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

住宅市場における情報整備の状況

§ 2.1　不動産価格の実務的な定義

2.1.1　取引価格と取引事例

一般に経済活動において「価格」といった場合,「取引価格」をさす. しかし, 不動産市場においては相対取引が前提となることから, 取引を前提として売り手が提示するいわゆる言い値（Asking Price）と「成約価格」が異なることを留意しなければならない.

日本では, 主要先進諸国と比較して実際の取引価格情報を得ることはきわめて難しい. しかしながら, 実は「取引事例」と呼ばれる取引価格に関する情報源は存在する. 取引事例は, 日本を代表する政府が提供する地価情報である公示地価および不動産鑑定評価の基礎的情報であり, 国土交通省によって地価公示法に基づき情報収集・整備が行われる.

取引事例の作成手続きとしては, 不動産取引の存在の確認から始めることが必要となる. 不動産の取引が行われると, 法務局に登記される. その情報を法務省から国土交通省へと送付されることとなっており, 国土交通省はその「買い手」に対してアンケート調査により「価格」を調べる. アンケート調査によって回収された価格情報は, 国土交通省によって開発された地理情報システムを用いたシステム, または不動産鑑定士及びそれに準ずる者によって「敷地条件（前面道路幅員等）」「街路条件」「最寄駅」「最寄駅までの距離（道路距離または直線距離）」等の利便性,「公法上の規制（都市計画用途・容積率等）」,「取引事情（売り進み・買い進み）」などが記載され,「取引事例カード」として整理される.

ただ, このように整備された「取引事例」は, 以下の点で注意が必要である. 一般に, 不動産の取引は更地として取引されることは少なく,「建付地」としての取引が中心である. つまり, 建物価格と土地価格の合計額としての不動産の取引価格が市場で観察されることになる.

ここで, 建物が完成した後の総費用は, 建物の延べ床面積 S と単位面積当たりの建築費 β_t, および土地の面積 L と単位面積当たりのコスト α_t に等しい.α_t, β_t は単位当たり（例えば m^2 あたり）の土地と建物の市場価格であり, 時間とともに変化していく.

ここで, 取引期 t における延べ床面積 $S_{t,n}$, 土地面積 $L_{t,n}$ で取引価格が $V_{t,n}$ であるような不動産を考える. この時, これらの価格が土地と建物価格の総和に誤差項 $\varepsilon_{t,n}$ を加えたものに等しいとする. すると, 取引期 t における土地価格 α_t と建物価格 β_t と不動産の総額 $V_{t,n}$ は, 以下のように表現することができる.

$$V_{t,n} = \alpha_t L_{t,n} + \beta_t S_{t,n} + \varepsilon_{t,n} \quad (t = 1, \ldots, 44;\ n = 1, \ldots, N(t)). \tag{2.1}$$

この式は取引期 t, 物件 n における土地面積 $L_{t,n}$ と建物の延べ床面積 $S_{t,n}$ という観察さ

れる量と, 時点 t における土地の市場取引価格単価 α_t と建築コストの平米単価 β_t という一定品質の価格から成り立っている. そのため, この式によって定義されるモデルは, 新築の場合に相当している.

一般的に市場で観察される不動産の取引価格は中古物件が多い. その場合には, 経年減価によって新築物件よりも価格が安くなる. そこで, 物件 n の取引期 t における建築後年数 $A(t,n)$ が分かっていれば, 以下の式のように変換できる.

$$V_{t,n} = \alpha_t L_{t,n} + \beta_t (1-\delta)^{A(t,n)} S_{t,n} + \varepsilon_{t,n}. \tag{2.2}$$

ここで,δ は正味の経年減価率を表している.

不動産鑑定士は, 取引事例を用いた「取引事例比較法」, 不動産の建築コストに注目した「コスト法」, そして将来収益の割引現在価値として求める「収益還元法」の 3 つの手法を用いて不動産の価値を決定している.(36) 式は, 取引事例比較法の重要な情報となる「取引事例」の生成と「コスト法」において根幹的なモデルとなる. そのため, 建物込みの価格として取引が行われた価格がアンケート調査によって収集される. そこで不動産鑑定士によって建物の価格が評価され, 取引価格総額から評価された建物価格を引いて, 取引事例地価が算定されるのである. この場合, 建物の価格をどのように評価するのかによって地価が変化することとなる.

市場で正確に観察可能となるのが, 不動産の取引総額である V_{tn} と土地面積 L_{tn} と建物の延べ床面積 S_{tn} であり, 時点 t における土地の市場取引価格単価 α_t と建築コストの平米単価 β_t, そして経年減価率 δ を推計しないといけないのである. そうすると, 土地の「取引事例」は建物込みで取引された場合には, 不動産鑑定士のフィルターを通した価格情報となっており, 純粋な意味での取引価格ではない.

現在は, わが国の不動産の取引価格の収集は, アンケート調査に依存していることから, 回収率にばらつきがあり, かつシェアもそれほど高くないことに加え, 申告された価格の正確度（accuracy）なども問題である. 買い手が虚偽の申告をしたとしても, それを確認する手段を持ち合わせていない. さらに, 現行のアンケートをベースにした調査方法に頼る以上, 情報入手に時間がかかり, 特に市場の変革期には情報鮮度といった意味でも問題がある. したがって取引事例に依拠して不動産市場を分析する場合, これらの問題を含むデータであることに留意する必要がある.

2.1.2 鑑定価格と課税価格

鑑定価格は, 不動産鑑定士によって評価された価格を意味する. わが国における不動産鑑定評価制度は, 昭和 38 年の「不動産鑑定評価に関する法律 昭和 38 年法律第 152 号」に基づき確立されたものであり, 費用から算定する原価法, 土地の収益を「適正な割引率」を設定した上で現在価値として求める収益還元法, 近隣の相応する土地の取引事例をもとに求める取引事例比較法の 3 手法を比較考慮した上で決定されることとなっている.

同制度は, 昭和 38 年 6 月 8 日に建設大臣から「最近における宅地価格の騰貴及び宅地の入手難が, 国民経済の健全な成長及び国民生活の安定に重大な障害を及ぼしている現状にかんがみ, 宅地価格の安定, 宅地流通の円滑化, 宅地の確保及び宅地の利用の合理化を図るため

に, いかなる制度上の措置を講ずるべきか」という諮問を受け, 宅地制度審議会において審議が開始され, 制度化に至った（小林, 1964）.

