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

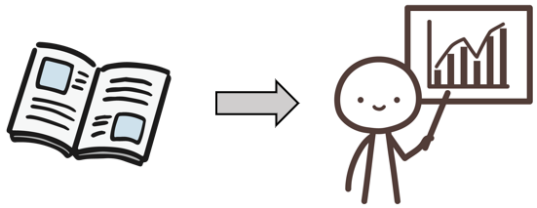
N212, 09:30-10:00 Tuesday, February 13, 2024.

1.1 本研究の背景

近年，世界各国で証拠に基づく政策立案（Evidence-Based Policy Making: EBPM）に注目が集まっている．日本においても例外ではなく，研究機関でも取り上げられている．EBPM を効果的に行うためには，社会の構造を適切に捉え，分析に活かす必要がある．

エビデンスベース

政策によって改善したい対象を明確化したうえでデータを収集し意思決定.



エピソードベース

住民によって役場に持ち込まれた問題に対して対面処理的に意思決定.

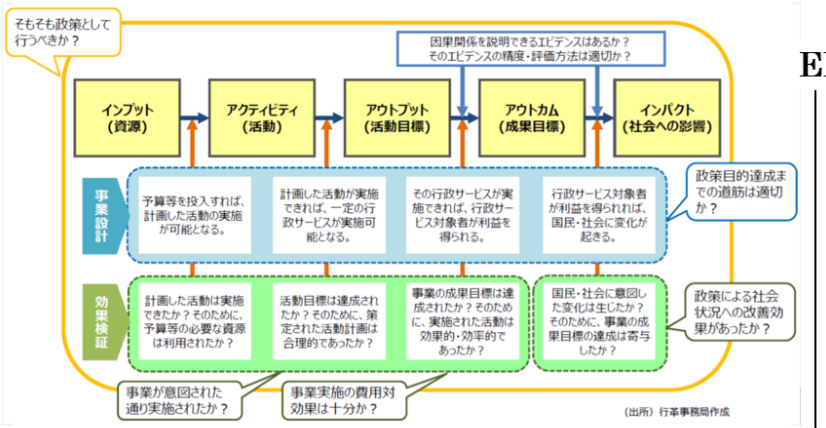


図1　EBPMのロジックモデル

EBPMの利点

- データに基づくため予測を行うことができ，問題が顕在化する前に対策を打つことが可能.
- その場限りのエピソードによるものではなく，明確な根拠があるため，住民の理解が得やすい.
- 政策による効果が事前に逆算できるため，状況に応じた微調整が可能.

1.2 本研究の目的

統計データを用いた機械学習によって，実世界に表出しない潜在的なクラスターを発見し，それに基づいて自治体をクラスタリングする．また，その結果を考慮しながらデータ間に成り立つ関係を数理モデルによって表す手法を提案する．加えて，これらの結果を可視化するシステムを開発する．

