2014年度　S-PLUS学生研究奨励賞応募論文

空間的相関を考慮する組成データ解析手法の社会経済データへの適用

吉田崇紘

筑波大学大学院システム情報工学研究科，博士前期課程２年

筑波大学
University of Tsukuba

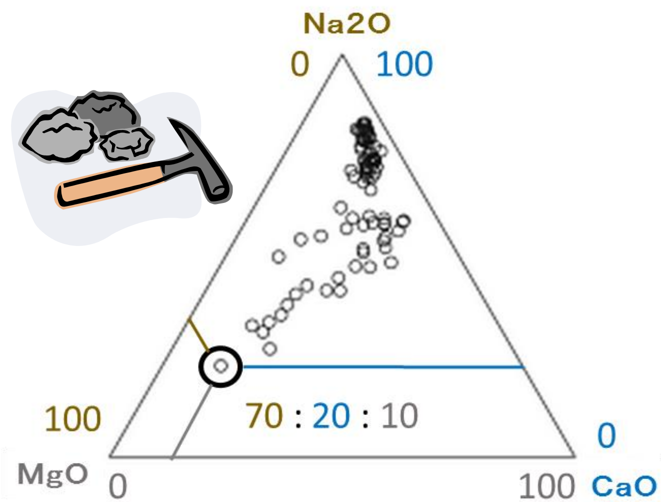
組成データ

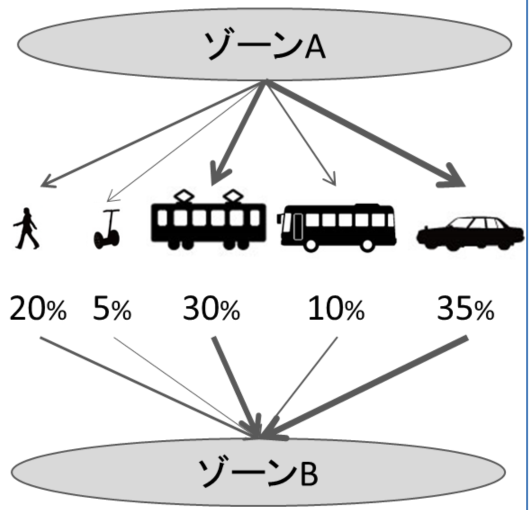
組成データ行列

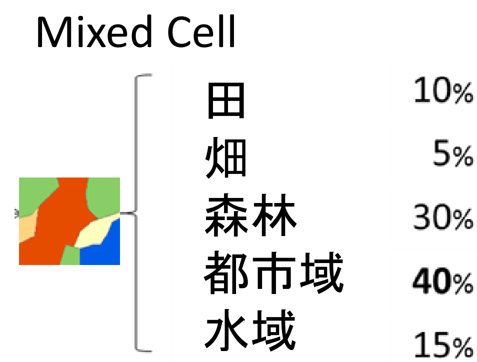
$$\mathbf{Y}_{n\times D}=\begin{pmatrix} \boxed{y_{11} \quad y_{12} \quad \cdots \quad y_{1D}} \\ \boxed{y_{21} \quad y_{22} \quad \cdots \quad y_{2D}} \\ \vdots \quad \vdots \quad \ddots \quad \vdots \\ \boxed{y_{n1} \quad y_{n2} \quad \cdots \quad y_{nD}} \end{pmatrix}$$

$$\boxed{\sum_{K=1}^{D} y_{iK} = \bar{a}}$$

- 定義（Aitchison, 1986）：
  - 全要素が非負の値であり，**定数和制約**を持つ多次元データ





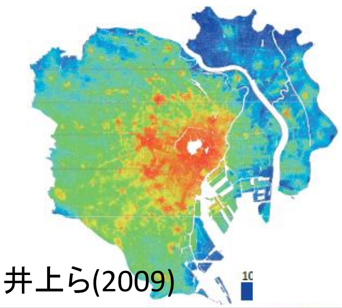
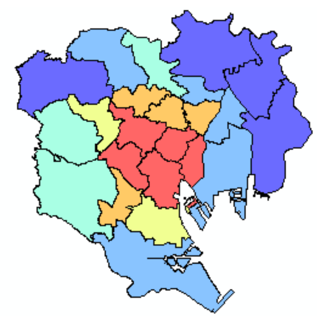


