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

冨田　健人[1]・秋山　祐樹[2]・馬場　弘樹[3]・谷内田　修[4]

[1]学生非会員　東京都市大学　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: g1818064@tcu.ac.jp

[2]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

[3]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

[4]非会員　前橋市未来政策課（〒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，日本では全国的に空き家が増加し続けており，その分布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からの目視が中心となっているため，調査に多大な時間・労力・予算を要することが全国の自治体において大きな課題となっている．そこで本研究は，群馬県前橋市を対象にデジタル住宅地図と自治体保有の個票データ（住民基本台帳・水道使用量・固定資産課税台帳）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，自治体の空き家の分布状況を迅速に推定する手法を開発した．その結果，空き家の空間分布を，建物単位の正解率で約 90%という高精度な水準で推定することが可能となった．

Key Words: *vacant house, estimation, municipal data, XGBoost, feature importance*

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]．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に大きな悪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」が施工された．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が自治体の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課題でなっている．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]．

(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」に関する研究と，大きく 2 つに分けられる[5]．そして，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．広域に渡って空き家の分布状況を把握しようとした研究としては，山下ら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，市域を 16 区分した上で，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，その根拠が明らかではない上に，水道データのみで空き家を特定することは困難となっている．

そこで自治体と民間が保有するデータを使用して，迅速・安価な空き家分布推定手法を開発する研究が取り組まれている．例えば，一部地域の空き家現地調査の結果と自治体保有のデータを組み合わせて空き家の分布推定を行った事例がある [4),7),8),9)]．これらの研究では，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・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，馬場ほか（2021）による群馬県前橋市の中心市街地における研究がある [10]．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た．しかし，同研究を含む数多くの既存研究では，調査対象地域が中心市街地のみなど限定的な場合が多く，また空き家の状態を説明する変数が全ての地域や建物で揃っているという理想的な状態において実施されている．しかし，実務上では自治体全域を欠損値を含んでいたり，使用可能なデータが限定された状態で処理できる方法を実現することが理想的である．すなわち，今後全国の自治体で空き家分布推定を可能にしていくためには，以上に挙げる既存手法の更なる改良を重ねながら手法の高度化を図ることが必要である．

(2) 本研究の目的

そこで本研究は，群馬県前橋市全域を対象に，自治体保有の個票データ（住民基本台帳・水道使用量・固定資産課税台帳）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，自治体全域という広域の空き家の空間分布状況を推定する手法の開発する．また，同手法の信頼性の検証を行うことで，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民間企業が保有するデータと，自治体が保有するデータを使用する．以下，それぞれのデータの詳細について説明する．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 194,229 棟であり，そのうち用途が一般住宅の建物数は 110,013 棟であった．

(2) 住民基本台帳（自治体保有データ）

2017 年の前橋市の全居住者（337,596 人）の住所，年齢，性別等を収録したデータである．また，同データは住所を持つため，アドレスマッチングを行うことで，位置情報（経度緯度）を与えた．なお，個人情報に抵触しないように，本研究で使用する同データは予め居住者名や個人番号等は削除されている．

(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）の水道使用量と住所が収録されたデータである．今回は他のデータの作成年を考慮し，2015 年と 2016 年の 2 年間の水道使用量の合計を使用した．また，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．

(4) 固定資産課税台帳（自治体保有データ）

2018 年の前橋市全域の建物ごとの構造や築年数，住所などが収録されているデータである．ただし，本研究で提供を受けたデータは中心市街地のみであるため，中心市街地外の空き家推定では使用しないものとする．なお，ここで言う「中心市街地」とは，前橋市中心市街地活性化基本計画 [11]に記載されている町丁名に該当する地域であり，図-1 に示す範囲となっている． また，固定資産

課税台帳に収録されている住所は住民基本台帳や水道で使用されている住居表示ではなく，地番となっているため，地番図（固定資産地籍図）から得られる位置情報（地番ポリゴンの重心座標）を位置情報として与えた．

(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）から入手した用途地域の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，用途地域種類コードを使用した．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 種類の用途地域が存在していた．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) 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．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）も収録しているが，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．

(7) 空き家データベースの整備

GIS を使用して， 建物データに対して，住民基本台帳，水道使用量データ，固定資産課税台帳，用途地域データ，空き家調査結果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．ただし，前述の通り固定資産課税台帳は中心市街地のみに付与される．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，必ずしも 1 つの建物に全ての自治体データが紐づけられるわけではなく，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．そこで，いずれの変数も付与されない建物は空き家データベースから排除した．その結果，説明変数の全部または一部を含む 60,301 件の戸建て住宅を取得した．