つまり, 初期の問題意識のなかには,「地価抑制」という考えが前提にあり, そのために「正常価格」という概念が登場する. こうした背景の元に,「あるべき価格 sollen」として評価すべきか「あるがままの価格 sein」として評価すべきかといった議論が長く続いた（門脇, 1981; pp.49-53）. このような議論に対して,1980 年 7 月に日本不動産鑑定協会は, 正常価格とは「市場性を有する不動産について合理的な自由市場で形成されるであろう市場価値を表示する適正な価格をいう」と定義し,「市場統制がなくて需要, 供給が自由に作用しうる市場において, 市場の事情に十分に通じ, かつ, 特別な動機を持たない多数の売り手と買い手とが存在する場合に成立する価格」であるとした. 後者の限定は, 市場が急速に変化する局面において,「正常価格」をどのようにして捉えるのかについて鑑定士の判断が強く働くことを意味し, そうした事態での鑑定価格の恣意性に対する疑いを呼び起こすことになった.

そのような中で,2000 年代初頭に始まった日本版不動産投資信託 J-REIT において, 不動産の鑑定評価額は市場と向き合うことが余儀なくされた.J-REIT に運営会社は, 物件の取得においては不動産鑑定評価額を参考にしなければならず, また定期的に各保有物件の鑑定価格を開示することが要求されたためである. そのような中で, 不動産鑑定評価制度は, 市場と対峙することが求められ,「あるがままの価格 sein」として評価を求められることになっていった.

一方で, 取引価格・鑑定評価価格に加え, 課税のための価格が存在している. 具体的には, 評価が必要とされる税の評価は, 固定資産税は市町村長が, 不動産取得税における評価は原則として固定資産税における評価に基づき都道府県知事が, 相続税・贈与税における評価は税務署長が, 登録税における評価は税務官吏が行うものとされている. それぞれの税の目的が鑑定評価でいう「正常価格」が想定する市場と異なることから, 公的評価間にもアンバランスがあることが指摘されてきた.

特に, 固定資産税評価と相続税のための路線価評価においては, 各自治体及び国税庁担当者が独自に評価をしていたことから, 固定資産税路線価においては自治体間や同一自治体内においても用途間などにおいて均衡が保たれておらず, さらに固定資産税路線価と相続税路線価との間にも大きな乖離が存在していた. そのため, 取引価格・公示地価とあわせてしばしば一物四価と揶揄され, 社会的な問題にまで発展した.

そのような中で,1989 年に制定された「土地基本法」,1991 年に閣議決定された「総合土地政策推進要綱」に基づき, 各公的土地情報間の整合性を確保することの必要性が指摘され, 相続税路線価は 1992 年以降公示地価の 8 割を目途に, 固定資産税路線価は 1994 年評価替え以降では公示地価の 7 割を目途に評価が行われている. ただし問題を更に複雑にしているのは, 固定資産税の場合, この評価額が必ずしも課税標準額でないことである. 急激な税額の上昇を抑えるために, 評価額は上に見た公示地価を基準としたものながら実際の課税標準額は緩やかに上昇させるように負担調整率が適用されている. そのため一物四価と呼ばれた状態は課税標準額で見る限り解消していない.

経済全体の縮退に伴い, 市町村の固定資産税収は減少してきている. そのような中で, 税の評価の在り方は抜本的に見直さないといけない時期に直面してきているものと考える.

§ 2.2　価格情報

住宅価格に関する情報は数多く, 表 1 に整理される. 価格情報のなかには, 市場において売買された「取引価格情報」以外に公示地価に代表される鑑定評価情報があり, さらに課税目的のために整備される相続税・固定資産税のための 2 つの「課税評価価格」といった複数の情報体系が存在している. それぞれの情報は, 目的に応じた特色を持っているが, データの性質に着目すると, 特定の土地・地域の価格水準を測る水準指標か, あるいは, 時系列的な価格変化の観察を目的とする価格指数か, に大別される.

まず, 非集計の水準指標は, 基本的には特定の土地・地域の価格水準を調べることを主目的としており, その場合には鑑定情報, 相場・取引情報かに大別される.

具体的にわが国で公的部門により公表される鑑定価格情報としては, 国土交通省による「地価公示」, 各都道府県による「地価調査」, 国税庁による「相続税路線価」, 各市町村による「固定資産税路線価」「固定資産税・標準宅地鑑定価格」が存在する.

公的な鑑定価格情報であっても, 相続税路線価, 固定資産評価額は課税目的の価格情報であり, 路線を単位として情報が提供されている. そのため相続税・固定資産税のそれぞれの課税目的の相違から, 違った価格査定基準を持って地価を捕捉している.

市場の相場・取引情報としては, 全国指定流通機構連絡協議会（不動産流通機構）による「REINS Market Information」（売り出し価格）があり, 戸建, マンションについて, 全国各地の情報が提供されている. 国土交通省による「不動産取引価格情報」（売買成約価格）では, 宅地（土地, 土地と建物）, 中古マンション等について, 全国各地の情報が提供されている. なお,Web 情報基盤の発達により, 昨今では, 不動産流通のポータルサイト（SUUMO,LIFULL HOME'S 等）上でも, 建て方・間取り別に, 価格相場情報（平均値）を得ることができる.

次に, 時間的な価格変化を観察することを目的として, 平均値等の単純集計の情報が公表されている. 国土交通省の「地価 LOOK レポート」では, 全国主要都市圏単位にとどまらず, 具体的な地点においても, 住宅地地価の変動動向を整理している.

不動産流通推進センターの「成約価格（不動産業統計集）」では, 住宅新報社「住宅新報」に基づき, 首都圏, 近畿圏における, 中古マンション, 戸建住宅, 土地の価格について, その平均値を公表している. 不動産経済研究所の「マンション・建売住宅市場動向」では, 首都圏・近畿圏において, マンション・建売住宅の平均価格が公表されており, さらに, 超高層マンション, コンパクトマンションについても情報が提供されている.

さらに, 品質調整を行い構築された指数情報として, 国土交通省の「不動産価格指数」がある. 全国・ブロック別・都市圏別・都道府県別に, 住宅総合, その内訳として住宅地・戸建住宅・マンション（区分所有, 主に中古）の各指数が公表されている.

民間の調査機関等による情報も多数存在する. 日本不動産研究所の「市街地価格指数」は, 鑑定評価の手法に基づき, 宅地価格を評価し指数化したものであり, 全国を対象に, 地方・都市圏等の単位で公表されている. 同じく日本不動産研究所の「不動研住宅価格指数（旧：東証住宅価格指数）」は, 東京証券取引所が公表してきた「東証住宅価格指数」（2011 年〜2014 年）を引き継ぐものである.

東日本不動産流通機構より提供された既存マンション（中古マンション）の成約価格情報を活用し, 同一物件の価格変化に基づき, 首都圏, その内訳として東京都, 神奈川県, 千葉県, 埼玉県について指数を算出している.

