2013年度 S-PLUS学生研究奨励賞 応募研究

局所探索法を用いた空間重み行列の構造決定

爲季和樹

筑波大学　大学院システム情報工学研究科

不動産・空間計量研究室

空間データ分析における従来手法の問題点

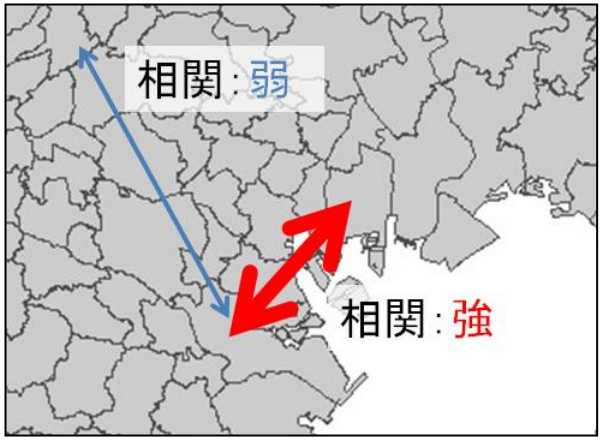
- 線形回帰モデル（LM）

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$$

$\mathbf{y}$：被説明変数	$n \times 1$ ベクトル	
$\mathbf{X}$：説明変数	$n \times q$ 行列	n：サンプル数 q：説明変数（定数項含む）の数
$\boldsymbol{\beta}$：回帰係数	$q \times 1$ ベクトル	
$\boldsymbol{\varepsilon}$：誤差項	$n \times 1$ ベクトル	

- 空間データ（地理的位置座標を持つデータ）を対象とした場合，そのデータの特徴を考慮することができない
  - 空間データの例：地価，１人当たり県民所得，地域別犯罪発生率，など…

空間データの特徴



- 地理学の第一法則（Tobler, 1970）
  - 「全ては他の全てに関連しているが，近いものほど密接に関連している」
  - 空間的自己相関
- 通常のLMは空間的自己相関を無視
  - データに空間的相関が存在する場合，パラメータ推定値の信頼性が低下（e.g., 塚井, 2005; 堤・瀬谷, 2010）

データの空間的相関を考慮したモデルの構築

『空間計量経済学』

- 地域格差分析（*e.g.* Rey and Dev, 2006）
- 空間ヘドニックアプローチ（*e.g.* Anselin and Lozano-Gracie, 2009）

空間計量経済学におけるモデル

- Anselin (1988), LeSage and Pace (2009)
- 例：空間ラグモデル（SLM）

$$\mathbf{y} = \rho \mathbf{W}^{*}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim N(0, \sigma^{2}\mathbf{I})$$

空間的相関を明示的に考慮

$\mathbf{W}^{*}$ ：$\mathbf{W}$ を行基準化した行列

$\mathbf{W}$ ：空間重み行列
データ（地点 or 地域）間における
地理的な近接性を表現した $n \times n$ の近接行列

$\mathbf{W}^{*}\mathbf{y}$ ：周辺地域の y の値の加重平均
≒近隣地域から受ける影響

W（データ間の近接性）をどのように定義するか？

W の与え方の代表例

- **境界の共有の有無**：境界が接していれば重みを与える

$$w_{ij} = \begin{cases} 1 & \text{地域 } i \text{ と} j \text{ が接している} \\ 0 & \text{else} \end{cases}$$

- **k近傍法**：最も近い k (正の整数) 地域に重みを与える

$$w_{ij} = \begin{cases} 1 & \text{地域 } j \text{ が } i \text{ の} k\text{近傍地域に含まれる} \\ 0 & \text{else} \end{cases}$$

- **距離の逆数**：地域間の距離 (d_{ij}) が近いほど大きな重みを与える

$$w_{ij} = 1/d_{ij}$$

- モデルの推定結果は **W** の与え方に大きく依存する
- 最適な **W** をどのように決定するか？

近年 **W** の構造決定に関する研究が活発化

W の与え方のアプローチの分類

Stakhovych and Bijmolt (2009), 堤・瀬谷(2012)

1. 完全に外生とする
   - 前頁参照

内生的に決定

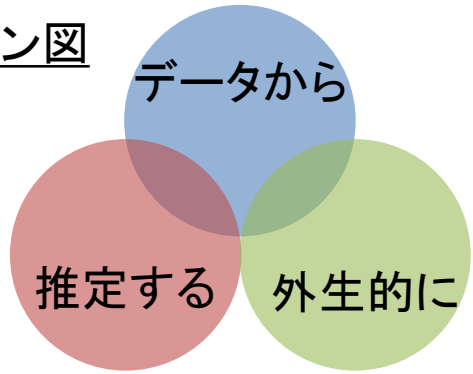
2. データから決定する
   - データからWを構築

     e.g., Getis & Aldstadt (2004), Mur & Paelinck (2011), Rogerson & Kedron (2012)
   - 複数のWの候補を考慮

     e.g., Kostov (2010), LeSage & Parent (2007), Seya et al. (2013)
3. 推定する
   - Wをパラメータとして推定

     e.g., Beenstock & Felsenstein (2012), Bhattacharjee & Jensen-Butler (2013)

