オープンデータを用いた戸建て住宅の分布推定手法の開発

齋藤　開*・秋山　祐樹・水谷昂太郎*****

Developing a Method for Estimating the Distribution of Detached House Using Open Data

Kai Saito*, Yuki Akiyama**, Kotaro Mizutani***

Abstract: In recent years, objective and data-driven policy-making has been required. Against this backdrop, detailed map databases containing attribute information for individual buildings are gaining attention as an indispensable form of social infrastructure. However, the difficulty in acquiring such data due to factors like personal information protection remains a challenge. Therefore, as a first step towards constructing a detailed and openly accessible map database, this study developed a method to estimate the distribution of detached houses using only open data. First, based on data from the entire Hiroshima Prefecture, we created a machine learning model that classifies each building as either a detached house or other building. Furthermore, we applied the constructed model to Saitama City, Saitama Prefecture, and validated its estimation results by comparing them with the statistical value at the small-area level. As a result, the coefficient of determination reached 0.99, making it possible to reproduce the distribution of detached houses.

Keywords: 建物データ（building database），オープンデータ（open data），機械学習（machine learning），
戸建て住宅（detached house），建物属性（building attribute）

１．はじめに

近年，都市空間の構造とその動態を精緻に把握した上で政策を立案する必要性が一層高まっている．この要請は，持続可能で，災害に強く，資源効率性が高い都市を構築しようとする国際的潮流によって強く後押しされている．その中で，国土交通省によると個々の建物の属性情報を含む詳細な地図データベース（以下「建物属性地図」）は，社会基盤として不可欠な役割を担うものとして注目されている．建物属性地図は，都市計画の高度化，エネルギー需給の最適化，災害リスク評価，将来人口推計など，多岐にわたる分野における活用可能性を有しているといえる．

建物属性に関するデータは，地方自治体や国の機関，さらには民間企業によって収集・整備されている．しかし，これらのデータセットがオープンデータとして広く公開される機会は限られている．その背景には，データの収集主体毎に異なる課題が存在する．行政機関が保有するデータは，個人情報保護上の制約や自治体ごとのフォーマットの不統一，さらには整備に要する膨大なリソースといった問題を抱えている．一方，民間企業が提供するデータは高価であることが多く，利用に際して経済的な障壁となり得る．このように，建物属性地図の入手と活用は依然として困難であり，この制約が多様な分野における実用・応用を阻んでいる．本研究ではこの問題に対応するため，詳細かつオープンデータとして利用可能な建物属性地図の開発を最終的な目標とする．本稿ではその第一歩として，建物の用途分類に焦点を当てる．

１.１　既存研究

オープンデータを活用した建物用途の推定に関する試みは，既にいくつかの研究が報告されている．例えば，Daniel et al. (2023) は OpenStreetMap（OSM）の建物形状のみを用いて，建物を居住用と非居住用に分類する手法を提案した．また，Fill et al. (2024) は，建物形状に加え，空間的関係性をグラフ構造として組み込むことで，高精度な建物用途分類手法を提案した．しかし，これらの既存研究には共通する課題がある．すなわち，建物用途分類の正解データとして OSM のラベルを利用している点である．OSM の属性情報は信頼性や品質が十分に検証されておらず，教師データとしての妥当性に懸念が残る．

以上のように，建物用途推定に関する研究には一定の蓄積があるものの，日本を対象とした事例はほとんど見られない．また，日本においては基盤地図情報など信頼性とアクセス性に優れたデータセットが整備されている．そのため，これらのデータセットを先行研究で提案され

*　学生会員　東京都市大学総合理工学研究科建築・都市専攻（Tokyo City University）
〒158-0087　東京都世田谷区玉堤１丁目２８－１　E-mail：g2481619@tcu.ac.jp
**　正会員　東京都市大学建築都市デザイン学部都市工学科（Tokyo City University）
〒158-0087 東京都世田谷区玉堤１丁目２８－１ E-mail：akiyamay@tcu.ac.jp
*** 正会員　東京都市大学建築都市デザイン学部都市工学科 准研究員（Tokyo City University）
〒158-0087 東京都世田谷区玉堤１丁目２８－１ E-mail：kmizu@tcu.ac.jp

