短経路探索に基づき簡易的に設定した．具体的には，国土地理院の Global Map Japan version 2.2 Vector data および国土数値情報の 2014 年空港間流通量データに基づく交通網を使用した．また，地方間輸送距離の算出では，移出入額を重みとする各都道府県庁の座標に対する加重平均により，発地・着地別に座標を設定したのち，都道府県と同様に最短経路探索を行った．

また，地域間距離に関連するデータとして，2015 年物流センサスの都道府県間物流時間および輸送単価を使用した．ただし，全産業に対して網羅的に説明変数を設定するため，いずれも代表輸送機関および品類を合計したデータのみを用いた．

以下では，本研究で用いた産業構造および輸送経路に関する特徴量の設定方法についてまとめる．

a) 産業構造に関する特徴量の設定

産業構造に関する説明変数として，以下の 2 種類の特徴量を作成した．

1 つ目は，財の供給地の移出額および需要地の移入額のコサイン類似度である．こうしたコサイン類似度は，ある地域間の経済的な結びつきの強さに影響を及ぼす可能性がある．

2 つ目は，移出入額の主成分得点である．他の産業の特徴量を活用するため，産業別の移出入額それぞれに対して主成分分析による次元削減を行い，第 5 主成分までを学習に使用した．ただし，主成分分析ではデータの標準化を行い，検証データへの主成分分析の適用時には学習データで使用したものと同等のスケーリングを行った．

b) 輸送経路に関する特徴量の設定

輸送経路に関する説明変数として，前述の交通網上の最短経路で使用されたリンクに関する特徴量を使用することが考えられる．そこで，最短経路で使用されるリンクに対して主成分分析による次元削減を行った．

主成分分析では，何らかの外れ値が存在する場合に特徴量スケーリングが有用であるとされている [12]．最短経路探索におけるリンクの使用の有無はバイナリデータであるため，通常，外れ値とはみなされないが，リンクの出現頻度に偏りが生じている可能性は高い．こうしたことは，テキストマイニングにおいて単語の重要度を算出する際にも問題となることが知られている．

そこで本研究では，テキストマイニングで利用される出現頻度および TF-IDF（term frequency-inverse document frequency）による特徴量スケーリングを行う [13]．これらの手法は，テキストマイニング以外の分野でも用いられている [7]．ただし，出現頻度として，バイナリデータをそのまま用いた値，およびそれらを各経路の使用リンク数で除した値の 2 種類を用いた．これらの手法と TF-IDF を合わせた 3 種類の特徴量スケーリングをそれぞれ行ったのち，移出入額の主成分得点と同様に第 5 主成分までを学習に使用した．

(4) モデルの概要・学習方法

a) ヌルモデル

地域間産業連関表作成時の地域間交易額の推計に特有の処理として，地域間交易額のバランス調整があり，地域内産業連関表から得られる移出額および移入額が制約条件となる．そのため本研究では，これらの制約条件のみを満たし，輸送距離などの他の説明変数を含まないヌルモデルを基準として設けることとする．

表-2 の分割表の周辺度数にあたる移出額と移入額が独立であると仮定すると，地域間交易額の期待度数$\tilde{t}_{rs}$は，式(10)となる．そこで本研究では，式(10)によって地域間交易額を推計するモデルをヌルモデルと呼ぶこととする．このモデルは，自動的に移出額および移入額の制約条件を満たすため，バランス調整は不要である．

$$\tilde{t}_{rs} = \frac{u_r n_s}{T} \tag{10}$$

b) 重力モデル

重力モデルでは，全ての産業の交易額を 1 つのモデルで推定するため，移出入額，輸送距離，および物流センサスの物流時間と輸送単価については，統合大分類別にダミー変数および各説明変数との交互作用項を導入し，産業別に切片および偏回帰係数を設定した．また，これらの変数については，従来手法と同様に対数変換を行った．

ただし，物流センサスの物流時間と輸送単価は，欠損値を含むため，物流時間については最短距離に基づく回帰分析，輸送単価については中央値によりそれぞれ補間を行った．さらに，過学習の発生の有無を確認するため，前述の産業構造および輸送経路に関する特徴量を加えて学習を行った．

また，パラメータ推定については，PPML で行った．具体的には，リンク関数を自然対数ln とした疑似ポアソン分布による一般化線形モデルを用いた．

c) XGBoost

GBDT は外挿に向かないとされているため，学習データの交易額と検証データの交易額において，目的変数および説明変数のとる範囲がおおよそ等しくなることが望ましいと考えられる．一方で，地方間交易額から都道府県間交易額を推計する場合，これらの値の範囲は大きく異なる可能性が高い．そこで本研究では，目的変数である地域間交易額および説明変数である移出入額に対してスケーリングを行うこととする．