各アプローチのベン図



内生的に決定するのが望ましいが、 既存研究（推定アプローチ）では
W が満たすべき性質（次頁参照）を満足させることが困難

W が満たすべき性質

- 対角要素は全てゼロ
  - **Wy** は周辺地域から受ける影響（p.4 参照）
- **W**は正則行列 (*i.e.*, det(**W**) > 0)
  - ランク落ちを防ぐ

既存研究では
この部分の制約を
導入することが困難

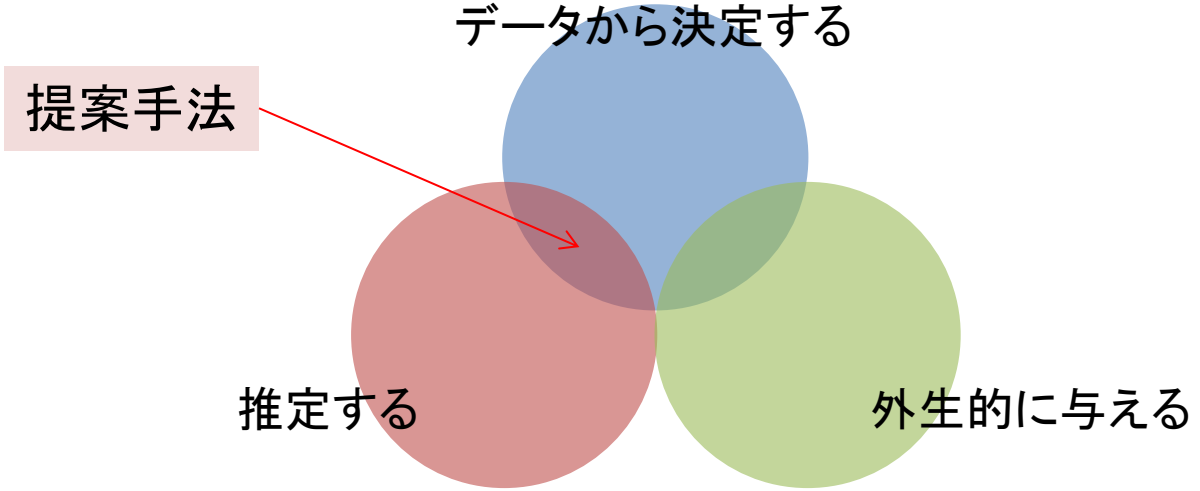
- データの地理的構造を表現していること
  - 「空間的相関」の考慮には **W** が地理的近接性を表している必要がある
  - 空間計量経済学の手法を用いる上で大前提となる仮定
- 非対角要素は非負
  - 負の値は地理的近接性の観点から解釈ができない

- 対称行列（無向グラフ）という制約は課さない
  - 非対称（有向グラフ）を許すことでより柔軟な表現が可能となる

本研究の目的

- 空間重み行列 **W** の構造を決定する新たな手法を提案
  - p.7 の性質を満足するような**W**の探索
  - 最適化問題における方法論を援用

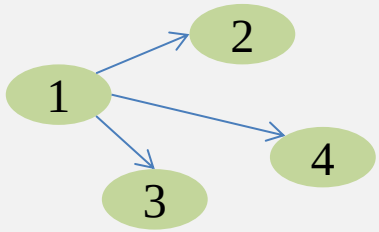
本研究の提案手法の位置づけ



提案手法の構築における考え方

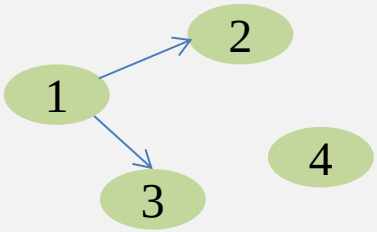
- 1 or 0 で構築される基本的な構造を仮定
  - グラフ理論の近接行列（有向グラフ）に等しい
- **W** の i 行は「地域 i がその他の各地域とリンクを繋ぐ（1）か否（0）か」の問題とみなせる

例：$n = 4$地域の場合（**W**は4×4行列）・$i = 1$ に着目すると



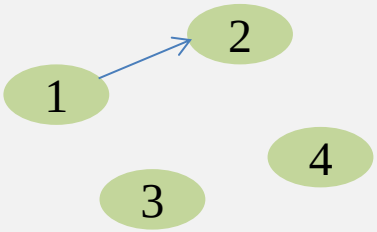
1 は 2 & 3 & 4 とリンク

$$W_1 = \begin{bmatrix} 0 & 1 & 1 & 1 \end{bmatrix}$$



1 は 2 & 3 とリンク

$$W_1 = \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix}$$



1 は 2 とリンク

$$W_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}$$

提案手法の構築
焼きなまし法を用いたアプローチ

- **W**の非対角要素すべてをパラメータとすると
- 組み合わせ数 $2^{n(n-1)}$ の中から最適な**W**を探す
  （正則条件により実際にはこれより少ない）
  - NP困難問題
- NP困難な問題に対するアプローチ
  - 効率的な探索法として局所探索法が有効
    - 山登り法，タブー探索法，焼きなまし法，…

