卒業論文

地価形成に関わるオープンデータの
ヘドニック・アプローチによる便益の予測

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5年（2023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

米国の心理学者である Steven A. Pinker は，著書『暴力の人類史』の中で，現代は最も暴力の少ない時代だと述べている [1]．確かに，国連薬物犯罪事務所（UNODC）によると，国ごとに差があるものの，世界全体として犯罪の発生件数は減少傾向にある [2]．それには，多くの理由が考えられるが，そのひとつに，コンピュータの発達により，さまざまなデータが蓄積・活用されるようになったことが挙げられる．すなわち，過去のデータから，犯罪に対する知見を得て，犯罪を事前に防止しようとする動きが強まったのである．

それの先駆けのひとつと言われている研究が，1980 年に発表された，複数のデータを説明変数として，米国における 1 年ごとの犯罪発生率を予測するものである．その後も，米国ミネソタ州セントポール市において，警察に通報された地点を調査したところ，全体のおよそ半数は，全地域の約 3 %から通報されていたことが分かった [4]．すなわち，犯罪は特定の地域に集中して発生し，そのような場所には何らかの脆弱性があると推察できる．また，2001 年に，過去の犯罪発生データと，その他の複数のデータを地理情報システム（Geographic Information System: GIS）上に重ね合わせ，犯罪発生と相関が強い要素を特定したうえで，予測をする手法を提案した [5]．これらのような，過去の犯罪発生データや，犯罪発生に関係があるデータを用いて，将来発生する犯罪を予測する研究は,「地理的犯罪予測」とも呼ばれる [6].

その後，地理的犯罪予測が実務で活用されるようになったのは，2010 年ごろである [7]. 米国では，2009 年と 2010 年に，国立司法研究所がシンポジウムを開き，犯罪発生に先駆けて予見的に警察活動を行う，予測型警察活動（Predictive Policing）について議論が行われた．また，2011 年には，その「Predictive Policing」が，Times 誌の「The 50 best Inventions of The Year 2011」に選定され，犯罪予測に対する社会的な関心も高いことがうかがえる．このように，米国をはじめ，英国，ドイツ，スイス，イタリア，フランスなど，欧米を中心に警察機関で地理的犯罪予測が導入された実績がある．

ここ数年は多くの研究分野で機械学習が注目されており，地理的犯罪予測においても例外ではない．機械学習は，ある目的変数と，それに関連する説明変数から，人間の介入少なく数理モデルを作成する手法である．犯罪学をはじめとする専門的な知識が必要なくとも，地理的犯罪予測に関する研究ができることから，特に計算機統計学の分野で目立って研究されている．

§ 1.2　本研究の目的

地理的犯罪予測に関する研究は，欧米を中心に研究が盛んに行われているものの，わが国においてはほとんど行われていない．実際に国内のデータを用いて地理的犯罪予測を行っている研究のうち，結果が報告されているものは，ごく限られている [1] [8] [9] [10] [11] [12]．また，そのうち，機械学習を用いているものはさらに限られ，いずれも過去の犯罪発生件数のみしか考慮しておらず，予測の時間的な解像度は 1 か月であり，必ずしも実用的とはいえない．

わが国が欧米と対照的な状況にある理由として，犯罪の発生件数が少ないことが挙げられるだろう．法務省によると，2014 年について，10 万人あたりの窃盗の発生件数は，米国で約 2,584 件のところ，日本は約 472 件であった [13]．このように，わが国は国際的にみても治安水準が高く，至急な対応が必要だとは必ずしもいえない．

しかしながら，わが国においても，将来的に犯罪の発生が増加する要因が潜在している [7]．たとえば，自治会などの住民組織加入率の低下や，人種・民族の多様化，単身世帯の増加は，治安の悪化を招くとされている．また，少子高齢化により警察官の人数も減少しており，警察の組織力そのものが弱まっていくことが危惧されている．そのため，地理的犯罪予測などの技術を用いて，限られた警察資源を効率的に配分し，治安を維持していく必要性は，今後ますます増加するだろう．

一方で，犯罪の発生頻度が少ないことは，それ自体が地理的犯罪予測の精度を低下させ，研究を難しくさせていることも考えられる．実際に，発生する頻度が異なる罪種間で，同一の予測手法を適用したところ，頻度が小さい罪種のほうが予測精度は小さくなると報告されている [14]．そこで，本研究では，犯罪が発生していないケースと比較して，犯罪が発生しているケースが極端に少ない，すなわち不均衡なデータに対して，適切なアプローチを行い，犯罪の発生頻度が小さいわが国においても，空間的・時間的に解像度を小さく予測することを試みる．

他方で，犯罪の発生を防止するためには，単純にパトロールを行うだけではなく，犯罪が発生する要因を認識し，根本的な抑止につなげることも大切であろう．しかし，機械学習を用いた地理的犯罪予測に関する研究は，犯罪が発生する要因まで言及しているものは少ない [7]．これは，機械学習によって作成されたモデルは，予測精度と引き換えに，なぜその予測値を出力したのか，その解釈性が小さくなってしまう性質をもっていることも，理由のひとつとして考えられる．

そこで，学習した予測モデルに対して，Shapley Additive Explanations（SHAP）を適用する．SHAP は，機械学習のアルゴリズムを問わず，ひとつの予測値に対して，それぞれの説明変数がどのように影響しているのかを算出することができる，解釈手法のひとつである．この場所ではなぜ犯罪が発生しやすいのか，または発生しにくいのか，その要因を GIS 上に可視化することで，根本的な抑止への知見を得られることを目標とする．

§ 1.3　本論文の概要

本論文は次のように構成される．

第 1 章 本研究の背景と目的について説明した．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．

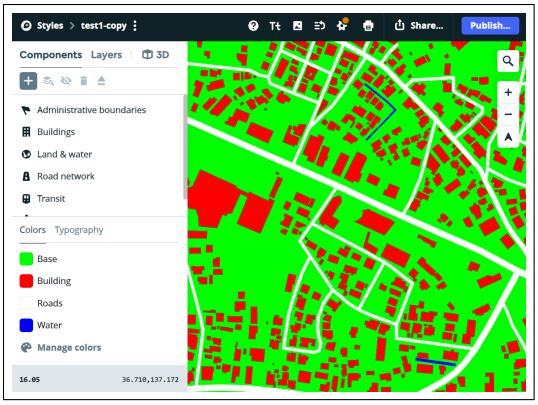


図 2.1: Mapbox Studio

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　ヘドニック・アプローチの理論モデル

本研究では.,Epple（1987）の議論に基づき., 住宅地地価の諸特性を取引する暗黙的な市場を想定したうえで., 市場均衡価格曲線としてのヘドニック関数を導出する. 以下では., 需要サイドおよび供給サイドにおける行動を定式化し., 市場均衡に至る過程を示す.

需要サイドの行動

地価を形成する特性の n 次元ベクトルを $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)$., ヘドニック価格関数を $p(\boldsymbol{z}) = p(z_1, z_2, \ldots, z_n)$ とする. また., 住宅地需要者の効用関数を $U(\boldsymbol{z}, x; \boldsymbol{\alpha})$ と定義する. ここで.,x はニュメレール（価値尺度財）.,$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ は需要者個人のテイストパラメータのベクトルである. 需要者の所得を y とした場合., 予算制約式は以下のように表される.

$$y = p(\boldsymbol{z}) + x \tag{2.3}$$

同時分布関数を $F(y, \boldsymbol{\alpha})$ と表す. この予算制約式のもと., 需要者が $\boldsymbol{z}$ および x について効用最大化行動を取ると., 次式に定式化される.

$$\max_{\boldsymbol{z}, x} \quad U(\boldsymbol{z}, x; \boldsymbol{\alpha}) \tag{2.4}$$

$$\text{s.t.} \quad y = p(\boldsymbol{z}) + x \tag{2.5}$$

この場合., 最適化のための 1 階条件（FOC）は以下の式で表される.

$$p_{\boldsymbol{z}} = \frac{U_{\boldsymbol{z}}(\boldsymbol{z}, y - p(\boldsymbol{z}); \boldsymbol{\alpha})}{U_x(\boldsymbol{z}, y - p(\boldsymbol{z}); \boldsymbol{\alpha})} = h(\boldsymbol{z}, y - p(\boldsymbol{z}); \boldsymbol{\alpha}) \tag{2.6}$$

ここで.,$p_{\boldsymbol{z}}$ はヘドニック価格関数の 1 階微分のベクトルであり.,$U_{\boldsymbol{z}}$ および U_x はそれぞれ特性ベクトル $\boldsymbol{z}$ およびニュメレール x の 1 階微分を示す.

