自治体保有データと国勢調査を用いた建物ごとの空き家確率推定モデルの構築 —群馬県前橋市における事例—

冨田　健人[1]・水谷　昂太郎[2]・秋山　祐樹[3]・馬塲　弘樹[4]・谷内田　修[5]

[1]学生会員　東京都市大学　総合理工学研究科建築都市専攻（〒158-8557 東京都世田谷区玉堤 1-28-1）
E-mail: g2281640@tcu.ac.jp

[2]学生会員　東京都市大学　工学部都市工学科（〒158-8557 東京都世田谷区玉堤 1-28-1）
E-mail: g1918082@tcu.ac.jp

[3]正会員　東京都市大学准教授　建築都市デザイン学部都市工学科（〒158-8557 東京都世田谷区玉堤 1-28-1）
E-mail: akiyamay@tcu.ac.jp

[4]非会員　京都大学特定助教　東南アジア地域研究研究所/白眉センター
（〒606-8501 京都府京都市左京区吉田下阿達町 46）
E-mail: hbaba@cseas.kyoto-u.ac.jp

[5]非会員　前橋市未来創造部（〒371-8601 群馬県前橋市大手町二丁目 12-1）
E-mail: o-yachida@city.maebashi.gunma.jp

近年，日本全国で空き家が増加し続けており，その空間分布の把握は自治体にとって重要な課題である．しかし、現在の空き家の分布調査の手法は外観目視が中心となっているため，調査に多大な時間・労力・予算を要している．そのため，迅速かつ安価に空き家の分布調査を実施する手法が求められている．そこで本研究では，群馬県前橋市を対象にデジタル地図，自治体保有データ（住民基本台帳・水道使用量データ），および国勢調査などのオープンデータを組み合わせたデータベースを作成し，実際の空き家分布情報を教師データとする機械学習（XGBoost）を行うことで，建物ごとの空き家確率を推定するモデルを構築した．なお，我々の既往研究では利用可能なデータの制約のため，中心市街地と中心市街地外モデルを分けて構築していたが，本研究ではその制約をなくし，さらにオープンデータを新たにモデルに組み込むことで，モデルの推定精度は約 90%に達した．また，SHAP という指標を用いることにより，各説明変数の空き家確率に対する寄与の程度を把握することが可能となった．

Key Words: *vacant house, municipality, national census, machine learning, shap value*

1. はじめに

近年，日本では人口減少や高齢化により全国的に空き家が増加し続けている．総務省の平成 30 年住宅・土地統計調査の集計結果によると，2018 年の全国の空き家数は約 850 万戸，空き家率は 13.6%に達している[1]．とりわけ管理が不十分な空き家が増加することにより，景観や治安の悪化，地震などの災害発生時における倒壊の危険性など，地域全体の魅力・活力の低下という影響を及ぼすことが指摘されている[2]．

このように空き家の適正な管理の必要性の高まりを受けて，2015 年 5 月より「空家等対策の推進に関する特別措置法」が施行された．これにより，自治体の調査に基づき管理が不十分な空き家を「特定空き家」に指定し，所有者への指導や改善の促進を行うことが可能となった．また，同法では「空き家の分布や状態等に関する現状把握」，「空き家に関するデータベースの整備」が全国の自治体で努力義務として定められている[3]．そのため，空き家の分布把握は現在，地方自治体にとって重要な業務の 1 つとなっている．しかし，現在の空き家の分布調査の手法は，現地調査（外観目視）が中心となっているため，調査に多大な時間，労力，費用を要してしまうことが大きな課題となっている[4]．そこで，自治体の所持しているデータ等を用いて，空き家の分布を迅速かつ安

価に把握する手法が求められている.

(1) 既存研究

益田・秋山（2020）によると，日本国内における空き家研究は，質的な現状の把握及び独自情報の獲得を目的とする「調査手法」に関する研究と，量的な情報の分析及び諸情報の関係の明示を目的とする「分析手法」に関する研究の大きく 2 つに分けられる[5].

空き家の空間分布を把握するための「調査手法」として最も数多く見られる手法が，前述した外観目視による現地調査および，関係者の空き家に関する所見を把握する聞き取り・アンケート調査となっている．これらの手法は，空き家の分布を建物 1 棟 1 棟の単位で高い信頼性を持って特定できるものの，いずれの研究においても調査対象範囲はごく限られた地区のみを対象としており，同手法を自治体全体といった広域調査に適用することは困難である[4].

一方，近年では様々な統計情報や空間情報を活用して，空き家の分布状況を把握・推定する「分析手法」に関する研究も増えつつある．広域に亘る空き家の分布状況を把握した研究としては，山下ら（2015）による水道の閉栓データを用いた空き家分布把握の例がある[6]．同研究では，栃木県宇都宮市を対象として，16 区分した各地区の 31 年分の空き家数および空き家率に関するデータベースを GIS により作成した上で，その経年変化を追っている．また，空き家率の経年変化（市域全体および 16 地区）を線形，対数，指数，ロジスティックの 4 関数による回帰分析を行い，その変化の性質を明らかにするとともに，空き家率予測の基礎を準備している．ただし，同手法は水道が「閉栓」あるいは「休止中」の物件を全て空き家と定義しており，その根拠が明らかにされていないため，同手法は空き家の空間分布を把握するために充分なであるとは言い切れない.

