修士論文

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

富山県立大学大学院 工学研究科 電子・情報工学専攻
2255013 長瀬永遠
指導教員 奥原 浩之 教授
提出年月: 令和6年（2024年）2月

目 次

図一覧

表一覧

記号一覧

以下に本論文において用いられる用語と記号の対応表を示す.

用語	記号
LiNGAM における i 番目の観測変数	x_i
LiNGAM における j 番目の観測変数から i 番目の観測変数へのパス係数	b_{ij}
LiNGAM における i 番目の観測変数に対する誤差（非観測変数）	e_i
主問題における各入力に対する重み	v^T
主問題における各出力に対する重み	u^T
主問題における対象 DMU の評価値	z
CCR モデルにおける DMU_o の入力	x_o
CCR モデルにおける DMU_o の出力	y_o
CCR モデルにおける DMU の入力	X
CCR モデルにおける DMU の出力	Y
双対問題における対象 DMU の評価値	w
入力指向モデルにおける対象 DMU の評価値	θ
入力指向モデルにおける各 DMU に対する重み	λ
出力指向モデルにおける対象 DMU の評価値	η
出力指向モデルにおける各 DMU に対する重み	μ
入力指向モデルにおける対象 DMU の i 番目の入力に対する改善案	$\hat{x_i}$
入力指向モデルにおける参照集合内の k 番目の DMU の i 番目の入力	x_{ik}
入力指向モデルにおける参照集合内の k 番目の DMU に対する重み	λ
出力指向モデルにおける対象 DMU の j 番目の出力に対する改善案	$\hat{y_j}$
出力指向モデルにおける参照集合内の k 番目の DMU の j 番目の出力	y_j
出力指向モデルにおける参照集合内の k 番目の DMU に対する重み	μ
提案手法における d 番目の市区町村の i 番目の入力	x_{id}
提案手法における d 番目の市区町村の i 番目の出力	y_{id}
提案手法における d 番目の市区町村に対する重み	λ_d
$robust\ Z-score$ における正規化後の値	ι
$robust\ Z-score$ を用いて正規化するデータ集合内の値	x
$robust\ Z-score$ を用いて正規化するデータ集合	X
$robust\ Z-score$ を用いて正規化するデータ集合の中央値	$median(x)$
$robust\ Z-score$ を用いて正規化するデータ集合の正規四分位範囲	$NIQR$
0〜1 変換の結果の値	ι'
0〜1 変換を行うデータ集合内の値の最大値	$max\|\iota\|$

第1章

はじめに

§ 1.1　本研究の背景

近年，世界各国の政府を中心に証拠に基づく政策立案（Evidence-Based Policy Making: EBPM）に対する取り組みの重要性が説かれている． EBPM とは，政策の立案をその場限りのエピソードに基づいて行うのではなく，政策によって改善したい対象を明確化したうえで，対象に関するデータを可能な限り収集し，合理的根拠に基づいて意志の決定を行うという考え方である [1]．EBPM を推進することは，政策の有効性を高め，国民の行政への信頼確保につながるとされる．

現在，日本政府における EBPM の取り組みとして，2017 年の官民データ活用推進戦略会議の決定のもと内閣府によって EBPM 推進委員会が発足され，内閣府の各部局によって EBPM の推進が図られている．また，EBPM を「科学的根拠に基づいた政策立案を推進する，アカデミズムと政治領域にまたがった運動」[2] と定義する論文もあることから，EBPM は単に行政のみが取り組むべき事柄ではなく，大学や民間の研究機関などと連携し，専門知識を活用しながら解決すべき課題であると考えられる．

特に効果的なデータ分析や適正な政策評価という観点では大学等の研究機関の寄与するところが大きく，現在の日本における EBPM に対する取り組みについての考察 [3] やエビデンスの質について言及し，システマティック・レビューを最も重要と位置づける書籍 [4] など EBPM に関する文献はさまざまな研究分野に属する研究者から出版されている．

以上のように，近年，日本において政府が積極的に推進し，研究機関においても多くの分野で多面的に考察がなされている EBPM であるが，現在でも全ての自治体，全てのケースにおいて EBPM に基づく意思決定を行うということは極めて困難である．そのため，現場における政策決定のいくらかは住民から行政機関に寄せられる問題に対して対面処理的な対応を行うエピソードベースの意思決定が用いられる．

このような事例の背景にある課題として，以下の二つの事柄における難しさがあると考えられる [5]．一つは政策における目的と手段の間に成り立つ関係を明確化することである．解決すべき目的とそれに対する手段である政策との論理的な関係性を示すことができない場合，政策の実施が課題解決にどうつながるのかを議論することが難しく，住民からの理解も得られにくい．

もう一つは，収集したデータを統計的手法に基づいて分析し，政策の実施と無関係の要因を取り除いた政策本来の効果を求めることである．政策立案の対象となるフィールドは様々な社会情勢の影響を受けているため，それらの影響を可能な限り排除した政策本来の効果を求めることは政策の有効性を議論するうえで非常に重要である．

また，これらの問題に付随して，政策立案の分野におけるデータ収集の難しさという問題が挙げられる．EBPM に用いられるデータはそのほとんどが市民の生活やそれに類する事象に関するものであるため，正しく収集し，蓄積するためには多くの費用や時間，労力が必要となる．

§ 1.2　本研究の目的

1.1 節で述べた課題に対して，現在，日本政府は政策立案における課題発見からその解決がなされるまでの各段階において行うべき事柄の詳細をその道筋に沿って記載したものであるロジックモデルなどを活用することによって目的と手段の間の関係性を整理する手法をとっている [6].

また，EBPM の研究分野では，それらによって整理された各段階においてそれぞれにあったデータの分析方法や政策の効果を効率的に算出する手法が研究されている．しかし，そのいずれにおいても，その根底にあるのはデータであり，データを収集するハードルの高い政策立案の分野では必要なデータが常に手に入ることは稀である．

そこで，本研究では，1.1 節で言及した EBPM がより多くの行政機関で広く普及するために考えるべき課題のうち，特に政策立案の分野におけるデータ収集の難しさに着目する．前述のとおり，政策立案の分野におけるデータ収集の難しさには対象となる事柄の規模が大きく，費用や時間等のコストが高いことが一番に挙げられる．また，これらは物理的に解決が困難な課題である．

このことから，本研究ではデータの収集には限度があり，その不足は仕方がないという前提を置いたうえで，それまでに収集されているデータを用いて統計的分析を行うことで収集されていないデータに対してなるべく確からしい値を予測し，データの補完を行う手法を提案する．

はじめに，任意の説明変数といくつかの説明変数との間に成り立つ関係を回帰分析によって多項式の形で求める多変量多項式回帰の一つである RF 法（Rule extraction method from Fact）[7] を用いて，既に収集されているデータから最新のデータを予測することを考える．

また，政策立案の分野で用いられるデータの特徴から，より精度高くデータの予測を行うために複数のデータ項目に基づいてサンプルをいくつかの潜在的なクラスに分類する手法である潜在プロファイル分析（Latent Profile Analysis: LPA）[8] の考え方を組み合わせた手法を提案する．

加えて，オープンデータを用いてそれぞれのデータ間の因果関係を分析し，その結果をもとに市区町村における運営の評価等を行う手法 [9] で提案されているデータ分析システムに対して，本研究の提案手法の結果を用いた機能の追加を行う．

最後に，以上のデータ予測手法について結果の予測精度に関する検証を行いその結果を示す．また，様々な観点から結果に対する考察を行い，今後考慮すべき事柄について言及する．

§ 1.3　本論文の概要

本論文は次のように構成される．

第 1 章 本研究の背景と目的について説明する．背景では，EBPM の重要性と日本国内での広がり，適切に導入する際に障壁となる課題について述べる．目的では，背景で述べた課題の中で本研究の対象としたいものを取り上げ，その解決に向けた新しいデータ分析手法を提案することについて述べる．

第 2 章 EBPM の概要とよく用いられる手法について解説し，それらを効率的に行う上での ICT の重要性に言及する．また，EBPM に向けたデータ分析の例を挙げ，その課題を解説する．加えて，それらと関係が深い回帰分析について，一般的な内容を示す．

第 3 章 本研究の提案手法を作成するにあたって参考としたデータ分析手法の先行研究を用いて，一般的な理論について解説する．

第 4 章 本研究の目的で挙げた課題を解決するための手法を提案する．また，提案手法によって求められる結果を用いて，先行研究で行われた分析を改善する手法について述べる．

第 5 章 人工野でもデータおよび実際のオープンデータを用いて実験を行うことで，本研究における提案手法の有意性を検証し，結果を考察する．

第 6 章 本研究に関する内容を簡潔にまとめ，本研究において実現できたことと今後の展望を示す．

第2章

EBPMとデータサイエンスの有用性

§ 2.1　EBPMとICTを用いた取り組み

経済社会構造が急速に変化するわが国において，限られた資源を有効活用しながら国民に信頼される行政施策を展開するために，政策の対象に関するデータを収集・分析し，それに基づいて政策における意思決定を行うという考え方であるEBPMに基づいて政策立案を行うことが重要視されている．

しかし，全ての政策において効果的なEBPMを適用するためには，膨大かつ多種多様なデータを収集・保存・管理し，それらのデータを適切かつ高速に高い信頼度を保って選択・統合・分析する必要があり，担当者に対する大きな負担となるため人手のみでそれらを行うことは困難となる．

そのため，特に地方自治体においてEBPMを政策の広範囲に適用することは人員の観点から見ても難しい課題であると考えられる．これらのことから，EBPMにおいて適切なエビデンスの収集・分析をおこなうには，ICTを用いることが欠かせない．また，そういった場合，ICTに対する専門知識が十分でないと考えられる一般的な職員でも不安なく業務にこれらの技術を活用できるように感覚的に理解しやすいシステムを提供するとともに，庁内全体で講習会を開催するなどしてICTに関する知識を醸成することが必要である．