需要者の効用水準が u の下でのビッド関数（bid function）を $\theta(\boldsymbol{z}; u, y)$ とすると.,$U(\boldsymbol{z}, y - \theta) = u$ が成立し., これを微分することで次式が得られる.

$$\frac{\partial \theta}{\partial z_i} = \frac{U_{z_i}}{U_x} > 0 \tag{2.7}$$

$$\frac{\partial^2 \theta}{\partial z_i^2} = \frac{U_{z_i}^2 U_{xx} - 2U_{z_i} U_x U_{z_i x} + U_x^2 U_{z_i z_i}}{U_x^3} < 0 \tag{2.8}$$

すなわち., ビッド関数は増加する凹関数である. 需要者の効用は., ヘドニック関数とビッド関数の接点において最大化されるため., 次の式が成立する.

$$\theta(\boldsymbol{z}^*; u^*, y) = p(\boldsymbol{z}^*) \tag{2.9}$$

$$\frac{\partial \theta}{\partial \boldsymbol{z}}(\boldsymbol{z}^*; u^*, y) = p_{\boldsymbol{z}}(\boldsymbol{z}^*) \tag{2.10}$$

図的には., ヘドニック関数がビッド関数のエンベロープ・カーブとなる.

供給サイドの行動

次に., 供給者の行動を定式化する. 供給者は自らの供給行動を決定する際., 住宅地地価を所与として利潤 π を最大化するように特性の束 $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)$ を選択する. 利潤関数は以下のように表される.

$$\max_{\boldsymbol{z}, M} \quad \pi = p(\boldsymbol{z})M - C(M, \boldsymbol{z}; \boldsymbol{\beta}) \tag{2.11}$$

ここで.,M は供給する住宅地の数.,$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_n)$ は供給者を特徴づけるパラメータベクトルであり., その分布関数を $G(\boldsymbol{\beta})$ とする. また.,$C(M, \boldsymbol{z}; \boldsymbol{\beta})$ は供給者の費用関数である. この場合., 利潤最大化の 1 階条件は以下のようになる.

$$p_{\boldsymbol{z}} = C_{\boldsymbol{z}}(M, \boldsymbol{z}; \boldsymbol{\beta}) \tag{2.12}$$

$$p(\boldsymbol{z}) = C_M(M, \boldsymbol{z}; \boldsymbol{\beta}) \tag{2.13}$$

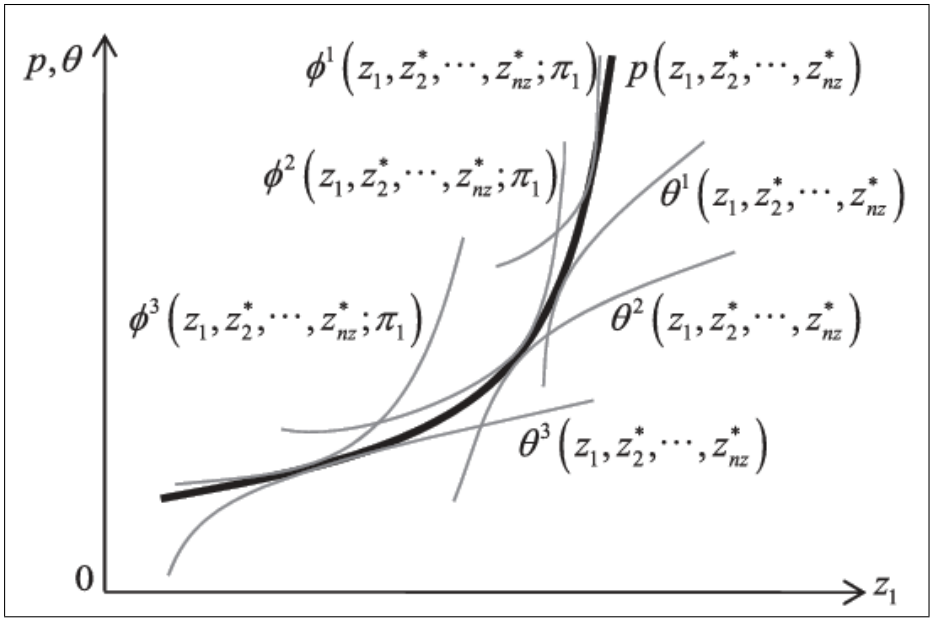


図 2.3: データセットを作成するまでの流れ

ここで., 供給者は各特性の限界的な価値と土地 1 単位当たり特性の限界費用が等しくなるように供給活動を行う. また., 土地の市場価値は供給限界費用に等しくなる. 供給者のオファー関数（offer function）を $\phi(\boldsymbol{z},\pi)$ とすると., 以下が成立する.

$$\phi_{\boldsymbol{z}} = \frac{C_{\boldsymbol{z}}}{M} > 0 \tag{2.14}$$

$$\phi_{\pi} = \frac{1}{M} > 0 \tag{2.15}$$

すなわち., オファー関数は増加する凸関数である. 市場均衡は以下の式を満たす.

$$p(\boldsymbol{z}^*) = \phi(\boldsymbol{z}^*,\pi^*) \tag{2.16}$$

$$p_{\boldsymbol{z}}(\boldsymbol{z}^*) = \phi_{\boldsymbol{z}}(\boldsymbol{z}^*,\pi^*) \tag{2.17}$$

このように., ヘドニック関数は需要者のビッド関数と供給者のオファー関数が市場均衡価格を挟んで接する形で決定される.

ヘドニック関数と市場均衡

ヘドニックアプローチでは., 特性 $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)$ を持つ住宅地に対する需要と供給が合致する点で市場均衡価格が決定される. この価格は需要者サイドの分布 $F(y,\boldsymbol{\alpha})$ と供給者サイドの分布 $G(\boldsymbol{\beta})$ に依存して決定される.

しかしながら.,$F(y,\boldsymbol{\alpha})$ や $G(\boldsymbol{\beta})$ が未知であるため., 一般的には $p(\boldsymbol{z})$ も未知であり., 需要サイドと供給サイドを同時推定することで市場均衡価格を導出する必要がある. この場合., 同時方程式バイアスや関数型の問題が生じることが., 清水・唐渡（2007）で指摘されている.

§ 2.3　説明変数の選定

2.1 節で述べた RTM のように，地理的犯罪予測において，GIS は欠かせない重要な要素となっている．本節では，その GIS について述べるとともに，GIS の地理的犯罪予測への応用について説明する．

GIS

GIS とは，位置に関する情報を持ったデータを総合的に管理・加工し，地理的な位置とデータを結び付けることによって，視覚的に表示することである．これにより，空間データの高度な分析や迅速な判断を可能にする技術である．地理空間データの例として，主題図（土地利用図，地質図，ハザードマップ等），都市計画図，地形図，地名情報，台帳情報，統計情報，空中写真，衛星画像などが挙げられる．GIS はいくつかの優れた特徴を持つため，さまざまな用途で用いられている．その中でも代表的なものとして，次のような 4 つの特徴と 3 つの利点が挙げられる [22].