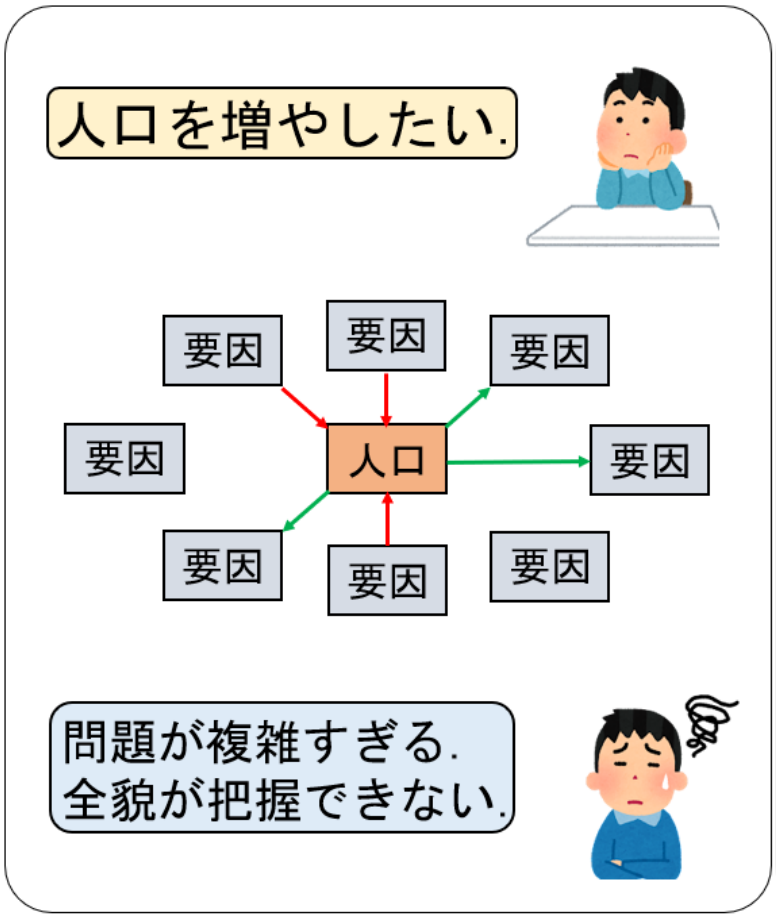


図2　ビッグデータ

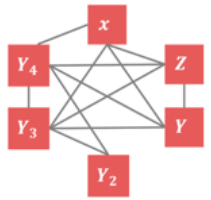


図3　因果グラフ

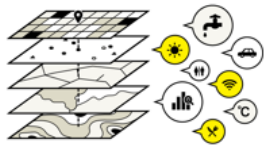


図5　GISのイメージ

地形情報・施設分布などと重ね合わせて新知見を発見.

図4　データ分析

政策決定支援

2.1 行政が持つ統計データとその公開

EBPM を実行するためには膨大なデータの収集・分析が必要になる．日本では，政府を中心に各自治体で様々なデータの収集が行われており，その一部はオープンデータサイトなどを通して民間にも共有されている．

EBPM推進の向けたシステム

- **内閣エビデンスシステム**
  研究機関の運営評価に特化．研究機関における「研究力」，「教育力」，「資金獲得力」などを見える化．
- **地域経済分析システム**
  経済に関するデータを中心に幅広いデータを扱う分析が可能．
- **V-RESAS**
  COVID-19による経済の変化を分析可能．
- **内閣府経済社会総合研究所**
  内閣府を支えるシンクタンク．政策研究を担う人材育成も行う．

図6　内閣エビデンスシステム

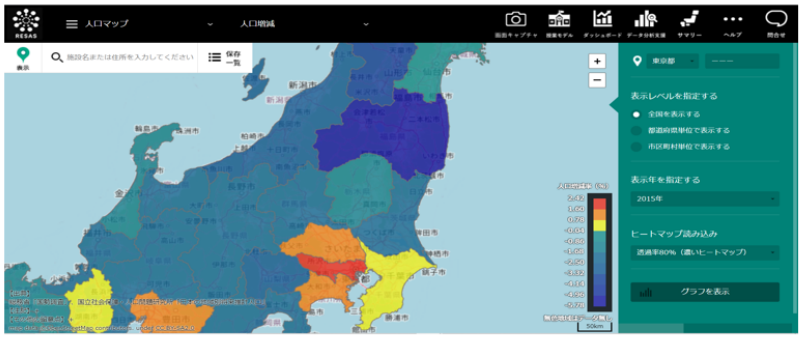


図7　地域経済分析システム

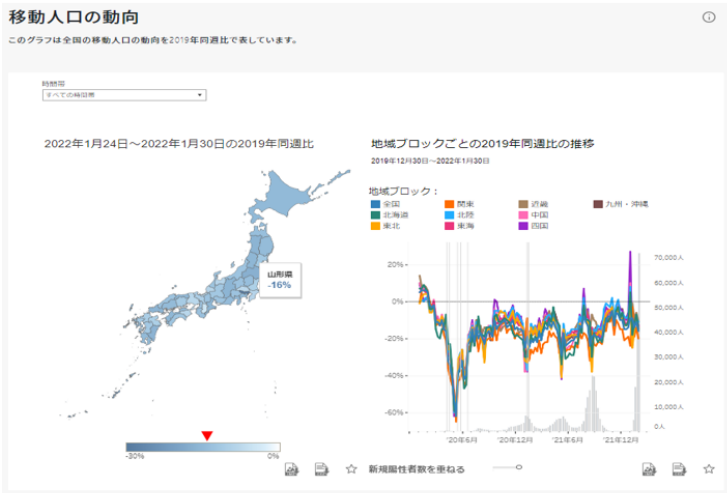


図8　V-RESAS

2.2 GIS を用いたデータ分析結果の可視化

GIS とは，地理空間データを総合的に管理・加工し，地理的位置とデータを結び付けて視覚化できる技術である．データ分析の結果を地図上にプロットし，地理的特徴と結び付けることで新たな知見が得られる可能性が示唆されている．

