# 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

## 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？

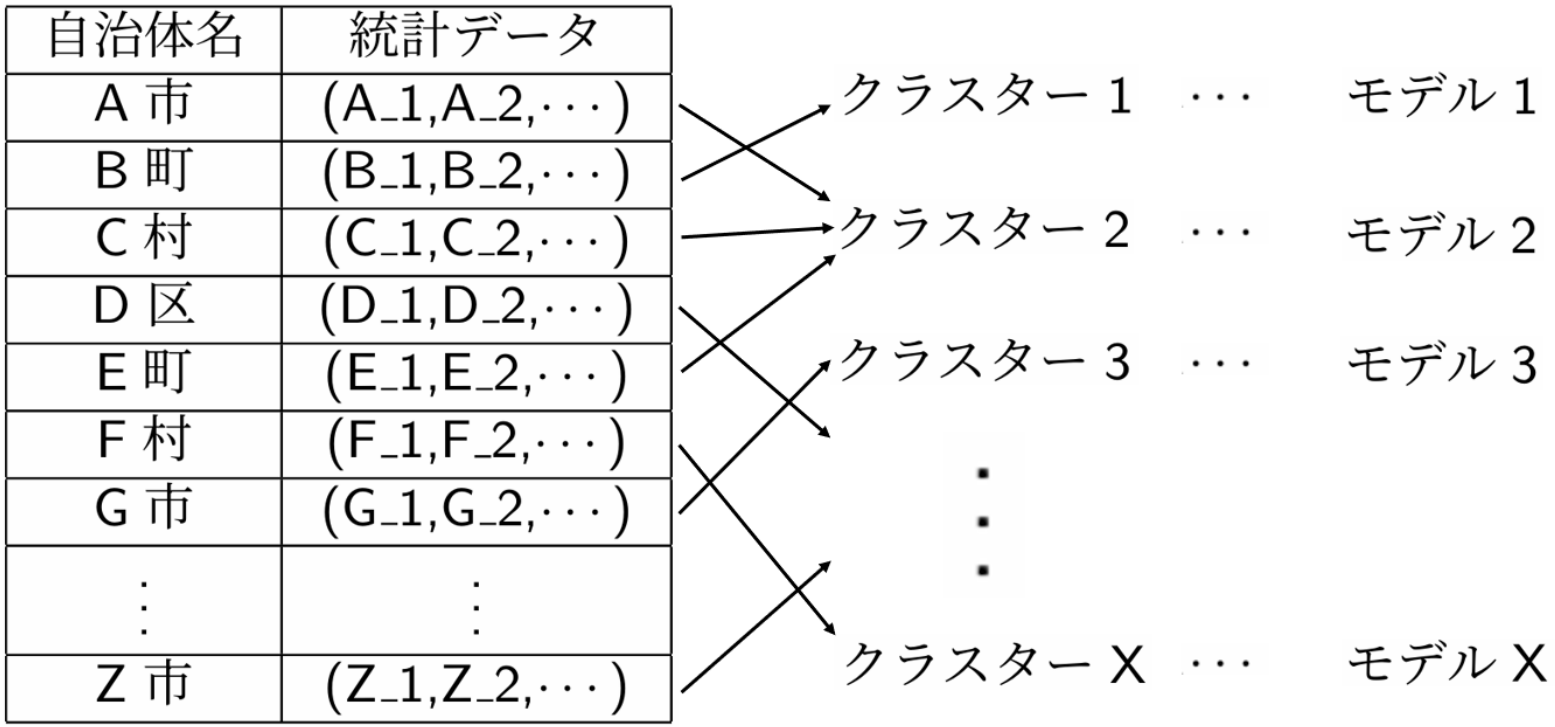
# クラスタリングと回帰分析を用いたモデル化





## 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．



# クラスタリングと回帰分析の概要

# 潜在変数を考慮したクラスタリング1





## 潜在プロファイル分析（Latent Profile Analysis: 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．

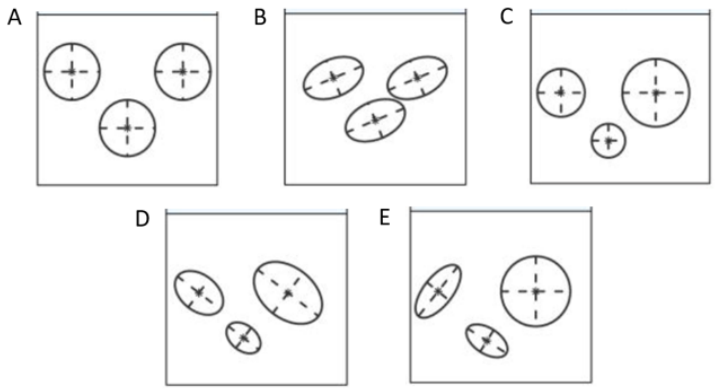
（混合ガウス分布）

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

$K$：潜在変数の数，$\pi_k$：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

$\Sigma_{rk}$

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．

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

# 潜在プロファイル分析 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（Rule extraction method from Fact 6.4）

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）

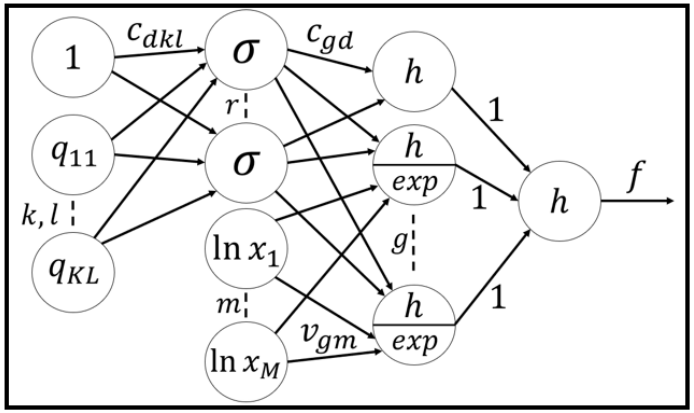
$$if \quad \bigwedge_{k} \bigvee_{q_{kl} \in Q_k^i} q_{kl} \quad then \quad y = v_0^i + \sum_{g=1}^{G} v_g^i \prod_{m=1}^{M} x_m^{v_{gm}^i} + \epsilon, \qquad i = 1, \cdots, I \quad (6)$$

