ゼロカウント切断分布を含む
空間計量混合分布モデルの開発

津田　敏明[1]・塚井　誠人[2]・林　遼平[3]

[1]学生会員　広島大学大学院　工学研究科社会基盤環境工学専攻（〒739-8527広島県東広島市鏡山1-4-1）
E-mail:m121388@hiroshima-u.ac.jp

[2]正会員　広島大学大学院准教授　工学研究院社会環境空間部門（〒739-8527広島県東広島市鏡山1-4-1）
E-mail:mtukai@hiroshima-u.ac.jp

[3]学生会員　広島大学大学院　工学研究科社会基盤環境工学専攻（〒739-8527広島県東広島市鏡山1-4-1）
E-mail:m146193@hiroshima-u.ac.jp

現在入手できる空間データは，集計単位の細密化が進んでいる．一方，従来の空間計量経済モデルは，単一の空間重み行列を用いて簡略的に空間構造を表現しているので，細密集計単位の空間依存性の表現が困難である．本研究では，ゼロカウントデータを含む地域異質性によって生じる統計分析上の問題を解消する，新たな空間計量経済モデルを開発する．具体的には，地域異質性の表現のために定義した標本重み変数を用いて，ゼロカウント切断分布を含む複数の確率分布を混合した空間計量混合分布モデルを定式化する．町丁目単位で集計されたGISデータを用いた実証分析の結果，提案モデルはゼロカウントを含む小地域データに対して精度の高いモデル推計を行えること，ならびに地域異質性を柔軟に表現できることが明らかとなった．

Key Words : *zero-count, areal heterogeneity, spatial econometric model, mixed distribution*

1. はじめに

近年，より詳細な小地域を対象とするきめ細やかな計量分析が求められるようになっている．たとえば，地域メッシュ統計では，従来の基準地域（3次）メッシュ（1km）をさらに細分化した1/2メッシュ（500m）の集計フォーマットが標準化されており，詳細地域分析に必要なデータが整備されつつある．一方で，データの集計単位が詳細になったことにより，非観測地域や観測の秘匿を要する小地域がより多く現れ，ゼロカウントとして公表される地域が増加している．また，地域数の拡大は分析対象地域間の異質性の増大をもたらすため，これに起因する新たな統計分析上の問題も無視できなくなっている．そこでゼロカウントデータを許容するモデリング手法が開発されている．

Lesage and Pace[1)]は，ゼロカウントデータの発生が避けられないODフローデータに対して，地域間の相互依存性を表現するために，空間重み行列を適用したOD空間計量経済モデルを提案した．爲季・堤[2)]は，Lesage and Paceが対数正規重力モデルを定式化していることに対して，1) フロー観測値の対数を用いなければならない，2) 非負制約を満足していない，3) 誤差項が等分散仮定となっている，4) フロー量がゼロの場合への対応が不適切，という4つの問題を指摘して，その解決法としてフロー観測値がポアソン分布に従うと仮定したポアソン重力モデルを提案した．さらに爲季・堤[3)]では，Haining et al.[4)]が提案した固有ベクトル空間フィルタリングによって空間依性を表現する負の二項分布重力モデルを拡張し，ゼロカウントが過剰に発生するデータへの対応が可能なゼロ過剰重力モデルを提案して，実証分析から，提案モデルの有効性を明らかとした．

本研究では，ゼロカウントデータを含む地域異質性の問題を解消するため，複数の確率分布を混合した新たな空間計量経済モデルを開発する．具体的には，ゼロカウント切断分布を含む観測データが複数のデータ生成過程（DGP）から生成されると仮定して，複数の確率分布の混合分布から成る空間計量混合分布モデルを開発する．本研究では，提案モデルの有効性を明らかにするため，実データに基づく実証分析を行い，提案モデルの推計精度と現況再現性を確認する．

