修士論文

証拠に基づく政策立案のための
数法則発見法と潜在プロファイル分析を用いた
社会事象モデル化とその可視化

Data Fusion through Web-GIS Visualization
Using Open Data 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．

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）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）の考え方を組み合わせた手法を提案する．

加えて，筆者が以前に行った研究である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に対して，本研究の提案手法の結果を用いたアップグレードを行う．

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，アップグレードした分析手法を用いて実際の市に対する分析を行い，以前との結果の違いを考察する．

§ 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

第 2 章

第 3 章

第 4 章

第 5 章

第 6 章

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 [9]．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 [10].

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 [11]．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

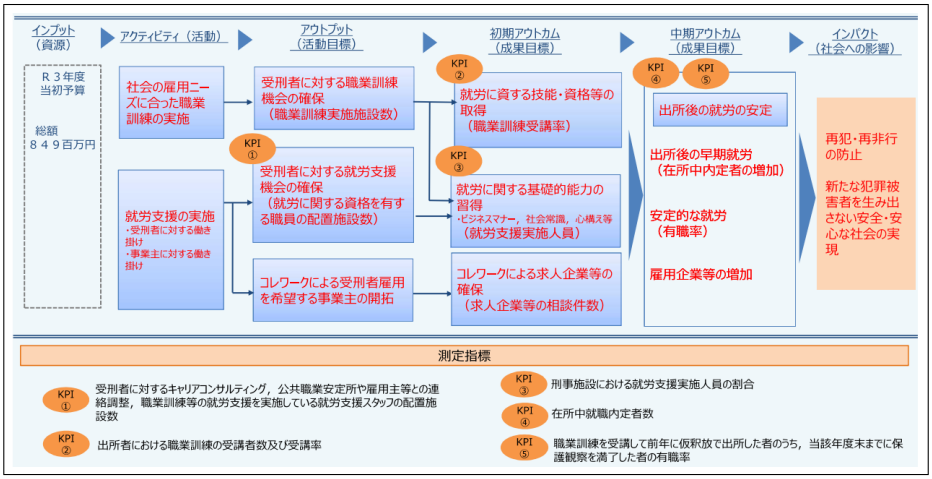

図 2.1: ロジックモデルの例

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 [12]．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）

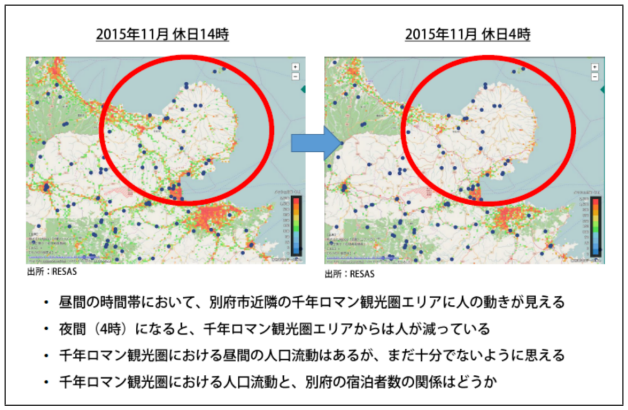

図 2.3: 別府市における RESAS 活用事例

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 [13]．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．

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，本研究では分析の結果を地図上に表示することを考える．

§ 2.2　EBPM の推進に向けたデータサイエンス

EBPM を効果的に実行するために，政策の目的を明確化し，それに沿ったデータを適切に収集・分析することが重要であるのは前述のとおりである．目的に対して必要なデータを収集し，正しく分析することはデータサイエンスと同義といえる．本節ではデータサイエンスの分野における EBPM の推進に向けた取り組みとして，実空間からデータを効果的に取得し分析した事例，政府によって公開されているデータに統計的分析を加えることで新たな知見を発見した事例 [14] の二つを挙げ，その内容を示す．

観光 EBPM に向けた大規模観光人流データ分析

EBPM 推進のためのデータ間における因果関係の分析

よって，次節では前述の課題解決のために本研究の提案手法で用いる数法則発見法について，一般的な概念と

§ 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 [15].