本節の以降では，本研究において重要な意味を持つEBPMの必要性を示すために，その概要と日本における動向，用いられる手法の例を解説する．また，内閣府と経済産業省が提供するEBPMのためのWebアプリケーションを取り上げ，EBPMにおけるICTや情報工学の重要性を明確にする．

EBPMの概要と日本における動向

EBPMとは前述のとおり，エビデンスに基づく政策立案であるが，元となった考え方の一つにエビデンスに基づく医療（Evidence-Based Medicine: EBM）というものがある [13]．これは，医療従事者が患者への医療行為に関する意思決定を行う際にその時点の医学において得ることの出来る最善の科学的根拠に基づいてそれらを行うというものである．

具体的なものとして，以下のような事例が挙げられる．従来の医療現場では心筋梗塞後に不整脈が多いと予後が悪いと考えられていたため，不整脈発生時には抗不整脈の薬を使用することが一般的であった．しかし，1989年に心筋梗塞の患者に対する抗不整脈の薬の影響を明らかにする実験が行われた結果，不整脈の薬を使用した場合，患者の死亡率が3〜5%ほど増加することが分かった．

このようにそれまでの経験から導かれ，半ば迷信のように信じられてきた方法ではなく，実際にデータを取り，それらを正しく分析することによって得られた結果をもとに新たに意思決定を行うという考え方を政策立案の分野に応用したものが EBPM である．これらの考えは英国では 1997 年からのブレア政権，米国では 2009 年からのオバマ政権で本格的な導入がされ始めた．

日本では 2010 年代からその必要性が議論されてきた．2017 年 2 月には，政府に「統計改革推進会議」が設置され，同年 5 月に「統計改革推進会議最終取りまとめ」が決定された．これが日本における本格的な EBPM の出発点といえる．同年 7 月に「官民データ活用推進戦略会議」の下に「EBPM 推進委員会」が設置され，この場で政府全体として EBPM を推進することとなった．

2018 年度からは各府省に組織内における EBPM 推進のモニタリング，指導などを行う「政策立案総括審議官」が配置され,「EBPM 推進委員会」はその取り組みを主導することとなった．また，2021 年 9 月にデジタル庁が設置されたことに伴い，同委員会はその下へ移行され，その活動は現在まで続いている．

EBPM の手法

ロジックモデル

ロジックモデルとは政策が立案され，それが遂行されることによって目的となる課題が解決されるまでの道筋を論理的に表したものであり，EBPM を構築するうえで重要なものである．ロジックモデルはインプット（資源），アクティビティ（活動），アウトプット（活動目標），アウトカム（成果目標），インパクト（社会への影響）の流れに沿って作成され，それぞれの段階において考えるべき事項やクリアすべき課題が明記されている．

実際に作成されたロジックモデルの例を図 2.1 に示す．このロジックモデルは法務省の「受刑者就労支援体制等の充実」事業において，受刑者が出所後に社会で安定した生活が送れず，再犯してしまうことを防ぐために在所中における就労支援体制を強化するという課題のために作成されたものである [14].

このようなロジックモデルの作成は EBPM の基本であり，それぞれの実情に応じたものを作成することについてその意義は大きい．一方，事業の性質によっては作成に向かないものも存在するため日本の各府省では政策立案総括審議官等が中心となってその意義についての精査も含めた作成にあたっている．

ランダム化比較試験

前述のロジックモデルにおけるアウトプットからアウトカムへの因果関係を分析する手法は複数存在し，それらは内閣府が定めた信頼度の目安によっていくつかのレベルに分けられる [15]．各レベルに属する分析手法を表 2.1 に示す．また，これらのレベル分けはエビデンスレベルと呼称される．

ここでは，表 2.1 に示される手法の中で最も信頼置ける手法とされるランダム化比較試験（Random Controlled Trial: RCT）について簡単に解説する．RCT とは対象者をランダムのグループに分け，政策を適用するグループ（介入群）と適用しないグループ（比較対照群）との比較によって政策効果を分析する手法である．RCT を行う際

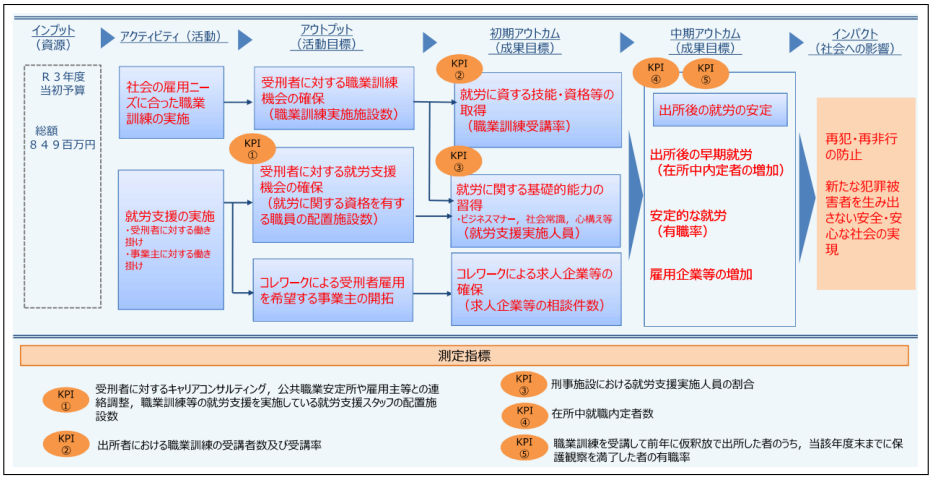


図 2.1: ロジックモデルの例 [14]

表 2.1: エビデンスレベル

↑	レベル1	ランダム化比較実験
↑	レベル2a	差の差分析、傾向スコアマッチング、操作変数法等
質が	レベル2b	重回帰分析、コーホート分析
高い	レベル3	比較検証、記述的な研究調査
	レベル4	専門家等の意見の参照

は政策の効果以外の条件が結果に影響する可能性を排除するため，グループ分けをランダムにするほか，対象者自身もどちらのグループに属しているか分からないようにするなどの条件設定が必要である．

以上のような条件を整えれば，非常に信頼度の高い評価が行える RCT であるが，実験を行うための費用，労力，時間などのコストが大きいほか，場合によっては個人の同意を得ずに実験を行わないと必要な条件がそろわないなど倫理的な課題もあり，実施が難しい場合も多い．そこで，既存のデータを実験を行った結果のように活用する手法である「自然実験」と呼ばれる手法がとられることもある．

EBPM の推進に向けた ICT の活用

EBPM の推進に向けた政府の取り組みの一つに，経済産業省と内閣官房デジタル田園都市国家構想実現会議事務局が協働で提供している Web アプリケーションである地域経済分析システム（Regional Economy Society Analyzing System: RESAS）がある [16]．このシステムは特に地方創生に関して効果的な施策の立案・実行・検証のために有用なデータの提供と可視化を目的として作成されており，経済産業省および内閣府が持つ経済に関する統計データを市町村，各年単位で表示できる．

また，目的の市区町村を指定することで単にデータを数値として表示するだけでなく，データのグラフ化，地図を用いた可視化，ほかに同様の傾向を持つ市区町村を自動的に検

図 2.2: RESAS の例（射水市）[16]

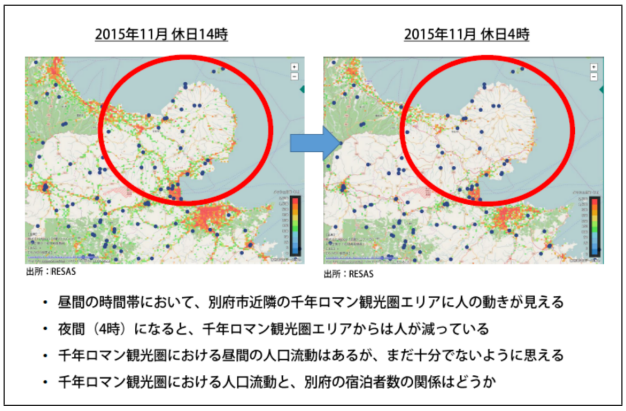


図 2.3: 別府市における RESAS 活用事例 [17]

索するなどができる．加えて，指定した市町村や比較となる市町村を選択し，それらに関するデータをファイルとしてダウンロードすることも可能である．RESAS を用いて富山県射水市の人口に関するデータを表示した結果の一部を 2.2 に示す．

以上のような機能を備える RESAS であるが，このシステムを利用してデータの分析を行い，新たな政策の立案に対する知見を得た事例として，大分県別府市で行われた RESAS を活用した政策立案ワークショップが挙げられる [17]．この事例では，当市における観光業振興における新たな政策の立案を目的として自治体職員，有識者のそれぞれが RESAS を用いた分析を行いその結果をもってディスカッションを行った．

RESAS を用いた分析では，別府市の産業におけるサービス業の比率が全国的に見ても圧倒的に高いことが示唆された．また，そのことに着目したうえで市の観光圏における休日昼夜それぞれの時間帯の人口流動を RESAS の地図機能を利用して描画し，それらの特徴をもとに議論が展開された．結果の一部を図 2.3 に示す．データを地図上にプロットし，見える化を行ったことで，人流などの特徴を感覚的に捉えることができ，新たな特徴の発見に繋がっている．

このような事例からも政策立案の際に ICT を活用し，グラフやマップなどを用いて蓄積されたデータを可視化することは新たな知見を得るために有効であると考えられる．一方，RESAS における可視化の主軸はあくまで統計データに関するものであり，EBPM の推進のためにはそれらのデータを用いた分析の結果を同様に可視化することがより効果的であると考えられる．よって，本研究では次節で示すデータ分析システムを踏襲し，統計的分析の結果を地図を用いて視覚的に提示することを考える．