2. 空間計量混合分布モデルの定式化

本研究では，地域規模（属性値）が大きな地域に対して「多変量正規分布 1」，中程度の地域に対して「多変量正規分布 2」，小さな地域に対して「負の二項分布」を仮定して，これらの確率分布を混合したモデルを提案する．同モデルでは，各確率分布において異なる空間過程を表現するために，それぞれに異なる空間パラメータを設定する．多変量正規分布 1（SMA 型）（k=1），多変量正規分布 2（SAR 型）（k=2），負の二項分布（SCR 型）（k=3）に関する提案モデルは，それぞれ式(1)~(3)で特定化する．

$$\begin{aligned}\boldsymbol{Y}_1 &= \boldsymbol{X}\boldsymbol{\beta}_1 + \boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon}_1 + \psi\boldsymbol{W}\boldsymbol{\varepsilon}_1 \\ &= \boldsymbol{X}\boldsymbol{\beta}_1 + \boldsymbol{\gamma}_1 + (\boldsymbol{I} + \psi\boldsymbol{W})\boldsymbol{\varepsilon}_1\end{aligned} \tag{1}$$

$$\begin{aligned}\boldsymbol{Y}_2 &= \rho\boldsymbol{W}\boldsymbol{Y}_2 + \boldsymbol{X}\boldsymbol{\beta}_2 + \boldsymbol{\gamma}_2 + \boldsymbol{\varepsilon}_2 \\ &= (\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{X}\boldsymbol{\beta}_2 \\ &\quad + (\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{\gamma}_2 + (\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{\varepsilon}_2\end{aligned} \tag{2}$$

$$\begin{aligned}\boldsymbol{Y}_3 = [\lambda_i] &= \exp\{\boldsymbol{X}\boldsymbol{\beta}_3 + \mu\boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta}_3 + \boldsymbol{\gamma}_3\} \\ &= \exp\{(\boldsymbol{I} + \mu\boldsymbol{W})\boldsymbol{X}\boldsymbol{\beta}_3 + \boldsymbol{\gamma}_3\}\end{aligned} \tag{3}$$

ここで，地点数を N，説明変数の数を K とすると，$\boldsymbol{Y}$：目的変数（N×1），$\boldsymbol{X}$：説明変数（N×K），$\boldsymbol{W}$：空間重み行列（N×N），$\boldsymbol{I}$：単位行列（N×N），$\boldsymbol{\beta}$：構造パラメータ（K×1），$\boldsymbol{\gamma}$：定数項（N×1），$\boldsymbol{\varepsilon}$：i.i.d 仮定を満足する誤差項（N×1），ψ, ρ, μ：空間相関パラメータである．

また，各確率分布の地域ごとの対数尤度関数は，式(4)~(6)で表される．

$$\ln L_{i|1} = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma_1^2) - \frac{1}{N}\ln|\boldsymbol{I} - \psi\boldsymbol{W}| - \frac{\varepsilon_{i|1}^2}{2\sigma_1^2} \tag{4}$$

$$\ln L_{i|2} = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma_2^2) + \frac{1}{N}\ln|\boldsymbol{I} - \rho\boldsymbol{W}| - \frac{\varepsilon_{i|2}^2}{2\sigma_2^2} \tag{5}$$

$$\begin{aligned}\ln L_{i|3} &= \ln\Gamma(y_i + \nu^{-1}) + \nu^{-1}\ln\left(\frac{\nu^{-1}}{\nu^{-1} + \lambda_i}\right) \\ &\quad + y_i\ln\left(\frac{\lambda_i}{\nu^{-1} + \lambda_i}\right) - \ln(y_i!) - \ln\Gamma(\nu^{-1})\end{aligned} \tag{6}$$

ここで，σ：誤差項$\boldsymbol{\varepsilon}$の標準偏差，$\Gamma(\cdot)$：ガンマ関数，ν：過分散パラメータである．

また本研究では，地域異質性を表現するため，式(7)の Logit モデルによって標本重み（地域 i の確率分布 k の重み）π_{ik} を定義する．

$$\pi_{ik} = \frac{\exp(V_{ik})}{\sum_k \exp(V_{ik})} \tag{7}$$

ここで，V_{ik}：地域 i の確率分布 k における重み関数である．標本重みπ_{ik}を用いると，標本帰属指標 z_{ik}（地域 i の確率分布 k への帰属を表す非観測変数）は式(8)で表される．

$$\hat{z}_{ik} = E[z_{ik}] = \frac{\pi_{ik} f_{ik}(x)}{\sum_{l=1}^{K} \pi_{il} f_{il}(x)} \tag{8}$$