そこで XGBoost では，実績値に基づく都道府県間交易額を式(10)に示した地域間交易額の期待度数$\tilde{t}_{rs}$で除した乖離度を目的変数とし，予測時には，推計値に$\tilde{t}_{rs}$を掛ける処理を行った．また，期待交易額の大きさを学習に反映させるため，学習時の重みとして，$\tilde{t}_{rs}$を用いた．移出入額については，地域区分によって値の範囲が変わるため，地域別に全産業に占める割合を算出し，説明変数として用いた．また，産業の違いを反映するため産業ダミーを用いたが，交差検証において十分なデータ量を確保するため，統合大分類ではなくひな形分類（10部門程度）に関するダミー変数を使用した．

表-6 に，9 地方間交易額データに基づく XGBoost のハイパーパラメータのチューニング結果をまとめる．本研究では，評価指標を MAE として，ベイズ最適化によるチューニングを行った．ただし，決定木の個数，学習率（learning rate）および早期終了（early stopping）回数についてはあらかじめ固定した．

d) Elastic net

Elastic net では，重力モデルと同様の説明変数，リンク関数および誤差分布による一般化線形モデルを用いた．ただし，XGBoost と同様に，交差検証において十分なデータ量を確保するため，統合大分類ではなくひな形分類ダミーを使用した．

表-7 に，9 地方間交易額データに基づく elastic net のハイパーパラメータのチューニング結果をまとめる．ここでは，XGBoost と同様に，評価指標を MAE として，ベイズ最適化によるチューニングを行った．$\alpha = 1$より，LASSO と同等なモデルとなった（**表-7**）．

4. 地域間交易額の精度検証の結果

(1) バランス調整前のモデル間精度比較

地域細分化におけるパラメータ移転可能性を検証するため，まず，IPF 法によるバランス調整前の地域間交易額に対する精度検証を行った．

図-1 に，実績値に基づく都道府県間交易額とモデルにより推計した都道府県間交易額（バランス調整前）の関係を示す．**図-1** より，一般化線形モデルに属する重力モデルおよび elastic net では，実績値に基づく都道府県間交易額に対する過大推計が生じる傾向がみられる一方で，ヌルモデルおよび XGBoost ではこうした傾向が弱いことが確認できる．XGBoost において過大推計が生じなかったのは，過大推計が生じづらい期待度数からの乖離度を目的変数とした学習方法や GBDT が内挿に適しているという性質に起因していると考えられる．

表-6　XGBoost のハイパーパラメータのチューニング結果

名称	内容	値
trees	決定木の個数	1,000（固定）
learn_rate	学習率	0.1（固定）
stop_iter	早期終了回数	100（固定）
min_n	ノード分割に必要な ノード内データの最小数	33
tree_depth	決定木の最大深さ	1
loss_reduction	ノード分割に必要な 損失関数の減少量	1.02×10^{-5}
sample_size	各反復における サンプリング割合	0.973
lambda	L2 正則化項	1.16×10^{-9}
alpha	L1 正則化項	0.0099

表-7　elastic net のハイパーパラメータのチューニング結果

名称	内容	値
λ	正則化項	0.00121
α	L1 正則化の割合	1

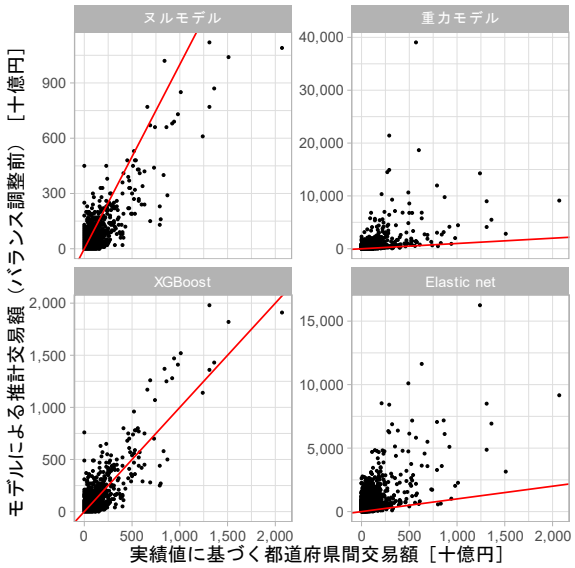


図-1　実績値に基づく都道府県間交易額と
モデル推計交易額（バランス調整前）の関係

また，図-5 に，バランス調整前におけるモデル間での精度比較の結果を示す．図-5 より，ヌルモデルおよび XGBoost の精度は，MAE・RMSE・WAPE においてほぼ同等であり，一般化線形モデルである重力モデルおよび elastic net の精度を大きく上回ることが確認できる．