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 [16]，サポートベクトル回帰 [17] 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をいくつか紹介する．

BACON システム

GMDH

RF 法

RF 法は GMDH と同じく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．また，GMDH との違いとして，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．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．

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 [18].

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 [19]．これらの手法のうち，RF5 法および 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 EBPM 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．

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	

第3章

数法則の発見とデータの潜在的分類

§ 3.1　RF5 法による量的変数からの数法則発見

3 層パーセプトロンを用いた多変量多項式回帰法である RF5 法の定式化を行う．まず，M 個の説明変数を $\mathbf{x} = (x_1, x_2, \cdots, x_M)$，目的変数を y とすると，RF5 法によって求められる多変量多項式は一般に以下のように表すことができる．

$$y = v_0 + \sum_{l=1}^{L} v_l \prod_{m=1}^{M} x_m^{v_{lm}} + \epsilon \tag{3.1}$$

ただし，定数項 v_0，各項の係数 v_l，各項の説明変数の指数 v_{lm} は未知の実数パラメータ，項数 L は未知の整数パラメータ，ϵ は誤差項である．さらに，全ての説明変数において $x_m > 0$ が成り立つと仮定すると，式 3.1 は以下のように変形できる．

$$y = v_0 + \sum_{l=1}^{L} v_l \exp\left(\sum_{m=1}^{M} v_{lm} \ln x_m\right) + \epsilon \tag{3.2}$$

ここで，$v_0, \{v_l\}, \{v_{lm}\}$ をまとめて ϕ と表し，$D(= ML + L + 1)$ を ϕ の次元（パラメータ数）とする．すると，式 3.2 の右辺の一部は以下のように $f(\mathbf{x}; \phi)$ と表すことができ，これは図 3.1 に示す 3 層パーセプトロンの前向き伝播と同義である．

$$f(\mathbf{x}; \phi) = v_0 + \sum_{l=1}^{L} v_l \exp\left(\sum_{m=1}^{M} v_{lm} \ln x_m\right) \tag{3.3}$$

なお，図 3.1 における中間ユニット数は式 3.1 の定数項を除いた項数 L に対応することが分かる．ここで，M の最適な値は未知であるため，何かの基準に基づいて決定する必要がある．最適な M の値を求める方法は後に示す．いま，サンプル数 N のデータ $\{(\mathbf{x}^n, y^n) : n = 1, \cdots, N\}$ が与えられており，最適な L が決定されているとすると，データを最もよく説明する ϕ は以下の式を最小化するものを求めることで決まる．また，これは 3 層パーセプトロンの学習問題と同義である．

$$E(\mathbf{x}; \phi) = \frac{1}{2} \sum_{n=1}^{N} (f(\mathbf{x}^n; \phi) - y^n)^2 \tag{3.4}$$

多層パーセプトロンの学習法

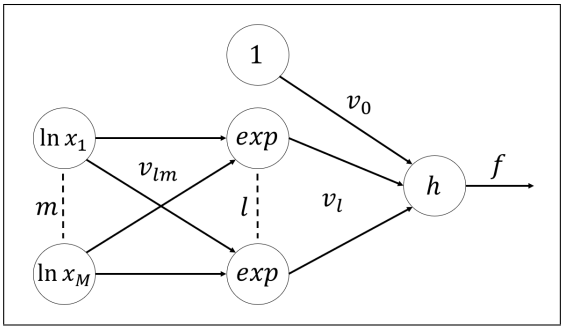

図 3.1: RF5 法の 3 層パーセプトロン

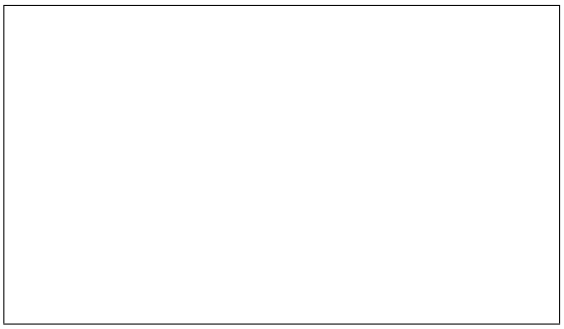

図 3.2: AAA

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，RF5 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 [20]．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，RF5 法における準ニュートン法のアルゴリズムを示す．