岩石の化学組成　　交通機関分担率　　土地利用組成

幅広い分野で一般的に存在するデータ形式

空間データと組成データモデル

表：空間データの分類※1（Cressie, 1993）と組成データモデルの適用例（※1: 点過程データを除く）

空間データ	例	領域（固定）	図	組成データモデルの適用例
地球統計データ	・標高 ・気温 自然科学データ	・連続空間（≒無限標本）		多数
地域／格子データ	・人口 ・所得 社会経済データ	・離散空間（≒有限標本）		**少数**

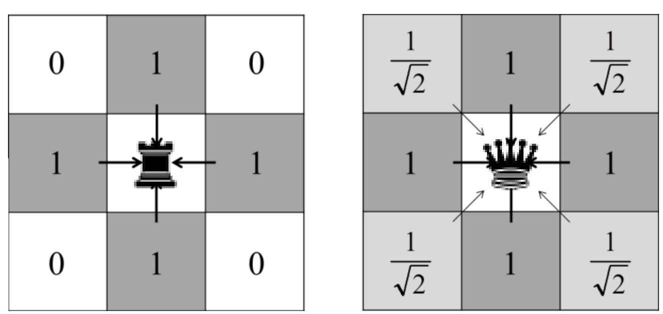
地域／格子データを対象として，空間データの特性（空間的相関）を考慮した研究は，Allen *et al*. (2013), Leininger *et al*. (2013) の**2 例のみ**（⇔地球統計データを対象とした例は多数存在）

目的

空間的相関を考慮する組成データモデルの地域／格子データ（社会経済データ）への適用

- 工夫の余地
  - ➢ 既往研究（Allen *et al.*, 2013; Leininger *et al.*, 2013）の空間的相関の考慮は，隣接の影響のみを対象としている
  - ➢ 社会経済データが持ちうる空間的相関の影響は距離に応じて減衰するのでは？（仮説）

Neighb.-based Spatial Model
（既往研究）



Distance-based Spatial Model
（本研究）

W: 空間重み行列
（空間的相関の影響関係を表現する行列．付録を参照）

図：中央のメッシュ（$W_{i\cdot}$）に対する **W** の要素の与え方

定数和制約の対処

- $\mathbf{y} \in \mathbb{S}^{D-1}$: 組成データ
- D: 次元
- $d = D - 1$

- 定数和制約
  - ➢ D 次元のうち，1 から $D - 1$ (= d) 次元までの変数の値が決まれば，残り 1 次元の変数の値は一意に決定される

$$\mathbf{y} = (y_1, \ldots, y_D)^T \mid y_k > 0 \ (k = 1, \ldots, D), \ \sum_k y_k = 1$$

  - ➢ 組成データを扱う際は必ず考慮する必要がある制約条件

◆対処法 : 対数比変換法 alr(.)

- alr: additive log-ratio
- $\boldsymbol{B}$: $(p + 1) \times d$ の係数行列
- p: 説明変数の数
- $\mathbf{x}_i$: $1 \times (p + 1)$ の説明変数ベクトル
- $\mathbf{V}$: $d \times d$ の共分散行列

$$\mathrm{alr}(\mathbf{y}_i) = \left(\ln \frac{y_{i1}}{y_{iD}}, \ldots, \ln \frac{y_{id}}{y_{iD}} \right)^T$$

➢ 利点 : $\mathrm{alr}(\mathbf{y}_i) \in \mathbb{R}^d$ は多次元正規分布 N_d に従いやすい

（Aitshison, 1986）

$\Rightarrow \quad \mathrm{alr}(\mathbf{y}_i) \sim N_d(\mathbf{B}^T \mathbf{x}_i, \mathbf{V})$ 　としてモデル化可能

空間的相関の考慮

- **Multivariate conditional autoregressive model (MCAR model)**（Mardia, 1988）
  - 階層ベイズモデル
    - ランダム効果の事前分布で空間的相関を考慮可能

