令和２年度
和歌山県における
空き家分布推定に関する研究成果報告書

令和 3 年 3 月

東京大学空間情報科学研究センター　特任准教授　秋山 祐樹
東京大学空間情報科学研究センター　特任助教　馬場 弘樹
東京大学空間情報科学研究センター　研究支援員　左右田敢太
東京大学空間情報科学研究センター　研究支援員　洪 義定
和歌山県データ利活用推進センター　主事　徳冨 智哉

目　次

1. 空き家分布推定研究の背景と目的

1.1. 背景と先行研究

近年、日本では人口減少や高齢化、都市部への人口移動などを背景に、全国的に空き家が増加している。総務省統計局「住宅・土地統計調査」によると、平成 30 年の日本全国の空き家数は約 846 万戸、空き家率は 13.6% に達しており、空き家数・空き家率ともに過去の調査から比較しても、一貫して増加が続いている状況にある。なかでも、「その他の住宅」（別荘などの一時的に利用実態がある住宅や、賃貸用・売却用の住宅以外の住宅）の増加は著しく、平成 20 年調査から平成30 年調査までの10 年間に約268 万戸から約347 万戸へと約1.3 倍に増加している。一部の管理不十分なその他の住宅は、腐朽・破損による倒壊危険性を有するだけでなく、地域の防犯性の低下や景観の悪化にもつながる。このような空き家は「特定空き家」と呼ばれ、近隣住民や地域全体に深刻な影響を及ぼす可能性が高いことから、特定空き家を含む空き家の実態把握はわが国にとって急務となりつつある[1, 2] 。

こうした背景を受け、平成 27 年 5 月から「空家等対策の推進に関する特別措置法（空家等対策特措法）」が全面施行された。同法の施行により、自治体は空き家の所有者への適切な管理の指導や、空き家跡地の活用促進、特定空き家に対する助言・指導・勧告・命令、さらには罰金・行政代執行も可能となり、空き家の活用・除却といった行動を法的根拠に基づいて実施することが可能になった[3] 。しかしこれらの行動を起こすためには、まずは自治体のどこにどの程度空き家が分布しているのか、という情報を把握する必要がある。しかし空き家の空間的分布を把握する手法は、現状では一棟一棟を個別に訪問し外観を見て判断する戸別目視が中心となっている。また、現地調査を実施する前に、空き家が数多く分布すると考えられる地域を予め把握する手法も確立されていないため、広域の空き家分布を継続的に把握し続けるには多大な労力と時間、そして費用を要する。これが自治体において空き家対策の取り組みを進めていく上で大きな障壁となっている。すなわち、これらの調査を「迅速」、「安価」かつ「継続的」に実施可能な手法の確立が期待されている。

広域を対象とした空き家の分布状況の把握を試みた研究としては、西山（2015）や Yamashita et al.(2015）による水道閉栓情報を用いた市域全域を対象とした例がある。しかし、これらの手法では水道が閉栓か休止中の建物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点に課題がある。また、秋山ほか（2018）により、水道閉栓の有無のみを使って空き家を特定することは困難であることが指摘されている。この問題を解決する方法としては、自治体が所有するデータ（以下「公共データ」）である住民基本台帳や、建物登記情報、水道使用量の情報などを活用して空き家を特定する方法もある（秋山ほか 2018, Akiyama et al., 2020）。特に、Akiyama et al.(2020) は、鹿児島県鹿児島市と福岡県朝倉市を対象に、上記 3 つのデータや地図情報から得られる建物属性等を説明変数とするクロス集計表を使用した分析モデルを提案した。このような公共データを活用した空き家分布推定の手法は、個別目視に頼る従来の手法に比べて、調査費用と調査時間を削減することができるだけでなく、広域を対象に迅速かつ安価な調査を、定期的かつ継続的に実施することができるという長所を持っている。

一方、和歌山市においても今後空き家がますます増加していくことが懸念されている。実際に、平成 25 年の住宅・土地統計調査によると、和歌山市の空き家率は 15.8%（全国平均は 13.5%）であったが、平成 30 年の調査によると和歌山市の空き家率は 18.9%（全国平均は 13.6%）となっており、全国平均と比べてもその値、また値の増加のペースも高い水準にある。こうした背景の中、

和歌山市は空き家対策の取り組みを進めており、平成 29 年 3 月には「和歌山市空家等対策計画」を策定し、空き家の空間分布の把握を目的に市内全域を対象とした「和歌山市空家実態調査」を平成 29 年度に完了させた。和歌山市空家実態調査は現地調査による目視判読で空き家か否かを確認しており、比較的信頼性の高い調査であることが期待される。したがって、Akiyama et al.(2020)に見られるような自治体が保有する各種公共データを活用して、和歌山市全域の空き家分布推定モデルを構築するとともに、その推定精度を和歌山市空家実態調査の空き家データを使って検証することで、和歌山市において有用な分析モデルの選定や、より推定精度の高い分析モデルの開発が実現できるものと期待される。その結果、和歌山市において今後は本研究で開発した手法を採用することで、迅速かつ安価な空き家分布調査を定期的かつ継続的に実施し、空き家対策の取り組みの効果的な推進とその支援を行うことが可能となるものと期待される。さらに、総務省統計局統計データ利活用センターや和歌山県データ利活用推進センターと協働することで、様々な公的統計のミクロデータ（国勢調査の個票データ等）が利用可能となるため、統計ミクロデータと和歌山市が持つ公共データを融合させることで、より推定精度の高い分析モデルの開発の実現が期待できる。

1.2. 前年度の研究と、今年度の研究の位置づけ

前年度の研究では、和歌山市の保有する公共データ（住民基本台帳、建物登記情報、水道使用量情報）を利用し、データの前処理など一連の過程を経て基本モデルを作成した。その結果、一定程度の精度を保証できるモデルの構築が可能であることが分かった。一方で、前年度の研究では自治体の保有データのみを用いて分析を行ったため、前述の自治体保有データを使用することが困難な自治体においては適用が困難であるという課題があげられた。

そこで、今年度の研究ではさらに公的統計（国勢調査）を追加してモデルを構築することで、公的統計から抽出した特徴量が空き家分布の推定精度向上にどれほど寄与するのかについて分析を行う。また自治体保有データを用いることなく国勢調査のみを用いた推定手法により、どれほどの推定精度を確保することができるのかを明らかにすることで、他の自治体への適用可能性についても検討する。

1.3. 本研究の目的

以上より、本研究の目的は前年度のモデルを引き継いだ精度の高い空き家分布予測モデルを構築するとともに、前年度のモデルが持つ社会実装上の課題を解決することとする。具体的には、以下の 3 点に整理される。

1. 前年度の研究成果を引き継ぎ、和歌山市全域の空き家分布状況をより迅速・安価にかつ高い精
   度で推定するモデルを構築する
2. 和歌山市の保有する公共データだけでなく、政府の公的統計である「国勢調査」を活用することで、モデルの推定精度にどのような影響が及ぼされるのかについて分析する。
3. 国勢調査のみを用いたモデルの推定精度を分析することで、他の自治体へのモデルの
   拡張可能性を検討する。