メタヒューリスティクスに基づく焼きなまし法に着目

局所探索法と焼きなまし法

- 局所探索法
  - 反復的に近傍解を探索しながら最適解の近似解を得る
  - 代表的手法は山登り法だが，局所最適解に陥りやすい

- 焼きなまし（SA）法の特徴
  - 遷移確率 $p = \left[1, \exp\left(-\frac{\Delta}{T}\right)\right]$ $\left(\begin{array}{r}\Delta: \text{目的関数の変化量} \\ T\,(>0): \text{温度パラメータ}\end{array}\right)$
    - 改悪解であっても確率的に受理 ←局所最適解から脱出可能
  - 温度関数（cooling schedule）
    - 時間が経つにつれ温度 T を徐々に下げていく
    - 反復回数が増えると改悪解への遷移確率が低くなる

本研究の提案手法のアルゴリズム

1. 初期値生成
2. 近傍解のランダム生成

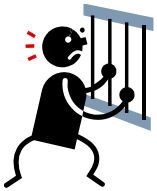
「地理的近接性の表現」(p.7)の制約を満たす

- ある地域 $i = 1, \ldots, n$ をランダムに選択
- 次のうちひとつをランダムに行う
  - Ⅰ. i とリンクされている地域群の中で最も i から遠い地域のリンクを切る
  - Ⅱ. i とリンクされていない地域群の中で最も i から近い地域へリンクを繋ぐ
- $\mathbf{W}^*$ を行基準化して $\mathbf{W}$ を構築

3. $\mathbf{W}$ と $\mathbf{y}, \mathbf{X}$ を用いてモデルを最尤推定し目的関数の値を算出
4. 遷移確率に基づいて近傍解を暫定解に置き換えるか選択
5. 温度関数に基づいて温度パラメータ T を減少
6. 終了条件を満たすまで 2 ~ 5 を繰り返す

実証分析
データセット

- 用いるデータ：Columbus, OH crime dataset
  - 米国オハイオ州コロンバス市49地区における1980年の犯罪データ

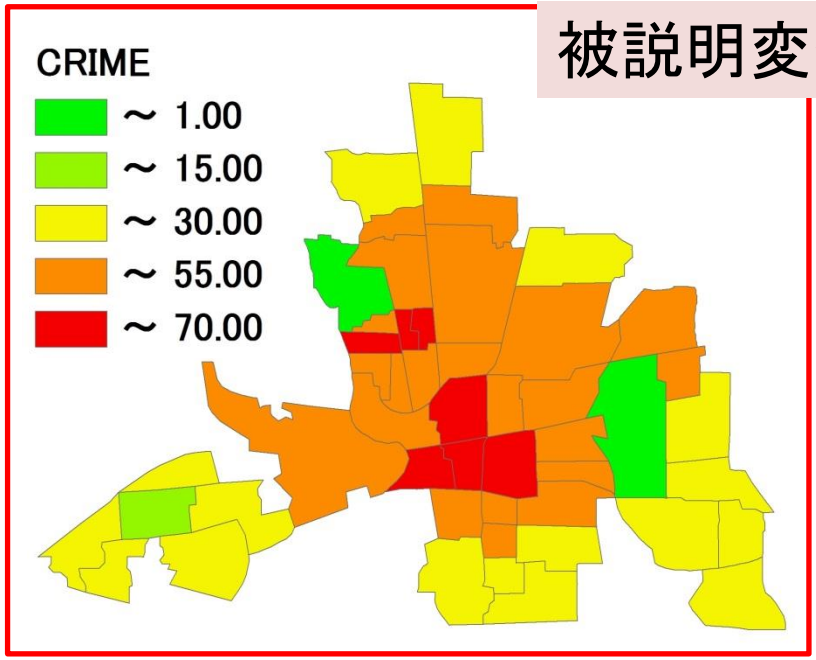


  - GeoDa Center（http://geodacenter.asu.edu）で入手可能（無償）
  - 空間データ分析において非常に有名なデータセット

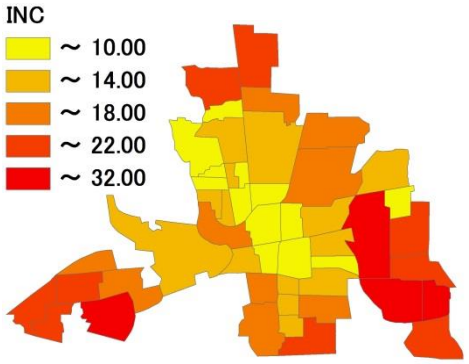
変数名	説明
CRIME	1000世帯当たりの住宅侵入窃盗・車両盗難の件数
INC	世帯収入（/$1000）
HOVAL	住宅価格（/$1000）
OPEN	空き地面積
PLUMB	配管未整備住宅の割合
DISCBD	中心業務地区までの距離