ここで，$f_{ik}(x)$：地域 i の確率分布 k における確率密度関数である．

以上より，本研究で提案する空間計量混合分布モデルの対数尤度関数は式(9)で表される．

$$\ln L_{mix} = \sum_i \sum_k \left(z_{ik} \cdot \ln(\pi_{ik} L_{i|k})\right) \tag{9}$$

式(9)の推計には欠測値を推計する Expectation-Step と，得られた欠測値を所与としてモデルパラメータを推計する Maximization-Step を交互に繰り返す EM アルゴリズムを適用する．具体的な推計手順は以下の通りである．

[1] 欠損値である標本帰属指標 z_{ik} に初期値を与え，各地域 i を確率分布 k=1~3 に帰属させる．

[2] 式(7)の Logit モデルの推計により，標本重みπ_{ik}を算出する．

[3] [1]の z_{ik} と[2]のπ_{ik}を用いて，式(9)の対数尤度関数の最尤推定を行う．

[4] 式(8)によって z_{ik}を更新する．

[5] [2]~[4]を収束条件 $L_{mix}^{(o+1)} - L_{mix}^{(o)} < 10^{-6}$ に達するまで繰り返す．

なお，本研究で提案する空間計量混合分布モデルによる予測値は 2 通りの表現が可能である．これらは，推計された標本重み $\hat{\pi}_{ik}$，または標本帰属指標 $\hat{z}_{ik}$ を用いてそれぞれ式(10)，(11)によって表される．

$$\begin{aligned}\hat{\boldsymbol{Y}} &= \sum_k (\hat{\boldsymbol{\pi}}_k \otimes \hat{\boldsymbol{Y}}_k) \\ &= \hat{\boldsymbol{\pi}}_1 \otimes \hat{\boldsymbol{Y}}_1 + \hat{\boldsymbol{\pi}}_2 \otimes \hat{\boldsymbol{Y}}_2 + \hat{\boldsymbol{\pi}}_3 \otimes \hat{\boldsymbol{Y}}_3\end{aligned} \tag{10}$$

$$\begin{aligned}\hat{\boldsymbol{Y}} &= \sum_k (\hat{\boldsymbol{Z}}_k \otimes \hat{\boldsymbol{Y}}_k) \\ &= \hat{\boldsymbol{Z}}_1 \otimes \hat{\boldsymbol{Y}}_1 + \hat{\boldsymbol{Z}}_2 \otimes \hat{\boldsymbol{Y}}_2 + \hat{\boldsymbol{Z}}_3 \otimes \hat{\boldsymbol{Y}}_3\end{aligned} \tag{11}$$

ここで，$\otimes$：クロネッカー積，$\hat{\boldsymbol{\pi}}_k$：確率 k の標本重みベクトル（N×1），$\hat{\boldsymbol{Z}}_k$：確率 k の標本帰属指標ベクトル（N×1），$\hat{\boldsymbol{Y}}_k$：確率分布 k の予測値である．

