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 における観測変数	x_i, x_j
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	c_{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
CCR モデルにおける各入力に対する重み	$\mathbf{v_{in}}^T$
CCR モデルにおける各出力に対する重み	$\mathbf{u_{out}}^T$
CCR モデルにおける対象 DMU の評価値	z_o
CCR モデルにおける対象 DMU の入力	$\mathbf{x}_o$
CCR モデルにおける対象 DMU の出力	$\mathbf{y}_o$
CCR モデルにおける DMU 集合の入力	$\mathbf{X}$
CCR モデルにおける DMU 集合の出力	$\mathbf{Y}$
入力指向モデルにおける対象 DMU の n_i 番目の入力に対する改善案	$\hat{x}_{n_i}$
入力指向モデルにおける参照集合内の r_d 番目の DMU の n_i 番目の入力	$x_{n_i r_d}$
入力指向モデルにおける参照集合内の r_d 番目の DMU に対する重み	λ_{r_d}
出力指向モデルにおける対象 DMU の n_j 番目の出力に対する改善案	$\hat{y}_{n_j}$
出力指向モデルにおける参照集合内の r_d 番目の DMU の n_o 番目の出力	$y_{n_o r_d}$
出力指向モデルにおける参照集合内の r_d 番目の DMU に対する重み	μ_{r_d}
RF6.4 法における質的説明変数	$\mathbf{q}$
RF6.4 法における量的説明変数	$\mathbf{x}$
RF6.4 法における目的変数	y
RF6.4 法における i 番目の回帰ルールのパラメータ	θ^i
RF6.4 法における誤差項	ϵ
RF6.4 法における回帰ルールの定数項	v_0
RF6.4 法における回帰ルールの各項に対する係数	v_g
RF6.4 法における回帰ルールの量的説明変数に対する次数	v_{gm}
RF6.4 法における質的説明変数にかかる係数	v_{kl}
RF6.4 法における学習パラメータ	ϕ
RF6.4 法における学習パラメータの要素	$c_0, c_{gd}, c_{dkl}, v_{gm}$
RF6.4 法における最適探索幅	α
BIC における最大対数尤度	L
BIC における自由度	k
BIC におけるデータ数	n
重み減衰法における正則化係数	λ
k-means における係数値ベクトル	$\mathbf{w}^n$

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．

また，特定の地域に対して議論を行う際にその地域の特色を正確に把握するためには，その地域が日本全体の中でどう位置づけられるかという視点を用いることが重要であるとされ，統計指標を用いて各地域の特徴を分析している研究もある [5]．

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．

このような背景には様々な課題が考えられるが，その中でも特に以下の 2 つが課題であると考えられる [6]．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．

§ 1.2　本研究の目的

1.1 節では，EBPM における研究動向と課題について述べた．これらに対し，日本では政府によってその推進に向けたアウトラインが作成されるとともに，様々な取り組みがなされている．しかし，政策立案の分野が扱う対象は非常に広範であり，様々な視点からの分析や知見が必要とされる．

そこで，本研究では前述の要素のうち，地域特性の分類と目的・手段間の関係性の 2 点に着目し，統計分析の手法を用いた新たなデータ分析システムを提案することでその解決を目指した．各要素に対する本研究のアプローチの概要を以下に示す．

1. 地域特性を捉える手段として，行政が持つ統計データを用いて，その背景に存在する潜在的要素を分析し，その結果を考慮して自治体をいくつかのクラスターに分ける手法を適用する
2. 目的・手段間の関係性を明確にするための指針の一つとして，データ間に成り立つ関係性を数理モデルとして表す手法を提案する

具体的な手法として，潜在的な要素を考慮した自治体のクラスタリングには潜在プロファイル分析（Latent Profile Analysis: LPA）[7] という手法を用いた．LPA では，実世界において観測可能なデータ（観測変数）がその背後に存在する観測不可能なカテゴリー変数（潜在変数）の影響を受けて顕在化していると仮定する．そのうえで，尤もらしい潜在変数の特徴を観測変数から推測することでデータのクラスタリングを行う．

次に，データ間に成り立つ関係性を数理モデルとして表す手法として，パーセプトロンの学習を用いた多変量多項式回帰の一つである RF 法（Rule extraction method from Fact）[8] を用いた．RF 法では，複数の説明変数と単一の目的変数の間に成り立つ関係を多項式の数理モデルとして表すことができる．また，行政が持つ統計データの特徴をより正確に捉えるために，前述の LPA の結果を考慮して RF 法の精度を向上させる手法について提案する．

また，これら二つの手法における分析結果を表現する手法として，地理情報システム（Geographic Information System: GIS）[9] を用いた Web アプリケーションシステムの開発を行った．ただし，システムの基本的な構造は本研究と同様に EBPM の推進に向けたデータ分析手法の提案を行った先行研究 [10] のものを踏襲し，そのシステムに機能を追加するという形で開発を行った．

最後に，本研究で提案する手法の精度を検証するために，オープンデータサイトを用いて収集した実際の統計データを用いて数値実験を行った．また，その結果について考察を行い，提案手法をより効果的に用いるために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 の推進に向けたデータ分析・可視化システムに関する先行研究を挙げ，その課題について述べる．加えて，本研究の目的である自治体のクラスタリングとデータの関係性のモデル化について，代表的な手法についてその概要を示す．

第 3 章 本研究の提案手法を構築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挙げ，一般的な理論について解説する．

第 4 章 第 3 章で挙げた分析手法を用いて自治体が属する潜在的なクラスターを分析する手法を提案する．また，その結果を考慮した数法則発見法を提案し，統計データの間に成り立つ関係を数理モデルとして表す手法を示す．加えて，これらの結果を GIS で提示するシステムの開発について，システムの概要と画面遷移を説明する．

第 5 章 提案手法における精度の検証を目的として行った数値実験について，その概要と結果を示す．また，結果に対して考察を行い，提案手法において考慮すべき事柄について言及する．

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

人口減少や自然災害をはじめとした様々な課題に直面する日本において，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，EBPM に基づく政策立案が重要視されている．EBPM とは，政策を立案する際にその対象に関するデータを適切に収集・分析し，その結果に基づいて意思決定を行うという考え方である．

しかし，政策における対象は多岐に渡るため，それらすべてに対して効果的な 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・分析する必要がある．この作業は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．

本節の前半では，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，EBPM に対する本研究の立ち位置を再度整理し，その理念を示す．後半では，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の略であるが，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 [11]．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，実際にデータを集め，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なされ始めた．

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 における手法と本研究の理念

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る．ロジックモデルを用いて，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を明確化することで課題に対してより的確にアプローチすることが可能になる．

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 [12].

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し，必要と判断した場合に作成を行っている．

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が，その目的であるアウトカムに対して適切であるかを分析する手法はこれまでに複数提案されている．また，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 [13]．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

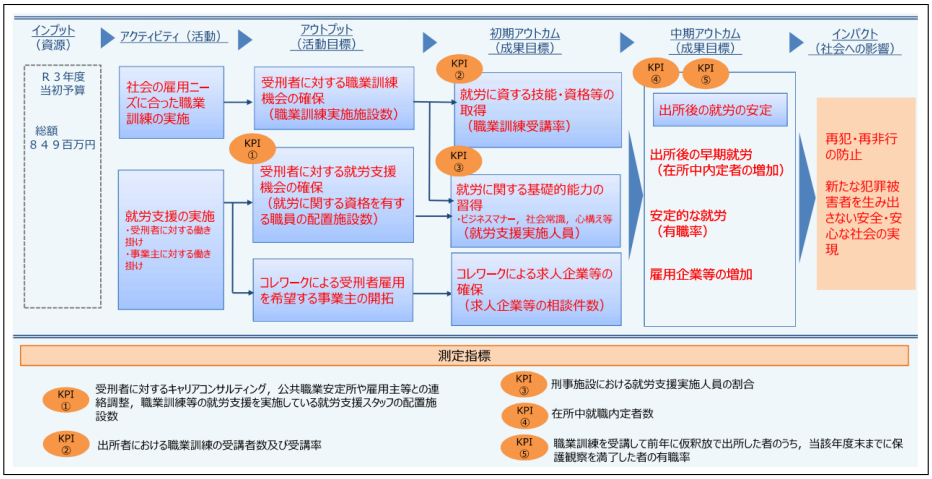

図 2.1: ロジックモデルの例 [12]

