卒業論文

スパーズ推定を用いた変数選択と
ヘドニック・アプローチによる
不動産価格形成要因の分析

Analysis of real estate price formation factors using variable selection and hedonic approach using Sparse estimation

富山県立大学 工学部 情報システム工学科
2120031 中島健希
指導教員 奥原 浩之 教授
提出年月: 令和7年（2025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
識別境界に直行している射影軸	$\boldsymbol{w}$
クラス間変動行列	$\boldsymbol{S}_B$
クラス内変動行列	$\boldsymbol{S}_W$
データセット内のクラス	C_n
クラスの平均値	$\boldsymbol{m}_n$
データセット内の多数クラス	C^{maj}
多数クラスのサンプル数	N^{maj}
データセット内の少数クラス	C^{min}
少数クラスのサンプル数	N^{min}
データセットの不均衡度	r
説明変数の集合	$\boldsymbol{X}$
学習済みのモデル	$\hat{f}(\boldsymbol{X})$
インスタンス i の説明変数の集合	$\boldsymbol{x}_i$
インスタンス i の予測値	$\hat{f}(\boldsymbol{x}_i)$
モデル $\hat{f}(\boldsymbol{X})$ の予測の期待値	$\mathbb{E}\left[\hat{f}(\mathbf{X})\right]$
インスタンス i の説明変数 $x_{i,j}$ の限界貢献度	$\Delta_{i,j}$
インスタンス i の説明変数 $x_{i,j}$ の SHAP 値	$\phi_{i,j}$
2 点 $(x_1, y_1), (x_2, y_2)$ の距離	D
2 点 $(x_1, y_1), (x_2, y_2)$ の緯度の差	D_y
2 点 $(x_1, y_1), (x_2, y_2)$ の経度の差	D_x
子午線曲率半径	M
卯酉線曲率半径	N
離心率	E
長半径	R_x
短半径	R_y
道路ネットワークにおける平均次数	k

第1章

はじめに

§ 1.1　本研究の背景

近年, ヘドニック・アプローチは不動産価格や消費財の価格形成を分析する手法として広く利用されている. ヘドニック・アプローチは, 物品やサービスの価格をその特性（特徴）に分解することにより, 市場価値をより正確に評価する方法として注目されている [1]. 具体的には, 不動産市場では土地や建物の立地, 面積, 築年数, 周辺のインフラ整備状況, さらには眺望や近隣施設の利便性といった多様な特徴が価格にどのように影響を与えるかを分析するために活用されてきた. また, 消費財の分野では, 製品の品質やデザイン, ブランドイメージ, 付加価値機能などが消費者の支払意欲にどのように結びつくかを評価する際に利用されている [2]. これらの応用は, 政策立案やマーケティング戦略の設計, さらには課税や価格設定の根拠を提供する上で, 有用性が高いとされる.

一方で, ヘドニック・アプローチによる推定結果が歪められる要因として, いくつかの問題が既存の研究において指摘されている. 特に, 多重共線性や欠落変数のバイアスが重要な課題とされている [3]. 多重共線性は, 説明変数間で高い相関がある場合に, 回帰モデルの推定を不安定にし, 結果として誤った推定値を生じさせる可能性がある. また, ヘドニックモデルにおいて重要な変数が欠落する場合, 価格形成に関する重要な側面を見逃し, 推定結果にバイアスが生じることがある. このような問題は, 不動産市場における物件の価値評価や消費財の価格設定に直接的な影響を与えるため, 解決すべき重要な課題とされる.

さらに, 交互作用効果や非線形効果を適切に捉えることができていないケースも多いと報告されている [3]. 例えば, 不動産の立地と面積がそれぞれ価格に影響を与えるだけでなく, それらが相互作用することで価格に与える影響が生じる可能性がある. 同様に, 消費財においても, ブランド力と機能性の組み合わせが単純な線形関係では説明できない影響を価格に及ぼす場合が少なくない. しかし, 既存のモデルではこれらの複雑な関係性を十分に反映することができていない. 加えて, ヘドニック・アプローチの推定結果は, データの質やモデル選択によっても大きく影響を受けることが明らかになっている [4]. 信頼性の高い分析を行うためには, データの特性に応じた柔軟な手法の適用が求められる.

近年, 機械学習や統計モデリングを活用した新たなアプローチが注目されている. これらの手法は, 多次元データの処理や非線形構造の解析に適しており, 従来の回帰モデルでは扱いきれなかった複雑な関係性を捉える可能性がある [5]. 特にランダムフォレストやサポートベクターマシンなどの手法は, 非線形性のモデル化において効果を発揮することが示されている [6]. また,ElasticNet や LASSO といった正則化法は, 多重共線性への対処や重要変数の選択において有用性が高いとされる [7]. これらの進展により, ヘドニック・アプローチの

推定精度を向上させるための新たな可能性が開かれている.

§ 1.2　本研究の目的

本研究の目的は, ヘドニック・アプローチにおける従来の限界を克服し, 価格形成におけるより正確で信頼性の高い評価手法を提供することである. 従来のヘドニックモデルでは, 多重共線性, 欠落変数のバイアス, 交互作用効果および非線形効果の取り扱いに問題があり, これらは不動産市場や消費財の価格評価において誤差を生じさせる要因となる [8]. 本研究は, これらの問題を解決するために, スパース推定法を活用した新たなアプローチを提案する.

従来のヘドニックモデルは, そのシンプルな線形構造と主要変数の関係を明確に示す点で強力であったが, 現実の市場では非線形な相互作用や高次の効果が絡むため, これらを十分に捉えることができないことが多かった. また, 異なる要因が相互作用を持つ場合, その効果を正確に評価するためには, より複雑なモデルが求められる. そこで本研究では,ElasticNet およびその進化形である AdaptiveElasticNet(AEN) を用い, 従来の限界を克服することを目指す.

ElasticNet は L1 および L2 正則化を組み合わせることで, 多重共線性の問題に対応すると同時に, 重要な変数の選択を行う能力を持つ. これにより, データ中で重要な特徴を捉え, 効果的な変数選択を実現できる. また,AEN は, 従来の ElasticNet をさらに発展させ, データの特性に応じて正則化パラメータを動的に調整する. この適応的なアプローチは, 複雑なデータ構造に対して非常に効果的であり, 非線形性や交互作用を持つデータにも適用できる. これにより, 不動産市場や消費財の価格形成における微細な変動や相互作用を捉え, 価格予測の精度を向上させることが可能となる.

さらに, 本研究では, 従来の線形回帰モデルでは捉えきれなかった非線形性や交互作用効果をスパース推定法を活用することで反映させることができる. 非線形回帰や機械学習技術と組み合わせることにより, より精緻な価格形成モデルが構築され, 従来の方法では見過ごされがちな要因を捉えることが可能となる. 例えば, 不動産市場における立地や面積の相互作用, 消費財におけるブランド力と機能性の複合的影響など, これらの複雑な関係性を反映させることで, より現実的な市場分析が行えるようになる.

また, こうした高度な統計手法や機械学習技術の活用により, 従来のヘドニック・アプローチを大きく強化することができる. 特に, 従来のアプローチでは捉えきれなかった価格形成における微細な要因を, スパース推定を用いることで明示化し, より高精度な価格評価を可能にする. これにより, 不動産市場や消費財市場における価格形成メカニズムの理解が深まり, 実務的な応用に貢献することが期待される.

§ 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した．背景では，特に欧米における地理的犯罪予測の歴史と事例について述べた．目的では，わが国における地理的犯罪予測の課題に

ついて述べ，本研究の意義について述べた．

第 2 章 地理的犯罪予測の概要と，その手法についてそれぞれ述べる．また，犯罪が発生するリスクについて述べる．さらに，地理的犯罪予測には欠かせない GIS について，その概要を述べる．

第 3 章 不均衡なデータに対するアプローチと，機械学習によって作成されたモデルを解釈する手法について述べる．また，さまざまな要因を考慮するため，サイバー空間から多様なデータを取得し，処理する方法について述べる．

第 4 章 データセットを作成し，不均衡に対処して犯罪発生予測モデルを作成する．さらに，その予測モデルに解釈手法を適用し，犯罪が発生する要因を可視化するまでの流れを説明する．

第 5 章 実際の犯罪発生データを用いて，第 4 章で述べた手法で，犯罪発生予測モデルを作成し，その予測精度を検証する．また，解釈手法によって可視化された要因が妥当なものであるかを確認する．

第 6 章 本論文における前章までの内容をまとめつつ，本研究で実現できたことと今後の展望について述べる．

第2章

多様な要因を考慮したデータセットの作成

§ 2.1　サイバー空間からのデータ取得

土地価格の変動には，数多くの要因が考えられる．そのため，土地価格を予測するモデルを作成するためには，それらを表現する説明変数を多く考慮する必要がある．しかし，我々が一般に取得できるデータ，すなわちオープンデータには，そのアクセスに限界がある．実際に，日本で公開されているオープンデータの数は，世界で最も公開されている台湾と比較して，約 67.7 %である [**?**, opendata] 国勢調査の結果など，統計的なデータは比較的公開されているものの，土地価格の要因として重要視される地理的なデータ，たとえば，特定の施設の位置などといったものは，依然として取得が容易ではない．そこで，本研究では，地理的なデータを地図画像やナビゲーションサービスから取得し，補うこととした．

地図画像の取得

地図画像は，その場所やその周囲の地理的な特徴を表す重要なデータである．そこで，本研究では，Mapbox から取得した地図画像から説明変数を抽出している．Mapbox は，機能 14 やデザインを自由にカスタムして，地図を自身の Web ページやアプリに埋め込むことができるサービスである．さまざまな API を公開しており，住所などから緯度・経度を算出する Geocoding API，ルートを検索する Directions API などがあるが，本研究では，地図をベクター画像として取得できる Mapbox Static Tiles API を用いて，地理的なデータを取得する．

Step 1: Mapbox Studio 上で，カスタムマップを作成する

Mapbox Studio では，地図上にあるさまざまな要素の色や表示の有無を自由に変更することができる．

Step 2: 緯度と経度から，取得するタイルを算出する

Mapbox Static Tiles API では，地球上のすべての範囲を正方形で仕切ったタイルごとに地図画像を取得できる．すなわち，緯度と経度から，特定のタイルを一意に決定することができる．対象の緯度と経度 (lat, lng) が含まれるタイル (X, Y) は，次のように算出できる．

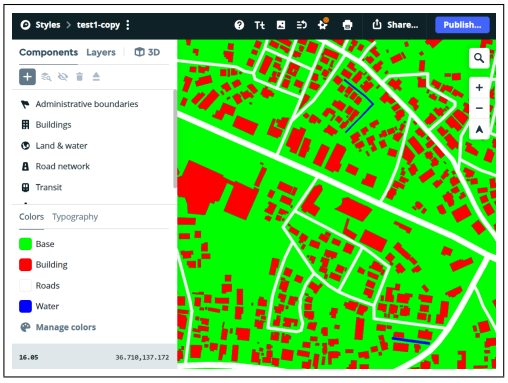

図 2.1: Mapbox Studio

図 2.2: NAVITIME

$$X = \lfloor \frac{lon + 180}{360} * 2^z \rfloor \tag{2.1}$$

$$Y = \lfloor \frac{\log_e tan \left(lat \frac{\pi}{180} + \frac{1}{cos\left(lat \frac{\pi}{180} \right)} \right)}{\pi} 2^{z-1} \rfloor \tag{2.2}$$

ここで，$\lfloor x \rfloor$ は，$n \leqq x < n+1$ を満たす整数 n を表す．また，z はズームレベルである．たとえば，$z = 17$ では 1 ピクセルあたり 1.194m，$z = 18$ では 1 ピクセルあたり 0.597m の地図画像を取得できる．Mapbox Static Tiles API で取得できる地図画像の大きさは 512×512 であるため，$z = 17$ では一辺が約 611m，$z = 18$ では約 306m である．

Step 3: Mapbox Static Tiles API を用いて，地図画像を取得する

以上により，タイル X, Y，およびズームレベル z を算出・決定したら，Mapbox Static Tiles API として指定されている URL に，それらをパラメータとして GET リクエストを行う．レスポンスされたデータはバイト列であるため，1 つの座標に RGB 値を格納する 3 次元配列に変換を行えば，画像として処理することができる．

施設データの取得

特定の施設やその近くは，犯罪の発生の要因となる可能性がある．施設データを取得できるサービスとして，Google Maps API が存在するが，無料で取得できる数に制限があるほか，たとえば遊園地や水族館など，レジャー施設としてジャンル分けできるものに対して，「レジャー施設」と検索しても，それらを網羅できるとは限らない点で，採用しなかった．そこで，ナビゲーションサービスのひとつである「NAVITIME」から施設データをスクレイピングして取得することとした．NAVITIME は，施設のジャンルごと，さらには都道府県ごとに一覧となって表示される．

図 2.3: Mapbox Studio

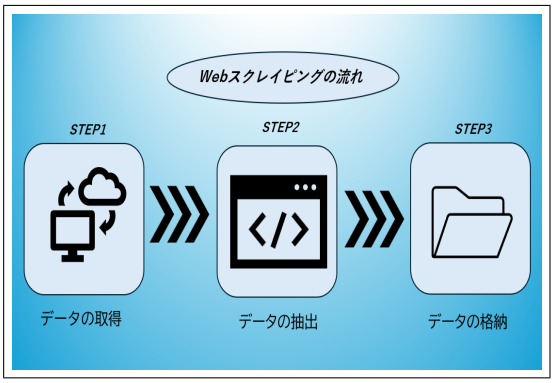

図 2.4: NAVITIME

スクレイピング

スクレイピングとは，データを収集し，かつ目的に合わせて加工することである．特に，Web 上から必要なデータを取得することを，Web スクレイピングと呼ばれている．Web スクレイピングの流れについて図 2.5 に示す．様々なツールやプログラミングでスクレイピングを自動化することで，Web データの収集にかかる手間や時間は大幅に削減が可能である．スクレイピングと似ている意味の言葉にクローリングがある．クロージングとは，Web 状で様々なサイトを巡回し，情報の保存や複製など様々なことを行うことを指す．クロージングとスクレイピングはともに情報を収集手段ではあるが，クロージングが巡回に焦点を当てている一方でスクレイピングは情報の抽出に焦点を当てている．また，企業や公共機関は，情報やデータを提供してくれることもあり，その際に使われている仕組みは API と呼ばれている．クローリングやスクレイピングをする前に，必要な情報が API によって提供されているかどうかまず確認することが大切になる．Web スクレイピングに主に用いられるツールとして，BeautifulSoup4 や，Selenium がある．ログインやボタンのクリックなどの，マウス操作が必要な Web サイトや，JavaScript で記述されている Web ページのスクレイピングするときは Selenium が用いられている，それらの処理を必要としない Web サイトには，高速でスクレイピングができる BeautifulSoup4 が使用されることが多い．