た方法論と組み合わせることで，より精度の高い建物用途分類の実現が期待される．

１.２　本研究の目的

以上の課題を踏まえ，本研究は，日本における信頼性の高いオープンデータを活用し，建物用途分類の精度向上と応用可能性の拡大を図ることを目的とする．特に，都市の建物ストックの大部分を占める戸建て住宅に着目し，その空間分布を明らかにすることで，インフラ整備の効率化や防災上の強靭性評価など，多角的な都市分析に資する基盤情報の整備を目指す．具体的には，公開されている建物ポリゴンデータと統計情報を活用し，戸建て住宅とそれ以外の建物（以下「非戸建て住宅」）を分類する手法を開発することによって，戸建て住宅の空間分布を把握することを目指す．

2. 研究手法

2.1　研究対象地域と使用したデータ

本研究の対象地域は，広島県全域および埼玉県さいたま市（以下「さいたま市」）全域の 2 地域とした．まず，モデルの学習には，都市的地域と非都市的地域が併存する広島県全域のデータを教師データとして用いた．次に，構築したモデルの汎化性能を検証するための適用先として，広島県から地理的には離れていながら，都市的地域と非都市的地域を併せ持つさいたま市を選定した．

また，本研究で使用したデータの一覧を表 1 に示す．これらは全て無償で入手可能なデータであり，中でも本研究の基盤となる建物データには，都市計画基礎調査の建物利用現況図を用いた．

２.２　教師データの作成

本研究で用いた特徴量を表 2 に示す．これらの特徴量は建物を戸建て住宅と非戸建て住宅に分類するために必要な 24 の変数で構成され，大きく「幾何学的特徴」と「周辺環境特徴」の 2 つのグループに分類できる．幾何学的特徴には，建物ポリゴンの面積，周長，頂点数などが含まれる．周辺環境特徴には，各建物が属する町丁字内の年齢階級別人口割合，居住期間，所得水準などの人口統計情報が含まれる．これらの周辺環境特徴は，市町村単位で集計された 2018 年住宅・土地統計調査と，町丁字単位で集計された 2020 年国勢調査のデータを用いて作成した．

また，本研究の機械学習モデルの構築には，網羅的かつ信頼性の高い教師データ，すなわち個々の建物の真の用途情報が不可欠である．本研究の対象地域である広島県は県全域の都市計画基礎調査の結果を公開している日本国内において数少ない地域である．そこで，同調査で収集・集計された建物用途情報を用い，各建物ポリゴンに戸建て住宅または非戸建て住宅のいずれかの正解ラベルを付与することで，教師データセットを構築した．

表 1. 使用したデータ

利用用途	データ	出典	作成年
建物ポリゴンデータ	都市計画基礎調査	各自治体	2021
年齢階級世帯数	国勢調査	総務省	2020
居住期間別世帯数	国勢調査	総務省	2020
所有形態別世帯数	国勢調査	総務省	2020
収入階級別人口	住宅・土地統計調査	総務省	2018
建物用途別	国勢調査	総務省	2020
POI データ（Point of Interest）	国土数値情報	国土交通省	2020

表 2. 使用した特徴量の概要

区分	特徴量	概要
幾何学的特徴	面積	建物ポリゴンの面積
	矩形度	建物ポリゴンの矩形度
	凸度	建物ポリゴンの凸度
	慣性モーメント	建物ポリゴンの慣性モーメント
	頂点数	建物ポリゴンの頂点数
	異方性指数	建物ポリゴンの異方性指数
	最長軸長	建物ポリゴンの長軸の長さ
	伸長度	建物ポリゴンの伸長度
	向き	建物ポリゴンの向き
周辺環境特徴	年齢階級別世帯割合（3 区分）	年少人口（<18），生産年齢人口（18–65），老年人口（ >65 ）
	居住期間別世帯割合（6 区分）	居住期間別の世帯割合
	世帯収入別世帯割合（9 区分）	世帯収入階級別の世帯割合
	建物用途別建物数（3 区分）	建物用途別の建物数（戸建て住宅，長屋，共同住宅）