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，自然発生した事象を実験の結果のように扱う「自然実験」と呼ばれる手法がとられることもある．

以上のように，日本では政策ごとにロジックモデルを作成することで，政策の立案から施行までの各段階における目標を明確化している．また，目標に対して活動が適切であるかを RCT をはじめとする調査と分析によって示すことで政策の論理的な妥当性を高めている．

しかし，RCT のような手法は条件を整えるために大きなコストがかかるという問題点が挙げられている．また，望ましい結果を得るためには，特定の政策に特化した実験を構築する必要がある．

一方，本研究では様々な政策に対して汎用的に使用でき，判断の材料とすることができる分析結果を提示する手法を目指す．そのためのアプローチとして，データ間の関係性を数理モデルによって表す．観測されているデータがどのように影響しあっているかを把握することは，政策の立案のほか，実験の構築の際にも役立つと考えらえれる．

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 [14]．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

表 2.1: エビデンスレベル [13]

↑ ↑ 質が 高い	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
	レベル2b	重回帰分析、コーホート分析
	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

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 [15]．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 GIS 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 [10]．また，そ

図 2.2: RESAS の例（射水市）[14]

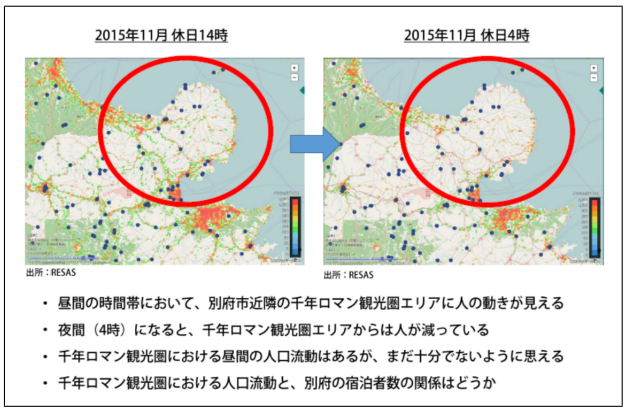

図 2.3: 別府市における RESAS 活用事例 [15]

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を確保すること

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 GIS 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．

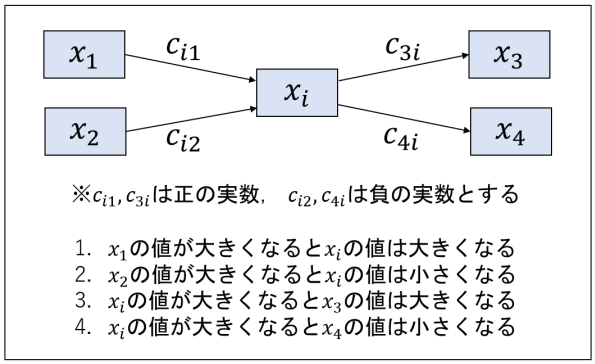

図 2.4: 4 パターンの因果関係

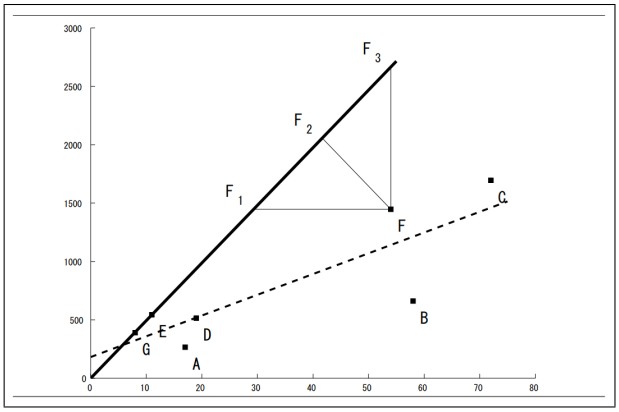

図 2.5: DEA における結果の例 [?]

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）[16] を用いている．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 [16]．

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）
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 における c_{ij} 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 [17] 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）[18] 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，対象の DMU における活動に対する評価値を同一集合内の他の DMU との相対評価によって求める．ここでいう同一集合とは，対象の DMU と同様の活動を行う DMU の集合であり，DMU が企業である場合，同業他社がそれにあたる．

DEA における研究は，今まで様々な研究者によって行われており，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 [19] 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 の評価値を同一集合内の他の DMU との相対評価によって求めるために，集合内の 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．

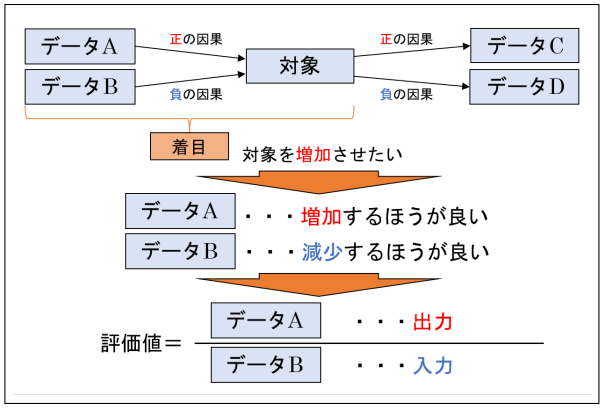

図 2.6: データの振り分け方法

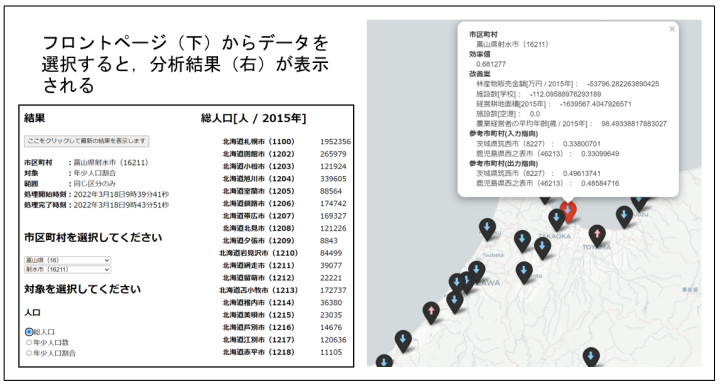

図 2.7: アプリケーションの概要 [10]

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

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

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．

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行政が持つ統計データを分析し，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．

GIS とは，地形情報や建造物の場所など空間的な情報とその他に考慮したい統計量などを同一の地図上にプロットし，重ね合わせや地図の切り替えなどを行うことができる技術である．なお，その他に考慮したい情報の例としては，感染症の拡大状況やリアルタイムな道路の混雑情報などが挙げられる．

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マーカーとして表示され，マーカー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，マーカーを一回クリックすると，その自治体に対する詳細な分析結果が表示される．

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オープンデータを用いてそれらの間に成り立つ因果関係を分析し，分析の対象としたいデータと因果関係のあるデータのみを抽出することで自治体の運営評価値の算出を行っている．しかし，提案手法における評価値の算出では，日本に存在するすべての自治体を DMU 集合として扱っており，それらすべてが重みに対する制約として考慮されているため，結果の精度が低下する可能性が考えられる．

本研究の提案手法の一つである自治体のクラスタリングを用いることで，DMU 集合の絞り込みを行うことが可能である．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　クラスタリングと回帰分析

本節では，1.1 節で挙げた二つの目的について，それぞれに対する手法であるクラスタリングと回帰分析について代表的な手法を示す．また，いくつかの手法を取り上げ，その概要を示す．