§ 2.2　EBPM の推進に向けたデータ分析・可視化システム

EBPM を効果的に行うにあたり，データを地図上にプロットするなどして可視化することが新たな知見の発見に有効なことは，2.1 節で述べたとおりである．そこで，本節では，EBPM の普及に向けたデータ分析の手法を提案し，その手法の結果を地図上に可視化する

ことで新たな知見の発見を支援した研究についてその目的と手法を詳述する [9]．また，その研究にて提案されている手法について，考えられる課題について言及する．

先行研究の背景と目的

現在，日本では，地方自治体における EBPM 推進に向けて，政府を中心に様々な取り組みがなされている．しかし，地方自治体の政策における意思決定には，しばしば過去の経験則に基づいたエピソードベースの決定がなされる．その原因には，様々な要素が考えられるが，先行研究では特に以下の 2 点における難しさに着目している．

1. 政策の対象となる問題を引き起こす原因を正しく特定すること
2. データ分析に特化した人員の確保

一つ目の項目について，政策の対象となる問題は何か一つの原因によって引き起こされるのではなく，複数の要因がお互いに影響し合った結果として起こるものである．そのため，人力によって要因間のつながりを正しく把握し，原因を特定することは非常に困難である．

二つ目の項目について，一般に，地方自治体における職員はデータ分析の専門家ではない．また，そのような人員を確保している自治体においてもその人数は十分とは言えない．そのため，観測されたデータを統計的手法で分析し，その結果を正しく解釈することはハードルが高い．

以上の課題に対して，先行研究では，統計的分析手法を用いて観測されているデータ間に成り立つ因果関係を分析し，その結果に基づいて各自治体の評価を行う手法を提案している．また，分析によって得られた評価値を向上させるために，改善すべき項目とその目標値を算出している．

加えて，それらの結果をプロットした地理情報システム（Geographic Information System: GIS）[18] を生成することで一般の自治体職員が感覚的に分析結果を理解できるような Web アプリケーションを作成している．

先行研究の分析手法

先行研究では，政策の対象における原因と結果の関係を分析するために，データ間の因果関係を独立主成分分析によって同定する手法である線形非ガウス非巡回モデル（Linear non-Gaussian acyclic model: LiNGAM）[19] を用いてい．LiNGAM のモデルにおける定式化の結果を以下に示す．

$$x_i = \sum_{i \neq j} c_{ij} x_j + e_i \quad i, j = 1, \cdot, N \tag{2.1}$$

ここで，x_i, x_j は観測されているデータ（内生変数），c_{ij} は x_i と x_j の間に成り立つ因果関係の向きと大きさを表す値，e_i は誤差項（外生変数）であり，N は観測されているデータの項目数である．LiNGAM では，式 2.1 を以下の仮定に基づいて同定する [19]．

1. 外生変数と内生変数をつなぐ関数は線形関数とする（内生変数とは実際に観測されている変数，外生変数とは内生変数以外の変数で内生変数のそれぞれに関する未知の値である）

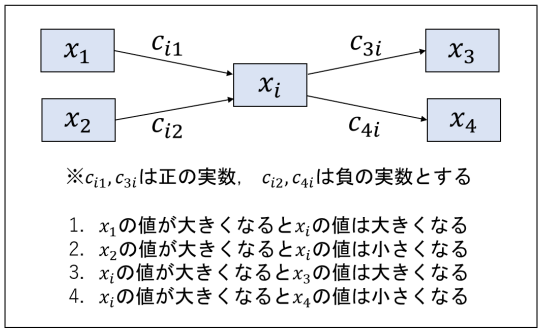


図 2.4: 4 パターンの因果関係

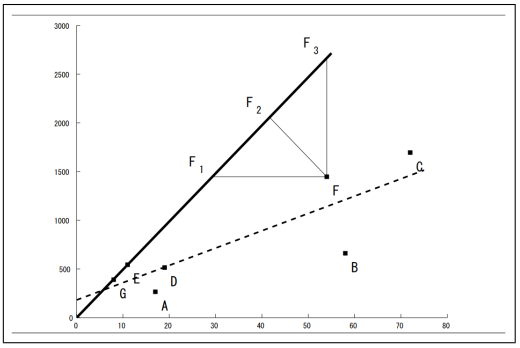


図 2.5: DEA における結果の例 [21]

2. 外生変数の分布は非ガウス連続分布とする
3. 因果グラフは非巡回とする
4. 外生変数は互いに独立とする

LiNGAM において，式 2.1 を求める手法はいくつか提案されているが，先行研究では回帰分析とデータ項目間の独立性評価を用いるアプローチである Direct-LiNGAM [22] を使用している．LiNGAM の結果，データ項目間の因果関係は式 2.1 の c_{ij} によって表され，一つの内生変数に着目すると関係する因果関係は 4 パターンに分けられる．ある内生変数 x_i に関係する 4 パターンの因果関係の例を図 2.4 に示す．

次に，先行研究では LiNGAM によって求められたデータ項目間の因果関係を用いてデータを整理し，整理されたデータを用いて対象の自治体に対する運営評価値を算出している．評価値の算出については，データ包絡分析法（Data Envelopment Analysis: DEA）[23] を用いている．

DEA とは，ある分野における組織の集合において，対象の組織の業績に対する評価値を算出するために生み出された手法である．ここでいう組織とは，その活動においていくつかの入力（投入）をいくつかの出力（産出）に変換することに携わる生産体（Decision Making Unit: DMU）のことを指す．

DMU における活動の例として，いくつかの材料を用いて製品を生産する工場における毎期の生産活動や，小売店における従業員や広告費にコストを支払って商品を販売することで利益を得る活動等が挙げられる．DEA ではこれらの活動に対して，対象としたい DMU を基準とし，集合内の他の DMU との比較によって，その業績を評価することが可能である．

DEA では，対象としたい DUM に対して，その DUM を含めた同様の活動を行う DUM 集合についてのデータを使用し，集合内の他の DMU における評価値を制約条件に加えながら，業績を評価する．評価する対象やデータの種類によって，いくつかのモデルが存在するが，先行研究では最も基本的なモデルである CCR モデル [24] を使用している．

CCR モデルにおける DMU の評価値は入力に対する出力の大きさによって求められる．また，対象の DMU の評価値を競合する他の DMU との相対評価によって求めるために，競合する DMU の評価値を制約として加える．具体的な方法としては，すべての DMU の入力および出力に対して，以下の二つの制約条件を満たす共通の重み変数を導入する．

- いずれの DMU においても評価値の最大は 1 である

- すべての入力・出力に対する重みは 0 以上である

以上の条件を満たす入力・出力に対する重みをそれぞれ $\mathbf{u_{in}}, \mathbf{u_{out}}$ とし，対象の DMU における入力・出力を $\mathbf{x_o}, \mathbf{y_o}$，競合する DMU 集合における入力・出力を $\mathbf{X}, \mathbf{Y}$ とすると，CCR モデルにおける評価値 z_o は以下のように定式化できる [20].

$$\text{maximize} \qquad z_o = \frac{\mathbf{u_{out}}^T \mathbf{y_o}}{\mathbf{v_{in}}^T \mathbf{x_o}} \tag{2.2}$$

$$\text{subject to} \qquad -\mathbf{v_{in}}^T \mathbf{X} + \mathbf{u_{out}}^T \mathbf{Y} \leq 0 \tag{2.3}$$

$$\mathbf{u_{out}} \geq 0 \tag{2.4}$$

$$\mathbf{v_{in}} \geq 0 \tag{2.5}$$

また，式 2.5 は，線形計画問題であるため，入力もしくは出力に対する重みを目的関数とした双対問題に書き換えることができる．さらに，目的関数における分子または分母のいずれかを 1 とおいて双対問題を解くことによって，入力・出力のいずれかに対してもっともよい評価をもつ DMU の集合である参照集合を求めることができる．

1 入力・1 出力の DEA における DEA の結果のイメージを図 2.5 に示す．ただし，横軸が入力，縦軸が出力を表す．図 2.5 の場合，参照集合は実線上にあるサンプル G，E となる．

参照集合に含まれる DMU のデータとそれに対する重みを用いれば，対象の DMU の評価値を制約の範囲内で最大化した値を求めることができる．参照集合における DMU の数を R_{DMU}，入力・出力の数をそれぞれ N_{in}, N_{out}，それらに対する重みを λ_{r_d}, μ_{r_d} とすると，対象の DMU における，各入力・出力の現実的な最大値 $\hat{x}_{n_i}, \hat{y}_{n_o}$ は以下のようになる．

$$\hat{x}_{n_i} = \sum_{DMU_{r_d}}^{R} x_{n_i r_d} \lambda_{r_d} \quad n_i = 1, \cdots, N_{in} \tag{2.6}$$

$$\hat{y}_{n_o} = \sum_{DMU_{r_d}}^{R} y_{n_o r_d} \mu_{r_d} \quad n_o = 1, \cdots, N_{out} \tag{2.7}$$

先行研究では，LiNGAM と DEA を合わせて用いることで自治体における運営評価値および評価値を最大化するための具体的な目標値の算出を行っている．以下に，分析のアルゴリズムを示す [9].

