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

Visualizing Crime Factors and Improving the Accuracy of Predictive Models Dealing with Imbalanced Data

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

不動産価格は「一物四価」とも言われ, 実勢価格, 公示地価, 路線価, 固定資産税評価額の 4 種類が存在する. 実勢価格は市場において現実に成立した価格であり, 一般的な売買の価格を決定する際に参考とされるが, 売主と買主の個別の事情によって価格が決定されるため, 必ずしも適正な価格が形成されているとは限らない. 次に, 公示地価は, 地価公示法に基づいて評価された毎年 1 月 1 日時点の $1m^2$ 当たりの価格であり, 毎年 3 月頃に公示され, 一般の土地の取引価格の指標とされる. また, 路線価は土地に対する相続税や贈与税算定のための基礎価格として, 国税庁が発表する価格のことであり, 公示地価の約 80%を目安に決定され, 毎年 1 月 1 日を評価時点として, 地価公示価格・基準地価・売買事例などをもとに算出される. 最後に, 固定資産税評価額は, 各市区町村が算定する固定資産税の基準となる評価額であり, 主に固定資産税計算時に用いられ, 公示地価の約 70%が目安となっている.

不動産の価値評価に関して，国土交通省の不動産鑑定評価基準では次のように記されている.

不動産の価格は，多数の要因の相互作用の結果として形成されるものであるが，要因それ自体も常に変動する傾向を持っている. したがって，不動産の鑑定評価を行うに当たっては，価格形成要因を市場参加者の観点から明確に把握し，かつ，その推移及び動向並びに諸要因間の相互関係を十分に分析して，前記三者（不動産の効用，相対的稀少性，不動産に対する有効需要）に及ぼすその影響を判定することが必要である. [国土交通省（2002）p.6]

具体的に価格形成要因は，一般的要因，地域要因及び個別的要因に分けられるとされている.

多くの要因ゆえに，一般的に住宅地地価を含む不動産はユニークな商品であるという側面を持つ. すなわち，自然的特性としての不動産，再生産不可能性，土地やその形状といった個別性，加えて人的特性としての用途の多様性ゆえに，同一の不動産というものは存在せず，それぞれは完全に差別化された財といえるのである. 要因それぞれの重要度を再整理したうえで定量化するために，本論ではヘドニック・アプローチ（hedonic approach）を採用する. ヘドニック・アプローチでは住宅地の価値を地域のアメニティ，環境質，利便性，規模等のような「特性（characteristics）」の合成物であると考える. この場合，共通の客観的性質を示す特性のレベルに依存して住宅地地価が決定されると考えることから，共通の客観的特性に基づく金額換算が可能となる. 概念的には，観測・非観測な情報分析を通じて，住宅地需要者が一定の予算制約のもとで効用最大化が図れる土地を選択し，一方住宅地供

給者は自らの利潤最大化が図れる場合に住宅地の供給を行うというものである. そうした結果として成立する均衡価格を想定するのがヘドニック・アプローチである. このようなアプローチは不動産市場分析でよく使われる手法である一方，社会資本等の非市場の経済評価や，日本銀行調査統計局（2007）のように品質調整法として活用されることも多い.

§ 1.2　本研究の目的

ヘドニックアプローチは, 物価指数の作成における品質調整方法の一つであり, 商品の品質をその商品の持つ個々の機能・性能に分解可能であるとの仮定に基づき, 各機能・性能と価格の関係を回帰分析を通じて推計する手法である. この手法は, 不動産のように多様な特性が価格形成に関与する場合に有効であり, 住宅地の価値が地域の特性や環境の質に依存するという考え方に基づいて, 各特性の価格への影響を評価することが可能となる. しかし, ヘドニックアプローチでは, 説明変数間の多重共線性の問題や, 変数の欠落によるバイアスの発生が課題とされている.

本研究では, これらの課題を克服するために, スパース推定の一つであるアダプティブエラスティックネット（Adaptive Elastic Net, AEN）を導入する.AEN は, 多数の説明変数の中から意味のある変数のみを選択し, それ以外の変数のパラメータを 0 とする特性を持つため, 変数選択の自動化を可能にする. さらに,AEN は「グループ効果」と呼ばれる多重共線性に対する頑健性や, 変数選択の適正性と係数の漸近的な不偏性を同時に担保する「オラクル性」を有している. これにより, 変数間の多重共線性や, 変数の欠落によるバイアスの問題を克服できる可能性がある.

本研究を行う理由は, 不動産価格の決定要因を明確化することが, 適正な不動産評価や価格予測の精度向上に寄与すると考えたからである. 特に, 土地取引の価格は, 地域経済や投資判断に大きな影響を与えるため, より客観的で信頼性の高い評価手法の開発が求められている. また, 不動産価格は地域社会の経済活動の指標の一つでもあり, 政策立案や不動産市場の安定化にも重要な示唆を与える. ヘドニックアプローチを用いた不動産価格の分析は, 不動産取引の透明性を高め, より公正な取引環境を構築するためにも必要不可欠な手法であるといえる.

本研究の重要性は, 従来の最小二乗法を用いたヘドニック分析の限界を克服する新たな手法を提示する点にある. 特に,AEN を用いたスパース推定による変数選択の自動化は, 多数の説明変数が存在する場合でも, 重要な変数のみを適切に選択できる可能性があるため, 予測精度の向上とともに, モデルの解釈性を高める効果が期待される. 本研究は, 不動産の価格形成メカニズムの理解を深めるとともに, 価格評価の公正性と透明性を高める新たなアプローチを提供するものであり, 不動産市場の発展に寄与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 [28]．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．

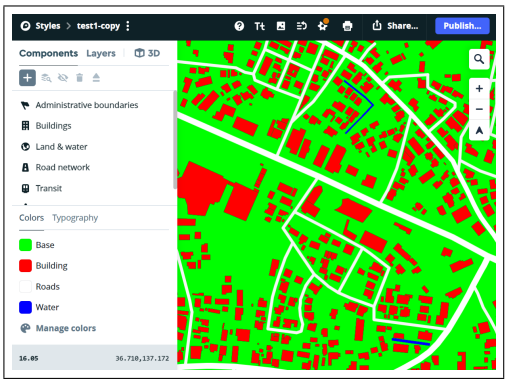


図 2.1: Mapbox Studio

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．

スクレイピングとは，データを収集した上で利用しやすいように加工をすることである．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．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り，スクレイピングとは違い，これは，単に Web 上のデータを収集することを意味する．データを活用するために，使いやすく抽出や加工をしたりするのがスクレイピングの特徴である．

BeautifulSoup4 とは，Web サイト上の HTML から，必要なデータを抽出することができるライブラリである．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 関数などがある．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ができ，そのオブジェクトから要素検索をすることで必要な情報を抽出する．

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．

§ 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.

