証拠に基づく政策立案のための潜在プロファイル分析と数法則発見法を用いた社会実情のモデル化と可視化

Modeling and Visualization of Social Reality
Using Latent Profile Analysis and Number Law Discovery Methods
for Evidence-Based Policy Making

長瀬 永遠 (Towa Nagase)
u255013@st.pu-toyama.ac.jp

富山県立大学大学院 工学研究科 電子・情報工学専攻
情報基盤工学講座

N212, 09:30-10:00 Tuesday, February 13, 2024.

本研究の背景 1

情報技術の発達により，社会における様々なデータを観測・収集することが可能に

→ 政策分野においても，EBPM に注目が集まる

EBPM とは

政策における意思決定をデータに基づいて行うという考え方

課題

大規模なデータから有用な情報を取り出すデータマイニングの技術が必要

本研究の背景 2

どんな情報を取り出すことが有効？

政策分野では，原因と結果の間に成り立つ関係性が重要

→ 複数の要因が複雑に影響しあうため，人間が把握するには限界がある

アプローチ

行政が持つ統計データを用いて分析を行い，データ間の関係性を数理モデルによって表す

行政が持つ統計データ

地域経済分析システム（RESAS）

経済に関する項目を中心に自治体単位でデータを公開するオープンデータサイト

Table 1: RESAS から自動取得可能なデータ

データ項目	
農地平均取引価格	農業産出額
林地平均取引価格	海面漁獲物等販売金額
住宅用地平均取引価格	林産物販売金額
商業用地平均取引価格	林作業請負収入
マンション等平均取引価格	企業数
一人あたりの固定資産税	事業所数
一人当たりの地方税	就業者数
一人当たりの法人住民税	総人口
製造品出荷額	経営耕地面積
年間商品販売額	etc