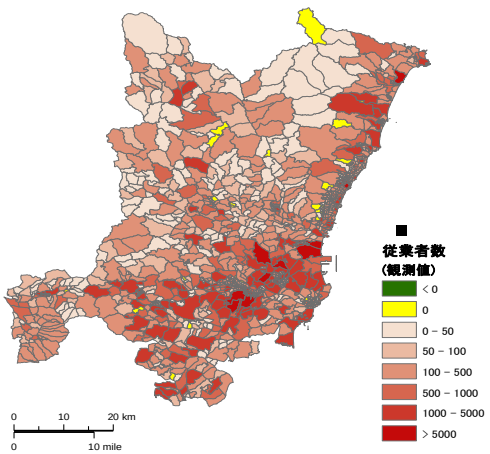

図-2　対象地域における従業者数の空間分布

表-2 モデル推計結果

変数	空間計量混合分布モデル		単一分布モデル							
	推計値	*t*値	推計値	*t*値	推計値	*t*値	推計値	*t*値	推計値	*t*値
標本重み	LOGIT									
建物用地	0.10	** 13.0	———		———		———		———	
都市地域割合	1.82	** 11.5	———		———		———		———	
世帯数	0.01	** 3.1	———		———		———		———	
定数項(k=1)	-3.80	** -21.7	———		———		———		———	
定数項(k=2)	-2.16	** -13.8	———		———		———		———	
確率分布 k=1	SMA		SMA						OLS	
建物用地	13.66	** 3.8	17.75	** 22.3	———		———		17.53	** 21.8
都市地域割合	1297.1	** 3.8	348.8	** 7.5	———		———		365.2	** 10.1
世帯数	2.54	+ 1.8	2.33	** 6.4	———		———		2.32	** 6.1
定数項	745.9	+ 1.9	-195.4	** -5.9	———		———		-191.2	** -7.6
σ_1（標準偏差）	1213.6	** 16.1	559.9	** 50.6	———		———		570.3	** 51.1
ψ（空間相関パラメータ）	0.26	0.9	0.62	** 9.8	———		———		———	
確率分布 k=2	SAR				SAR					
建物用地	4.49	** 8.8	———		17.05	** 21.5	———		———	
都市地域割合	78.50	** 4.4	———		313.4	** 8.6	———		———	
世帯数	0.47	* 2.1	———		2.29	** 6.2	———		———	
定数項	105.5	** 5.4	———		-293.2	** -10.1	———		———	
σ_2（標準偏差）	189.4	** 36.8	———		558.7	** 50.9	———		———	
ρ（空間相関パラメータ）	0.18	** 5.9	———		0.34	** 6.7	———		———	
確率分布 k=3	NB						NB			
建物用地	0.05	** 6.9	———		———		0.03	** 16.7	———	
都市地域割合	0.37	** 2.9	———		———		1.07	** 11.9	———	
世帯数	-9.E-04	-0.4	———		———		-1.E-03	-1.3	———	
定数項	3.11	** 25.4	———		———		4.27	** 38.9	———	
ν（過分散パラメータ）	1.13	** 15.4	———		———		1.35	** 28.4	———	
μ（空間相関パラメータ）	-0.05	-0.5	———		———		0.24	+ 1.9	———	
推計精度指標	$\hat{Y}_\pi$	$\hat{Y}_z$								
Adj. R^2	0.465	0.798	0.546		0.532		0.130		0.547	
RMSE	683.1	382.8	570.4		579.8		8.78.E+04		570.3	
GMI	8.541	1.485	0.689		4.805		0.777		10.399	
予測負値数	0	0	225		236		0		224	
1標本あたり最終対数尤度	-6.829		-7.733		-7.749		-6.564		-7.765	
AIC	1.79.E+04		2.02.E+04		2.03.E+04		1.72.E+04		2.03.E+04	
サンプル数	1307		1307		1307		1307		1307	

+ 10%有意　* 5%有意　** 1%有意

3. GIS データを用いた実証分析

3.1 データの概要

本章では，提案した空間計量混合分布モデルの有効性の検証のために，GIS データを用いた実証分析を行う．分析対象地域は，図-2 に示す茨城県北部である．データの集計単位は町丁目，分析対象地域数 N は 1307 である．モデルの目的変数として，従業者数を用いる．従業者数は，ゼロカウント地域を 51 地点（3.9%）含み，また検証用データセットの全変数においてグローバルな空間的自己相関が存在していることを確認している．

3.2 モデル推計結果

空間計量混合分布モデルは，式(10)，(11)のように 2 通りの再現精度の表現が可能である．本分析では，式(10)，(11)のそれぞれについて推計精度を算出して，それらを比較する．さらに推計精度の比較のため，従来の単一分布モデルも推計する．モデルの目的変数には従業者数，説明変数には建物用地，都市地域割合，世帯数の 3 変数を用いる．空間重み行列 $\boldsymbol{W}$ には，距離変数 d_{ij} を地点間直線距離，空間減衰パラメータを α=2 として，行基準化を行った距離減衰行列を用いる．

表-2 にモデル推計結果を示す．提案した空間計量混合分布モデルの標本帰属指標 z_{ik} の初期値は，試行錯誤の結果，閾値 θ_1=100，θ_2=1000 として，z_{i1}=0，z_{i2}=1，z_{i3}=0，

