2012年度 S-PLUS 学生研究奨励賞応募

GeoSOMによるセグメンテーションを用いた不動産データのヘドニック分析

筑波大学 理工学群 社会工学類（主専攻：経営工学）

黒田　翔　（S. Kuroda）

不動産・空間計量研究室

Real **E**state & **S**patial **S**tatistics Laboratory

University of Tsukuba, Japan

発表概要

- 本研究の概要
  - 空間クラスタリングの一手法である GeoSOM による
    セグメンテーションを用いた不動産データのヘドニック分析を行う
  - S-PLUS を使用することで，多変量解析を容易に実行できるだけでなく
    SOM を柔軟に拡張し実装することが可能である
- 研究の意義
  - セグメントごとにヘドニック関数を推定し，モデルの精度を向上させる
- 発表内容
  - 背景： セグメンテーションの意義，既存研究
  - 方法： GeoSOM とその拡張
  - 実証： Boston Housing Data を用いた実証分析
  - 考察： 実証の考察と，今後の展望

セグメンテーションの意義と活用

- 不動産データのセグメンテーション
  - 地価・賃料予測モデルの精度を向上させる
    - 誤差項の空間的自己相関 (spatial auto-correlation) を防ぐ
    - 階層モデルに拡張する
  - データの対象地域における市場構造の理解に資する
- 既存のセグメンテーション
  - 地名（大字等）などを基準としている
    セグメンテーション（エリア分割）は
    実際の現況を反映しているのか不明

    （例：右図のセグメンテーション@千代田区）

* セグメンテーション (segmentation)： 本研究においては，データのカバーする地域を，地理空間的に連続な小地区に分割（≒クラスタリング）することを指す

空間クラスタリングの概要

- クラスタリング (clustering)
  - 類似したデータを分類する教師なし (unsupervised) 学習
  - タイプ： 階層，分割，密度ベース，グラフベース，格子ベース
- 空間クラスタリング (spatial clustering)
  - 地理空間（座標）的に連続な（飛び地のない）クラスタを生成
    （位相的な意味での「連結空間 (connected space)」への分割）
  - 分析の対象地域を，幾つかのクラスタに分割する
  - *spatial cluster*: “geographically bounded group” (Knox, 1989)

✓ *cf.* ホットスポット (hot spot) の検出

- 空間疫学，犯罪学，空間計量経済学の分野では “spatial clustering” という語がホットスポット検出の意味で用いられることが多い（本研究とは異なる）

空間クラスタリングに関する既存研究

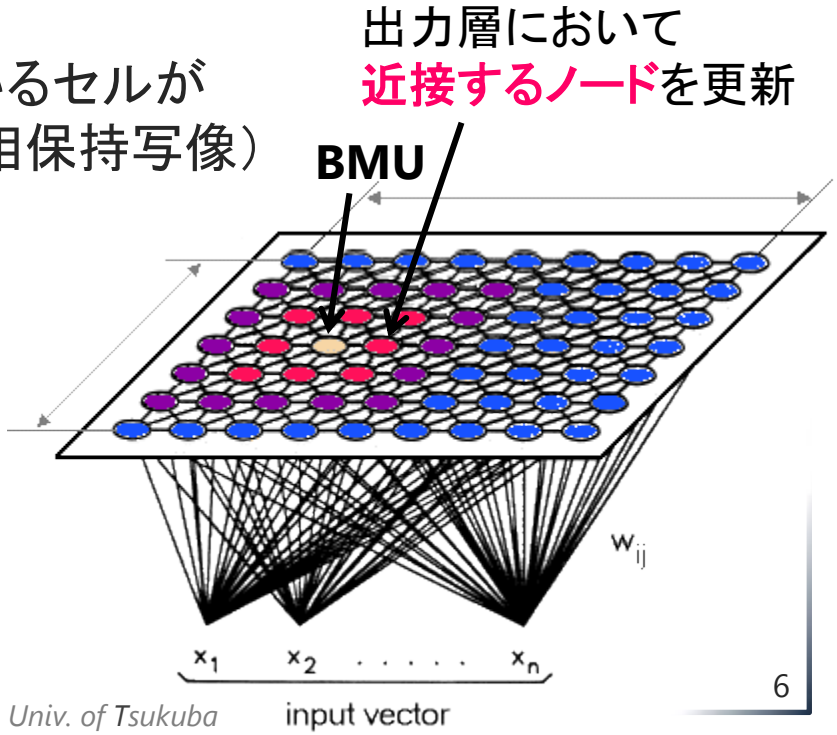
- 分野ごとのレビュー論文の例
  - 地理情報的側面 (regionalization)　Liu *et al.* (2012)
  - 不動産市場の分割 (market segmentation)　Islam & Asami (2009)
  - 行政的側面 (都市計画，選挙区 / zoning problem)　増山 (2009)
- Liu *et al.* (2012) によるアルゴリズムの比較　（抜粋）