自治体区分	総数
市	792
町	743
村	183
特別区	23
合計	1741

疑問

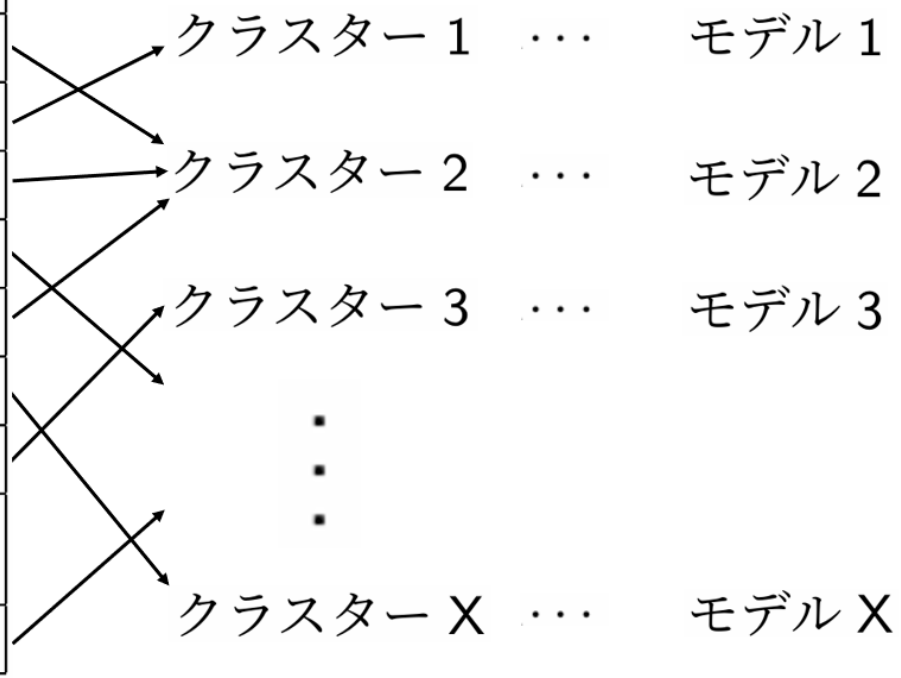
すべての市区町村を同じ尺度のサンプルとして扱うのは適切？

クラスタリングと回帰分析を用いたモデル化

目的

行政が持つ統計データを用いて自治体をクラスタリングし，その結果を考慮しながら統計データ間の関係性をモデル化する手法を提案する．

自治体名	統計データ
A 市	$(A_1, A_2, \cdots)$
B 町	$(B_1, B_2, \cdots)$
C 村	$(C_1, C_2, \cdots)$
D 区	$(D_1, D_2, \cdots)$
E 町	$(E_1, E_2, \cdots)$
F 村	$(F_1, F_2, \cdots)$
G 市	$(G_1, G_2, \cdots)$
⋮	⋮
Z 市	$(Z_1, Z_2, \cdots)$



本研究におけるクラスタリングと回帰分析の概要

クラスタリング手法

潜在プロファイル分析（Latent Profile Analysis: LPA）を採用する．一般的なクラスタリング分析はデータ間の距離に基づいてサンプルを分割するため，クラスター自体に意味を持たない．一方，LPA ではサンプルの潜在的な特徴に基づいてクラスターを作成する．

回帰分析

多変量多項式回帰の一つである RF6.4 法（Rule extraction method from Fact 6.4）を採用する．RF6.4 法を用いることで汎化性と可読性を両立した数理モデルを求めることができる．

潜在変数を考慮したクラスタリング1

LPA

観測されたデータを混合分布モデルとして扱い，モデルの背後に存在する潜在変数を分析．その結果を考慮しながらデータをクラスタリングする手法である．

（混合ガウス分布）

$$f(\mathbf{y}_i|\mathbf{\Theta}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i|\mu_k, \mathbf{\Sigma}_{rk}) \tag{1}$$

K：潜在変数の数，π_k：母集団における潜在変数の構成割合，
$\mathbf{y}_i = (y_1, \cdots, y_r)_i \quad i = 1, \cdots, I$：観測変数，
$\mu_k = (\mu_{1k}, \cdots, \mu_{rk})$：観測変数の平均，$\mathbf{\Sigma}_{rk}$：各観測変数の共分散行列

Σ_{rk}

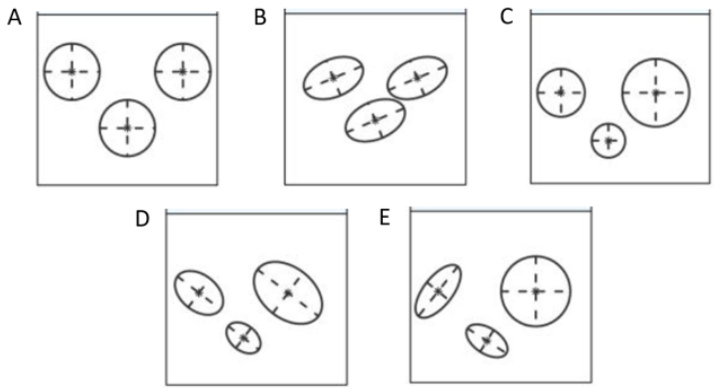
A $\begin{bmatrix} \sigma_1^2 & & & \\ 0 & \sigma_2^2 & & \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & \sigma_r^2 \end{bmatrix}$

B $\begin{bmatrix} \sigma_1^2 & & & \\ \sigma_{21} & \sigma_2^2 & & \\ \vdots & \vdots & \ddots & \\ \sigma_{r1} & \sigma_{r2} & \cdots & \sigma_r^2 \end{bmatrix}$

C $\begin{bmatrix} \sigma_{1k}^2 & & & \\ 0 & \sigma_{2k}^2 & & \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & \sigma_{rk}^2 \end{bmatrix}$

D $\begin{bmatrix} \sigma_{1k}^2 & & & \\ \sigma_{21} & \sigma_{2k}^2 & & \\ \vdots & \vdots & \ddots & \\ \sigma_{r1} & \sigma_{r2} & \cdots & \sigma_{rk}^2 \end{bmatrix}$