- **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: 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.

Step 6: 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 7: 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.

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.

§ 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.
- **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（Lasso Regression） 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.

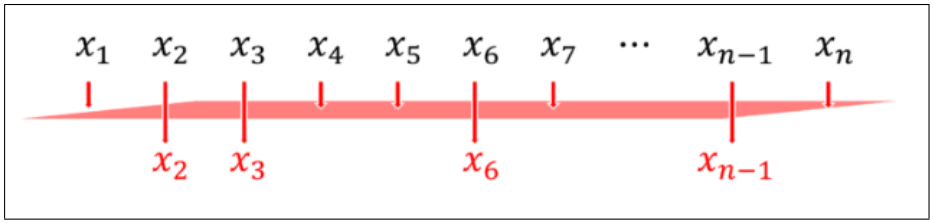


図 2.3: データセットを作成するまでの流れ

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（Forward Selection,Backward Elimination） 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（Forward Selection）**：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（Backward Elimination）**：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（Stepwise Selection）**：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（PCA） 主成分分析（PCA）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（Variance Inflation Factor）**：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（Condition Number）**：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

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 [1]。

ヘドニック法の数式

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。

$$P = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \epsilon$$

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。

ヘドニック法の問題点

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**
- **欠落変数によるバイアス**
- **交互作用の存在**

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します。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ります [**?**]。

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す。

- **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ります。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ります [?]。
- **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ります。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ります [?]。

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ります。

- **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ります。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ります [?]。
- **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ります。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ります [?]。
- **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ります。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す [?]。

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、以下のものがあります。

- **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ます。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ります [?]。
- **分散拡大係数（VIF）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す。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ります [?]。

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ります。

- **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す [?]。

- **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きます [?].
- **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きます [?]。
- **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きます [?].

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ます。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す [9]。

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ます。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ります。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ります [11]。

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ます。

- **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ります。

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。

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) + \epsilon \tag{3.1}$$

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 [?]。

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 [13]。

§ 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.

$$\underset{\beta}{\text{minimize}} \quad \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3.2}$$

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.

$$\underset{\beta}{\text{minimize}} \quad \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3.3}$$

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.

$$\underset{\beta}{\text{minimize}} \quad \frac{1}{2n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=1}^{p}\beta_j^2 \tag{3.4}$$

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.

スパース推定の効果と利点

スパース推定には, 以下の利点がある.

- **変数選択の自動化**：Lasso 回帰や Elastic Net は, 不要な変数を自動的に除去するため, モデルの解釈性が向上する.
- **過学習の防止**：Ridge 回帰は, すべての変数をモデルに残しながら係数を小さく抑えるため, 過学習のリスクを低減できる.
- **高次元データへの対応**：Lasso や Elastic Net は, 変数の数が観測データ数よりも多い場合でも使用可能である.

グループ効果

グループ効果（Group Effect）は, 関連する変数がグループとしてまとめて選択される現象を指す.Lasso では一部の変数のみが選択されるが, 関連する変数を同時に選択したい場合には,**Group Lasso** が用いられる.Group Lasso は, 変数を事前に定義されたグループに分割し, グループごとに ℓ_2 ノルムを適用しながら, 全体として ℓ_1 ノルムの正則化を行う.

$$\underset{\beta}{\text{minimize}} \quad \frac{1}{2n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\sum_{g=1}^{G}\|\beta^{(g)}\|_2 \tag{3.5}$$

ここで,$\beta^{(g)}$ はグループ g に属する回帰係数ベクトルである. これにより, 関連する変数がグループ単位で選択される.

オラクル性

オラクル性（Oracle Property）とは, スパース推定手法が真のモデル構造を正しく推定できる性質を指す. オラクル性を満たす手法は, 真に有効な変数を正しく特定し, 不要な変数の係数を 0 にできる.Lasso は特定の条件下でオラクル性を満たすとされるが,Elastic Net や Ridge 回帰は必ずしもオラクル性を満たすわけではない. これを達成するためには, 非ゼロの係数が十分に大きいことや, サンプルサイズが十分に大きいことが必要とされる.

スパース推定の応用事例

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	×	×	○	○	○	×	○	×

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

2.2 節，2.3 節では，すべての GIS に対して一般的な名称として GIS と表記したが，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 [?].