GISの特徴

- **データの可視化**
  統計データ等を2Dや3D，アニメーションなど多様な表現方法で地図上にプロット．
- **データ間の関係性把握**
  複数データを同一画面に表示し比較．
- **データの統合と分析**
  位置情報をキーとして異なる特徴を持つデータを統合．
- **データの作成と更新**
  データ更新の負担を軽減し鮮度を保つ．

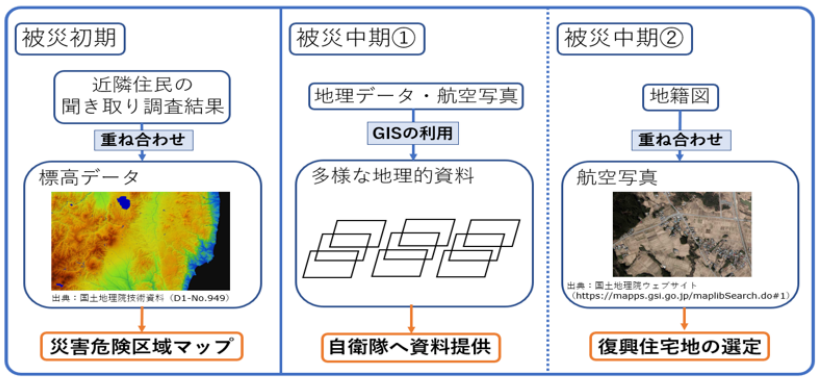


図9　被災下でのGIS

GISの利点

- **業務効率化による情報管理のコスト削減**
  電子化により作業時間や人員を縮小．
- **最適な意思決定の促進**
  同一画面上に複数データを表示．
- **コミュニケーション性の向上**
  視覚的な情報共有や議論が可能．

図10　地方自治体におけるGIS

2.3 クラスタリングと回帰分析

データの特徴や関係性を求める方法として様々な分野で機械学習の技術が用いられている．データを何らかのルールに基づいて分類するクラスタリングとデータ間の関係性をモデル化する回帰分析は機械学習の代表的な手法である．

3.1.1 潜在プロファイル分析 1

潜在プロファイル分析（Latent Profile Analysis: LPA）とは，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考慮しながらデータをクラスタリングする手法である．

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

Table 1: r 個の顕在変数に対する共分散行列 Σ_k のパラメータ

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

3.1.2 潜在プロファイル分析 2

モデル選択

顕在変数 $\mathbf{y}$ 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．ただし，潜在変数 z は観測できない．

$$\begin{aligned} p(\mathbf{y}, z|\pi, \mu, \sigma) &= p(z|\pi, \mu, \sigma)p(\mathbf{y}|z, \pi, \mu, \sigma) \\ &= \prod_i \prod_k [\pi_k f(y_i|\mu_k, \sigma_k)]^{z_{ik}} \end{aligned} \tag{2}$$

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

$$\begin{aligned} E_{z_{ik}}[z_{ik}] &= \sum_{z_{ik}=0,1} z_{ik} p(z_{ik}|y_i, \pi_k, \mu_k, \sigma_k) = 1 \times p(z_{ik} = 1|y_i, \pi_k, \mu_k, \sigma_k) \\ &= \frac{p(z_{ik} = 1)p(y_i|z_{ik} = 1)}{\sum_k p(z_{ik} = 1)p(y_i|z_{ik} = 1)} \\ &= \frac{\pi_k f(\mathbf{y}|\mu_k, \sigma_k)}{\sum_k \pi_k f(\mathbf{y}|\mu_k, \sigma_k)} \end{aligned} \tag{4}$$

3.1.3 潜在プロファイル分析 3

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 (4) に代入することによって $E_{z_{ik}}[z_{ik}]$ の値を求める.
- **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.