2. 本研究で使用したデータの概要

本研究を実施するためには、まず和歌山市から提供された各種公共データの変数の概要やデータのレイアウト構造を把握する必要がある。そこで、本章では空き家分布推定の際に利用したデータの変数と構造について概説する。和歌山市から提供された公共データは、住民基本台帳（以下、「住基データ」）、建物登記情報（以下、建物登記データ）、水道利用量情報（以下、水道情報データ）、和歌山市空き家実態調査である。住基、建物登記、水道情報の各データのファイルは CSV 形式（コンマ区切りのテキスト形式）であり、住基データは、表側に住民、表頭に各変数が並ぶ形式、水道データは表側に住民、表頭に各変数が並ぶ形式である。しかし、建物登記データはそのような形式になっていないため、別途レイアウトの構成について解説する。なお、本章は前年度報告書の一部を再校正したものである。

2.1. 公共データの概要

まず、公共データの概要を説明する。

１）住基データ（2019 年 4 月現在のデータ）

住基データは住所がキー変数となる。すなわち、当該データは住所によって他のデータと結び付けられる。一つの住所には、同じ住所を持つ住民が複数存在する場合や、住所に複数の世帯番号が対応している場合がある。そのため、実際の分析の際には世帯ごとに集計し直し、各データに対してユニークな住所を与えている。その際、世帯人数等は合計に直すなどのデータ整形を行う必要がある。住基データの一覧は次の通りである。

変数一覧

- 住所（C）…○○番地△△号
- 年齢（V）
- 世帯識別の番号（V）…7 桁の数字
- 住定日（V）…1 桁目は元号(１：明治、２：大正、３：昭和、４：平成)、2～3 桁目は年、4～5 桁目は月、6～7 桁目は日を表す。

２）建物登記データ（2018 年 10 月現在のデータ）

建物登記データは特殊な形になっており、取り扱いに注意が必要である（**表 1**）。1 列目の番号は建物ごとに割り振られた建物番号（項番）である。表 1 の 1 行目を見ると、項番、物件情報、物件種別、物件状態…と項目が並んでいる。2 行目以降も同じ規則が成立するが、4 行目は主である建物情報の項目が並んでおり、最も重要な部分である。なお各キー変数の 5 行目において、付属建物情報がついている箇所が見られるが、物置等の居住者が存在しない建物と考えられるため無視する。各建物番号の i 行 j 列目の項目を(i, j)とおく。建物登記データは分析する際に建物構造等の文字列をカテゴリー化する必要があるため、データの変数情報についての説明は、後述するデータ変数作成に譲る。

表 1　建物登記データのレイアウト

	行	1 列	2 列	3 列	4 列	5 列	6 列	7 列	8 列	9 列
物件情報	1 行	項番	物件情報	物件種別	物件状態	地番区域	地番家屋番号	不動産番号		
一般建物表題部登記事項	2 行	項番	所在	所在（実際の所在地）				原因及びその日付	登記の日付	その他
	3 行	項番	家屋番号						登記の日付	その他
	4 行	項番	主である建物の表示		種類	構造	床面積	原因及びその日付	登記の日付	その他
	5 行	項番	付属の建物の表示	符号	種類	構造	床面積	原因及びその日付	登記の日付	その他

3）水道情報データ（2019 年 5 月現在のデータ）

水道情報データは住所をキー変数としたリレーショナルな形式になっているため、取り扱いは容易である。以下、離散値であるカテゴリー変数は（C）、連続変数は(V)の記号で表す。水道情報データの変数は次の通りである。

変数一覧

- 住所（C）…○○番地△△号
- 開栓区分（C）…開栓、閉栓の 2 値データ
- 開栓日(V)…1 桁目は元号(１：明治、２：大正、３：昭和、４：平成)、2～3 桁目は年、4～5 桁目は月、6～7 桁目は日を表す。4040706 は、平成 04 年 07 月 06 日を表す。
- 閉栓日(V)…閉栓している場合は、開栓年月日を 7 桁の数字で表す。開栓中の場合は 0 で表す。
- 月ごとの水道使用量（V）…単位は、立法メートル。変数名は、429-1 のような形となっている。429-1 は平成 29 年 1 月を表す。

4）空き家実態調査データ（2016～2017 年に現地調査実施）

空き家実態調査データは、水道閉栓情報、平成 24 年和歌山県廃墟建築物調査、和歌山市危険家屋台帳（苦情をピックアップした台帳）を基に空き家候補を割り出し、その候補に対して現地調査（外観目視による判定）を行うことにより、空き家を特定したものである。そのため、上記のデータで空き家候補とならなかったものは調査対象となっていないため、過小推計になっている可能性があることに注意が必要である。空き家実態調査データはシェープファイル形式であるため、建物の位置を表す座標情報が予め付与されている。空き家実態調査データの中で必要な変数は次の 2 つである：

- 位置座標（建物重心の経度緯度）
- 空き家判定…和歌山市空き家実態調査で空き家と判定されたものは 1、空き家でないものは 0 とする 2 値変数

2.2. 国勢調査の概要

前年度の研究においては自治体が保有する公共データのみを用いて空き家を推定したが、今年度は前年度に利用したデータに加えて、国勢調査を追加的に用いた。なお、本研究で使用した国勢調査は基本単位区単位で集計が行われているものを使用した。

2.3. データの処理

2.3.1 データの空間結合

住基・建物登記・水道情報及び空き家実態調査データは住所をキー変数としており、住所情報を使って 4 つのデータを 1 つのデータに統合できる。しかし、住所には表記ゆれ（漢数字とローマ数字など）が存在するため、同じ住所に対して複数の住所表記が存在する場合がある。例えば、住所には「ヶ」「ケ」「が」、「ノ」「の」、「1 丁目 2-3」「1-2-3」のようにいくつかの表記方法がある。従って、住所文字列の完全一致によるデータの統合は一般に困難であり、他の方法によって文字列の表記ゆれに対応する必要がある。

住所の表記ゆれに対応する一手法として、「ジオコーディング」が挙げられる。同手法を用いることで、「住所」を「緯度・経度」という定量的な位置座標に変換することができる。座標は住所のような表記ゆれを含まないユニークな値であるため、前述のような問題は発生しない。その結果、座標というユニークな情報をキー変数とすることが可能になる。

そこで本研究では、ジオコーディングの中でも研究目的として利用可能な東京大学空間情報科学研究センターが提供している「CSV アドレスマッチングサービス」を利用して緯度経度座標を付与した。なお、住所が適切に記述・設定されていない場合や、CSV アドレスマッチングサービスの住所の参照元となる住宅地図が、新規開発などに伴う住所の変更・新設を捕捉しきれていない場合があるため、一定程度の誤差が生じてしまう点には注意が必要である。

住所データを座標に変換する前に注意すべき点としては､CSV アドレスマッチングサービスは○○県○○市○○町○○番地という形式の住所でなければ正確には読み込めないため、水道情報・住基・建物登記データの住所を整形する必要がある。住所の整形作業の例として、省略されている都道府県名の付加や、入力ミス等によって住所が文字化けするなどの個別の問題に対処する等の作業がある。これらの作業の後、各データの住所情報を CSV アドレスマッチングサービスを用いて座標に変換した。