- 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．

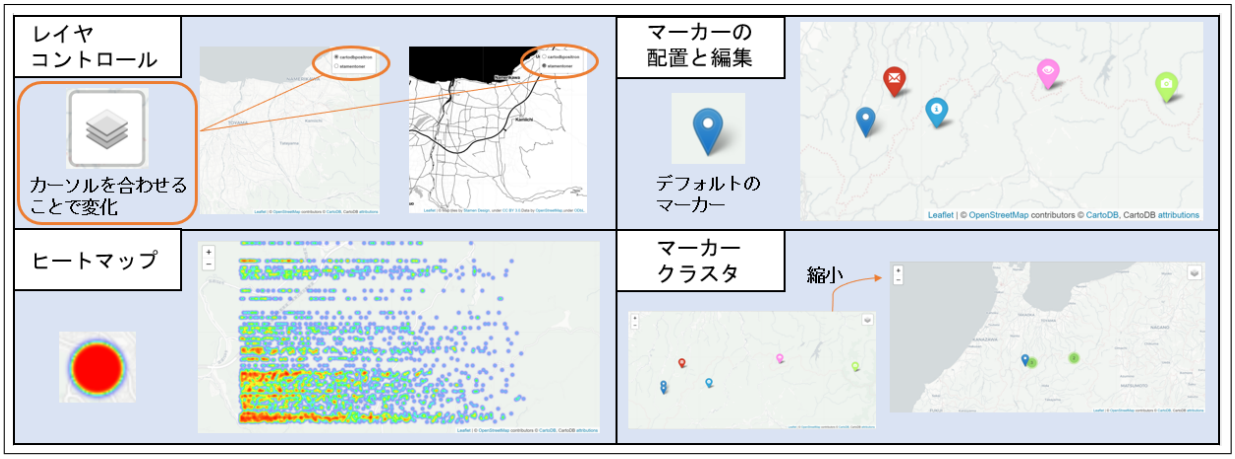


図 3.1: Folium による Web-GIS 実装例

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．

以上で解説した各種機能の実装例については図 3.1 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と変数選択

本研究では，予測する対象地域を富山県とし，過去の犯罪発生データから，特定の日にどこで犯罪が発生するかを予測するモデルを作成する．犯罪発生予測モデルを作成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.

説明変数の選定

犯罪の発生にはさまざまな要因が考えられる．そのため，できるだけ多くの説明変数を考慮することが望ましい．しかしながら，むやみやたらに目的変数と相関がない説明変数を追加しても，予測精度が上がるどころか，計算コストが増大するだけであろう．また，本システムで統計データを取得するために利用している e-Stat は，グリッドセル・小地域ごとのデータに絞っても，200 以上公開されている．さらに，それらを別のデータと計算し，新たな説明変数を作成することを許せば，その組み合わせは考慮しきれない.

そこで，機械学習による犯罪予測モデルを作成する際に，説明変数を容易に選定することができるようにしている．選定できるデータは，e-Stat で公開されているグリッドセル・小地域ごとの統計データ，および，NAVITIME で公開されている施設データである．前者については，異なるデータ間で計算した結果を説明変数として使用できるようになっている.

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.

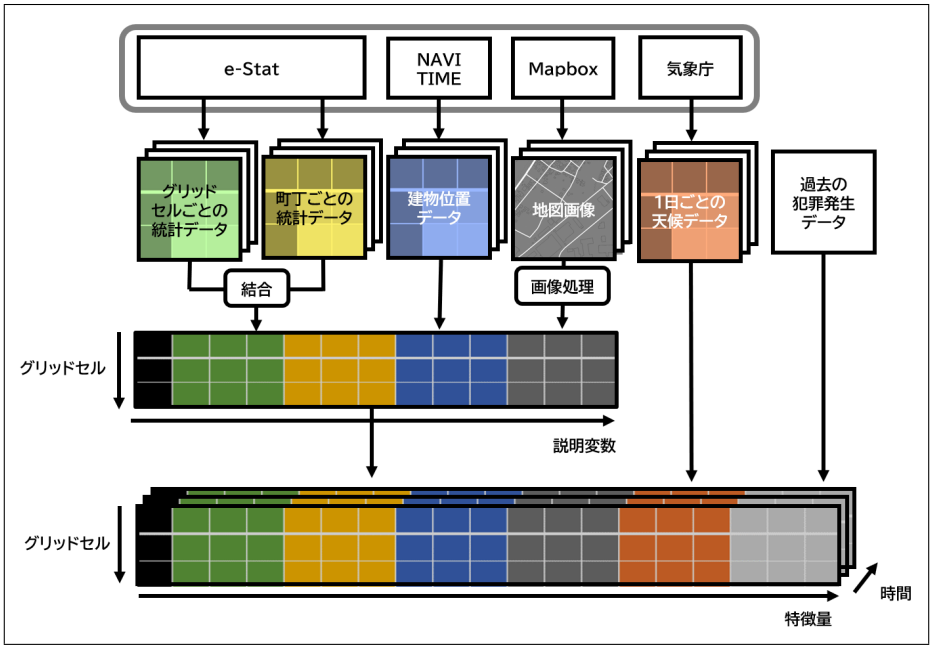


図 4.1: データセットを作成するまでの流れ

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

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，犯罪発生の要因となり得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

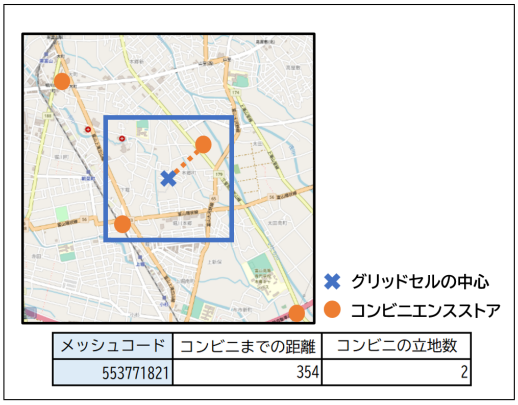
さまざまなジャンルの施設について，NAVITIME からスクレイピングを行い，施設名と，その緯度と経度を取得する．本システムでは，それぞれのジャンルごとに，対象のグリッドセルに含まれる数と，最も近くにある施設までの距離を説明変数とする．

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 [?] 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．

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

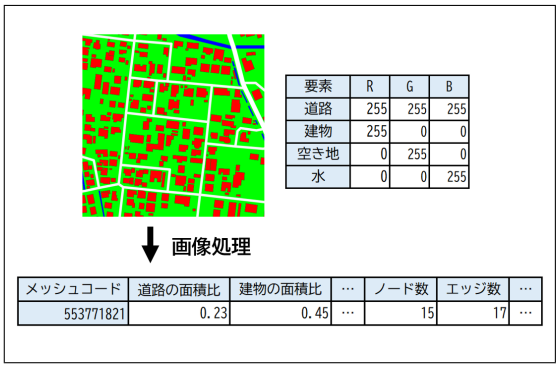
$$M = \frac{R_x(1 - E^2)}{W^3} \tag{4.3}$$