また，国勢調査を用いて空き家の分布状況を把握しようとした研究としては，石河ら（2017）による例がある[7]．同研究では，日本全国を対象に国勢調査の小地域集計に基づく世帯数と住宅地図の住戸数を組み合わせて町丁・字単位の空き家分布を推定する手法を提案している．ただし，2015 年の国勢調査の小地域集計が利用できなかったことから，2010 年の世帯数をベースに推計した 2015 年の世帯数を用いたため，実際の 2015 年の世帯数とは差があることや両データの作成年度が異なることから，空き家率が負の値を取る場合もあるため，地域によって精度にばらつきが生じるといった課題が残った.

こうした中で，自治体と民間企業が保有するデータを使用して，迅速かつ安価に空き家分布を推定する手法を開発する研究が取り組まれている[4),8),9),10)]．同研究では，鹿児島県鹿児島市や福岡県朝倉市を対象に，空き家の現地調査結果を教師データとする空き家データベースを作成し，機械学習を実装することにより，最終的に 500m メッシュごとの空き家数・空き家率及び，建物ごとの空き家・非空き家の推定結果を得ることが可能となった．また，一部地域の現地調査結果をもとに市全域の空き家数および空き家率の推定が可能となった.

さらに，本研究の先行的研究として，馬場ら（2021）による群馬県前橋市の中心市街地を対象とした研究がある[11)]．過年度データから将来の空き家分布推定モデルを構築し，将来の空き家予測確率地図を作成している．しかし，同研究の研究対象地域は前橋市全域ではなく同市の中心市街地のみであった.

そこで，著者らは先行研究として，群馬県前橋市全域を対象に自治体保有データ（水道使用量データ，住民基本台帳，固定資産課税台帳）を使用し，空き家の空間分布を推定する手法を開発した[12)]．しかし，前橋市から提供を受けたデータの制約として，固定資産税台帳が中心市街地のみしか利用できなかった．そのため，中心市街地（固定資産データ有り）と中心市街地外（固定資産データ無し）でモデルを分ける必要があった．また，分割後の両地域の建物数に大きな差があったため，モデル間で推定精度や予測に効く変数にばらつきが生じるといった課題が残った.

なお，空き家の空間分布と地域特性の観点から，双方の関連性を検討する研究も見られる．例えば，空き家の発生要因を明らかにした研究として水澤ら（2021）の研究がある[13)]．同研究では，広島県呉市の斜面市街地を対象として，建築属性（建築年代等）や地域特性（標高や傾斜度、前面道路幅員等）の両面から，空き家が発生しやすい地域の分析を行った．その結果，空き家の分布は標高，道路幅員，傾斜度，建築年代と関連を持つことが明らかになった．また，空き家期間と地域特性の関連を分析した研究として馬場ら（2022）による例がある[14)]．同研究では，群馬県前橋市を対象に，スマートメータによる空き家期間に基づいた空き家数と地域特性（生活利便施設への近接性）の関係を明らかにするため，回帰分析を実施した．その結果，市街化区域では，最寄り駅またはバス停までの距離などの公共交通施設への利便性や，小学校までの距離などの施設利便性は重要因子になることが明らかになった.

(2) 本研究の目的

そこで本研究は著者らの先行研究の課題を解決するために，固定資産課税台帳を使用することなく，住民基本台帳と水道使用量の 2 つの自治体保有データと，新たに国勢調査等のオープンデータを組み合わせて空き家分布推定を行うためのデータベース（以下「空き家データベース」）を構築することで，前橋市全域を 1 つのモデル

に統一し，実際の空き家分布情報を教師データとする機械学習を行うことで，前橋市全域の空き家の空間分布状況を推定する手法を開発する．また，同手法の信頼性の検証や空き家と地域特性の関連を調査することで，同手法の利点や課題を明らかにする．

2. 本研究で使用したデータと空き家データベースの整備

本研究では後述する自治体および民間企業が保有するデータと，国土数値情報や国勢調査などのオープンデータを使用した．また，これらのデータをそれぞれの位置情報に基づいて空間結合することで，空き家の分布推定を行うための分析用データである「空き家データベース」を整備した．以下，それぞれのデータの詳細について紹介するとともに，空き家データベースの構築方法について説明する．

(1) 建物データ（民間保有データ）

株式会社ゼンリンの 2016 年の住宅地図を使用した．同データは建物 ID や建物 1 棟ごとの住所，用途，面積，周長，階数などの情報を保有している．また，本研究では戸建て住宅を対象として空き家推定を行うため，用途が一般住宅（戸建て住宅）の建物のみを抽出したデータを使用した．なお，前橋市の 2016 年の住宅地図に掲載されている市全域の建物棟数は 194,269 棟であり，そのうち用途が一般住宅の建物数は 83,533 棟であった．なお，本稿でこれ以降使用する「建物」という表現は，「戸建て住宅」を意味するものとする．

(2) 住民基本台帳（自治体保有データ）

2017 年 3 月 31 日現在の前橋市の全居住者（337,595 人）の住所，年齢，性別等を収録したデータである．なお，個人情報に抵触しないように，予め居住者名や個人番号等は削除されている．同データは住所情報を持つため，アドレスマッチングを行うことで，位置情報（経度緯度座標）を与えた．また，本研究では（1）建物データと空間結合させるため，号レベルで位置情報が特定されたデータのみ使用した．さらに，本研究では世帯内の最高年齢および最低年齢，世帯人員，若年人口・生産年齢人口・老年人口別の世帯人員，世帯内の男性率を予測に使用する説明変数とした．