1. 評価値を算出する際，特に注目したいデータ項目をデータセットから 1 つ選択する
2. データセットにおけるすべてのデータを用いて LiNGAM による因果探索を行い，注目したいデータ（ターゲット）に対して影響を与える方向に因果関係を持つデータを抜き出す
3. 因果探索によって得られた因果の正負に着目しながら，データを入力・出力に振り分け DEA を用いた評価値および目標値の算出を行う

手順 3 において，データを入力・出力に振り分ける方法を図 2.6 に示す．DEA の説明でも述べたとおり，DEA の基本的な考え方は，いかに少ない入力から多くの出力を生み出せ

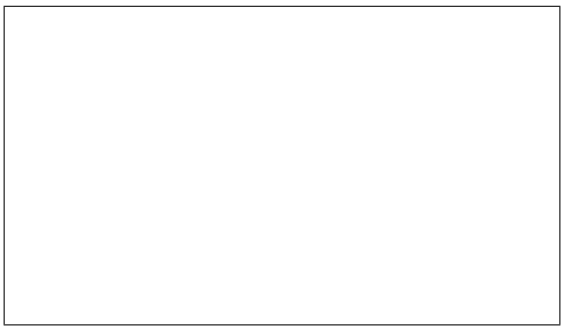


図 2.6: データの振り分け方法

図 2.7: アプリケーションの概要 [9]

るかである．このことから，式 2.5 を最大化するためには，分母をより小さく，分母をより大きくするべきであるといえる．

また，ターゲットに対して影響を与える方向の因果関係を持つデータに着目した場合，ターゲットを最大化するためには負の因果性を持つデータを小さく，正の因果性を持つデータを大きくすることが望ましいといえる．

以上の関係から，先行研究における手法では，ターゲットに対して負の影響を与える因果性を持つデータを DEA の入力に，正の影響を与える因果性を持つデータを DEA の出力に配置することで DEA における分析を行っている．

先行研究のシステム

先行研究では，前述の手法を用いて政策立案分野のデータを分析し，GIS を用いた Web アプリケーションによって視覚的に分析結果を提示するシステムを提案している．分析に用いるデータはオープンデータサイトである RESAS から API を用いて収集したものを用いる．

先行研究のアプリケーションにおける画面遷移の様子を図 2.7 に示す．はじめに，ユーザは画面に出力されるドロップダウンリストから，対象としたい自治体名を選択する．また，ラジオボタンからデータ項目名を選択する．このとき，データ項目を選択すると画面の右側に分析に用いるデータが表示される．自治体名の隣に表示されている値は政府から市町村に割り当てられたコードである．

次に，実行ボタンを押すと，選択した自治体と項目をターゲットとしてデータが分析され，データベース内に存在するすべての自治体に対する評価値およびデータの目標値が算出される．結果は各自治体の役所が存在する地点にピンとして表示され，ピンには評価値の大きさに応じて 3 段階の矢印が表示される．また，ピンを 1 回クリックすると，その自治体に対する詳細な分析結果を表示される．

先行研究における課題

先行研究では，これまで述べたように行政機関が実際に調査を行った結果であるオープンデータを用いて分析を行っている．しかし，対象としている政策決定の分野はデータの更新が頻繁ではない．そのため，分析を行うタイミングによっては数年前のものを最新のデータとして扱って分析を行う必要があり，これは好ましくない．

よって，本研究の手法では，既存のデータを用いて最新のデータを予測することで分析全体の精度の向上を目指す．また，先行研究の手法ではシステムを用いて実行できる分析

が 1 種類に限られており汎用性が乏しいため，本研究でデータ予測を行う上で用いるデータ分析の結果をシステムで行える分析の種類に加えることで，システムの機能を拡張する．

§ 2.3　回帰分析と社会事象への適用

近年，コンピュータに関する技術の発達とそれに伴う普及により，社会の様々な分野において大量に情報があふれるようになった．それに伴い，大量のデータの中から新たに有益な情報を生み出すデータマイニングの技術が研究されている．その中でも，観測されているいくつかの変数に基づいて別の変数の実数値を予測する回帰分析は，歴史が古く，最もポピュラーな手法の一つといえる．

しかし，単に回帰分析といっても現在までに考えられているモデルは様々あり，手段や目的，データの種類などによって使い分けられる．代表的な回帰モデルを出力の形，扱うことのできる説明変数の種類で分類したものを表 2.2 に示す．説明変数の種類には量的変数と質的変数がある．量的変数とは，その値の増減自体に意味がある変数で，身長や気温などがこれにあたる．一方，質的変数とは，数値で表現することもできるが，その値自体に意味はない変数で，性別や選択式のアンケート結果などが代表的である．

回帰分析のうち，最も単純なものの一つである重回帰は，目的変数が説明変数に対する線形関数で表現されるという仮定の下，その線形関数を事前に与えられた学習データに基づいて予測する手法である．また，重回帰分析では，説明変数に量的変数のみを用いるのに対して，質的変数も含めて分析を行う手法が数量化理論一類である．これらの手法はまとめて線形回帰モデルと呼ばれる [25].

線形回帰モデルに対して，非線形回帰モデルと呼ばれる手法は，説明変数と目的変数の関係をより柔軟に表現することができる．非線形回帰の代表例としては，多変量多項式回帰やニューラルネット回帰 [26]，サポートベクトル回帰 [27] などが挙げられる．

また，線形回帰モデルや非線形回帰モデル以外にも，出力を数式という形ではなく木構造で得る回帰木モデルや現在のデータを過去のデータを用いて回帰することで時系列データの予測を行うことができる自己回帰モデルなどが存在する．

このように，これまで様々な種類のモデルが研究されている回帰分析であるが，これらを行う上で重要となる要素に，計算量，回帰式の可読性，回帰式の汎化性が挙げられる．回帰式の可読性とは，回帰分析によって得られた回帰式の解釈における難しさを意味し，汎化性とは，未知のデータに対していかに精度の良い予測値を推定できるか意味する．

これらの要素のうち，計算量は少なく，可読性と汎化性は高いほうが良いのだが，回帰分析の研究においては特に可読性と汎化性が重視される．これは，回帰分析が用いられる多くの場合においてリアルタイム性がもとめられることは稀であるため，多少時間がかかったとしても可読性に富み，精度が高い結果を得ることがより重要だからである．

このことを踏まえて表 2.2 におけるそれぞれの手法を考えた場合，以下のような特徴が挙げられる．線形回帰モデルである重回帰分析や数量化理論一類は，回帰式が線形であるため，その可読性は非常に高いが，非線形の関係を持つデータに対して高い汎化性は期待できない．一方，サポートベクトル回帰やニューラルネット回帰などは，非線形であるため多くのデータに対して高い汎化性が期待できるが，入出力関係がブラックボックスなため，回帰式の可読性が悪く，得られた回帰式の解釈が難しい．

このように，回帰分析における可読性と汎化性は一般的にトレードオフの関係にある場合が多いが，多変量多項式回帰，回帰木などの手法はこれら二つの要素のバランスが比較的に優れていることが知られている．

そこで，様々な種類のデータを用いる必要があり，結果の解釈が明確である必要があるEBPM の分野を扱う本研究においては，提案手法に多変量多項式回帰を用いることとする．また，回帰木ではなく多変量多項式回帰を選んだ理由は，出力が木構造よりも多項式であった方が一般的に理解しやすいと考えたからである．以下に，代表的な多変量多項式回帰の先行研究における手法をいくつか紹介する．

BACON システム

GMDH

発見自己組織化法（Group Method of DataHandling：GMDH）とは，多入力・1 出力のデータに対して，発見的自己組織化法の原理を用いて非線形の多項式をモデリングすることができる手法である [28]．特徴として，以下のような要素が挙げられる．

- 他の手法と比較して少ないデータからモデリングを行うことができる
- 結果として得られる多項式を自己選択できる
- 求められる結果の複雑さに対して，計算量が少ない

GMDH は，その汎用性の高さと結果の可読性から様々な分野でデータ予測の手法として用いられている．例として，地価の予測や河川流量の予測などが挙げれられる．一方，得られる結果に関する制約として，各説明変数における次数は整数で与えられるという条件がある．GMDH のアルゴリズムの概要を以下に示す．

1. 説明変数のうち，独立性が強く，目的変数と相関が高いものを順に N 個選択する
2. N 個の説明変数を 2 個ずつ取り出した組合せを作り，それらに係数パラメータをかけたものの総和を部分記述式とする
3. 最小二乗法を用いて部分記述式における係数パラメータを推定し，係数パラメータと説明変数の組合せを中間変数とする
4. 中間変数と目的変数との二乗平均誤差が小さい順に中間変数を選択する
5. 中間変数と目的変数との二乗誤差が更新されなくなった際の部分記述式を最終的な完全記述式として採用して終了する
6. 誤差が更新された場合は選択基準を最適化して 2 へ戻る

RF 法

RF 法は数値的アプローチを採用した多変量多項式回帰の手法である．特徴として，各項における指数に実数を用いることができる点が挙げられる．また，モデル構造を事前に決定することを必要とせず，アルゴリズムの中で最終的なモデルを自己選択できる．これによって，より複雑な特徴を持つ多項式においても事前知識なしで容易に発見することができる．

表 2.2: 代表的な回帰モデル

出力の形	説明変数	
	量的変数のみ	量的・質的変数
線形	重回帰	数量化理論一類
		質的条件付き重回帰
多項式	多変量多項式回帰	質的条件付き多変量多項式回帰
非線形	サポートベクトル回帰，ニューラルネット回帰	
木構造	回帰木	
その他	自己回帰，ロジスティック回帰	

RF 法では，多層パーセプトロンの学習によって解を求める．また，考慮する説明変数の種類やパーセプトロンの層の数の違いによって回帰式が異なる 3 つの手法が存在する．最も基本的な手法である RF5 法は，説明変数が全て量的変数の場合に 3 層パーセプトロンの学習によって解を求める [29].