表 2.2: 代表的なクラスタリング手法

手法のカテゴリ	分類の種類	手法名
階層型	ハードクラスタリング	群平均法，ウォード法，最短距離法
非階層型	ハードクラスタリング	k-means 法，x-meand 法
	ソフトクラスタリング	ソフト k-means 法，潜在プロファイル分析

クラスタリングの概要

クラスタリングとは，与えられたデータを特定のルールに基づいていくつかのクラスターに振り分ける教師なし学習の手法である．教師なし学習とは，機械学習における手法の一つであり，正解データを与えずにデータセットの中に存在するパターンを発見する手法である．正解データとは，何らかの入力とそれに対する出力が両方含まれたデータである．

機械学習の分野では，いままでに様々なクラスタリング手法が提案されており，その研究は現在も続いている．代表的なクラスタリング手法を表 2.2 に示す．

まず，クラスタリング手法は階層型と非階層型に分けられる．階層型とは，データ全体の中で最も類似度が高いもしくは低いデータ同士を一つずつグループにしていき，すべてのデータがグループに属するまで，それを繰り返す手法である．一方，非階層型ではデータ全体でバランスをとりながらクラスターを作成していく手法である．

次に，分類の種類であるが，ハードクラスタリングとソフトクラスタリングに分けられる．前者はデータ内の各サンプルに対して，必ず一つのカテゴリーが決定される手法である．一方で，後者はサンプルが複数のクラスターに属することを許容しており，各カテゴリーに対する存在確率が求められる．

回帰分析の概要

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．

なお，以降では，回帰分析における変数に対して，目的変数および説明変数という呼称を用いる．目的変数とは，回帰分析によってえられる結果である回帰式において説明される側の変数のことを指す．反対に，説明する側の変数を説明変数とする．

回帰分析に関する研究の歴史は古く，現在までに多くのモデルが考えられている．そのため，適切な結果をえるためには，手段や目的，データの種類などによって手法を使い分けられる必要がある．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 2.3 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 [21].

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 [22]，サポートベクトル回帰 [23] 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 2.3 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．

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 [24]．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

表 2.3: 代表的な回帰モデル

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

- 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．

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 [8].

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 [25].

第3章

潜在クラスタリングと数法則の発見

§ 3.1　LPA によるデータの潜在クラスタリング

実世界に顕在化しており実際に観測することができる変数（観測変数）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ー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）[26] や潜在プロファイル分析（Latent profile analysis: LPA）[7] がある．

一般に，観測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のアプローチではなく，変数の分布に対する最尤推定を用いたモデル指向のアプローチであるという点である．

例えば，クラスタリング手法として代表的な k-means 法では，各サンプルにおけるデータ間の距離を基準にしてデータを複数のクラスターに分割する．一方，LPA では，各変数の分布の重なりを考慮してサンプルが属するクラスターを決定する．そのため，単位やスケールが異なる複数観測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クラスタリングを行うことができるという利点がある．

また，LPA では各サンプルが属するクラスターを完全に一つに決定するのではなく，各クラスターへの所属確率という形で求める．そのため，前者がハードクラスタリングと呼称されるのに対して，LPA のような手法はソフトクラスタリングと呼ばれる．

本節では，LPA において用いられる混合分布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がどの程度の確率で所属しているかを求める方法について順に示す．

モデルの概要

混合分布モデルとは，いくつかの分布の重ね合わせによって，より複雑な分布を表現する分布モデルである．混合分布モデルを説明するために，まず観測変数が一つの場合について考える．連続変数 y_i をサイズ $N(i = 1, \ldots, N)$ のサンプルの i 番目の要素における観測変数とする．また，本来は未知であるが，仮にサンプルの母集団が 2 つの異なるクラスター $(K = 2)$ によって構成されているとする．

$K = 1$ のとき平均が μ_1，分散が σ_1^2，$K = 2$ のとき平均が μ_2，分散が σ_2^2 のように，各クラスターに属するデータがそれぞれ異なる分布に基づくとすると，母集団の分布は二つの分布の混合分布によって表される．このとき，母集団に置ける各クラスターの混合割合が

表 3.1: 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	$\begin{bmatrix} \sigma_{1k}^2 & & & \\ \sigma_{21k} & \sigma_{2k}^2 & & \\ \vdots & \vdots & \ddots & \\ \sigma_{r1k} & \sigma_{r2k} & \cdots & \sigma_{rk}^2 \end{bmatrix}$

$\pi_1 : \pi_2$ だとすると，この例のモデルは各クラスターの分布 f_1, f_2 を用いて以下のように表すことができる．

$$f(y_i|\mathbf{\Theta}) = \pi_1 f_1(y_i|\mu_1, \sigma_1^2) + \pi_2 f_2(y_i|\mu_2, \sigma_2^2) \tag{3.1}$$

ただし，$\Theta = \{\pi_1, \mu_1, \sigma_1^2, \pi_2, \mu_2, \sigma_2^2\}$ はモデルにおけるパラメータベクトルである．また，混合モデルにおける各分布の重み π_1, π_2 は非負であり，その和は 1 となる．つまり，式 3.1 における重みが $\pi_1 = 0.6, \pi_2 = 0.4$ とした場合，母集団の 60% はクラスター 1 で，母集団の 40% はクラスター 2 で説明できることになる．また，式 3.1 では二つの分布におけるそれぞれの平均，分散を別の値として表記しているが，平均または分散を共通とするような制約をかけるなどすることで，様々なデータに対して適合することができる．

次に，観測変数が r 個の多変量分布について考える．この多変量分布が K 個のクラスターから構成されるとすると，式 3.1 は各クラスター k の平均ベクトル μ_k と共分散行列 $\mathbf{\Sigma}_k$，母集団における各クラスターの構成割合 π_k を用いて以下のように表すことができる．

$$f(\mathbf{y}_i|\mathbf{\Theta}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i|\mu_k, \mathbf{\Sigma}_k) \tag{3.2}$$

式 3.1 において，各クラスターの平均や分散を変化させることでモデルを様々なデータに適合させることができることについて言及したが，式 3.2 においても同様である．特に，共分散を考慮した場合，より多くのモデルを考えることができる．r 種類の観測変数が観測されており，クラスター数が k 個の場合における共分散行列 Σ_k のパラメータの例を五つ表 3.1 に示す．

モデル A では，各クラスター内における観測変数間での分散は異なる一方，クラスター間においてはすべての分散が共通である．また，観測変数間の共分散を考慮しない．これは表 3.1 の中で最も単純なモデルである．このモデルに対して，クラスター間の分散においても異なることを考慮したモデルがモデル C にあたる．このモデルにおいても観測変数間の共分散は考慮しない．

一方で，モデル A に対して観測変数間の共分散を考慮するようにしたモデルがモデル B である．このモデルでは，クラスター間の分散，共分散に対しては共通であるという制約を設けている．また，これに対してクラスター間の共分散は共通であるという制約を残しつつ，クラスター間の分散が共通であるという制約のみをなくしたモデルがモデル D になる．

最後に，モデル E は表 3.1 の中で最も複雑なモデルであり，このモデルは各クラスター内における観測変数間とクラスター間いずれにおいても分散，共分散が異なることを許容している．以上が，LPA において用いられる混合分布モデルであるが，LPA では観測変数の値に基づいてパラメータを推定し，モデルを決定するために，以下の手法を用いている．

モデル決定の手法

LPA では，モデルにおけるパラメータを推定するための手法として，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，その値を最大値を求める EM アルゴリズムという手法を用いている [27]．ただ，z は潜在変数であるため，現実には観測することができない．

そのため，EM アルゴリズムでは求めたい観測変数 $\mathbf{y}$ と潜在変数 z の同時確率を以下のような目的関数を置き換えることによってその値を求める．まず，本来求めたい同時確率を以下に示す．

$$\begin{aligned} p(\mathbf{y}, z|\pi, \mu, \mathbf{\Sigma}) &= p(z|\pi, \mu, \mathbf{\Sigma}) p(\mathbf{y}|z, \pi, \mu, \mathbf{\Sigma}) \\ &= \prod_i \prod_k [\pi_k f(y_i|\mu_k, \mathbf{\Sigma}_k)]^{z_{ik}} \end{aligned} \tag{3.3}$$