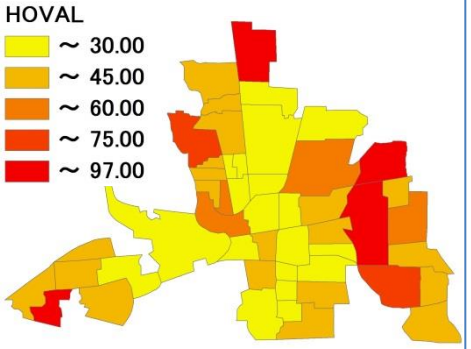
実証分析
データの空間分布

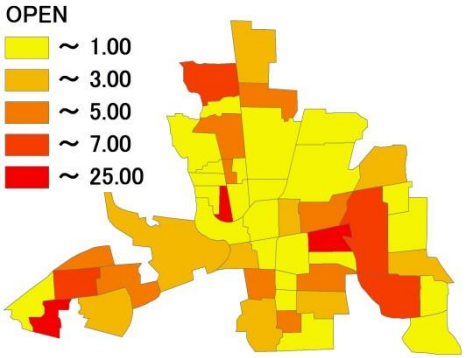
被説明変数

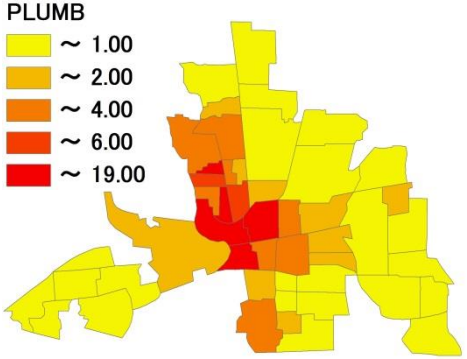


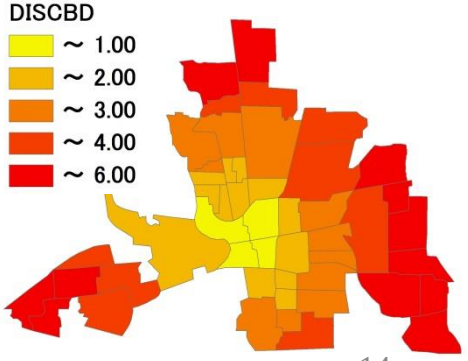
説明変数











実証分析
モデルと目的関数

- **空間計量経済モデル**（ロバスト性チェックのため2つのモデルを使用）
  - 空間ラグモデル（SLM）

$$\mathbf{y} = \rho \mathbf{W}^{*}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim N(0, \sigma^{2}\mathbf{I})$$

  - 空間エラーモデル（SEM）

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \qquad \mathbf{u} = \lambda \mathbf{W}^{*}\mathbf{u} + \boldsymbol{\varepsilon} \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^{2}\mathbf{I})$$

- **目的関数**：－（対数尤度）

$$\text{SLM}: \quad -\ln L = \frac{n}{2}\ln(2\pi\sigma^{2}) - \ln|\mathbf{A}| + \frac{(\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^{2}}$$

$$\text{SEM}: \quad -\ln L = \frac{n}{2}\ln(2\pi\sigma^{2}) - \ln|\mathbf{B}| + \frac{[\mathbf{B}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]'[\mathbf{B}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]}{2\sigma^{2}}$$

$$\left(\mathbf{A} = \mathbf{I} - \rho \mathbf{W}^{*}, \ \mathbf{B} = \mathbf{I} - \lambda \mathbf{W}^{*} \right)$$