4. 機械学習を用いた空き家の分布推定

(1) 空き家分布推定に使用する説明変数の決定

まず，表-1 に示すように空き家分布推定に使用する説明変数を決定した．なお，住民基本台帳や固定資産課税台帳からは様々な説明変数を作成することが可能であるが，表-1 に示した変数を採用した理由は，それぞれの変数が既存研究[4),7),8)]において空き家判定を行う上で説明力が高いことが示されているためである．また本研究では「女性率」という変数を導入した．この変数を導入した理由としては，前橋市では単身の女性世帯の割合が高く，一人暮らしをする当人が死亡すれば空き家になる可能性が高いものと予想されるためである．

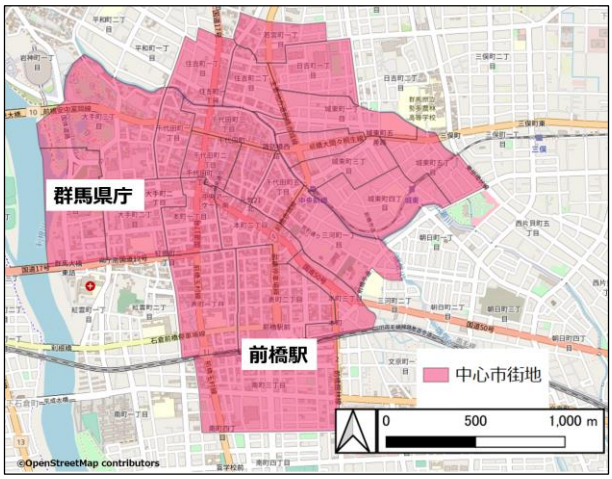

図-1 前橋市の中心市街地の範囲

表-1 空き家分布推定に使用する説明変数

データ	変数	中心市街地	
		内	外
建物データ	建物面積（㎡）	有	有
水道データ	水道使用量（t）	有	有
	閉栓期間（月）	有	有
住民基本台帳	居住者の最高年齢（歳）	有	有
	居住者の最低年齢（歳）	有	有
	居住者の平均年齢（歳）	有	有
	世帯人員（人）	有	有
	女性率（%）	有	有
固定資産課税台帳	築年（年）	有	無
	構造判定	有	無
用途地域データ	11 種類（中心市街地は 5 種類）	有	有

(2) 機械学習の実装

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）[12)]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．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，例えば，がん細胞の遺伝子発現データの解析や[13]，オンラインコマースの行動評価[14]，中古住宅価格の予測[15]などで応用された実績がある．

本研究では固定資産課税台帳が中心市街地のみで利用可能なことを考慮し，中心市街地と中心市街地外で別々のモデル構築を行った．まず，教師データ 60,301 件を中心市街地（3,347 件）と中心市街地外（56,954 件）に分け，訓練データとテストデータに 1:1 の割合で分割して学習した．続いて，学習したモデルにより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．

図-2 に中心市街地における建物ごとの推定空き家率の空間分布を示す．中心市街地では中央前橋駅の北部や西部，また JR 前橋駅の北部の地区で推定空き家率が高い，すなわち空き家と判定された建物が多く分布することが分かった．また，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．中心市街地では建物数が多いため，一般的に空き家数が多く，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い．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は注意が必要である．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出来ることから，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言える．

5. モデルの推定精度の検証

学習したモデルの推定精度を検証するため，テストデータに対して，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）および，F 値（実際に空き家のうち正しく空き家と推定された割合および，空き家と推定されたもののうち実際に空き家であるものの割合の調和平均）を算出した．また，どの説明変数が推定に寄与したか把握するため，各説明変数の重要度を分析した．

表-3 に，テストデータ 30,151 件における推定結果を示す．中心市街地では正解率が 0.921，F 値が 0.544 となり，中心市街地外では正解率が 0.941，F 値が 0.325 となった．正解率はいずれも 90%を超えた．一方，F 値は中心市街地の方が大きくなったものの一概に高いとは言えない結