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



メッシュコード	コンビニまでの距離	コンビニの立地数
553771821	354	2

図 4.2: 施設データに基づく説明変数



要素	R	G	B
道路	255	255	255
建物	255	0	0
空き地	0	255	0
水	0	0	255

メッシュコード	道路の面積比	建物の面積比	…	ノード数	エッジ数	…
553771821	0.23	0.45	…	15	17	…

図 4.3: 地図画像に基づく説明変数

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

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

多くの人が集まりやすいレジャー施設やショッピングモールは，犯罪生成・誘引要因となりやすい．逆に，人気が少ない駐車場は，犯罪可能要因となり得る．施設に関する説明変数を自由に選択できるようにしたことは，犯罪要因の特定に役立つことが期待できる．

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．他人による自然な監視は犯罪を抑制する．たとえば，道路や建物の面積比率が大きいほど，監視の量が増え，犯罪が起こりにくく，逆に空き地の面積比率が大きいほど，犯罪が発生しやすい傾向があるならば，それぞれの説明変数は有用なものとなるだろう．

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 [?] 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．

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

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

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 [?]．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．

動的データと静的データの組み合わせ

上で述べたものはすべて，1 日ごとに変化しない静的データであった．それらと，1 日ごとに変化する動的データから，空間（グリッドセル）軸 × 特徴量 × 時間軸の，3 次元のテーブルを作成する．本システムでは，動的データとして，対象のグリッドセルやその周囲で，過去一定期間に発生した犯罪発生件数や，平均気温，日照時間，降水量，降雪量などの天候データを用いる．前者については，犯罪の発生には近接反復被害効果があることから採用した．また，後者については，たとえば，雨や雪が降っている日は，自転車を使う人が減少し，それと同時に自転車盗難も減少するなど，天候も少なからず犯罪発生に寄与すると考え，採用した．

以上により，機械学習に用いるデータセットの作成が完了する．

§ 4.2　価格形成要因の分析

作成したデータセットを用いて，犯罪発生予測モデルを作成する．犯罪が発生するデータが少なく，不均衡であることを考慮し，本システムでは XGboost を機械学習のアルゴリズムとして用いることとする．

XGBoost は，アンサンブル学習のひとつであり，2 値分類の場合は，以下のような流れで学習する．

1 予測値 y を求める．ただし，$0 \leqq y \leqq 1$ であり，初期値は $y = 1$ である．

2 目的変数と予測値との誤差を最小にする決定木を構築する．

3 目的変数と予測値の誤差から各ノードの出力値を求める．

4 各ノードの出力値から予測値を計算する．

5 予測値と目的変数との誤差を計算する．

6 それを目的変数にとし、目的変数と予測値との誤差を最小にする決定木を構築する．

7 2〜5 を繰り返し，誤差を最小化するように逐次的に学習が進む．

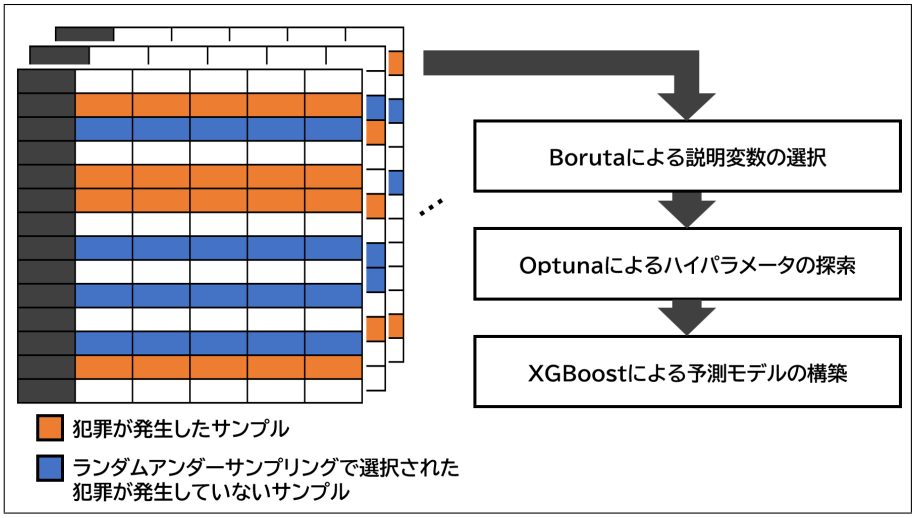


図 4.4: 予測モデルを作成する流れ

適切なランダムダウンサンプリングの探索

不均衡なデータに対するアプローチのひとつにアンダーサンプリングがある．本システムでは，犯罪が発生していないデータからランダムに抽出する，ランダムアンダーサンプリングを行う．

しかしながら，犯罪にはホットスポットと呼ばれる概念があり，特定の数少ない地域に多く発生する傾向がある．そのため単純に，犯罪が発生していないデータの数を，犯罪が発生したデータの数を同一になるようにダウンサンプリングを行った場合，予測対象の地域全体に対して，データセットに含まれる地域の割合は小さくなり，予測精度が小さくなる可能性が考えられる．

本システムで用いるデータセットは，1 日単位の時間軸をもっているため，1 日ごとにランダムダウンサンプリングを行い，新たなデータセットを作成する．このとき，犯罪が発生していないデータからサンプリングする数は，同日に発生したデータの数と同数にしたものとする．

Boruta による説明変数の選択

本システムでは，ユーザが自由に説明変数を選択することができるが，過度に説明変数の数が大きかったり，目的変数と相関がない説明変数があると，過学習などによって，かえって予測精度が低下してしまう可能性がある．そこで，本システムでは，犯罪発生予測モデルを作成する前に，Boruta [?] と呼ばれるアルゴリズムを用いて，適切な説明変数を選択する．