GIS における 4 つの特徴

- データの可視化

  2D や 3D，アニメーションなど多様な表現方法で地図上にデータを可視化することによって，数値のみを見るだけでは気づくことができないようなデータの傾向やデータ間の関連性など様々な情報を一目で把握できるようになる．

- データ間の関係性の把握

  複数のデータによる重ね合わせを行うことでデータ間の対比が容易になり，データ同士の関係性を直感的に把握することができる．また，地図上の位置関係からデータを特定することによって，定量的な情報を把握することができる．

- データの統合と分析

  位置情報をキーとしてそれぞれ別の特徴を持つデータを統合したり，複数のデータを重ね合わせて分析することによって，独自のアプローチで課題の解を導き出すことができる．

- データの作成と更新

  今日の日本では新しいビルの建設や合併による行政界の変更など，現実世界が日々変化しているため地理情報データもそれらに基づいて定期的に更新していく必要がある．この際，GIS を利用することでデータの作成・更新における負担が軽減され，常に鮮度を保ってデータの管理・提供を行うことができる．

GIS を用いることの 3 つの利点

- 業務効率化によるコスト削減

  GIS は日常の業務を最適化するために幅広く利用されている．紙地図から GIS を利用したデジタルな地図へ移行することで，現地調査や設備管理，統計分析などを行う際

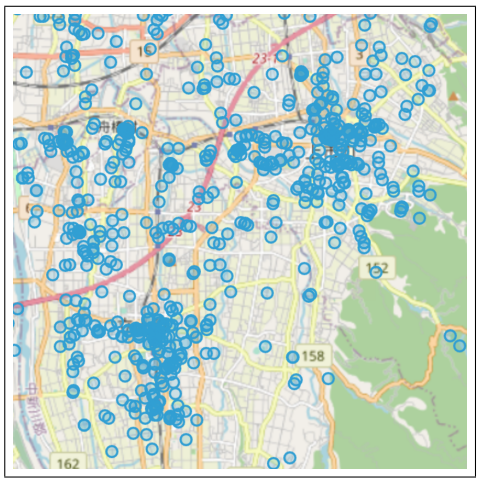


図 2.4: 犯罪発生地点を可視化した例

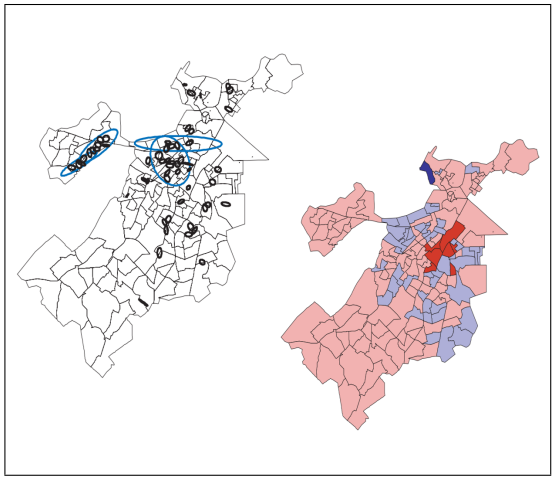


図 2.5: GIS によるホットスポット検出 [23]

の物質的な制約が減少し，より簡単かつ効率的な作業の実現が可能になる．これにより作業時間や人員など業務にかかるコストを大幅に削減することができるという利点を持つ．

- 最適な意思決定

  組織における活動において，場所に関する意思決定を正しく行うことは成功のための重要な要素の 1 つといえる．場所に関する決定に視覚を用いることができるという利点を持つ GIS は店舗の出店場所や配送ルート，避難地域・経路，天然資源の採取地点など多様な分野で最適な場所を策定するために使用されている．

- コミュニケーションの向上

  GIS を使用することでさまざまな表現方法を用いて位置情報を地図上に可視化することができる．可視化された位置情報は状況を効果的に伝え，的確な理解を促すことができ，グループや組織間，社会におけるコミュニケーションの向上を図ることが可能になるという利点を持つ．

地理的犯罪予測への GIS の応用

犯罪にはホットスポットと呼ばれる概念が存在する．ホットスポットは，犯罪が時空間的に集積して発生する現象であり，すなわち，ホットスポットは何らかの脆弱性が存在し，その同定は，将来に発生する犯罪を予測するためにも有用である．このようなホットスポットを分析するうえで，犯罪発生状況を地図上に可視化するクライムマッピングは重要となる．

特に，ホットスポットを特定するためによく用いられるカーネル密度推定は，GIS の発達により容易になったといっても過言ではない．このように，GIS が発達し，誰でも容易に用いることができるようになったことで，ホットスポットの同定が行えるようになり，ホットスポットに関する理論や分析が，地理的犯罪予測の基礎となったといえる [7]．また，地理的犯罪予測の発展には，警察機関の動向も大きく関わっている．

1990 年代に米国の警察が，GIS を日常的に使用するようになったことで，地理的犯罪予測に関する研究が一気に加速したと述べている [24]．その例として，当時のニューヨーク市警察が取り入れた「コムスタット」が挙げられる．これは，犯罪データを週単位で集計して地図上で可視化し，どのような戦略を行うのか，その立案を行うものである．この GIS を取り入れた警察活動は，地理的犯罪予測と，その結果に基づく警察活動の基盤となった．

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　未観測の交絡因子への対処

機械学習における分類問題で，学習に用いるデータセットが，そのクラスの比率に極端な偏りが生じているものを，一般に「不均衡データ」という．

不均衡なデータが予測精度に及ぼす影響

実世界で観測されるデータは，そのクラス比率が不均衡になっているものが多い．分かりやすい例として，医療データがある．身体検査のデータから特定の疾患を発病しているかどうかを出力するモデルを作成しようとするとき，健常者のデータは比較的容易に収集できるが，それと比較して，その患者のデータは，健常者よりも母数が小さく，収集できるデータ数が少なくなってしまう．

しかしながら，一般的な機械学習アルゴリズムは，それぞれのクラスのデータ数が同一であることを仮定していることが多く，不均衡な学習データを用いて機械学習を行うと，少数派に対する分類精度が著しく低下する [26]．この理由を，線形判別分析（Linear Discriminant Analysis: LDA）を例に説明する．LDA は，2 つのクラスを，最も識別できる直線（識別境界）で分離する手法である．多次元であるときを考慮し，サンプル $\boldsymbol{x}_n \in \mathbb{R}^M$ を，識別境界に直行している軸 $\boldsymbol{w}$ で線形変換した

$$y_n = \boldsymbol{w}^\top \boldsymbol{x}_n \tag{3.1}$$

を用いて，クラスを分類する．このとき，目的関数 $J(\boldsymbol{w})$ について，

$$\text{maximize} \quad J(\boldsymbol{w}) = \frac{\boldsymbol{w}^\top \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{S}_W \boldsymbol{w}} \tag{3.2}$$

となる射影軸 $\boldsymbol{w}$ を求める．なお，クラス間変動行列 $\boldsymbol{S}_B$ は，

$$\boldsymbol{S}_B = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^\top \tag{3.3}$$

であり，クラス内変動行列 $\boldsymbol{S}_W$ は，

$$\boldsymbol{S}_W = \sum_{n \in C_1} (\boldsymbol{x}_n - \boldsymbol{m}_1)(\boldsymbol{x}_n - \boldsymbol{m}_1)^\top + \sum_{n \in C_2} (\boldsymbol{x}_n - \boldsymbol{m}_2)(\boldsymbol{x}_n - \boldsymbol{m}_2)^\top \tag{3.4}$$

である．いま，クラス C_1, C_2 の平均 $\boldsymbol{m}_1, \boldsymbol{m}_2$ は変化せずに，C_2 のサンプル数が十分小さいとき，式 3.4 の第 2 項は，第 1 項と比べて非常に小さくなる．したがって，射影軸 $\boldsymbol{w}$ は，ほ