- η_i: $d\times 1$ のランダム効果ベクトル
- Σ: $d\times d$ の共分散行列
- w_{ij}: $n\times n$ の空間重み行列 $\mathbf{W}$ の要素
- S_i: $\mathbf{W}$の行和

$$\mathrm{alr}(\mathbf{y}_i) \sim N_d\left(\mathbf{B}^T\mathbf{x}_i + \boldsymbol{\eta}_i, \mathbf{V}\right)$$

係数行列　ランダム効果　分散共分散行列

多次元正規分布

多次元正規分布

$$\boldsymbol{\eta}_i \mid \{\boldsymbol{\eta}_j\}_{j\neq i} \sim N_d\left(\frac{1}{S_i}\sum_{j=1}^{n} w_{ij}\boldsymbol{\eta}_j, \frac{1}{S_i}\boldsymbol{\Sigma}\right)$$

逆ウィシャート分布

：共役事前分布

逆ウィシャート分布

パラメータ推定法

- ギブス・サンプラーを用いたMCMC法

  条件付事後分布が全て標準的な分布に従う → 効率的なサンプリングが可能

条件付事後分布

$$vec(\mathbf{B}) \mid \mathbf{z}_i, \mathbf{X}, \mathbf{V}, \boldsymbol{\eta}_i \sim \underline{N_1}\left(\boldsymbol{\Omega}\left(\mathbf{V}^{-1} \otimes \mathbf{X}^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i^T\left(\mathbf{z}_i - \boldsymbol{\eta}_i\right), \boldsymbol{\Omega} \right)$$

$$\mathbf{V} \mid \mathbf{z}_i, \mathbf{X}, \mathbf{B}, \boldsymbol{\eta}_i \sim \underline{IW_d}\left(m_V + S_i, \mathbf{M}_V + \sum_{i=1}^{n} \mathbf{E}_i \mathbf{E}_i^T \right)$$

$$\boldsymbol{\eta}_i \mid \mathbf{z}_i, \mathbf{X}, \mathbf{B}, \{\boldsymbol{\eta}_j\}_{j \neq i}, \boldsymbol{\Sigma} \sim \underline{N_d}\left(\mathbf{A}^{-1}\left(\mathbf{V}^{-1}\left(\mathbf{z}_i - \mathbf{B}^T\mathbf{x}_i\right) + \boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n} w_{ij} \boldsymbol{\eta}_j \right) \right), \mathbf{A}^{-1} \right)$$

$$\boldsymbol{\Sigma} \mid \boldsymbol{\eta}_i \sim \underline{IW_d}\left(m_\Sigma + S_i, \mathbf{M}_\Sigma + \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\mathbf{D}_w - \mathbf{W}\right)_{ij} \boldsymbol{\eta}_i \boldsymbol{\eta}_i^T \right)$$

where

$$\boldsymbol{\Omega} = \left(\lambda \mathbf{I}_{d \times p} + \mathbf{V}^{-1} \otimes \mathbf{X}^T\mathbf{X}\right)^{-1},$$

$$\mathbf{E} = \mathbf{z}_i - \mathbf{B}^T\mathbf{x}_i - \boldsymbol{\eta}_i, \quad \mathbf{A} = \mathbf{V}^{-1} + S_i \boldsymbol{\Sigma}^{-1}$$

- $\mathbf{D}_w$: $n \times n$ の対角行列（$(\mathbf{D}_w)_{ii} = S_i$）
- λ, m_V, M_V, m_Σ, M_Σ: ハイパーパラメータ

実証分析

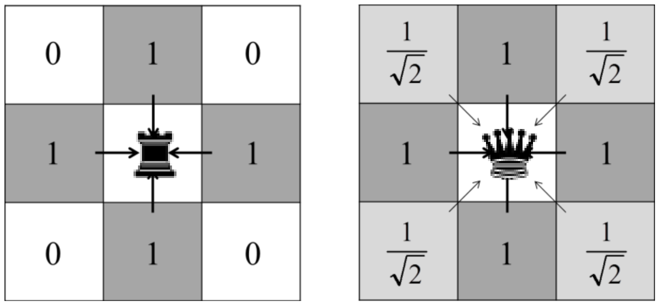
■ 空間重み行列 W の設定を **Neighb.-based** から **Distance-based** に拡張し，予測精度の比較を行う