Beautiful Soup4

BeautifulSoup4 とは，Web サイト上の HTML から，必要なデータを抽出するための Python のライブラリである．Beautifulsoup4 でスクレイピングする際，最初に対象の Web ページから HTML を取得する必要がある．HTML を取得する方法として，同じく Python のライブラリである，Requests の get 関数や，Selenium の page source 関数を使うなどの方法がある．上記の方法によって取得された HTML テキストを，BeautifulSoup4 の BeautifulSoup 関数に渡すことで，BeautifulSoup オブジェクトを作成することができる．また，そのオブジェクトから class を検索することで Web サイトの必要な情報を抽出する．

Selenium

Selenium は, ウェブアプリケーションの自動化とテストを目的としたオープンソースツールであり,Python を含む多言語で利用可能である. その中核を成す「Selenium WebDriver」は,

主要なウェブブラウザ（Chrome,Firefox など）をプログラムで制御し, ボタン操作やフォーム入力, 動的コンテンツの処理, クロスブラウザテストを可能にする. ウェブアプリケーションの動作検証や動的データの抽出に広く活用されており, スクレイピングや定型作業の自動化にも応用される.

§ 2.2　データセットに対する前処理

土地価格の変動には，数多くの要因が考えられる．そのため，土地価格を予測するモデルを作成するためには，それらを表現する説明変数を多く考慮する必要がある．しかし，生のデータには欠損値や外れ値が含まれていることが多く，そのままではモデルの精度が低下する恐れがある．本節では，回帰分析におけるデータセットの前処理手法について説明する．

データの前処理の概要

前処理は，モデルの精度向上や学習の安定性を確保するために不可欠な手順である．データセットの前処理では，以下の手順を実施する．

Step 1: 欠損値の処理　データセットには，しばしば欠損値が含まれている．欠損値の処理には，以下のような方法がある．

- **削除**：欠損のあるデータポイントを削除する．ただし，サンプルサイズが減少するリスクがある．
- **補完**：欠損値を他の値で補完する方法．平均値，中央値，最頻値での補完や，k-近傍法（kNN）を用いた補完がある．
- **モデルを用いた補完**：機械学習モデルを用いて，欠損部分を予測する方法もある．

Step 2: 外れ値の処理　外れ値は，モデルの予測精度を低下させる要因となる．外れ値の検出方法には，以下の手法がある．

- **四分位範囲（IQR）による検出**：四分位範囲（IQR）の 1.5 倍を超えるデータポイントを外れ値とみなす方法.
- **標準偏差を用いた検出**：平均からの標準偏差が一定の範囲を超えるデータを外れ値とする方法.
- **視覚的検出**：散布図や箱ひげ図を用いて視覚的に外れ値を確認する方法.

外れ値は, そのまま維持するか, 削除するか, または他の値に変換する（Winsorization）かのいずれかを選択する．

Step 3: カテゴリ変数のダミー変数化　ダミー変数とはカテゴリカルデータを「0」または「1」の数値データに変換した変数のことである. カテゴリ変数は，機械学習モデルでは直接使用できないため，数値データに変換する必要がある．代表的な方法は，**ワンホットエンコーディング**である．

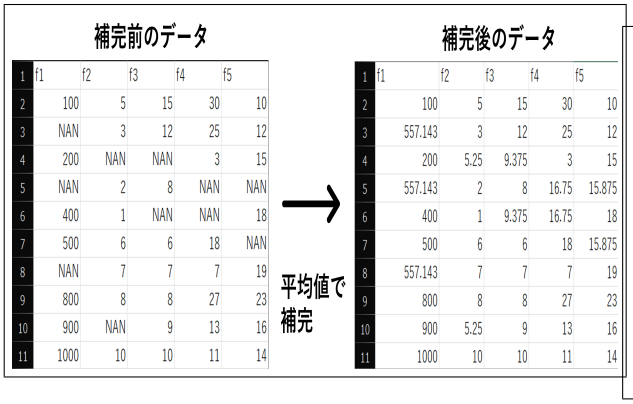

図 2.5: 欠損値の補完の例

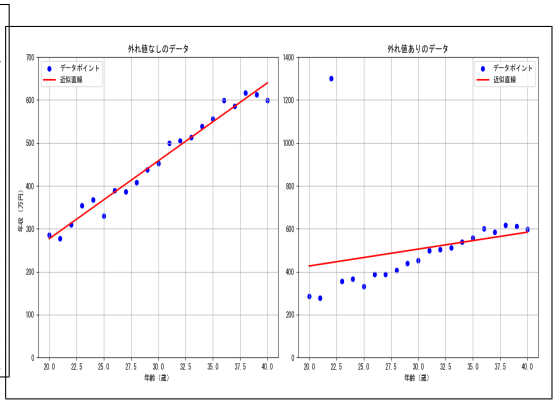

図 2.6: 外れ値の影響の例

- **ワンホットエンコーディング**：カテゴリ変数を 0 または 1 の二値変数に変換する方法．たとえば，`A`，`B`，`C` という 3 つのカテゴリがある場合，これを `[1, 0, 0]`, `[0, 1, 0]`, `[0, 0, 1]` といったベクトルに変換する．
- **ラベルエンコーディング**：カテゴリ変数に対して整数を割り当てる方法．ただし，この手法はカテゴリの大小関係が生じてしまうため，回帰分析には不向きな場合がある．

- **3 つ以上のカテゴリを持つ変数のダミー変数化**：3 つ以上のカテゴリを持つ変数の場合, ワンホットエンコーディングにより, 各カテゴリに対応するダミー変数を作成する．例えば,A, B, C, D というカテゴリ変数がある場合, これを [1, 0, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0], [0, 0, 0, 1] のように変換することができる．
- **ダミー変数の落とし込み（ダミー変数の落とし込み問題）**：ダミー変数化の際, 冗長な変数が生じることを避けるため,1 つのカテゴリ変数を削除することが一般的である．例えば,4 つのカテゴリ変数がある場合,3 つのダミー変数を作成し, 残り 1 つを基準カテゴリとして扱う．この基準カテゴリは, 回帰分析において「参照カテゴリ」となり, 他のカテゴリとの相対的な影響を示す.
- **注意点**：ダミー変数を多く生成しすぎると, モデルの計算負担が大きくなることがあるため, カテゴリ数が多い場合には, 次元削減技術（例えば, 主成分分析（PCA）など）を検討することが望ましい.

Step 4: 二次項の作成　モデルの性能を向上させるために，特徴量間の非線形関係を捉えるために二次項を作成することが有効な場合がある．

- **二次項の作成方法**：特徴量 $x_1, x_2, \ldots, x_n$ に対して，$x_1^2, x_2^2, \ldots, x_n^2$ を新たに特徴量として追加することができる．
- **二次項の有用性**：線形モデルにおいて, 二次項を加えることで, モデルがより複雑なデータのパターンを学習できるようになる．

Step 5: 交互作用項の作成　特徴量間の相互作用を捉えるために交互作用項を作成する．交

互作用項を使うことで，ある特徴量が別の特徴量に与える影響をモデルに組み込むことができる．

- **交互作用項の作成方法**：交互作用項は以下のパターンで作成することができる：
  - 連続変数同士の交互作用：$x_1 \times x_2$（連続変数同士の相互作用）
  - 連続変数とダミー変数の交互作用：$x_1 \times D$（連続変数とダミー変数の相互作用）
  - ダミー変数同士の交互作用：$D_1 \times D_2$（ダミー変数同士の相互作用）
- **交互作用項の有用性**：交互作用項を加えることで，線形回帰や他の機械学習モデルがより正確に相関関係を捉えることができる．

Step 6: 標準化

1. **Z スコアによる標準化** 各変数の値から平均を引き, 標準偏差で割ることで, 平均が 0, 標準偏差が 1 となるように変換する. これにより, 異なるスケールの変数を均一な基準に揃えることができる. 標準化の数式は次の通りである.

$$z = \frac{x - \mu}{\sigma} \tag{2.3}$$

ここで,x は元の値,μ は平均,σ は標準偏差である.

2. **Min-Max スケーリング** 変数の最小値を 0, 最大値を 1 に変換する方法である. すべてのデータが 0 から 1 の範囲に収まるため, ニューラルネットワークのようなモデルでよく用いられる. スケーリングの数式は以下の通りである.

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{2.4}$$

ここで,x は元の値,$x_{\min}$ は最小値,$x_{\max}$ は最大値である.

このように，データセットの前処理は，モデルの精度向上に欠かせない手順であり，各手法の選択は問題の特性に合わせて慎重に行う必要がある．

前処理の重要性

前処理は，モデルの予測精度や安定性に大きな影響を与えるため，回帰分析の成功にとって極めて重要である．欠損値や外れ値がある状態でモデルを学習させると，モデルのパフォーマンスが低下することがあるため，適切な前処理が求められる．また，カテゴリ変数をダミー変数に変換しないままモデルに投入すると，予測誤差の増大を招く可能性がある．そのため，データセットの内容を把握し，必要な前処理を慎重に行うことが不可欠である．

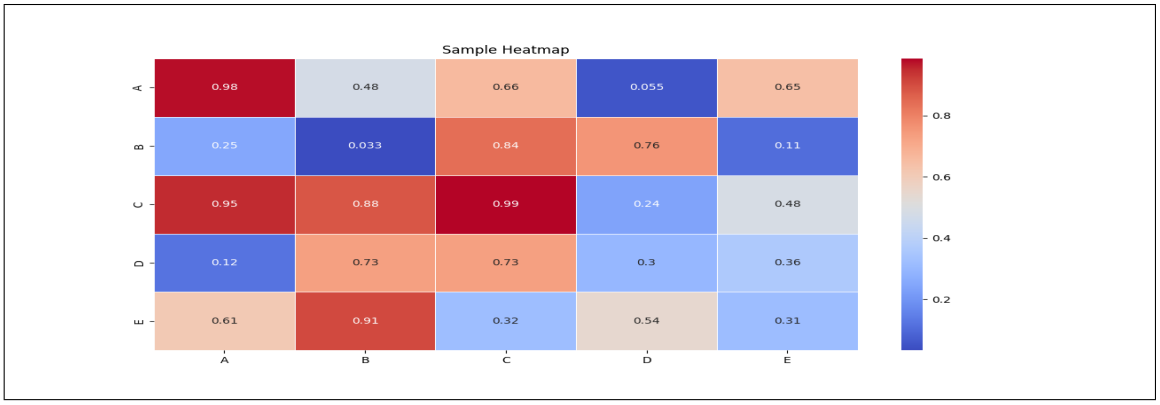

図 2.7: ヒートマップの例

§ 2.3　説明変数の選定手法

説明変数の選定は, 回帰モデルや機械学習モデルの性能を最大化するために欠かせない工程である. 適切な説明変数を選ぶことにより, モデルの予測精度が向上し, 過剰適合（オーバーフィッティング）を防ぐことができる. また, 説明変数の選定はデータの理解を深め, モデルの解釈性を高める重要なプロセスでもある. この章では, 説明変数選定に用いられる代表的な手法を詳述する.

説明変数選定の概要

説明変数選定手法には, 以下のような代表的なものがある.

- フィルタ法
- ラッソ回帰（Lasso Regression）
- ステップワイズ法（Forward Selection,Backward Elimination）
- 主成分分析（Principal Component Analysis,PCA）
- 共分散選択基準（VIF, 条件数）

これらの手法はそれぞれ異なるアプローチで変数選定を行うが, 共通して目的変数に対する説明変数の重要度を評価し, 不必要な変数を削除することでモデルのパフォーマンスを最適化する. 以下で, それぞれの手法について詳しく解説する.

フィルタ法　フィルタ法は, 目的変数との相関関係や統計的有意性を基に説明変数の選定を行う手法である. 最も一般的な方法として, 相関係数や統計的検定を用いて変数の関連性を評価する. フィルタ法はモデル構築前に変数を選定するため, 計算が非常に速く, 簡単に実行できるが, 説明変数間の相互作用や非線形な関係を捉えることはできない.

- **相関係数**：目的変数との相関が強い説明変数を選ぶ. 相関係数が 0.8 以上の場合, 説明変数間で多重共線性が発生する可能性があり, その場合は片方の変数を削除することが推奨される.