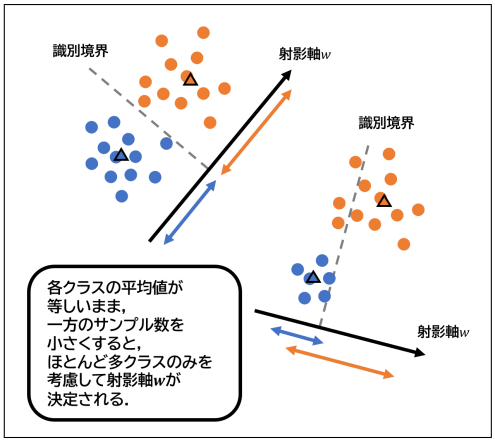


図 3.1: クラス比率による分類精度

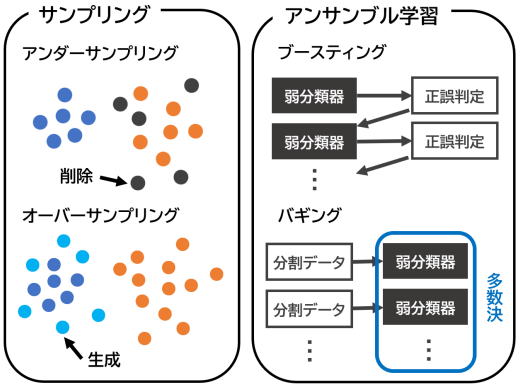


図 3.2: サンプリングとアンサンブル学習

とんどクラス C_1 の変動のみを考慮することとなり，クラス分類の精度は小さくなる（図?? 参照）．このような現象は，LDA に限ったことではない．

多数クラス C^{maj} のサンプル数を N^{maj}，少数クラス C^{min} のサンプル数を N^{min} とすると，データの不均衡度 r は

$$r = \frac{N^{min}}{N^{maj} + N^{min}} \tag{3.5}$$

と定義できる．$r < 0.2$ であると，そのデータは十分不均衡であり，機械学習に用いるときは，何らかの工夫が必要である [27]．以降，不均衡データに対するアプローチとして，サンプリングとアンサンブル学習を説明する．

サンプリング

サンプリングとは，多数クラス C^{maj} のデータを間引いたり，少数クラス C^{min} のデータを生成することで，式 3.5 における不均衡度 $r = 0.5$ に近づける手法である．

i. アンダーサンプリング

アンダーサンプリングとは，多数クラス C^{maj} のデータを間引く手法であり，選択型と生成型の 2 つのアプローチがある．

選択型は，既存のデータから間引くものを選択するアプローチであり，ランダムに選択するランダムアンダーサンプリング（Random Under Sampling: RUS），クラスをいくつかに分割して，それぞれからランダムに選択するクラスタ基準アンダーサンプリング，トメクリンク（距離が近い C^{maj} と C^{min} のデータのペア）の多数クラス C^{maj} 側のデータを間引くアンダーサンプリングがある．

生成型は，多数クラス C^{maj} の既存のデータをそのまま使用せずに，新たなデータを少数クラス C^{min} と同数生成するアンダーサンプリングである．新たなデータの生成には，k-平均法が用いられることが多い．

ii. オーバーサンプリング

オーバーサンプリングとは，少数クラス C^{min} のデータを新たに生成する手法であり，そのアルゴリズムは，さまざまなものが提案されている．

Synthetic Minority Oversampling Technique（SMOTE）は，注目している i 番目の少数クラス C^{min} のデータを $\boldsymbol{x}_i^{min}$ として，$\boldsymbol{x}_i^{min}$ と k 番目までに距離が近い k 個の $\boldsymbol{x} \in C^{min}$ を取り出す．これを，k-近傍サンプル $\boldsymbol{x}_k^{min}$ と呼ぶ．$\boldsymbol{x}_k^{min}$ のなかからランダムにひとつ選択し，これを $\boldsymbol{z}$ とする．$\boldsymbol{x}_i^{min}$ と $\boldsymbol{z}$ の内挿となる点に，$\boldsymbol{x}_i^{min}$ の新たなデータ $\boldsymbol{y}_i$

$$\boldsymbol{y}_i = \boldsymbol{x}_i^{min} + r(\boldsymbol{x}_i^{min} - \boldsymbol{z}) \tag{3.6}$$

を生成する．これを，すべての $\boldsymbol{x}_i^{min} \in C^{min}$ に対して行い，生成された $\boldsymbol{y}_i$ を学習データに追加する．ほかにも，SMOTE から派生した Adaptive Synthetic（ADASYN）や，ボーダーライン SMOTE などの手法が存在する．

アンサンブル学習

アンサンブル学習とは，複数のモデルを学習し，それらをもとに最終的な出力を決定する手法である．最終的な出力を決定するために学習させる複数のモデルを弱学習器と呼び，それぞれをどのように学習させるかによって，バギングとブースティングの 2 つに分けることができる．

バギングとは，ブートストラップ法（データからランダムに一部のサンプルを取り出すことを繰り返し，ひとつのデータセットから複数のデータセットを生成する手法）を行い，それぞれのデータセットで弱学習器を作成し，それぞれの出力をもとに，最終的な出力を決定する手法である．

ブースティングとは，n 個の弱学習器をそれぞれ直列に学習させ，それぞれの出力をもとに最終的な出力を決定するが，k 番目（ただし，$k = 1, 2, 3 \cdots, n-1$）に学習させた弱分類器の出力を，$k+1$ 番目に学習させる弱分類器に応用する手法である．

§ 3.2　土地価格決定要因の分析

2.2 節で述べたように，犯罪の発生には，数多くの要因が考えられる．そのため，犯罪の発生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].

国税調査の結果など，統計的なデータは比較的公開されているものの，犯罪発生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

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．

$$X = \lfloor \frac{lon + 180}{360} * 2^z \rfloor \tag{3.7}$$

$$Y = \lfloor \frac{\log_e tan \left(lat \frac{\pi}{180} + \frac{1}{cos\left(lat \frac{\pi}{180} \right)} \right)}{\pi} 2^{z-1} \rfloor \tag{3.8}$$

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．

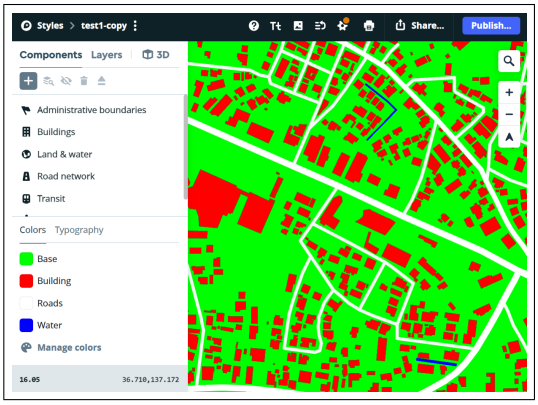


図 3.3: Mapbox Studio

図 3.4: NAVITIME

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．

§ 3.3　構造推定

機械学習，たとえば，近年急速に注目されている深層ニューラルネットワークといったアルゴリズムは，複雑かつ非線形な性質であってもモデリングすることができる．すなわち，より予測精度の大きいモデルを作成することができる．しかしながら，一般にモデルの精度が大きくなるほど，その解釈性は小さくなる性質がある．

予測モデルを解釈する重要性

近年，深層ニューラルネットワークなどの表現力の高いモデルを作成できるアルゴリズムの登場により，多くの分野で機械学習が活用されるようになってきた．医療分野では，網膜の画像から，糖尿病網膜症かどうかを診断するシステムが，米国で認可されている [29]．そのような責任が大きい判断の場合は，予測精度が大きいことはもちろん，なぜその予測値を出力したのか，その根拠も人間が知る必要がある．そのため，総務省が示している「AI 利活用原則」や，EU が施行している「一般データ保護規則（General Data Protection Regulation: GDPR）」においても，機械学習モデルの説明責任について言及しており，予測モデルを解釈することは，国際的に重要視されているといえるだろう．

予測精度と解釈性のトレードオフ

以下のような線形回帰モデルを考える．

$$f(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{3.9}$$

このとき，X_1 が 1 だけ大きくなると，$f(X_1, X_2)$ は β_1 倍だけ大きくなることが明示的に分かる．このように，線形回帰モデルは，目的変数と説明変数とのあいだに単純な関係を仮定しており，モデルに対する透明性が高いと言える．これを，一般に「解釈性が高い」と言う．

一方で，比較的近年に発表されたアルゴリズム，例えば深層ニューラルネットワークやランダムフォレストなどは，目的変数と説明変数とのあいだに線形性などの仮定を置いていない．よって，より複雑な関係をモデリングできるようになり，一般に線形回帰モデルよりも予測精度は大きくなりやすい．しかしながら，線形回帰モデルと違い，その複雑さから，なぜその予測値を出力するのかを理解することができず，その中身はブラックボックスとなりやすい．これを，一般に「解釈性が低い」と言う．