リクルート住まいカンパニー・MSCI INC. による「IPD/リクルート日本住宅指数（RRPI）」

種別	調査名	調査機関	性格	周期	開始時点
非集計	地価公示	国土交通省	鑑定価格	年 1 回	1970
	地価調査	都道府県	鑑定価格	年 1 回	1975
	相続税路線価	国税庁	査定価格	年 1 回	1963
	固定資産税路線価	市町村	査定価格	3 年毎	1950
	固定資産税・標準宅地鑑定価格	市町村	査定価格	3 年毎	1994
	REINS Market Information	全国指定流通機構連絡協議会（不動産流通機構）	市場	日次	2007
	不動産取引価格情報	国土交通省	取引価格	四半期	2005
集計・指数	地価 LOOK レポート	国土交通省	変動動向	四半期	2007
	不動産業統計集	不動産流通推進センター	取引価格（平均）	月次	1989
	マンション・建売住宅市場動向	不動産経済研究所	取引価格（平均）	月次	1988
	不動産価格指数（住宅）	国土交通省	取引価格（ヘドニック指数）	月次	2008
	市街地価格指数	日本不動産研究所	鑑定（指数）	年 2 回	1955
	不動研住宅価格指数（旧：東証住宅価格指数）	日本不動産研究所	取引価格（リピート・セールス指数）	月次	1993
	IPD/リクルート日本住宅指数（RRPI）	リクルート住まいカンパニー、MSCI INC.	取引価格（ヘドニック指数）	月次	1986
	住宅マーケットインデックス	日本不動産研究所、アットホーム、ケン・コーポレーション	補正（指数）	年 2 回	2001
	LIFULL HOME'S PRICE INDEX	LIFULL	ヘドニック指数	月次	2010
	J-REIT 不動産価格指数	三井住友トラスト基礎研究所	取引価格（ヘドニック指数）	週次	2002

図 2.1: 価格情報（住宅）

では, 首都圏（一都三県）の中古マンションを対象に, リクルート住まいカンパニーの発行する「SUUMO」への登録物件のうち, 成約等を理由に登録を抹消した物件の価格情報より, 品質調整済みのヘドニック指数を月次で公表している.

日本不動産研究所・アットホーム・ケン・コーポレーションの「住宅マーケットインデックス」では, 東京 23 区内の新築マンション・中古マンション事例を対象に, 統計的手法を用いて築年数についての補正を行い, エリア別, 面積別, 期間別に集計し公表されている.

LIFULL の「LIFULL HOME ' S PRICE INDEX」では, 中古マンション, 中古戸建てを対象に, 関東（東京 23 区, 東京都下, 横浜市, さいたま市, 千葉市）, 関西・中部（大阪市, 京都市, 神戸市, 名古屋市）, 札幌市, 福岡市において, 品質調整済みのヘドニック価格指数を月次で公表している.

三井住友トラスト基礎研究所による「J-REIT 不動産価格指数（住宅）」は,J-REIT による実物不動産・不動産信託受益権の取引事例をもとに, 週次で価格指数を算出する試みである.

§ 2.3　賃貸とその他の情報

2.3.1　賃貸情報

同様に, 家賃に関しても様々な情報が存在する（表 2）.Web 情報基盤の発達により, 昨今では, 不動産流通のポータルサイト SUUMO,LIFULL HOME'S 等上でも, 建て方・間取り別に, 家賃相場（平均値）を得ることができる.

種別	調査名	調査機関	性格	周期	開始時点
集計・指数	賃料相場	全国宅地建物取引業協会連合会	平均賃料	日次	－
	マンション家賃相場（不動産業統計集）	公益財団法人不動産流通推進センター	平均賃料	年 2 回	1989
	消費者物価指数（家賃、持家の帰属家賃）	総務省統計局	指数	月次	1970
	IPD/リクルート日本住宅指数（RRPI）	リクルート住まいカンパニー、MSCI INC.	ヘドニック指数	月次	1989
	マンション賃料インデックス	アットホーム、三井住友トラスト基礎研究所	ヘドニック指数	年 4 回	2009
	全国賃料統計	日本不動産研究所	実質賃料（査定）	年次	1995
	国際不動産価格賃料指数	日本不動産研究所	実質賃料（鑑定）	年 2 回	2013
	住宅マーケットインデックス（指数）	日本不動産研究所、アットホーム、ケン・コーポレーション	築年数等を補正	年 2 回	2001
	J-REITNOI 指数	三井住友トラスト基礎研究所	NOI（ヘドニック指数）	週次	2003

図 2.2: 賃料情報（住宅）

集計された情報については, 民間の調査機関等からも多数公表されている. 全国宅地建物取引業協会連合会の「賃料相場」では全国の市区町村を単位として, 駅徒歩 10 分以内の賃貸アパート・マンション・一戸建ての平均賃料を間取り別に算出している.

公益財団法人不動産流通推進センターの「マンション家賃相場（不動産業統計集）」では, 住宅新報社「住宅新報」に基づき, 東京圏, 大阪圏, 名古屋圏, 福岡圏におけるマンション家賃の下限・平均・上限を, 間取り別に公表している.

統計処理をもとに指数化された情報については, 政府統計としては, 総務省統計局の「消費者物価指数（家賃, 持家の帰属家賃）」がある. これは, 全国や, 大都市圏・都市単位で公表されているが, 住宅の経年減価を反映していない等の課題が指摘されている.

民間の調査機関等による情報も, 多数公表されている. リクルート住まいカンパニー・MSCI INC. の「IPD/リクルート日本住宅指数（RRPI）」では, 首都圏（一都三県）の中古マンションを対象に, リクルート住まいカンパニーの発行する「SUUMO」への登録物件のうち, 成約等を理由に登録を抹消した物件の賃料情報より, 品質調整済みのヘドニック賃料指数を月次で公表している.

アットホーム・三井住友トラスト基礎研究所の「マンション賃料インデックス」では, 主要都市（東京 23 区, 東京都下, 大阪市, 大阪広域, 札幌市, 仙台市, 埼玉東南部, 千葉西部, 横浜・川崎市, 名古屋市, 京都市, 福岡市）について, 賃貸マンションの成約事例より, 品質調整済みのヘドニック賃料指数を年 4 回公表している.

日本不動産研究所の「全国賃料統計」では, 全国の主要都市の共同住宅を対象に, モデル建物の新規賃料を査定し, それに市場規模を示すウェイトを乗じて, 主要都市（圏）における賃料指数を作成している.

同「国際不動産価格賃料指数」では, 対象物件（マンション）の新築・新規契約を前提とした賃料単価を不動産鑑定士が評価したものであり, 国内の都市としては東京・大阪を対象に, 公表されており, 国際的な主要都市（ニューヨーク, ロンドンや, アジア主要都市）との国際比較が可能である.

日本不動産研究所・アットホーム・ケン・コーポレーションの「住宅マーケットインデックス（指数）」では, 東京 23 区内の賃貸マンションを対象に, 統計的手法を用いて築年数についての補正を行い, エリア別, 面積別, 期間別に集計し公表されている.

三井住友トラスト基礎研究所の「J-REIT NOI 指数」は,J-REIT の決算データより, 週次で指数を算出する試みである.