Step 1:パラメータの初期化

ϕ_1 を任意の値に初期化し，$\mathbf{H}_1 = \mathbf{I}, s = 1$ 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 $\Delta\phi_s = -\mathbf{H}_s \nabla E(\mathbf{x}; \phi_s)$ 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．

$$\nabla E = \frac{\partial E}{\partial \phi} = \sum_{n=1}^{N} (f^n - y^n) \frac{\partial f^n}{\partial \phi} \tag{3.5}$$

$$\frac{\partial f^n}{\partial v_0} = 1 \tag{3.6}$$

$$\frac{\partial f^n}{\partial v_l} = \exp\left(\sum_{m=1}^{M} v_{lm} \ln x_m \right) \tag{3.7}$$

$$\frac{\partial f^n}{\partial v_{lm}} = v_l \exp\left(\sum_{m=1}^{M} v_{lm} \ln x_m \right) \ln x_m \tag{3.8}$$

Step 3: 最適探索幅 α_s の計算

$E(\mathbf{x}; \phi_s + \alpha_s \Delta\phi_s)$ を最小にする最適探索幅 α_s 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

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

Step 5: 二次微分の逆行列の近似値 H の更新

$s \equiv 0 (mod\ D, D$: 全パラメータ数) のとき，$\mathbf{H}_{s+1} = \mathbf{I}$ とし，それ以外のとき，$\mathbf{H}_{s+1}$ を

更新する．$\mathbf{H}_{s+1}$ 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 [21]．また，$s \leftarrow s+1$ として，Step2 の戻る．

$$\begin{aligned}
\mathbf{H}_{s+1} &= \mathbf{H}_s + \left(1 + \frac{\mathbf{q}^T\mathbf{H}_s\mathbf{q}}{\mathbf{p}^T\mathbf{q}}\right)\frac{\mathbf{p}\mathbf{p}^T}{\mathbf{p}^T\mathbf{q}} - \frac{\mathbf{p}\mathbf{q}^T\mathbf{H}_s + \mathbf{H}_s\mathbf{q}\mathbf{p}^T}{\mathbf{p}^T\mathbf{q}} \\
\mathbf{p} &= \alpha_s \Delta\phi_s \\
\mathbf{q} &= \nabla E(\mathbf{x};\phi_{s+1}) - \nabla E(\mathbf{x};\phi_s)
\end{aligned} \tag{3.9}$$

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 s を省略する．また，Step3 では，変数が α しか存在しないため，$E(\mathbf{x};\phi+\alpha\Delta\phi)$ 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．

$$g(\alpha) \approx g(0) + g'(0)\alpha + \frac{1}{2}g''(0)\alpha^2 \tag{3.10}$$

$$g'(0) = \sum_{n=1}^{N}(f(\mathbf{x}^n;\phi) - y^n)f'(\mathbf{x}^n;\phi) \tag{3.11}$$

$$g''(0) = \sum_{n=1}^{N}((f'(\mathbf{x}^n;\phi))^2 + (f(\mathbf{x}^n;\phi) - y^n)f''(\mathbf{x}^n;\phi)) \tag{3.12}$$

$$f'^n(\mathbf{x}^n;\phi) = \Delta v_0 + \sum_{l=1}^{L}(\Delta v_l m_l^n + v_l m_l'^n) \tag{3.13}$$

$$f''^n(\mathbf{x}^n;\phi) = \sum_{l=1}^{L}(2\Delta v_l m_l'^n + v_l m_l''^n) \tag{3.14}$$

$$m_l^n = \exp\left(\sum_{m=1}^{M} v_{lm}^n \ln x_m^n\right) \tag{3.15}$$

$$m_l'^n = m_l^n \sum_{m=1}^{M} \Delta v_{lm} \ln x_m^n \tag{3.16}$$

$$m_l''^n = m_l^n \left(\sum_{m=1}^{M} \Delta v_{lm} \ln x_m^n\right)^2 \tag{3.17}$$

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．

$$\alpha = \begin{cases} -\frac{g'(0)}{g''(0)} & (g''(0) > 0) \\ -\frac{g'(0)}{\sum_{n=1}^{N}(f'^n(\mathbf{x}^n;\phi))^2} & (g''(0) \leq 0) \end{cases} \tag{3.18}$$