E $\begin{bmatrix} \sigma_{1k}^2 & & & \\ \sigma_{21k} & \sigma_{2k}^2 & & \\ \vdots & \vdots & \ddots & \\ \sigma_{r1k} & \sigma_{r2k} & \cdots & \sigma_{rk}^2 \end{bmatrix}$



潜在プロファイル分析 2

パラメータ推定

観測変数 $\mathbf{y}$ と潜在変数 z の同時確率の尤度を目的関数とし，その最大値を求める．値の最大化には EM アルゴリズムを用いる．

（EM アルゴリズム）

- **E ステップ**
  その時点で最適と考えれるパラメータを式 (2) に代入することによって $E_{z_{ik}}[z_{ik}]$ の値を求める．

$$E_{z_{ik}}[z_{ik}] = \frac{\pi_k f_k(\mathbf{y}_i|\mu_k, \Sigma_{rk})}{\sum_k \pi_k f_k(\mathbf{y}_i|\mu_k, \Sigma_{rk})} \tag{2}$$

- **M ステップ**
  E ステップで求めた $E_{z_{ik}}[z_{ik}]$ を式 (3) の右辺に代入し，式 (3) が最大となるようなパラメータを求める．求めたパラメータをその時点で最適なパラメータとし，再度 E ステップに戻る．

$$E_z[\ln p(\mathbf{y}, z|\pi, \mu, \sigma)] = \sum_i \sum_k E_{z_{ik}}[z_{ik}](\ln \pi_k + \ln f_k(\mathbf{y}_i|\mu_k, \Sigma_{rk})) \tag{3}$$

モデル選択

ベイズ情報量基準 (Bayesian information criterion: BIC) を用いて，最適なパラメータとクラスター数を決定.

$$BIC = -2L + k \ln n \tag{4}$$

L: EM アルゴリズムで最大化した値，k: クラスター数，n: サンプル数

各サンプルにおける存在確率の算出

最適なモデルが決定した後，各サンプルにおける観測変数ベクトル $\mathbf{y_i}$ と最適なパラメータを用いて存在確率を算出.

$$\pi_{k|\mathbf{y}_i} = \frac{\pi_k f_k(\mathbf{y_i}|\mu_k, \mathbf{\Sigma}_{rk})}{\sum_k \pi_k f_k(\mathbf{y}_i|\mu_k, \mathbf{\Sigma}_{rk})} \tag{5}$$

パーセプトロンを用いた数法則発見法 1

RF6.4 法

目的変数と M 個の量的説明変数，K 個の質的説明変数からなるデータセットを与え，4 層パーセプトロンの学習を用いてパラメータ $c_{0d}, c_{gd}, c_{dkl}, v_{gm}$ を最適化することで質的条件付き多変量多項式を発見する．

（質的条件付き多変量多項式）

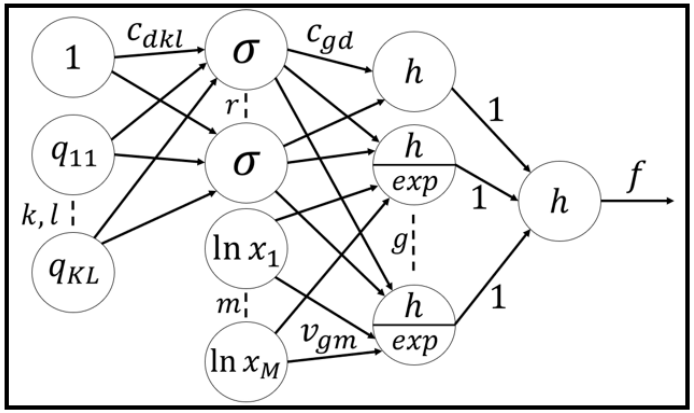
$$if \quad \bigwedge_{k} \bigvee_{q_{kl} \in Q_k^i} q_{kl} \quad then \quad y = v_0^i + \sum_{g=1}^{G} v_g^i \prod_{m=1}^{M} x_m^{v_{gm}^i} + \epsilon, \qquad i = 1, \cdots, I \quad (6)$$