2.3.2 その他の情報

価格・家賃については比較的統計情報が充実しているものの, 市場が置かれている状況を的確に把握するためには必ずしも十分でなく, その他の情報も含めて包括的に判断することが求められる（カクテル・アプローチ）. とりわけ 2000 年以降に,J-REIT が誕生し, 不動産市場の透明性確保に向けた市場指標整備の必要性が高まっている.

その他の情報の例として, 法人価値, 利回り, 取引量, ストック量とその変化, 需給バランス（賃貸）, 物件の特徴量, 不動産業の見通し DI 等があり, 表 3 に整理される.

REIT の法人価値として, 東京証券取引所の「東証 REIT 指数」がある. 当該市場に上場する不動産投資信託全銘柄を対象とした浮動株ベースの時価総額加重型の株価指数である. さらに, 各 REIT が保有する物件の用途に着目して構成銘柄を選定した株価指数「東証 REIT オフィス指数」,「東証 REIT 住宅指数」,「東証 REIT 商業・物流等指数」も公表されている. 同種の指数は, 三井住友トラスト基礎研究所からも「SMTRI J-REIT Index®」として公表されている.

不動産投資の観点から, 個別不動産の利回りに関する情報も重要である. 日本不動産研究所の「不動産投資家調査」では, 専門家アンケートを通して期待利回り・投資利回りを公表している. 東京都内では詳細地区, その他全国主要都市において, 賃貸住宅の種類（ワンルーム・ファミリー向け等）毎に公表されている. 三井住友トラスト基礎研究所では,「J-REIT インプライド・キャップレート」として, 住宅系についても指数を公表している. インプライド・キャップレートは, 資本市場が（株価を通じて）示す J-REIT 運用不動産に対する要求利回りであり,J-REIT に対する投資判断指標として, また J-REIT の不動産投資運用における一種のハードルレートとしての意味をもつ指標である.

ARES（不動産証券化協会）では, 米国で最も普及し世界的知名度が高い NCREIF インデックスの方式に準拠した利回り指数を公表している. 実物不動産インデックスとして,「AJPI ARES Japan Property Index」（インカム /キャピタル /総合）を, 不動産ファンドインデックスとして,「AJFI ARES Japan Fund Index」（インカム /キャピタル /総合）がある. 日本不動産研究所・アットホーム・ケン・コーポレーションの「住宅マーケットインデックス」では, 東京 23 区内の新築マンション・中古マンション事例を対象に, 統計的手法を用い

種別	調査名	調査機関	性格	周期	開始時点
法人価値	東証 REIT 指数	東京証券取引所	時価総額(加重平均)	日次	2003
	SMTRI J-REIT Index®	三井住友トラスト基礎研究所	時価総額(加重平均)	日次	2001
利回り	不動産投資家調査 期待利回り/投資利回り	日本不動産研究所	専門家アンケート(中央値)	年 2 回	1999
	J-REIT インプライド・キャップレート	三井住友トラスト基礎研究所	ヘドニック指数	月次	2005
	AJPI (ARES Japan Property Index)インカム/キャピタル/総合	ARES(不動産証券化協会)	NCREIF 型(加重平均)	月次	2002
	AJFI(ARES Japan Fund Index)インカム/キャピタル/総合	ARES(不動産証券化協会)	NCREIF 型(加重平均)	月次	2002
	住宅マーケットインデックス(指数)	日本不動産研究所、アットホーム、ケン・コーポレーション	新築平均利回り(査定値)、築 10 年平均利回り(査定値)	年 2 回	2001
取引量	登記統計	法務省	取引件数合計	年 1 回	1957
	不動産取引件数・面積	国土交通省	取引件数・面積合計	月次	2005
	不動産業統計集	不動産流通推進センター	新規登録件数、成約報告件数		1989
ストック量・変化	住宅・土地統計調査	総務省統計局	戸数等	5 年毎	1948
	建築物ストック統計(住宅)	国土交通省	延床面積合計	年 1 回	1991
	マンション・建売市場動向	不動産経済研究所	供給戸数、年末在庫数、販売初月契約率等	月次	1988
	建築着工統計調査	国土交通省	着工面積合計	月次	1950
	建築物滅失統計調査	国土交通省	建築物数、戸数、床面積	月次	1951
需給バランス(賃貸)	タス空室インデックス、募集期間、更新確率・中途解約確率	タス	空室状況	月次	—
物件の特徴量	購入・賃貸可能な住宅の平均像	全国宅地建物取引業協会連合会	面積等の特徴量	日次	—
見通し	不動産市況 DI 調査	全国宅地建物取引業協会連合会	不動産事業者へのアンケート	年 4 回	2016

図 2.3: その他の情報（住宅）

て築年数についての補正を行い, エリア別, 面積別, 期間別に集計し公表されている. 利回りについては, 新築平均利回り査定値, 築 10 年平均利回り（査定値）の情報が公表されている.

不動産価格と取引量には, 一定の相関が指摘されている. すなわち, 価格上昇期には取引量も増加し, 価格下落には取引量の減少が伴う. 法務省の「登記統計」では, 建物・土地の売買による取引件数を法務局及び地方法務局単位で公表しており, 国土交通省も「不動産取引件数・面積」として, 市区町村単位で公表している. 不動産流通推進センターの「不動産業統計集」では, 売り物件新規登録件数, 売り物件成約報告件数等を, 物件種別や地域毎に公表している.

ストック量として, 総務省統計局の「住宅・土地統計調査」がわが国における住戸に関する実態並びに現住居以外の住宅及び土地の保有状況, その他の住宅等に居住している世帯に関する実態を調査している. 住宅・土地統計調査等を基に, 国土交通省の「建築物ストック統計（住宅）」では, 使途別, 構造別, 竣工年代別等に床面積の総量を推計している. 株式会社不動産経済研究所の「マンション市場動向」では, 供給戸数, 年末在庫数, 販売初月契約率等を公表している. ストック量の変化として, 国土交通省の建築動態統計調査として,「建築着

工統計調査」「建築物滅失統計調査」が実施されている. 前者は, 建築物の着工状況について建築主別の建築物の数, 床面積の合計, 工事費予定額などの結果を, 全国, 都道府県, 市区町村の地域で公表している. 後者は, 建築物の滅失状況について構造別の建築物の数, 住宅の戸数, 床面積の合計などの結果を, 全国, 都道府県の地域で公表している.

賃貸住宅市場における需給バランスとして, タスの「タス空室インデックス, 募集期間, 更新確率・中途解約確率」がある. 市況のレポートとして, 首都圏・関西圏・中京圏・福岡県において, 構造別に公表している.

市場に流通している物件の統計量として, 全国宅地建物取引業協会連合会の「購入・賃貸可能な住宅の平均像」では, 面積等の平均値を公表している.