RF6.3 法は，RF5 法と同じく 3 層パーセプトロンの学習によって解を求めるが，説明変数に質的変数を考慮することができる．RF6.4 法は，RF6.3 法と同じく質的変数を考慮しつつ，発見できる多項式の表現能力を向上させるために 4 層パーセプトロンによって学習を行う手法である [30]．これらの手法のうち，RF6.4 法における定式化と学習方法の詳細は 3.1 節，3.2 節に示す.

本研究が対象とする政策立案の分野では，考慮すべき説明変数の数が非常に多く，本来求めるべき多項式の形が複雑になると考えられる．そのため，本研究における数法則の発見にはこれら二つの条件に比較的強い RF 法を用いた手法を提案する.

第3章

数法則の発見とデータの潜在的分類

§ 3.1　LPA によるデータの潜在的分類

顕在化しており実際に観測することができる変数（顕在変数）の背後にカテゴラルな潜在変数が存在するという仮定に基づいてサンプルをいくつかのクラスタに分類する潜在変数モデリング手法に潜在クラス分析（Latent class analysis: LCA）や潜在プロファイル分析（Latent profile analysis: LPA）がある．

一般に，顕在変数がカテゴラルな場合には LCA，連続変数の場合には LPA が用いられる．LPA が他の一般的なクラスタリング手法と異なる点は，各サンプルにおける変数の値に基づいたデータ指向ではなく，最尤推定を用いたモデル指向の手法であるという点である．そのため，単位やスケールが異なる複数の顕在変数からなるデータであっても，スケーリングを気にすることなくラスタリングを行うことができるという利点がある．

本節では，LPA について，モデルの基本的な概念とモデル選択の方法，選択したモデルに対して各サンプルを振り分ける方法について順に示す．

モデルの概要

§ 3.2　RF6.4 法におけるパーセプトロンの学習

いくつかの量的変数からなる説明変数と 1 つの量的変数からなる目的変数の間に成り立つ関係式を多変量多項式で表現し，3 層パーセプトロンの学習によって最適な重みを求めることによって最適な関係を導く方法は RF5 と呼ばれる．RF6.4 法とは，RF5 法を基本的な枠組みとして，入力変数に質的変数を考慮できるように拡張した手法である．

また，RF6.4 法では，重みの学習を 4 層パーセプトロンによって行う．これは，RF5 法と同様に 3 層パーセプトロンを用いて学習を行った場合，質的変数の重みの表現が線形になり，質的変数の組合せによってはうまく表現できないからである．本節では，RF6.4 法における定式化と学習方法について述べる．

まず，4 層パーセプトロンを用いた質的条件付き多変量多項式回帰法である RF6.4 法の定式化を行う．いま，解析対象とするデータが $(q_{11}, \cdots, q_{kl}, x_1, \cdots, x_m, y)$ で与えられているとする．ただし，$\mathbf{q}$ は質的説明変数，$\mathbf{x}$ は量的説明変数，y は目的変数である．

また，q_{kl} における k は質的説明変数の数，l は各質的説明変数が取りうるカテゴリーの数である．q_{kl} の値は 1 または 0 の二進法で表され，q_k がカテゴリー l に該当する場合は $q_{kl} = 1$，それ以外の場合は $q_{kl} = 0$ となる．

これらを用いて q_{kl} の組合せによって適用される回帰ルールが異なる，回帰ルール集合からによって表される質的条件付き多変量多項式は以下のようになる．

$$if \bigwedge_{k} \bigvee_{q_{kl} \in Q_k^i} q_{kl} \quad then \quad y = h(\mathbf{x}; \theta^i) + \epsilon, \quad i = 1, \cdots, I \tag{3.1}$$

ここで，Q_k^i は i 番目の回帰ルールに該当する際の q_{kl} の集合，$\theta^{\mathbf{i}}$ は i 番目の回帰ルールのパラメータベクトル，I は回帰ルールの数である．式 3.1 における $h(\mathbf{x}; \theta)$ は以下のように表せる．

$$h(\mathbf{x}; \theta) = v_0 + \sum_{g=1}^{G} v_g \prod_{m=1}^{M} x_m^{v_{gm}} \tag{3.2}$$

ただし，v_0, v_j, v_{gm} は未知の実数パラメータ，G は多変量多項式の項の数にあたる整数パラメータであり，θ は v_0, v_j, v_{gm} で構成されるパラメータベクトルである．また，式 3.1 は関数 a を用いて以下のように近似表現を行うことができるため，単一のパーセプトロンの結果として表すことができる．関数 a における $\sigma(o)$ はシグモイド関数である．

$$a(\mathbf{q}; \mathbf{v}) = \sigma\left(\sum_{k=1}^{K} \sum_{l=1}^{L_k} v_{kl} q_{kl}\right) \tag{3.3}$$

$$F(\mathbf{q}, \mathbf{x}; \mathbf{v^1}, \cdots, \mathbf{v^I}, \theta^{\mathbf{1}}, \cdots, \theta^{\mathbf{I}}) = \sum_{i=1}^{I} a(\mathbf{q}; \mathbf{v}) h(\mathbf{x}; \theta^i) \tag{3.4}$$

よって，式 3.2 は 3 層パーセプトロンを用いた学習によってもパラメータベクトルの値を求めることができる．しかし，RF6.4 法では，質的変数部分を 2 層の非線形ネットワークで表現することで回帰ルール条件部の表現能力を向上させるため，4 層パーセプトロンの学習によってパラメータベクトルの値を求める．RF6.4 法における 4 層パーセプトロンの出力結果となる式を以下に示す．

$$\begin{aligned}
&f(\mathbf{q}, \mathbf{x}; \phi) = w_0 + \sum_{g=1}^{G} w_g s_g, \\
&w_0 = \sum_{d=1}^{D} c_{0d} \sigma_d, \quad w_g = \sum_{d=1}^{D} c_{gd} \sigma_d, \\
&\sigma_d = \sigma\left(\sum_{k=1}^{K} \sum_{l=1}^{L_k} c_{dkl} q_{kl}\right), \quad s_g = \exp\left(\sum_{m=1}^{M} v_{gm} \ln x_m\right)
\end{aligned} \tag{3.5}$$

ここで，ϕ は $c_{0d}, c_{gd}, c_{dkl}, v_{gm}$ で構成されるパラメータベクトルである．また，G および D は隠れユニットの数であり，G は結果として得られる多変量多項式の項数にあたる．D は結果の式には現れない．式 3.5 を出力する 4 層パーセプトロンを図 3.1 に示す．何らかの方法によって，隠れユニット数 G および D を決定し，図 3.1 の学習を行うことで式 3.5 が得らえれる．

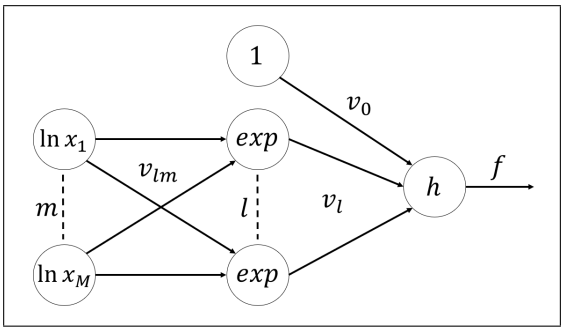


図 3.1: RF6.4 法の 4 層パーセプトロン

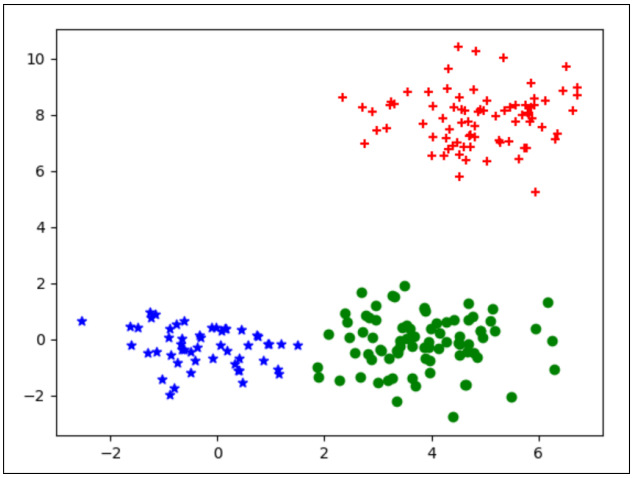


図 3.2: 二次元の k-means

いま，サンプル数 N のデータ $\{(\mathbf{q}^n, \mathbf{x}^n, y^n) : n = 1, \cdots, N\}$ が与えられているとする．このとき，図 3.1 の学習結果は，以下の二乗和誤差を求めることによって得られる．二乗和誤差および最適な G および D の求め方は以降に記述する．

$$E(\mathbf{q}, \mathbf{x}; \phi) = \frac{1}{N} \sum_{n=1}^{N} (f(\mathbf{q}, \mathbf{x}; \phi)^n - y^n)^2 \tag{3.6}$$

多層パーセプトロンの学習法

式 3.1 で示した多変量多項式は表現能力が高く，非線形なデータに対しても高い精度で数法則の表現できるが，一方で多くの局所解を含むという特徴を持つ．そのため，RF6.4 法では 2 次の学習アルゴリズムである BPQ 法を用いることで効率的に学習を行う [31].