Algorithm	Problem				
	arbitrary shaped	uneven density	robust to noise	NOT rely on prior knowledge	attribution and spatial proximity
k-means	×	×	×	√	×
Single-link	√	×	×	√	×
Complete-link	×	×	×	√	×
GDBSCAN	√	×	√	√	√
Geo-SOM	×	×	×	√	√

- 本研究では，空間的近接性を考慮したクラスタリングが可能な数少ない手法の一つである GeoSOM を用いる

SOM（自己組織化写像）

- Self-Organizing Maps
  - Kohonen (1982) などで提案
  - 入力データを任意の次元（2Dが主流）に写像する
  - 応用： クラスタリング，視覚化，画像処理，データマイニング，*etc.*
- 特徴
  - 学習後に得られるマップで隣接しているセルが入力データ空間上でも隣接する（位相保持写像）
- アルゴリズム （詳細は次頁；#7）
  - 入力ベクトルを出力層に学習させる（出力層のノードの値を更新する）
  - 勝者ユニット (BMU) から近いほど学習する度合いを強める



Source:
Self-Organizing Maps http://www.sis.pitt.edu/~ssyn/som/som.html

SOM のアルゴリズム

- 使用する変数とパラメータ
  - w_{ij}：出力層（マップ）の (i,j) 要素のノード（ニューロン）に関連付けられた重みベクトル
  - x_k：第 k 番目の入力ベクトル
  - h：近傍を定義する関数　　$e.g.\ h = \exp\left(-\frac{\text{勝者ユニット(BMU)と各セルの距離}^2}{2\sigma^2(t)}\right)$
  - α：学習割合を決める係数

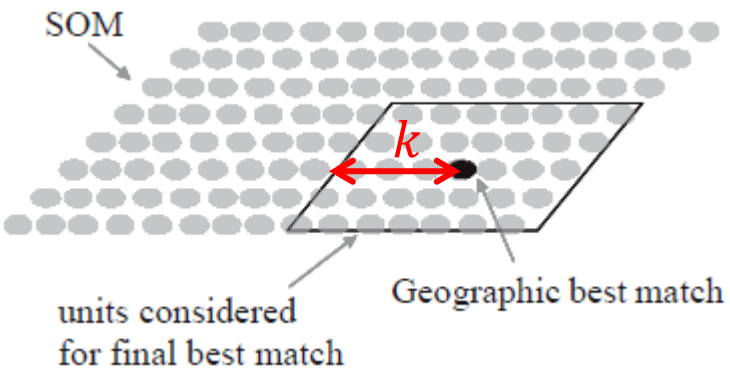
時間 t に関する減衰関数

- アルゴリズム
  1. 重みベクトル w_{ij} の初期化（ランダムに値を割り当てる）
  2. 繰り返し（規定回数 or 重みベクトルが収束するまで）
     1. 全入力データについて繰り返し（$k = 1,2, \ldots$）
        1. 全ての重みベクトルとの距離 $d_{ij} \coloneqq ||x_k - w_{ij}||$ を計算
        2. 値が最小となった重みベクトル $w_{ij}: d_{ij} \coloneqq \min(d_{mn})$ を勝者ユニットとする
        3. $w_{ij} \leftarrow w_{ij} + \alpha \cdot h \cdot ||x_k - w_{ij}||$ によって重みベクトルを更新する
     2. 学習係数 α を時間（ステップ）に対して減衰させる
        （例えば，0.9 から線形に 0 に収束させる）

前頁の BMU (best matching unit) のこと

GeoSOM（Geo＋SOM）

- Bacao *et al.* (2004, 2005) が提案
  - GeoSOM = “geographical SOM”
  - SOMを拡張したもので，
    空間的な連続性をもつクラスタリングが可能
- アルゴリズム
  - BMUを決定する前に，座標データのみの
    類似度によるgeo-BMUから地理的許容度 k の範囲で，
    BMUを選択する
  - $k = 0$ にすることで，
    空間的連続性が満たされる
    （地理座標のみによってBMUを選択）



Source:
Bacao *et al.* (2005) Fig. 2. Structure of a Geo-SOM.