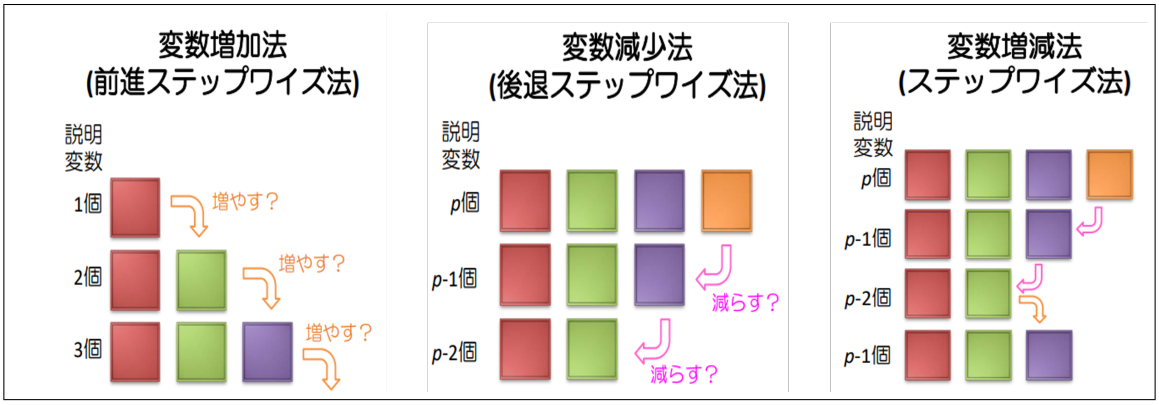

図 2.8: ステップワイズ法の手法

- **t 検定や F 検定**：t 検定は単一の変数が目的変数と有意に関連しているかを評価し,F 検定は複数の変数が有意であるかを評価する. これらを用いて, 統計的に有意な変数を選択することができる.

フィルタ法の利点は計算が高速である点だが, 変数間の複雑な相互作用を無視するため, 重要な変数が選ばれない可能性もある.

ラッソ回帰 ラッソ回帰は回帰分析において正則化を施すことで, 不要な説明変数の影響を制御する手法である.Lasso（Least Absolute Shrinkage and Selection Operator）は, 回帰係数に L1 ノルム（絶対値の合計）をペナルティとして加え, 係数がゼロに近づくように最適化する. これにより, 自動的に説明変数の選定が行われ, 冗長な変数が排除される.

ラッソ回帰の目的関数は以下の通りである：

$$\hat{\beta} = \underset{\beta}{\text{minimize}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{2.5}$$

ここで,λ は正則化パラメータであり,β_j は回帰係数である.λ が大きいほど, 回帰係数がゼロに近づき, モデルが単純化される. ラッソ回帰は特に, 多くの変数を扱う場合や, 変数間に相関がある場合に有効な手法である.

ラッソ回帰は過剰適合を防ぎ, 簡素なモデルを得るために役立つが, 正則化パラメータ λ の選定が重要である. 過剰にペナルティをかけると, モデルが過度に単純化され, 重要な変数を排除してしまうことがある.

ステップワイズ法 ステップワイズ法は, 変数選定を段階的に行う方法である. 最初に空のモデルから始め, 変数を 1 つずつ追加または削除しながら最適なモデルを見つける. この過程で,AIC（赤池情報量基準）や BIC（ベイズ情報量基準）を用いてモデルの適合度を評価する.

- **前進選択法**：最初に空のモデルから始め, 最も目的変数との関連が強い説明変数を追加していく. この方法では, 追加された変数がモデルにどれだけ貢献するかを, 赤池情報量基準（AIC）や決定係数（R^2）を基準に評価する.
- **後退消去法**：最初にすべての説明変数を含むフルモデルから始め, 最も影響の少ない変数を 1 つずつ削除していく. 変数削除の基準としては,p 値や AIC がよく使用される.
- **双方向法**：前進選択法と後退消去法を組み合わせた手法で, 変数を追加しながら, 同時に不適切な変数を削除するプロセスを繰り返す.

ステップワイズ法は直感的で使いやすい手法だが, 変数の順番や基準に依存するため, 異なる結果が得られることがある. また, モデルが複雑になりすぎる場合や, 変数間に強い相関がある場合には, 過剰適合を引き起こす可能性がある.

主成分分析 主成分分析は, 説明変数間の相関を低減し, 次元を削減するための方法である.PCA は, 説明変数の線形結合である主成分を抽出し, これを新しい説明変数として使用する.PCA を使用することにより, 変数間の相関がなくなるため, モデルの計算が効率的になり, 過剰適合を防ぐことができる.

PCA で得られる主成分は以下の式で計算される：

$$z_k = \sum_{j=1}^{p} w_{kj} x_j \tag{2.6}$$

ここで,w_{kj} は第 k 主成分の係数で,x_j は元の説明変数である.PCA によって得られる主成分は, 元の変数よりも少ない次元数であり, 計算の効率化を図ることができる. 主成分分析は, 特に多くの変数が相関している場合に有効で, 説明変数の数が多すぎてモデルが複雑になる場合に有用である.

共分散選択基準（VIF, 条件数） 共分散選択基準は, 説明変数間の多重共線性を評価し, それを防ぐために変数選定を行う手法である. 多重共線性とは, 説明変数が互いに強い相関を持ち, 回帰分析で不安定な結果を引き起こす現象である.VIF（分散拡大係数）や条件数を用いて多重共線性を評価し, 問題がある場合は変数を削除または標準化する.

- **VIF**：VIF は, 各説明変数が他の変数とどの程度相関しているかを示す指標である.VIF が 10 を超える場合, その変数は多重共線性を引き起こす可能性が高いとされ, 除外が検討される.
- **条件数**：条件数は, 設計行列の特異値分解を用いて評価される指標で, これが高い場合も多重共線性が発生している可能性を示唆する. 一般的に, 条件数が 30 以上であれば, 多重共線性が問題となる可能性がある.

共分散選択基準を用いることで, モデルが不安定になるのを防ぎ, より頑健な回帰モデルを構築することができる.

説明変数選定の重要性

説明変数の選定は, モデルの精度に大きな影響を与える. 適切な変数を選定することにより, モデルの予測能力が向上し, 過剰適合を防ぐことができる. 不要な変数を排除することにより, 計算効率が高まり, モデルが解釈しやすくなる. 各手法はデータや目的に応じて使い分けるべきであり, 変数選定が適切に行われることが, 良い予測モデルを作成するための鍵である.

第3章

ヘドニック・アプローチによる土地価格決定要因の分析

§ 3.1　ヘドニック法とその問題点

ヘドニック法は, 製品の特性や属性がその価格に与える影響を定量化する手法である. 例えば, 住宅価格を分析する際, その住宅の広さ, 立地, 築年数, 設備等が価格に与える影響をヘドニック法を用いて測定する. このような分析は, 不動産市場における価格変動を理解するために不可欠である [10].

ヘドニック法は回帰分析を基盤とし, 価格を説明変数（製品の特性）に基づいて回帰する形で表現される. 一般的な数式は以下の通りである.

$$P = \beta_0 + \sum_{i=1}^{k} \theta_i X_i + \epsilon \tag{3.1}$$

ここで,P は価格,X_i は製品の特性,β_i はそれぞれの特性に対応する回帰係数,ϵ は誤差項である.

ヘドニック法の適用には以下のような問題点が存在する.

多重共線性

多重共線性は, 重回帰分析において 2 つ以上の説明変数が高い線形関係にある状況を指す. これは, ある説明変数が他の説明変数によって説明される場合に発生し, 回帰係数の推定値が不安定になるため, 重回帰分析の結果が誤解を招く可能性がある [**?**].

多重共線性の原因

多重共線性が生じる主な原因は以下の通りである.

- **説明変数の選択による原因**: 同じ種類のデータを複数の指標で表現する場合や, 指標の計算方法が類似している場合に相関が高くなり, 多重共線性が生じることがある. また, 説明変数が多すぎる場合にも多重共線性が生じる可能性がある.
- **データ収集時の問題による原因**: サンプルサイズが小さい場合や, 説明変数が取りうる値の範囲が狭い場合に相関が高くなり, 多重共線性が生じることがある. また, 説明変数が一部欠損している場合にも, 欠損していない説明変数との相関が高くなり, 多重共線性が生じることがある.

多重共線性の影響

多重共線性が生じると, 以下のような影響がある.

- **回帰係数の推定値の不安定化**: 説明変数間に高い相関があると, その相関に応じて回帰係数の値が大きく変化することがある. これにより, 回帰係数の推定値に対する信頼性が低下し, 説明変数の効果を正確に評価できなくなる.
- **モデルの解釈における問題点**: 説明変数同士が強く相関しているため, モデルの解釈が困難になる. 例えば, ある説明変数が目的変数に影響を与えていると考えられた場合でも, 実際には他の説明変数と相関していることが原因で, その影響を正確に評価できない場合がある.
- **過学習の問題**: 説明変数の数が多くなるため, モデルが複雑になり過ぎて過学習の問題が生じることがある. 過学習とは, 学習データに過剰に適合したモデルを構築し, 新しいデータに対して予測精度が低下する現象である.

多重共線性の検出

多重共線性が生じているかどうかを検出する方法には, 以下のものがある.

- **相関行列や散布図行列による検出方法**: 相関行列や散布図行列を用いて, 説明変数間の相関を確認することができる. 相関係数が高い説明変数がある場合には, 多重共線性が生じている可能性がある.
- **分散拡大係数による検出方法**: 分散拡大係数は, ある説明変数の回帰係数の標準誤差を, その説明変数の標準偏差で割った値である. 分散拡大係数が大きい説明変数がある場合には, 多重共線性が生じている可能性がある.

多重共線性の対処法

多重共線性が生じた場合には, 以下のような対処法がある.

- **不要な説明変数の削除**: 相関が高い説明変数のうち, モデルにとって不要となる変数を削除することが効果的である.
- **変数選択法による説明変数の選択**: 前向き選択法, 後退的除去法, ステップワイズ法などを用いることで, モデルに必要な説明変数のみを残すことがである.
- **相関が強い説明変数同士を組み合わせて新しい説明変数を作成する方法**: 相関が強い説明変数を組み合わせて新しい説明変数を作成することで, 多重共線性が生じるリスクを軽減することがである.
- **主成分分析による次元削減法**: 主成分分析は, 説明変数をより少ない数の主成分に圧縮することで, 多重共線性が生じるリスクを軽減することがである.

欠落変数によるバイアス

欠落変数によるバイアスは, 回帰分析において説明変数の一部がモデルから除外されることによって生じる. このようなバイアスは推定結果を歪める原因となり, 特に欠落変数が予測に重要な影響を与える場合に顕著である [18].

欠落変数によるバイアスの影響

欠落変数バイアスが生じた場合, 回帰係数の推定は不正確となり, 実際の変数間の関係性が歪められる. 特に, 欠落した変数が他の変数と強い相関関係を持っている場合, 回帰係数が過大に評価されたり, 過小に評価されたりすることがある. このバイアスを無視して推定を行うと, モデルの予測性能が低下し, 結果の信頼性が損なわれる可能性がある [20].

欠落変数バイアスの処理方法

欠落変数バイアスを避けるためには, 以下の方法が考えられる.

- **変数の選定を慎重に行う**：分析に必要な変数を慎重に選び, 欠落する可能性のある重要な変数を事前に特定する.
- **補完法の使用**：欠落データがランダムに発生している場合, 補完法（例えば, 多重補完）を使用してデータを補う.
- **固定効果回帰モデルの利用**：特定の変数が欠落している場合に, 固定効果モデルを用いることで, 時間やグループに依存するバイアスを軽減する.

これらの方法を適切に適用することで, 欠落変数バイアスの影響を最小限に抑え, より信頼性の高い推定結果を得ることが可能となる.

交互作用の存在

交互作用項は, 複数の説明変数が同時に目的変数に与える影響が変化する場合に用いられる分析手法であり, 特に回帰モデルにおいて重要な役割を果たす. 交互作用項を導入することで, 説明変数の効果が他の変数の水準によって異なることを考慮でき, より精緻なモデルの構築が可能となる.

交互作用の導入方法

交互作用項を回帰モデルに組み込む場合, 基本的な回帰式に交互作用項を追加する. 例えば,2 つの説明変数 X_1 と X_2 の交互作用を考慮する場合, 以下のようにモデルを構築する.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) + \epsilon \tag{3.2}$$

ここで,$X_1 \times X_2$ は交互作用項を表し, この項が回帰係数 β_3 に与える影響を評価することで,X_1 と X_2 の組み合わせによる効果を捉えることができる [22].

交互作用項の解釈

交互作用項を含むモデルの解釈は, 単純な回帰モデルと異なり複雑である. 交互作用項が有意である場合, その効果は他の説明変数の値に依存するため, 単独の説明変数の影響を解釈する際には慎重を要する. 例えば, 交互作用項 $X_1 \times X_2$ の係数が有意であれば,X_1 が X_2 の水準に応じて異なる影響を目的変数に与えることを意味する [23].

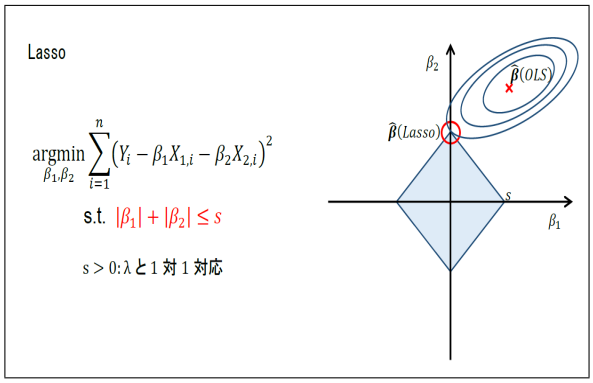

図 3.1: Lasso のイメージ図

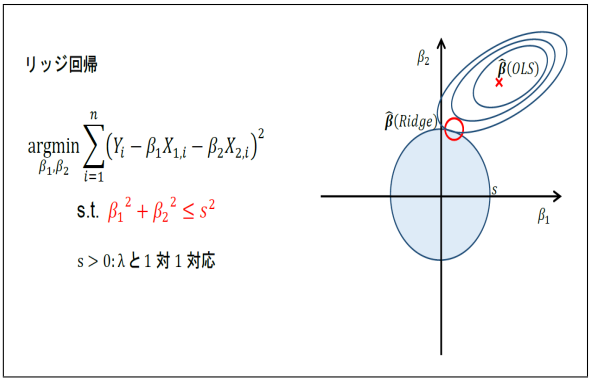

図 3.2: Ridge のイメージ図

§ 3.2　スパース推定

スパース推定は, モデルの複雑さを抑制しつつ, 予測性能を向上させるための手法である. 特に高次元データにおいては, すべての説明変数をモデルに含めるのではなく, 一部の重要な変数のみを選択することが望ましい. このような背景から, スパース推定は回帰分析において重要な役割を果たす.