BPQ 法とは，準ニュートン法の考え方に基づき，最適ステップ幅を二次近似の最小点として計算する手法である．以下に，RF6.4 法における準ニュートン法のアルゴリズムを示す．なお，以降では簡略化のために $f(\mathbf{q}, \mathbf{x}; \theta)$ を f と表記する．

Step 1:パラメータの初期化

ϕ_1 を任意の値に初期化し，$\mathbf{H}_1 = \mathbf{I}, b = 1$ とおく．ただし，$\mathbf{I}$ は ϕ と次元の等しい単位行列である．

Step 2: 探索方向 $\Delta\phi_s$ の計算

探索方向 $\Delta\phi_s = -\mathbf{H}_s \nabla E(\mathbf{q}, \mathbf{x}; \phi_s)$ を計算する．∇E の計算式は以下のとおりである．ここで，任意の停止条件を満たした場合，反復を終了する．

$$\nabla E = \frac{\partial E}{\partial \phi} = \sum_{n=1}^{N} (f^n - y^n) \frac{\partial f^n}{\partial \phi} \tag{3.7}$$

$$\frac{\partial f^n}{\partial c_{gd}} = \sigma_d^n s_g^n \tag{3.8}$$

$$\frac{\partial f^n}{\partial c_{dkl}} = \sigma_d^n(1-\sigma_d^n)q_{kl}^n\left(\sum_{g=0}^{G} c_{gd}s_g^n\right) \tag{3.9}$$

$$\frac{\partial f^n}{\partial v_{gm}} = w_g^n s_g^n x_m^n \tag{3.10}$$

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 を求める．最適探索幅の求め方についての詳細は後に示す.

Step 4: 結合重み ϕ_s の更新

$\phi_{\mathbf{b+1}} = \phi_{\mathbf{b}} + \alpha_{\mathbf{s}}\mathbf{\Delta}\phi_{\mathbf{b}}$ に従って結合重み ϕ_b を更新する.

Step 5: 二次微分の逆行列の近似値 H の更新

$b \equiv 0(mod\ Z, Z$: 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，$\mathbf{H}_{b+1}$ を更新する．$\mathbf{H}_{b+1}$ を更新する際の計算方法はいくつかあるが，ここでは以下の BFGS 公式を用いる [32]．また，$b \leftarrow b+1$ として，Step2 の戻る．

$$\begin{aligned}
&\mathbf{H}_{b+1} = \mathbf{H}_b + \left(1+\frac{\mathbf{q}^T\mathbf{H}_b\mathbf{q}}{\mathbf{p}^T\mathbf{q}}\right)\frac{\mathbf{p}\mathbf{p}^T}{\mathbf{p}^T\mathbf{q}} - \frac{\mathbf{p}\mathbf{q}^T\mathbf{H}_b + \mathbf{H}_b\mathbf{q}\mathbf{p}^T}{\mathbf{p}^T\mathbf{q}} \\
&\mathbf{p} = \alpha_b\Delta\phi_b \\
&\mathbf{q} = \nabla E(\mathbf{q},\mathbf{x};\phi_{b+1}) - \nabla E(\mathbf{q},\mathbf{x};\phi_b)
\end{aligned} \tag{3.11}$$

最適探索幅の計算方法

BPQ 法の Step3 における最適探索幅の計算方法を示す．なお，以降では添え字 b を省略する．また，Step3 では，変数が α しか存在しないため，$E(\mathbf{q},\mathbf{x};\phi+\alpha\Delta\phi)$ を単に $g(\alpha)$ で表す．このとき，$g(\alpha)$ の二次近似式は以下のようになる.

$$g(\alpha) \approx g(0) + g'(0)\alpha + \frac{1}{2}g''(0)\alpha^2 \tag{3.12}$$

$$g'(0) = \sum_{n=1}^{N}(f^n - y^n)f'^n \tag{3.13}$$

$$g''(0) = \sum_{n=1}^{N}((f^n)^2 + (f^n - y^n)f''^n) \tag{3.14}$$

$$f'^n = \sum_{g=0}^{G}(w_g'^n s_g^n + w_g^n s_g'^n) \tag{3.15}$$

$$f''^n = \sum_{g=0}^{G}(w_g''^n s_g^n + 2w_g'^n s_g'^n + w_g^n s_g''^n) \tag{3.16}$$

$$w_g'^n = \sum_{d=1}^{D}(\Delta w_{gd}\sigma_d^n + w_{gd}\sigma_d'^n) \tag{3.17}$$

$$w_g''^n = \sum_{d=1}^{D}(2\Delta w_{gd}\sigma_d'^n + w_{gd}\sigma_d''^n) \tag{3.18}$$

$$\sigma_g'^n = \sigma_g^n(1-\sigma_g^n)\left(\sum_{k=1}^{K}\sum_{l=1}^{L}\delta w_{drk}q_{kl}^n\right) \tag{3.19}$$

$$\sigma_g''^n = \sigma_g^n(1-\sigma_g^n)(1-2\sigma_g^n)\left(\sum_{k=1}^{K}\sum_{l=1}^{L}\delta w_{drk}q_{kl}^n\right)^2 \tag{3.20}$$

$$s_g'^n = s_g^n\sum_{m=1}^{M}\Delta v_{gm}\ln x_m^n \tag{3.21}$$

$$s_g''^n = s_g^n\left(\sum_{m=1}^{M}\Delta v_{gm}\ln x_m^n\right)^2 \tag{3.22}$$

ここで，$g'(0) > 0$ の場合，目的関数 $g(\alpha)$ の値を最小化することはできないため，$g'(0) < 0$ となるように探索方向と $\mathbf{H}$ の値をそれぞれ $\Delta\phi = -\nabla E, \mathbf{H} = \mathbf{I}$ と設定する．これにより，最適探索幅 α は $g''(0)$ の正負それぞれの場合において，以下の式で求められる．

$$\alpha = \begin{cases} -\frac{g'(0)}{g''(0)} & (g''(0) > 0) \\ -\frac{g'(0)}{\sum_{n=1}^{N}(f'^n(\mathbf{x}^n;\phi))^2} & (g''(0) \leq 0) \end{cases} \tag{3.23}$$

また，明らかに $g'(0) < 0$ であるとき，式 3.23 で求められる値は正となる．この場合，探索位置が鞍点付近にある可能性があるので，$b = Z, \mathbf{H} = \mathbf{I}$ とする．以上の方法で最適探索幅 α の値を計算することができるが，この方法では α を近似によって求めているため，目的関数 $g(\alpha)$ 値が常に減少するとは限らない．

よって，$g(\alpha) \geq 0$ 場合，以下に示す式にしたがって α の値を更新し，この作業を $g(\alpha) < g(0)$ となるまで繰り返すことで常に $g(\alpha) < g(0)$ となるような α の値を求める．

$$\tilde{\alpha} = -\frac{g'(0)\alpha^2}{2(g(\alpha) - g(0) - g'(0)\alpha)} \tag{3.24}$$

§ 3.3　RF6 法におけるモデル選択とルール復元

最適な中間ユニット数 G, D の判定

図 3.1 に示した 4 層パーセプトロンにおける中間ユニット G の数は式 3.5 における項数と一致する．また，中間ユニット D は結果に直接現れないが，結果の精度に関わる．そのため，目的変数と説明変数の間に成り立つ関係をもっともよく表現する式 3.5 を求めるためには，最適な中間ユニットの数 G, D を何らかの方法で求める必要がある．

また，前述の BPQ 法を用いた学習おいては G, D の値が未知であるため，任意の自然数を G, D と置いて計算を行う必要がある．G, D の値を大きくすれば，パーセプトロンの学習による訓練誤差の値は小さくなるが，その場合，データに含まれるノイズにオーバーフィットした結果を得てしまう可能性が懸念される．

そこで，RF 法では最適な G, D の値を決定するためのモデル評価尺度として，主にベイズ情報量基準（Bayesian information criterion: BIC）または，交差検証法による結果を用いる．これらの手法のうち，BIC は計算量が比較的少なく効率的に中間ユニットの数を判定することができる．BIC の基本形を以下に示す [33].

$$BIC = -2L + k \ln n \tag{3.25}$$

ただし，L は最大対数尤度，k は自由度，n はデータ数である．これを RF6.4 法における中間ユニット数 G, D，サンプル数 N，説明変数 $\mathbf{q}, \mathbf{x}$，目的変数 y，学習によって得られたパラメータベクトル ϕ，パラメータ数 Z を用いて表すと以下のようになる．

$$BIC(G, D) = \frac{N}{2} \ln \left(\frac{1}{N} \sum_{n=1}^{N} (f(\mathbf{q}^n, f(\mathbf{x}^n; \hat{\phi}_{G,D}) - y^n)^2 \right) + \frac{Z}{2} \ln N \tag{3.26}$$

ただし，$\hat{\phi}_{G,D}$ は中間ユニットの数が G, D のパーセプトロンの学習によって得られたパラメータベクトルを表す．RF6.4 法では，パーセプトロンの学習を行う前に任意の自然数を中間ユニット数 G, D とし，それによって学習を行う．そして，学習によって得られたパラメータベクトルを用いて BIC を計算し，BIC の値がより小さいときのパラメータベクトルを最適な学習結果とする．

正則化法

RF6.4 法のような多変量多項式を用いた学習モデルは非常に強力な学習モデルである．そのため，特に制約なくこのようなモデルの学習を行うと学習結果が訓練データに含まれるノイズに適合しすぎてしまい，汎化性能が低下する．これをオーバーフィッティングという．

そのため，RF6.4 法ではオーバーフィッティングを防ぐために，あえて制約をかけて学習を行うことで汎化性能を向上させる．学習において目的関数に本来求めたい値とは別のパラメータ（正則化項）を加えることで学習に制約をかける方法は正則化と呼ばれる．目的関数に新たなパラメータを加えることによって，学習に対するパラメータベクトルの影響を小さくすることができ，オーバーフィッティングを防ぐことができる．

正則化の方法の 1 つである重み減衰法では，パラメータベクトルの平方和を新たな正則化項として加える．正則化項を加えた目的関数を $e(\mathbf{q}, mathbfx; \phi)$ として以下に示す．

$$e(\mathbf{q}, \mathbf{x}; \phi) = E(\mathbf{q}, \mathbf{x}; \phi) + \lambda \left(\frac{1}{2} \sum \phi^2 \right) \tag{3.27}$$