予測モデルを解釈する主な手法

機械学習によって作成されたモデルに対して，何らかの解釈を与える手法はいくつか存在するが，特に有用なものとして，以下の 4 つが挙げられるだろう．

- Permutation Feature Importance（PFI）
- Partical Dependence（PD）
- Individual Conditional Expectaion（ICE）
- Shapley Additive Explanations（SHAP）

それぞれは何を解釈できるのかが異なり，用途によって使い分ける必要がある（図 3.5 参照）．例えば，モデル全体の傾向など，マクロな視点から解釈する場合は PFI を，出力されたひとつの予測値に対する根拠など，ミクロな視点を知りたい場合は ICE を用いるべきだろう．本研究では，ミクロな視点から解釈できるものの，マクロな視点からの解釈も可能な SHAP [31] を用いることとする．

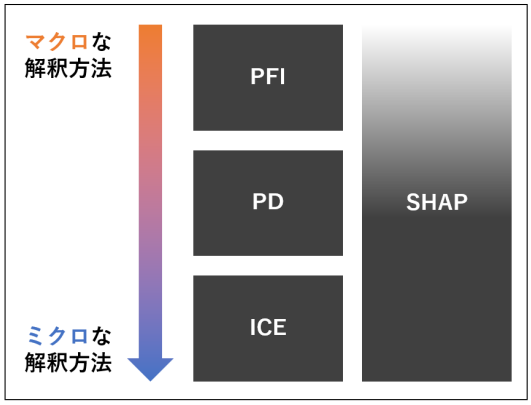


図 3.5: 機械学習モデルの解釈手法 [30]

表 3.1: アルバイトゲームの例

参加者	報酬	参加者	報酬
A	6	A・B	20
B	4	A・C	15
C	2	B・C	10
		A・B・C	24

SHAP

$\mathbf{X} = (X_1, \ldots, X_J)$ を説明変数とする学習済みのモデルを $\hat{f}(\mathbf{X})$ とする．インスタンス i の説明変数が $\mathbf{x_i} = (x_{i,1}, \ldots, x_{i,J})$ とすると，インスタンス i の予測値は $\hat{f}(\mathbf{x}_i)$ である．ここで，予測の期待値を $\mathbb{E}\left[\hat{f}(\mathbf{X})\right]$，インスタンス i の説明変数 $x_{i,j}$ の貢献度 $\phi_{i,j}$ としたとき，

$$\hat{f}(\mathbf{X}_i) - \mathbb{E}\left[\hat{f}(\mathbf{X})\right] = \sum_{j=1}^{J} \phi_{i,j} \tag{3.10}$$

のように，期待値からの差分を貢献度の総和で表現できるように，貢献度を分解することが，SHAP の基本的な考え方である．線形モデルであれば，比較的容易に分解することができるが，非線形モデルではこのままでは難しい．そのため，SHAP では，協力ゲーム理論の Shapley 値の考え方を用いて，貢献度を分解する．

ここで，協力ゲーム理論のひとつであるアルバイトゲームを説明する．アルバイトの参加者として，A，B，C の 3 つのプレイヤーを仮定し，アルバイトの参加者とそのときに得られる報酬には，表 3.1 のような関係があるとする．

A・B・C の 3 プレイヤーが参加したときの報酬は 24 である．より貢献度が大きいプレイヤーに，より多くの報酬を配分するとすれば，その貢献度はどのように算出すべきだろうか．ここで，限界貢献度という概念を導入する．限界貢献度とは，あるプレイヤーが参加したときの報酬と，参加する直前の報酬との差を表す．例えば，B・C がすでに参加しているときに A が参加した場合の限界貢献度は，$24 - 10 = 14$ である．しかし，各プレイヤーがどのような順序で参加するかにより，限界貢献度は異なる．例えば，A の限界貢献度について，A，B，C という順番で参加したときは 6 であるが，B，C，A という順序で参加したときは 14 である．

この影響を解消するため，考えられるすべての順序で限界貢献度を算出し，その平均を求めることにする．例えば，A の限界貢献度の平均値は，$(6+6+16+14+13+14)/6 = 11.5$ である．この限界貢献度の平均値を Shapley 値といい，これをもとに報酬を分配する．一般に，J つのプレイヤーが存在するとき，プレイヤー j の Shapley 値 ϕ_j は以下のように算出される．

$$\phi_j = \frac{1}{|\mathcal{J}|!} \sum_{\mathcal{S} \subseteq \mathcal{J} \setminus \{j\}} (|\mathcal{S}|!(|\mathcal{J}| - |\mathcal{S}| - 1)!)(v(\mathcal{S} \cup \{j\}) - v(\mathcal{S})) \tag{3.11}$$

SHAP は，この Shapley 値の考え方を機械学習のモデルに適用している．例えば，説明変数が X_1, X_2 であるモデルにおいて，インスタンス i の予測値 $v(\{1,2\})$ の，説明変数を $x_{i,1}, x_{i,2}$ とすると，

$$v(\{1,2\}) = \hat{f}(x_{i,1}, x_{i,2}) \tag{3.12}$$

である．また，$x_{i,1}$ と $x_{i,2}$ のいずれも未知の場合は，予測値の期待値とし，

$$v(\emptyset) = \mathbb{E}\left[\hat{f}(X_1, X_2)\right] \tag{3.13}$$

である．では，$x_{i,1}$ は既知であり，$x_{i,2}$ は不明であるときの予測値 $v(\{1\})$ は，後者について周辺化を行い，

$$v(\{1\}) = \mathbb{E}\left[\hat{f}(x_{i,1}, X_2)\right] = \int \hat{f}(x_{i,1}, x_2) p(x_2) dx_2 \tag{3.14}$$

である．よって，$x_{i,1}$，$x_{i,2}$ という順序で説明変数が判明したときの，それぞれ時点における限界貢献値 $\Delta_{i,1}, \Delta_{i,2}$ は，

$$\Delta_{i,1} = \mathbb{E}\left[\hat{f}(x_{i,1}, X_2)\right] - \mathbb{E}\left[\hat{f}(X_1, X_2)\right] \tag{3.15}$$

$$\Delta_{i,2} = \mathbb{E}\left[\hat{f}(x_{i,1}, x_{i,2})\right] - \mathbb{E}\left[\hat{f}(x_{i,1}, X_2)\right] \tag{3.16}$$

である．Shapley 値と同様に，考え得るすべての順番で算出し，それらを平均する．すなわち，説明変数 $x_{i,1}, x_{i,2}$ について，その平均値 $\phi_{i,1}, \phi_{i,2}$ は，

$$\phi_{i,1} = \frac{1}{2}\left(\left(\mathbb{E}\left[\hat{f}(x_{i,1}, X_2)\right] - \mathbb{E}\left[\hat{f}(X_1, X_2)\right]\right) + \left(\hat{f}(x_{i,1}, x_{i,2}) - \mathbb{E}\left[\hat{f}(X_1, x_{i,2})\right]\right)\right) \tag{3.17}$$

$$\phi_{i,2} = \frac{1}{2}\left(\left(\hat{f}(x_{i,1}, x_{i,2}) - \mathbb{E}\left[\hat{f}(x_{i,1}, X_2)\right]\right) + \left(\mathbb{E}\left[\hat{f}(X_1, x_{1,2})\right] - \mathbb{E}\left[\hat{f}(X_2, X_2)\right]\right)\right) \tag{3.18}$$

である．このとき，$\phi_{i,1}, \phi_{i,2}$ は，協力ゲーム理論においては Shapley 値と呼ぶが，SHAP においては SHAP 値と呼ぶ．式 3.17 と式 3.18 より，$\phi_{i,1}$ と $\phi_{i,2}$ を足すと，

$$\phi_{i,1} + \phi_{i,2} = \hat{f}(x_{i,1} + x_{i,2}) - \mathbb{E}\left[\hat{f}(X_1, X_2)\right] \tag{3.19}$$