## 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**

$\phi_1$ を初期化．$\mathbf{H}_1 = \mathbf{I}, b = 1$ とおく．$\mathbf{I}$ は $\phi$ 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 $\alpha_s$ の計算**

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

**Step 4: 結合重み $\phi_s$ の更新**

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

以降のスライドは未編集です

# データベース作成と因果探索によるデータの選定





データベース内には地理情報を持たない数値データ，地理情報を持つ数値データ，施設等の位置データがある．データは RESAS API および国土交通省のウェブサイトを用いて収集した．それらを因果探索で絞りこみ，DEA の入力・出力に振り分ける．

表1　地理情報を持たない数値データ

| データ項目 | 単位 | データ項目 | 単位 |
|---|---|---|---|
| 耕作放棄地率 | % | 経営耕地面積 | 1 畝 / 経営体 |
| 農業産出額 | 千万円 | 労働生産性 | なし |
| 企業数 | 社 | 従業員数 | 人 |
| 歳出決算額 [総務費] | % | 農地平均取引価格 | 円 /$m^2$ |
| 歳出決算額 [民生費] | % | 商業用地平均取引価格 | 円 /$m^2$ |
| 歳出決算額 [衛生費] | % | 住宅用地平均取引価格 | 円 /$m^2$ |
| 歳出決算額 [農林水産業費] | % | 林地平均取引価格 | 円 /$m^2$ |
| 歳出決算額 [商工費] | % | マンション等平均取引価格 | 円 /$m^2$ |
| 歳出決算額 [土木費] | % | 1 人あたりの地方税 | 千円 |
| 歳出決算額 [警察費・消防費] | % | 製造品出荷額 | 万円 |
| 歳出決算額 [教育費] | % | 事業所数 | 事業所 |
| 歳出決算額 [公債費] | % | 総人口 | 人 |
| 歳出決算額 [労務費] | % | 老年人口 | % |
| 歳出決算額 [その他（雑費）] | % | 生産年齢 | % |
| 農業就業人口平均年齢 | 歳 | 年少人口 | % |
| 農業経営者平均年齢 | 歳 | 年間商品販売額 | 百万円 |
| 林作業請負収入 | 万円 | 海面漁獲物等販売額 | 万円 |
| 林産物販売金額 | 万円 | 付加価値額 | 万円 |
| 一人当たりの法人住民税 | 千円 | 1 人あたりの固定資産税 | 千円 |