(3) 水道使用量データ（自治体保有データ）

2014 年から 2019 年の 5 年間の 2 か月ごとの前橋市内の全水道栓（251,562 本）の水道 ID，水道使用量と住所が収録されたデータである．住民基本台帳と同じく住所に基づいてアドレスマッチングを行うことで，位置情報を与えた．今回は空き家の現地調査が実施された 2016 年と他のデータの作成年を考慮し，2015 年，2017 年の各年の最大使用量を使用した．さらに 2015 年の水道最大使用量を基準とした 2016 年比，2016 年を基準とした 2017 年比を計算し予測に使用する説明変数とした．

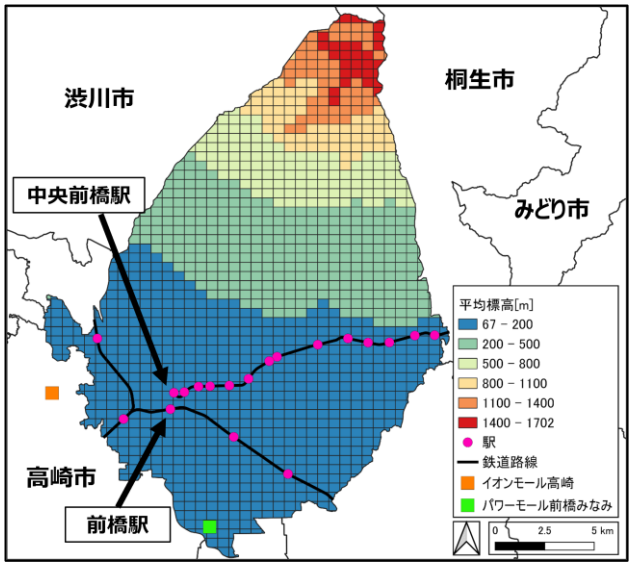

図-1 前橋市における 500m メッシュごとの平均標高と駅および大型商業施設の立地

(4) 国勢調査（オープンデータ）

e-Stat（総務省統計局）から入手した 2015 年の前橋市全域の小地域単位の国勢調査を使用した．データ内には，小地域ごとに人口や配偶関係，就業状態や住居の種類など各項目 の総数と区分別 の集計結果が記載されている．本研究では各項目の総数当たりの区分別の割合を計算することで，その地域の特徴を説明する説明変数とした．具体的には，人口・就業状態・世帯構造等基本集計に関する集計，従業地・通学地による人口・就業状態等集計に関する集計などが挙げられる．

(5) 用途地域データ（オープンデータ）

国土数値情報（国土交通省）から入手した 2019 年の用途地域ポリゴンデータを使用した．用途地域データには，行政コード，都道府県名，市区町村名，用途地域種類コード，用途地域名などが記載されているが，本研究では各建物が立地する用途地域を知ることが目的となるため，（1）建物データの各建物に用途地域種類コードを空間結合した．前橋市の場合，第二種低層住居専用地域と田園住居地域を除く 11 種類の用途地域が分布していた．なお，本研究ではそれぞれの用途地域でダミー変数化することで，分析可能な状態にした．

(6) 大型商業施設

本研究では 2 つの大型商業施設の立地にも着目した．

一つは，前橋市南部にある「パワーモール前橋みなみ」，もう一つは市外に立地するものの，前橋市民の利用の多い「イオンモール高崎」である（**図-1**）．（1）建物データから得られる両施設の建物形状ポリゴンから建物重心を求め，全ての建物から両大型商業施設までの直線距離を算出し，予測に使用する説明変数とした．

(7) その他オープンデータ

さらに本研究では**表 1** に示すオープンデータ（元データは何れも国土数値情報）を加工することにより，複数の説明変数を作成した．

「最寄り駅までの最短距離」は，建物ごとに前橋市内にある全 19 駅の中から最も直線距離が近い駅を最寄り駅として算出した．

「住宅と駅の直線平均勾配」は，先ほどの最寄り駅のデータと各建物に 4 次メッシュごとの平均標高を付与し，その差分と最寄り駅までの最短距離から平均勾配を算出した．**図-1** に示すように，前橋市は市内の北部に位置する赤城山の山頂に向けて，標高・勾配が大きくなる傾向にある．

「最寄りの医療機関までの距離」は，総合病院，一般診療所，内科の全 250 施設を対象として，建物ごとに最も近い施設までの距離を算出した．

「小学校区に該当する小学校までの距離」は，各小学校区のポリゴン内に該当する建物を抽出し，その区内の小学校との距離をそれぞれ算出した．前橋市内には 49 の小学校区（2016 年時点）が存在していた．

なお，（6）と（7）で作成した説明変数はいずれも生活利便性に関連する変数であり，馬場ら（2022）で生活利便施設への近接性や立地密度が空き家率と相関することが明らかとなっている[14]．そのため生活利便性に関する情報も説明力を持つことが期待されたため，予測に使用する変数として投入した．