Boruta のアルゴリズムは，以下のような流れである．

1 もともとのテーブルをコピーし，各列をシャッフルする．もとのテーブルの説明変数を Original features，シャッフルした説明変数を Shadow features と呼ぶことにする．このとき，Shadow features は，なんら目的変数に寄与しないはずである．

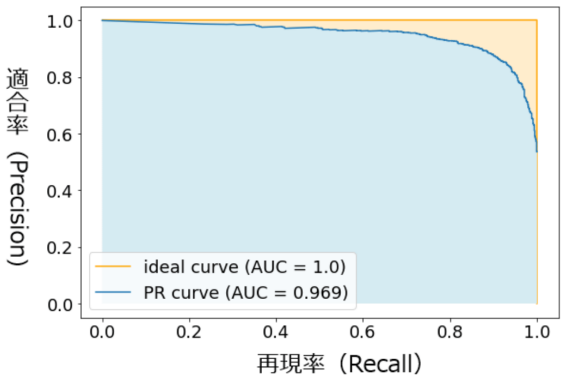


図 4.5: PR-AUC のグラフ例 [?]

		予測値	
		0	1
実測値	0	TN	FP
	1	FN	TP

$$\text{再現率}=\frac{TP}{TP+FN} \qquad \text{適合率}=\frac{TP}{TP+FP}$$

$$F_1\text{スコア}=\frac{2TP}{2TP+FP+FN}$$

図 4.6: 混同行列と評価指数

2 Original features と Shadow features を結合し，ランダムフォレストでモデルを作成する．

3 そのモデルにおいて，それぞれの説明変数の重要度を算出し，Shadow features における最大値よりも大きい Original features を見つける（hit する）．

4 ランダムフォレストの性質により，モデルを作成するごとに重要度は変化するため，1〜3 を n 回くり返す．

5 各 Original features について，Shadow features の重要度と同じことを帰無仮説，より大きい・より小さいことを対立仮説とし，hit した合計を検定統計量 T，$p = 0.5$ としたときの二項分布を用いて検定を行う．

検定の結果，説明変数が，Confirmed，Tentative，Rejected の 3 つに分類される．本システムでは，Confirmed，Tentative の 2 つを，説明変数として用いることとする．

Optuna によるハイパラメータの探索

本研究では，機械学習のアルゴリズムとして XGboost を用いる．XGBoost は，学習する際，決定木の数や最大深度，学習率などといった，事前に決定しなければならない項目（ハイパラメータ）が存在する．ハイパラメータは，一律に最適解は存在せず，データごとに最も予測精度が大きくなるものを探索する必要がある．

ハイパラメータの探索は，さまざまなアルゴリズムが提案されているが，本システムでは Tree Parzen Estimator（TPE）を用いることとし，それを利用できるフレームワーク “Opuna” を用いることにする．

TPE は，学習したモデルの評価基準をもとに，ベイズ最適化によってハイパラメータを探索するものである．モデルの評価基準には，Area Under the Precision-Recall Curve（PR-AUC）を用いる．XGBoost によって作成された 2 値分類を行うモデルは，$0 \leqq y \leqq 1$ を出力する．どれぐらい正例・負例を正しく予測できているかどうかは，しきい値をどのように決定するかどうかに左右される．PR-AUC は，適合率（Precision）と再現率（Recall）をしきい値ごとに算出し，適合率を縦軸，再現率を横軸としてプロットしたときの，下側にある面積である．なお，適合率と再現率は，それぞれ以下のように算出される．

$$Precision = \frac{TP}{TP + FP} \tag{4.10}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.11}$$

ここで，観測値が正例のものについて，正例として予測した数を True Positive（TP），負例として予測した数を False Negative（FN），観測値の負例のものについて，正例として予測した数を False Positive（FP），負例として予測した数を True Negative（TN）と表す．同様の指標として，真陽性率（True Positive Rate）を縦軸，偽陽性率（False Positive Rate）を横軸とする Area Under the ROC Curve（AUC）があるが，AUC-PR は正例の予測に焦点を当てた指標であるため，不均衡なデータにおけるモデルの性能差をより明確に捉えることが期待できる．

以上により，適切なダウンサンプリング，および説明変数の選択を行ったデータセットを用いて，探索によって発見したハイパラメータで XGboost による犯罪発生予測モデルを生成する．

§ 4.3　システムの流れ

犯罪発生予測モデルを使用し，グリッドセルごとに犯罪が発生する予測したマップと，発生する要因を示したマップを作成する．

予測精度を最大化する閾値の決定

犯罪発生予測モデルに，前日までの犯罪発生データを入力する．それらをもとに，当日に犯罪が発生するかを予測し，その結果をマップとして表示する．ここで，モデルから出力される予測値 y は $0 \leqq y \leqq 1$ であり，どの予測値 y から 0 または 1 とみなすか，その閾値 t を決定する必要がある．そこで，過去に発生した犯罪発生データを検証用とし，$t = 0.001, 0.002, \cdots, 0.999$ としたときの F_1 スコアを算出する．最終的に，最も F_1 スコアが大きくなったときの閾値 t を採用する．なお，F_1 スコアは，以下のとおり算出する．

$$F_1 = 2\frac{Precision \cdot Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \tag{4.12}$$

すべてのグリッドセルに対する予測値 y に対して，採用した閾値 t が $y < t$ であれば予測値を $y = 0$，$y \geqq t$ であれば $y = 1$ とし，マップ上に表示する．

犯罪発生要因の可視化

犯罪発生予測モデルに対して，SHAP を適用する．SHAP は，マクロな解釈手法であり，ひとつのインスタンスについて，それぞれの説明変数の SHAP 値が算出される．犯罪発生要因マップを作成するときには，過去の犯罪発生データを入力とする．過去の犯罪発生データは，空間軸 × 特徴量 × 時間軸の 3 次元であった．そのため，これらすべてを SHAP に適