ヘドニック分析

- 財の価格を，その属性の価格の和で表現する
  - Rosen (1974) が経済理論に基づいて展開
- 基本となるモデル

正規性を仮定することが多い

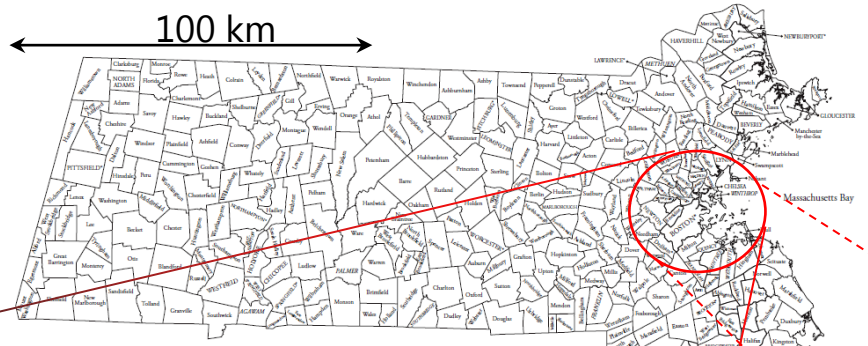
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I})$$

  - ただし，$\boldsymbol{y}$ は応答変数ベクトル，$\boldsymbol{X}$ は説明変数行列，$\boldsymbol{\beta}$ はパラメータベクトル，$\boldsymbol{\varepsilon}$ は *i.i.d.* の誤差ベクトル，σ^2 は誤差項の分散パラメータ
- モデルの評価
  - 誤差 (*e.g.* RMSE) や説明力 (*e.g.* R^2) ，パラメータの統計的有意性検定 (*t* test, *F* test) による

実証分析

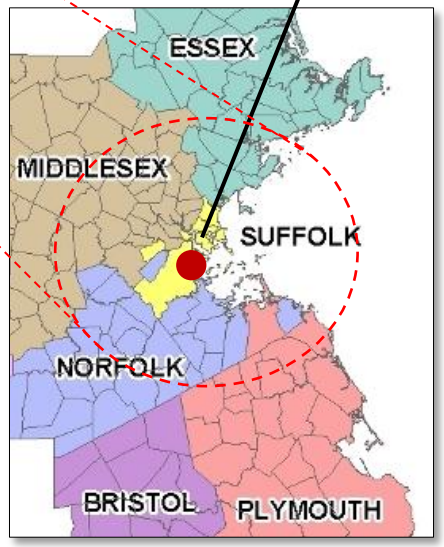
- データ： Boston Housing Data (Harrison & Rubinfeld, 1978)
  - 1970年 US Census 等に基づく住宅価格データ
  - 多くの実証で用いられてきた有名なデータセットで，様々なモデルや手法のベンチマークとして適切
  - Pace & Gilley (1997) によって地理座標が付加される
  - census tract（全506地域）ごとに集計されたデータを使用
    （可変地区単位問題が存在しうるが，本研究では考慮しない）
- 分析
  - セグメントごとのヘドニック関数推計
  - Cross Validation による予測精度（誤差）の推定