スパース推定の代表的な手法

スパース推定の代表的な手法には, 以下のものがある.

Lasso 回帰

Lasso (Least Absolute Shrinkage and Selection Operator) 回帰は, 目的関数に ℓ_1 正則化項を加えることで, 不要な変数の係数を 0 にする手法である. これにより, 変数選択が自動的に行われ, モデルの解釈性が向上する.

$$\hat{\boldsymbol{\beta}}^{(Lasso)} = \arg\min_{\boldsymbol{\beta}} \left(\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \right) \tag{3.3}$$

ここで,y_i は目的変数,x_i は説明変数,β は回帰係数,λ は正則化パラメータである.λ の値が大きいほど, より多くの変数の係数が 0 になる.

Ridge 回帰

Ridge 回帰は, 目的関数に ℓ_2 正則化項を加える手法である.ℓ_2 正則化は, すべての係数を小さな値に抑制するが,Lasso のように係数を 0 にはしない.

$$\hat{\boldsymbol{\beta}}^{(Ridge)} = \arg\min_{\boldsymbol{\beta}} \left(\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right) \tag{3.4}$$

Ridge 回帰は, 多重共線性の問題を緩和する効果があり, すべての変数をモデルに残しながら過学習を防ぐ役割を果たす.

Elastic Net

Elastic Net は,Lasso と Ridge の正則化項を組み合わせた手法である.ℓ_1 と ℓ_2 の両方の効果を取り入れることで, 変数選択と多重共線性の両方の問題に対処できる.

$$\hat{\boldsymbol{\beta}}^{(EN)} = \arg\min_{\boldsymbol{\beta}} \left(\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \right) \tag{3.5}$$

ここで,λ_1 と λ_2 はそれぞれ ℓ_1 と ℓ_2 の正則化パラメータである.Elastic Net は,Lasso と Ridge の利点を同時に享受できる手法として広く用いられている.

グループ効果と多重共線性に対する頑健性

Lasso については，多重共線性の強いデータでは変数選択の結果が不安定になることが知られている．たとえば，ある二つの変数にかかるパラメータの真の値が β_1^*, β_2^* であるとする．このとき，極端な例として，この二つの変数のデータセット上の値が全く同一であるとすると，Lasso による最適化の解は，以下のように無数に存在し，一意に定まらない.

$$\underset{\beta}{\text{Lasso}} = (s(\beta_1^* + \beta_2^*), (1-s)(\beta_1^* + \beta_2^*)) \quad \text{for any } s \in [0,1] \tag{3.6}$$

このように，相関が強い二変数が存在するとき，Lasso の変数選択はデータセットの僅かな変化に強く影響され，どちらの変数がモデルに取り込まれるかが安定しない.

ヘドニック法で用いるデータは相関が強いことが多いため，上記のような多重共線性の問題に対して頑健な性質を持つスパース推定を用いる必要がある．こうした性質を持つ代表的なスパース推定の一つが「エラスティック・ネット（Elastic Net, EN）」である．EN は，回帰式の誤差二乗和に正則化項として ℓ_2 ノルムと ℓ_1 ノルムの双方を加えた関数を最小化する β を推定する手法である.

その結果，変数選択できる Lasso の長所と，多重共線性に強いリッジ回帰の長所を併せ持った推計手法となっている．EN が有する多重共線性に対する頑健性は「グループ効果」と呼ばれる.

グループ効果とは，説明変数間の相関が強い場合に，それらの変数にかかる係数の差が小さくなるような推計結果を与える性質である．極端な例として，二つの説明変数のデータセット上の値が全く同じである場合，EN は，その二つの変数にかかるパラメータを，以下のように全く等しい値として推定する.

$$\underset{\beta}{\text{EN}} = \left(\frac{1}{2}(\beta_1^* + \beta_2^*), \frac{1}{2}(\beta_1^* + \beta_2^*) \right) \tag{3.7}$$

このため，多重共線性が強く，どの変数が真に説明力を有しているのかデータから識別するのが困難な状況でも，安定的な変数選択，パラメータの推定が可能となる.

オラクル性

スパース推定により得られる推定量が満たすべきもう一つの性質として「オラクル性」がある. 具体的には, 真の係数をとしたとき, スパース推定による推定量が以下の二つの条件を満たすことを, オラクル性を持つと言う.

- **変数選択の一致性**:

$$\beta_j^* = 0 \text{ の場合} \quad \lim_{n\to\infty} P(\hat{\beta}_j = 0) = 1 \tag{3.8}$$

これは, 真の係数がゼロである変数について, その係数の推定量が一致性を満たすことを意味する.

- **非ゼロ係数の漸近正規性**:

$$\beta_j^* \neq 0 \text{ の場合} \quad \lim_{n\to\infty} \frac{(\hat{\beta}_j - \beta_j^*)}{\sigma(\hat{\beta}_j)} \sim N(0,1) \tag{3.9}$$

ここで,$\sigma^2(\hat{\beta}_j)$ は推定量の漸近分散を表す. この条件は, 真の係数がゼロでない変数について, その推定誤差が漸近的に正規分布に従うことを意味する.

オラクル性は, スパース推定が同時に行う「変数選択」と「係数の推定」の両者の適正性を漸近的に保証する重要な性質である. しかし, スパース推定の中でも, 上述した Lasso や Elastic Net は, データセット次第では, どれほど適切に正則化パラメータを選択してもオラクル性が満たされないことが知られている. スパース推定におけるオラクル性は, 統計的推論の観点からも重要であり, モデルの信頼性を評価する上で不可欠な要素となっている.

スパース推定の応用事例

スパース推定は, 以下のような分野で活用されている.

- **金融工学**：リスク管理やポートフォリオ最適化において, 不要な資産を選択的に除外するために使用される.
- **医療分野**：ゲノムデータ解析において, 重要な遺伝子を特定するために Lasso や Elastic Net が活用されている.
- **マーケティング**：顧客の購買行動を予測するためのモデルにおいて, 影響力の大きい要因を特定するために利用されている.
- **不動産評価**：土地価格の評価モデルにおいて, 不要な説明変数を除外し, 重要な要因を特定するために用いられる.

§ 3.3 Folium を用いた Web-GIS の開発

GIS には動作するプラットフォームや形態，提供している団体によって複数種類のソフトウェアが存在する．代表的な GIS アプリケーションを表 3.1 に示す．また，本節では，Web アプリケーションとして World Wide Web 上で機能する GIS を Web-GIS と表記するものとする.

html 形式で記述され，Web-GIS は多様なプラットフォーム上で動作する GIS の中でも代表的なフォーマットであるといえる．計算機を用いて広く利用することができ，処理の大部分は html が置かれているサーバ上で行われることで参照する端末自体のスペックに依存

表 3.1: 代表的な GIS ソフトウェア

GIS ソフトウェア	無料	オープンソース	Windows	MacOS	Linux	BSD	Unix	Web
Microsoft MapPoint	×	×	○	×	×	×	×	○
ArcGIS	×	×	○	○	×	×	○	○
GRASS GIS	○	○	○	○	○	○	○	○
QGIS	○	○	○	○	○	○	○	○
MapInfo	×	×	○	×	×	×	×	○
TNTmips	×	×	○	○	○	×	○	×

しづらいことから，利用者も多い．また，html 形式で実装されるがゆえに作成者が直接作成する方法のほか，いくつかのプログラミング言語を用いて自動的に生成することも可能である．

本節では，以上のような特徴を有する Web-GIS の開発において，現在，一般に広く用いられているプログラミング言語の一つである Python のライブラリを用いた方法を解説する．また，Web-GIS において実装することができる代表的な機能とその役割を解説する．

Web-GIS の実装方法

Python を用いた Web-GIS の開発には「Folium」という Python 用のライブラリを用いる．Folium を用いてメソッドに対して初期位置，ベースタイル，初期縮尺などを引数として与えてプログラムを実行することでことで Web-GIS のベースとなるマップを表示する html が自動的に生成される．ベースタイルとして指定することの出来るマップタイルには代表的な例として以下のようなものがある [24].

- CartoDB (positron and dark_matter)
- OpenStreetMap
- Mapbox Bright
- Cloudmade
- Mapbox

Folium による Web-GIS の開発はこのベースマップに対して Folium のライブラリ内に含まれる様々なメソッドを用いることで Web-GIS における各種機能や実際に表示した情報を追加するという形で行われる．ここからは，Folium によって実装することの出来る各種機能とその内容について代表的なものを取り上げる．

ベースマップの切り替えとレイヤの重ね合わせ

2.3 で言及したような GIS によるデータの重ね合わせは FeatureGroup 関数によるレイヤの作成とそれらを制御する LayerControl メソッドによって実現される．また，レイヤとは Web-GIS 上でマップの重ね合わせを行う際のそれぞれの層のことを表す．

まず，FeatureGroup 関数によってベースマップとは別のマップ（レイヤ）を任意の個数作成する．次に，後述する様々なメソッドを用いてベースマップや各レイヤに対して各種機能や情報を追加する．

この時，常に表示され，レイヤの切り替えに左右されないようにする必要のある情報に関してはベースマップに，それ以外の情報に関しては切り替えによって表示したい各レイヤーに追加するようにする．最後に，LayerControl メソッドを用いてレイヤを管理する機能を Web-GIS に付与することによって，html を生成した際に自由にレイヤを切り替える機能を持ったものが生成される．

マーカーを置く

folium.Marker メソッドに対して引数としてマーカーを置く位置の座標を与えることで地図上の任意の位置にマーカーを立てることができる．なお，マーカーは Folium 内に組み込まれているものの中から色やアイコンのマークを自由に切り替えて使用することができるほか，CustomIcon 関数によってアイコンのマークを自作し，独自のマーカーとして使用することもできる．また，任意のマーカーに対して popup 機能を追加し，テキストを付与しておくことでマーカーをクリックした際にポップアップとしてテキストが表示されるようになる．

ヒートマップを描く

ヒートマップとは，二次元データの数値を色やその濃淡で表したものである．広義におけるヒートマップは「マップ」と付いてはいるが必ずしも地図で表現する訳ではなく，テーブルを値で色分けしたものなど数値データを色分けによって可視化したものすべてがこれにあたる．ただ，Web-GIS におけるヒートマップは地図上にプロットされた色の濃淡で数値の大小を示すものである．

また，地図の色分けによって数値の大小を表現する方法として，ヒートマップとは様式が異なるものとして，コロプレス図 (階級区分図) がある．コロプレス図とは，例えば地図を都道府県ごとに境界線で分けて，各都道府県における統計データの大きさによって色分けするなどのものがある．具体的な実例としては，アメリカ大統領選の際の州ごとに赤と青で色分けされた地図などが挙げられる．

folium.plugin.HeatMap メソッドを用いることで地図上の任意の位置を中心としたヒートマップを作成することができる．引数として値を与えることで半径や色の透明度，グラデーション，ぼかしの程度が設定できるほか，前述の LayerControl と組み合わせることで表示の切り替えも行うことができる．

大量のマーカーをまとめて表示

前述の folium.Marker メソッドでは，1 つのマーカーに対して多くの情報を付与する方法について解説したが，Plugins.MarkerCluster メソッドでは，大量のマーカーを点として地図上にプロットし，一定の閾値を定めることで地図の縮尺によってその付近にあるマーカーを 1 つのマーカーとして表示することができる．これによって，大量のマーカーを一度にプロットした際でも見やすく情報を提供することができる．また，このような機能はマーカーの密度という観点でヒートマップのような表し方と考えることもできる．

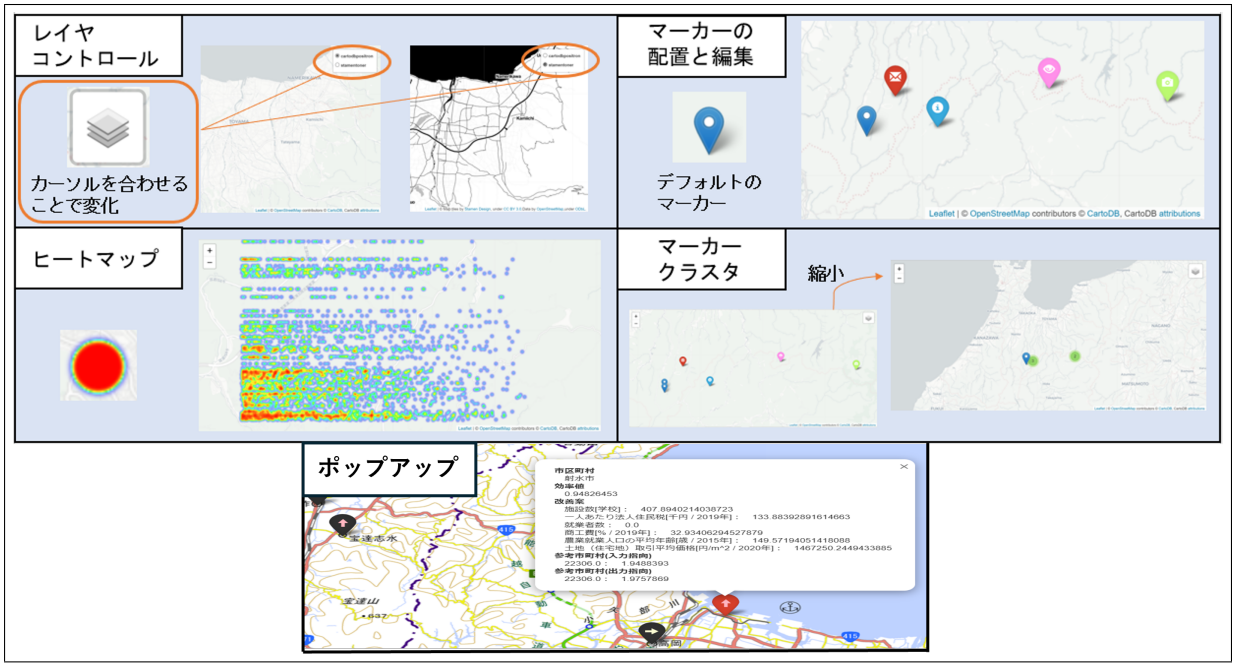

図 3.3: Folium による Web-GIS 実装例

ポップアップ機能の追加と応用

Folium では, 地図上に配置したマーカーやレイヤに対してポップアップ機能を追加することができる. ポップアップとは, 指定した要素をクリックした際に表示される小さなウィンドウであり, 利用者に追加情報を提供する際に有用である. ポップアップに表示できる内容はテキストだけに限らず, 画像や HTML 形式のコンテンツも含めることができる. また,GeoJSON 形式のデータと組み合わせることで, 複数の地点やエリアに対して統一的にポップアップを付与することも可能である. これにより, 広範囲にわたる地理データを効率的に可視化しつつ, 詳細情報を個別に提供できる. さらに, ポップアップはカスタマイズ性が高く, フォントや背景色の調整, リンクやボタンの埋め込みを行うことで, 利用者にとって直感的かつ実用的なインターフェースを提供することができる. この機能を活用することで, 地図上の情報提供がより充実し, 利用者にとって有益なデータ閲覧体験を実現できる.