不動産業の見通しとして, 全国宅地建物取引業協会連合会の「不動産市況 DI 調査」があり, 専門家に対し, 不動産価格や取引動向の 3 か月前と現状の比較,3 か月後の見通しを調査している.

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　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．

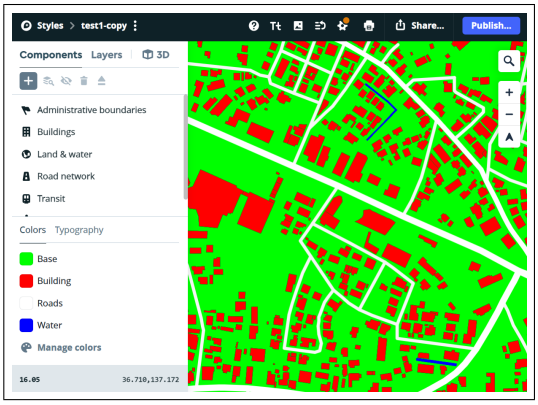


図 3.1: Mapbox Studio

図 3.2: NAVITIME

$$X = \lfloor \frac{lon + 180}{360} * 2^z \rfloor \tag{3.1}$$

$$Y = \lfloor \frac{\log_e tan \left(lat \frac{\pi}{180} + \frac{1}{cos \left(lat \frac{\pi}{180} \right)} \right)}{\pi} 2^{z-1} \rfloor \tag{3.2}$$

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．

§ 3.2　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.

$$y = p(\boldsymbol{z}) + x \tag{3.3}$$

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.

$$\max_{\boldsymbol{z}, x} \quad U(\boldsymbol{z}, x; \boldsymbol{\alpha}) \tag{3.4}$$

$$\text{s.t.} \quad y = p(\boldsymbol{z}) + x \tag{3.5}$$

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.

$$p_{\boldsymbol{z}} = \frac{U_{\boldsymbol{z}}(\boldsymbol{z}, y - p(\boldsymbol{z}); \boldsymbol{\alpha})}{U_x(\boldsymbol{z}, y - p(\boldsymbol{z}); \boldsymbol{\alpha})} = h(\boldsymbol{z}, y - p(\boldsymbol{z}); \boldsymbol{\alpha}) \tag{3.6}$$

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.

$$\frac{\partial \theta}{\partial z_i} = \frac{U_{z_i}}{U_x} > 0 \tag{3.7}$$

$$\frac{\partial^2 \theta}{\partial z_i^2} = \frac{U_{z_i}^2 U_{xx} - 2U_{z_i} U_x U_{z_i x} + U_x^2 U_{z_i z_i}}{U_x^3} < 0 \tag{3.8}$$

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.

$$\theta(\boldsymbol{z}^*; u^*, y) = p(\boldsymbol{z}^*) \tag{3.9}$$

$$\frac{\partial \theta}{\partial \boldsymbol{z}}(\boldsymbol{z}^*; u^*, y) = p_{\boldsymbol{z}}(\boldsymbol{z}^*) \tag{3.10}$$

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.

$$\max_{\boldsymbol{z}, M} \quad \pi = p(\boldsymbol{z})M - C(M, \boldsymbol{z}; \boldsymbol{\beta}) \tag{3.11}$$

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.

$$p_{\boldsymbol{z}} = C_{\boldsymbol{z}}(M, \boldsymbol{z}; \boldsymbol{\beta}) \tag{3.12}$$

$$p(\boldsymbol{z}) = C_M(M, \boldsymbol{z}; \boldsymbol{\beta}) \tag{3.13}$$

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.

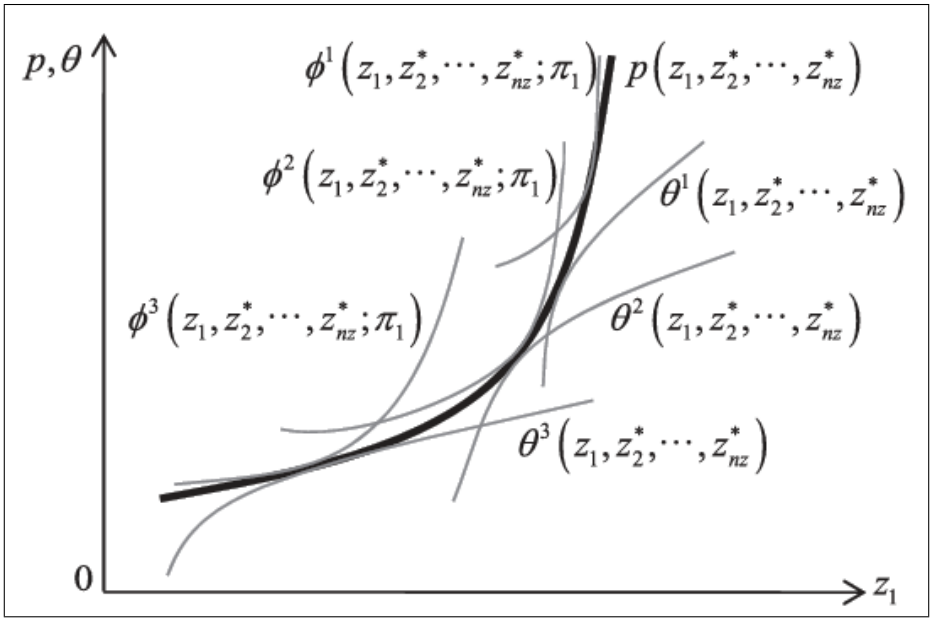


図 3.3: データセットを作成するまでの流れ

$$\phi_{\boldsymbol{z}} = \frac{C_{\boldsymbol{z}}}{M} > 0 \tag{3.14}$$

$$\phi_{\pi} = \frac{1}{M} > 0 \tag{3.15}$$

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.

$$p(\boldsymbol{z}^*) = \phi(\boldsymbol{z}^*, \pi^*) \tag{3.16}$$

$$p_{\boldsymbol{z}}(\boldsymbol{z}^*) = \phi_{\boldsymbol{z}}(\boldsymbol{z}^*, \pi^*) \tag{3.17}$$

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.