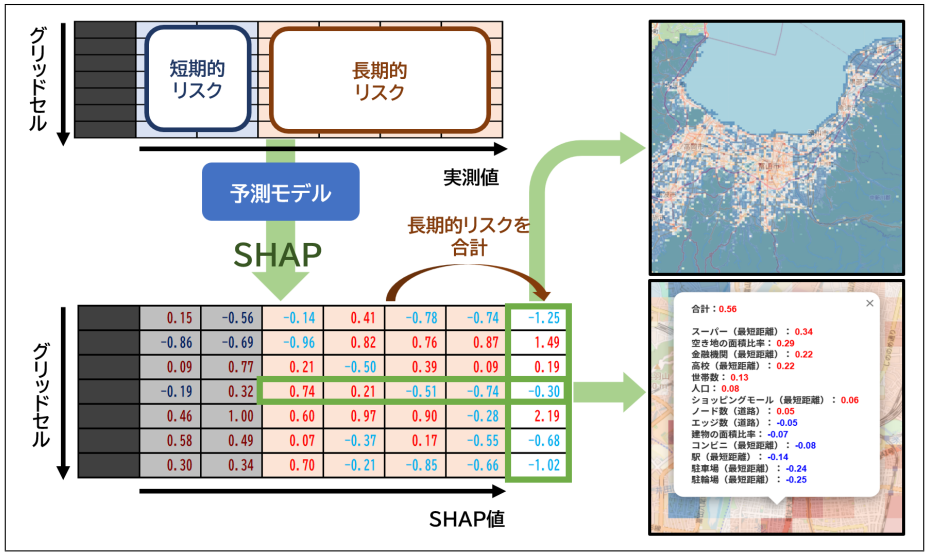


図 4.7: 要因マップを作成する流れ

表 4.1: SHAP 値テーブルの例

メッシュコード	要素1	要素2	要素3	要素4	要素5	合計
00001	0.32	0.42	−0.2	0.1	0.04	0.68
00002	−0.23	−0.1	0.23	−0.31	−0.5	−0.91

用すると，すべてのグリッドセルに対して，予測値に対する説明変数の SHAP 値が，1 日ごとに算出されることになる．

このとき，1 日ごとに変化しない静的データ，すなわち長期的リスクとなり得る要素は，SHAP 値も変化しない．そこで，出力された SHAP 値テーブルから動的データを削除し，さらに，グリッドセルごとに静的データをまとめる．これにより，それぞれのグリッドセルの長期的リスクが，犯罪発生にどれぐらい影響しているのかを表したテーブルが完成する．

SHAP 値は加法性をもっている．つまり，SHAP 値テーブルの各行の合計は，それぞれのグリッドセルが長期的リスクによって，どれだけ犯罪が発生するリスクが大きいかを表している．そして，それぞれの要素の値は，その長期的リスクがどれぐらい犯罪発生に影響しているのかを表している．

たとえば，表 4.1 を例としてみてみる．まず合計の列に着目すると，上のグリッドセルは犯罪が発生しやすく，下のグリッドセルは犯罪が発生しにくいと予測していることが分かる．また，各要素の SHAP 値に着目すると，上のグリッドセルは要素 2 が最も犯罪の発生に影響しており，下のグリッドセルでは要素 4 が最も犯罪の発生を抑制する傾向があることが分かる．これをもとに，GIS 上に可視化する．図 4.8 に可視化した例を示す．SHAP 値の合計が大きくなるほどグリッドセルを赤く着色し，小さくなるほど青く着色することで，どこで犯罪が発生しやすいのかを分かりやすく可視化する．また，それぞれグリッドセルをクリックすることで，各要素の SHAP 値が大きい順で表示される．たとえば，図 4.9 の例では，スーパーとの最短距離が最も犯罪発生に影響しており，次に金融機関との最短距

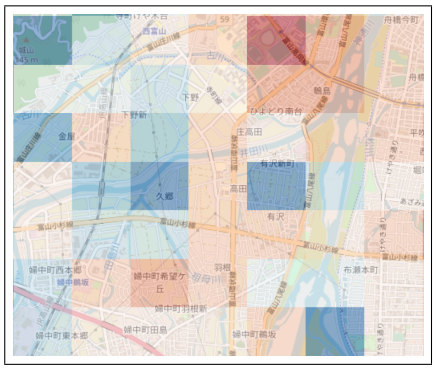


図 4.8: 犯罪発生要因マップの例

図 4.9: 各要素の SHAP 値の例

離，世帯数と続いている．逆に，駐車場との最短距離，駅との最短距離は，犯罪の発生に負の影響を与えていることが分かる．

本研究ではこのように，機械学習によって作成した犯罪発生予測モデルに対して，解釈手法の一種である SHAP を適用することによって，それぞれのグリッドセルはそれぐらい犯罪が発生しやすいのか，それぞれの要因がどれぐらい影響しているのかを GIS 上に可視化することによって，ただ単に予測の結果をもとにパトロールを強化するだけではなく，別のアプローチから犯罪を抑止することを支援する．

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，実際の犯罪発生データを用いて，4 章で述べた手法で犯罪発生予測モデルを作成し，その精度を確認，および考察を行う．また，その予測モデルを用いて，要因を可視化したマップを作成し，考察を行う．

犯罪発生データ

今回の数値実験で使用する犯罪発生データは，富山県警が公開している「犯罪発生マップ」[?] から取得したものを用いる．犯罪発生データには，データ項目として「発生時刻」,「罪種」,「発生場所（緯度，経度）」が含まれており,「自転車盗難」,「ひったくり」,「車上ねらい」,「部品ねらい」,「わいせつ」,「声かけ，つきまとい」,「タイヤ盗難」の 7 種類の路上犯罪が収録されている．また，今回使用するレコードは，2010 年 9 月 1 日から 2020 年 9 月 31 日の約 10 年間とし，発生時刻や発生場所が不正（欠損や範囲外など）なものを除外した 25,814 件である．なお,「犯罪発生マップ」は，このような分析を目的としておらず，掲載されている情報に誤差や欠損が存在する可能性があることに注意されたい．