２.３　戸建て住宅モデルの構築

本研究で構築する戸建て住宅または非戸建て住宅の分類を行うモデルには，eXtreme Gradient Boosting（以下「XGBoost」）を採用した．XGBoost (Chen and Guestrin., 2016) を選択した理由は以下の通りである．第一に，本研究で使用する教師データには欠損値が一部含まれている．XGBoost は欠損値を含む特徴量から直接学習することができ，我々のデータセットに対して頑健なモデルを

構築することが可能である．第二に，学習速度が非常に速く，モデル構築に必要な計算コストを大幅に削減できる．第三に，XGBoost はツリーベースの手法であるため，データの標準化を必要としない．これは，特徴量の相対的な順序に基づいて分割を行うため，値のスケールや分散の違いに影響されないためである．そのため，データの前処理を大幅に効率化できる．これらの理由から，本研究では XGBoost を分類モデルとして採用した．

また，モデルの性能評価指標には，以下の式(1)で定義される対数損失（Logarithmic Loss: logloss）を採用した．logloss は予測確率と真のラベル（$y_i \in \{0,1\}$）との間の距離を定量化するものであり，N はデータ数を表す．logloss の値が小さいほど，モデルの性能が優れていることを示す．

$$logloss = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i) + (1-y_i)\log(1-p_i) \quad (1)$$

ここで，N はデータ数，y_iは真値，p_iは戸建て住宅である確率をそれぞれ示す．

モデル構築にあたり，まず全データセットを訓練・検証用（70%）とテスト用（30%）のサブセットに分割した．また，ハイパーパラメータを最適化し，汎化性能を向上させるため，訓練/検証用サブセットに層化 K 分割交差検証（Stratified K-Fold Cross-Validation）を適用した．モデルの最適化は，F 値が最大化されることを目的とし，決定木の最大深度，学習率，データインスタンスのサブサンプリング率，特徴量の列サブサンプリング率を最適化した．

２．４　モデルの外挿

本研究で構築した分類モデルを，さいたま市の全建物に適用し，戸建て住宅と非戸建て住宅の分類を行った．ただし，本モデルは町丁字レベルで集計された統計データを特徴量として利用しているため，外挿に際して地域的特性に依存した空間的偏りが生じる可能性がある．その結果，標準的に用いられる予測確率に閾値を適用する分類手法では，個々の建物の形状的特徴よりも地域全体の平均的傾向に予測が引きずられ，推定結果に歪みが生じる恐れがある．この課題に対応するため，本研究では推定された戸建て住宅数を 2020 年の国勢調査における町丁字レベルの統計値に整合させる修正を加えた．具体的には，各町丁字においてモデルが算出した戸建て住宅確率の高い建物から順に選択し，国勢調査に基づく戸建て世帯数に達するまでを戸建て住宅と判定した．この方法により，地域的偏りを補正し，より現実に即した建物用途分布の推定を可能にした．

３．結果

３．１　戸建て住宅モデルの評価

テストデータに対する推定結果を表 3 に示す．全体の正解率は約 86%，F 値は 0.913 であり，本研究で開発した分類モデルが高い精度で戸建て住宅を分類可能であることが示された．

一方で，一部の建物は誤分類されることがわかった．具体的には，2 種類の誤分類が確認された．一つは，面積が小さく形状が単純な非戸建て住宅を戸建て住宅と誤分類するケース（偽陽性）であり，もう一つは，逆に面積が広く形状が複雑な戸建て住宅を非戸建て住宅と誤分類するケース（偽陰性）である．特に前者の偽陽性が本モデルにおける主要な誤差要因であることが示された．この結果は，図 1 の特徴量重要度が示すように，モデルが建物の大きさや形状を，分類を行う上で手がかりとしている一方で，その単純なルールだけでは捉えきれない例外的なケースが存在することを示唆している．