次にデータ統合作業について説明する。本研究では、住基・建物登記・水道の緯度経度座標を用いて地理空間上のポイントデータとして表現し、対応する建物ポリゴンデータ（Zmap TOWN Ⅱ（株式会社ゼンリン））にそれぞれのデータを対応させることで、データ同士の統合を行った。ただし、建物ポリゴンの座標と水道・住基・建物登記データに付与した座標が完全に一致する保証はなく、むしろ完全一致しない座標が一定数存在すると考えるべきである。そのため、本研究では地理情報システム（GIS）を利用し、水道情報・住基・建物登記及び空き家実態調査データの各座標から距離が最も近い建物ポリゴンにデータの値を付加する方法で、建物ポリゴンに各データの値を与えた。一方、空き家実態調査データは既に建物を特定できるレベルの緯度経度座標を保有しているため、その座標と重複する建物ポリゴンに空き家判定の結果を与えた。以上の作業の結果、建物ポリゴンの固有 ID をキー変数とした表形式のデータを作成することで、データ統合作業を完了させた。これらの作業を実施して得られたデータから、分析モデルの作成、空き家推定、推定精度の検証等が可能となる。以上の作業フローをまとめたものを**図 1** に示す。

なお、本分析で問題となるのが、建物登記データや住基、水道情報の一部が地番表記となっており、CSV アドレスマッチングを利用しても緯度経度座標を特定できない場合がある点である。そのため、住居表示の地区では CSV アドレスマッチングを利用する一方で、地番表示の地区では地番図の文字列を対応させて住所を特定した。地番図は和歌山市から提供いただいたもので、既に地番の割り振られた土地がポリゴンとして存在する。各ポリゴンの重心点を対応する地番住所の緯度経度とすることで対応づけを行った。

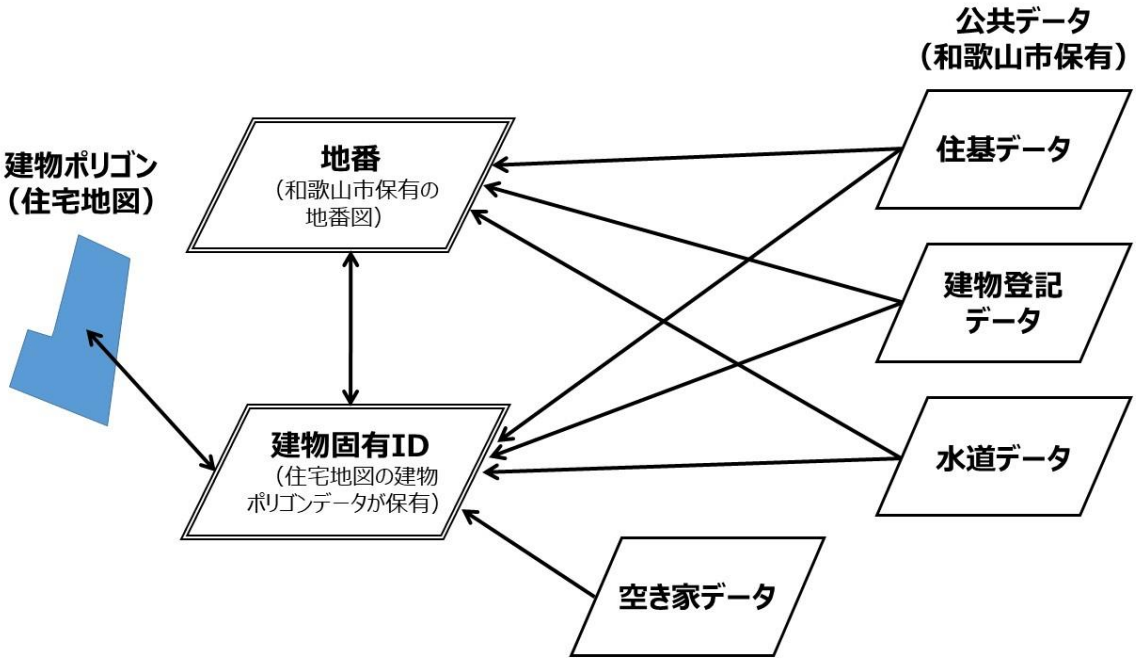

図 1 データ統合のフロー

2.3.2 変数の抽出

続いて、各自治体保有データからどのような変数を抽出するかについて述べる。なお、以下の変名末尾の（C）は定性的情報（文字列等のカテゴリデータ）、（V）は定量的情報（数値等のバリューデータ）である。

１）住基データ

住基データから作成した変数は以下の通りである。

- 住所（C）
- 建物内人員数（V）
- 建物内最高年齢（V）
- 建物内最少年齢（V）

これらの変数は住所ごとに居住者情報を集計することで作成することができる。建物内人員数の値は同じ建物に住む人員の合計、建物内最高年齢の値は同じ建物に住む人員の年齢の最大値、建物内最少年齢の値は同じ建物に住む人員の年齢の最小値とした。これらの情報は居住家屋が空き家になる確率と相関を持つと仮定できるため抽出した。

２）水道使用量データ

水道使用量データから作成した変数は以下の通りである。

- 水道使用量(V)…1 年間（偶数月に 2 か月分の使用量が記載）の使用量の合計値
- 開栓ダミー(C)…開栓か閉栓かの 2 値データ
- 閉栓月数(V)…現在の年月日と閉栓日との差から計算

水道使用量は、月別水道使用量を合計して年度別に集計した。開栓ダミーは、開栓の状態にあるものを 1 とし、それ以外の場合を 0 とした。閉栓月数について、閉栓日が 0 である場合（すなわち「開栓」の状態にあるもの）は値を 0 とし、それ以外の場合は「水道停止」変数の値を、閉栓日を西暦年月日に直したうえで現在年月日（西暦）から閉栓日を引くことで求めた。

３）建物登記データ
建物登記データから作成した変数は以下の通りである。

- 住所（C）…○○番地△△号
- 建物用途（C）…居宅系、非居宅系の 2 値データ
- 建物構造（C）…木造、鉄骨造、RC/SRC のカテゴリーデータ
- 築年数（V）…現在の西暦年から建築時期（西暦年）を引くことで計算
- 延床面積（V）
- 階数（V）

建物登記データの各項目の文字列情報からカテゴリー変数を作成する方法を説明する。表 1 のレイアウトを参考にすると、「住所」変数は(1,5)の地番から番地を(1,6)から号を文字列として抽出し、結合して作成した。「建物用途」変数は、建物用途を表す(4,4)の文字列が「居宅・○○」であれば、「居住系」とし、そうでなければ「非居宅系」とした（「○ ○・居宅」は、居宅は主な用途でないと判断したため、「非居宅系」としたがこの分類は要検討）。建物構造を表す(4,5)は先頭から 2 文字までの文字列が、「木造」であれば「建物構造」変数の値を「木造」とし、「鉄筋」または「軽量」であれば「鉄骨造」とし、「RC」、「SRC」、「コンクリート造」などの場合は「RC/SRC」とした。床面積を表す(4,6)は各階の床面積が記載されているため、そこから建物階数と総床面積を抽出した。建築された時期を表す(4,7)は先頭から 2 文字目が元号、3 から 4 文字目が年となっているため、西暦年に変換した。そのうえで、「築年」変数の値は現在年月日から建築時期を引くことで求めた。

４）空き家実態調査データ
空き家実態調査データから作成した変数は以下の通りである。

- 空き家ダミー(C)…空き家かそうでないかの二値データ

空き家ダミーはすでに元データである空き家実態調査データに格納されているため、特にデータ加工の必要はない。なお、空き家の地理的分布は予測値との対応をさせる際に重要であるため、事前にその分布傾向を把握しておく必要がある。そこで GIS を用いて 500m メッシュ四方で空き家数を集計した地図を作成した。**図 2** にその結果を示す。