§ 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．

$$f(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{3.18}$$

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）

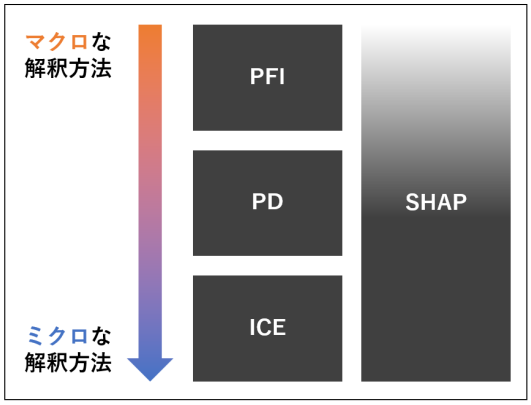


図 3.4: 機械学習モデルの解釈手法 [30]

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

- 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 3.4 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．

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，

$$\hat{f}(\mathbf{X}_i) - \mathbb{E}\left[\hat{f}(\mathbf{X})\right] = \sum_{j=1}^{J} \phi_{i,j} \tag{3.19}$$

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．

$$\phi_j = \frac{1}{|\mathcal{J}|!} \sum_{\mathcal{S} \subseteq \mathcal{J} \setminus \{j\}} (|\mathcal{S}|!(|\mathcal{J}| - |\mathcal{S}| - 1)!)(v(\mathcal{S} \cup \{j\}) - v(\mathcal{S})) \tag{3.20}$$

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，

$$v(\{1,2\}) = \hat{f}(x_{i,1}, x_{i,2}) \tag{3.21}$$

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，

$$v(\emptyset) = \mathbb{E}\left[\hat{f}(X_1, X_2)\right] \tag{3.22}$$

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，

$$v(\{1\}) = \mathbb{E}\left[\hat{f}(x_{i,1}, X_2)\right] = \int \hat{f}(x_{i,1}, x_2) p(x_2) dx_2 \tag{3.23}$$

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，

$$\Delta_{i,1} = \mathbb{E}\left[\hat{f}(x_{i,1}, X_2)\right] - \mathbb{E}\left[\hat{f}(X_1, X_2)\right] \tag{3.24}$$

$$\Delta_{i,2} = \mathbb{E}\left[\hat{f}(x_{i,1}, x_{i,2})\right] - \mathbb{E}\left[\hat{f}(x_{i,1}, X_2)\right] \tag{3.25}$$

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，

$$\phi_{i,1} = \frac{1}{2}\left(\left(\mathbb{E}\left[\hat{f}(x_{i,1}, X_2)\right] - \mathbb{E}\left[\hat{f}(X_1, X_2)\right]\right) + \left(\hat{f}(x_{i,1}, x_{i,2}) - \mathbb{E}\left[\hat{f}(X_1, x_{i,2})\right]\right)\right) \tag{3.26}$$

$$\phi_{i,2} = \frac{1}{2}\left(\left(\hat{f}(x_{i,1}, x_{i,2}) - \mathbb{E}\left[\hat{f}(x_{i,1}, X_2)\right]\right) + \left(\mathbb{E}\left[\hat{f}(X_1, x_{1,2})\right] - \mathbb{E}\left[\hat{f}(X_2, X_2)\right]\right)\right) \quad (3.27)$$

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，

$$\phi_{i,1} + \phi_{i,2} = \hat{f}(x_{i,1} + x_{i,2}) - \mathbb{E}\left[\hat{f}(X_1, X_2)\right] \quad (3.28)$$

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．

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．

$$y = 15x_1 + 10x_2 + 5x_3 + x_4 + 0.3x_5 + b \quad (3.29)$$

$$y = 10x_1 + 10x_2 + 5x_3 + x_4 + 0.3x_5 + b \quad (3.30)$$

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．

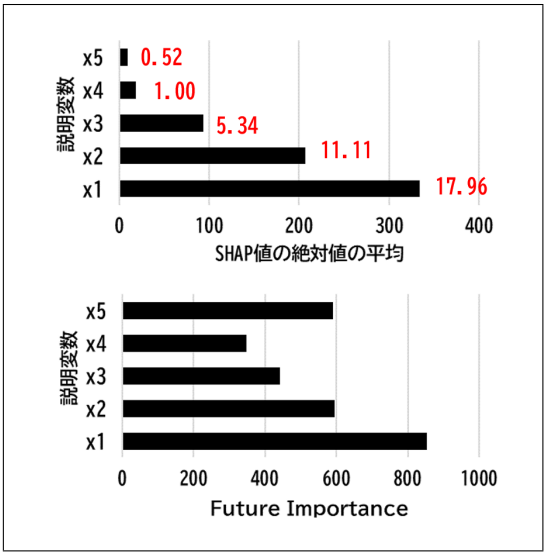


図 3.5: 等解像度データの結果

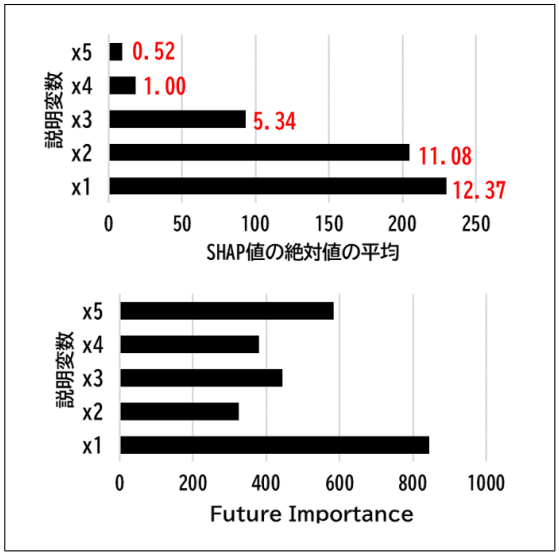


図 3.6: 非等解像度データの結果

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.