ルール復元の方法

RF6.4 法では，パーセプトロンの学習結果を用いて，式 3.1 の形で回帰ルール集合を復元する必要がある．また，式 3.1 を用いれば，質的説明変数の全カテゴリーに対する全ての組合せに対応した回帰ルール集合を表すことができるが，その場合，それぞれの回帰ルール

は質的変数の組合せが全く同じである少数のサンプルのみにしか対応せず，特化しすぎている．

そこで，ここでは，複数の組合せに対応したより一般的な回帰ルールの集合を復元する方法について示す．パーセプトロンの学習によって得らえた最適なパラメータを $\hat{c}_{gd}, \hat{c}_{dkl}$ とすると，式 3.5 における，あるサンプル n に対する第 g 項の係数は，以下のようになる．

$$w_g^n = \sum_{d=1}^{D} \hat{c}_{gd} \hat{\sigma}_d^n, \quad \hat{\sigma}_d^n = \sigma \left(\sum_{k=1}^{K} \sum_{l=1}^{L_k} \hat{c}_{dkl} q_{kl}^n \right) \tag{3.28}$$

ここで，$k-means$ 法を用いることで N 個の係数値ベクトル $\{\mathbf{w^n}: n = 1, \cdots, N\}$ を I 個の代表点 $\{\mathbf{u^i}: i = 1, \cdots, I\}$ にベクトル量子化することを考える [34]．$k-means$ 法では，はじめに各データをランダムなクラスタに割り振り，それぞれのクラスタの重心を求める．次に，各データと求めた重心との距離が最も近くなるように各データを再度クラスタに割り振る．これをクラスタの重心が動かなくなるまで繰り返す．

例として，二次元データに対する k-means の結果を図 3.2 に示す．この例では，各サンプルが二次元の値を持ち，そのサンプル数が 240 のデータを用いて，それらを 3 つのクラスタに分類している．

RF6.4 では，この k-means を用いてルール復元を行うために，それぞれ N_i 個のデータを互いに素であるクラスタ $\{R_i, i = 1, \cdots, I\}$ に含めた場合に以下の式 3.29 によって求められる値が最小となる場合の代表点 $\{\mathbf{u^i}: i = 1, \cdots, I\}$ を求める．また，ここでの代表点とは，各クラスタの重心のことである．

$$dis = \sum_{i=1}^{I} \sum_{n \in R_i} \|\mathbf{w^n} - \mathbf{u^i}\|^2, \quad \mathbf{u^i} = \frac{1}{N_i} \sum_{n \in R_i}^{N_i} \mathbf{w^n} \tag{3.29}$$

式 3.29 によって代表点 $\{\mathbf{u^i}: i = 1, \cdots, I\}$ が得られた場合，復元された回帰ルール集合は以下の式 3.30 のように与えられ，あるサンプル n が属する代表点の番号を求める式は以下の式 3.31 のようになる．

$$if i(\mathbf{q}) = i \quad then \quad \hat{f} = u_0^i + \sum_{g=1}^{G} u_g^i \prod_{m=1}^{M} x_m^{\hat{v}_g m}, \quad i = 1, \cdots, I \tag{3.30}$$

$$i(\mathbf{q}) = \arg \min_i \|\mathbf{w^n} - \mathbf{u^i}\|^2 \tag{3.31}$$

次に，N 個のサンプルをいくつの代表点 $\{\mathbf{u^i}: i = 1, \cdots, I\}$ に分けるかであるが，これには交差検証法と呼ばれる手法を用いる [35]．交差検証法では，与えられたデータをランダムに A 個に分割し，$A-1$ 個を学習用に，残りの 1 個をテスト用に用いて平均二乗誤差を算出する．

この処理を A 個全てのデータセグメントがテストに用いられるように繰り返し行い，それら A 回の結果の総和が最も小さいものを最適なモデルとして採用する．今回の例では，代表点の個数をあらかじめいくつか定め，それらの中で最適な個数を交差検証法によって求める．

第4章

提案手法

§ 4.1　LPA によるデータのクラスタリング

§ 4.2　潜在的クラスと RF6.4 法を用いた数法則発見

本節では，社会における実情の間にどのような関係が成り立っているかを把握するために，観測されたデータのうちの 1 つが他のデータによってどのような説明されるかをデータの観点から求める手法について述べる．また，得られた結果を用いて観測されたデータを更新することについても言及する．

社会実情データを用いた数法則発見

既存の観測データを用いてそれらの間に成り立つ法則を求める手法は，2.3 節でも述べた通り複数存在する．また，本研究において対象とする社会実情の法則化では，以下の 3 つの条件を満たす手法が必要と考えられる．

1. 事前に式の項数などの形を指定しなくても結果を得ることができる
2. 各説明変数が目的変数に与える影響の大小の表現力が高い
3. 結果の解釈が可能であり，データ間の関係性が認識できる

1 については，社会における事象は複雑であり，事前に前提知識として式の形を与えることが困難であるからである．2 については，各説明変数がどのように目的変数に寄与しているかをより詳細に表現することが好ましいからである．3 については，結果の解釈が難解な手法は，関係性を把握するという目的があるからである．

以上のことから，本研究では事前に結果の項数を指定することなく結果を求めることができ，各説明変数における次数に制限の少ない RF 法を用いて観測データ間の関係性を求める手法を提案する．また，RF 法における結果は説明変数が目的変数に与える影響を判読可能な数式として得られるため，上記の 3 つのすべてに適している．

次に，観測された社会実情データを用いてどのように RF 法を行うかであるが，その前に観測される社会実情データにおける特徴について考える．本研究で想定する社会実情データは，各自治体における人口や税収などのため，量的変数である．そのため，一見すると量的データのみを考慮して数法則の発見を行う RF5 法を用いて分析を行うことが最適なように思える．

しかし，社会実情データは，数年に一度しか観測されないデータである．そのため，それらのデータを用いて学習を行い，各変数間の関係を数式化するためには，時間軸ではなく空間軸で大量に収集したデータを用いることになる．

つまり，あるデータ A を他のデータ B，C，D で説明する数式を求めることによってそれらの間に成り立つ関係を把握したい場合，ある自治体におけるデータセットの A から D の項目はそれぞれ 1 つずつしかデータが存在しないため，自治体を増やすという方法でしか分析に用いる大規模な学習データを確保できないということである．本研究における学習データの構造を図 4.1.

このようなデータを用いて学習を行う場合，新たに懸念される課題が発生する．それは，すべての自治体が同一な特徴を持つサンプルとして扱うことは正しいのかということである．例えば，2024 年現在，日本には約 1700 の市区町村が存在するがそれらの中には首都圏とその周辺にあるような世界有数の大規模都市から，人口の少ない地方の村までさまざまな人口規模の自治体が含まれる．

このような特徴は人口のみに限ったものではなく，そこに住む人々の構成や主要産業など扱うデータセットの項目数に応じて無数に存在し，そのすべてを人間の手によって事前に把握することは不可能といえる．

そのため，それらの自治体を対象として集められたすべてのサンプルを特に処理することなく学習サンプルとして用いて RF5 法を行った場合，得られる結果は日本全国における特徴を一般化したものとなり，それを用いて特定の自治体について言及する場合，精度が落ちることが考えられる．

よって，本研究では空間軸で収集した大規模な学習データに対して，一度それらの中に存在する特徴をいくつかのパターンに分類し，そのパターンを用いて各自治体をいくつかのグループに振り分ける．また，その結果を説明変数に含めて RF 法による分析を行うことで各変数間に成り立つ数法則を発見する手法を提案する．

まず，学習データに対して，それらを特徴ごとにいくつかのパターンに分類する手法だが，これには 4.1 節と同様に LPA を用いる．つまり，4.1 節の手法によって求めた日本全国における自治体ベースの潜在クラスとそれぞれの自治体における存在確率を用いて RF 法に用いる学習データに背景知識を与え，それぞれの潜在クラスに特化した数法則を求める．

どのようにして存在確率を組み込んだ RF 法を行うかであるが，これには RF5 法ではなく RF6.4 法を用いる．3.1 節および 3.2 節でも述べたとおり，RF6.4 法では説明変数に質的変数を用いてサンプルを部分空間に切り分けることによって各サンプルの状況に応じた数法則を発見することができる．

よって，LPA によって求めた各潜在クラスに対するサンプルの存在確率を RF6.4 法における質的説明変数として扱う．これにより，潜在クラスを用いてサンプルを部分空間に切り分け，それぞれにあった数法則を発見することが可能になる．

ここで，存在確率を質的変数として扱うと述べたが，存在確率は 0 以上 1 未満の実数で与えられる．そのため，そのままでは RF6.4 法における質的変数として扱うことができない．よって，何らかの方法で実数である存在確率をカテゴリーに置き換え，質的変数として扱うことができる形にすることが必要である．

本研究の提案手法では，存在確率が保持する情報をなるべく減らさない形で量子化することを考える．まず，単一のサンプルにおける存在確率について考える．存在確率は LPA

図 4.1: 学習データの一例

図 4.2: 存在確率のカテゴリー化

によって得られたいくつかのクラスに対して各サンプルがどれくらいの確率で属するかを表した値であるため，単一のサンプルにおけるすべての存在確率の合計は 1 となる．

このことから，LPA によって得られたクラス数が仮に A から E の 5 つであった場合，一番存在するクラスがあいまいなサンプル（5 つのクラスに均等に存在するとされたサンプル）が持つ存在確率は A から E のすべてに対して 0.2 となる．よって，このような場合，RF6.4 法における質的変数の個数とそれらがとりうるカテゴリー数を 5×5 とすると，質的変数によってすべての存在確率のパターンが表せることになる．存在確率のカテゴリー化の例を図 4.2 に示す．