(8) 空き家調査結果（自治体保有データ）

前橋市が 2016 年に実施した市全域の空き家分布調査の結果である．空き家の状態と位置情報（経度緯度）を収録するデータである．全データ件数は 7,086 件である．なお，空き家の状態とは，空き家の損壊の程度（流通中や流通可能といった良好な状態から，除却が必要なほどの損壊といった状態の違い）であるが，本研究では様々な状態の空き家を含めて，全ての空き家の空間分布を推定することを目的とするため，空き家の詳細な状態は考慮せず，空き家および非空き家の 2 種類でダミー変数化し，目的変数とした．

(9) 空き家データベースの整備

最後に GIS を使用して（1）建物データに対して，（2）から（8）を空間結合することで 1 つのデータベース（空き家データベース）を整備した．なお，後述する機械学習を行う際には，少なくとも 1 つ以上の説明変数が必要となるが，建物によっては何かしらの自治体データが欠損する場合もある．なお，住民基本台帳データは約 79.6%（66,489 件），水道使用量データは約 78.3％（65,402 件）の建物と紐づいた．そこで，これらが紐づかなかった建物については，紐づかなかった情報を「欠損値」として処理した． また，正解データとなる空き家調査結果のうち，建物用途が戸建て住宅ではない建物に紐づいたものは除外した．その結果，7,086 件のデータのうち約 32.3%（2,287 件）が建物に紐づいた．一方，オープンデータは全ての建物に各情報を与えることができた．

表-1　使用したオープンデータと作成した説明変数

説明変数	元となるオープンデータ
最寄り駅までの最短距離	鉄道時系列ポイントデータ
住宅と駅の直線平均勾配	鉄道時系列ポイントデータ
	標高・傾斜度 4 次メッシュデータ
最寄りの医療機関までの距離	医療機関ポイントデータ
小学校区に該当する小学校までの距離	小学校区ポリゴン
	学校ポイントデータ

3. 機械学習による空き家確率推定モデルの構築

(1) 本研究で使用した機械学習手法

本研究では，建物ごとの空き家確率（以下「空き家率」）を推定するために，決定木ベースの機械学習モデルである「XGBoost (eXtreme Gradient Boosting)」を採用した．XGBoost は，Chen and Guestrin（2016）[15]によって提案された手法であり，「勾配ブースティング」と呼ばれる手法の 1 つである．

まず，勾配ブースティングの仕組みについて詳説する．はじめに，目的変数と予測値から計算される目的変数を改善するように決定木を作成し，モデルを学習させ予測値を算出する．次に，その予測値と目的変数の誤差を算出し，その誤差を埋めるように新たな決定木を作成し，学習させる．これを指定した木の本数分繰り返す．木を作成するうちにモデルの予測値が目的変数に近づいていくため，作成される決定木の重みは小さくなっていく．そして予測対象のデータがそれぞれの決定木で属する葉の重みの和を予測値とする（**図-2**）．例えば，世帯内最高年齢が 75 歳，水道使用量が 300m^3，築年数が 40 年の住宅を得た場合，**図-3** の木の構造から，推定空き家率 0.31 を得る．そこから決定木を逐次作成しながら，各決

定木の結果を合計することで最終的な予測値，すなわち推定空き家率を算出する．

同手法を採用した理由は，推定精度が高く，欠損値を扱うことができることから，本研究において他の学習モデルを採用するよりも有利なためである．また，XGBoost は様々な分野の先行研究において欠損値を含むデータの分類に用いられており，高い実績をあげていることが知られている．例えば，がん細胞の遺伝子発現データの解析や[16]，オンラインコマースの行動評価[17]，中古住宅価格の予測[18]などが挙げられる．

(2) 機械学習モデルの構築

まず，元となるデータベースを訓練データとテストデータに分けた．これは全体のデータを全て訓練データに使用した場合，訓練データのみに適応した学習済みモデルが出来上がってしまい，そのモデルで未知のデータを予測させると精度が大幅に落ちてしまうためである．そこで本研究では，訓練データとテストデータを 8:2 の割合で分割した．建物数 83,533 棟中，訓練データは 66,826 棟，テストデータは 16,707 棟である．また，正解データの空き家調査結果は，空き家評価有（空き家）と評価無（非空き家）のデータの数に大きな差がある不均衡データである．そのため，学習用データの非空き家：空き家を 3：1 になるようにアンダーサンプリングを行ってから学習を行った．また，XGBoost はパラメータチューニングを必要とするため，「Optuna」を用いてチューニングを行った[19]．木の深さの最大値を示す max_depth は 3-10，子ノードにおいて観察されるデータの重み付けの合計値の最小の値 min_child_weight は 1-5，各木においてランダムに抽出される割合 subsample を 0.5-1，各木においてランダムに抽出される列の割合を表す colsample_bytree は 0.5-1，学習率である learning_rate を 0-1 の範囲でそれぞれ試行した．さらに，クロスバリデーションを実装して学習させ，建物ごとの推定空き家率を算出した．そして，推定空き家率が 0.5 以上の場合は空き家，0.5 未満の場合は非空き家と分類した．