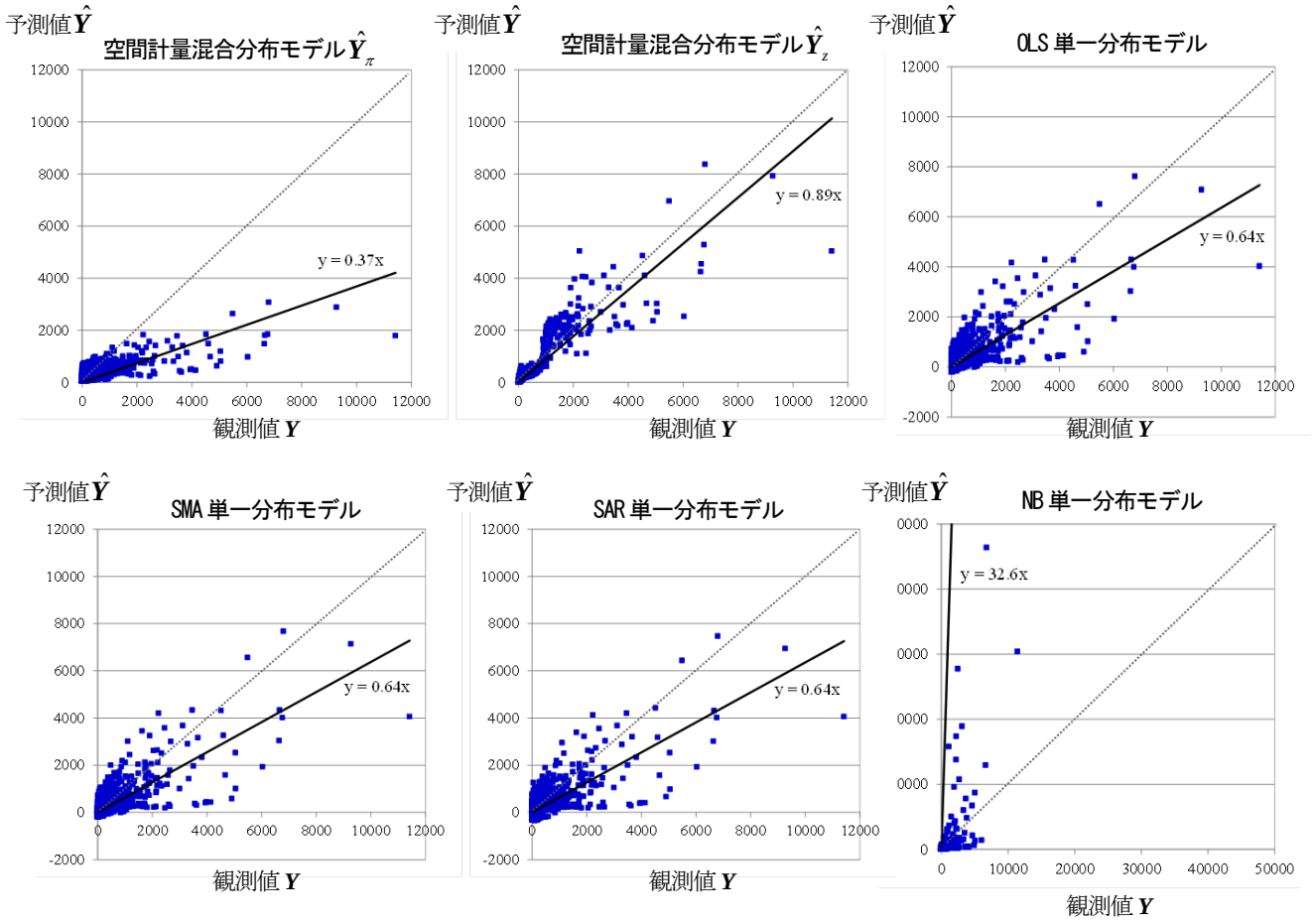

図-3　観測値と予測値の散布図と回帰直線

$y_i \le \theta_1$ のとき z_{i1}=0，z_{i2}=0，z_{i3}=1 となるように与えた．提案した空間計量混合分布モデルでは，確率分布 k=1，及び確率分布 k=3 の空間相関パラメータが有意とならなかった．これは，地域規模（属性値）が著しく大きい，もしくは小さい地域が相互に離れた位置に点在しているため，それらへの帰属確率が高い地域について近隣地域との相互依存関係が弱くなったことが原因と考えられる．また，確率分布 k=3 の世帯数パラメータも有意とならなかったが，その他のパラメータは全て有意となり，また推計されたパラメータの符号も期待された条件を満たしている．