設定

乱数発生回数：20,000回
Burn-in期間：2,000回
$d = 4,\ \lambda = 1,000$
$m_{\mathrm{V}} = m_{\Sigma} = (d+2)\ ,\ M_{\mathrm{V}} = M_{\Sigma} = 2\mathbf{I}_d$

用いるデータ

- 対象範囲：　茨城県（n = 5,904）
- 集計単位：　３次メッシュ
- 被説明変数（組成データ）：
  - ➢ 土地利用データ（国土数値情報）(D =) 5 種
- 説明変数：
  - ➢ 地理的条件（標高など），社会経済的条件（人口など）→ 次ページ



Neighb.-based Spatial Model
（既往研究）

Distance-based Spatial Model
（本研究）

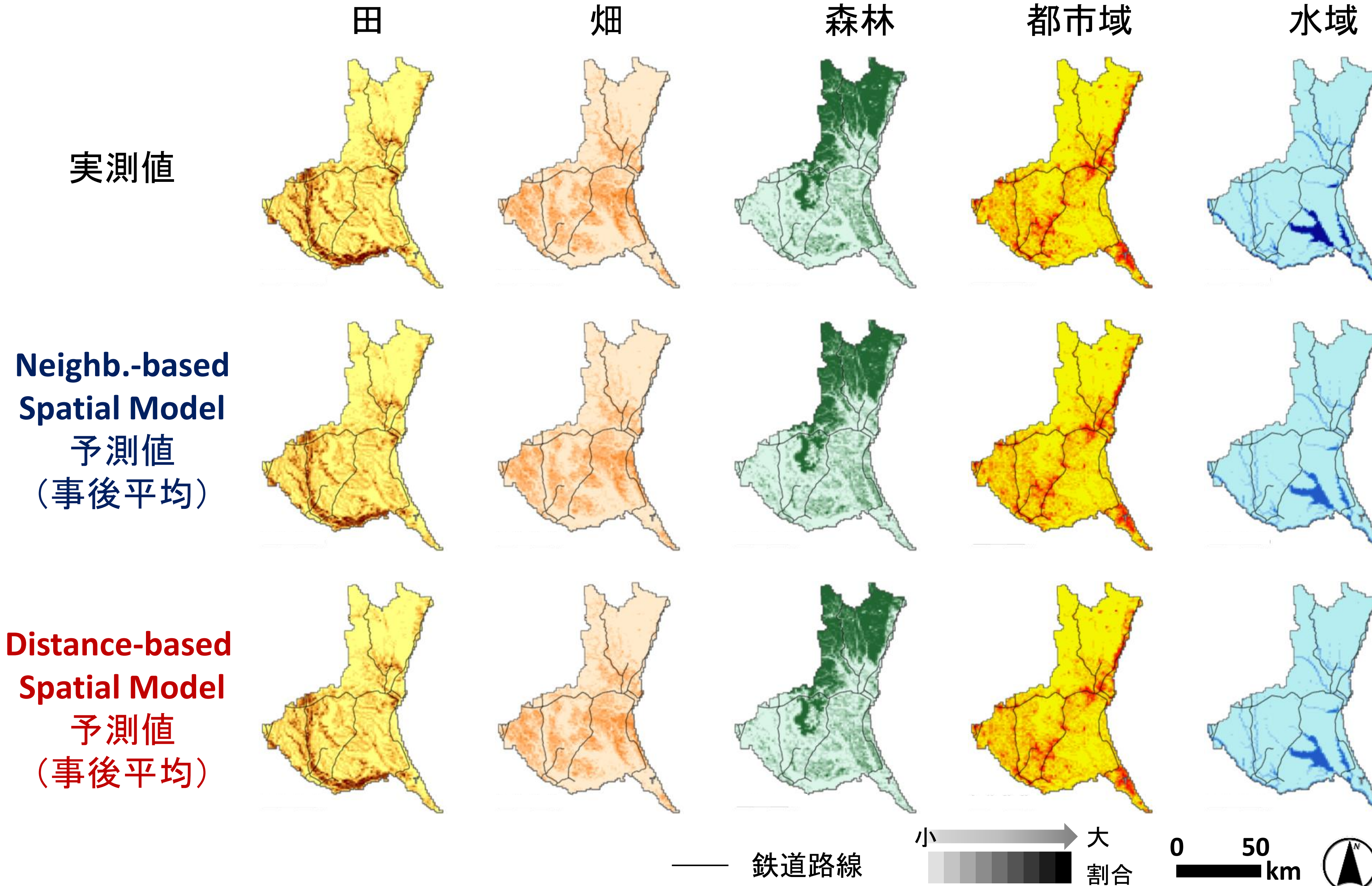
図：中央のメッシュ（$W_{i\cdot}$）に対する Wの要素の与え方

説明変数

変数名	内容
lnPOP	人口密度（人/km^2） の自然対数値
lnPOP_2	人口密度（人/km^2） の二乗の自然対数値
Avg_Elv	平均標高（m）
Avg_Slope	平均傾斜（度）
TRL	道路総延長（Total Road Length）（km）
Dist_Sta	最寄駅までの直線距離（km）
Dist_River	最寄一級河川までの直線距離（km）
D_AF	扇状地（Alluvial Fun）（該当：1，該当しない：0）