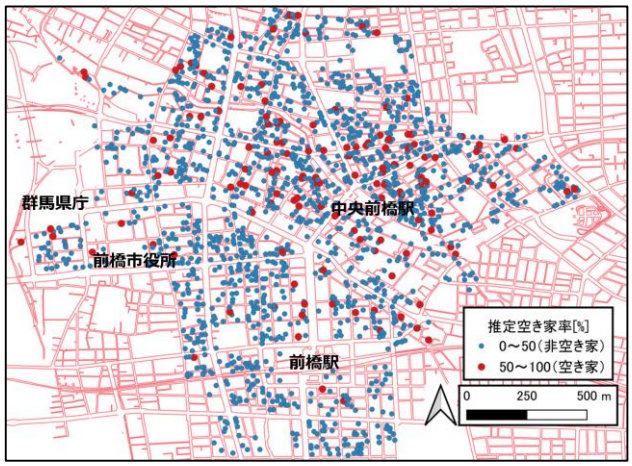

図-2　中心市街地における建物ごとの推定空き家率

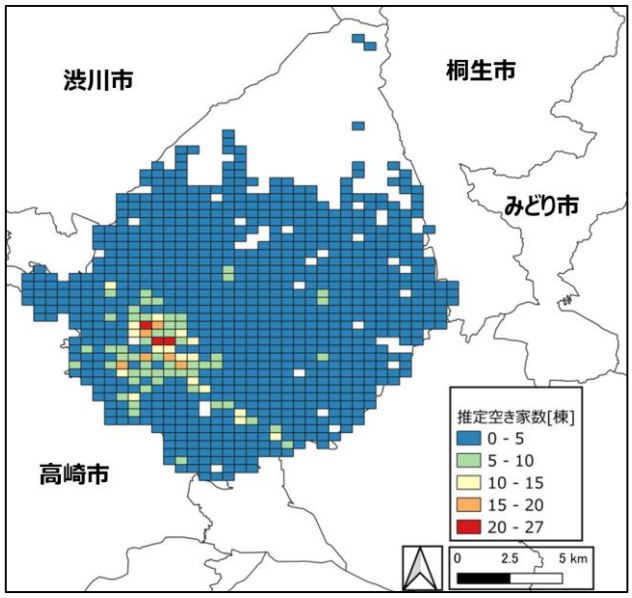

図-3　前橋市全域の推定空き家数（500m メッシュ）

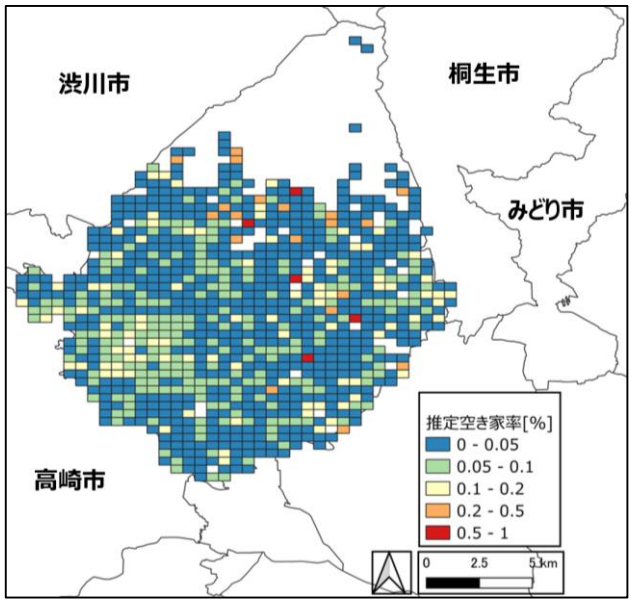

図-4　前橋市全域の推定空き家率（500m メッシュ）

果となった．また，これはあくまでもテストデータに対する結果であるが，中心市街地外では空き家の見落としも一定の数（556 件）発生していることから，今後は説明変数やモデルの改良を行い，正しく空き家と予測できなかった建物数を減らしていくことが課題である．

続いて，構築したモデルにおいてどの説明変数が目的変数に対してどの程度影響を与えているかについて，その重要度を算出した．図-5 に中心市街地における説明変数の重要度評価を，図-6 に中心市街地外における説明変数の重要度評価を，それぞれ重要度が大きい順に 10 位までの説明変数を示す．中心市街地，中心市街地外ともに水道の使用量が空き家推定に大きく寄与しており，中心市街地外では特に顕著に表れていることが分かった．また，中心市街地では水道の他に近隣商業地域に指定されていることや，築年数が大きく影響を与えていることが分かった．一方，中心市街地外では水道の閉栓期間や建物面積も大きく影響を与えていることが分かった．

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を用いて，空き家データベースを作成し，前橋市の中心市街地と中心市街地外で分けて，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，同モデルは市域全体で正解率 90%以上という高い精度で空き家の分布を推定することが可能となった．

今後の課題としては，まず中心市街地外では実際には空き家である建物を非空き家と判定する見落としが一定の数発生していることから，説明変数の追加やモデルの改良を進めることでさらなる精度向上を図っていく．また，本研究では機械学習手法として 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）を利用することで，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，自治体における空き家関連業務でにおける活用を考慮すると，空き家の状態別での推定（例えば特定空き家やその候補の分布を推定するなど）が出来るのが望ましいため，空き家の状態ごとの分布推定手法の開発も検討していきたい．