以上で解説した各種機能の実装例については図 3.3 にて示す. また, 以上のような機能のほかにも, Folium によって実現可能な Web-GIS の機能は多くあるが, 本研究では特にマーカーによる情報のフィードバックおよび重ね合わせを中心に行っていく. システムの詳細については 4 章にて示す.

第4章

提案手法

§ 4.1　データセットの作成

本研究では，分析する対象地域を富山市とし，実際に取引が行われた不動産情報から，不動産価格に影響を与える要因を分析する．不動産価格要因を分析する際に，必要なデータセットを作成するまでの流れを図 4.1 に示す.

一物四価

不動産の価格には「一物四価」と呼ばれる異なる 4 つの価格が存在する．すなわち，公示地価，基準地価，固定資産税評価額，実勢価格である．これらのうち，公示地価は国土交通省によって毎年公表される指標であり，地価の一般的な動向を示す重要な指標として利用されてきた．しかしながら，公示地価はあくまで評価額であり，実際の取引における価格（実勢価格）とは異なる可能性がある．

近年のヘドニック・アプローチによる研究では，公示地価を用いた分析が行われてきた．[25] [26] 公示地価に影響を与える要因を分析することにより，地価の形成要因を明らかにしようとする試みがなされている．しかし，公示地価は市場での実際の取引価格を直接反映しているわけではないため，実勢価格に基づいた分析を行うことが重要であると考えられる.

異なる空間的解像度のデータの結合

e-Stat で提供されている統計データは，集計されている区分ごとに，全国ごと，都道府県ごと，市区町村ごと，”…丁目”といったの小地域ごと，グリッドセルごとの５種類が存在する．本システムでは，予測する空間的な単位をグリッドセルとしているため，グリッドセルごとの統計データを使用するが，より考慮できる要因を増やすため，小地域ごとの統計データも使用できるようにした．小地域ごとのデータはそのまま用いることはできないため，グリッドセル単位に変換する必要がある．そのため，小地域ごとのデータについて，小地域全体に均等に分布していると仮定し，対象のグリッドセルに重なっている割合だけを足し合わせる．すなわち，対象のグリッドセル C に重なる小地域 $A_1, A_2, \cdots, A_n$ について，それぞれの全体の面積を $S_1, S_2, \cdots, S_n$，対象のグリッドセルと重なる面積を $s_1, s_2, \cdots, s_n$，データ値を $x_1, x_2, \cdots, x_n$ とすると，対象のグリッドセル C のデータ値 X を次のように算出する．[27]

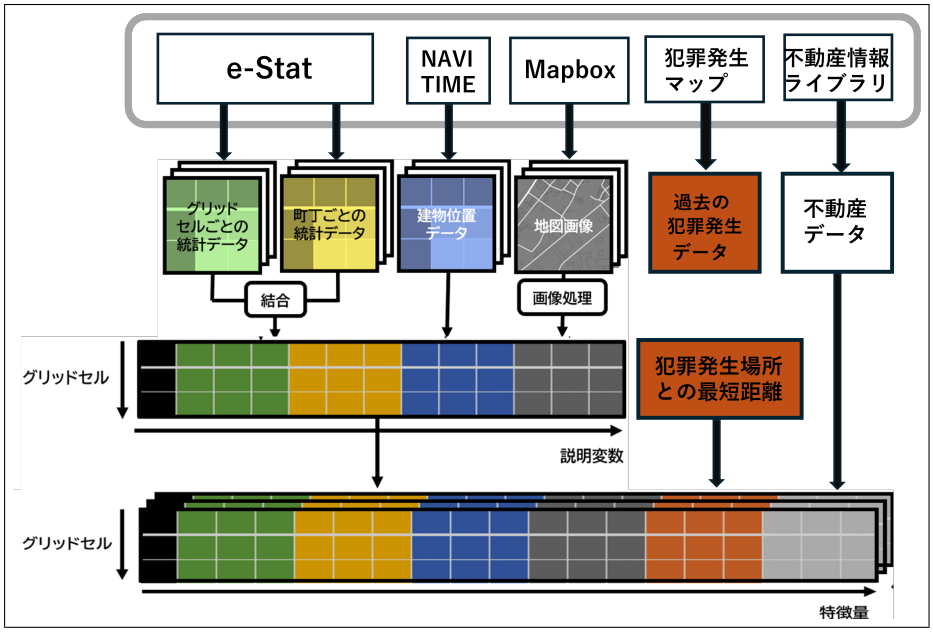

図 4.1: データセットを作成するまでの流れ

$$X = \sum_{k=1}^{n} x_k \frac{s_k}{S_k} \tag{4.1}$$

統計データとして公開されることの多い要素のなかには，不動産価格に影響を与えるものが多く存在し，それらが豊富に公開されている e-Stat から，ドメイン知識をもとに自由に説明変数を選択できるようにしたことは，大きな利点と考える．

施設の最短距離と立地数の算出

さまざまなジャンルの施設について，NAVITIME からスクレイピングを行い，施設名と，その緯度と経度を取得する．本システムでは，それぞれのジャンルごとに，対象のグリッドセルに含まれる数と，最も近くにある施設までの距離を説明変数とする．

なお，地球は楕円体であるため，単純なユークリッド距離では誤差が生じてしまう．そこで，本システムではヒュベニの公式 [28] を用いて，距離を算出している．対象のグリッドセルの中心を $P_o(x_o, y_o)$，注目する施設を $P_n(x_n, y_n)$ とすると，それら 2 点間の距離 D は以下で求まる．

$$D = \sqrt{(D_y M)^2 + (D_x N cosP)^2} \tag{4.2}$$

$$M = \frac{R_x(1 - E^2)}{W^3} \tag{4.3}$$

$$N = \frac{R_x}{W} \tag{4.4}$$

$$W = \sqrt{1 - E^2 sinP^2} \tag{4.5}$$

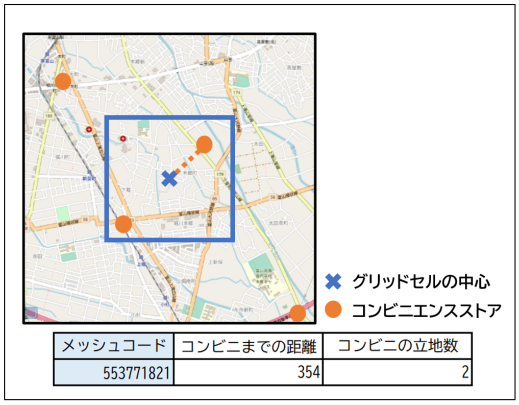

図 4.2: 施設データに基づく説明変数

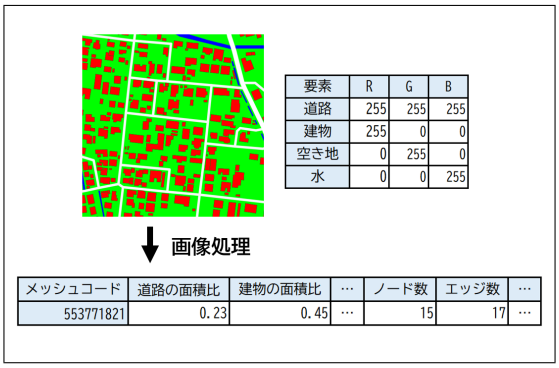

図 4.3: 地図画像に基づく説明変数

$$E = \sqrt{\frac{R_x^2 - R_y^2}{R_x^2}} \tag{4.6}$$

これにより，不動産価格に影響を与える立地要因の影響を精度高く分析することが可能となる．

地図画像にもとづく説明変数の抽出

Mapbox Static Tile API を利用して，それぞれのメッシュに対応する地図画像を取得する．本研究では，建物，道路，水，空き地の 4 つを色分けした地図画像を取得し，それぞれの画像の大きさに対する面積の比率を説明変数としている．これにより，各メッシュ内の土地利用状況を数値化し，不動産価格への影響を分析する．

対象の要素の面積比率 p_a は，地図画像の大きさを $n \times m$，その要素と同一の RGB 値をもつピクセル数を x とすると，以下のように算出する．

$$p_a = \frac{x}{nm} \tag{4.7}$$

なお，Mapbox Static Tile API によって取得する地図画像は，本来は画像処理を目的としていない．そのため，Mapbox Studio 上で指定した RGB 値と誤差があるピクセルがある．そのため，対象のピクセルの RGB 値と，それぞれの要素の RGB 値とのユークリッド距離を算出し，最も小さい要素を指定する．

また，地図画像から道路ネットワークを抽出し，道路に関連する属性を説明変数として抽出する．まず，道路とそれ以外の 2 値画像に変換し，ノイズを削除するためにオープニング処理を行う．その画像に対して，ネットワークを抽出するアルゴリズム [29] を使用し，ネットワークの属性であるノード数 N，エッジ数 E を取得する．また，それらから密度 d，平均次数 k を以下のとおり算出する．

$$d = \frac{2E}{N(N-1)} \tag{4.8}$$

$$k = \frac{2E}{N} \tag{4.9}$$

表 4.1: ダミー変数化の基準

ダミー変数（質的変数）	最寄り駅名_富山	最寄り駅名が富山なら1，そうでないならば0
	都市計画_市街化調整区域	不動産が市街化調整区域内なら1，そうでないならば0
	土地の形状_不整形	土地の形状が不整形なら1，そうでないならば0
	建物の構造_木造	建物の構造が木造なら1，そうでないならば0
	前面道路：種類のダミー	前面道路の種類を国道・市道・私道・県道に分けてダミー変数化
	前面道路：方位のダミー	前面道路の方位を8つに分けてダミー変数化（北・北東・東・南東・南・南西・西）

なお，密度 d と平均次数 k は，道路のネットワークとしてみたとき，それぞれ次のような特徴をもつ [30]．密度 d が大きい道路ネットワークは，道路が網目状に相互に接続された状態であり，幅員の狭い生活道路であると考えられる．また，平均次数 k が小さい道路ネットワークは，交差点の少ない直線的な道路が多いと考えられる．

以上により，不動産価格形成要因の分析に用いるデータセットの作成が完了する．

§ 4.2　データの前処理と変数選択

作成したデータセットに対して、前処理を行い、多量の説明変数から AEN によって選択を行う．

前処理は以下のような流れで行う．

1 欠損値を含む行は分析の精度を確保するために削除する．

2 外れ値を含む行も同様に削除し，データの信頼性を向上させる．

3 最寄り駅名，土地の形状，建物の構造，前面道路の方位，前面道路の種類，都市計画の 6 つのカテゴリ変数に対し，ワンホットエンコーディングを適用する．ただし，落とし込みを防ぐために基準カテゴリを削除する．ダミー変数化の基準は表 4.1 で示す．

4 連続変数の非線形性を考慮し，二次項を作成する．

5 変数間の相互作用を考慮し，連続変数×連続変数，連続変数×ダミー変数，ダミー変数×ダミー変数の交互作用項を作成する．これにより，説明変数は合計で 2329 個となる．

6 すべての連続変数に対し，Z スコアによる標準化を施し，スケールの影響を排除する．

これでデータセットに対する前処理とする．次に，多量の説明変数からの変数選択を行う．その手法として AEN を用いる．

AEN による説明変数の選択

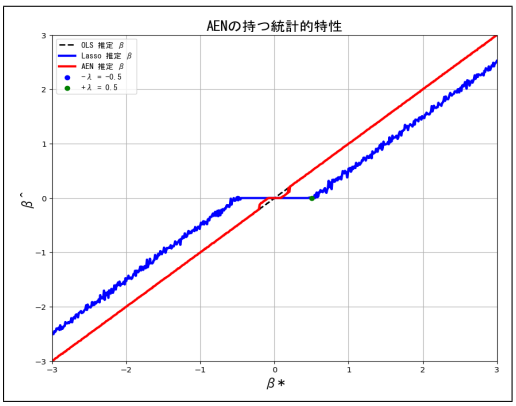

図 4.4: AEN の持つ統計的特性

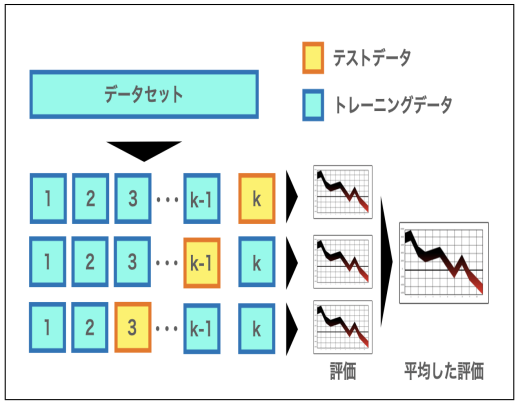

図 4.5: k 分割交差検証のイメージ図 [31]