4. 結果

(1) 構築モデルの精度評価

学習したモデルの推定精度を検証するため，テストデータ合計 16,707 件における推定結果を算出し，さらにそこから全データに対する予測を行った．正解率（正しく空き家あるいは非空き家と推定された割合）はテストデータでは 0.8944，全体では 0.8980 といずれも同程度の精度となった．また，**表-2** に全データにおける推定結果を示す．実際に非空き家のうち，正しく非空き家と予測できた割合（特異度）は 0.9058，実際に空き家のうち正しく空き家と予測できた割合（再現率）は 0.6187 となり，特に非空き家の予測精度は高い水準となった．一方，空き家と予測した結果，非空き家であった件数は 7,651 件，非空き家と予測した結果，空き家であった件数は 872 件と一定数の誤差が生じた．これらの誤差が生じた原因としては，空き家現地調査結果の戸建て住宅への結合率が約 2.9%（83,533 件中 2,287 件）と低く，教師データとしてのデータ数が十分に確保できなかったことが原因であると考えられる．また，空き家現地調査結果は調査員による外観目視によって調査が実施されているため，非空き家・空き家の判定結果に調査員の間でばらつきが発生する．このばらつきが誤差となった可能性が考えられる．例えば，本当は空き家であったにも関わらず，空き家の状態（外観）が良いため，調査員が非空き家と判定してしまうケースなどが挙げられる．

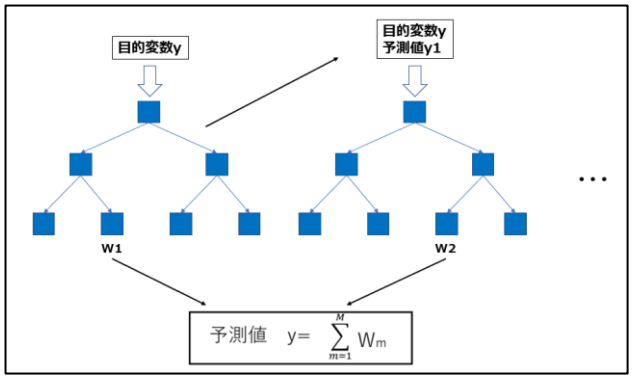

図-2 XGBoost の予測値の算出のイメージ

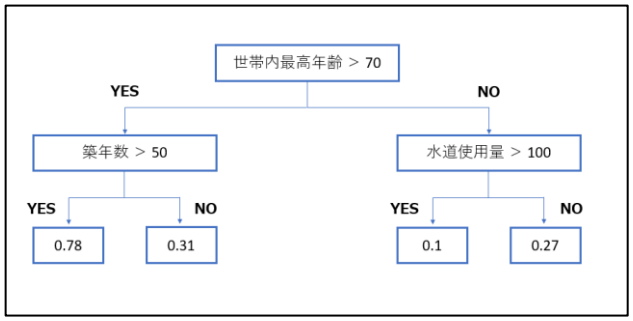

図-3 XGBoost における決定木のイメージ

表-2 全データ 83,533 件の検証結果

		推定値［棟］	
		非空き家	空き家
真値［棟］	非空き家	73,595	7,651
	空き家	872	1,415

(2) 空き家分布推定結果

図-3，**図-4** に前橋市全域における 500m メッシュごとの推定空き家数と推定空き家率を示す．中心市街地（JR

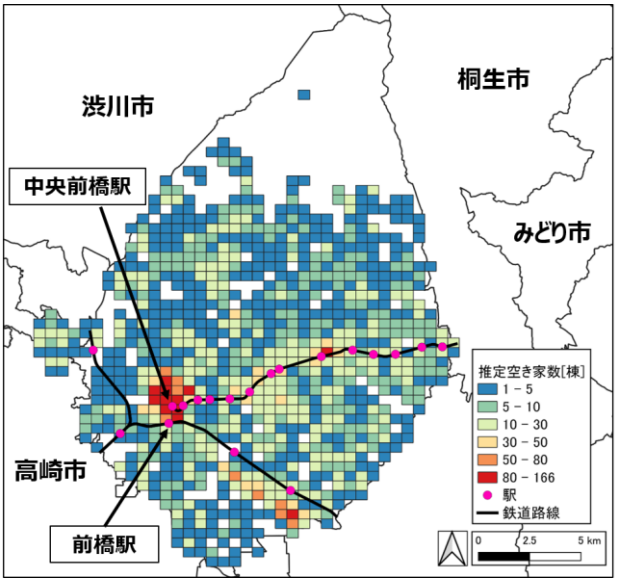

図-4　前橋市全域の推定空き家数（500m メッシュ単位）

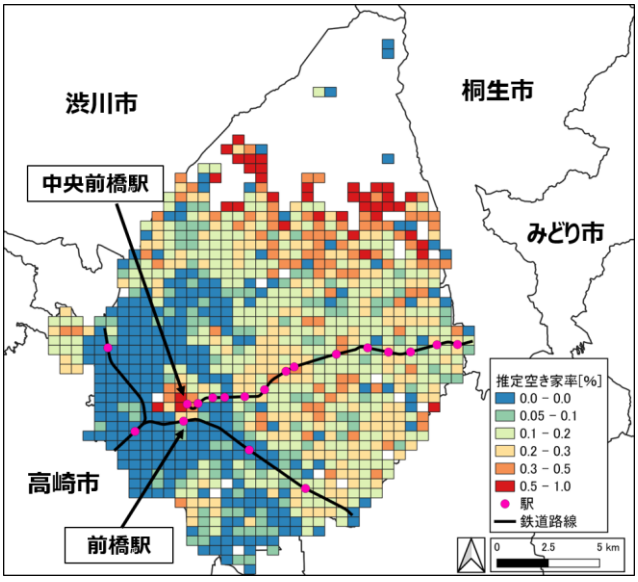

図-5　前橋市全域の推定空き家率（500m メッシュ単位）

前橋駅，上毛電気鉄道中央前橋駅周辺の地域）では建物数が多いため空き家数が多く，中心市街地から離れた農村部および中山間地域では空き家数が少なくなった．一方，空き家率に注目すると中心市街地および，中心市街地から離れた農村部・中山間地域，さらに前橋市東部で高くなっている．