予測の空間的な解像度は一辺が約 500m のグリッドセル，時間的な解像度は 1 日とする．グリッドセルの基準としては，日本標準地域メッシュの 2 分の 1 地域メッシュを採用した．なお，富山県に含まれるグリッドセルは 17,031 個であるが，データセットに含まれる 10 年間で犯罪が発生したグリッドセルは約 18.4 %の 3,126 個であった．それ以外のグリッドセルでは，犯罪が発生する可能性は今後も限りなく小さいと考え，予測する対象のグリッドセルは，その 3,126 個のみとした．

データセット

予測に用いる説明変数は，表 5.3 に示す計 65 個とした．

3,126 個のグリッドセルについて，静的データをまとめたテーブルを作成し，それに動的データを追加した 3 次元のテーブルを作成する．予測モデルの精度を検証するため，データセットのうち，2020 年 8 月 31 日までの 10 年間を学習用，それ以降の 1 か月間を検証用とする（図 5.1 参照）．

よって，この時点におけるデータセットのレコード数は，$N = 11419278$ であり，式??における不均衡度 r は，$r \approx 0.0023$ である．そこで，学習用データについて，ランダムサンダーサンプリングを行い，$N = 51065$，$r \approx 0.477$ となった．

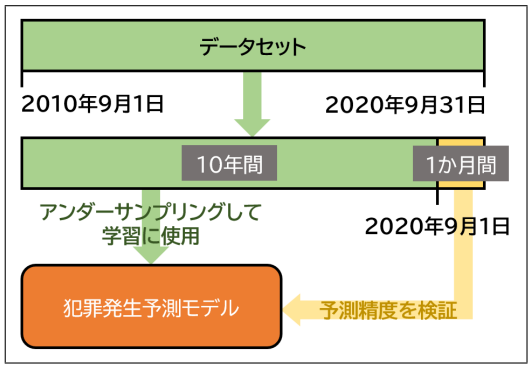


図 5.1: 予測モデルを検証する流れ

表 5.1: データセットに含まれる罪種

カテゴリ	発生件数
自転車盗難	13121
声かけ・つきまとい	6760
車上荒らし	5211
タイヤ盗難	3968
部品ねらい	690
わいせつ	401
ひったくり	59

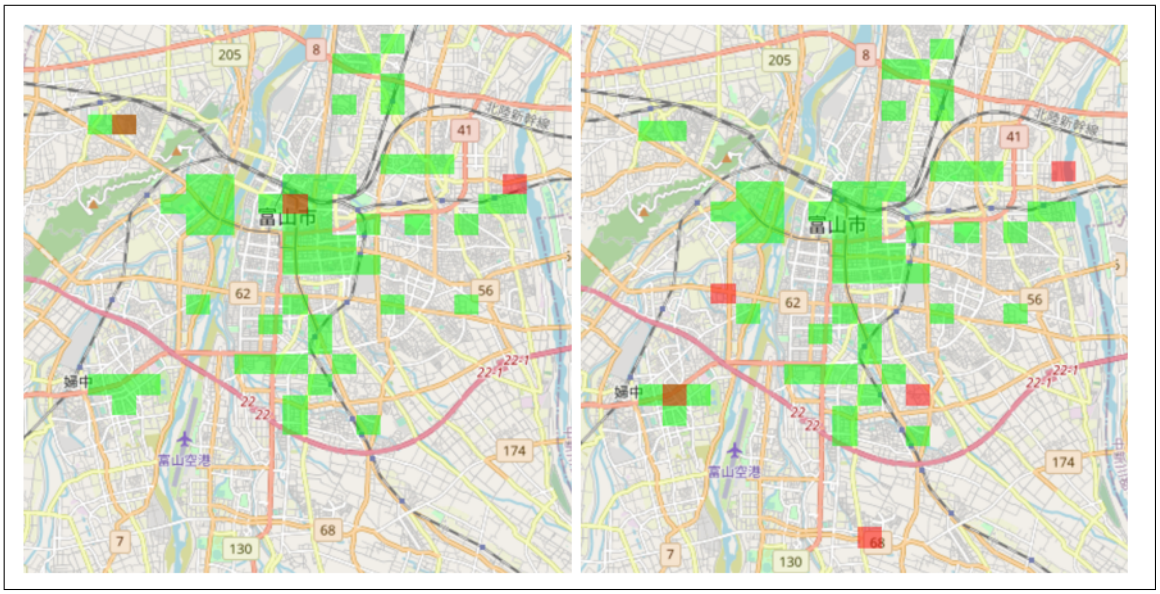


図 5.2: 2020 年 9 月 1 日（左）と 2 日（右）の予測結果

§ 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 5.2 に，同地域における複数日の予測結果を図 5.2 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 5.2 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 5.3 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

表 5.2: 検証用データによる予測結果

		予測値	
		0	1
実測値	0	90946	2677
	1	107	50

適合率	0.018
再現率	0.318
F1スコア	0.035

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 5.2 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．

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 5.3 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 5.3 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 5.4 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