次に，両辺の対数をとることによって，式 3.3 以下のように表すことができる．

$$\ln p(\mathbf{y}, z|\pi, \mu, \mathbf{\Sigma}) = \sum_i \sum_k z_{ik} (\ln \pi_k + \ln f(y_i|\mu_k, \mathbf{\Sigma}_k)) \tag{3.4}$$

さらに，前述のとおり潜在変数 $\mathbf{z}$ の値は直接与えられていないため，$\mathbf{z}$ の事後分布の期待値を考えると以下のように表すことができる．

$$E_z[\ln p(\mathbf{y}, z|\pi, \mu, \mathbf{\Sigma})] = \sum_i \sum_k E_{z_{ik}}[z_{ik}] (\ln \pi_k + \ln f(y_i|\mu_k, \mathbf{\Sigma}_k)) \tag{3.5}$$

EM アルゴリズムでは，この式 3.5 を目的関数として，これを最大化するパラメータ $\pi_k, \mu_k, \mathbf{\Sigma}_k$ を推定する．また，$z_{ik} \in \{0, 1\}$ であるので，式 3.5 の右辺における $E_{z_{ik}}[z_{ik}]$ は各パラメータを用いて以下のように求めることができる．

$$
\begin{aligned}
E_{z_{ik}}[z_{ik}] &= \sum_{z_{ik}=0,1} z_{ik} p(z_{ik}|y_i, \pi_k, \mu_k, \mathbf{\Sigma}_k) = 1 \times p(z_{ik}=1|y_i, \pi_k, \mu_k, \mathbf{\Sigma}_k) \\
&= \frac{p(z_{ik}=1)p(y_i|z_{ik}=1)}{\sum_k p(z_{ik}=1)p(y_i|z_{ik}=1)} \\
&= \frac{\pi_k f(\mathbf{y}|\mu_k, \mathbf{\Sigma}_k)}{\sum_k \pi_k f(\mathbf{y}|\mu_k, \mathbf{\Sigma}_k)}
\end{aligned}
\tag{3.6}
$$

EM アルゴリズムでは，式 3.5 を最大化するパラメータを推定するために，以下に示す期待値（Expectation: E）ステップと最大化（Maximization: M）ステップを繰り返す．

- **E ステップ**
  その時点で最適と考えれるパラメータを式 3.6 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 3.5 の右辺に代入し，式 3.5 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．

これら二つのステップを繰り返し行い，潜在変数の期待値 z_{ik} と各パラメータ π_k, μ_k, Σ_k を更新していくことで，最終的に目的関数を最大化する尤もらしいパラメータの値を求めることができる．

また，実世界で観測された値からモデルを決定する場合，潜在変数の数 k の値は未知数である．そのため，いくつかの指標を用いて適切な k の値を判断する．ここでは，その指標であるベイズ情報量基準（Bayesian information criterion: BIC）を示す．BIC の基本形は以下のように定義される [28].

$$
BIC = -2L + k \ln n \tag{3.7}
$$

ただし，L は最大対数尤度，k はパラメータ数，n はデータ数である．BIC は，パラメータ数にペナルティを設けることでモデルの自由度に対して制限をかける指標である．

各サンプルにおけるクラスターへの存在確率の算出

EM アルゴリズムを用いてクラス数 k ごとのモデルを求め，BIC に基づいて最適なモデルが選択された後，LPA では各サンプルにおけるそれぞれのカテゴリーへの存在確率が算出される．算出方法は各クラスター数に対する事後確率で求められ，以下の式で表される．

$$
\pi_{k|y_i} = \frac{\pi_k f_k(\mathbf{y_i}|\mu_k, \mathbf{\Sigma}_k)}{\sum_k \pi_k f_k(\mathbf{y}_i|\mu_k, \mathbf{\Sigma}_k)} \tag{3.8}
$$

§ 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一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．

$$if \bigwedge_{k} \bigvee_{q_{kl} \in Q_k^i} q_{kl} \quad then \quad y = h(\mathbf{x}; \theta^i) + \epsilon, \quad i = 1, \cdots, I \tag{3.9}$$

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 3.9 における $h(\mathbf{x}; \theta)$ は以下のように表せる．

$$h(\mathbf{x}; \theta) = v_0 + \sum_{g=1}^{G} v_g \prod_{m=1}^{M} x_m^{v_{gm}} \tag{3.10}$$

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 3.9 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．

$$a(\mathbf{q}; \mathbf{v}) = \sigma\left(\sum_{k=1}^{K} \sum_{l=1}^{L_k} v_{kl} q_{kl}\right) \tag{3.11}$$

$$F(\mathbf{q}, \mathbf{x}; \mathbf{v^1}, \cdots, \mathbf{v^I}, \theta^{\mathbf{1}}, \cdots, \theta^{\mathbf{I}}) = \sum_{i=1}^{I} a(\mathbf{q}; \mathbf{v}) h(\mathbf{x}; \theta^i) \tag{3.12}$$

よって，式 3.10 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

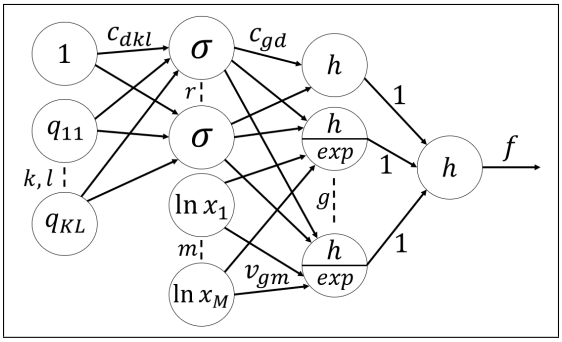

図 3.1: RF6.4 法の 4 層パーセプトロン

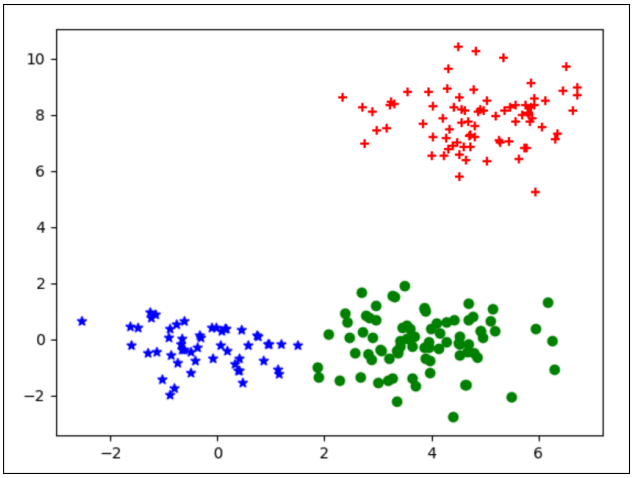

図 3.2: 二次元の k-means

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．

$$
\begin{aligned}
f(\mathbf{q},\mathbf{x};\phi) &= w_0 + \sum_{g=1}^{G} w_g s_g, \\
w_0 = \sum_{d=1}^{D} c_{0d}\sigma_d, \quad w_g &= \sum_{d=1}^{D} c_{gd}\sigma_d, \\
\sigma_d = \sigma\left(\sum_{k=1}^{K}\sum_{l=1}^{L_k} c_{dkl} q_{kl}\right), \quad s_g &= \exp\left(\sum_{m=1}^{M} v_{gm}\ln x_m\right)
\end{aligned}
\tag{3.13}
$$

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 3.13 が得らえれる．

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．

$$
E(\mathbf{q},\mathbf{x};\phi) = \frac{1}{2}\sum_{n=1}^{N}(f(\mathbf{q},\mathbf{x};\phi)^n - y^n)^2 \tag{3.14}
$$

多層パーセプトロンの学習法

式 3.9 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 法では二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 [29]．