表2　地理情報を持つ数値データ

| データ項目 | 単位 |
|---|---|
| 施設位置 [空港] | 経度・緯度 |
| 施設位置 [工業団地] | 経度・緯度 |
| 施設位置 [都市公園] | 経度・緯度 |
| 施設位置 [道の駅] | 経度・緯度 |
| 施設位置 [学校] | 経度・緯度 |

表3　位置データ

| データ項目 | 単位 |
|---|---|
| 施設数 [空港] | 箇所 |
| 施設数 [工業団地] | 箇所 |
| 施設数 [都市公園] | 箇所 |
| 施設数 [道の駅] | 箇所 |
| 施設数 [学校] | 箇所 |

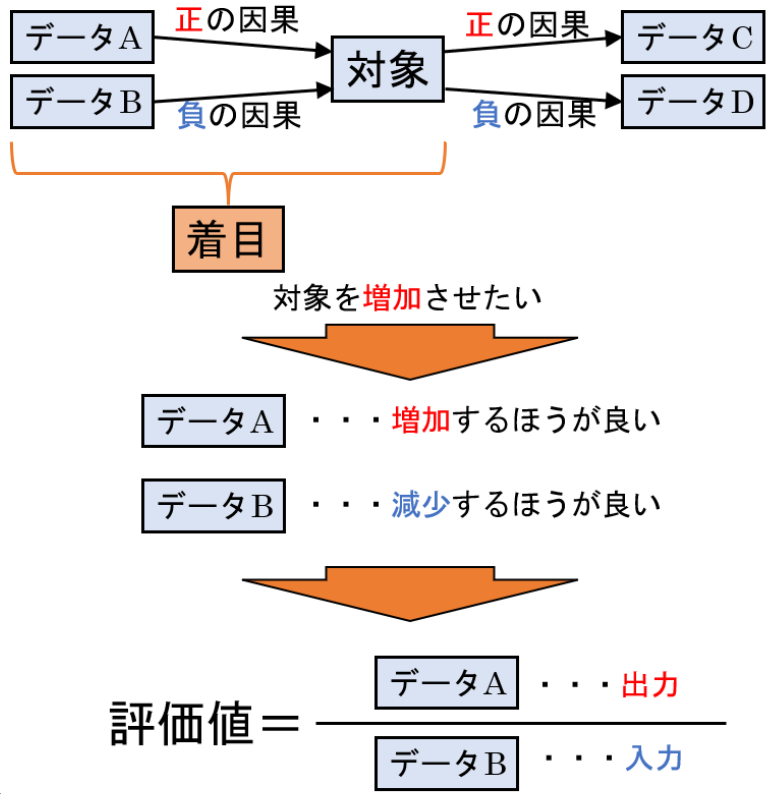
**データの正規化**

<robust Z-score>

$$x' = \frac{x - median(x)}{IQR}$$

$median(x)$ …データの四分位範囲　$IQR$ …データの中央値

**入力出力への振り分け**



# 選定されたデータに基づく DEA 分析





因果探索によって振り分けられた入力・出力を用いて DEA を行うことで評価値，入力・出力改善案，参照集合に属する市区町村とそれらにかかるウェイトを算出する．種類ごとの計 4 つの csv ファイルで結果を出力する．