このように市域全体の空き家数，空き家率が把握できることにより，自治体が実際に現地調査を行う際に，重点的かつ早期に空き家調査を実施するべき地域を検討する際に有益な情報となるものと期待される．ただし，本研究の値はいずれも推定値であることから，必ずしも実際の空き家数・空き家率と正確に一致するものではない点に注意が必要である．とはいえ，一度推定モデルを構築してしまえば，市全域の空き家データの更新を迅速かつ安価に実施することが可能なため，前述の空き家調査における課題の解決に貢献できるものと期待できる．

(3)　特徴量の重要度と構築モデルの解釈

本研究では構築したモデルに対して，どの説明変数がどの程度の影響を与えているか，すなわち特徴量の重要度の把握や，各説明変数が予測値をどのように変化させているのか，ということについての検証を行った．以上の検証には「SHAP（SHapley Additive exPlanations）」と呼ばれる手法を用いた．これは，協力ゲーム理論のシャープレイ値（Shapley Value）を機械学習に応用したオープンソースのライブラリであり，機械学習モデルの解釈手法の 1 つである [20]．SHAP を使うことで，各変数の重要度だけではなく，モデルが導き出した予測値に対して，各説明変数の影響の度合いを詳細に把握可能となる．すなわち，ミクロ的な解釈からマクロ的な解釈までを一貫して行える点で優れた解釈手法である．

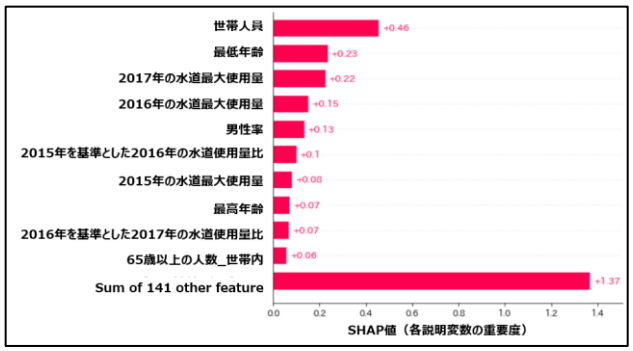

図-6　説明変数の重要度評価（上位 10 項目）

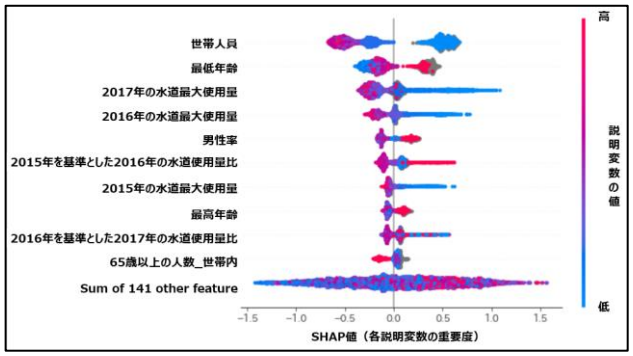

図-7　説明変数のごとのビースウォーム図
（重要度上位 10 項目）

図-5 および図-6 に SHAP を用いて得られた結果を示す．**図-5** は説明変数の重要度を大きい順に 10 項目並べたヒストグラムであり，**図-6** は説明変数の重要度上位 10 項目ごとのビースウォーム図を示した．なお，ビースウォーム図とは，説明変数の値の大小によって，SHAP 値（各説明変数の寄与度）との相関を把握するために使用する．本研究では SHAP 値の正の値が大きいほど，空き家である確率をより大きくし，負の値が大きいほど空き

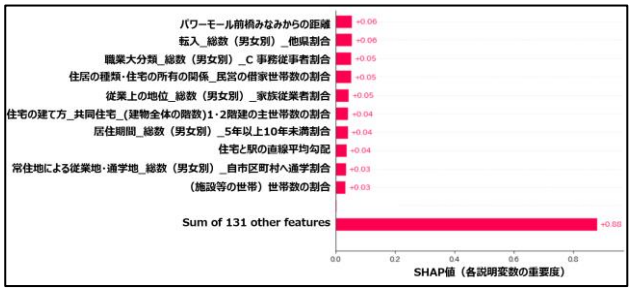

図-8　本研究により新たに付与した説明変数の重要度評価（上位 10 項目）

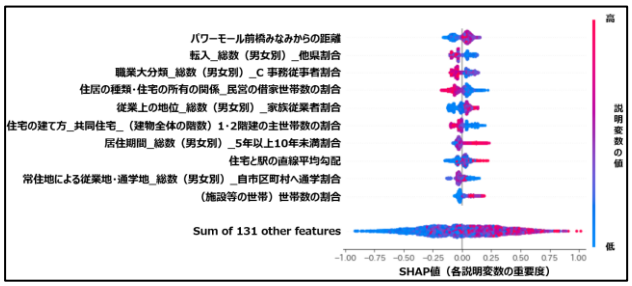

図-9　本研究により新たに付与した説明変数ごとのビースウォーム図（重要度上位 10 項目）

家である確率をより小さくしたことを示している．また各点の色は説明変数の値の大小関係を示しており，赤くなるほど値が大きく，青くなるほど値が小さいことを示している．灰色は値が欠損している建物を表す．例えば，「2017 年の水道最大使用量」を見ると，水道使用量が少ない（青い分布）ほど，SHAP 値が大きくなる，すなわち予測値を正に押し上げている．つまり，水道使用量が少ないほど空き家である確率を高めていることが分かる．