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．

$$\nabla E = \frac{\partial E}{\partial \phi} = \sum_{n=1}^{N} (f^n - y^n) \frac{\partial f^n}{\partial \phi} \tag{3.15}$$

$$\frac{\partial f^n}{\partial c_{gd}} = \sigma_d^n s_g^n \tag{3.16}$$

$$\frac{\partial f^n}{\partial c_{dkl}} = \sigma_d^n (1 - \sigma_d^n) q_{kl}^n \left(\sum_{g=0}^{G} c_{gd} s_g^n \right) \tag{3.17}$$

$$\frac{\partial f^n}{\partial v_{gm}} = w_g^n s_g^n x_m^n \tag{3.18}$$

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

$\boldsymbol{\phi}_{\mathbf{b+1}} = \boldsymbol{\phi}_{\mathbf{b}} + \alpha_{\mathbf{s}} \boldsymbol{\Delta} \boldsymbol{\phi}_{\mathbf{b}}$ 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 [30]．また，$b \leftarrow b + 1$ として，Step2 の戻る．

$$\begin{aligned} \mathbf{H}_{b+1} &= \mathbf{H}_b + \left(1 + \frac{\mathbf{q}^T \mathbf{H}_b \mathbf{q}}{\mathbf{p}^T \mathbf{q}} \right) \frac{\mathbf{p}\mathbf{p}^T}{\mathbf{p}^T \mathbf{q}} - \frac{\mathbf{p}\mathbf{q}^T \mathbf{H}_b + \mathbf{H}_b \mathbf{q}\mathbf{p}^T}{\mathbf{p}^T \mathbf{q}} \\ \mathbf{p} &= \alpha_b \Delta\phi_b \\ \mathbf{q} &= \nabla E(\mathbf{q}, \mathbf{x}; \phi_{b+1}) - \nabla E(\mathbf{q}, \mathbf{x}; \phi_b) \end{aligned} \tag{3.19}$$

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．

$$g(\alpha) \approx g(0) + g'(0)\alpha + \frac{1}{2}g''(0)\alpha^2 \tag{3.20}$$

$$g'(0) = \sum_{n=1}^{N}(f^n - y^n)f'^n \tag{3.21}$$

$$g''(0) = \sum_{n=1}^{N}((f^n)^2 + (f^n - y^n)f''^n) \tag{3.22}$$

$$f'^n = \sum_{g=0}^{G}(w_g'^n s_g^n + w_g^n s_g'^n) \tag{3.23}$$

$$f''^n = \sum_{g=0}^{G}(w_g''^n s_g^n + 2w_g'^n s_g'^n + w_g^n s_g''^n) \tag{3.24}$$

$$w_g'^n = \sum_{d=1}^{D}(\Delta w_{gd}\sigma_d^n + w_{gd}\sigma_d'^n) \tag{3.25}$$

$$w_g''^n = \sum_{d=1}^{D}(2\Delta w_{gd}\sigma_d'^n + w_{gd}\sigma_d''^n) \tag{3.26}$$

$$\sigma_g'^n = \sigma_g^n(1 - \sigma_g^n)\left(\sum_{k=1}^{K}\sum_{l=1}^{L}\delta w_{drk}q_{kl}^n\right) \tag{3.27}$$

$$\sigma_g''^n = \sigma_g^n(1 - \sigma_g^n)(1 - 2\sigma_g^n)\left(\sum_{k=1}^{K}\sum_{l=1}^{L}\delta w_{drk}q_{kl}^n\right)^2 \tag{3.28}$$

$$s_g'^n = s_g^n\sum_{m=1}^{M}\Delta v_{gm}\ln x_m^n \tag{3.29}$$

$$s_g''^n = s_g^n\left(\sum_{m=1}^{M}\Delta v_{gm}\ln x_m^n\right)^2 \tag{3.30}$$

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．

$$\alpha = \begin{cases} -\frac{g'(0)}{g''(0)} & (g''(0) > 0) \\ -\frac{g'(0)}{\sum_{n=1}^{N}(f'^n(\mathbf{x}^n;\phi))^2} & (g''(0) \leq 0) \end{cases} \tag{3.31}$$

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．

$$\tilde{\alpha} = -\frac{g'(0)\alpha^2}{2(g(\alpha) - g(0) - g'(0)\alpha)} \tag{3.32}$$

§ 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 3.13 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 3.13 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 3.1 節の式 3.15 でも示した 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．

RF 法では，学習によって得らえられたパラメータを用いて算出した $f(\mathbf{q}^n, f(\mathbf{x}^n; \hat{\phi}_{G,D})$ と観測値である y の値との誤差がガウス分布に基づくと仮定しているため，式 3.15 は残差平方和 (RSS) とサンプル数 n，パラメータ数 k を用いて以下のように表すことができる．

$$BIC = n \ln \left(\frac{RSS}{n} \right) + k \ln n \tag{3.33}$$

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．

$$BIC(G, D) = \frac{N}{2} \ln \left(\frac{1}{N} \sum_{n=1}^{N} (f(\mathbf{q}^n, f(\mathbf{x}^n; \hat{\phi}_{G,D}) - y^n)^2 \right) + \frac{Z}{2} \ln N \tag{3.34}$$

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．

よって，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．

正則化の方法の一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 $e(\mathbf{q}, \mathbf{x}; \phi)$ として以下に示す．

$$e(\mathbf{q}, \mathbf{x}; \phi) = E(\mathbf{q}, \mathbf{x}; \phi) + \lambda \left(\frac{1}{2} \sum \phi^2 \right) \tag{3.35}$$

ここで，λ は正則化係数と呼ばれ，非負の値である．λ の値を大きくすると，目的関数に対する正則化項の割合が大きくなり，各パラメータは 0 に近くなる．逆に，λ の値を小さくすると正則化項が効かなくなる．そのため，適度な大きさの λ において汎化性能の向上がみられる．

また，重み減衰法ではすべてのパラメータに均等な正則化を行うが，パラメータの中には重要なものと不要なものが存在することが一般的である．そのため，不要なパラメータにより大きい正則化係数をかけ，重要なパラメータにかかる正則化係数を小さくして自由度を与えることが望ましい．そこで，各重みごとに異なる正則化係数を用いる方法が考えられる．その場合，目的関数は以下のようになる．

$$\epsilon(\phi) = E(\phi) + \frac{1}{2} \phi^{\mathbf{T}} \mathbf{\Lambda} \phi \tag{3.36}$$

ただし，Λ はパラメータベクトル ϕ の要素数 M と同じ次元の対角行列であり，以下のように表せる．なお，対角要素に $\exp(\lambda)$ を用いているのは係数を常に正にするためである．

$$\Lambda = \begin{pmatrix} \exp(\lambda_1) & & 0 \\ & \ddots & \\ 0 & & \exp(\lambda_1) \end{pmatrix} \tag{3.37}$$

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 3.9 の形で回帰ルール集合を復元する必要がある．また，式 3.9 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 3.13 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．

$$w_g^n = \sum_{d=1}^{D} \hat{c}_{gd} \hat{\sigma}_d^n, \quad \hat{\sigma}_d^n = \sigma \left(\sum_{k=1}^{K} \sum_{l=1}^{L_k} \hat{c}_{dkl} q_{kl}^n \right) \tag{3.38}$$

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 [31]．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 法の結果を図 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，それらを三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．

$$dis = \sum_{i=1}^{I} \sum_{n \in R_i}^{N_i} \|\mathbf{w}^n - \mathbf{u}^i\|^2, \quad \mathbf{u}^i = \frac{1}{N_i} \sum_{n \in R_i}^{N_i} \mathbf{w}^n \tag{3.39}$$

式 3.39 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 3.40 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 3.41 のようになる．

$$if \quad i(\mathbf{q}) = i \quad then \quad \hat{f} = u_0^i + \sum_{g=1}^{G} u_g^i \prod_{m=1}^{M} x_m^{\hat{v}_g m}, \quad i = 1, \cdots, I \tag{3.40}$$

$$i(\mathbf{q}) = \arg \min_i \|\mathbf{w}^n - \mathbf{u}^i\|^2 \tag{3.41}$$

次に，N 個のサンプルをいくつの代表点 $\{\mathbf{u}^\mathbf{i} : i = 1, \cdots, I\}$ 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 [32]．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 による市町村のクラスタリングと可視化