(2) バランス調整後のモデル間精度比較

次に，IPF 法によるバランス調整後の地域間交易額に対する精度検証を行った．

図-2 に，実績値に基づく都道府県間交易額とモデルにより推計した都道府県間交易額（バランス調整後）の関係を示す．バランス調整により，図-1 でみられた重力モデルおよび elastic net の過大推計が緩和されることが確認できる（図-2）．

図-3 に，バランス調整後におけるモデル間での精度比較の結果を示す．ただし，ヌルモデルの精度はバランス調整前後で同等である．図-3 より，重力モデルおよび elastic net の精度が大幅に改善し，他のモデルの精度を上回ることが確認できる．以上より，バランス調整を行う場合には，一般化線形モデルの精度が高くなることが示唆される．また，いずれの評価指標においても，重力モデルの精度は，elastic net の精度をわずかに上回った．そのため本検証では，機械学習手法による精度向上効果は確認されなかった．

さらに，WAPE により，バランス調整後におけるひな形分類およびモデル別の精度比較を行った（図-4）．図-4 より，鉱業，農林漁業および製造業では地域間交易額の精度が低いことが確認できる．

(3) 地域間交易額に関する説明変数の重要度

次に，地域間交易額の精度向上に寄与する特徴量を明らかにするため，過学習が起こりづらいとされる elastic net の標準偏回帰係数により説明変数の重要度を算出した（表-8）．ただし，ひな形分類ダミーとの交互作用項に対しては，それぞれの絶対平均および平均を用いた要約を行った[注3)]．

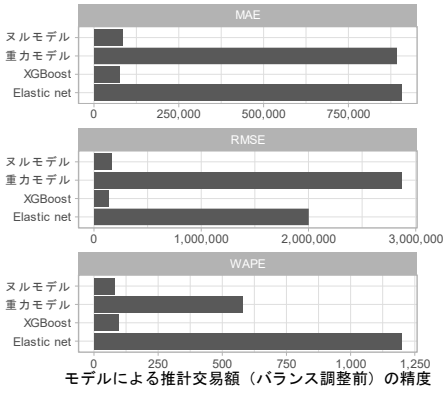


図-5 バランス調整前のモデル間精度比較結果

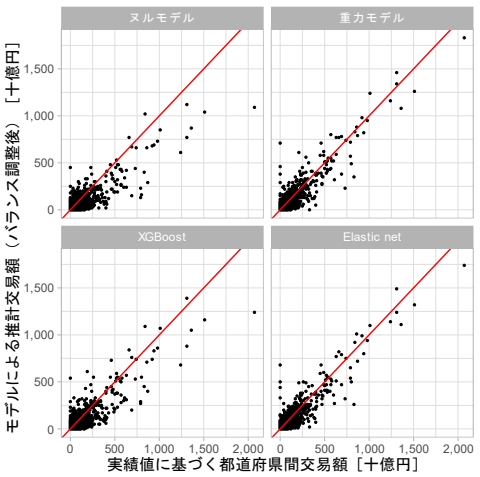


図-2 実績値に基づく都道府県間交易額とモデル推計交易額（バランス調整後）の関係

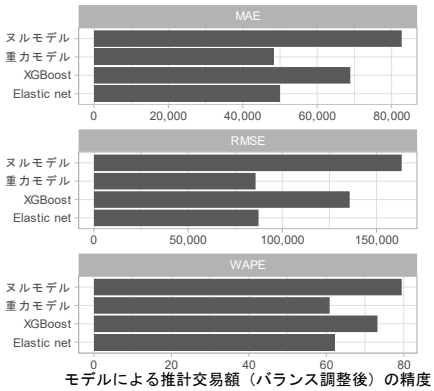


図-3 バランス調整後のモデル間精度比較結果

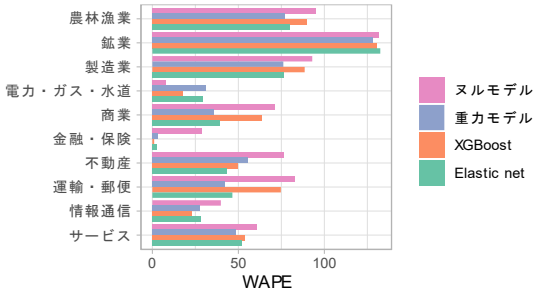


図-4 バランス調整後のひな形分類・モデル別の WAPE

表-8 elastic net の説明変数の重要度（上位 10 変数）

説明変数	標準偏回帰係数	
	絶対平均	平均
移出入額のコサイン類似度	0.918	-0.452
移出額	0.788	0.788
物流時間（物流センサス）	0.771	-0.771
移入額	0.564	0.564
輸送距離	0.401	-0.372
輸送単価（物流センサス）	0.310	-0.287
第 4 主成分_バイナリ_最短経路	0.263	0.263
第 2 主成分_出現頻度_最短経路	0.234	0.234
第 1 主成分_移出額	0.218	-0.218
第 2 主成分_移入額	0.174	-0.174