であり，式 3.10 と同様に，インスタンス i の予測値と，予測の期待値との差分になっていることが分かる．

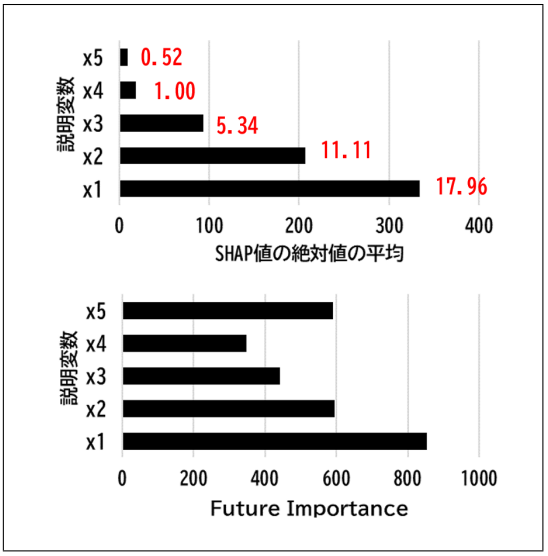


図 3.6: 等解像度データの結果

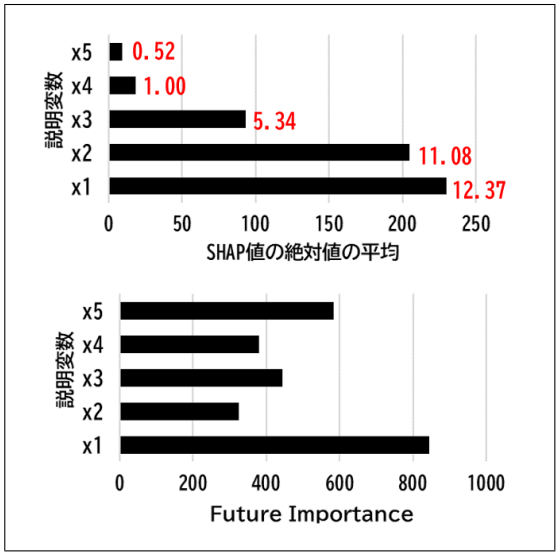


図 3.7: 非等解像度データの結果

SHAP の有用性

SHAP は，ブラックボックスなモデルであっても，なぜその予測値を出力したのか，説明変数ごとにその貢献度を出力できる．その貢献度がどれほど的確に推定できているかを検証したところ，回帰問題について，既存の解釈手法より正しく貢献度が推定できることが示された [32].

平均 0，標準偏差 30 の正規分布 $N(0, 30^2)$ に従う 5 つの説明変数 x_1, x_2, x_3, x_4, x_5 と，平均 0，標準偏差 10 の正規分布 $N(0, 10^2)$ に従うノイズ b を生成し，式 3.20 および式 3.21 で教師データをそれぞれ 10000 行作成する．式 3.21 で作成した教師データについては，x_2 だけ十の位で四捨五入し，ほかの説明変数とデータの解像度が異なるケースを再現する．

$$y = 15x_1 + 10x_2 + 5x_3 + x_4 + 0.3x_5 + b \tag{3.20}$$

$$y = 10x_1 + 10x_2 + 5x_3 + x_4 + 0.3x_5 + b \tag{3.21}$$

それぞれの教師データを，決定木をベースとした機械学習アルゴリズムである XGBoost でモデルを学習した．そのモデルを SHAP で解釈した結果と，従来手法である Future Importance で解釈した結果を比較する．式 3.20 による教師データの結果を図 3.6，式 3.21 の結果を図 3.7 に示す．なお，Future Importance は，ある説明変数が予測精度をどれだけ向上させたかを，その説明変数の「重要度」として示した値である．図中の赤字で書かれた値は，x_4 の大きさを 1 としたときの各説明変数の比率であり，Future Importance と比較しても，おおむね正確に貢献度を推定できていることが分かる．

第4章

提案手法

§ 4.1　多様な要因を考慮したデータセットの作成

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.

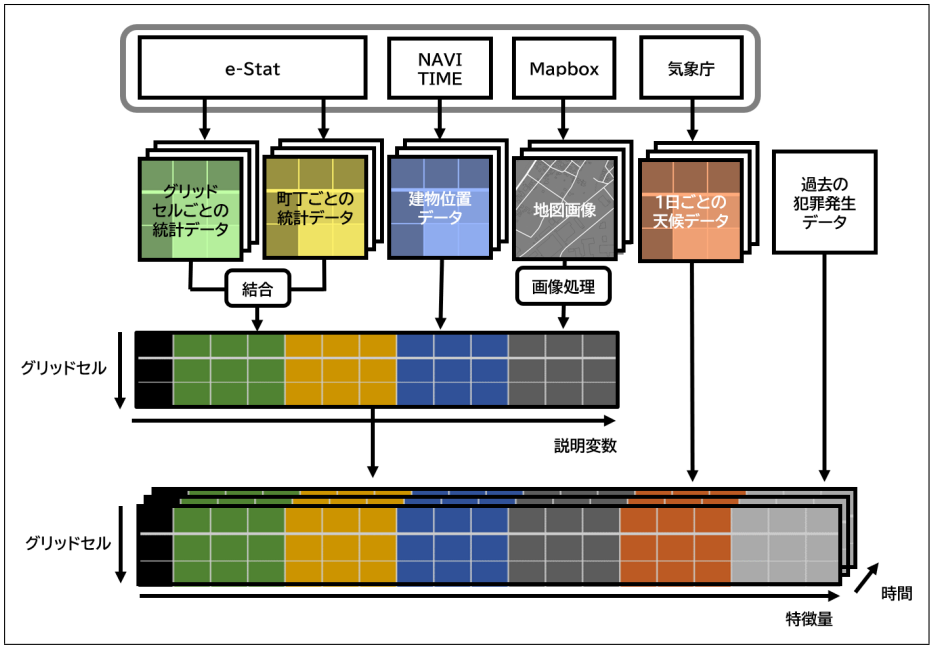


図 4.1: データセットを作成するまでの流れ

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

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

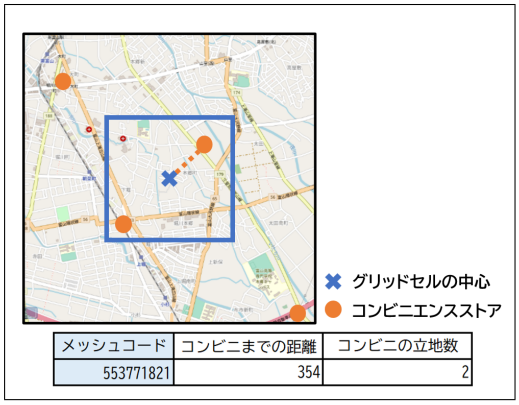
さまざまなジャンルの施設について，NAVITIME からスクレイピングを行い，施設名と，その緯度と経度を取得する．本システムでは，それぞれのジャンルごとに，対象のグリッドセルに含まれる数と，最も近くにある施設までの距離を説明変数とする．

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 [33] 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．

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

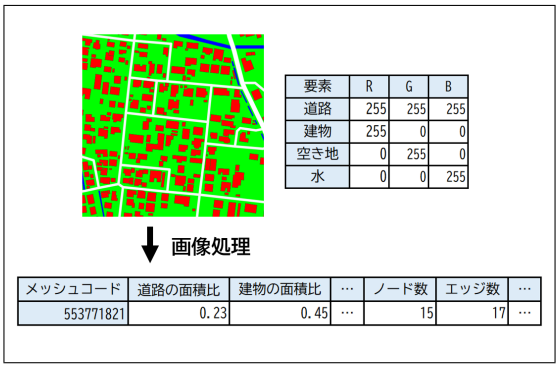
$$M = \frac{R_x(1 - E^2)}{W^3} \tag{4.3}$$

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



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 [34] 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 [35]．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　不均衡なデータに対処した予測モデルの構築