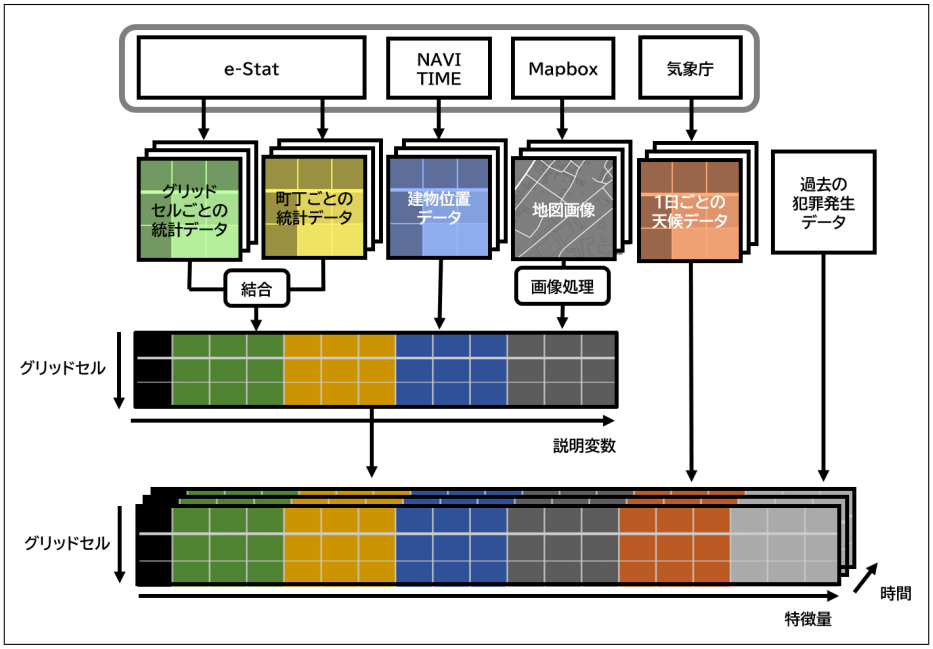


図 4.1: データセットを作成するまでの流れ

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

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

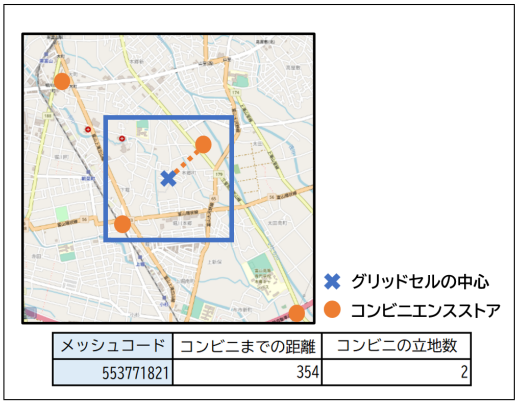
さまざまなジャンルの施設について，NAVITIME からスクレイピングを行い，施設名と，その緯度と経度を取得する．本システムでは，それぞれのジャンルごとに，対象のグリッドセルに含まれる数と，最も近くにある施設までの距離を説明変数とする．

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．

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

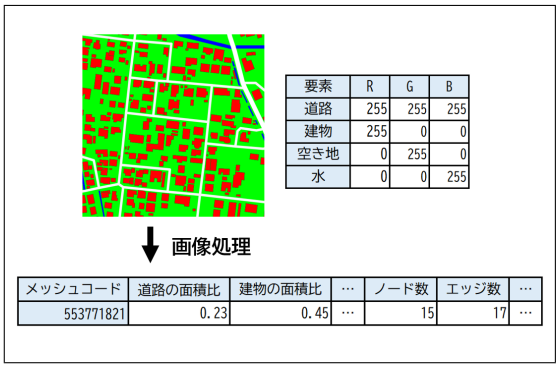
$$M = \frac{R_x(1 - E^2)}{W^3} \tag{4.3}$$

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



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．

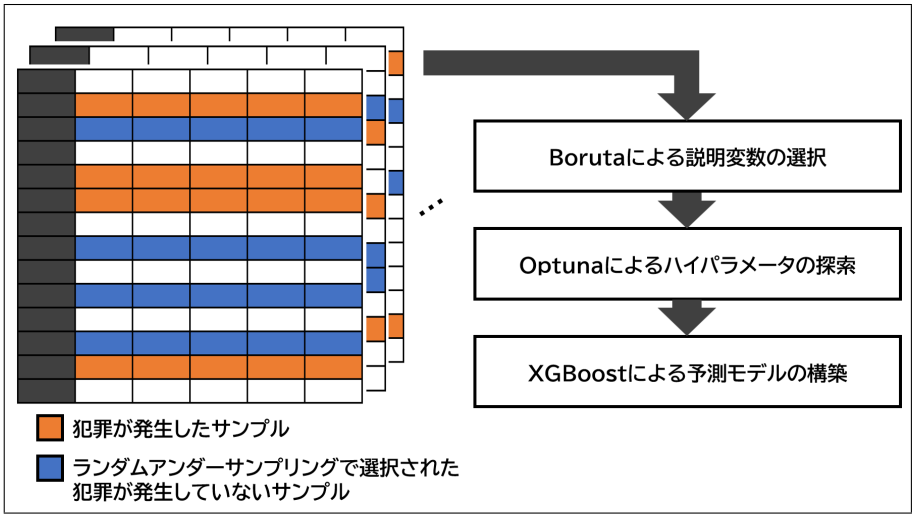


図 4.4: 予測モデルを作成する流れ

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．

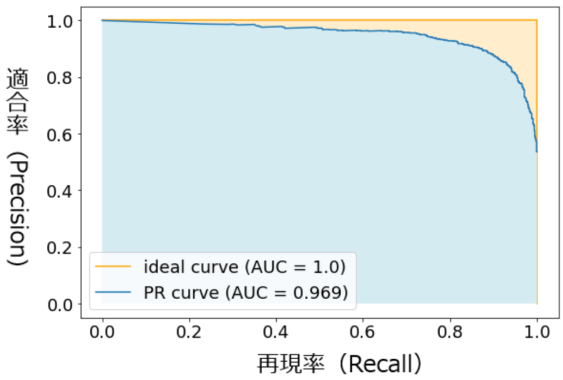


図 4.5: PR-AUC のグラフ例 [37]

		予測値	
		0	1
実測値	0	TN	FP
	1	FN	TP

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

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

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　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

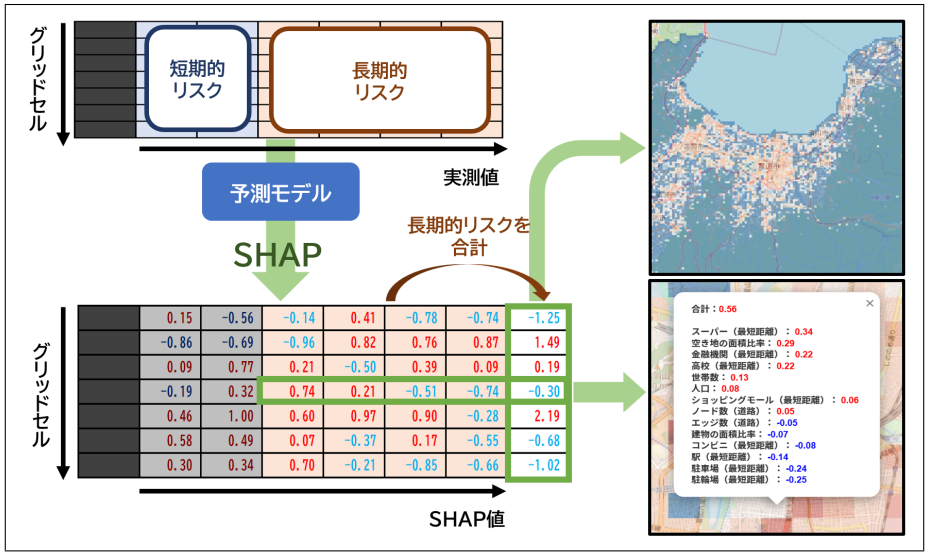


図 4.7: 要因マップを作成する流れ

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

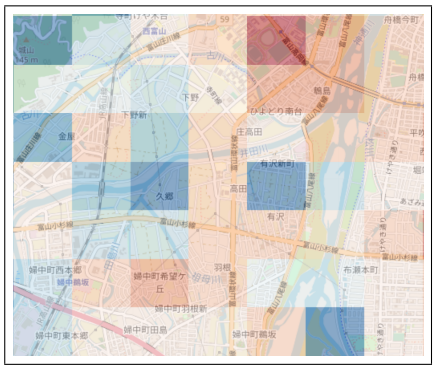


図 4.8: 犯罪発生要因マップの例

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??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．