D_NL	自然堤防（Natural Levee）（該当：1，該当しない：0）
D_BM	後背湿地（Back Marsh）（該当：1，該当しない：0）
D_Delta	三角州・海岸低地（Delta）（該当：1，該当しない：0）
D_SD	砂州・砂礫州（Sandbar）（該当：1，該当しない：0）
D_Lake	湖沼内（Lake）（該当：1，該当しない：0）

各パラメータの収束はGeweke の方法によって確認している

予測結果の比較


田
畑
森林
都市域
水域
実測値
Neighb.-based
Spatial Model
予測値
（事後平均）
Distance-based
Spatial Model
予測値
（事後平均）
鉄道路線
小
大
割合
0
50
km

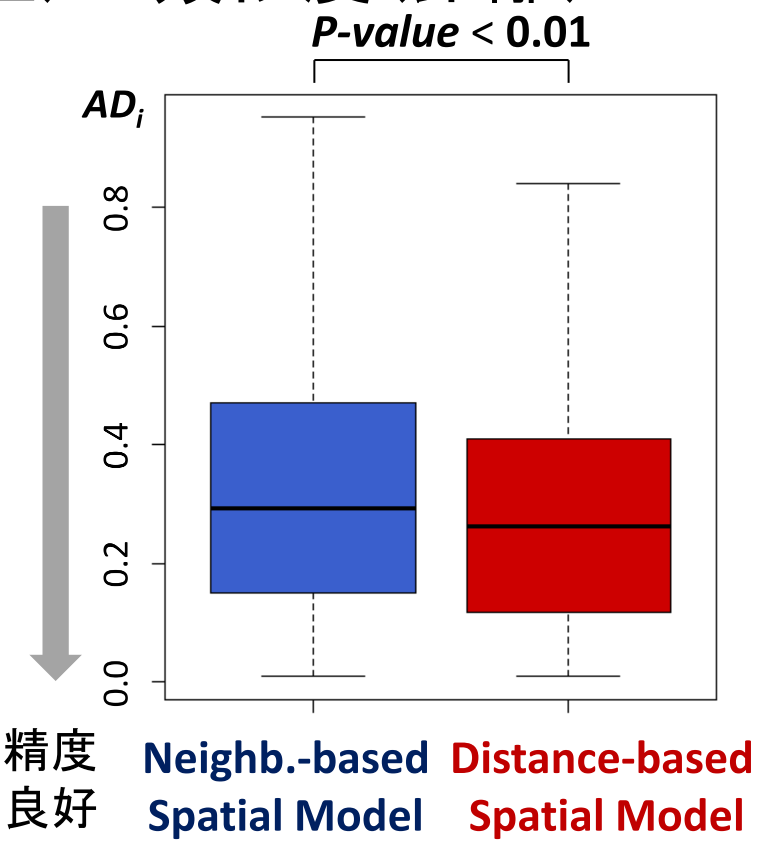
予測精度の評価

- 指標：Aitchison 距離 ***AD$_i$***
  - 組成データ間（実測値と予測値）の類似度（距離）

$$AD_i = \sqrt{\sum_{K=1}^{D}\left[\ln\left\{\frac{y_{iK}}{\left(\prod_{K=1}^{D} y_{iK}\right)^{1/D}}\right\} - \ln\left\{\frac{\hat{y}_{iK}}{\left(\prod_{K=1}^{D} \hat{y}_{iK}\right)^{1/D}}\right\}\right]^2}$$