次に，提案モデルと従来モデルの推計精度について考察する．Adj.R^2 は提案モデル $\hat{Y}_z$ が最も大きな値となり，従来モデルと比較して大きく改善されている．RMSE についても，$\hat{Y}_z$ が最も小さな値を示しており，Adj.R^2 と同様に大きく改善されている．一方で，NB 単一分布モデルは両指標において，精度は劣悪である．GMI は，従来モデルでは SMA 単一分布モデル，NB 単一分布モデルにおいて残差の空間的自己相関なしとなっているが，NB 単一分布モデルは再現精度が著しく低いため，結果の信頼性が乏しい点に注意が必要である．

予測負値数は，提案モデルと NB 単一分布モデルでは 0 となっているが，その他のモデルでは約 230 地点（17.6%）発生しており，仮定した分布が不適切だった（モデルの特定化が誤っていた）と考えられる．AIC は NB 単一分布モデルが最も小さいが，提案モデルは NB 単一分布モデルに次いで小さい値であり，NB 単一分布モデルと近い値となった．提案モデルはパラメータ数が従来モデルより多いものの，簡潔で当てはまりの良好なモデルと考えられる．

以上から，提案モデルは，モデルの当てはまりが良好であった．特に標本帰属指標 z_{ik} を用いた予測値 $\hat{Y}_z$ による推計精度は，単一分布モデルと比較して大きく改善される結果が得られた．

3.3　提案モデルの適用可能性

図-3 に観測値と提案モデルの予測値 のプロットを示す．同図より，$\hat{Y}_\pi$ について予測値と観測値の回帰直線の傾きは 0.37 だが，$\hat{Y}_z$ について回帰直線の傾きは 0.89 であり，45 度線に近接している．したがって，データの再現には式（11）を用いることが妥当といえる．他の従来モデルについても同様に予測値の散布図を作成して比較した