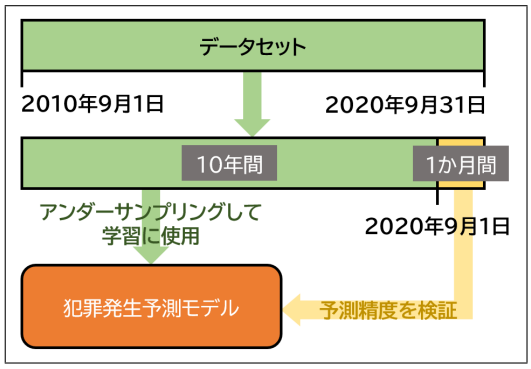


図 5.1: 予測モデルを検証する流れ

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

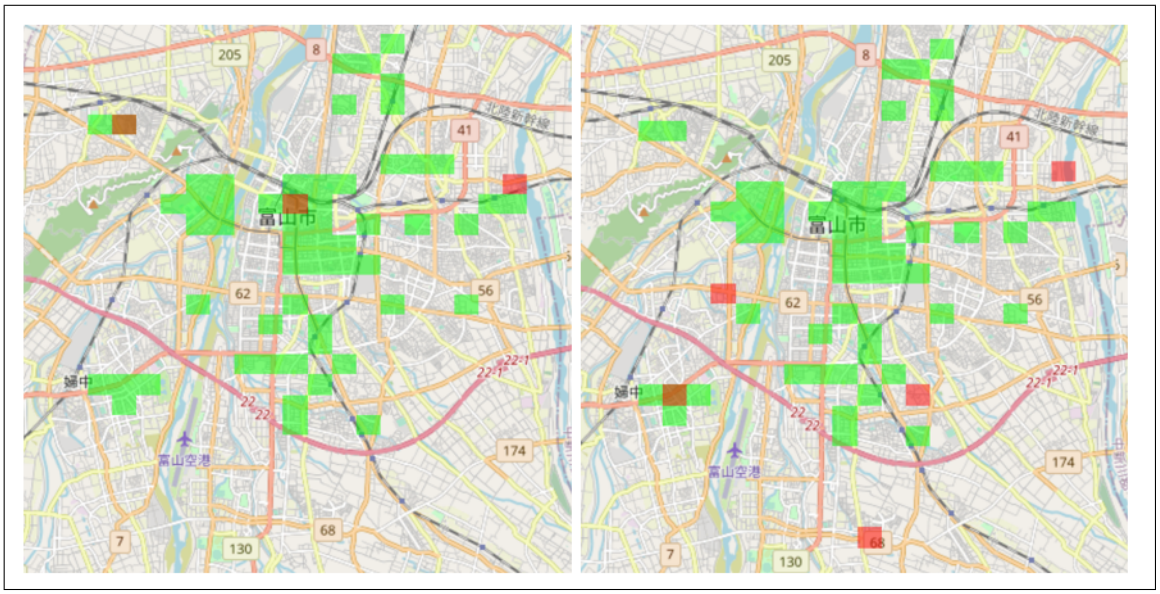


図 5.2: 2020 年 9 月 1 日（左）と 2 日（右）の予測結果

§ 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

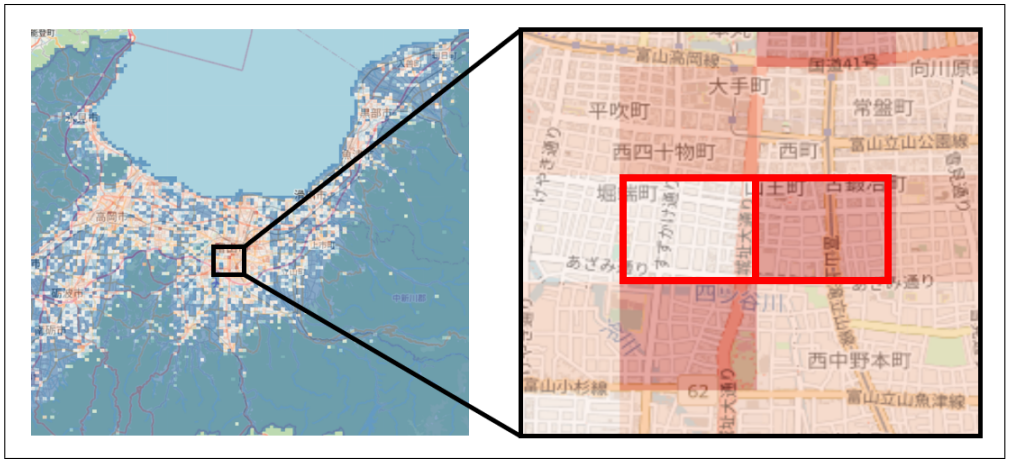


図 5.3: 予測モデルを可視化した結果

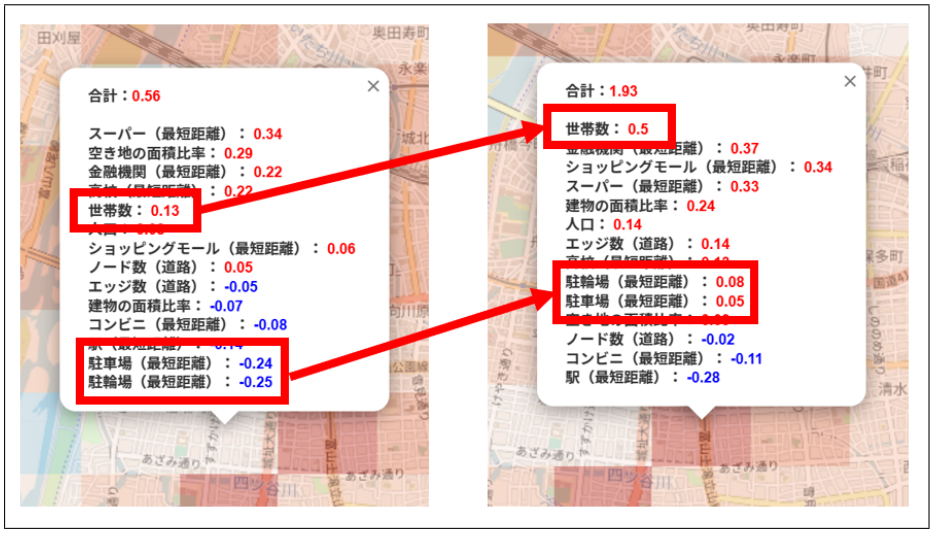


図 5.4: 特定のグリッドセルの要因を可視化した結果

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.