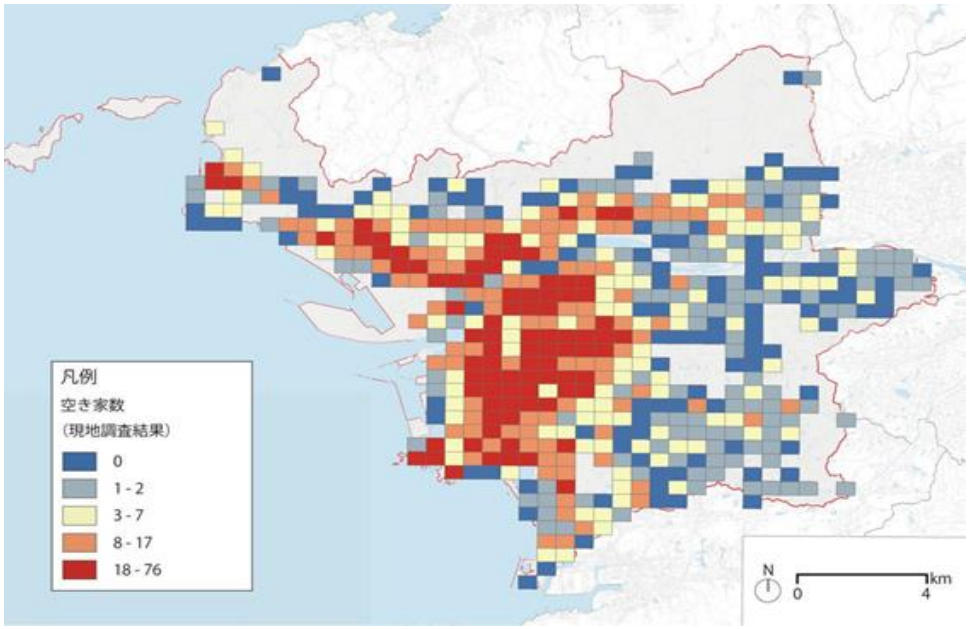

図２　500m メッシュ別空き家現地調査結果（和歌山市全域）

５）国勢調査

以上の公共データは、何れも建物単位のデータであるため、建物固有 ID をキーにして、すなわち建物単位で結合させることができた。一方、国勢調査のデータは建物単位ではなく、複数の建物を含む「基本調査区」の単位で集計が行われている。そこで、本研究では建物 ID と基本調査区同士を、GIS を用いて空間結合することで建物 ID に国勢調査のデータを結合させた。

2.4. データの結合結果

以上の手法により、和歌山市保有のデータを住宅地図の建物ポリゴンに結合させた。結合作業には大きく分けて二つのステップを踏む必要がある。ひとつはジオコーディングであり、もうひとつは建物へのデータ集約である。

2.4.1 ジオコーディングの結果

前述の通り住居表示のものは CSV アドレスマッチングを用いて文字列住所から緯度経度座標を付与し、地番表示のものは地番図の住所と対応させ、地番ポリゴンの重心点を緯度経度座標として得た。なお、CSV アドレスマッチングは緯度経度座標の特定した精度（例えば号レベルの精度や丁目・大字レベルの精度など）を出力させることができる。本研究では住宅地図の建物ポリゴンと空間結合させる必要があるため、建物を完全に特定可能な精度である「号レベル」で位置情報が特定されたデータのみを利用した。**表２**にデータ統合の結果を表す。住基データは 173,769 件のうち、141,719 件（81.6%）に、建物登記データは 142,401 件のうち、96,451 件（67.7%）に、水道使用量データは 160,230 件のうち、122,035 件（76.2%）に号レベルの緯度経度座標を付与させることができた。一見、これらの値は小さいように見えるが、最終的には高い割合で建物に吸着しているため、十分なデータが得られていると考える。一方、空き家実態調査はそもそも緯度経度の情報を得ているため、5,510 件（100%）が緯度経度付きデータとして利用可能である。

2.4.2 建物へのデータ集約の結果

続いて、データがいずれかの建物に吸着した件数を説明する。なお、当該作業は緯度経度座標を付与されたポイントデータと建物ポリゴンとの対応を見るため、データはジオコーディングに成功したデータのみを用いる。住基データは 141,719 件のうち 67,616 件（47.7%）が建物に吸着した。このように比較的低い割合になった理由は、本研究では共同住宅を研究対象から除外しており、アパートやマンションなどの共同住宅に住む世帯のデータが除外されたためである。建物登記データは 96,451 件中 68,330 件（70.8%）が建物と吸着し、比較的高い割合でデータを利用出来ることが分かった。水道情報データは 122,035 件のうち 69,915 件（57.3%）が建物と吸着した。空き家実態調査のデータは 5,510 件中 4,494 件（81.6%）であった。このように空き家実態調査データの捕捉率が高くなった理由は、空き家の多くが戸建て住宅であることに起因するものと考えられる。

表２は利用可能なデータの中で、住基、建物登記、水道情報データがどのような割合で付与されているのかについて表している。分析可能データの件数のうち、1 種類のみのデータが付与されているのは 16,048 件（18.2%）であり、一方で 3 種類全てのデータが付与されているのは 49,789 件（56.3%）であることから、概ね良好に分析データを構築できたといえる。次章以降では、本章で構築したデータを利用して空き家の分布推定を行う。

表２　和歌山市保有データの統合結果

	住基	建物登記	水道情報	空き家調査
(A)元データの件数	173,769	142,401	160,230	5,510
(B)ジオコーディング成功件数	141,719	96,451	122,035	5,510
ジオコーディング成功比率 ((B)/(A))[%]	81.6	67.7	76.2	100.0
(C)建物に吸着したデータの件数	67,616	68,330	69,915	4,494
建物吸着データの比率 ((C)/(B)) [%]	47.7	70.8	57.3	81.6
住基＼(建物登記∪水道情報)				3,229
建物登記＼(住基∪水道情報)				7,659
水道情報＼(住基∪建物登記)				5,160
住基∩建物登記				55,046
住基∩水道情報				55,414
建物登記∩水道情報				59,130
住基∩建物登記∩水道情報				49,789
住基∪建物登記∪水道情報 (分析可能データの件数)				88,363

注: A＼B は B を含まない集合 A、A∪B は A または B を含む集合、A∩B は A かつ B を含む集合を表す。

3. 空き家分布推定のモデルの改良

3.1. 空き家分布推定モデルの考え方

本節では、空き家分布推定モデルの変遷について概説を行うとともに、本報告書における基本モデルの提案を行う。

これまでに空き家分布推定の手法を提案した研究はいくつか見られる。例えば秋山ほか(2018)による空き家得点を算出することで空き家率を推定する手法や、Akiyama et al. (2020)によるクロス集計を使用する手法などが挙げられる。これらの手法は欠損値に対応し、精度の高い結果を得られた一方で、変数を離散化する際の閾値設定が恣意的にならざるを得ないという問題点があった。