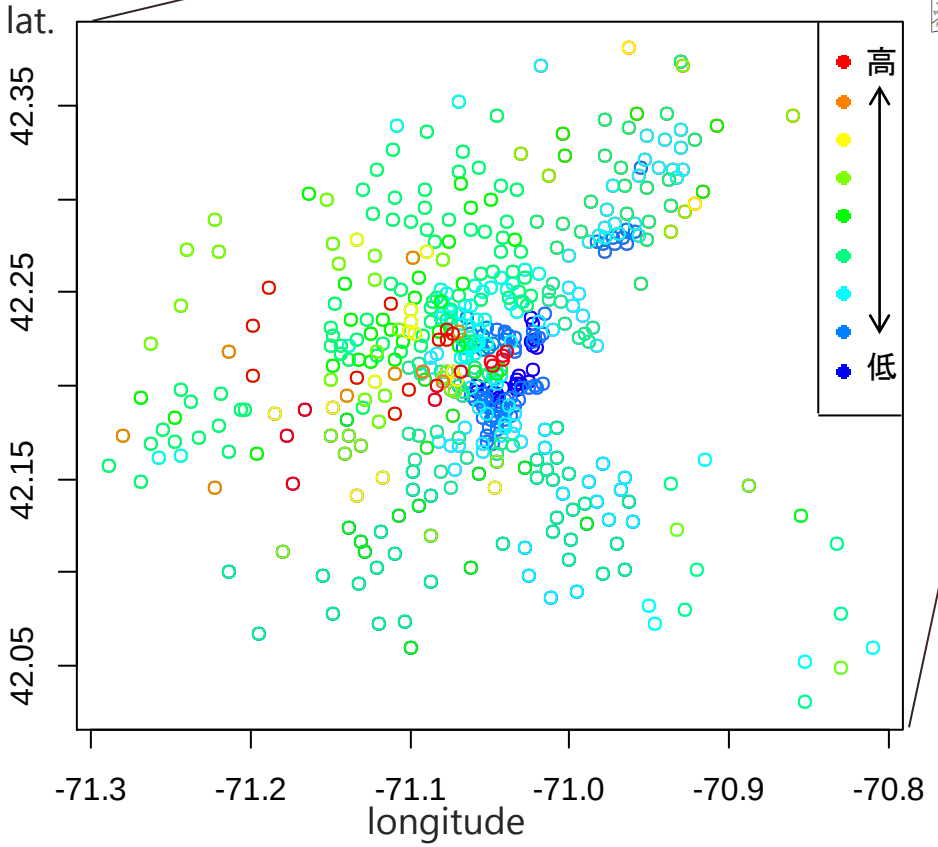
住宅価格の分布図



マサチューセッツ州

ボストン市





Source:
Massachusetts Cities & Towns
http://www.sec.state.ma.us/cis/cispdf/City_Town_Map.pdf
MassDEP
http://www.mass.gov/dep/water/resources/grtpond.htm

使用する変数

	desc.	Mean	S.D.	Min.	Max.
CMEDV	持家の価格（中央値，USD）	22,530	9,182	5,000	50,000
RM	部屋数	6.29	0.70	3.56	8.78
AGE	1940年以前建築の物件割合	68.57	28.15	2.90	100.00
LSTAT	lower status の割合	12.65	7.14	1.73	37.97
CRIM	犯罪率（詳細不明）	3.61	8.60	0.01	88.98
ZN	25,000 sq. ft. / lot を超える宅地割合	11.36	23.32	0.00	100.00
INDUS	（小売業以外の）商用地割合	11.14	6.86	0.46	27.74
TAX	固定資産税 ($/$10,000)	408.2	168.54	187.0	711.0
PTRATIO	児童の教師に対する割合	18.46	2.16	12.60	22.00
CHAS	Charles 川が tract の境界 (dummy)	該当しない: 471, 該当: 35			
DIS	employment centers までの距離	3.80	2.11	1.13	12.13
RAD	道路のアクセシビリティ指標	9.55	8.71	1	24
NOX	窒素酸化物の濃度	0.55	0.12	0.39	0.87
b	$1000(\text{黒人割合}-0.63)^2$	356.67	91.29	0.32	396.90