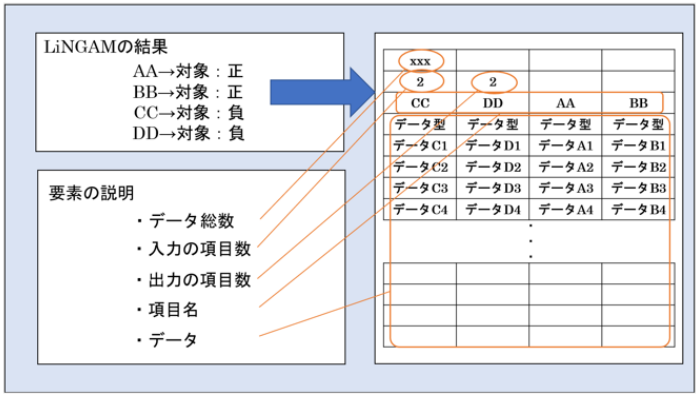



図21　DEAの入力・出力ファイル

表4　各都道府県におけるDMUの内訳

| 都道府県 | 市 | 区 | 町 | 村 | 都道府県 | 市 | 区 | 町 | 村 | 都道府県 | 市 | 区 | 町 | 村 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 北海道 | 35 | 0 | 130 | 20 | 石川県 | 11 | 0 | 8 | 0 | 岡山県 | 14 | 4 | 10 | 2 |
| 青森県 | 10 | 0 | 22 | 8 | 福井県 | 9 | 0 | 8 | 0 | 広島県 | 13 | 8 | 0 | 9 |
| 岩手県 | 14 | 0 | 15 | 4 | 山梨県 | 13 | 0 | 8 | 6 | 山口県 | 13 | 0 | 0 | 6 |
| 宮城県 | 13 | 5 | 18 | 1 | 長野県 | 19 | 0 | 23 | 35 | 徳島県 | 8 | 0 | 15 | 1 |
| 秋田県 | 13 | 0 | 9 | 3 | 岐阜県 | 21 | 0 | 19 | 2 | 香川県 | 8 | 0 | 9 | 0 |
| 山形県 | 13 | 0 | 19 | 3 | 静岡県 | 21 | 10 | 12 | 0 | 愛媛県 | 11 | 0 | 9 | 0 |
| 福島県 | 13 | 0 | 31 | 15 | 愛知県 | 37 | 16 | 14 | 2 | 高知県 | 11 | 0 | 17 | 6 |
| 茨城県 | 32 | 0 | 10 | 2 | 三重県 | 14 | 0 | 15 | 0 | 福岡県 | 27 | 14 | 29 | 2 |
| 栃木県 | 14 | 0 | 11 | 0 | 滋賀県 | 13 | 0 | 6 | 0 | 佐賀県 | 10 | 0 | 10 | 0 |
| 群馬県 | 12 | 0 | 15 | 8 | 京都府 | 14 | 11 | 10 | 1 | 長崎県 | 13 | 0 | 8 | 0 |
| 埼玉県 | 39 | 10 | 22 | 1 | 大阪府 | 31 | 31 | 9 | 1 | 熊本県 | 13 | 5 | 23 | 8 |
| 千葉県 | 36 | 6 | 16 | 1 | 兵庫県 | 28 | 9 | 12 | 0 | 大分県 | 14 | 0 | 3 | 1 |
| 東京都 | 26 | 23 | 5 | 8 | 奈良県 | 12 | 0 | 15 | 1 | 宮崎県 | 9 | 0 | 14 | 3 |
| 神奈川県 | 16 | 28 | 13 | 1 | 和歌山県 | 9 | 0 | 20 | 1 | 鹿児島県 | 19 | 0 | 20 | 4 |
| 新潟県 | 19 | 8 | 6 | 4 | 鳥取県 | 4 | 0 | 14 | 1 | 沖縄県 | 11 | 0 | 11 | 19 |
| 富山県 | 10 | 0 | 4 | 1 | 島根県 | 8 | 0 | 10 | 1 | 合計 | 733 | 188 | 727 | 203 |