まず**図-6** より，上位 10 項目には，水道使用量や年齢，世帯人員，65 歳以上の世帯人数など自治体保有データの重要度が大きくなり，特に世帯人員は顕著に表れている．よって，地域単位で集計した国勢調査よりも，各建物に対応するピンポイントなデータの方が，空き家予測に大きく寄与することが確認できた．

また**図-7** より，水道使用量の他にも住民基本台帳から，世帯人員が少なく男性率が高い建物ほど，また，世帯の最低年齢が高いほど空き家になる傾向にあった．さらに，最低年齢と世帯内の 65 歳以上の人数の灰色の散布図は欠損を表している．つまり，住民基本台帳が欠損している建物は空き家になりやすい特徴にあることが分かった．

続いて，**図-8**，**図-9** は上位 10 項目には含まれていないものの，本研究により新たに付与した説明変数の重要度上位 10 個のヒストグラムとビースウォーム図を示したものである．**図-8** から，変数の寄与度にあまり変化はないものの，「パワーモール前橋みなみからの距離」や，国勢調査の「居住期間」および「住居の種類」などが，他の変数に比べて僅かながらに寄与していることが分かった．また**図-9** より，「パワーモール前橋みなみからの距離」は，遠くも近くもない中間くらいの距離であるほど，空き家になりやすい傾向にあった．「住宅と駅の直線平均勾配」は大きいほど，つまり赤城山麓付近の地域ほど，空き家になりやすい傾向にあった．さらに，国勢調査から「居住期間が 5 年以上 10 年未満の割合」が多い地域や，「共同住宅_（建物全体の階数）1・2 階建の主世帯数の割合」が少ない地域ほど空き家になりやすい傾向にあることが分かった．加えて，前橋市は日本の多くの地方都市と同様に車社会であるため，本研究で導入した表-1 の変数 の寄与度は**図-6** から**図-9** に該当するものはなく，は空き家予測にあまり寄与しないことが明らかとなった．

6.　おわりに

本研究では，主に自治体が保有しているデータと，著者らの先行研究では使用していなかった国勢調査等のオープンデータを用いて，空き家データベースを作成し，実際の空き家分布情報を教師データとする機械学習を実装することで，同市全域において空き家分布推定を行うモデルを開発した．その結果，同モデルを用いることで前橋市全域の空き家および非空き家の空間分布を，約 90%という高い精度で推定することが可能となった．また，構築したモデルにおいてどの変数がどの程度影響を及ぼすのか，ということについて SHAP を用いて可視化することにより，地域特性と空き家の関連についても把握可能となった．これらの指標は自治体における立地適正化計画の策定や空き家に関連した都市・地域計画などの立案とその支援に貢献できるものと考えられる．

今後の課題や方針は以下の通りである．まず，建物ごとあるいは地域単位で変数の SHAP 値が推定できたため，異なる変数ごとに SHAP 値をメッシュ集計し，マッピングすることにより，地域ごとに空き家確率を高めている要因を明らかにすることが出来る．そこで，こうした結果が地域ごとの空き家の発生を抑制する政策に活かすことができないか，自治体の担当者等へのヒアリングを通して検討を行う．

また，本研究では機械学習の手法として XGBoost を採用したが，他の機械学習手法（例えば，同じ決定木ベースの機械学習モデルである LightGBM など）を採用することで，さらなる精度の向上を図ることが出来る可能性について検討を行う．さらに，自治体によっては空き家調査結果を保有していない場合が考えられるため，ある自治体で学習したモデルを，同じような地理的条件を有する他の自治体に対してどれほど外挿することが可能か，検証を進めたいと考えている．加えて，本研究の結果か

ら国勢調査の一部の変数は空き家予測に寄与するものの，データの集計単位が小地域であるため，建物単位ではあまり大きく寄与しないことが分かった．そこで，国勢調査など政府統計のみを使用し，小地域単位で空き家数や空き家率を推定するモデルの開発にも取り組みたい．

謝辞：本研究は東大CSIS共同研究（No.880）の一環として実施した．また，本研究は前橋市における超スマート自治体研究協議会および前橋市未来政策課より，前橋市の自治体保有データ（住民基本台帳，水道使用量データ，空き家現地調査結果）の提供を受けることで実現した．さらに，東京都市大学総合研究所デジタル都市空間情報研究開発ユニットの成果の一部でもある．ここに記して謝意を表したい．