実証分析
その他アルゴリズムに関する設定

- 繰り返し回数$(k = 1, 2, \ldots, K)$：$K = 30{,}000$

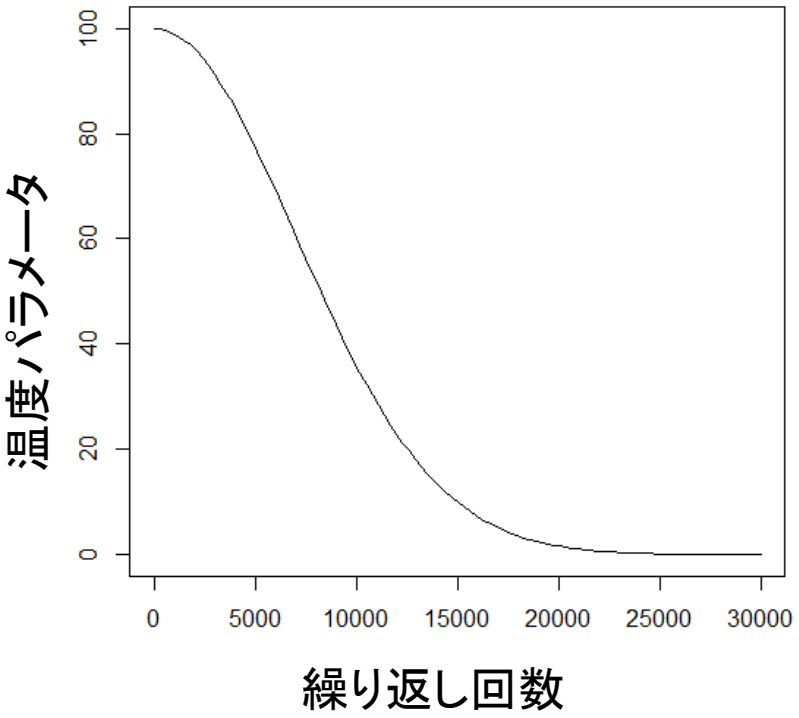
- 温度関数$(T_k = T_0, T_1, \ldots, T_K)$：

$$T_k = T_0 e^{-\alpha k^2}$$

$$\alpha = \left(\frac{1}{K^2}\right) \ln\left(\frac{T_0}{T_K}\right)$$

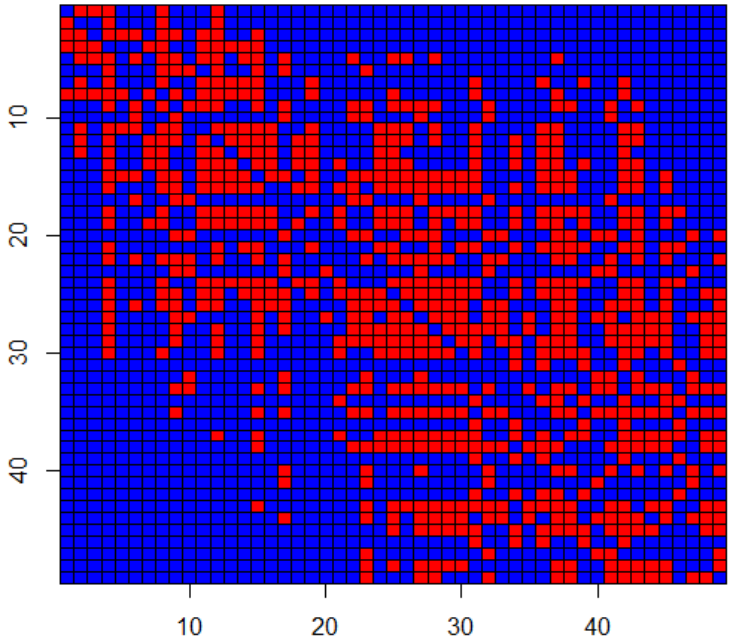
  - 初期温度：$T_0 = 100$
  - 終了温度：$T_K = 0.01$