セグメントごとの分析：郡別

pooled: all of sample
each county
- Middlesex
- Essex
- Suffolk
- Norfolk

* Plymouth は除外
（サンプル数14）

red coef.: sig. at 5%

有意な変数の符号は
直観に整合する

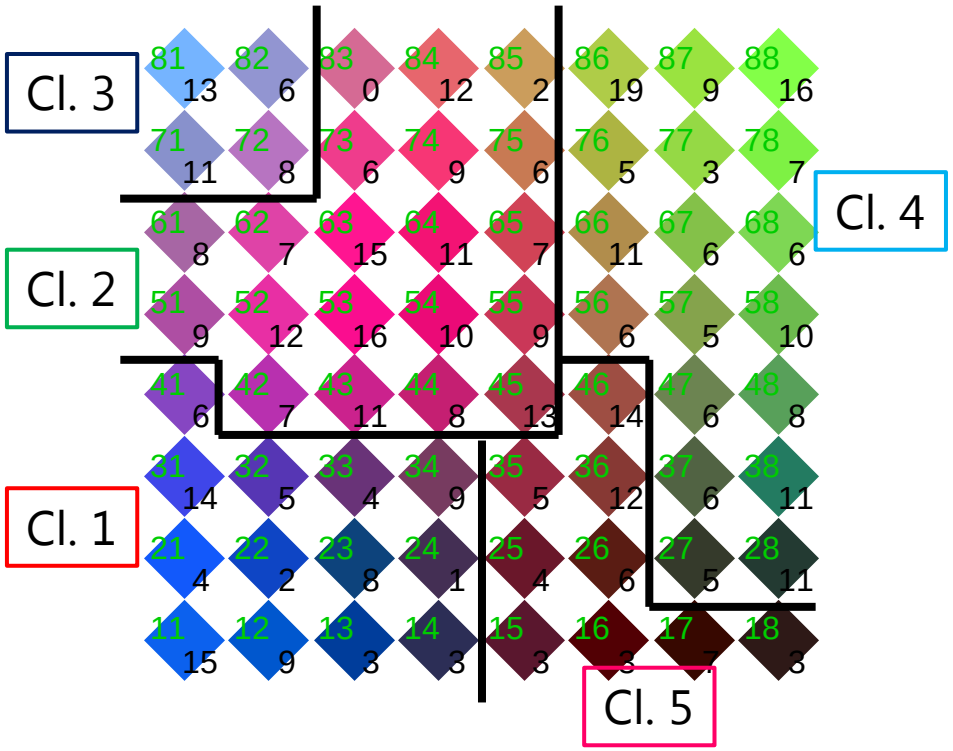
郡 (county) によって
価格形成要因が
大きく異なる

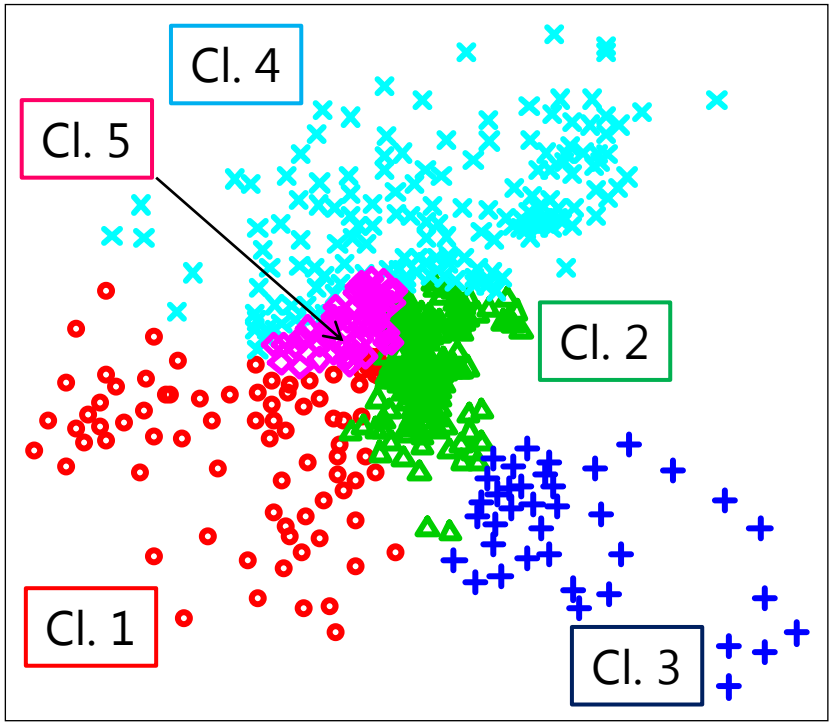
サフォーク郡では
モデルの精度が低い

	pooled	Mid.	Ess.	Suf.	Nor.
切片	6.99	4.63	8.47	5.59	1.85
CRIM	-0.011	0.039	-0.16	-0.0098	0.013
ZN	0.00088	0.00030	0.000088	NA	-0.00022
INDUS	0.0040	-0.00052	-0.012	-0.042	0.0027
CHAS	0.091	0.017	–	0.32	0.031
NOX	-0.77	-0.62	1.53	-1.70	-0.40
RM	-0.86	-0.34	-2.10	-0.28	0.50
RM^2	0.074	0.044	0.19	0.017	-0.014
AGE	0.000072	-0.0013	-0.0031	0.0013	-0.0024
DIS	-0.041	-0.039	0.012	-0.058	-0.042
RAD	0.013	0.0054	-0.0070	0.0034	0.0057
TAX	-0.00061	-0.00064	-0.00030	0.0012	-0.00095
PTRATIO	-0.031	-0.024	-0.0015	NA	-0.030
b	0.00036	0.00059	0.00049	0.00020	0.00012
LSTAT	-0.030	-0.012	-0.0094	-0.042	-0.0057
# of sample	506	192	65	150	85
R^2	0.82	0.89	0.93	0.73	0.96
Adj. R^2	0.82	0.88	0.91	0.70	0.96
RMSE	0.0296	0.0105	0.00510	0.0527	0.00302

GeoSOM による Segm.

- パラメータ
  - 地理的許容度 $k = 0$, 学習係数 $\alpha = 0.1$から0.01 まで線形に減少
    出力層：格子状（8×8），繰り返し：50回
- 学習には，使用できる変数（#12）を全て使用 （セグメントは5つとする）
  - ただし，$k = 0$ であることから，BMUは位置座標によってのみ決定することに留意する





セグメントごとの分析 （GeoSOMを使用）

前頁 (#14) で決めた5つのセグメントごとにヘドニック関数を推計

※ Cl. 4 では大気汚染の指標であるNOXの符号が正で直観にそぐわないが，それ以外は概ね可