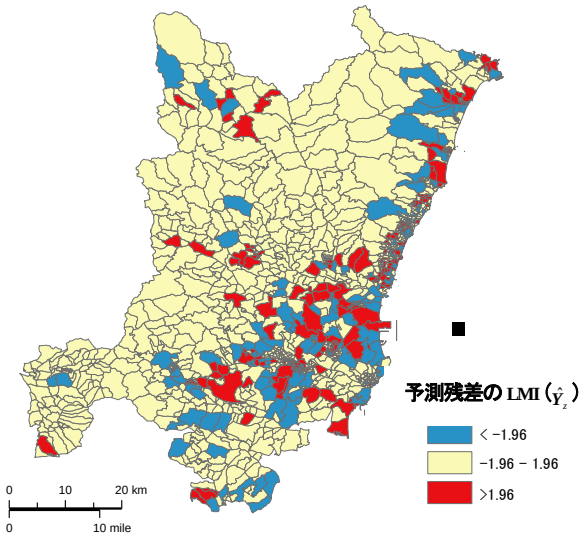

図-4　予測値$\hat{\boldsymbol{Y}}_z$ の予測残差の LMI

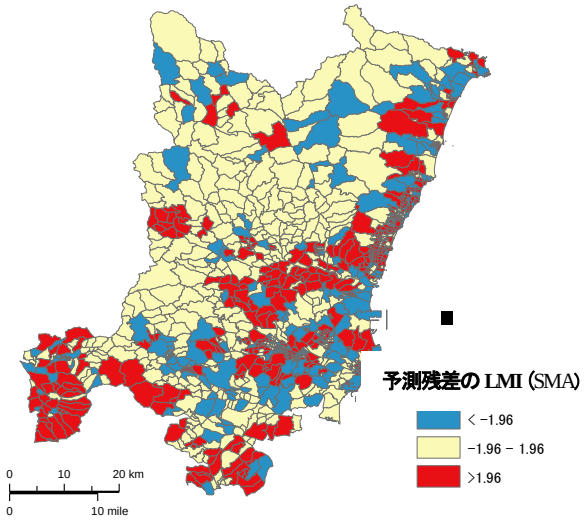

図-5　SMA 単一分布モデルの予測残差の LMI

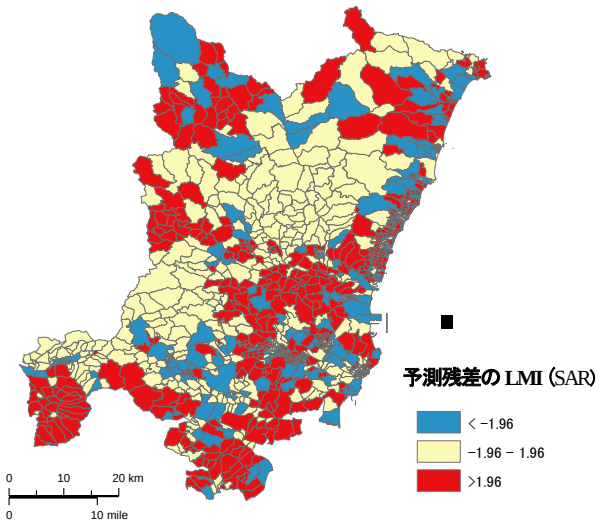

図-6　SAR 単一分布モデルの予測残差の LMI

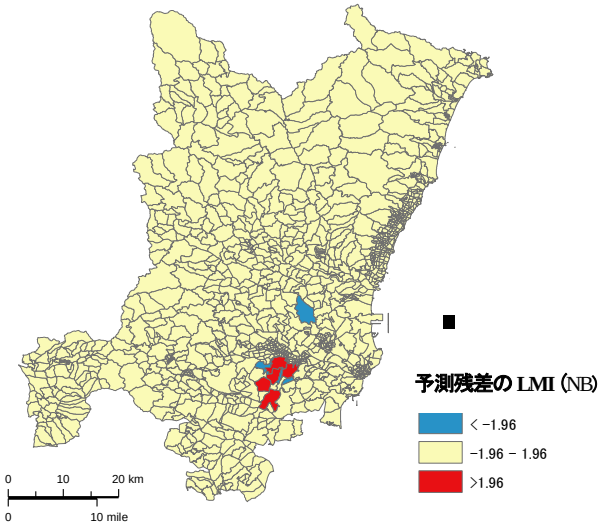

図-7　NB 単一分布モデルの予測残差の LMI

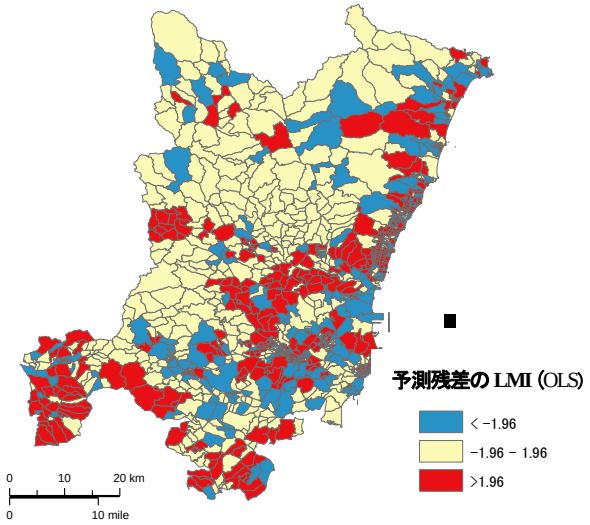

図-8　OLS 単一分布モデルの予測残差の LMI

結果，提案モデルの予測値では，観測値の小さな地域における 45 度線からのばらつきが小さくなっていることが確認できた．以上より，単一の確率分布を仮定する従来モデルでは，規模（属性値）の大きな地域への当てはまりを高めると規模の小さな地域の再現性が悪化するのに対して，提案モデルではその問題の回避が可能なため，地域全体の再現性が高いことが明らかとなった．