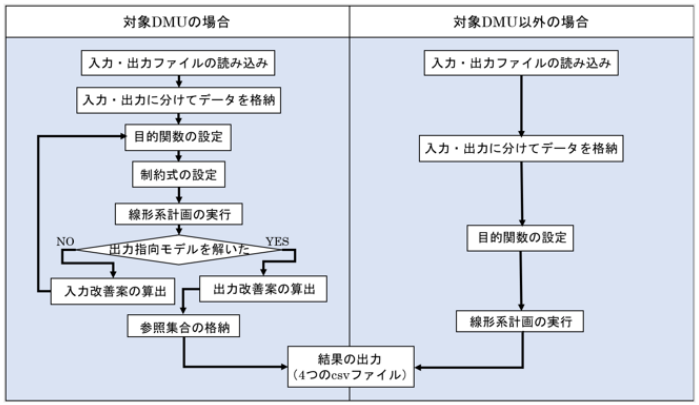



図22　DEA部分のフロー

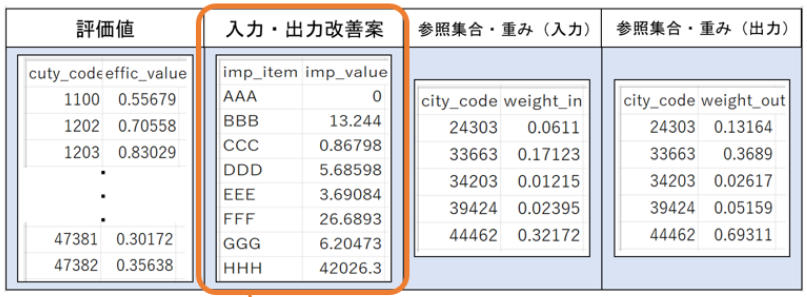



図23　DEA部分のアウトプット

改善案は正規化されていた値を逆変換して表示．

## 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





富山県射水市における少子高齢化問題について，その解決に向けた政策を行うというモデルケースに提案手法を適用することによって本研究の有効性を検証・考察する．

## 射水市における人口問題

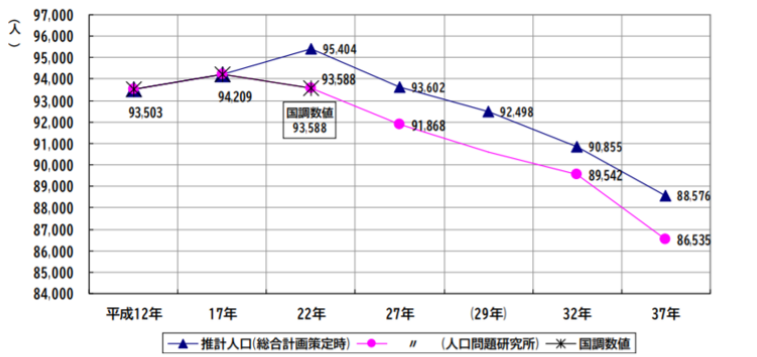



図24　射水市の推計人口

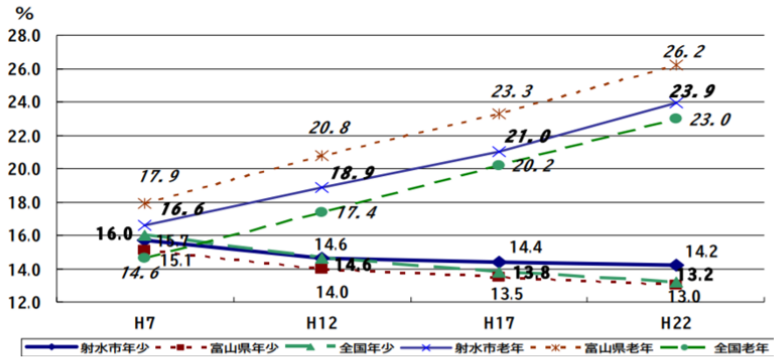



図25　年少・老年人口割合推移

## 数値実験のながれ

対象データを「年少人口[割合]」，
対象市区町村を「射水市」と入力．

↓

「年少人口[割合]」をターゲットとして因果探索．

↓

結果をcsvファイル①として排出．

↓

csvファイル①を入力・出力としてDEA分析

↓

全市区町村の評価値をcsvファイル②
射水市の入力・出力改善案をcsvファイル③
参照集合と射水市に対する重みをcsvファイル④，⑤
として排出．

↓

csvファイル②～⑤を用いてWeb-GISを作成．

↓

Web-GISをもとに考察．

# 数値実験の結果と考察





システムの各部分における結果を示す.