そこで、本報告書で提案する基本モデルでは、決定木ベースの機械学習モデルを採用することで、欠損値への対応を行いながら以上の既存研究で課題となっている最適な閾値設定を行えるようにした。また今後、国勢調査や住宅土地統計調査等の公的統計データを活用することも踏まえ、特徴量の拡張が可能な、採用事例の多い一般的な手法を用いることとした。本年度は基本モデルの構築と精度検証までを行うが、今後訓練データの割合を変更するなどすることで、精度検証するプロセスを繰り返し、社会実装が可能なモデルの構築を目指す（**図 3**）。

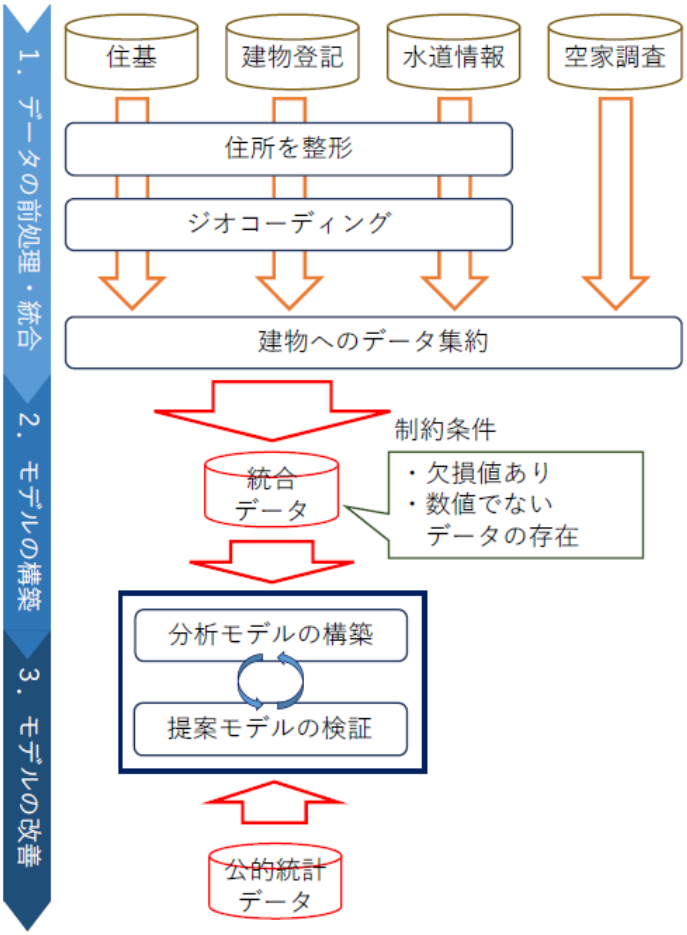

図 3　空き家分布推定モデル構築の全体像

3.2. 推定値の算出方法

モデルの構築には、機械学習的分類手法のひとつである XGBoost (eXtreme Gradient Boosting)を用いた。これは、Chen and Guestrin (2016)によって提案された手法であり、後述する決定木を基本として、それをいくつも作成し、各決定木から得られた値を足し合わせることで最終的な予測値を得る手法である。本節では概説に留めるが、詳しくは付録を参照されたい。

まず、本研究のデータに XGBoost が適合する理由について述べる。まず、当該手法は推定精度が一般的に高く出る点において他手法よりも有利である。一般的に利用されている回帰分析などの手法は、その解釈が容易な一方で、予測のためのモデルではないため、本研究の目的に対しては必ずしも有用とはいえない。例えば、回帰分析は空き家の多寡が延床面積などの変数に対して単調増加（または減少）することを仮定しているが、本研究で扱う現象の場合、それは必ずしも当てはまらない。当該手法は変数の変化に応じて柔軟に推定空き家確率を予測できるため、そのような問題は生じない。さらに、XGBoost では欠損値を扱うことができるが、他の多くの手法では扱うことができない。本研究は複数種類のデータを統合しており、必然的に欠損値の多いデータとなっている。欠損値を削除して推定値を算出すると、真の推定値を得られない可能性があり、その問題を解決することは重要である。以上のような利点を有するものの、当該手法は複雑な事象の場合、上手く予測値を得られない可能性がある。ただしこの点は予測モデルに対して検証データを投入し、予測値の適合度合いから問題の有無を判断することができる。

以下、XGBoost について概説する。当手法は決定木という手法をベースにして考える。決定木は、ある住宅が空き家であるかどうかを予測する際、いくつかの条件を設定して住宅を分岐させた構造をいう。例えば、条件に世帯内最高齢者の年齢、築年数、年間水道使用量を設定すると、**図 4** のようなイメージとなる。ここでは、はじめに世帯内最高齢者の年齢で分岐させ、続いて築年数と年間水道使用量で分岐させている。予測用のデータをこの構造に従って分岐させ、当てはまるデータが多い木の最終分岐点で高い値となる。例えば、世帯内最高齢者の年齢が 80 歳、築年数 30 年、年間水道使用量 100 ㎥の住宅を得た場合、**図 4** の木構造から、0.23 を得る。

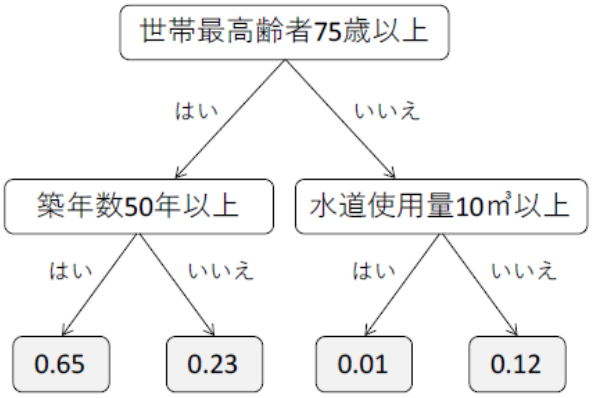

図 4 決定木のイメージ

空き家の推定確率を算出する際、上記の決定木を逐次作成していき、各決定木の結果を足し合わせることで予測値を算出する。1 番目の決定木では、現地調査空き家のデータに基づいて木の構造を作成する。その際に算出される推定空き家確率は、2 番目の決定木の作成に利用される。予測値を改善するため、2 番目の決定木では、1 番目で生じた誤差を評価してその決定木の重み（重要度）を定める。このような操作を予測値が改善しなくなるまで繰り返す。

以上のような流れは**図 5** のように表せる。モデルの複雑さに依存するが、最終的に何千本もの決定木が構築され、それぞれの決定木から算出された値を足し合わせることで最終的な推定空き家確率を得ることができる。このような構造を採用することで、未知のデータにもうまく適合して予測値を返すことができるようになる。