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．

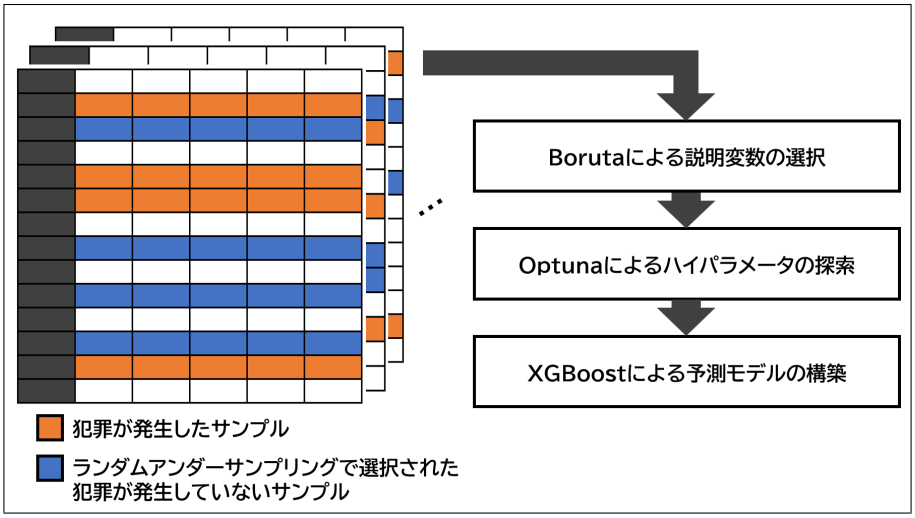


図 4.4: 予測モデルを作成する流れ

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 [36] 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．

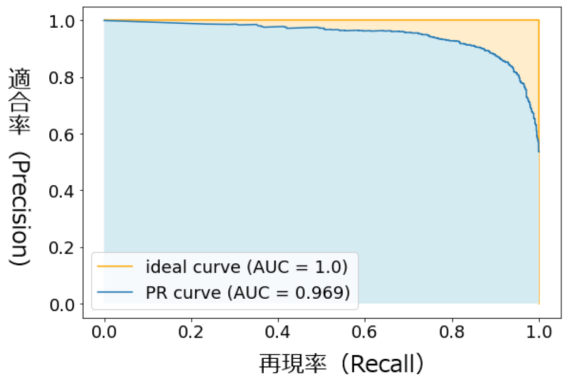


図 4.5: PR-AUC のグラフ例 [37]

		予測値	
		0	1
実測値	0	TN	FP
	1	FN	TP

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

$$F_1スコア=\frac{2TP}{2TP+FP+FN}$$

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　犯罪発生要因の可視化

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

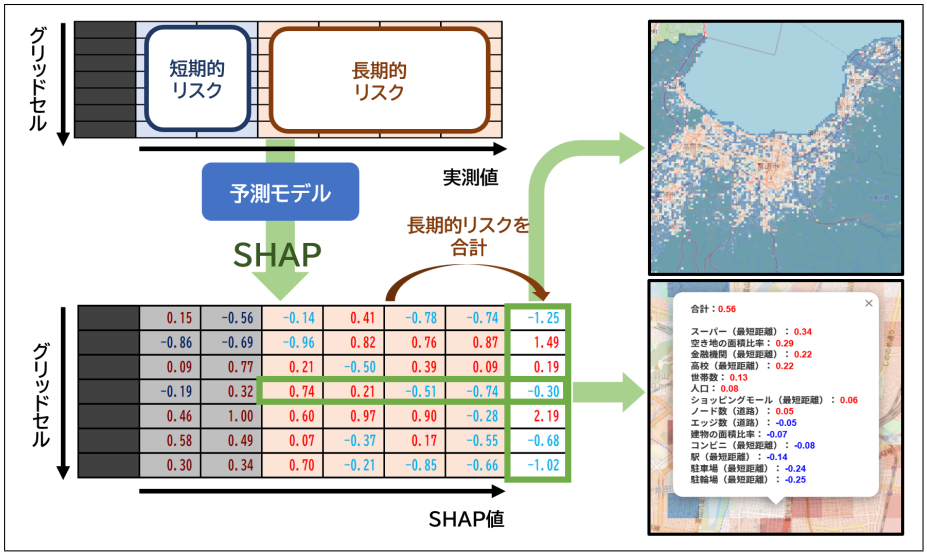


図 4.7: 要因マップを作成する流れ

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

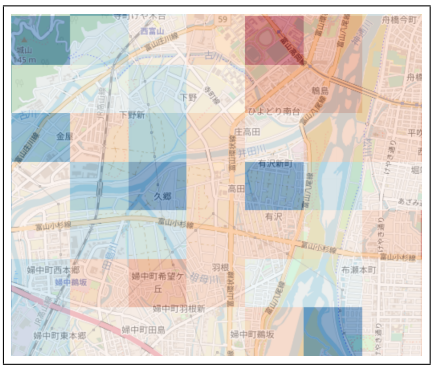


図 4.8: 犯罪発生要因マップの例

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」[38] 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 3.5 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．