※ サフォーク郡を含むCl. 2の決定係数が悪いことは，郡別の分析 (#13) とも整合

	pooled	Cl. 1	Cl. 2	Cl. 3	Cl. 4	Cl. 5
切片	6.99	2.70	6.94	8.90	4.68	4.60
CRIM	-0.011	-0.011	-0.010	0.21	-0.20	0.0070
ZN	0.00088	0.00095	0.0051	0.00052	0.00072	-0.0036
INDUS	0.0040	0.0020	0.013	0.011	-0.0079	0.0095
CHAS	0.091	0.039	0.22	-	0.16	-0.045
NOX	-0.77	0.27	-1.46	-1.00	0.80	-0.68
RM	-0.86	-0.075	-0.44	-2.10	-0.69	-0.12
RM^2	0.074	0.026	0.033	0.20	0.075	0.024
AGE	0.000072	-0.0018	0.0020	-0.0022	-0.0034	-0.0039
DIS	-0.041	-0.023	-0.061	-0.0074	-0.027	-0.15
RAD	0.013	0.0090	0.019	0.032	0.0046	0.024
TAX	-0.00061	-0.00049	-0.00076	-0.00052	-0.00060	-0.0014
PTRATIO	-0.031	-0.013	-0.066	0.0097	-0.0048	-0.019
b	0.00036	0.00094	0.00023	-0.00081	0.00035	0.00084
LSTAT	-0.030	-0.0098	-0.037	0.0061	-0.0037	-0.024
# of sample	506	83	170	38	158	57
R^2	0.82	0.93	0.72	0.94	0.90	0.93
Adj. R^2	0.82	0.91	0.69	0.91	0.89	0.91
RMSE	0.0296	0.00553	0.0543	0.00173	0.00676	0.00800

考察

- 郡別での分析 (#13) との比較
  - R^2, RMSEともに僅かな向上しか観察されなかったが，これは郡内で価格決定要因が同質的で郡区分け（郡境の決定）が適切であることを示唆
  - モデルの精度が大幅に改善したわけではなく，アルゴリズムの大幅な改良や開発が必須であることが明らかとなった
- GeoSOMによるセグメントごとの分析
  - 地理的に隣接していないセグメント (Cl. 3 と Cl. 4) では価格構造が類似しており，隣接するセグメント同士では価格構造にかなりの相違があることから，同質な価格形成要因ごとセグメントの導出としてGeoSOMは適切であったと考えられる
  - ただしSOMの学習部分には位置座標以外の属性は用いられておらず，その他の属性は出力層（マップ）からセグメンテーションをする際のみに用いられているので，空間的連続性を満たし且つその他の属性の類似性も学習プロセスに組み込むような拡張が要請される