スパース推定の中でもオラクル性を満たす手法の一つとして，AEN がある．AEN は，Elastic Net に適応ウエイトを加えることで，変数選択の一致性と非ゼロ係数の漸近正規性の両方を満たす手法である．

$$\hat{\boldsymbol{\beta}}^{1\mathrm{st}} = \left(1 + \frac{\lambda_2}{n}\right)\left\{\arg\min_{\boldsymbol{\beta}}\left(\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_2 \sum_{k\geq j\geq 0} \beta_{jk}^2 + \lambda_1 \sum_{k\geq j\geq 0} |\beta_{jk}|\right)\right\} \tag{4.10}$$

$$\hat{w}_{jk} = \left(\left|\hat{\beta}_{jk}^{1\mathrm{st}}\right|\right)^{-\gamma} \tag{4.11}$$

$$\hat{\boldsymbol{\beta}} = \left(1 + \frac{\lambda_2}{n}\right)\left\{\arg\min_{\boldsymbol{\beta}}\left(\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_2 \sum_{k\geq j\geq 0} \beta_{jk}^2 + \lambda_1^* \sum_{k\geq j\geq 0} \hat{w}_{jk}|\beta_{jk}|\right)\right\} \tag{4.12}$$

AEN の推計は，二段階で行われる．まず，予備推計として EN で係数を推定する．そのうえで，係数の絶対値が小さい変数に対してより強い罰則を与えるよう L1 ノルムの正則化項を変数ごとに調整したうえで，改めて EN を実施する．このように二段階推計を行うことで，データセットの性質に殆ど依存することなくオラクル性を得ることができる．

AEN がオラクル性を満たす理由

図 4.4 では，AEN がオラクル性を満たす理由について直観的に示している．ここでは，説明変数の行列 $\mathbf{X}$ と撹乱項のベクトル $\boldsymbol{\varepsilon}$ を人工的に生成し，真のモデル $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ に基づいて被説明変数のベクトル $\mathbf{Y}$ を算出した．そのうえで，$\mathbf{Y}$ と $\mathbf{X}$ を観測値としたとき，OLS，Lasso，AEN が $\boldsymbol{\beta}$ をどのように推定するかを確認している．図中の横軸は真の係数 $\boldsymbol{\beta}^*$ を表しており，縦軸には $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$，$\hat{\boldsymbol{\beta}}_{\mathrm{Lasso}}$、$\hat{\boldsymbol{\beta}}_{\mathrm{AEN}}$ をプロットしている．

まず，Lasso を見てみると，$|\boldsymbol{\beta}^*| < \lambda$ のとき，$\hat{\boldsymbol{\beta}} = 0$ となり，スパース性を有していることがわかる．一方で，$|\boldsymbol{\beta}^*| \geq \lambda$ のとき，$\hat{\boldsymbol{\beta}}$ は真の値である $\boldsymbol{\beta}^*$ と比較して，絶対値で見て λ だけ小さく推定されていることがわかる．つまり，正則化パラメータ λ とオラクル性の条件との関係は，λ が大きくなるほど，①ゼロ係数を推定しやすくなり変数選択の一致性を

満たしやすくなる一方，②推定値が絶対値で見てλだけ小さくなり非ゼロ係数の漸近正規性を満たしにくくなるというトレードオフの関係となっていることがわかる．

これに対して AEN を見てみると，$|\boldsymbol{\beta}^*|$ が小さい場合，一段階目の推計における係数の小ささを反映して罰則が強く与えられることで $\hat{\boldsymbol{\beta}} = 0$ が導かれる．一方で，$|\boldsymbol{\beta}^*|$ が大きい場合には，罰則があまり与えられない結果として，$\hat{\boldsymbol{\beta}}$ は $\boldsymbol{\beta}^*$ に漸近していく関係となっている．このように，罰則の大きさを一段階目の推計値に応じて調整することで，係数が小さいときはゼロが推定されやすくなると同時に，係数が大きいときは絶対値で見た推計値の縮小幅が最小限に抑えられ，オラクル性の二条件が同時に満たされやすくなることがわかる．

k 分割交差検証によるハイパーパラメータの探索

AEN で推計を行う際に，適切なハイパーパラメータの選定が不可欠である．特に，ElasticNet 正則化における λ_1，λ_2，および二段階目の ElasticNet で使用される L1 ノルムの正則化項の係数 λ_1^* の値は，モデルの精度と変数選択に大きな影響を与える．そのため，最適なハイパーパラメータを探索するために，k 分割交差検証を用いることが一般的である．

k 分割交差検証は，データセットを k 個の部分集合に分割し，1 つを検証データ，残りの k-1 個を訓練データとして使用する方法である．このプロセスを k 回繰り返すことで，すべてのデータが一度は検証データとして使用され，モデルの汎化性能を高精度に評価することができる．各回の結果を平均化することで，モデルの過学習を防ぎ，最も適切なハイパーパラメータの設定を選定することが可能となる．

AEN モデルのチューニングでは，以下の 3 つのハイパーパラメータを探索する：

- λ_1：1 段階目の L1 正則化項の係数
- λ_2：1 段階目の L2 正則化項の係数
- λ_1^*：二段階目の ElasticNet で使用される L1 ノルムの正則化項の係数

1 段階目のハイパーパラメータは以下の式で定義される：

$$\lambda_1 = \alpha_1 \cdot l1\_ratio_1 \tag{4.13}$$

$$\lambda_2 = \alpha_1 \cdot (1 - l1\_ratio_1) \tag{4.14}$$

2 段階目のハイパーパラメータは以下のように定義される：

$$\lambda_1^* = \alpha_2 \cdot l1\_ratio_2 \tag{4.15}$$

- 1 段階目のハイパーパラメータ（λ_1, λ_2）の探索範囲：$\alpha_1 \in [10^{-5}, 10^0]$（対数スケールで 25 分割），$l1\_ratio_1 \in [0.0001, 1.0]$（線形スケールで 25 分割）
- 2 段階目のハイパーパラメータ（λ_1^*）の探索範囲：$\alpha_2 \in [10^{-5}, 10^0]$（対数スケールで 25 分割），$l1\_ratio_2 \in [0.0001, 1.0]$（線形スケールで 25 分割）

これらの式に従って，α_1，$l1\_ratio_1$（1 段階目）および α_2，$l1\_ratio_2$（2 段階目）の組み合わせを探索し，最適な λ_1，λ_2，λ_1^* を決定する．

k 分割交差検証を実行し，各組み合わせのパフォーマンスを評価した後，最も低い平均誤差を持つ λ_1，λ_2，および λ_1^* のペアが最適な設定とされる．

このようにして得られた最適なハイパーパラメータは，AEN モデルの性能を最大化し，過剰適合を防ぐために非常に重要である．最適なパラメータ設定を見つけることによって，モデルは訓練データに対する過剰な適合を避け，未知のデータに対しても高い汎化能力を持つようになる．さらに，交差検証を用いることで，モデルの性能を異なるデータ分割に基づいて評価することができる．このプロセスにより，データセットの特定の分割に依存せず，安定した予測性能を確認できるため，モデルが様々なデータの変動に対して適切に対応できることが確認できる．これにより，最適なハイパーパラメータが得られると同時に，モデルの信頼性や適用範囲の広さも向上することが期待される．

以上の手順により，データの前処理を行い，k 分割交差検証を用いて最適なハイパーパラメータを探索した．その結果得られたハイパーパラメータを用いた AEN による推計を行い，係数が非ゼロとなった変数を説明変数として選択した．最終的に，これらの変数を用いたデータセットを基に，不動産価格の形成要因を分析する．

§ 4.3　不動産価格形成要因の分析と GIS による可視化

選択された説明変数を用いて，ヘドニック・アプローチによる不動産価格形成要因の分析を行い，Web-GIS への可視化を行う．

不動産価格形成要因の分析

従来のヘドニック法に加え，一次項のみならず二次項および交互作用項も回帰式に含めた分析を行う．これにより，非線形効果や複数の要因が相互に影響を与える場合の価格形成要因をより精緻に評価することが可能となる．具体的な回帰式は以下のように表される．

$$Y_i = \hat{\Theta}_0 + \sum_{j=1}^{p} \hat{\Theta}_{0j} x_{j,i} + \sum_{j=1}^{p} \hat{\Theta}_{jj} x_{j,i}^2 + \sum_{k>j\geq 1} \hat{\Theta}_{jk} x_{j,i} x_{k,i} \tag{4.16}$$

次に，回帰係数が有意であるかどうかを検定するために，t 値，p 値，および決定係数（R^2）を用いる．回帰係数 θ_j の t 値は以下の式で計算される．

$$t_j = \frac{\hat{\theta}_j}{SE(\hat{\theta}_j)} \tag{4.17}$$

t 値が大きいほど，回帰係数が有意である可能性が高くなる．次に，p 値は t 値に基づいて計算され，通常，p 値が 0.05 以下であれば回帰係数が有意であるとみなされる．

$$p_j = P(|t_j| > t_{\alpha/2,df}) \tag{4.18}$$

さらに、以下の有意水準に基づく検定が行われる：

	変数	回帰係数	t値	p値	有意性
2	定数項	75768.14131	91.15533845	0	1% 有意
3	最寄駅：距離（分）	1676.163248	1.662002028	0.096593609	10% 有意
4	面積（㎡）	-14398.76034	-16.883913	9.27E-62	1% 有意
5	犯罪発生率(件/25000)	-2562.00744	-2.767919864	0.005668454	1% 有意
6	人口	11281.18954	11.44427778	7.57E-30	1% 有意
7	コンビニ^2	1949.01754	1.630256234	0.103128984	有意でない
8	スーパー^2	-7700.017504	-6.127862154	9.80E-10	1% 有意
9	前面道路：幅員（m）×小学校	6629.483392	6.231240595	5.12E-10	1% 有意
10	前面道路：幅員（m）×中学校	3148.34509	2.901659999	0.003732926	1% 有意
11	前面道路：幅員（m）×高校	-7769.916257	-6.732683059	1.91E-11	1% 有意
12	駐車場×道路の面積比率	-2673.442274	-1.883404113	0.05972078	10% 有意
13	駐車場×建物の面積比率	-2710.711273	-2.076722578	0.037893208	5% 有意
14	駐車場×水の面積比率	-1221.972873	-1.415455536	0.157015692	有意でない

図 4.6: 分析結果の例

- **1%有意水準**（$\alpha = 0.01$）：p 値が 0.01 未満であれば、回帰係数は非常に有意とみなされる。つまり、帰無仮説が正しい場合、1%の確率で誤った結論を下す可能性を認めるという非常に厳格な基準である。
- **5%有意水準**（$\alpha = 0.05$）：p 値が 0.05 未満であれば、回帰係数が有意であるとみなされる。最も一般的に使用される基準であり、通常の統計分析において標準的な基準として広く採用されている。
- **10%有意水準**（$\alpha = 0.10$）：p 値が 0.10 未満であれば、回帰係数が有意であるとみなされる。これは比較的寛容な基準であり、初期の探索的な分析や追加的な証拠がない場合に使用されることがある。

また，回帰モデルの決定係数 R^2 は，モデルがどれだけ目的変数 Y の分散を説明しているかを示す指標であり，次の式で計算される．

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{4.19}$$

R^2 が高いほど，モデルはデータに適合していることを示す．

一般的に，R^2 の値が 0.7 以上 であれば，モデルはデータに良好に適合していると見なされる．0.5 から 0.7 の範囲では，モデルはある程度の適合度を示すが，追加的な要因を考慮することが望ましい．0.5 未満 の場合，モデルの説明力が低いとされ，再評価や変数の見直しが必要となることが多い．

これらの統計量（t 値，p 値，決定係数）を用いて回帰係数の有意性を検定し，モデルの適合度を評価することによって，価格形成要因に対する影響を正確に把握し，信頼性の高い分析を実施することができる．

不動産価格形成要因の可視化

Folium を用いた Web-GIS におけるベースマップは cartodbpositron および地理院タイルの 2 種類であり，cartodbpositron は白色でマーカーが視認しやすいという理由，地理院タイルは等高線や色分けによって地形が表されており，土地の地理的状況が視覚的に理解で

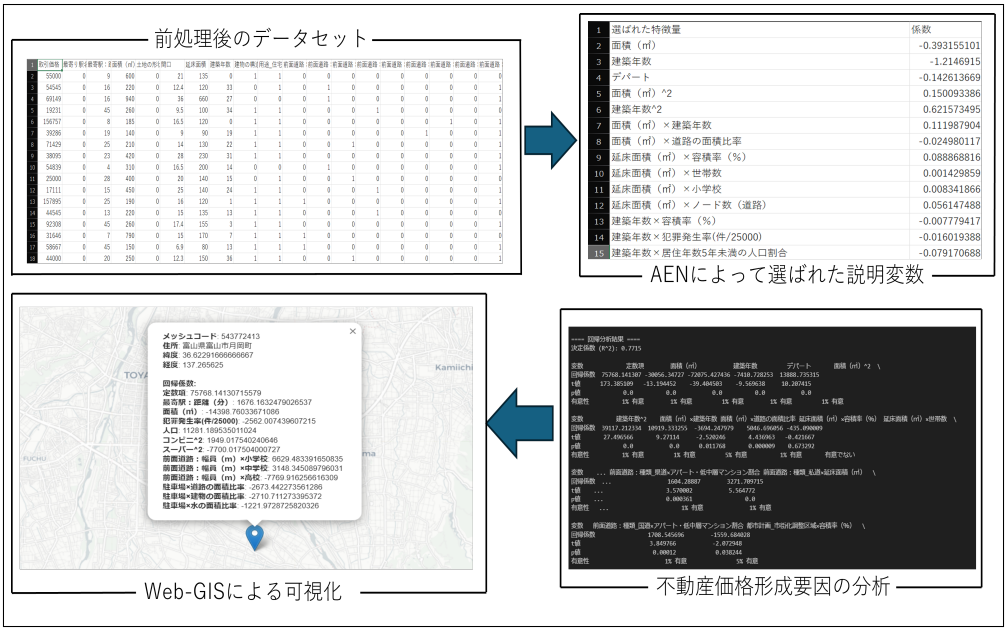

図 4.7: Web-GIS を作成する流れ

きるという理由で採用した．Web-GIS 上で、各メッシュの中心にマーカーをプロットし、ポップアップには以下の情報を表示することを考えている：

- 町までの住所
- 緯度，経度
- メッシュコード
- 不動産価格を形成する要因とその回帰係数

これにより，各メッシュに関連する情報を視覚的に表示することができ，ユーザーは地図上で直接情報を確認できるようになる．ポップアップを活用することで，ユーザーは特定のメッシュに関する詳細なデータに簡単にアクセスでき，町までの住所や緯度・経度といった基本情報から，不動産価格を形成する要因とその回帰係数に至るまで，重要な分析結果を迅速に把握することができる．このように，地理的情報と分析結果を組み合わせることで，データの解釈が容易になり，地域ごとの特性をより深く理解できるようになる．