Wilcoxon の符号順位検定

⇒ 1 %水準で有意

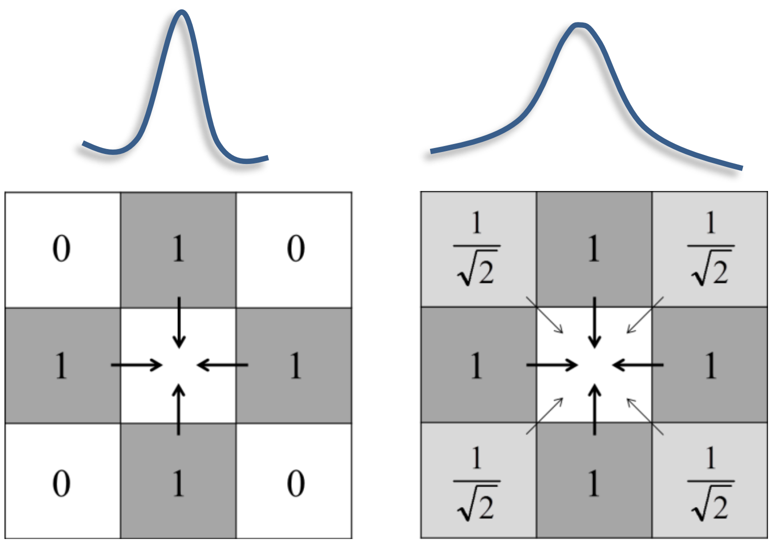


図：実測値と両モデルの予測値の
AD_i の比較（箱ひげ図）

Distance-based に拡張することで
統計的有意に予測精度が向上

考察

- 予測精度の向上：
  - Distance-based は距離に応じてスムージング
    ⇒ 土地利用データの空間的相関を
    Neighb.-based に比べ良く表現している可能性



まとめ

- **MCAR**モデルにおける **W** を，**Neighb.-based**から**Distance-based**に拡張
- 実データを用いて，**Neighb.-based**と**Distance-based**の予測精度を比較 ⇒ 統計的有意に精度が向上