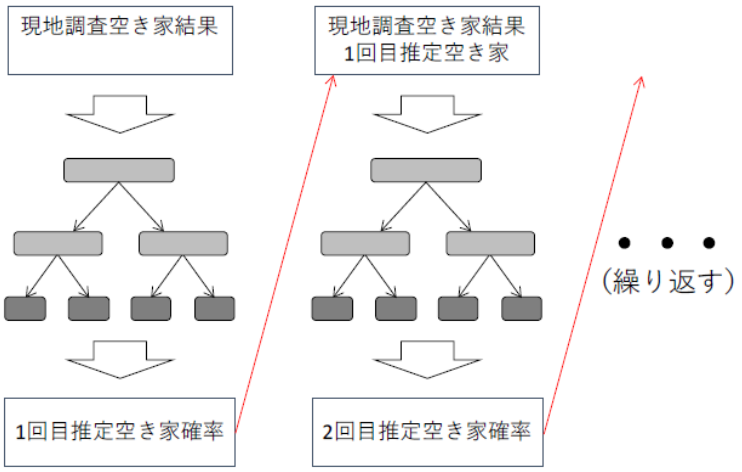

図 5 分析手法のイメージ

本研究では以上で紹介した XGBoost を自治体保有データと国勢調査データから抽出した変数に対して適用することで、建物ごとの空き家確率を推定した。さらにその結果を任意の空間単位（例えば町丁目やメッシュなど）で集計することで、集計単位ごとの空き家数や空き家率を推定することが可能になる。本研究では 4 章で紹介するように、国勢調査における最小の地域単位として、20～30 の世帯から成る基本単位区で結果を作成した。

3.3. 昨年度のモデル開発との相違点

昨年度の分析においては、オープンソース・フリーソフトウェアの統計解析向けのプログラミング言語である R 言語を用いて分析を行った。しかし、今年の分析においてはより汎用性の高い Python 言語を用いて分析環境を構築した。

3.4. 使用データによるモデルの分類

前年度は公共データのみをもとに空き家の推定を行ったが、今年度は公共データに加えて、国勢調査も追加して空き家の推定を行った。本報告書では国勢調査の有用性を検証するために３つのモデルを構築し、比較検討を行った。

モデル１：公共データのみを用いて分析したモデル（以下、「mu」）
モデル２：公共データと国勢調査の両方を用いて分析したモデル（以下、「mu+na」）
モデル３：国勢調査のみを用いて分析したモデル（以下、「na」）

4. モデルごとの空き家分布推定の結果と信頼性の検証

4.1. 検証データを用いた信頼性の検証

まず、本モデルの結果の妥当性について検証を行う。本研究の手法では前年度の手法と同じく、予め訓練用のデー タと検証用のデータを 7:3 の比率で分けておき、訓練用のデータのみでモデルを構築した。 **表 3** は自治体と国勢調査両方の検証データを用いて、チューニングされたモデルの推定空き家と現地調査結果による空き家とをクロス集計したものである。例えば、推定結果が空き家であり、そのうち実際に空き家である 件数は 835 件となる。

なお、精度の検証で多用されるのが「正解率」、「適合率」、「再現率」、「F 値」である。それぞれの値について結果を見ていく。

1）正解率

ここでいう「正解率」とは、全データ件数のうち、正解で合ったものの比率、すなわち真陽値（推定値が空き家で実際に空き家であったもの）と真陰値（推定値は空き家ではなく、実際も居住中の建物、すなわち空き家では無かったもの）の和を全体数で除した値である。本モデルの場合、真陽値が 835 件、真陰値が 25,100 件なので、正解率は 97.8% (25,935/26,509)と良好な結果が得られたといえる。

2）適合率

ここでいう「適合率」とは、モデルにより空き家と予測された建物のうち、現地調査でも実際に空き家であった割合のことである。本モデルの場合、空き家と予測された 896 件の内,実際に空き家だったのは 835 件だったので、適合率は 93.2%と非常に高い値となった。

3）再現率

ここでいう「再現率」とは、現地調査で実際の空き家であった建物のうち、モデルにより空き家と予測された割合である。本モデルの場合、実際に空き家であった 1,348 件の内、空き家と予測したのは 835 件だったので、再現率は 61.9%と適合率に比べて低い値となった。

4）F 値

「F 値(F-measure)」とは再現率と適合率の調和平均を表す指標であり、以下の式で表される。

$$\text{F-measure} = \frac{2 * \text{正解率} * \text{再現率}}{\text{正解率} + \text{再現率}}$$

本モデルの場合、適合率が 93.2%、再現率が 61.9%であったため、F 値は 0.744 となった。

表 3　検証データを用いた推定結果の検証

検証データ		推定空き家数		
		Yes	No	合計
現地調査結果による空き家数	Yes	835	513	1,348
	No	61	25,100	25,161
	合計	896	25,613	26,509

4.2. 3 つのモデルの比較

続いてモデル構築に使用したデータによって分類された、3 つのモデルを対象にそれぞれ比較分析を行った。

4.2.1 公共データのみを用いたモデル（mu）と、公共データに国勢調査を加えたモデル（mu+na）との推定精度の差を比較した結果

まず、公共データのみを用いたモデル（mu）と、公共データに国勢調査を加えたモデル（mu+na）との推定精度の差を比較した結果を**表 4** に示す。この結果より、双方のモデルに大きな推定精度の違いは見られないことが分かった。すなわち、公共データを用いることが可能である場合には、国勢調査を追加することよってモデルの精度が大きく改善するとは言えないことが分かった。

モデルの推定精度を見ると、両方とも 97.8%という高い正解率となった。適合率も高く、空き家であると予測した建物が実際に空き家であることの信頼性は非常に高いと言える。一方で再現率は 6 割程度と、それほど高くない。実際に空き家である建物を、空き家であると正確に予測する精度に関しては、改善の余地があるといえる結果となった。

続いて、基本単位区毎の予測空き家率の誤差を集計すると、二乗平均平方根誤差（以下、RMSE）は 0.083 であった。**図 6**、**図 7** は、その誤差を地図上に可視化したものである。これらの結果より、どちらのモデルも誤差 10%以内の基本単位区（白色の地区）が多くなっていることが分かる。なお、実際より空き家率を高く予測した基本単位区（青色の地区）よりも、低く予測した基本単位区（赤色の地区）が多くなっている理由は、モデルが実際よりも多少低く空き家率を予測していることを表しており、これはすなわちモデルの再現率が低いことと対応している。

表 4　公共データのみを用いたモデル（mu）と公共データに国勢調査を加えたモデル（mu+na）との推定精度の差を比較した結果

	正解率[%]	適合率[%]	再現率[%]	F 値
mu	97.8	91.1	63.1	0.746
mu+na	97.8	93.2	62.0	0.744

4.2.2　国勢調査のみを用いたモデル（na）の推定精度

次に、国勢調査のみを用いたモデルの予測空き家率と、実際の空き家率との誤差を集計した結果、基本単位区毎の RMSE は 0.18 であった。**図８**は、その誤差を基本単位区毎に色分けしたものである。mu や mu+na と比べると白色の基本単位区の数が減り、予測と実際の空き家率との誤差が大きくなったことが分かる。しかし、他の 2 つのモデルと異なり、青色と赤色の基本単位区が同数程度見られ、予測と実際の空き家率との誤差には偏りがないことが分かる。この結果から、空間的により広域な単位（例えば基本調査区や 1km メッシュ等）で集計することで、より精度の高い空き家率の予測を行うことが可能であるものと期待される。