ベイズ情報量基準 (Bayesian information criterion: BIC)

$$BIC = -2L + k \ln n \tag{5}$$

3.1.4 潜在プロファイル分析 4

各サンプルにおける存在確率の算出

$$\pi_{k|y_i} = \frac{\pi_k f_k(\mathbf{y_i}|\mu_k, \mathbf{\Sigma}_k)}{\sum_k \pi_k f_k(\mathbf{y}_i|\mu_k, \mathbf{\Sigma}_k)} \tag{6}$$

3.2 パーセプトロンを用いた数法則発見法

RF6.4 法（Rule extraction method from Fact 6.4）

4.1 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

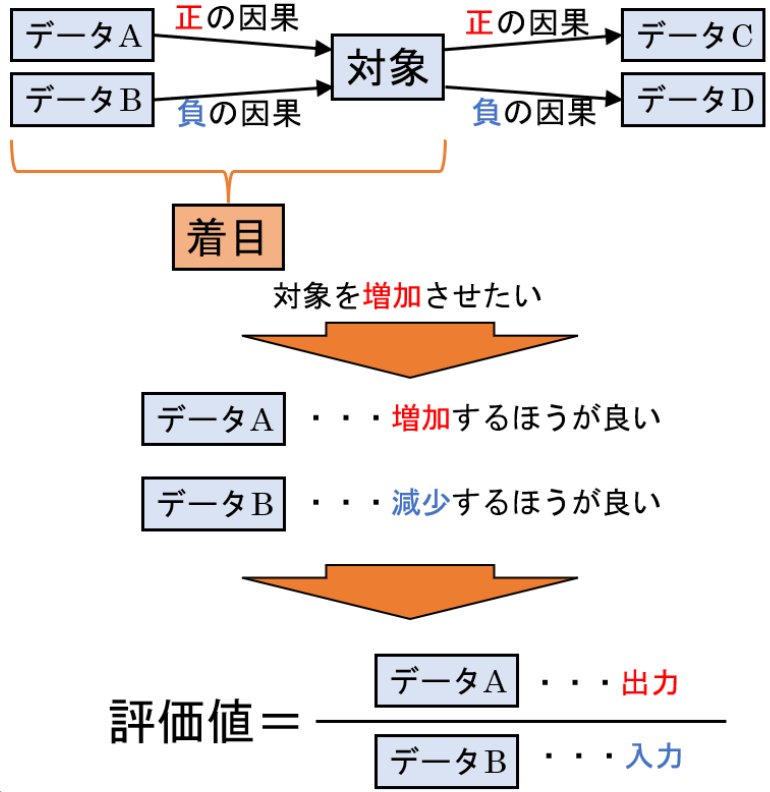
データの正規化

<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



4.2 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．

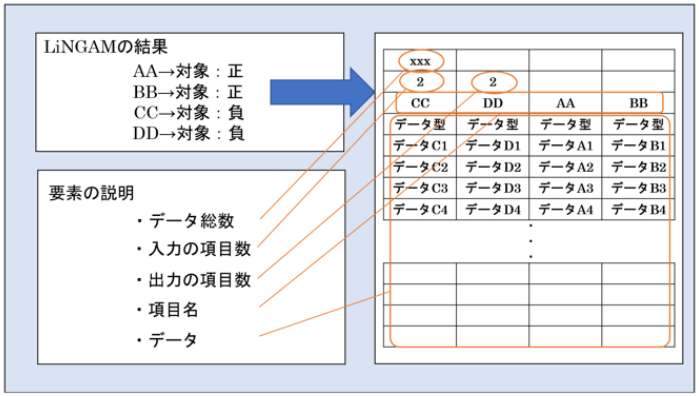


図21　DEAの入力・出力ファイル

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