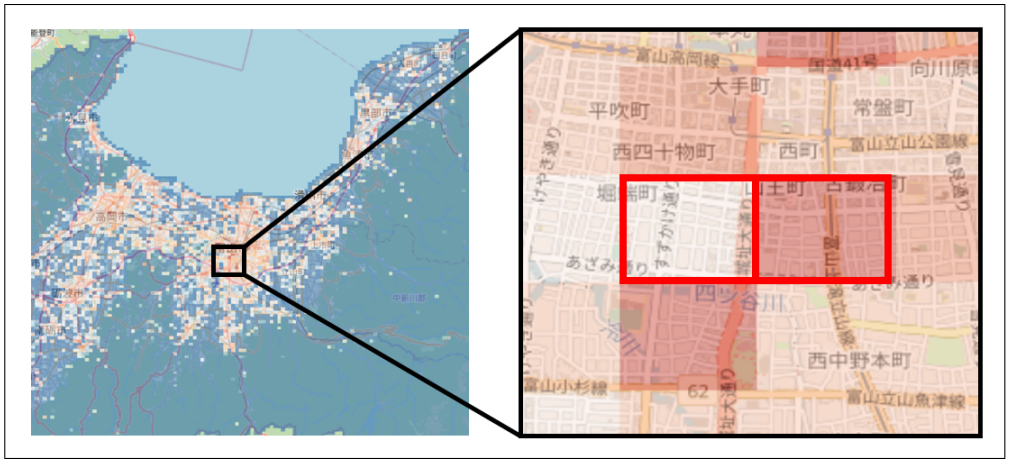


図 5.3: 予測モデルを可視化した結果

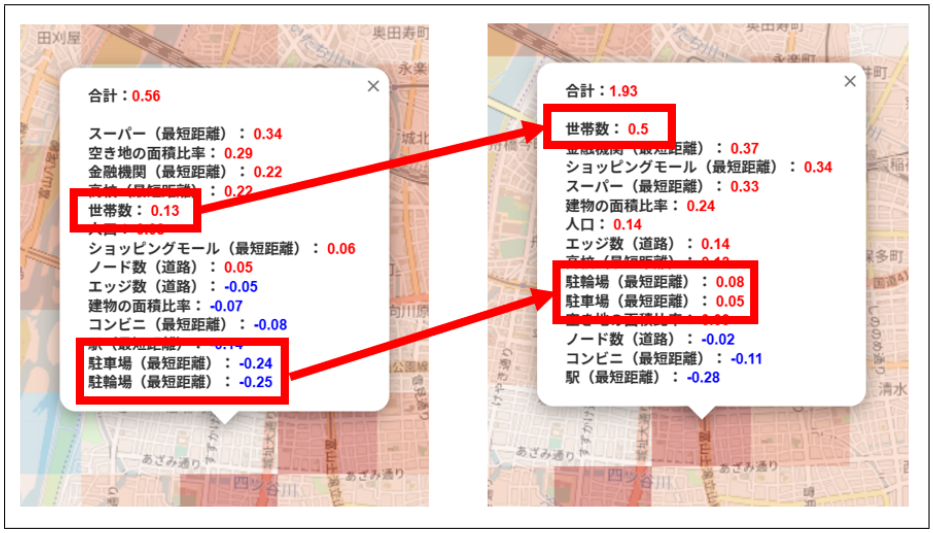


図 5.4: 特定のグリッドセルの要因を可視化した結果

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．

表 5.3: 使用，および選択された説明変数一覧

長期的リスク		短期的リスク
人口	コンビニエンスストア(立地数)	平均気温
18歳未満人口割合	**コンビニエンスストア(最短距離)**	日照時間
65歳以上人口割合	駅(立地数)	降水量
外国人人口割合	**駅(最短距離)**	降雪量
世帯数	駐車場(立地数)	過去7日間の犯罪発生件数
単身世帯割合	**駐車場(最短距離)**	**過去1か月間の犯罪発生件数**
核家族世帯割合	駐輪場(立地数)	休日(ダミー変数)
正規労働者割合	**駐輪場(最短距離)**	曜日(ダミー変数)
非正規労働者割合	金融機関(立地数)	
最終学歴が中学以下の人口割合	**金融機関(最短距離)**	
最終学歴が高校の人口割合	旅館(立地数)	
最終学歴が大学以上の割合	旅館(最短距離)	
居住年数5年未満の人口割合	ホテル(立地数)	
居住年数20年以上の人口割合	ホテル(最短距離)	
一戸建て世帯割合	スーパーマーケット(立地数)	
アパート・低中層マンション世帯割合	**スーパーマーケット(最短距離)**	
高層マンション割合	ショッピングモール(立地数)	
道路の面積比率	ショッピングモール(最短距離)	
建物の面積比率	デパート(立地数)	
空き地の面積比率	デパート(最短距離)	
水の面積比率	警察署／交番(立地数)	
ノード数(道路)	警察署／交番(最短距離)	
エッジ数(道路)	小学校(立地数)	
密度(道路)	小学校(最短距離)	
平均次数(道路)	中学校(立地数)	
	中学校(最短距離)	
	高校(立地数)	
	高校(最短距離)	
	大学(立地数)	
	大学(最短距離)	
	保育園／幼稚園(立地数)	
	保育園／幼稚園(最短距離)	
	レジャー施設(立地数)	
	レジャー施設(最短距離)	

■ ■ Borutaによって選択された説明変数

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．

2023 年 2 月

島部 達哉

参考文献

[1] 総務省統計局（2021）「ヘドニック法について」。オンライン。 https://www.stat.go.jp/info/ronbun/pdf/cpi0611.pdf

[2] 日本銀行（2020）「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」。オンライン。 https://www.imes.boj.or.jp/research/papers/japanese/kk14-3-5.pdf

[3] 一橋大学（2018）「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」。オンライン。 https://pubs.iir.hit-u.ac.jp/admin/ja/pdfs/file/1599

[4] IBM. (2025). 多重共線性 — IBM. https://www.ibm.com/jp-ja/topics/multicollinearity

[5] Bellcurve. (2025). 多重共線性とは. https://bellcurve.jp/statistics/glossary/1792.html

[6] Qiita. (2025). 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

[7] Shibuya, T. (2025). 多重共線性のはなし. https://tjo.hatenablog.com/entry/2025/01/13/120000

[8] TJO Blog. (2025). 【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

[9] Yukiyanai, T. (2020). 「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

[10] 「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

[11] 「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

[12] 「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

[13] 「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?utm_source=chatgpt.com Reference Service, 2002.