本研究の提案手法では，このような方法を用いて存在確率を質的変数に変換し，それによってサンプルを部分空間に切り分けることで，学習を行う．これにより，空間軸で収集したために特徴にばらつきがあると考えられるサンプルに対しても，ある程度それらの特徴を考慮しながら数法則を発見することができる．

また，従来の RF6.4 法では，パーセプトロンの学習結果を用いてルールを復元する際に最適なルール数が分からない．そのため，あらかじめルール数をいくつか定め，それぞれのルール数で復元を行った後，その結果に対して交差検証法を用いることで最適なルール数を発見している．

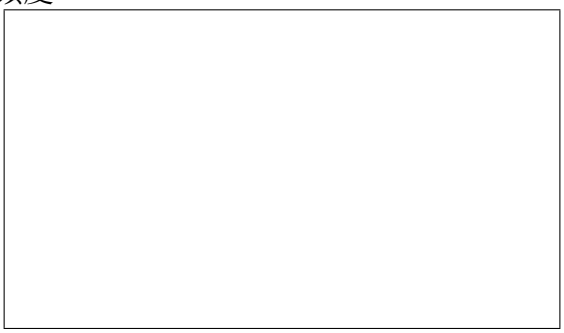
一方，本研究の提案手法では，事前に LPA を用いてデータサンプルがいくつに分けられるかを求めているため，その結果によって得られたクラス数をルール数として採用する．これにより，大規模なデータサンプルに対しても交差検証法を用いずにルール数が決定できるため，計算量削減につながると考えられる．

数法則を用いた観測データの更新

EBPM の分野で分析に用いられるデータは，費用や労力などコスト面の負担が大きいことから，高い頻度で収集することが非常に困難であるという特徴を持つ．そのため，それらのデータには数年に一度しか更新されないものも多く，更新されたとしても項目によってタイミングにばらつきが見られる．

しかし，これらの課題をデータの収集方法のみの改善を用いて解決することは現実的ではない．なぜなら，前述のコストに関する問題に加えて，データの収集にあたる機関がそれぞれによって異なるという構造があるからである．このようなデータの一部とその更新頻度について，オープンデータとして公開されているものを例として，表 4.1 に示す．

表 4.1: EBPM に用いられるデータの更新頻度



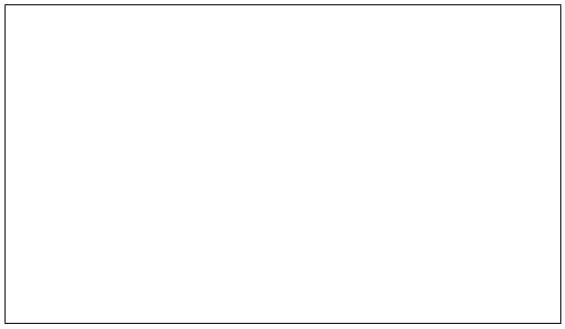


図 4.3: 数法則発見によるデータ予測

このような高頻度に収集することが難しいデータに対して，本研究では，前述の RF6.4 法の結果である数式と量的説明変数それぞれのデータ項目における可能な限り最新の値を用いて最新の目的変数の値を予測しその結果によってデータの更新を行うこと考える．

表 4.1 のとおり，社会実情データは数年に一度しか更新されないものも多いが，それらの更新周期が重なる一年に限定すれば最新のデータがそろっている年は存在する．まず，その年のデータを用いて社会実情データの間に成り立つ関係性を本研究の提案手法によって数法則化する．

次に，その分析に用いた量的説明変数の全データ項目について，可能な限り最新の値を収集する．収集した値を提案手法によって求めた数法則に代入することによって目的変数とした項目に関する最新の予測値を算出する．数法則発見を用いたデータ予測のながれを図 4.3 に示す．

§ 4.3　Web-GIS 描画による潜在的な法則の可視化

第5章

数値実験並びに考察

§ 5.1　数値実験の概要

　メモ：一度数値実験を回した後に，最も重要そうな項目を取り除いて再度分析を行った際にどのような結果が得られるか

§ 5.2　実験結果と考察

第6章

おわりに

謝辞

本研究を遂行するにあたり，多大なご指導と終始懇切丁寧なご鞭撻を賜った富山県立大学工学部電子・情報工学科情報基盤工学講座の奥原浩之教授，António Oliveira Nzinga René 講師に深甚な謝意を表します．また，システム開発および数値実験にあたり，ご助力いただいた富山県立大学電子・情報工学科 3 年生の島部達哉氏に感謝の意を表します．最後になりましたが，多大な協力をしていただいた研究室の同輩諸氏に感謝致します．

2022 年 2 月

長瀬 永遠

参考文献

[1] 内閣府, ”内閣府における EBPM への取組”, 閲覧日 2022-02-08, https://www.cao.go.jp/others/kichou/ebpm/ebpm.html.

[2] 杉谷和哉, ”行政事業レビューにおける EBPM の実践についての考察”, 日本評価学会, Japanese journal of evaluation studies, Vol. 21, No. 1, pp. 99-111, 2021.

[3] 中泉拓也, ”英国の EBPM（Evidence Based Policy Making）の動向と我が国への EBPM 導入の課題”, 関東学院大学経済経営研究所年報, Vol. 41, pp. 3-9, 2019.

[4] 井伊雅子, 五十嵐中, ”新医療の経済学：医療の費用と効果を考える”, 日本評論社, 2019.

[5] 中村圭, ”地方自治体における EBPM の進め方とは？【基本編】地域課題の解決に繋がる EBPM に向けて”, 閲覧日 2024-02-06, https://www.fujitsu.com/jp/group/fjm/business/mikata/column/local-government/fri-nakamura/.

[6] Slack 参照

[7]

[8] 論文探す

[9]

[10] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, ”DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model”, Journal of Machine Learning Research, Vol. 12, pp. 1225-1248, 2011.

[11] 末吉俊幸, ”DEA-経営効率分析法-”, 朝倉書店, 2001.

[12] 国土交通省国土地理院, ”GIS とは”, 閲覧日 2022-02-08, https://www.gsi.go.jp/GIS/whatisgis.html.

[13]

[14]

[15]

[16]

[17]

[18] 卒論から

[19] 卒論から

[20] 金成賢作, 篠原正明, ”DEA における入力指向と出力指向の比較（その 1）”, 日本大学生産工学部第 42 回学術講演会, 2009.

[21] 刀根薫, ”包絡分析法 DEA”, 日本ファジィ学会誌, Vol. 8, No. 1, pp. 11-14, 1996.

[22] 論文探す

[23] 卒論から 藤井さんの方

[24] 卒論から 刀根さんの方

[25]

[26]

[27]

[28] 参考論文ファイルから

[29]

[30]

[31]

[32] Slack 参照

[33]

[34] Slack

[35]

[36] 佐藤主光, ”税財政分野における EBPM の基礎と活用”, 閲覧日 2022-02-08, https://www.ipp.hit-u.ac.jp/satom/lecture/localfinance/2019_local_note07.

[37] esri ジャパン,” GIS（地理情報システム）とは” , 閲覧日 2022-02-08, https://www.esrij.com/getting-started/what-is-gis/.

[38] 国土交通省国土地理院, ”基盤地図情報の利活用事例集”, 閲覧日 2022-02-08, https://www.gsi.go.jp/common/000062939.

[39] esri ジャパン, ”東日本大震災対応における政策形成支援に GIS を活用”, 閲覧日 2022-02-08, https://www.esrij.com/industries/case-studies/35859/.

[40] 田中貴宏, 佐土原聡, ”都市化ポテンシャルマップと二次草原潜在生育地マップの重ね合わせによる二次草原消失の危険性の評価：ー福島県旧原町市域を対象として”, 環境情報科学論文集, Vol. 23, pp. 191-196, 2009.

[41] 坪井利樹, 西田佳史, 持丸正明, 河内まき子, 山中龍宏, 溝口博, ”身体地図情報システム”, 日本知能情報ファジィ学会誌, Vol. 20, No. 2, pp. 155-163, 2008.

[42] 杉原豪, 塚井誠人, ”統計的因果探索による社会基盤整備のストック効果の検証”, 土木学会論文集 D3（土木計画学）, Vol. 75, no.6, pp. 583-589, 2020.

[43] Dentsu Digital Tech Blog, ”Google Colab で統計的因果探索手法 LiNGAM を動かしてみた”, 閲覧日 2022-02-08,
https://note.com/dd_techblog/n/nc8302f55c775.

[44] 藤井秀幸, 傅靖, 小林里佳子, ”データ包絡分析を用いたふるさと納税の戦略提案-K 市のふるさと納税への適用事例-”, 日本経営工学会論文誌, Vol. 71, No. 4, pp. 149-172, 2021.

[45] 日本オペレーション・リサーチ, ”第 4 章 包絡分析-入力と出力と”, 閲覧日 2022-02-08,
http://www2.econ.tohoku.ac.jp/ ksuzuki/teaching/2006/ch4.

[46] pork_steak, ”folium 事始め”, 閲覧日 2022-02-08,
https://qiita.com/pork_steak/items/f551fa09794831100faa.

[47] 保母敏行ほか, ”日本分析学会における標準物質の開発”, 日本分析化学会誌, vol. 57, No. 6, pp. 363-392, 2008.

[48] 射水市役所, ”総合戦略-射水市”, 閲覧日 2022-02-08,
https://www.city.imizu.toyama.jp/appupload/EDIT/054/054185.

[49] 射水市役所, ”共通課題-射水市”, 閲覧日 2022-02-08,
https://www.city.imizu.toyama.jp/appupload/EDIT/024/024383.