RF6.4 法の 4 層パーセプトロン

$$f(\mathbf{q}, \mathbf{x}; \phi) = v_0 + \sum_{g=1}^{G} v_g s_g,$$

$$v_0 = \sum_{d=1}^{D} c_{0d}\sigma_d, \quad v_g = \sum_{d=1}^{D} c_{gd}\sigma_d,$$

$$\sigma_d = \sigma\left(\sum_{k=1}^{K}\sum_{l=1}^{L_k} c_{dkl} q_{kl}\right), \quad s_g = \exp\left(\sum_{m=1}^{M} v_{gm} \ln x_m\right)$$



パーセプトロンを用いた数法則発見法 2

パーセプトロンの学習方法

RF6.4 法の学習には，準ニュートン法を基本枠組みとする学習アルゴリズムである BPQ（Back propagation based on Partial Quasi-Newton）法を用いる．

（最小化する関数）

$$E(\mathbf{q}, \mathbf{x}; \phi) = \frac{1}{2} \sum_{n=1}^{N} (f(\mathbf{q}, \mathbf{x}; \phi)^n - y^n)^2 \tag{7}$$

Step 1:パラメータの初期化

ϕ_1 を初期化．$\mathbf{H}_1 = \mathbf{I}, b = 1$ とおく．$\mathbf{I}$ は ϕ と次元の等しい単位行列

Step 2: 探索方向 $\Delta\phi_s$ の計算

探索方向 $\Delta\phi_s = -\mathbf{H}_s \nabla E(\mathbf{q}, \mathbf{x}; \phi_s)$ を計算．$\nabla E(\mathbf{q}, \mathbf{x}; \phi_s)$ の全要素が 0.0001 未満の場合，反復を終了．

Step 3: 最適探索幅 α_s の計算

$E(\mathbf{x}; \phi_b + \alpha_b \Delta\phi_b)$ を最小にする最適探索幅 α_b を計算．

Step 4: 結合重み ϕ_s の更新

$\boldsymbol{\phi_{b+1}} = \boldsymbol{\phi_b} + \boldsymbol{\alpha_s \Delta \phi_b}$ に従って結合重み ϕ_b を更新．

Step 5: 二次微分の逆行列の近似値 $\mathbf{H}$ の更新

$b \equiv 0 \pmod{Z}, Z$：全パラメータ数) のとき，$\mathbf{H}_{b+1} = \mathbf{I}$ とし，それ以外のとき，BFGS 公式で $\mathbf{H}_{b+1}$ を更新．$b \leftarrow b + 1$ として，Step2 の戻る．

パーセプトロンを用いた数法則発見法 3

パーセプトロンの最適中間層数

BPQ 法では，事前に G, D の値が必要なため，$G = 1, \cdots, 5$，$D = 1, \cdots, 5$ として 25 回試行し，BIC（ベイズ情報量基準）が最小の結果を採用する．

$$BIC(G, D) = \frac{N}{2} \ln \left(\frac{1}{N} \sum_{n=1}^{N} (f(\mathbf{q}^n, f(\mathbf{x}^n; \hat{\phi}_{G,D}) - y^n)^2 \right) + \frac{Z}{2} \ln N \quad (8)$$

N：サンプル数，Z：パラメータ数

ルールの復元

パーセプトロンの学習結果では，各サンプルに対するパラメータベクトルが得られるが，より一般的な数法則として記述するために，以下の処理を行う．

1. 各サンプルに対する数法則の係数値ベクトル $\mathbf{c}^\mu = (c_0^\mu, \cdots, c_J^\mu)$ を求める
2. k-means 法を用いて I 個のクラスターを求め，それぞれの重心ベクトル $\mathbf{a}^I = (a_0^i, \cdots, c_J^i)$ を数法則の係数値とする