本節では，日本全国の市区町村について，それらの調査によって観測された統計データに基づいて分析を行うことでいくつかのクラスターに分類する手法について提案する．また，クラスタリングの結果を可視化することの利点について述べ，その手法を提案する．

クラスタリング手法

2.3 節で述べたとおり，観測されたいくつかの量的変数を用いてサンプルを分類し，適切なクラスターに振り分ける手法は現在までに複数提案されている．また，それらの多くは各サンプルにおけるデータ間の類似度を距離などを用いて算出し，類似度の高いデータを同じクラスターとして扱うものである．

このような手法を用いて自治体をクラスタリングすることも可能であるが，本研究の提案手法では，以下に挙げる 3 つの要素を考慮し，クラスタリング手法に 3.1 節で述べた LPA を用いる．

1. データの背景にある潜在的な特徴を捉えること
2. スケールが異なるデータを直接扱えること
3. モデルの評価に基づいてクラスター数を決定できること

一つ目の要素は，EBPM の推進に向けて新しい視点で自治体を分類する手法を提案するという本研究の目的のためである．日本における自治体は，その規模や主要産業など観測できる要素だけをとっても非常に多種多様である．そこで，本研究ではその背後に潜在的な共通要素が存在するという仮定に基づいて自治体を分類する．

二つ目および三つ目の要素は，社会から顕在するデータの特徴から必要と考えた．社会において観測されるデータは単位やスケールが多様である．また，複数種類のデータを用いてクラスタリングを行った場合，経験やデータの傾向に基づいてクラスター数を判断するのは不可能である．よって，単位やスケールを考慮する必要がなく，モデルの観点から妥当なクラス数を決定できるクラスタリング手法を用いることが効果的であると考えた．

次に，社会において観測されたデータをどのような仮定に基づいてクラスタリングするかであるが，本研究では各データが複数の正規分布の重ね合わせによって表されているという仮定を置いて LPA によるクラスタリングを行う．つまり，3.1 節で述べた混合分布モデルにおいて，混合ガウス分布を仮定する．

近年の研究では，現実世界で起こる事象のいくつかはべき乗分布などの分布に従うということが指摘されている [33]．しかし，人口においては都市の規模によってその分布に検討の余地があることも示されている [34]．また，すべての事象においてそれを表す分布を把握することは不可能である．混合ガウス分布は，構成するガウス分布の数やパラメータによって様々なデータに対応することができると考えられる．そのため，本研究ではすべてのデータは混合ガウス分布によって表せるという仮定で分析を行う．

行政が持つデータをどのように用いて具体的なモデルの推定を行うかを以降で示す．まず，行政が持つ複数種類の統計データをそれぞれ観測変数として考え，各変数が混合ガウス分布に基づいているという仮定のもと，3.1 節で述べた EM アルゴリズムを用いて，データに最も適合するモデルのパラメータを求める．なお，一つの自治体をを一つのサンプルとし，共分散行列には 3.1 節の表 3.1 に示した五つの行列をそれぞれ用いる．

これを潜在変数の数，つまり，データを分けるカテゴリーの数を変化させながら複数回行う．結果として，共分散行列とカテゴリー数ごとに最適なパラメータベクトルが得られる．次に，それらのパラメータを用いて式 3.7 の BIC を算出し，BIC の値が最も望ましいものを適切なモデルとして採用する．最後に，そのモデルを用いて各自治体の存在確率を算出する．

クラスタの可視化

前述のクラスタリング手法における結果は，各市町村がどのクラスターにどれくらいの確率で所属するかの確率によって求められる．また，日本全国の市町村を対象として分析を行った場合，サンプル数は 1700 以上となる．そのため，結果を生データから確認することも可能であるが，そのままでは情報がうまくまとまっておらず，解釈にはある程度の時間が必要になる．

クラスタリングの結果をどのような形式で表示することが適切であるかは，その後の目的によって異なるが，本研究では結果を感覚的に捉え，大まかな特徴を確認するために GIS を用いて各市町村が存在するクラスターを描画する手法を提案する．また，GIS を用いた描画では，クラスターの様子が地図上にプロットされるため，地理的な要因を含めた議論が可能になるという利点もあると考えられる．GIS を用いたクラスターの描画方法について以下で述べる．

まず，各自治体に対してそれぞれクラスタの所属確率という複数の連続値で与えられるクラスタリングの結果を地図上にどのような形でプロットするかであるが，本研究では単純に各市区町村における最も所属確率が高いクラスターを表示する方法を用いる．LPA では，各市町村が異なる複数のクラスターに属することを認めているが，多くの場合でどれか一つのクラスターに高い所属率を示すため，本研究ではそのように扱う．

次に，実装についてであるが，本研究では「folium」という Python のライブラリを使用している [35]．folium を用いると，いくつかの関数を使用することで GIS に機能を追加することができ，実行結果として作成された GIS を含む html ファイルが返される．folium によって追加できる機能には主に以下のものが挙げられる．また，機能の例を図 4.1 に示す．

- 指定した緯度・経度にマーカーを配置する
- 指定した範囲を中心にヒートマップを描画する

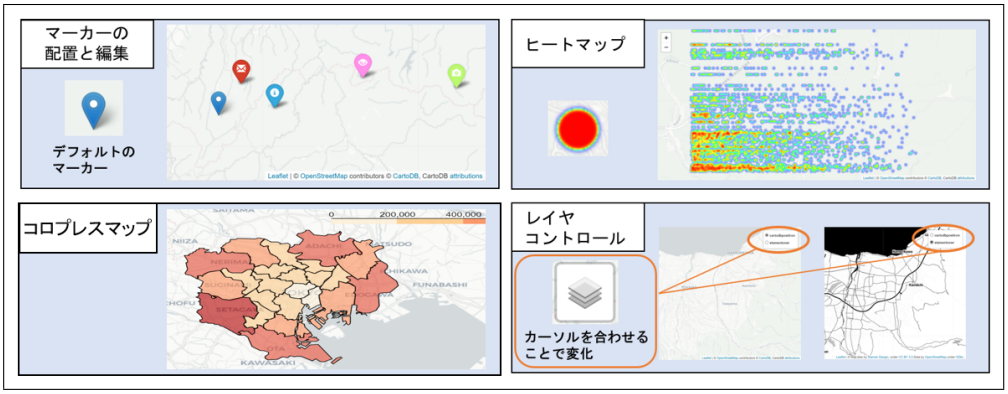

図 4.1: folium ので追加できる機能の例

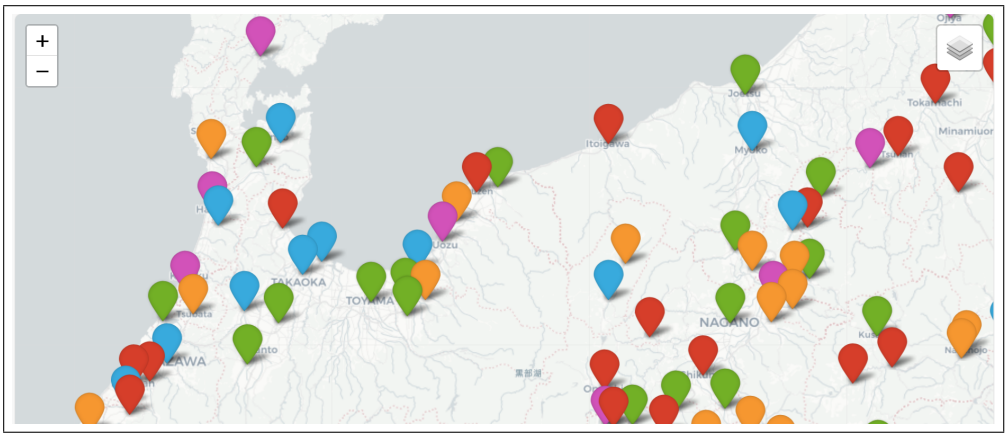

図 4.2: GIS を用いた潜在的クラスターの描画

- 複数の緯度・経度を用いてコロプレスマップを描画する
- レイヤーを複数重ね合わせる

本研究の提案手法の場合，各市町村とそれに紐づくクラスターを GIS で可視化する必要がある．そのため，図 4.1 におけるマーカーまたはコロプレスマップを用いて各市町村の位置を示し，それを用いて各市町村が属するクラスターを表示する方法が適していると考えられる．