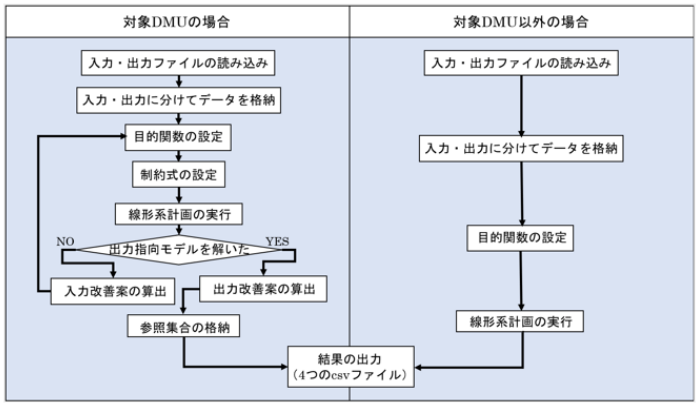


図22　DEA部分のフロー

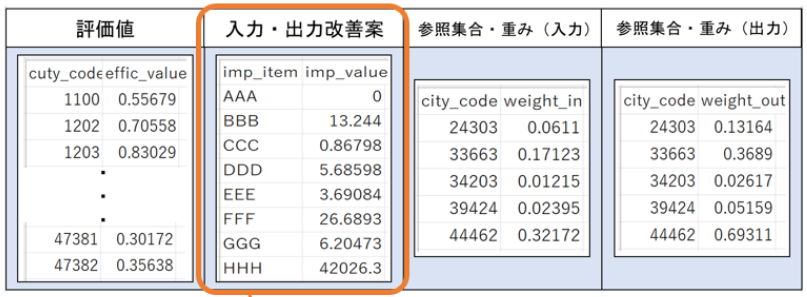


図23　DEA部分のアウトプット

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.

5.1 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

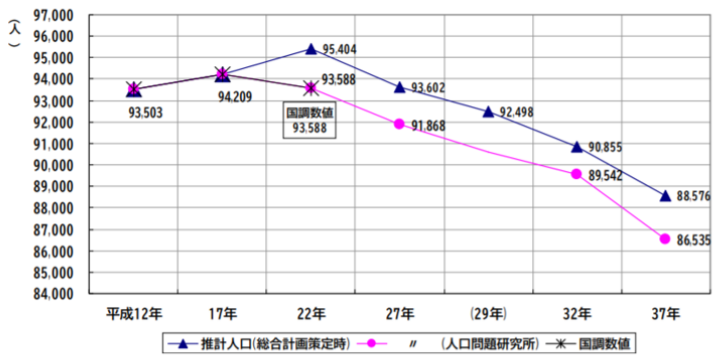


図24　射水市の推計人口

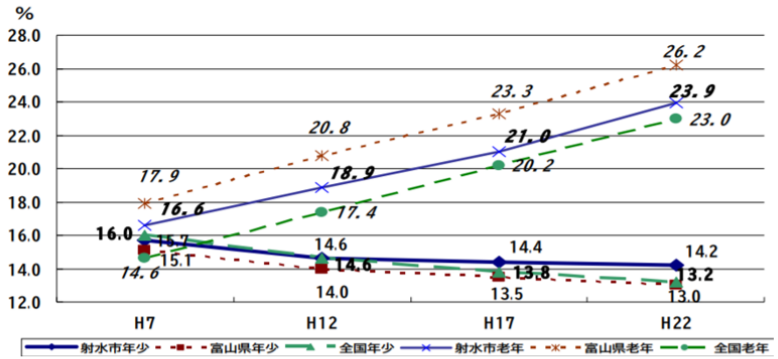


図25　年少・老年人口割合推移

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．

5.2 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	

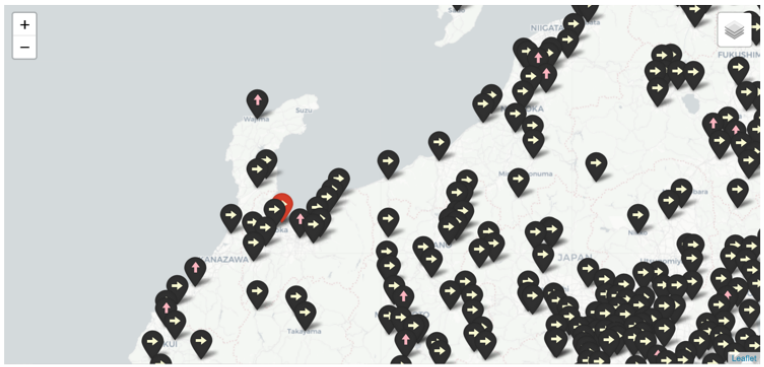


図26　EBPM-GISの結果

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

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

6 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.