クーリング・スケジュール



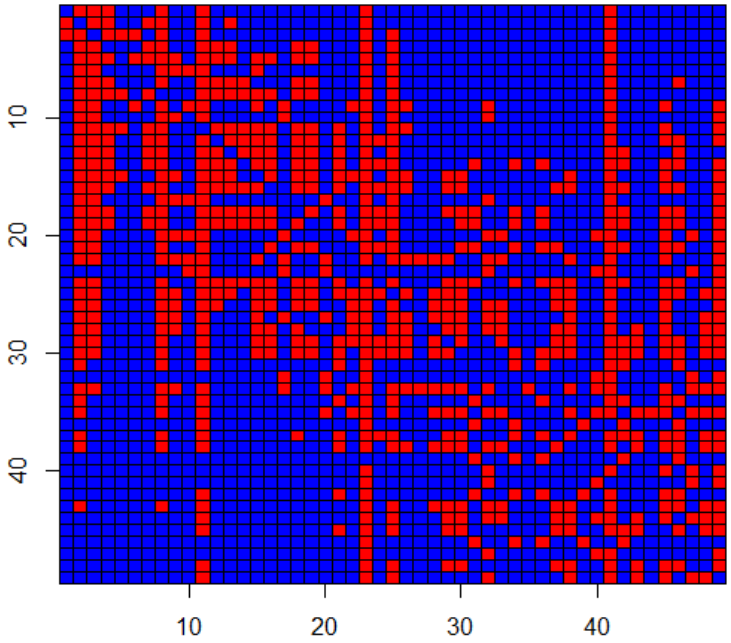
分析結果
得られた最適な W の視覚化

SLM



平均リンク数：16.6
密度：33.8%

SEM



平均リンク数：16.3
密度：33.2%

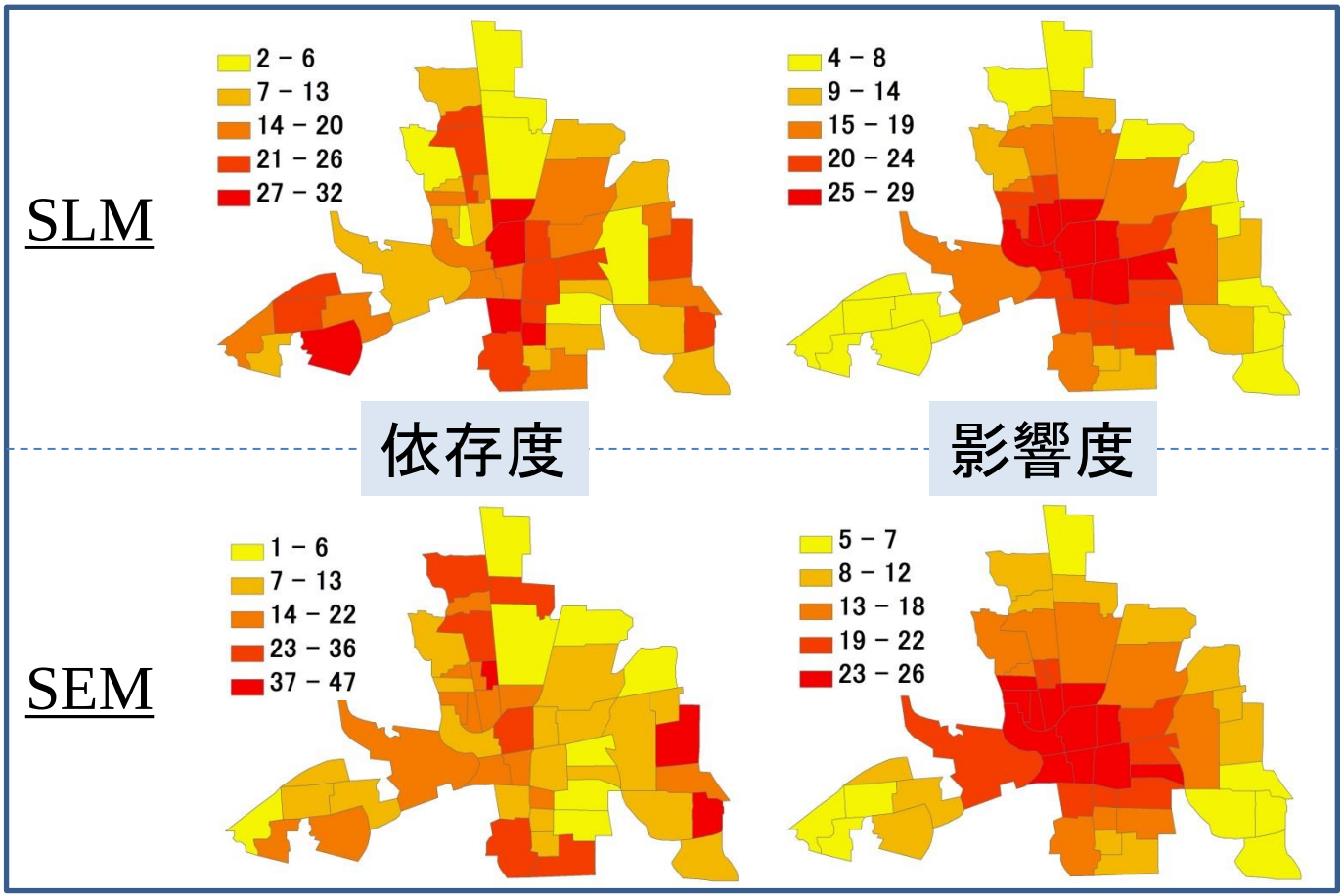
1：リンク有
0：リンク無

- SLM・SEMで W の構造は異なるが，平均リンク数・密度は類似

分析結果
リンク数から見る地域の依存・影響度分析

空間重み行列 **W**:

- 行和 ＝ 地域 i に影響を与えている地域数 ⇒ 依存度
- 列和 ＝ 地域 i が影響を及ぼしている地域数 ⇒ 影響度



空間分布の傾向

- 依存度:
  全体的に分散して分布
- 影響度:
  中心地域ほど高い

影響度は空間的な配置に依存
⇒ネットワーク中心性

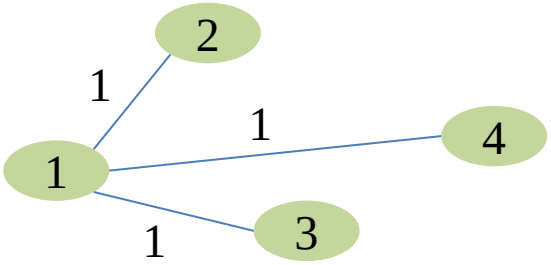
分析結果
パラメータ推定結果

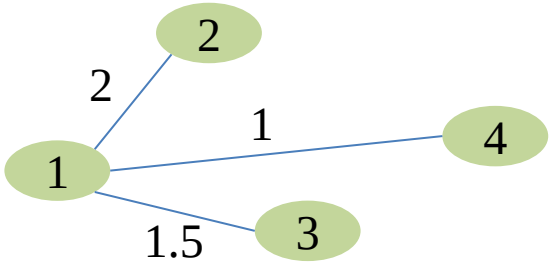
	LM		SLM		SEM	
	係数	p 値	係数	p 値	係数	p 値
定数項	67.37	0.00	105.22	0.00	68.07	0.00
INC	-0.99	0.00	-1.06	0.00	-1.04	0.00
HOVAL	-0.22	0.03	-0.24	0.01	-0.22	0.01
OPEN	0.10	0.76	0.17	0.55	0.11	0.72
PLUMB	0.61	0.21	0.63	0.14	0.59	0.14
DISCBD	-3.91	0.02	-6.75	0.00	-3.72	0.00
ρ			-0.71	0.01		
λ					-0.08	0.03

- ρ, λ（空間パラメータ）はいずれも p 値 < 0.05
  - 5%水準で負に有意　←**W** が 近接性をうまく表現できていない可能性
- 各リンクの重みに関する仮定を変更して再度分析