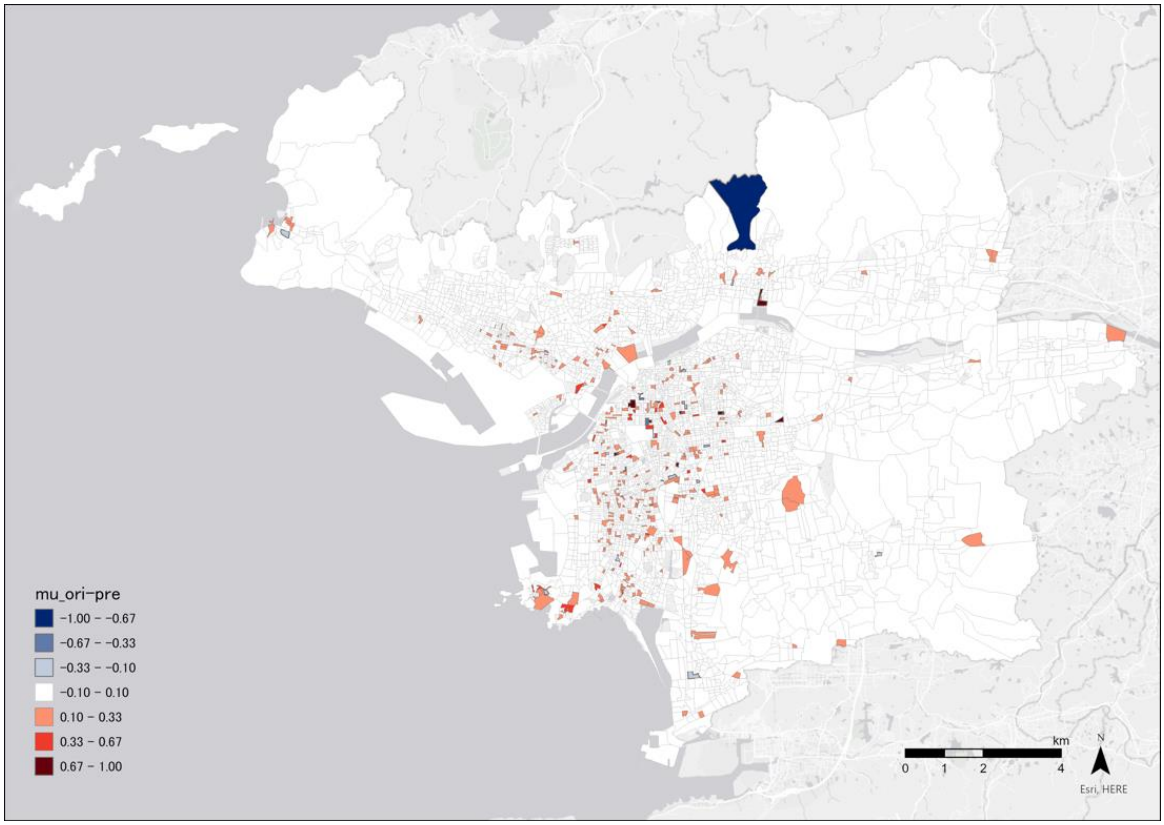

図 6　mu モデルの基本単位区毎の実際の空き家率と予測空き家率との差
（青いほど実際より高く空き家率を予測しており、赤いほど低く予測している。）

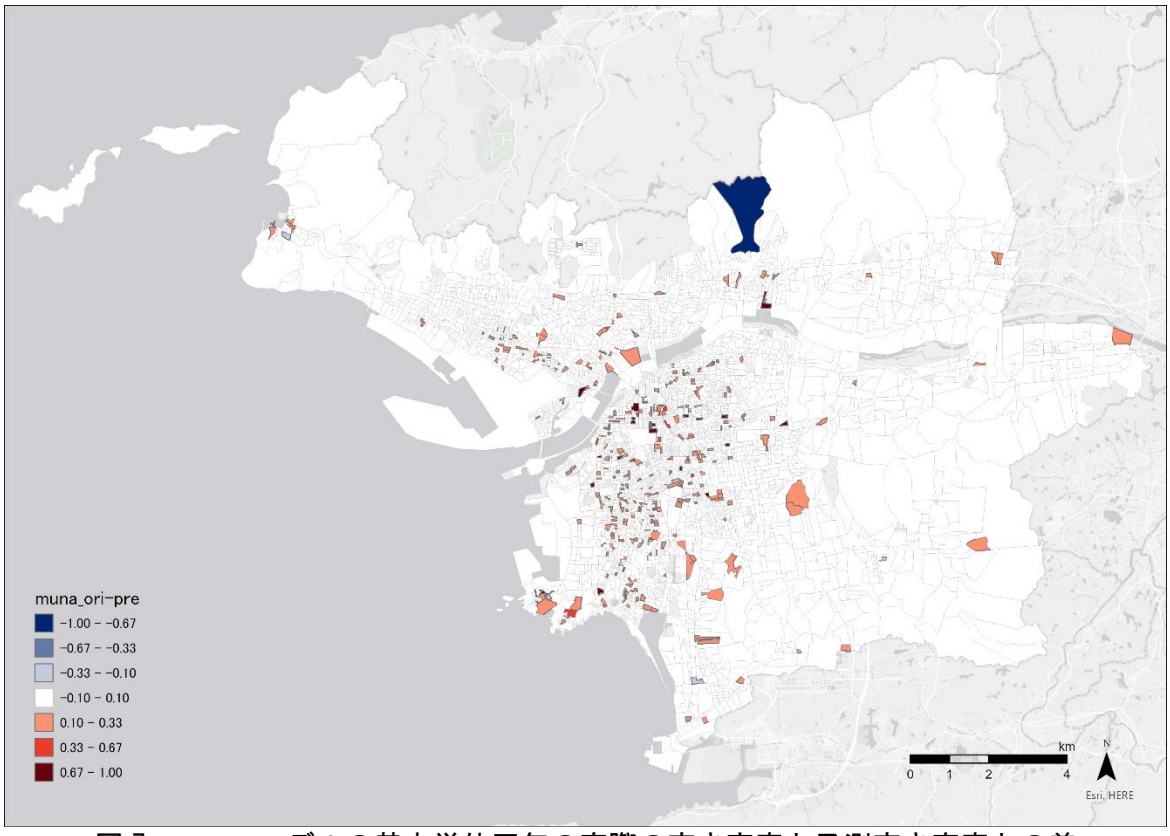

図 7　mu+na モデルの基本単位区毎の実際の空き家率と予測空き家率との差
（青いほど実際より高く空き家率を予測しており、赤いほど低く予測している。）

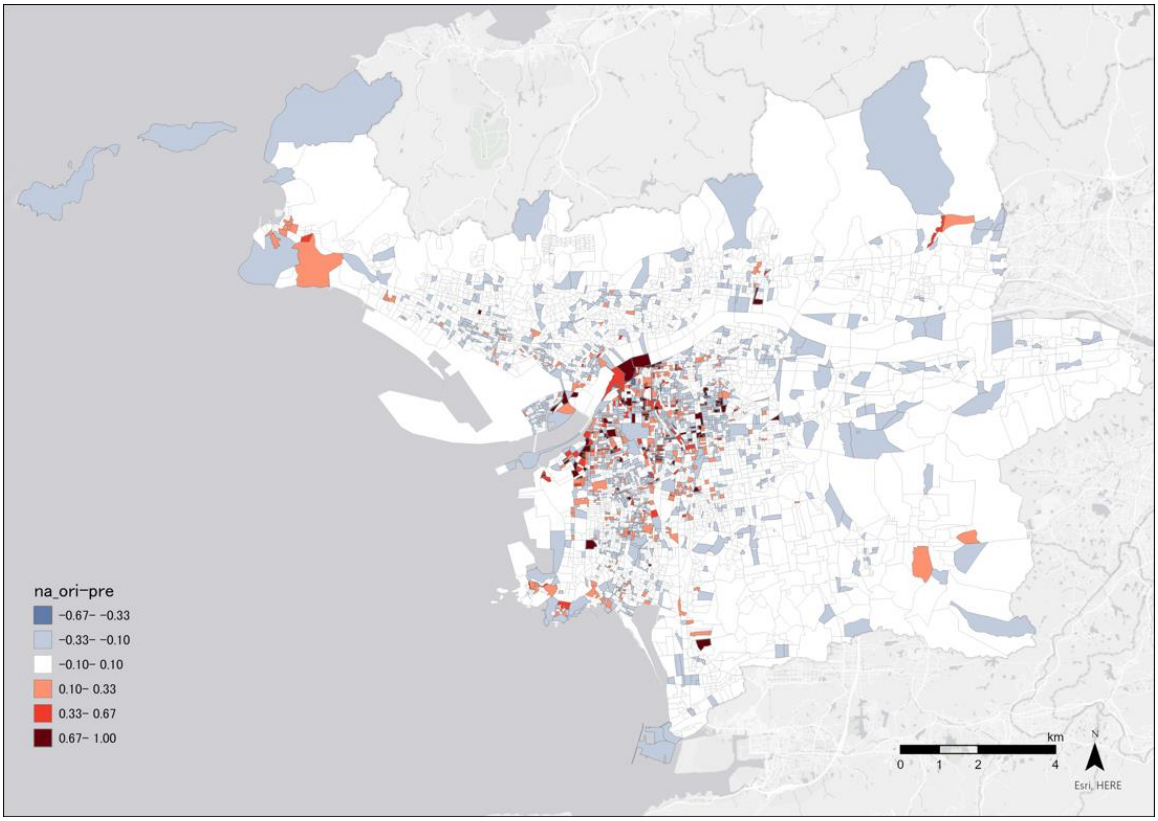

図 8　na モデルの基本単位区毎の実際の空き家率と予測空き家率との差
（青いほど実際より高く空き家率を予測しており、赤いほど低く予測している）

以上の結果より、3 つのモデルの比較分析結果をまとめる。まず、公共データを用いたモデルは 9 割以上の正解率と適合率となった。モデル内で空き家と予測されたものに関しては、実際に空き家であることの信頼性が高いといえる。次に、公共データに国勢調査を加えて分析を行ったが、国勢調査を加えることによってモデルの推定精度が大きく向上することはなかった。これは、機械学習において重要な特徴量の多くに公共データの変数が用いられており、国勢調査の特徴量がモデルに対して及ぼす影響が小さかったためであると考えられる。最後に、公共データを用いずに国勢調査のみで分析を行った結果、公共データを用いたモデルほどではないが、比較的良好な予測精度を得ることができた。本研究で行った基本単位区毎の予測では誤差がバラついてしまうが、基本調査区や 1km メッシュといったより大きなスケールで空き家率の集計を行うことで、さらに誤差の小さな結果が得られるものと考えられる。したがって、公共データを使用できない対象地域においては国勢調査による推定を検討することが有効であろう。

4.3. 水道データ吸着有無によるモデルの分離

3 つのモデルの比較検討の結果、いずれのモデルにおいても影響力が最も大きかったのは水道データであった。しかし、この場合水道データが吸着していないサンプルは、水道データが吸着していないことそのものが、推定結果の信頼性において大きなファクターになってしまう恐れがある。そこでこの問題を解決するために、水道データの吸着有無でサンプルを分けた各々のサンプルにおいて機械学習を行い、それぞれのモデルから得られる推定結果を 4.2 章の mu モデルと比較した。なお、全体のデータのうち水道データが吸着しているものは 20,975 件で、水道データが吸着していないものは 5,535 件であった。

まず、表 5 で、水道データが吸着しているモデルで機械学習を行なった結果を表す。水道データが吸着しているサンプルのみで機械学習を行なったモデルの真陽値は 192 件、真陰値は 20,376 件であったことから、正解率は 98.1%(20,568/20,975)となった。また、適合率は 71.6%(192/268)、再現率は 36.7%(192/523)で、適合率と再現率の調和平均である F 値は 0.485 であった。