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，東京都市大学総合研究所未来都市研究機構都市マネジメント研究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．

表-3　テストデータ 30,151 件の検証結果

		中心市街地 推定値［棟］		中心市街地外 推定値［棟］		合計［棟］
		非空き家	空き家	非空き家	空き家	
真値［棟］	非空き家	1,465	56	26,423	1,090	29,056
	空き家	75	78	566	398	1,095

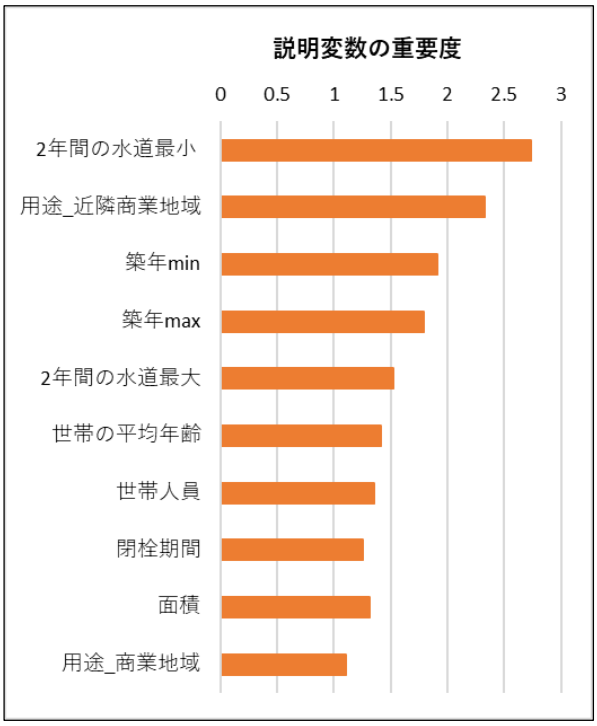

図-5　中心市街地における説明変数の重要度評価

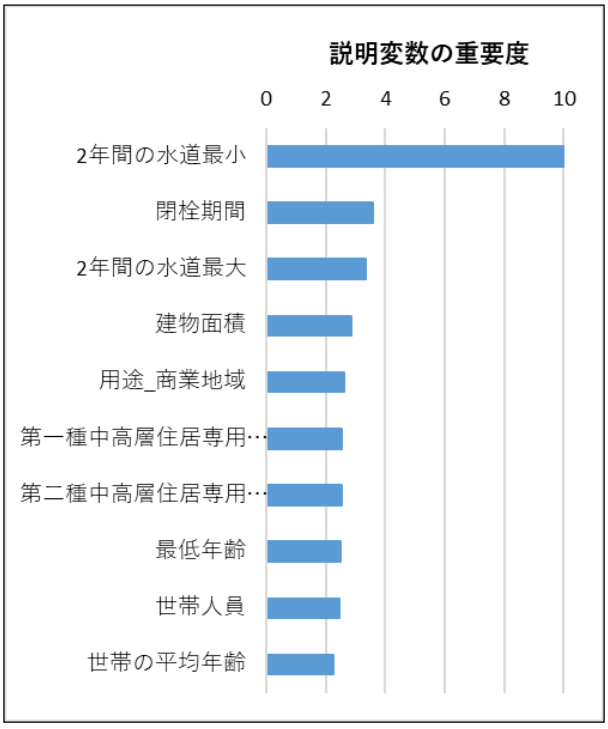

図-6　中心市街地外における説明変数の重要度評価

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) 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.
8) 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.
9) 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
10) 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.
11) 前橋市：前橋市中心市街地活性化基本計画（平成 29 年 3 月），<https://www.city.maebashi.gunma.jp/material/files/group/54/master_plan_honbun.pdf>，（最終閲覧日 2022 年 1 月 26 日）
12) 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.
13) 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.
14) 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.
15) 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.

(Received ？？, 2022)
(Accepted ？？, 2022)

DEVELOPMENT OF A METHOD FOR ESTIMATING THE SPATIAL DISTRIBUTION OF VACANT HOUSES BY MACHINE LEARNING USING MUNICIPAL DATA

Kento TOMITA, 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 developed a method to quickly estimate the distribution of vacant houses in a municipality by developing a database of vacant houses in Maebashi City, Gunma Prefecture: a typical local city of Japan by integrating a digital housing map and pinpoint data (basic resident register, water usage, and fixed asset taxation register) owned by the municipality, and performing machine learning (XGBoost) using actual information on the distribution of vacant houses as ground truth data. As a result, it was possible to estimate the spatial distribution of vacant houses with a high accuracy of approximately 90% in terms of the percentage of correct answers per building.