参考文献

1) 国土交通省：平成 30 年住宅・土地統計調査の集計結果（住宅及び世帯に関する基本集計）の概要，<https://www.mlit.go.jp/common/001314574.pdf>，（最終閲覧日 2022 年 1 月 25 日）
2) 国土交通省：空き家等の現状について，<https://www.mlit.go.jp/common/001172930.pdf>，（最終閲覧日 2022 年 1 月 25 日）
3) 国土交通省：空家等対策特別措置法について，<https://www.mlit.go.jp/policy/shingikai/content/001385948.pdf>，（最終閲覧日 2022 年 1 月 25 日）
4) 秋山祐樹，上田章紘，大野佳哉，髙岡英生，木野裕一郎，久冨宏大：鹿児島県鹿児島市における公共データを活用した空き家の分布把握．「自治体の公共データを活用した空き家の分布把握手法に関する研究（その 1）」，日本建築学会計画系論文集，Vol.83，No.744，pp.275-283，2018.
5) 益田理広・秋山祐樹：日本国内における近年の空き家研究の動向．地理空間，13-1，pp.1-26，2020.
6) 山下伸・森本章倫：地方中核都市における空き家の発生パターンに関する研究．都市計画論文集，Vol.50，No.3，pp.932-937，2015.
7) 石河正寛，松橋啓介，金森有子，有賀敏典：住戸数と世帯数に基づく空き家の詳細地域分布の把握手法．都市計画論文集，Vol.52，No.3，pp.689-695，2017.
8) 秋山祐樹，上田章紘，大内健太，伊藤夏樹，大野佳哉，髙岡英生，久冨宏大：公共データを活用した空き家の分布把握手法の高度化．「自治体の公共データを活用した空き家の分布把握手法に関する研究（その 2）」，日本建築学会計画系論文集，Vol.84，No.764，pp.2165-2174，2019.
9) 秋山祐樹，馬場弘樹，大野佳哉，高岡英生：機械学習による空き家分布把握手法の更なる高度化．「自治体の公共データを活用した空き家の分布把握手法に関する研究（その 3）」，日本建築学会計画系論文集，Vol.86，No.786，pp.2136-2146，2021.
10) 秋山祐樹：ビックデータは何を語るか？ 地理空間，12-3，pp159-178，2019
11) 馬場弘樹，秋山祐樹，谷内田修：自治体保有データを活用した空き家の空間分布の将来予測モデル構築 ―群馬県前橋市を対象として―．土木学会論文集 D3（土木計画学），vol.77，No2，pp.62-71，2021.
12) 冨田健人，秋山祐樹，馬場弘樹，谷内田修：自治体保有データ用いた機械学習による空き家の分布推定手法の開発，第 65 回土木計画学研究発表会・講演集，p211，2022.
13) 水澤克哉，田村将太，田中貴宏：斜面市街地における空き家の発生要因に関する研究．都市計画論文集，Vol.56，No.3，pp.897-904，2021.
14) 馬場弘樹，秋山祐樹，清水千弘：スマートメータを利用した空き家機関と地域特性との関係分析−群馬県前橋市を対象として−．「GIS −理論と応用」，30（1），pp.39-50，2022.
15) Chen, T. and Guestrin, C.: Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785-794, 2016.
16) M. A. Latief, A. Bustamam, and T. Siswantining, "Performance Evaluation XGBoost in Handling Missing Value on Classification of Hepatocellular Carcinoma Gene Expression Data," *2020 4th International Conference on Informatics and Computational Sciences (ICICoS)*, pp.1-6, 2020, doi: 10.1109/ICICoS51170.2020.9299012.
17) Y. Yang, "Market Forecast using XGboost and Hyperparameters Optimized by TPE," *2021 IEEE International Conference on Artificial Intelligence and Industrial Design (AIID)*, pp. 7-10, 2021, doi: 10.1109/AIID51893.2021.9456538.
18) Z. Peng, Q. Huang and Y. Han, "Model Research on Forecast of Second-Hand House Price in Chengdu Based on XGboost Algorithm," *2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT)*, pp. 168-172, 2019, doi: 10.1109/ICAIT.2019.8935894.
19) Takuya, A., Shotaro, S., Toshihiko, Y., Takeru, O.and Masanori, K. "Optuna: A Next-generation Hyperparameter Optimization Framework," *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623-2631,2019
20) Scott, M, L. and Su-In, L. "A Unified Approach to Interpreting Model Predictions," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768-4777,2017

（2022. 9. 30 受付）

DEVELOPMENT OF A MODEL FOR ESTIMATING THE PROBABILITY OF A VACANT HOUSE FOR EACH BUILDING USING MUNICIPALITY OWNED DATA AND CENSUS DATA

Kento TOMITA, Kotaro MIZUTANI, Yuki AKIYAMA, Hiroki BABA and Osamu YACHIDA

In recent years, the number of vacant houses in Japan has continued to increase throughout the country, and understanding their distribution is an important issue for local governments. However, the method of surveying the distribution of vacant houses is mainly based on visual inspection from the outside, which requires a lot of time, labor, and budget for the survey. In this study, we created a database that combines building point data, municipally owned data (basic resident registers and water consumption data), and open data such as the national census for Maebashi City, Gunma Prefecture, and used machine learning (XGBoost) to estimate the probability of vacant houses for each building using actual vacant house distribution information as training data. In our previous study, the model was constructed separately for the CBD area and the outside the CBD area due to the limitation of available data, but by eliminating this limitation and incorporating new open data into the model, the accuracy of the model was approximately 90%, By using the indicator SHAP, it was possible to determine the degree of contribution of each explanatory variable to the probability of vacant houses.