今後の展望

- データから **W** を構築･決定する方法の検討
  - たとえば，地球統計学のバリオグラムを用いて，空間的相関の影響が及ぶ範囲（距離）を推定

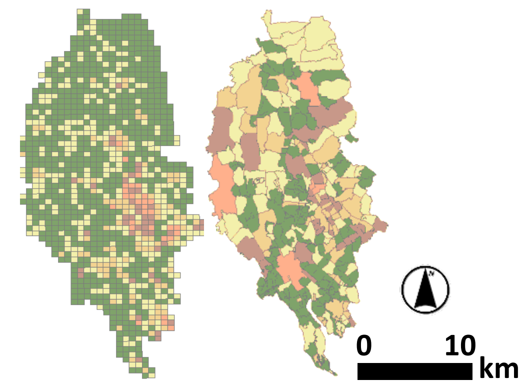
参考文献

- Aitchison, J.: *The statistical analysis of compositional data*, Chapman and Hall, 1986.
- Allen, J., Leininger, T., Hurd, J., Civico, D., Gelfand, A., and Silander, J.: Socioeconomics drive woody invasive plant richness in New England, USA through forest fragmentation, *Landscape Ecology*, **28** (9), 1671–1686, 2013.
- Cressie, N.: *Statistics for Spatial Data, Revised Edition*, Wiley, 1993.
- Leininger, T., Gelfand, A., Allen, J., and Silander, J.: Spatial Regression Modeling for Compositional Data With Many Zeros, *Journal of Agricultural, Biological, and Environmental Statistics*, **18** (3), 314–334, 2013.
- Mardia, V.: Multi-dimensional Multivariate Gaussian Markov Random Fields with Applications to Image Processing, *Journal of Multivariate Analysis*, **24** (2), 265–284, 1988.
- 井上 亮，清水英範，吉田雄太郎，李勇鶴：時空間クリギングによる東京23区・全用途地域を対象とした公示地価の分布と変遷の視覚化，『GISー理論と応用』，**17** (1), 13–24, 2009.
- 小荒井衛，中埜貴元：地理空間情報の時空間化の検討とつくば市における試作，『GIS-理論と応用』，**21** (1), 1–7, 2013.

付録：
組成データが生じる場面の例(1)

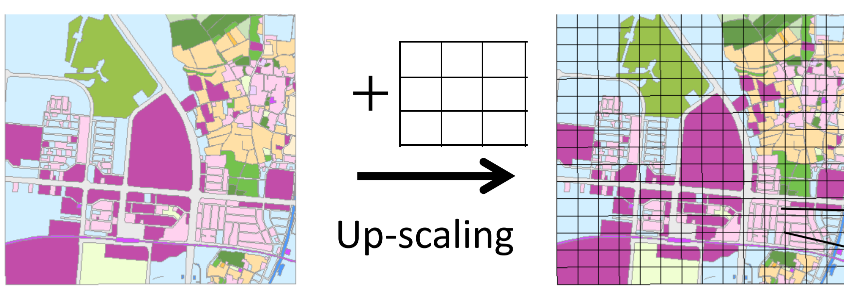
- 土地利用データ
  - ➢ 航空写真・衛星画像の撮影精度向上
    ⇒　空間詳細なデータが入手**可能**
- 社会経済データ(*e.g.* 人口データ)
  - ➢ 秘匿・特定防止
    ⇒　空間詳細なデータは入手**困難**

図：研究学園駅周辺の土地利用分布
(小荒井・中埜, 2013)



図：つくば市の人口分布
(左：500mMesh, 右：小地域)