実証分析2
距離関数によるリンクへの重みづけ

- 以上の分析では全てのリンクの重みが等しい
- しかし距離が近い地域は遠い地域よりも与える影響は大きいはず





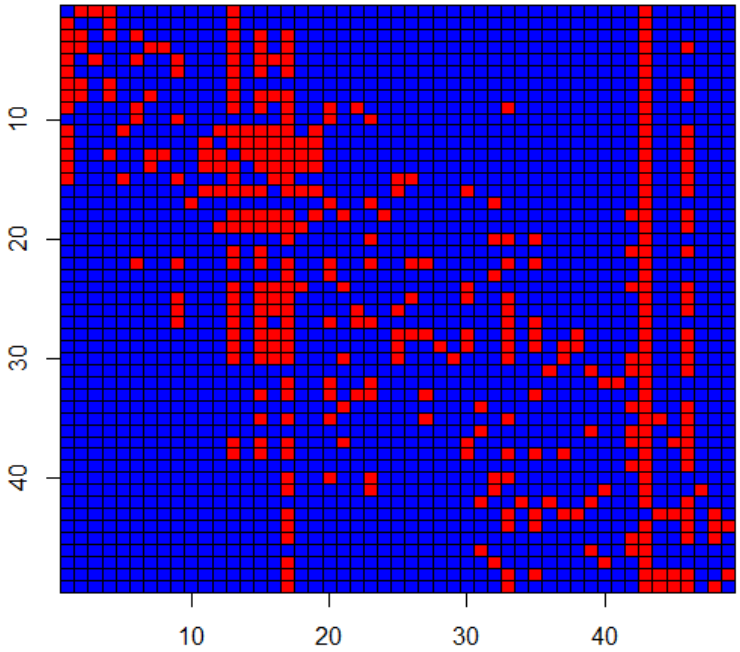
与える影響は
距離に関わらず<u>同じ</u>

与える影響は
<u>距離に反比例</u>する
（空間的相互作用の考え方）

リンクの重みを距離逓減関数（$1 / d_{ij}$）で与える

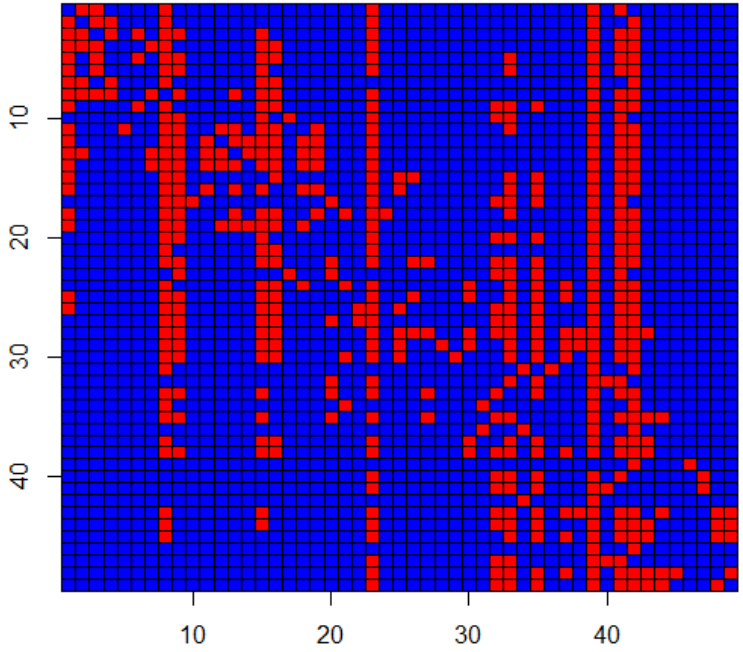
分析結果2
得られた最適な W の視覚化

SLM



平均リンク数：8.2
密度：16.7%

SEM



平均リンク数：10.8
密度：22.0%

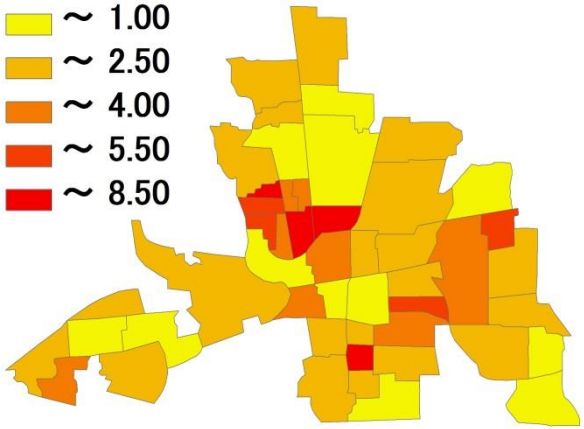
1：リンク有
0：リンク無

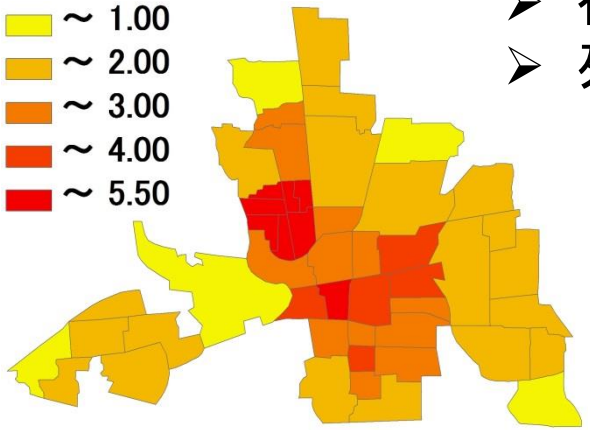
- リンクに距離関数の重みを導入することで密度が減りスパースな行列に

分析結果2
地域の依存・影響度分析

- 行和 ＝ 依存度
- 列和 ＝ 影響度

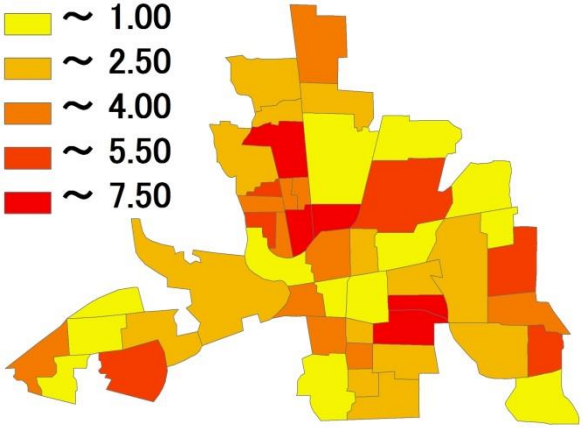
SLM

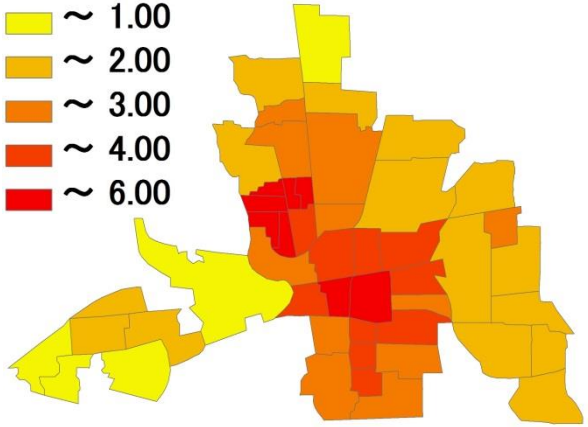




依存度　　影響度

SEM





- 依存・影響度の空間分布は分析1 と同様の傾向を示している

分析結果2
パラメータ推定結果

	LM		SLM		SEM	
	係数	p 値	係数	p 値	係数	p 値
定数項	67.37	0.00	23.31	0.00	61.16	0.00
INC	-0.99	0.00	-0.51	0.04	-0.65	0.01
HOVAL	-0.22	0.03	-0.38	0.00	-0.30	0.00
OPEN	0.10	0.76	0.14	0.55	0.39	0.04
PLUMB	0.61	0.21	0.96	0.01	0.46	0.17
DISCBD	-3.91	0.02	1.58	0.22	-3.65	0.01
ρ			0.69	0.00		
λ					0.72	0.00

- ρ, λ は1%水準で正に有意
- 距離による重みづけで近接性を適切に捉えられたことが示唆された

まとめと今後の課題

- 空間重み行列の構造を決定する新たなアプローチ法を提案
  - モデルのパラメータ推定と同時に決定可能
  - メタヒューリスティクスに基づく手法の援用により制約の柔軟な導入が可能
  - 得られたWを用いてクラスター分析や影響度分析が可能となる点も本提案手法の特徴
- 今後の課題
  - 遺伝的アルゴリズム等のその他メタヒューリスティック手法の構築と比較検討

主な参考文献

- Anselin L (1988) *Spatial Econometrics: Methods and Models*, Kluwer, Dordrecht.
- Beenstock M, Felsenstein D (2012) Nonparametric estimation of the spatial connectivity matrix using spatial panel data, *Geographical Analysis*, 44, pp.386–397.
- Bhattacharjee A, Jensen-Butler C (2013) Estimation of the spatial weights matrix under structural constraints, *Regional Science and Urban economics*, 43, pp.617–634.