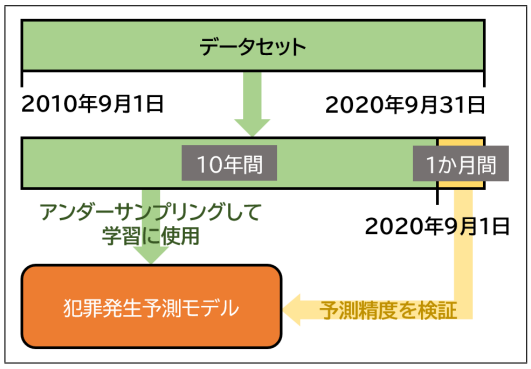


図 5.1: 予測モデルを検証する流れ

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

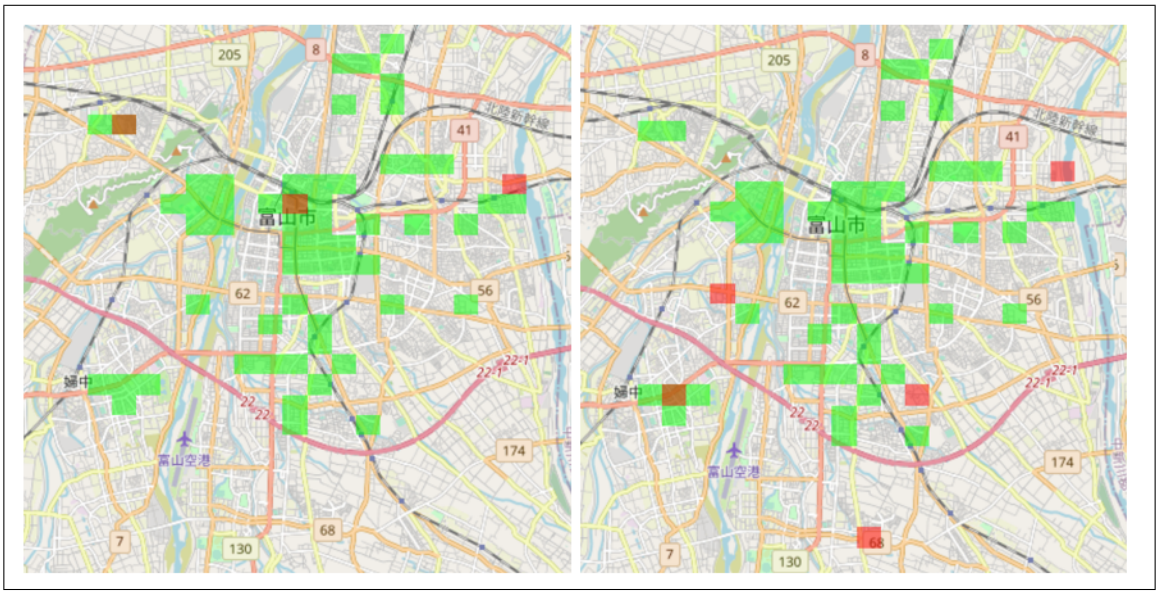


図 5.2: 2020 年 9 月 1 日（左）と 2 日（右）の予測結果

§ 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

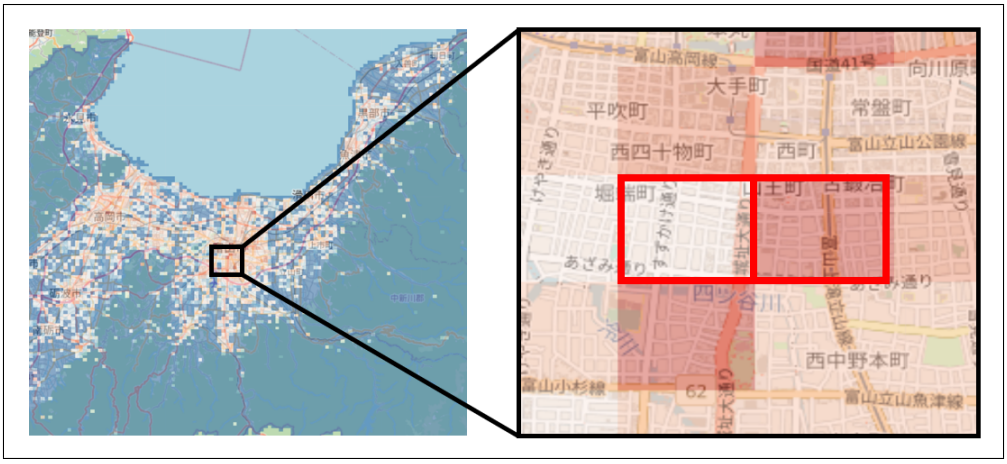


図 5.3: 予測モデルを可視化した結果

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] 張暁斉, 米澤剛, 吉田大介, “オープンデータと LSTM を用いた犯罪発生の予測及び時間的近接性における考察”, 情報学, Vol. 16, No. 1, 2019

[2] 国連薬物犯罪事務所（UNODC）, ”dataUNODC”, https://dataunodc.un.org/, 2023 年 1 月 10 日閲覧

[3] Cohen, L., Felson, M., and Land, K., “Property crime rates in the united states: Amacrodynamic analysis, 1947-1977; with ex ante forecasts for the mid-1980s.”, *American Journal of Sociology*, Vol. 86, No. 1, pp. 90-118, 1980.

[4] Sherman, L., Gartin, P., and Buerger, M., “Hot spots of predatory crime: Routine activities and the criminology of place.”, *Criminology*, Vol. 27, No. 1, pp. 27-56, 1989.

[5] Groff, E., and La Vigne, N., “Forecasting the future of predictive crime mapping”, *Crime Prevention Studies*, Vol. 13, pp. 29-58, 2002.

[6] 雨宮護, “犯罪心理学辞典”, 丸善出版, 2016.

[7] 大山智也, “日本における地理的犯罪予測手法の開発に関する研究”, 筑波大学システム情報工学研究科博士論文, 2020.

[8] 野貴泰, 糸井川栄一, “犯罪多発地点の予測に基づく防犯パトロール経路に関する提案”, 地域安全学会論文集, No.31, 2017

[9] 花岡和聖, “公然わいせつに関連する犯罪発生場所の時間的・地理的特徴 : 地理情報システムを活用した空間分析”, 立命館大学人文学会, Vol. 649, pp. 197-205, 2017

[10] 花岡和聖, “大阪府における不審者遭遇情報の地理的分布: Risk Terrain Model を用いた犯罪リスクのマッピング”, 立命館大学人文学会, Vol. 656, pp. 708-720, 2018

[11] 森本修介, 川向肇, 申吉浩, “情報伝搬モデルとガウス過程に基づく犯罪予測”, 人工知能基本問題研究会, No. 109, 2019

[12] 西颯人, 樋野公宏, “オープンデータを用いた深層学習による犯罪発生予測の試み”, 公益社団法人日本都市計画学会 都市計画報告集, No. 16, 2017

[13] 法務省, “平成 30 年版犯罪白書”, https://hakusyo1.moj.go.jp/jp/65/nfm/mokuji.html, 2023 年 1 月 10 日閲覧.

[14] Giménez-Santana, A., Caplan, J., and Drawve, G., “Risk terrain modeling and socioeconomic stratification: identifying risky places for violent crime victimization in Bogotá, Colombia”, *European Journal on Criminal Policy and Research*, Vol. 24, pp 417-431, 2018.

[15] Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., and Pentland, A., “Once upon a crime: towards crime prediction from demographics and mobile data”, *In Proceedings of the 16th international conference on multimodal interaction*, pp. 427-434.

[16] Gorr, W., and Harries, R., “Introduction to crime forecasting. International Journal of Forecasting”, Vol. 19, No. 4, pp. 551-555, 2003.

[17] Ratcliffe, J., Taylor, R., and Perenzin, A., “Predictive Modeling Combining Short and Long-Term Crime Risk Potential: Final Report”, https://www.ncjrs.gov/pdffiles1/nij/grants/249934.pdf, 2023 年 1 月 31 日閲覧.

[18] Taylor, R., Ratcliffe, J., and Perenzin, A., “Can we predict long-term community crime problems? The estimation of ecological continuity to model risk heterogeneity”, *Journal of research in crime and Delinquency*, Vol. 52, No. 5, pp. 635-657.

[19] Sampson, R., Lauritsen, J. L., “Violent victimization and offending: Individual situational- and community-level risk factors.” *Understanding and Preventing Violence*, Vol. 3, No. 1

[20] Crowe, T., “Crime prevention through environmental design: Applications of architectural design and space management concepts.”, *Boston: Butterworth-Heinemann*, 1991.

[21] Gorr, W., and Olligschlaeger, A., “Crime hot spot forecasting: Modeling and comparative evaluation, final project report. Washington”, DC: National Criminal Justice Reference Service, 2002.

[22] 長瀬永遠, “証拠に基づく政策立案のためのオープンデータを利活用した Web-GIS 可視化によるデータフュージョン”, 富山県立大学学士論文, 2022.

[23] Alberto R. Gonzales, Regina B. Schofield, Sarah V. Hart, “Mapping Crime: Understanding Hot Spots”, https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=06b4dc6a56092493439faf0f5367293cca3f00b7#page=7, 2023 年 1 月 31 日閲覧.

[24] Gorr, W., and Harries, R., “Introduction to crime forecasting”, *International Journal of Forecasting*, Vol. 19, No. 4, pp.551-555.

[25] Government Technology Magazine, “The Role of Data Analytics in Predictive Policing”, https://www.govtech.com/data/Role-of-Data-Analytics-in-Predictive-Policing.html, 2023 年 1 月 31 日閲覧.

[26] Nathalie Japkowicz, “The class imbalance problem: Significance and strategies”, *In: Proc. of the Int’l Conf. on Artificial Intelligence*, 2000.

[27] 藤原幸一, “スモールデータ解析と機械学習”, オーム社, 2022

[28] Open Knowledge Foundation, “Place overview - Global Open Data Index”, http://index.okfn.org/place.html, 2023 年 2 月 23 日閲覧.

[29] 亀谷由隆, “説明可能 AI 技術のこれまでとこれから”, 電子情報通信学会 基礎・境界ソサイエティ Fundamentals Review, No. 16, Vol. 2, pp. 83-92, 2022.

[30] 森下光之助, “機械学習を解釈する技術”, 技術評論社, 2021

[31] Scott M. Lundberg, Su-In Lee, “A Unified Approach to Interpreting Model Predictions”, *Neural Information Processing Systems*, Vol. 31, 2017

[32] 吉田秀穂，田嶋優樹，今井優作，“決定木ベースモデルの解釈における SHAP 値の有用性の検討”, 人工知能学会全国大会論文集, Vol. 34, 2020

[33] 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.

[34] “danvk/extract-raster-network: Extract a network graph (nodes and edges) from a raster image”, GitHub, https://github.com/danvk/extract-raster-network, 2023 年 2 月 7 日閲覧.

[35] 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

[36] Kursa, M. B., Rudnicki, W. R., “Feature Selection with the Boruta Package”, *Journal of Statistical Software*, Vol. 36, No. 11, pp. 1–13, 2010.

[37] 一色政彦, “第 11 回　機械学習の評価関数（二値分類／多クラス分類用）を理解しよう：TensorFlow 2 ＋ Keras（tf.keras）入門 - @ IT”, https://atmarkit.itmedia.co.jp/ait/articles/2103/04/news023.html, 2023 年 1 月 31 日閲覧.

[38] 富山県警察，“富山県警察　犯罪発生マップ”, http://www.machi-info.jp/machikado/police_pref_toyama/index.jsp, 2023 年 1 月 31 日閲覧.