表 3. 戸建て住宅モデルにおける混合行列と評価指標

テストデータ		予測	
		非戸建て住宅	戸建て住宅
真値	非戸建て住宅	32,898	34,416
	戸建て住宅	7,427	218,531
精度			0.857
適合率			0.864
再現率			0.967
F 値			0.913

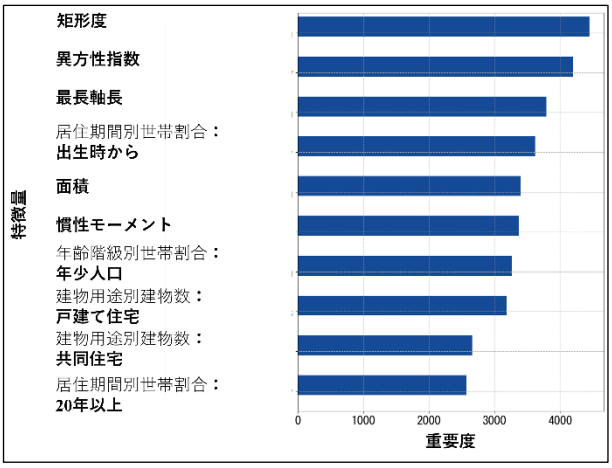


図 1. 戸建て住宅モデルの特徴量重要度

３．２　外挿結果の検証

３．２．１　外挿結果の検証方法

外挿結果の精度を定量的に評価するため，正解データとしてさいたま市の都市計画基礎調査に含まれる建物用途情報を利用した．まず，町丁字単位で外挿結果による戸建て住宅数と都市計画基礎調査による実際の戸建て住宅数をそれぞれ集計した．次に，両者の数値を比較し，平均絶対誤差（MAE），決定係数（R^2），二乗平均平方根誤差（RMSE）を算出した．ただし，検証に用いる都市計画基礎調査の建物用途ラベルには，一部不確実性がある．

例えば，戸建て住宅の敷地内に存在する倉庫などが主たる建物と同じ用途として記録されている場合がある．このようなラベルの不確実性が検証に与える影響を低減するため，戸建て住宅については住生活基本計画における最低居住面積水準を参考に，面積が 25 ㎡以下の建物を検証対象から除外した．

３.２.２　外挿結果の検証

上記の検証方法に基づき，戸建て住宅分類モデルの外挿結果を評価した．その結果，MAE は 24.7，R^2 は 0.99，RMSE は 38.7 という良好な精度が得られた（図 2）．この結果は，本モデルが実際の居住パターンをよく反映しており，他地域においても戸建て住宅の分布を高い精度で再現できる一定の汎化性能を持つことを示している．また，図 3 に外挿結果の具体的な例として，東武野田線七里駅周辺の建物ごとの外挿結果を示す．

ただし，学校や工場の敷地内にある比較的面積が大きい建物を戸建て住宅と誤分類した例も存在する．この問題に対処するためには，各建物の周辺に存在する建物群の平均面積や密度などを特徴量として追加し，建物が属する地域全体の特性をモデルに学習させることで，このような誤差を抑制できるものと期待される．