- Getis A, Aldstadt J (2004) Constructing the spatial weights matrix using a local statistic, *Geographical Analysis*, 36, pp.90–104.
- Kostov P (2010) Model boosting for spatial weight matrix selection in spatial lag models, *Environment and Planning B*, 37, pp.533–549.
- LeSage JP, and Parent O (2007) Baeysian model averaging for spatial econometric models, *Geographical Analysis*, 39, pp.241–267.
- LeSage JP, Pace RK (2009) *Introduction to Spatial Econometrics*, Chapman & Hall/CRC, Boca Raton.
- Mur J, Paelinck JH (2011) Deriving the W-matrix via p-median complete correlation analysis of residuals, *The Annals of Regional Science*, 47, pp.253–267.
- Rogerson PT, Kedron P (2012) Optimal weights for focused tests of clustering using the local Moran statistic, *Geographical Analysis*, 44, pp.121–133.
- Seya H, Yamagata Y, Tsutsumi M (2013) Automatic selection of a spatial weight matrix in spatial econometrics: Application to a spatial hedonic approach, *Regional Science and Urban Economics*, 43, pp.429–444.
- Stakhovych S, Bijimolt THA (2009) Specification of spatial models: A simulation study on weights matrices, *Papers in Regional Science*, 88, pp.389–408.
- Tobler W (1970) A computer movie simulating urban growth in the Detroit region, *Economic Geography*, 46, pp.234–240.
- 塚井誠人（2005）空間統計モデルのフロンティア，土木計画学研究・論文集，Vol.22，pp.1–13.
- 堤盛人・瀬谷創（2010）便益計測への空間ヘドニック・アプローチの適用，土木学会論文集D，Vol.66, pp.178–196.
- 堤盛人・瀬谷創（2012）応用空間統計学の二つの潮流：空間統計学と空間計量経済学，統計数理，60，pp.3 –25.