図-4～8 に，提案モデルの予測値$\hat{Y}_z$，および従来モデルの標準化された予測誤差の Local Moran's I 検定統計量（以下，LMI）の空間分布を示す．予測残差における空間的自己相関ありとされた地域数が最も少なかったのは，NB 単一分布モデルであった．しかし 4.2 節でも述べたように，NB 単一分布モデルの再現精度は劣悪であり，その予測残差を用いて算出した LMI の結果は信頼性が低い．NB 単一分布モデルに次いで検出地域数が少なかったのは，提案モデルの標本帰属指標に基づく予測値

$\hat{\boldsymbol{Y}}_z$ であり，図-4～8 より，他のモデル（NB 単一分布を除く）と比較して，$\hat{\boldsymbol{Y}}_z$ で空間的自己相関ありと検出される地域が少ないことがわかる．これをふまえると，$\hat{\boldsymbol{Y}}_z$ は誤差成分における空間的自己相関をうまく除去し，回帰成分で空間依存性を説明できていることがわかる．

4. 結論

本研究では，ゼロカウントデータを含む地域異質性によって，小地域データ分析で発生する問題を解消するため，新たな空間計量経済モデルを開発した．具体的には，ゼロカウント切断分布を含む，複数の確率分布の混合分布のそれぞれに対して，空間依存性を考慮した空間計量混合分布モデルを定式化した．また，地域異質性をモデル上で表現するために定義した，標本重みと標本帰属指標を用いたモデル推計手順及び 2 種類の予測値算出法を示し，それぞれの予測値を比較して，提案モデルの有効性を検証した．

実証分析では，茨城県北部（町丁目単位）の 1307 地域を対象地域として，提案モデルの小地域データ分析への適用可能性を検証した．目的変数を従業者数（ゼロカウント割合 3.9%），説明変数を建物用地，都市地域割合，世帯数の 3 変数として，提案モデルと従来モデルの推計を行い，各モデル診断指標によって推計精度を確認した．

その結果，提案モデルの推計精度は各指標で優れており，提案モデル（$\hat{\boldsymbol{Y}}_z$）の小地域データ再現性は非常に高かった．これは，提案モデルの標本重みによって地域異質性を柔軟に捉えられたためと考えられる．ただし，標本重みでは，地域間のばらつきが小さくなる一方で，標本帰属指標では地域間の重みの違いが明確で，ばらつきが大きくなる相違点があることがわかった．

以上から，提案モデルは高精度の小地域データ分析への適用できることが明らかとなった．

参考文献

1) LeSage, J. P. and Pace, R. K. : Spatial econometric modeling of origin-destination flows, Journal of Regional Science, Vol. 48, No.5, pp.941–967, 2008.
2) 爲季和樹，堤盛人：フロー間の空間的相関を考慮した負の二項重力モデル，統計数理，Vol.60，No.1，pp.121-130, 2012.
3) 爲季和樹，堤盛人：固有ベクトル空間フィルタリングを用いたゼロ過剰重力モデル，土木計画学研究・講演集，Vol.45，2012.
4) Haining, R. , Law, J. , Griffith, D. : Modelling small area counts in the presence of overdispersion and spatial autocorrelation, Computational Statistics and Data Analysis, Vol.53, pp.2923-2937, 2009.

(?)

Development of Mixed Distribution Model in Spatial Econometrics Including Truncated Distribution at “Zero-count”

Toshiaki TSUDA, Makoto TSUKAI and Ryohei HAYASHI

Recently, we can get various geogrophical data with finer spatial scale due to a development of GIS. On the other hand, fine spatial data raises new statistical problems related to zero-count in some zones and areal heterogeneity over the sample set. Conventional spatial econometric models cannot handle the above statistical problems because those models simplify areal heterogeneity by assuming a single spatial weights matrix. This study develops a novel spatial econometric model considering areal heterogeneity and the zero-count problem in fine areal data analysis. The proposed model considers multiple data generation processes (DGPs). By assuming multiple distributions including a truncated distribution at zero-count. In other words, we expanded the mixed distribution model of non-spatial version into spatial dataset. In this study, we checked the performance of proposed model by an empirical analysis about the spatial dataset of employees. As a result, the proposed model can perform well for fine area data analysis including zero-count, due to the successful introduction of heterogeneity in given dataset.