表-8 より，従来手法でも用いられる移出額，移入額，輸送距離などの重要度が高いことが確認できる．一方で，最も標準偏回帰係数の絶対平均が高くなったのは，従来手法で用いられることは少ない移出入額のコサイン類似度であった．

移出入額のコサイン類似度の標準偏回帰係数の平均は負であるため，全体的には，産業別の移出額と移入額が類似する地域で地域間交易額が減少する傾向があることがわかる．一方で，ひな形分類によって標準偏回帰係数の正負が異なるため，標準偏回帰係数の絶対平均と平均（の絶対値）の値の乖離が大きいことが確認された．

また，産業構造および輸送経路に関する特徴量のうち，主成分分析に基づく次元削減により作成したものについては，他の説明変数と比べて重要度が低くなった．

5. おわりに

本研究では，地域間産業連関表の作成に必要となる，多産業の地域間交易額の精度向上に有効なモデルおよび特徴量を明らかにするため，重力モデル，GBDT の代表的な学習アルゴリズムである XGBoost，および elastic net を対象とした精度検証を行った．また，IPF 法によるバランス調整の前後での精度比較を行った．

検証の結果，バランス調整前では，XGBoost の精度が一般化線形モデルに属する重力モデルや elastic net より高くなるという結果が得られた．一方で，バランス調整後では，一般化線形モデルの精度が XGBoost よりも高くなるという結果が得られた．そのため，IPF 法によるバランス調整を行う場合には，一般化線形モデルが有効である可能性が示された．また，バランス調整後では，機械学習手法（XGBoost および elastic net）の精度が従来手法である重力モデルの精度を上回らないという結果が得られた．

さらに，地域間交易額の推計に有効な特徴量について検討を行い，従来手法で用いられることの多い移出額，移入額，距離関連データ（輸送距離・物流時間），輸送単価に加えて，移出入額のコサイン類似度が有効となる可能性が示された．

本研究では，上記に示す成果を得られたものの，今後，以下に示す 3 つの課題に取り組む必要がある．

1 つ目は，データサイズの多寡による影響把握である．本研究で用いた 9 地方間交易額は約 2,000 行のデータであったため，XGBoost といった GBDT の学習データのサイズとして十分でなかった可能性がある．また本研究では，交差検証において十分な学習データを確保するため，XGBoost および elastic net では重力モデルよりも粗い部門分類の産業ダミーを用いた．こうしたデータ制約は，精度低下につながった可能性がある．そのため今後，さらに詳細な地域区分の地域間交易額を対象とした精度検証を行う必要がある．

2 つ目は，移出入額に関する制約条件を考慮したハイパーパラメータのチューニング・モデル学習である．本研究では，移出入額に関する制約条件を満たすためのバランス調整により精度が大きく変化することが明らかとなったが，機械学習手法のハイパーパラメータのチューニングやモデル学習では，こうしたバランス調整の影響が考慮されていない．そのため今後，移出入額に関する制約条件を考慮したハイパーパラメータのチューニングやモデル学習の可能性を検討する必要がある．

3 つ目は，産業構造に関する特徴量が地域間交易額に与える影響分析である．本研究で高い説明力を有した移出入額のコサイン類似度は，産業構造に関する特徴量の一種と捉えることができる一方で，こうした特徴量が地域間交易額の推計に有効な理由については明らかとなっていない．そのため，産業構造の類似度に関する特徴量が地域間交易額に与える影響について，今後さらに詳細に分析を行う必要がある．

NOTES

注1) Erdös ら [9]は，対角成分が所与の分割表が非負の解をもつ必要十分条件を，3 × 3以上の大きさの分割表を対象に示したものである．ただし，表 4 の分割表のように対角成分がゼロの場合，2 × 2分割表においても Erdös ら [9]の示した条件が必要十分条件となることは明らかである．