両データの関係を分析したい…



図：土地利用分布(小荒井・中埜, 2013)と新たな集計単位

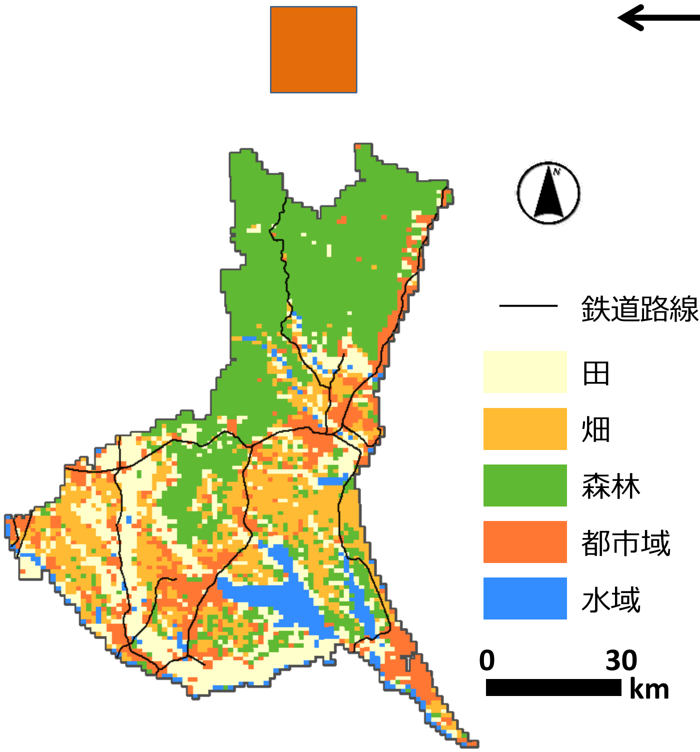
新たな集計単位における
属性値をどう与えるか？

？

Mixed cell

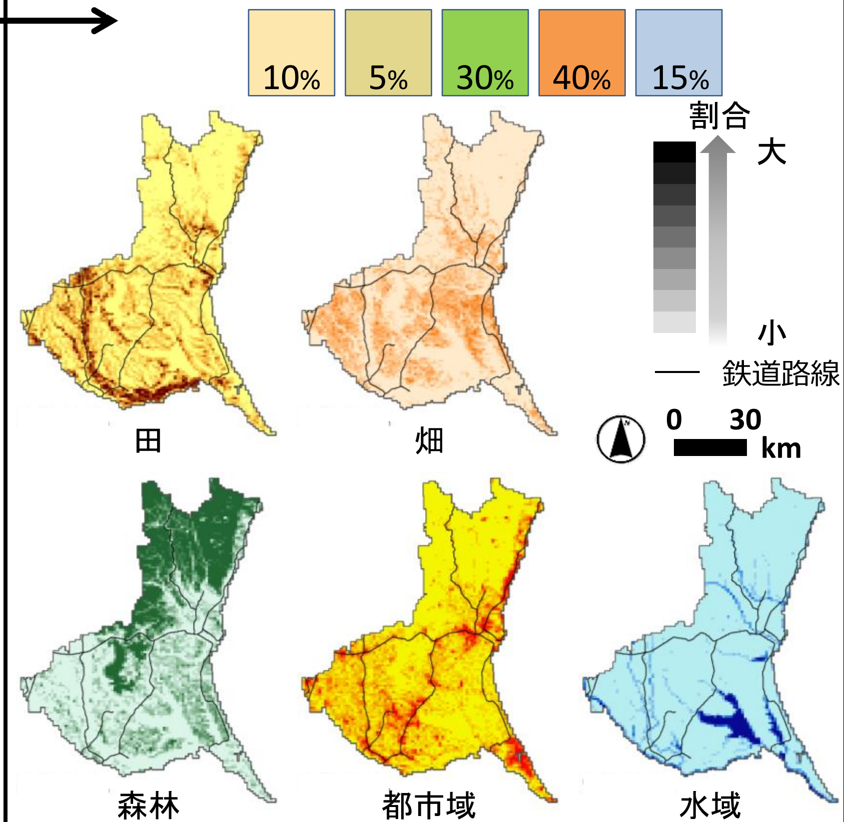
付録： 組成データが生じる場面の例（2）

代表値を1つ与える



図：茨城県の土地利用分布

各属性の割合を与える



図：土地利用組成の分布

分析モデル

- ➢ 離散選択モデル
- ➢ 組成データモデル に着目
  （⊃集計ロジットモデル）

付録： 空間データとその特性

- 空間データ
  - ➢ 地理的な位置情報をもつデータ
  - ➢ 例： 地価，標高，土地利用，人口，*etc.*

◆地理学の第一法則（Tobler, 1970）

- ➢ 「空間上の事物や現象は，互いの距離が近いほど強く影響し合う」
- ➢ 空間的相関

Tobler W.: A computer movie simulating urban growth in the Detroit region, *Economic Geography*, **46**, pp. 234–240, 1970.

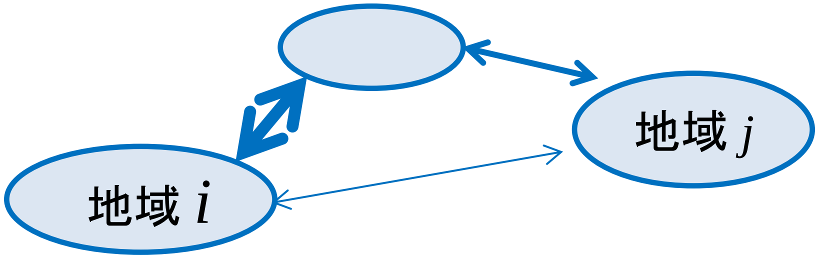
付録： 空間重み行列 W

- データ(地域)間における地理的な近接性を表現する $n \times n$ の行列

- 行列の要素 w_{ij} の与え方の例

$$w_{ij} = \left(\frac{1}{d_{ij}} \right)^2$$

$$w_{ij} = \begin{cases} 1, & \text{if } i \text{ is contiguous with } j \\ 0, & \text{otherwise} \end{cases}$$



線の太さの大小：関係性の強弱を表現