潜在的クラスターを考慮した数理モデル化手法

提案手法

自治体名	統計データ
A 市	$(A_1, A_2, \cdots, Y_A)$
B 町	$(B_1, B_2, \cdots, Y_B)$
C 区	$(C_1, C_2, \cdots, Y_C)$
⋮	⋮
Z 村	$(Z_1, Z_2, \cdots, Y_Z)$

LPA

所属確率 1	所属確率 2	⋯	所属確率 X
0.9	0.1	⋯	0
0.3	0	⋯	0.1
0.5	0.5	⋯	0
⋮	⋮	⋮	⋮
0	0	⋯	1.0

$$\tilde{y} = \frac{y - mean(y)}{std(y)} \quad \tilde{x} = \frac{x}{max(x)}$$

量子化

目的変数	量的説明変数
$\widetilde{Y_A}$	$(\widetilde{A_1}, \widetilde{A_2}, \cdots)$
$\widetilde{Y_B}$	$(\widetilde{B_1}, \widetilde{B_2}, \cdots)$
$\widetilde{Y_C}$	$(\widetilde{C_1}, \widetilde{C_2}, \cdots)$
⋮	⋮
$\widetilde{Y_Z}$	$(\widetilde{Z_1}, \widetilde{Z_2}, \cdots)$

質的説明変数 1	質的説明変数 2	⋯	質的説明変数 X
[1, 0, ⋯,0]	[0, 0, ⋯,1]	⋯	[0, 0, ⋯,1]
[0, 0, ⋯,0]	[0, 0, ⋯,1]	⋯	[0, 0, ⋯,1]
[0, 1, ⋯,0]	[0, 1, ⋯,0]	⋯	[0, 0, ⋯,1]
⋮	⋮	⋮	⋮
[0, 0, ⋯,1]	[0, 0, ⋯,1]	⋯	[1, 0, ⋯,0]

RF6.4法

$$Y_1 = v_{1\_0} + \sum_{g=1}^{G} v_{1\_g} \prod_{m=1}^{M} x_m^{v_{gm}} \quad Y_2 = v_{2\_0} + \sum_{g=1}^{G} v_{2\_g} \prod_{m=1}^{M} x_m^{v_{gm}} \quad \cdots \quad Y_X = v_{X\_0} + \sum_{g=1}^{G} v_{X\_g} \prod_{m=1}^{M} x_m^{v_{gm}}$$

GIS による分析結果の可視化 1

各自治体が所属するクラスターと対応する数理モデルを視覚的に提示するため，GIS を用いたシステムを作成

GIS とは

地理情報システム（Geographic Information System）の略称．地図上にデータを表示し，可視化や重ね合わせなどを行うことが可能

GIS を用いる狙い

- 分析結果全体を俯瞰的に把握
- クラスターの分布と地理的特徴の関係を可視化

動画

提案システムのながれを動画でお見せします.

数値実験の概要

実験目的

自治体を潜在的な要素に基づいてクラスタリングすることがモデルの精度向上に有効か

比較する手法

- 提案手法（LPA ＋ RF6.4 法）
- 量的説明変数用いた手法（RF5 法）

データ

項目 ： 目的変数と 10 個の説明変数
サンプル数 ： データ欠損のない 650 の自治体におけるオープンデータ
対象年 ：2020 年
参照元 ： 地域経済分析システム-RESAS-

比較方法

- サンプルをランダムに 520 個の学習データと 130 個のテストデータに分割
- 学習データを用いてモデルを作成し，テストデータとの二乗誤差の平均を比較

数値実験の結果と考察

おわりに

まとめ

行政が持つ統計データを用いて，自治体をクラスタリングし，データの関係を数理モデルで表す手法を提案した.

- クラスタリングを用いない場合と比較してモデルの精度が向上した
- 一部のクラスターに対して大きい誤差が見られた

今後

- 項目数，サンプル数ともにより大規模なデータを用いた議論が必要
- 得られたクラスターやモデルに対する解釈が必要