機能面ではどちらの方法を用いても実現できるが，コロプレスマップの作成には地図上に境界線を描画するために膨大な量の緯度・経度データが必要となる．そのため，今回のように市町村単位で細かく地図を細分する際には作成，表示ともに機器への負担が大きくなり不向きである．以上のことから，本研究ではマーカーを用いて市町村の位置を表示し，属するクラスターによってマーカーの色に変化をつけるという方法で GIS の実装を行った．

本研究における GIS を用いた潜在的なクラスターの描画の実装例を図 4.2 に示す．この例では，市区町村が五つの潜在的なクラスターに分かれるという結果が得られた場合を描画しており，五つの色を用いてそれぞれのクラスターを表している．マーカーが配置されている位置は各市町村における代表地点である．

§ 4.2　潜在クラスタリングとRF6.4法を用いた数法則発見

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．

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一つが他の複数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．

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．一方，本研究において対象とする統計データ間に成り立つ関係性のモデル化には，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

各条件に対して必要と考えた理由を以下に示す．1について，社会における事象は複雑であり，前提知識を用いてモデルの概形を定義してからモデルのパラメータを求める手法の適用は現実的ではないと考えたからである．2について，各説明変数がどのように目的変数に寄与しているかを高い自由度で表現できることが好ましいと考えたからである．3について，結果の解釈が難解な手法は，関係性を把握するという当初の目的にそぐわないからである．

以上のことから，本研究では求めるモデルの項数を事前に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．空間軸でのデータ収集の例を以下に示す．

単一の年の統計データにを用いて分析を行うことを考える．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，一つの自治体におけるAからDの項目におけるデータは，それぞれ一つずつしか存在しない．そのため，回帰分析のように大量のサンプルが必要な場合，自治体の数を増やす

という方法でしか十分な数の学習データを確保できないということになる．本研究における学習データの構造を図 4.3 に示す．

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すべての自治体に対するデータを一つの数式で無理やり記述したものになる．このような手法では，社会の実情を正しくとらえることが難しいと考える．

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，各サンプルをクラスタリングし，その結果を説明変数に含めて RF 法を行うことで，より現実に即した数法則を発見する手法を提案する．

学習データに対して，それらをクラスタリングする手法だが，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，それぞれのクラスターにおける数法則を求める．

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 によって求めた各クラスター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，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，存在確率を質的説明変数として扱うと述べたが，存在確率は 0 以上 1 未満の連続した値で与えられる．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 によって得られたいくつかのクラスター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ター数が仮に A から E の 5 つであった場合，所属するクラスターが最もあいまいなサンプル（五つのクラスター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 4.4 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，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と考えた.

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.

最後に，提案手法におけるパーセプトロンの学習について，パラメータの初期値と各変数のスケーリングの方法を述べる．まず，初期値についてはすべてのパラメータに対して -1 〜1 の一様乱数を与えることとする．次に，各変数のスケーリングについてだが，本研究では実世界において観測された統計データを用いて分析を行うため，単位やスケールが大幅に異なる．そのため，データを何らかの形でスケーリングする必要があると考えられる.

本研究では，目的変数，質的説明変数のそれぞれに対して異なるスケーリングの方法を用いる．目的変数に対するスケーリングを式 4.1，量的説明変数に対するスケーリングを式 4.2 に示す．なお，スケーリング後の目的変数，量的説明変数を $\tilde{y}, \tilde{x}$，元の目的変数，量的説明変数を y, x，サンプルにおける目的変数全体の中央値，標準偏差，量的説明変数の最大値を $mean(y), std(y), max(x)$ とする.

$$\tilde{y} = \frac{y - mean(y)}{std(y)} \tag{4.1}$$

$$\tilde{x} = \frac{x}{max(x)} \tag{4.2}$$

§ 4.3 Web-GIS 描画による潜在的な法則の可視化

4.1 節では，行政が持つ統計データを用いて，自治体を潜在的な特徴に基づいてクラスタリングする手法とその可視化について述べた．また，4.2 節では，クラスタリングの結果を考慮した多変量多項式回帰を行うことで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について提案した.

本研究では，以上の結果を統合して GIS を用いた可視化を行う Web アプリケーションを開発した．本節では，分析に用いたデータと提案手法について再度整理し，開発したアプリケーション全体の構成を示す．また，以下では本研究におけるシステムを本システムと呼ぶ.

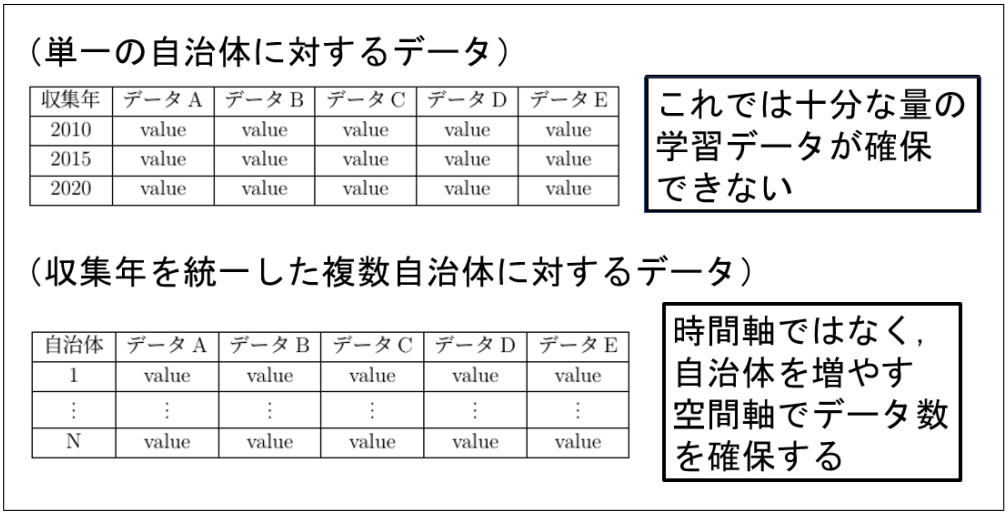

図 4.3: 学習データの一例

表 4.1: 本システムのデータベース

データ項目	単位	データ項目	単位
1 人あたりの固定資産税	円	総人口	人
1 人あたりの地方税	円	住宅用地平均取引価格	円 $/m^2$
1 人当たりの法人住民税	円	商業用地平均取引価格	円 $/m^2$
経営耕地面積	1 畝 / 経営体	農地平均取引価格	円 $/m^2$
製造品出荷額	円	林地平均取引価格	円 $/m^2$
年間商品販売額	円		

まず，使用したデータの参照元について，本システムでは 2.1 節でも述べた RESAS から API を使用してデータを取得している．RESAS は各自治体単位の経済に関するデータを公開しているオープンデータサイトであり，人口などの基本的な項目のほかに各産業における収入などを公開している．RESAS から自動的に取得し，本システムに使用しているデータの一覧を表 4.1 に示す．なお，表 4.1 の各データ項目について本システムでは 2010, 2015, 2020 年のデータを扱っている．次に，本システムにおける基本的な処理の遷移を説明する．