表5　「年少人口[割合]」に対するDirect-LiNGAMの結果

| データ項目 | パス係数 | データ項目 | パス係数 |
|---|---|---|---|
| 施設数 [空港] | 0.059 | 企業数 | −0.006 |
| 衛生費 | −0.019 | 商工費 | −0.024 |
| 警察・消防費 | −0.038 | 教育費 | 0.017 |
| 住宅用土地平均取引価格 | −0.043 | 生産年齢人口 | 0.249 |
| 老年人口 | −0.559 | | |

表6　DEAにおける入力・出力

| 入力 | 出力 |
|---|---|
| 住宅用土地平均取引価格 | 生産年齢人口 |
| 警察費・消防費 | 教育費 |
| 商工費 | 施設数 [空港] |
| 衛生費 | |
| 企業数 | |
| 老年人口 | |

表7　元データと改善案の比較

| | 元データ | 改善案 | |
|---|---|---|---|
| 企業数 | 3075 | -27769.854 | 入力：減少 |
| 衛生費 | 10.08 | -2.954 | |
| 商工費 | 2.92 | -1.647 | |
| 警察費・消防費 | 2.88 | -1.470 | |
| 住宅土地平均取引価格 | 19507 | -368560.385 | |
| 老年人口 | 28.6 | 22.821 | |
| 施設数 [空港] | 0 | 0.143 | 出力：増加 |
| 教育費 | 12.5 | 32.502 | |
| 生産年齢人口 | 57.5 | 66.692 | |

表8　参照集合に属する市区町村とウェイト

| 入力指向モデル | | 出力指向モデル | |
|---|---|---|---|
| 参照市区町村 | ウェイト | 参照市区町村 | ウェイト |
| 山形県東置賜郡川西町 | 0.268 | 山形県東置賜郡川西町 | 0.315 |
| 長野県北佐久郡御代田町 | 0.197 | 長野県北佐久郡御代田町 | 0.231 |
| 滋賀県米原市 | 0.108 | 滋賀県米原市 | 0.127 |
| 滋賀県蒲生郡日野町 | 0.158 | 滋賀県蒲生郡日野町 | 0.186 |

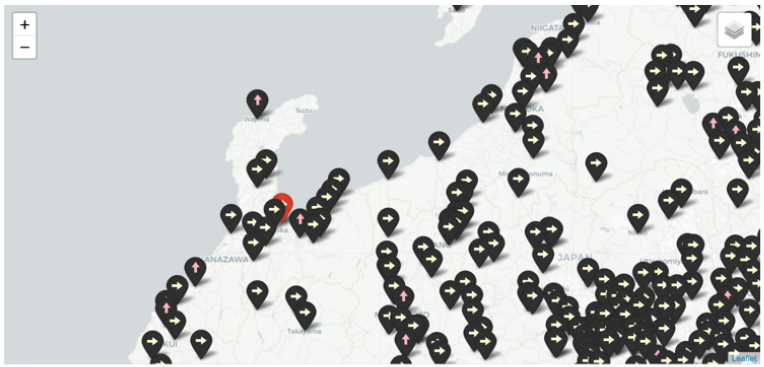


図26　EBPM-GISの結果

**考察**

・生産年齢人口が出力に振り分けられた.
　・・・年少人口の親世代にあたるため妥当.

・警察費・消防費が入力に振り分けられた.
　・・・意外性がある結果＝普通では気づかない関係性

# おわりに





## まとめ

- **データ間の因果関係に基づくデータ選択**
  - ・・・因果探索の結果をもとにデータの絞り込み，入力・出力を選定を行った.
- **DEAを用いた評価と改善案の算出**
  - ・・・振り分けられたデータを用いてDEAを行った.
- **GISによる結果の表示と重ね合わせによるデータフュージョン**
  - ・・・分析結果と地形データ，施設分布データを重ね合わせて分析が行われるようにした.

## 今後の課題

- **データベース内の情報量の充実**
  - ・・・扱える問題の汎用性と分析結果の確からしさ両方の向上につながる.
- **因果探索における手法の洗練**
  - ・・・因果グラフにおけるパス係数をどの程度まで有意とするか検討する.
    Direct-LiNGAMに限らず，より効果的な因果探索手法を模索する.
- **DEAにおけるモデルの深化**
  - ・・・本研究のテーマによりフィットしたモデルを模索・提案する.