第5章

数値実験並びに考察

§ 5.1　数値実験の概要

本章では，実際の不動産取引データを用いて，4 章で述べた手法で不動産価格形成要因を分析し，その精度を確認，および考察を行う．また，その分析を用いて，要因を可視化したマップを作成する．

不動産取引データ

今回の数値実験で使用する不動産取引データは，国土交通省が公開している不動産情報ライブラリ [32] から取得したものを用いる．不動産取引データには，住所，経度，緯度，最寄駅の名称および最寄駅までの距離（分），取引価格（総額），面積（m^2），取引価格（m^2 単価），土地の形状，間口，延床面積（m^2），建築年，建物の構造，用途，前面道路の方位，種類，幅員（m），都市計画，建ぺい率（%），容積率（%），および取引時点が含まれている．また，今回使用するレコードは，2007 年から 2023 年第 2 四半期とし，欠損値や外れ値がある行を除外した 3872 サンプルである．

データセット

分析に用いる説明変数は，表 5.1 に示す計 67 個とした．さらに，そこから二次項と交互作用項の作成を行うと，連続変数の二次項が 51 個，連続変数同士の交互作用項が 1,275 個，連続変数とダミー変数の交互作用項が 816 個，ダミー変数同士の交互作用項が 120 個生成され，合計で 2,262 個の新たな変数が追加された．したがって，最終的に分析に用いる説明変数の総数は，元の一次項 67 個と合わせて 2,329 個となる．ここで，連続変数間の相関係数を図 5.1 に示す．

§ 5.2　実験結果と考察

予測モデルの精度検証

検証用のデータを用いて，作成した犯罪発生予測モデルの精度を検証した結果を表 5.3 に，同地域における複数日の予測結果を図??に示す．犯罪が発生したグリッドセル，発生しなかったグリッドセルともに，正しく予測した確率（正解率）は約 0.970 であったが，発生したグリッドセルを，発生すると予測した確率（再現率）は約 0.318，発生すると予測したグ

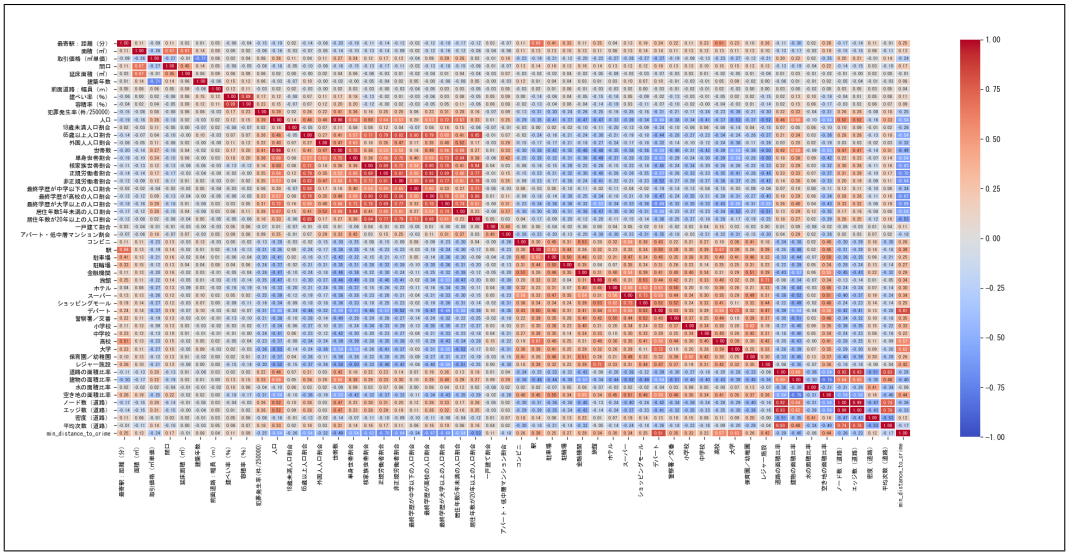

図 5.1: 連続変数間の相関係数

リッドセルで，実際に発生した確率（適合率）は約 0.018 であった．すなわち，犯罪が発生しないグリッドセルは比較的正しく予測できているものの，犯罪が発生しているグリッドセルについては，それに多少のランダム性を持っていたとしても，実用的な精度であるとは到底いえない結果となった．

この理由として，図??で分かるように，この 2 日間で犯罪が発生すると予測したグリッドセルが変化していない．表 5.1 のうち，灰色で着色した説明変数は，Boruta によって選択されたものであることを示しているが，これから分かるように，1 日ごとに変化する説明変数のうち，選択されたものは「過去 1 か月間の犯罪発生件数」のみであった．このため，そのような静的データに対して，動的データが予測値に寄与する絶対量が小さくなり，この予測モデルは，1 日ごとに予測値が変化しにくい可能性が考えられる．

短期的リスク，特に近接反復という犯罪の特性は，大きな犯罪発生の要因となり得る．そのため，静的データと動的データと分けて，それぞれに対して予測モデルを構築し，前者の予測モデルの予測値に対して重みづけを行うことによって，1 日ごとに予測値を大きく変化させることによって，予測精度が改善する可能性がある．

また，図??に示している地域に含まれるグリッドセルは約 550 個であり，実際に犯罪が発生したグリッドセルは 3〜5 個である．すなわち，その割合は 0.01 を下回る．本研究では，適切なアプローチを行い，不均衡なデータであっても，時空間的に解像度の大きい予測を行うことを目指したが，今回の犯罪発生データは，その限界を超えており，それでもなお精度の向上が見込めなかった可能性がある．

そのため，この改善案として，犯罪が発生する例を異常な例として，異常検知問題として取り扱うことが挙げられる．異常検知問題とは，検知したい異常な例がかなり少ないか，まったくないときに，正常な例のみを用いて，異常な例か正常な例かを判断することである．異常検知問題を取り扱うために用いるアルゴリズムは，異常な例を必要としないため，極端な不均衡，もしくは，まったく例がないデータを前提としていることである．そのため，犯罪が発生する例を異常な例として，異常検知問題として予測を行うことで，精度の向上が期待できるだろう．

表 5.1: 説明変数の一次項の候補一覧

最寄り駅名_富山	前面道路：種類_国道	最終学歴が中学以下の人口割合	駅（最短距離）
最寄駅：距離	前面道路：種類_市道	最終学歴が高校の人口割合	小学校（最短距離）
面積	前面道路：種類_私道	最終学歴が大学以上の人口割合	中学校（最短距離）
土地の形状_不整形	前面道路：種類_県道	居住年数5年未満の人口割合	高校（最短距離）
間口	都市計画_市街化調整区域	居住年数が20年以上の人口割合	大学（最短距離）
延床面積	建ぺい率	一戸建て割合	
建築年数	容積率	アパート・低中層マンション割合	ホテル（最短距離）
建物の構造_木造	犯罪発生率	ショッピングモール（最短距離）	道路の面積比率
用途_住宅	人口	警察署／交番（最短距離）	建物の面積比率
前面道路：方位_北	18歳未満人口割合	駐車場（最短距離）	水の面積比率
前面道路：方位_北東	65歳以上人口割合	駐輪場（最短距離）	空き地の面積比率
前面道路：方位_南	外国人人口割合	金融機関（最短距離）	ノード数（道路）
前面道路：方位_南東	世帯数	保育園／幼稚園（最短距離）	エッジ数（道路）
前面道路：方位_南西	単身世帯割合	レジャー施設（最短距離）	密度（道路）
前面道路：方位_東	核家族世帯割合	スーパー（最短距離）	平均次数（道路）
前面道路：方位_西	正規労働者割合	コンビニ（最短距離）	平均次数（道路）
前面道路：幅員	非正規労働者割合	デパート（最短距離）	min_distance_to_crime

犯罪発生要因の可視化

次に，学習用データを用いて，予測モデルに SHAP を適用することにより，予測モデルを可視化した結果を，図??に示す．4.3 節で述べたように，それぞれのグリッドセルに描画されている色は，各説明変数（なお，長期的リスクのみ）のもつ SHAP 値の合計を示しており，0 を基準に，大きくなるほど濃い赤色に，小さくなるほど濃い青色となっている．長期的リスクのみを抽出しているため，SHAP 値の合計は，そのグリッドセルの潜在的な犯罪発生リスクと捉えることができるだろう．

まず，学習用データを入力したときの予測モデルの精度を表??に示す．学習用データは，犯罪が発生する例としない例がほとんど同数となるようにアンダーサンプリングを行っていること，予測モデルを構築するときに使用したため，再現率は約 0.842 と大きかった．

また，図??において，赤枠で示したグリッドセルの犯罪発生要因を可視化した結果を，一例として図??に示す．これらのグリッドセルは，隣接しているのにもかかわらず，右のグリッドセルの SHAP 値の合計，すなわち犯罪が発生するリスクの合計は，左のグリッドセルの約 3.45 倍であると算出された．その内訳を確認すると，右のグリッドセルでは，左のグリッドセルと比較して，特に「世帯数」,「駐車場（最短距離）」,「駐輪場（最短距離）」の SHAP 値が増加しており，その差は，それぞれ 0.37，0.3，0.32 であった．すなわち，右のグリッドセルは，左のグリッドセルと比べて，世帯数が多いこと，また，駐車場，駐輪場が近い（もしくは，そのグリッドセル内にある）ことが犯罪の発生に寄与している，という知見を得ることができるだろう．

なお，本研究では，データセットとして使用した，または可視化された説明変数を「要因」と仮定し分析を行ったが，あくまでも予測値に対する説明変数の「関係性」を示して

表 5.2: 検証用データによる予測結果

選ばれた説明変数	係数	選ばれた説明変数	係数
面積（㎡）	-0.393155102	最寄り駅名_富山×建築年数	-0.087846746
建築年数	-1.214691499	最寄り駅名_富山×18歳未満人口割合	0.001648931
デパート	-0.142613669	最寄り駅名_富山×ホテル	-0.000641945
面積（㎡）^2	0.150093386	最寄り駅名_富山×スーパー	0.022251487
建築年数^2	0.621573495	最寄り駅名_富山×建物の面積比率	0.071468663
面積（㎡）×建築年数	0.111987903	土地の形状_不整形×平均次数（道路）	-0.006620294
面積（㎡）×道路の面積比率	-0.024980115	建物の構造_木造×単身世帯割合	-0.036652893
延床面積（㎡）×容積率（%）	0.088868817	用途_住宅×面積（㎡）	-0.004510996
延床面積（㎡）×世帯数	0.00142986	用途_住宅×エッジ数（道路）	0.030296567
延床面積（㎡）×小学校	0.008341866	前面道路：方位_北×単身世帯割合	0.007814123
延床面積（㎡）×ノード数（道路）	0.056147485	前面道路：方位_南×世帯数	0.010202565
建築年数×容積率（%）	-0.007779418	前面道路：方位_南×アパート・低中層マンション割合	0.006518005
建築年数×犯罪発生率(件/250000)	-0.016019388	前面道路：方位_南×密度（道路）	0.002371489
建築年数×居住年数5年未満の人口割合	-0.079170688	前面道路：方位_南×min_distance_to_crime	0.00006655
18歳未満人口割合×金融機関	0.007023098	前面道路：種類_県道×アパート・低中層マンション割合	0.019030997
18歳未満人口割合×水の面積比率	0.004618301	前面道路：種類_私道×延床面積（㎡）	0.031765495
世帯数×居住年数5年未満の人口割合	0.025456813	前面道路：種類_国道×アパート・低中層マンション割合	0.014563587
世帯数×デパート	-0.024713889	都市計画_市街化調整区域×容積率（%）	-0.042752176
単身世帯割合×ショッピングモール	0.015399415	都市計画_市街化調整区域×居住年数が20年以上の人口割合	-0.004820905
正規労働者割合×高校	-0.035466746	最寄り駅名_富山×用途_住宅	0.034125269
最終学歴が大学以上の人口割合×居住年数5年未満の人口割合	0.120149306	最寄り駅名_富山×前面道路：方位_南東	0.001299833
居住年数5年未満の人口割合×高校	-0.008516534	最寄り駅名_富山×前面道路：種類_市道	0.046550904
一戸建て割合×警察署／交番	-0.004852252	建物の構造_木造×前面道路：種類_私道	-0.024858249
アパート・低中層マンション割合×min_distance_to_crime	0.012657824	前面道路：方位_南東×前面道路：種類_市道	0.003533599
ホテル×道路の面積比率	-0.021506799		