また，明らかに $g'(0) < 0$ であるとき，式 3.18 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，$s = D, \mathbf{H} = \mathbf{I}$ 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．

$$\tilde{\alpha} = -\frac{g'(0)\alpha^2}{2(g(\alpha) - g(0) - g'(0)\alpha)} \tag{3.19}$$

最適な中間数 L の判定

図 3.1 に示した 3 層パーセプトロンにおける中間ユニットの数は式 3.1 における項数と一致する．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 3.1 を求めるためには，最適な中間ユニットの数 L 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 L の値が未知であるため，任意の自然数を L と置いて計算を行う必要がある．L 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 L 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．ここでは，これらの手法のうち，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 による手法を示す [22].

$$BIC = -2L + k \ln n \tag{3.20}$$

式 3.20 は BIC の基本形である．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 RF5 法における中間ユニット数 L，サンプル数 N，説明変数 $\mathbf{x}$，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 D を用いて表すと以下のようになる．

$$BIC(L) = \frac{N}{2} \ln \left(\frac{1}{N} \sum_{n=1}^{N} (f(\mathbf{x}^n; \phi_L) - y^n)^2 \right) + \frac{D}{2} \ln N \tag{3.21}$$

正則化法

§ 3.2 RF6 法による質的変数を考慮した数法則発見

§ 3.3 LPA によるデータの潜在的分類

本節では，本研究の提案手法において，RF 法の精度向上のために用いる LPA という手法について

第4章

提案手法

§ 4.1　LPA によるデータのクラスタリング

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例とその更新頻度を表 4.1 に示す．

本研究では，過去に収集されている複数項目のデータに対して，いずれか 1 項目を目的変数，残りのすべてを説明変数と置いて，3.1 節および 3.2 節で示した RF 法を適用することでデータ間に成り立つ一般的な数法則を発見する．また，その数法則を用いることで目的変数に選択した項目における可能な限り現状に近いデータを予測し，前述した高頻度にデータを収集することが難しいという課題を補完する手法を提案する．数法則の発見によるデータ補完のイメージを図 4.1 に示す．

2.3 節で示した通り，RF 法にはパーセプトロンや分析に用いることの出来る変数の特徴などによって，いくつかの手法が存在する．特に，変数の特徴に着目すると，説明変数に量的変数のみを用いることができる RF5 法と量的変数，質的変数の両方を用いることができる RF6 法に分けられる．

本研究では行政機関によって収集された様々なデータのうち，量的変数である項目のみを扱うことを想定している．なお，データの収集方法や具体的なデータベースの作成手順は 4.3 節に記載する．このような前提があるため，本研究の提案手法では RF5 法を用いて数法則の発見を行うことが妥当なように思える．

しかし，各データにおける特徴のみに限らず，データベース全体を

表 4.1: EBPM に用いられるデータの更新頻度

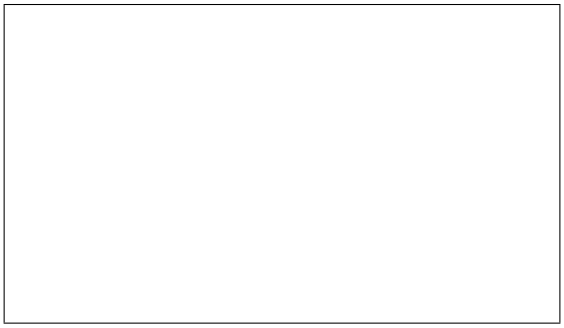

図 4.1: 数法則発見によるデータ補完のイメージ

§ 4.2　潜在的クラスとRF6.4を用いた数法則発見

§ 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] 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.

[7] 末吉俊幸, ”DEA-経営効率分析法-”, 朝倉書店, 2001.

[8] 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.

[9]

[10]

[11]

[12]

[13]

[14]

[15]

[16]

[17]

[18]

[19]

[20]

[21] Slack 参照

[22]

[23] 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.

[24] 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/.

[25] 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.

[26] 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/.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

[38] 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.