参考文献

- Bacao, F., Lobo, V., Painho, M. (2004) Geo-Self-Organizing Map (Geo-SOM) for Building and Exploring Homogeneous Regions, *Geographic Information Science*, Proceedings. Lecture Notes in Computer Science, **3234**, 22-37.
- Bacao, F., Lobo, V., Painho, M. (2005) The self-organizing map,the Geo-SOM, and relevant variants for geosciences, *Computers & Geosciences*, **31**, 155-163.
- Rosen, S. (1974) Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition, *The Journal of Political Economy*, **82** (1), 34-55.
- Harrison, D., Rubinfeld, D.L. (1978) Hedonic housing prices and the demand for clean air, *Journal of Environmental Economics and Management*, **5** (1), 81-102.
- Pace, P.K., Gilley, O.W. (1997) Using the Spatial Configuration of the Data to Improve Estimation, *The Journal of Real Estate Finance and Economics*, **14** (3), 333-340.
- Knox EG (1989) Detection of clusters, In *Methodology of enquiries into disease clustering*; London. (ed: Elliott P.), 17-20.
- Liu, Q., Deng, M., Shi, Y., Wang, J. (2012) A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity, *Computers & Geosciences*, **46**, 296-309.
- Islam, K.S., Asami, Y. (2009) HOUSING MARKET SEGMENTATION: A REVIEW, *Review of Urban & Regional Development Studies*, **21**, 93-109.
- 増山篤 (2009) 都市計画およびその周辺分野における地域区分方法, 都市計画報告集, **8** (2), 106-113.

【補】S-PLUSコード（GeoSOM部分のみ）

引数　input_data：入力データ（データフレーム型を想定）
xycoord：入力データの何列目が位置座標であるか
iteration：繰り返しの数　※ この実装では収束判定は行わない
k：地理的許容度（$k = 0$ で地理的連続性が満たされる）

```
GeoSOM <- function(input_data, xycoord, iteration, k) {
  sizeN <- 8
  alpha <- seq(.9, .01, length = iteration)
  dimention <- c(sizeN, sizeN, dim(input_data)[2])
  neuron <- array(rnorm(prod(dimention)), dim = dimention)
  neuron_x <- matrix(rep(1:sizeN, sizeN), byrow = T, nrow = sizeN)
  neuron_y <- matrix(rep(1:sizeN, sizeN), nrow = sizeN)

  for (i in 1:iteration) {
    for (j in 1:dim(input_data)[1]) {
      dst <- (input_data[j, xycoord[1]] - neuron[,, xycoord[1]])^2 + (input_data[j, xycoord[2]] - neuron[,, xycoord[2]])^2
      mindst <- dst == min(dst)
      geoBMU <- c(sum(neuron_x * mindst), sum(neuron_y * mindst))

      cand_BMU <- sqrt((neuron_x - geoBMU[1])^2 + (neuron_y - geoBMU[2])^2) <= k
      dst <- matrix(0, nrow = sizeN, ncol = sizeN)
      for(attr in 1:dim(input_data)[2]) dst <- dst + sqrt((input_data[j, attr] - neuron[,, attr])^2)
      dst[!cand_BMU] <- 999
      mindst <- dst == min(dst)
      BMU <- c(sum(neuron_x * mindst), sum(neuron_y * mindst))
      dst <- sqrt((neuron_x - BMU[1])^2 + (neuron_y - BMU[2])^2)

      for (attr in 1:dim(input_data)[2]) neuron[,, attr] <- neuron[,, attr] + alpha[i]/(dst+1)*(input_data[j, attr] - neuron[,, attr])
    }
  }
  return(neuron)
}
```

sizeN：出力層（マップ）の一辺のノードの数
※ 出力層は正方形で，格子型を実装
alpha：学習係数 α　～　0.9 から 0.01 まで線形に減少
dimention：出力層（マップ）neuron の次元
neuron：出力層（マップ）
※ この実装において，重みベクトルを直接表現

dst：重みベクトル w_{ij} の計算に相当
geoBMU：位置座標のみを用いたときのBMU

cand_BMU：geoBMUから
半径＝地理的許容度の範囲内（BMUの候補）
BMU：（真の）勝者ユニットBMU

この部分で学習（重みベクトル = neuron の更新）

返り値　出力層（マップ）