表 5.3: 検証用データによる予測結果

変数	回帰係数	t値	p値	有意性	変数	回帰係数	t値	p値	有意性
定数項	75768.14131	173.3851089	0	1% 有意	ホテル×道路の面積比率	-1245.129657	-2.06116926	0.039354408	5% 有意
面積（㎡）	-30056.34727	-13.19445228	6.62E-39	1% 有意	最寄り駅名_富山×建築年数	-6516.240316	-8.1472518	4.99E-16	1% 有意
建築年数	-72075.42744	-39.40450256	2.75E-285	1% 有意	最寄り駅名_富山×18歳未満人口割合	1648.832446	1.337979771	0.180982651	有意でない
デパート	-7410.728253	-9.56963807	1.86E-21	1% 有意	最寄り駅名_富山×ホテル	-3534.913745	-4.60382123	4.28E-06	1% 有意
面積（㎡）^2	13888.73531	10.20741475	3.74E-24	1% 有意	最寄り駅名_富山×スーパー	2739.34915	2.601589555	0.009315085	1% 有意
建築年数^2	39117.21233	27.49656567	5.01E-152	1% 有意	最寄り駅名_富山×建物の面積比率	3385.830703	2.707705686	0.006805167	1% 有意
面積（㎡）×建築年数	10919.33326	9.271139987	3.00E-20	1% 有意	土地の形状_不整形×平均次数（道路）	-1331.555222	-2.966052524	0.003035169	1% 有意
面積（㎡）×道路の面積比率	-3694.247979	-2.520245695	0.011767713	5% 有意	建物の構造_木造×単身世帯割合	-4465.517146	-6.755629203	1.64E-11	1% 有意
延床面積（㎡）×容積率（%）	5046.696056	4.436962976	9.38E-06	1% 有意	用途_住宅×面積（㎡）	-455.8098027	-0.73861672	0.460185133	有意でない
延床面積（㎡）×世帯数	-435.0900994	-0.421667246	0.673291635	有意でない	用途_住宅×エッジ数（道路）	2366.093438	3.209214634	0.001341991	1% 有意
延床面積（㎡）×小学校	1173.771797	1.589277134	0.112080571	有意でない	前面道路：方位_北×単身世帯割合	1443.127295	3.085057392	0.002049684	1% 有意
延床面積（㎡）×ノード数（道路）	5287.120057	3.717395307	0.000204172	1% 有意	前面道路：方位_南×世帯数	1532.231262	2.644650411	0.00821099	1% 有意
建築年数×容積率（%）	-3184.050708	-2.60092905	0.009333006	1% 有意	前面道路：方位_南×アパート・低中層マンション割合	719.8363042	1.257209283	0.208754701	有意でない
建築年数×犯罪発生率(件/25000)	-2041.890012	-3.805401357	0.000143791	1% 有意	前面道路：方位_南×密度（道路）	507.3301136	0.996296967	0.319168964	有意でない
建築年数×居住年数5年未満の人口割合	-3861.404861	-4.129784092	3.71E-05	1% 有意	前面道路：方位_南×min_distance_to_crime	953.9434148	2.009767019	0.044526026	5% 有意
18歳未満人口割合×金融機関	1291.852612	2.483465457	0.013053667	5% 有意	前面道路：種類_県道×アパート・低中層マンション割合	1604.28887	3.570001548	0.000361372	1% 有意
18歳未満人口割合×水の面積比率	618.0756973	1.104554865	0.269422067	有意でない	前面道路：種類_私道×延床面積（㎡）	3271.709715	5.564772286	2.81E-08	1% 有意
世帯数×居住年数5年未満の人口割合	3246.998129	3.429532823	0.000611022	1% 有意	前面道路：種類_国道×アパート・低中層マンション割合	1708.545696	3.849766398	0.000120165	1% 有意
世帯数×デパート	-3088.599915	-4.166441079	3.16E-05	1% 有意	都市計画_市街化調整区域×容積率（%）	-1559.684028	-2.072947879	0.038244011	5% 有意
単身世帯割合×ショッピングモール	4418.736892	6.635219049	3.70E-11	1% 有意	都市計画_市街化調整区域×居住年数が20年以上の人口割合	-2534.445449	-3.170643342	0.00153305	1% 有意
正規労働者割合×高校	-402.9315642	-0.461481599	0.644479389	有意でない	最寄り駅名_富山×用途_住宅	2817.132191	3.286392184	0.001023963	1% 有意
最終学歴が大学以上の人口割合×居住年数5年未満の人口割合	7837.249237	7.288911045	3.78E-13	1% 有意	最寄り駅名_富山×前面道路：方位_南東	1025.820377	2.167899805	0.030227851	5% 有意
居住年数5年未満の人口割合×高校	-4378.94621	-4.254707309	2.14E-05	1% 有意	最寄り駅名_富山×前面道路：種類_市道	3238.430038	2.561366045	0.010464201	5% 有意
一戸建て割合×警察署／交番	-1307.911092	-2.684265245	0.007300305	1% 有意	建物の構造_木造×前面道路：種類_私道	-2705.378954	-4.78130845	1.81E-06	1% 有意
アパート・低中層マンション割合×min_distance_to_crime	1546.296617	3.169646004	0.001538307	1% 有意	前面道路：方位_南東×前面道路：種類_市道	1168.810742	2.519254181	0.011800849	5% 有意

おり，実際に因果関係を検証するためには，因果推論などの手法を用いる必要があることに注意する必要がある．さらに，解釈された結果は，予測モデルの精度に左右されるため，その精度が大きいことが前提となっていることにも留意すべきである．

しかしながら，単純にその場所で過去に発生した犯罪の件数を蓄積し，「ここは犯罪が発生しやすい」と判断するだけではなく，予測値に対する各説明変数の貢献度を算出し，どのような要素が犯罪の発生に寄与しているのか，または寄与していないのか，その傾向を可視化することで，上記のような新たな知見を得られる可能性があることは，本研究における有意性のひとつと言えるだろう．一方で，可視化される要因の精度も少なからず考慮しなければならない．そこで，今後の課題として，予測モデルを介さず，実際のデータから各説明変数の貢献度を算出し，可視化することが挙げられるだろう．

第6章

おわりに

本研究では，欧米を中心に研究や実用されている地理的犯罪予測について，犯罪が発生する頻度が小さいわが国においても，適切なアプローチを行うことによって，時空間的に解像度の大きい予測を行う手法を検討した．また，予測モデルに対して，解釈手法を用いることによって，特定の地域ごとに犯罪が発生する要因を算出し，GIS 上に可視化する手法を提案した．また，予測の精度をさらに高めるために，統計データなどのオープンデータのほかに，地図画像という非構造データを処理することによって，そのエリアの地理的な特徴量を抽出した．また，ナビゲーションサービスからスクレイピングをすることにより，さまざまなジャンルの施設データを取得し，犯罪発生予測モデルの構築に使用した．

数値実験では，富山県警察が公開している「犯罪発生マップ」から犯罪発生データを取得し，本研究で提案している手法を用いて，犯罪発生予測モデルを構築した．学習に使用しなかった検証用データによる予測モデルの精度の検証では，満足のいく予測精度ではなかったものの，予測モデルを構築する際の特徴量選択により，地図画像から抽出した建物や空き地の面積比率，道路のエッジ数が犯罪の発生に寄与していることが分かり，地図画像からの特徴量は，予測精度に正の影響を与えることが分かった．

また，予測モデルを解釈することにより，予測モデルはどの地域で犯罪が発生しやすいと予測しているのか，また，その要因はどのようなものなのかを分かりやすく可視化した．そのため，たとえば，この地域では犯罪が発生しやすく，特に金融機関の最短距離が強く影響しているから，その地域にある金融機関を特にパトロールしたり，また金融機関に注意喚起の張り紙をつけるなど，犯罪抑止への新たな知見が得られることが期待できる．

今後の課題として，まず予測精度の向上が挙げられる．本研究の手法により作成した予測モデルは，必ずしも実用的だとは言えず，改善が必要である．たとえば，さまざまなサンプリング手法や，アンサンブル学習のアルゴリズムで比較する必要があるだろう．また，犯罪が発生することを異常だと仮定し，異常検知問題として予測することも考えられる．

また，要因の可視化は，予測モデルに基づくものであった．すなわち，その要因の精度は予測モデルの精度に左右される．そこで，予測モデルを介せず，実際のデータのみで要因を可視化する手法の開発も課題のひとつであるだろう．

さらに，警察関係者が簡単に犯罪を予測したり，要因を確認できるよう，本研究で提案した手法をバックエンドにもつシステムを作成することも，今後の課題として挙げる．

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．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．

2025 年 2 月

中島 健希

参考文献

[1] S. Rosen, "Hedonic prices and implicit markets: Product differentiation in pure competition," *Journal of Political Economy*, vol. 82, no. 1, pp. 34-55, 1974.

[2] D. J. Gallagher and M. McLoughlin, "Modeling house price indices using hedonic regression," *Journal of Property Research*, vol. 23, no. 3, pp. 203-220, 2006.

[3] G. Chamberlain and G. W. Imbens, "Nonlinear models of hedonic price functions and the impact of hedonic variation on demand," *Econometrica*, vol. 71, no. 1, pp. 1-23, 2003.

[4] P. Cheshire and S. Sheppard, "On the Price of Land and the Value of Amenities," *Economica*, vol. 62, no. 246, pp. 247-267, 1995.

[5] H. R. Varian, "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, vol. 28, no. 2, pp. 3-28, 2014.

[6] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[7] H. Zou and T. Hastie, "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301-320, 2005.

[8] Y. Wang, J. Kawakami, Y. Hatayama, and S. Furuta, "スパース推定を用いた新しいヘドニック法について," 日本銀行ワーキングペーパーシリーズ, WP 20-J-05, 2020.

[9] Open Knowledge Foundation, "Place overview - Global Open Data Index," http://index.okfn.org/place.html, 2025 年 1 月 10 日閲覧.

[10] 総務省統計局, "ヘドニック法について", https://www.stat.go.jp/info/ronbun/pdf/cpi0611.pdf. 閲覧日 2025-01-10.

[11] 日本銀行, "ヘドニック・アプローチによる品質変化の計測と CPI への影響", https://www.imes.boj.or.jp/research/papers/japanese/kk14-3-5.pdf. 閲覧日 2025-01-10.

[12] 一橋大学, "価格比較サイトデータに基づくヘドニック法による分析", https://pubs.iir.hit-u.ac.jp/admin/ja/pdfs/file/1599. 閲覧日 2025-01-10.

[13] IBM, "多重共線性 — IBM", https://www.ibm.com/jp-ja/topics/multicollinearity. 閲覧日 2025-01-10.

[14] Bellcurve, "多重共線性とは", https://bellcurve.jp/statistics/glossary/1792.html. 閲覧日 2025-01-10.

[15] Qiita, "多重共線性について難しいので専門書の記述をまとめた", https://qiita.com/aokikenichi/items/8b72965010e21cc1a004. 閲覧日 2025-01-10.

[16] T. Shibuya,“多重共線性のはなし”, https://tjo.hatenablog.com/entry/2025/01/13/120000. 閲覧日 2025-01-10.

[17] TJO Blog,“【共線性解決!?】python で主成分分析 (PCA) をやってみた”, https://tjo.hatenablog.com/entry/2025/01/13/120000. 閲覧日 2025-01-10.

[18] T. Yukiyanai,“欠落変数バイアスと処置後変数バイアスの検討”, https://yukiyanai.github.io/jp/classes/econometrics2/contents/R/omitted-variable-bias.html. 閲覧日 2025-01-10.

[19] “省略変数バイアスとは”, https://ja.statisticseasily.com/%E7%94%A8%E8%AA%9E%E9%9B%86/%E6%AC%A0%E8%90%BD%E5%A4%89%E6%95%B0%E3%83%90%E3%82%A4%E3%82%A2%E3%82%B9%E3%81%A8%E3%81%AF%E4%BD%95%E3%81%8B. 閲覧日 2025-01-10.

[20] “欠落変数バイアスと回帰分析の影響”, https://www.goodnalife.com/entry/2020/02/26/230731. 閲覧日 2025-01-10.

[21] “回帰分析における欠落変数バイアス”, https://nufs-nuas.repo.nii.ac.jp/record/1142/files/B-NUFS01_01.pdf. 閲覧日 2025-01-10.

[22] 交互作用の定義と具体例. ビジネスリサーチラボ. URL: https://www.business-research-lab.com/220606/?utm_source=chatgpt.com

[23] “交互作用とは？主効果との関係や交互作用の有無を判定するやり方”, https://gmo-research.ai/research-column/interaction?. 閲覧日 2025-01-10.

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

[25] 得田雅章, ”ヘドニック・アプローチによる滋賀県住宅地の地価形成要因分析”, 山崎一眞教授退職記念論文集, 第 381 号, 2009.

[26] 尾崎正憲, 福山博文, ”ヘドニック・アプローチによる鹿児島県住宅地の地価形成要因分析”, 地域政策科学研究, 9 巻, pp. 17-37, 2012.

[27] 島部達哉, ”犯罪発生要因の可視化と不均衡なデータに対処した予測モデルの精度向上”, 富山県立大学学士論文, 2023.

[28] 三浦英俊, ”緯度経度を用いた 3 つの距離計算方法”, オペレーションズ・リサーチ, December 2015.

[29] ”danvk/extract-raster-network: Extract a network graph (nodes and edges) from a raster image”, GitHub, https://github.com/danvk/extract-raster-network, 2025 年 1 月 30 日閲覧.

[30] 向直人, ”愛知県の犯罪オープンデータと地理的特徴量を利用した機械学習による犯罪種別の学習と予測”, 椙山女学園大学文化情報学部紀要, Vol. 21, pp. 109-119, 2022

[31] Boritaso Blog, ”交差検証の基礎をわかりやすく解説！”, https://boritaso-blog.com/cross_validation/, 閲覧日 2025-02-03,

[32] 国土交通省, ”不動産情報ライブラリ”, https://www.reinfolib.mlit.go.jp/realEstatePrices/,2024 年 10 月 20 日閲覧.