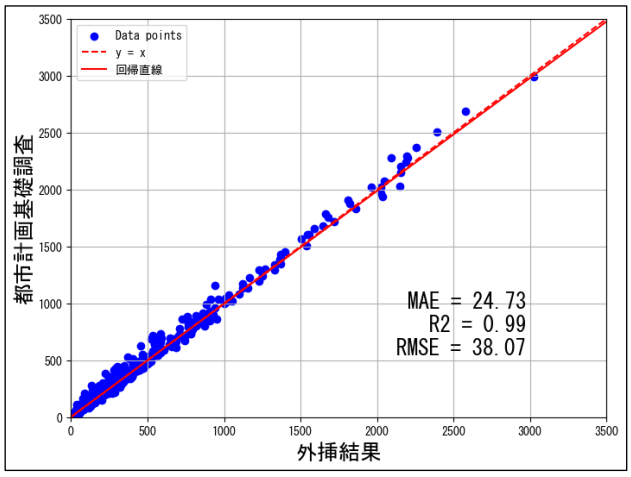


図 2. 外挿結果と都市計画基礎調査の比較

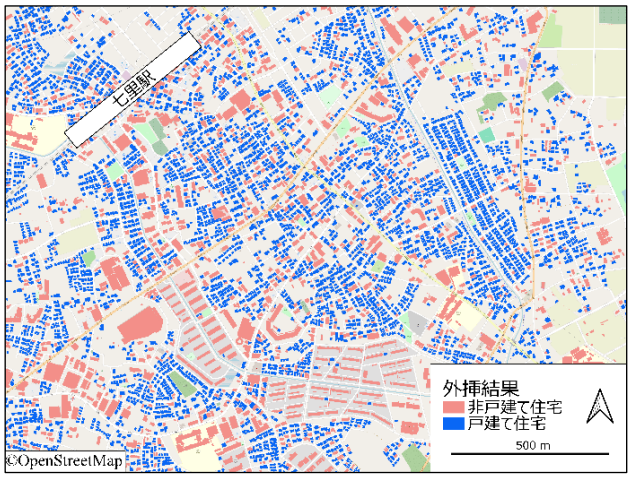


図 3. 七里駅周辺の外挿結果

４．まとめと今後の展望

本研究では，都市計画基礎調査の建物利用現況図に収録された建物ポリゴンデータと統計情報を活用して，建物を戸建て住宅，または非戸建て住宅に分類する手法を開発した．その結果，戸建て住宅の空間分布を高精度に特定することが可能になった．

今後は共同住宅の分布推定技術の開発も行うことで，戸建て住宅と共同住宅を併せた居住用建物の分布を把握する技術の確立を検討している．そこで Fill et al. (2024) によって提案された，建物間の空間的関係を考慮するグラフニューラルネットワーク（GNN）分類器の手法を応用し，日本の地域特性に合わせた用途分類モデルを開発することも検討している．

本研究の最終目標は，詳細な全国規模の建物データベースを構築することである．この目標を達成する上で，我々の方法論の主要な利点は，全国で利用可能なオープンな統計情報と建物ポリゴンデータを活用することで，このデータベースを効率的かつ低コストで構築できることにある．このデータベースの構築には，主要な建物属性（建築年，建物構造，高さ，階数など）を推定し，それらを組み合わせて個々の建物の分布とその用途を推測する作業が含まれる．以上の推定を通じて，個々の建物の詳細な情報を全国規模で統合したオープンな地図データベースの構築を目指す．

謝辞

本研究は、JSPS 科研費（24K00243）の助成を受けて実施した．ここに記して謝意を表す．

参考文献

国土交通省（2023）都市計画情報のデジタル化・オープン化ガイダンス ＜ https://www.mlit.go.jp/toshi/city_plan/content/001618095.pdf ＞ (2025,8,30 閲覧)

Adams, D. S., Moehl, J., Stipek, C., Hauser, T., Lit, P., 2024: SIGHT: Stacked Integration of Geospatial Hierarchical Typologies for Inferring Building Characteristics. 2024 IEEE International Conference on Big Data (Big Data).

Fill, J., Eichelbeck, M., Ebner, M., 2024: Predicting building types and functions at transnational scale.

Chen, T., Gue strin, C., 2016: XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

総務省統計局（2022）令和 2 年国勢調査 ＜ https://www.stat.go.jp/data/kokusei/2020/index.html ＞ (2025,8,26 閲覧)

総務省統計局（2020）平成 30 年住宅土地統計調査 ＜ https://www.stat.go.jp/data/jyutaku/2018/tyousake.html ＞ (2025,8,26 閲覧)

国土交通省 (2023) 住生活基本計画（全国計画）<https://www.mlit.go.jp/common/001392030.pdf> (2025,8,26 閲覧)