Step 1: 分析方法と対象年の指定

本システムは 2.2 節で述べたシステムと同一の枠組みによって作成されているため，システムにアクセスした際に表示されるフロントページでは，どの分析を行うか選択することになる．そのページで LPA＋RF を選択すると，分析を行うデータと試行回数に関する条件を選択するページに遷移する．

分析を行いたい年と最終的に求める数理モデルにおける目的変数，RF の試行回数を選択し，その下に表示されている実行ボタンを選択すると分析待機画面に遷移する．ユーザに表示されている画面の遷移は分析が終了して結果が得られるまでその画面で待機することになる．システム側では，指定された年におけるデータを用いて Step2 以降の分析が開始される．ここまでの画面的な遷移を図 4.5 に示す．

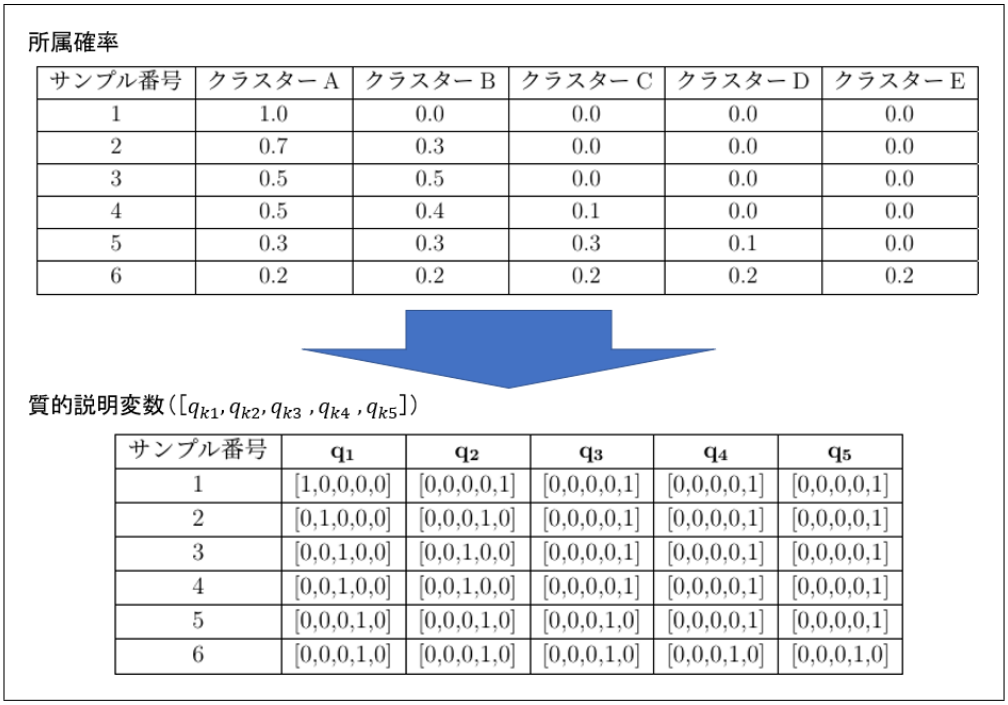

図 4.4: 存在確率のカテゴリー化

Step 2: LPA を用いた自治体のクラスタリング

Step2 では，Step1 で選択された年のデータに基づいて LPA を用いた自治体のクラスタリングを行う．まず，3.1 節の表 3.1 における 5 つの共分散行列を仮定してそれぞれに適したモデルを求める．なお，クラスター数の数は 1〜7 個を考慮している．結果として，共分散行列の数 × クラスター数の結果である計 35 個のモデルにおける最適なパラメータが算出される．

次に，これらに対して BIC を算出し，最も良好な値を示したモデルをクラスタリングに用いるモデルとして採用する．最後に，採用したモデルを用いて各サンプルにおけるクラスターへの存在確率を求め，その結果を用いて Step3 の分析に進む．

Step 3: PF6.4 を用いたデータ間の関係性のモデル化

Step3 では，Step2 で得られた各自治体に対するクラスタリングの結果を考慮して，RF6.4 法を用いたデータ間の関係性のモデル化を行う．求める数理モデルの目的変数には，Step1 で指定されたデータ項目が用いられる．

量的説明変数には本システムのデータベースにある目的変数以外のすべての項目が使用される．質的説明変数には Step2 で得られたクラスタリングの結果がカテゴラルな形で使用される．なお，目的変数および量的説明変数に関しては，4.2 節の式 4.1 および式 4.2 に基づく正規化を施した値が使用される．

RF6.4 法におけるパーセプトロンの学習に用いる各パラメータの初期値は，-1 以上 1 未満の一様乱数を使用し，ユニット数 G および D の数をそれぞれ 1〜5 まで変化させながらそれぞれに対するパラメータの最適化を行う．この処理をパラメータの初期値を変化させながら Step1 で選択された回数試行する．最後に，すべての結果の中で

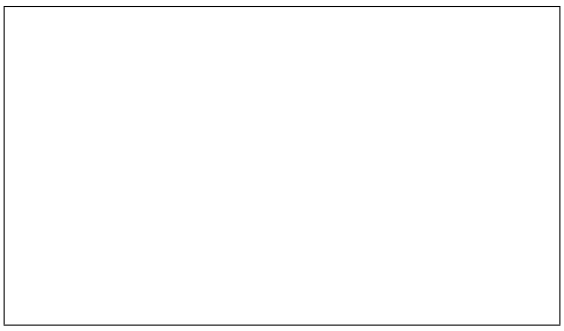

図 4.5: 本システムの画面遷移

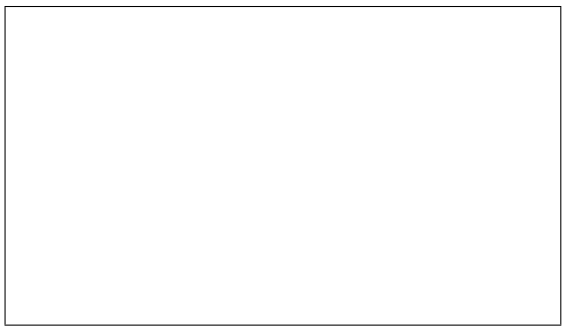

図 4.6: 実装した GIS

BIC が最も優れているパラメータベクトルの値を用いて数法則を復元し，データ間の関係性を表す最適な数理モデルとして採用する．

Step 4: EBPM-GIS の作成・データの重ね合わせ

Step4 では，Step2 および 3 で得られた結果を用いて GIS を作成し，結果を可視化して画面に表示する．まず，LPA を用いた自治体のクラスタリング結果について，4.1 節で述べたとおり，各自治体の所在地にマーカーをプロットし，その色を各クラスターに対応したものに塗分けるという方法で表現する．

次に，RF6.4 法の結果である数理モデルについて，それぞれの自治体とそれに対する潜在的クラスター，数理モデルはそれぞれ一対一で対応しているため，LPA の結果を表示する際にプロットしたマーカーに数理モデルを紐づけるという形で表示を行う．

具体的には，各マーカーに対してポップアップ機能を付与し，そのポップアップの中に自治体の名称と数理モデルを表示するという方法を用いる．GIS 上に表示されたマーカーのいずれかを選択すると，ポップアップが表示されるというような仕組みとなる．実装した GIS を図 4.6 に示す．

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考察メモ：今回はLPAのモデル選択にBICのみを用いている，外れ値を除去せずにLPAを行っている

§ 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]

[6] 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/.

[7] 生徒の学習

[8] コネクショニスト

[9] 国土交通省国土地理院, ”GIS とは”, 閲覧日 2024-02-06, https://www.gsi.go.jp/GIS/whatisgis.html.

[10]

[11]

[12]

[13]

[14]

[15]

[16] 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.

[17] 論文探す

[18] 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.

[19] 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.

[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]

[22]

[23]

[24] 参考論文ファイルから

[25]

[26] COVIT-19

[27] 論文探す

[28]

[29]

[30] Slack 参照

[31] Slack

[32]

[33] 論文探す

[34]

[35] 卒論から

[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] 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.

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

[46] 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.

[47] 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.

[48] 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.