注2) 宮城ら [1]では，IPF 法と異なる手法を用いて交易係数が推定されている．しかし，宮城ら [1]の示した中間需要の列部門に関する費用バランス式には，本来含まれるべきではない最終需要額が含まれていることは問題であると考えられる．そのため本研究では，IPF 法のみを用いることとする．

注3) 一般化線形モデルにおける交互作用項の設定では，カテゴリー数より 1 少ない個数のダミー変数との交互作用項を作成することが多い．しかし，この場合，どのカテゴリーを基準とするかによって，標準偏回帰係数の推定結果が異なってくる．そのため本研究では，産業別に交互作用項を設ける説明変数については，もともとの項を削除し，全産業に対して交互作用項を設けた．

REFERENCES

1) 宮城俊彦，石川良文，由利昌平，土谷和之：地域内産業連関表を用いた都道府県間産業連関表の作成，土木計画学研究・論文集，20 巻 1 号，pp. 87-95，2003．[MIYAGI, T., ISHIKAWA, Y., YURI, S. and TSUCHIYA, K.: The Construction of Interregional Input-Output Table at Prefecture Level Using Intraregional Input-Output Tables, *INFRASTRUCTURE PLANNING REVIEW*, Vol. 20, No. 1, pp. 87-95, 2003.]

2) 唐渡広志，山野紀彦，人見和美：電力供給地域に対応する 1995 年全国 10 地域間産業連関表の開発，電力中央研究所報告，Y01019，2002．[Karato, K.,

Yamano, N. and Hitomi, K.: Estimating the inter-regional input-output table of 10 power-supply regions:1995, *CRIEPI Report*, Y01019, 2002.]

3) 中野諭，西村一彦：地域産業連関表の分割における多地域間交易の推定，産業連関，15 巻 3 号，pp. 44-53，2007. [Nakano, S. and Nishimura, K.: Estimation of multi-regional trade in regional input-output table partitioning, *Input-output analysis : innovation & I-O technique*, Vol. 15, No. 3, pp. 44-53, 2007.]
4) 塚本高浩，小見山尚子，根本二郎：愛知県内全市町村間産業連関表の作成と市町村間の空間的相互依存関係に関する分析，産業連関，26 巻 1 号，pp. 1-20，2018. [Tsukamoto, T., Komiyama, N. and Nemoto, J.: Creation of an inter-municipal input-output table for all municipalities in Aichi Prefecture and analysis of spatial interdependencies between municipalities, *Input-output analysis : innovation & I-O technique*, Vol. 26, No. 1, pp. 1-20, 2018.]
5) Miller, R. E. and Blair, P. D.: *Input-Output Analysis: Foundations and Extensions*, Cambridge University Press, 2009.
6) Silva, J. M. C. S. and Tenreyro, S.: The Log of Gravity, *The Review of Economics and Statistics*, Vol. 88, No. 4, pp. 641-58, 2006.
7) 門脇大輔，阪田隆司，保坂桂佑，平松雄司：Kaggle で勝つデータ分析の技術，技術評論社，2019. [Kadowaki, D., Sakata, R., Hosaka, K. and Hiramatsu, Y.: *Data Analysis Techniques to Win Kaggle*, Gijutsu-Hyohron, 2019.]
8) 鈴木譲：スパース推定 100 問 with R，共立出版，2020. [Suzuki, J.: *Sparsity in Statistics with 100 Math & R Problems*, Kyoritsu Shuppan, 2020.]
9) Erdös, P. and Minc, H.: Diagonals of Nonegative Matrices, *Linear and Multilinear Algebra*, Vol. 1, No. 2, pp. 89-95, 1973.
10) Deming, W. E. and Stephan, F. F.: On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known, *Annals of Mathematical Statistics*, Vol. 11, No. 4, pp. 427-444, 1940.
11) Little, R. J. A. and Wu, M. M.: Models for Contingency Tables with Known Margins When Target and Sampled Populations Differ, *Journal of the American Statistical Association*, Vol. 86, No. 413, pp. 87-95, 1991.
12) Chen, T. and Guestrin C.: XGBoost: A Scalable Tree Boosting System, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-94, 2016.
13) Alice Zheng，Amanda Casari（株式会社ホクソエム訳）：機械学習のための特徴量エンジニアリング その原理と Python による実践，オライリー・ジャパン，2019. [Zheng, A.,Casari, A.: *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*, O'Reilly, 2018.]

SIMULTANEOUS ESTIMATION OF INTER-REGIONAL TRADE VALUE OF MULTIPLE INDUSTRIES BY MACHINE LEARNING

Mizuki UCHIDA, Tatsuya SUGIMOTO, Daisuke MURAKAMI and Yuki TAKAYAMA