次に、水道データが吸着していないサンプルのみで機械学習を行なった結果を表 6 で表す。水道データが吸着していないサンプルのみで機械学習を行なったモデルの真陽値は 681 件、真陰値は 4,710 件であったことから、正解率は 97.4%となった。また、適合率は 100%(681/681)、再現率は 82.5%(681/825)で、適合率と再現率の調和平均である F 値は 0.904 であった。

さらに、表 7 は表 5 と表 6 の結果を足し合わせたものである。水道データ吸着有無で分けることで、全体としての正解率は 97.9%(25,959/26,510)、適合率は 92.0%(873/949)、再現率は 64.8%(873/1,348)、F 値は 0.760 となった。

最後に表 8 に 4.2 でもっとも性能が良かった mu モデルの結果と、水道データ吸着有無により分けて機械学習を行なった表 7 のモデルの結果を比較した結果を示す。水道データが吸着しているサンプルと、水道データが吸着していないサンプルを分けてモデリングをすることで、モデルの精度、適合率、再現率が上昇したことが分かる。

表 5　水道データありのサンプルを用いた推定の結果の検証

検証データ		推定空き家数		
		Yes	No	合計
現地調査結果による空き家数	Yes	192	331	523
	No	76	20,376	20,707
	合計	268	20,452	20,975

表 6　水道データなしのサンプルのみを用いた推定結果の検証

検証データ		推定空き家数		
		Yes	No	合計
現地調査結果による空き家数	Yes	681	144	825
	No	0	4,710	4,710
	合計	681	4,854	5,535

検証データ		推定空き家数		
		Yes	No	合計
現地調査結果による空き家数	Yes	873	475	1,348
	No	76	25,086	25,162
	合計	949	25,561	26,510

表 7　水道データ有無で分けたモデルの推定結果の検証

表 8　mu モデルと水道データの吸着有無で分離したモデルとの比較

	mu モデル	水道データ吸着有無で分けたモデル
適合率[%]	91.1	92.0
再現率[%]	63.1	64.8
正解率[%]	97.8	97.9
F 値	0.746	0.760

5. まとめと今後の検討事項

5.1. 考察

まず、4.2 で行った 3 つのモデルの比較分析について考察する。今年度は前年度の研究を拡張し、政府統計である国勢調査を加えて分析を行った。公共データのみをサンプルに入れて分析したモデルと、公共データと国勢調査両方をサンプルに入れて分析したモデルを比較した結果、国勢調査を加えたとしても推定結果が更に良くなることはなかった。この原因としては、公共データは建物単位で集計が行われている、すなわち建物ごとの結果を収録する非常に空間解像度の高いデータである一方、国勢調査は基本調査区単位で集計が行われていたため、公共データに比べてデータの空間解像度が低く、そのため推定モデルにはうまく反映されなかったものと考えられる。

一方で、国勢調査のみで基本調査区ごとの空き家率を推定するモデルの予測性能は比較的高かった。すなわち、国勢調査のみを用いた空き家分布推定も、建物単位という高い解像度では困難ではあるが、基本調査区、さらには町丁目や地域メッシュなどの単位における空き家率や空き家数の推定においては、十分に利用価値があることが明らかとなった。

次に、4.3 で行った分類モデルについて考察する。水道データの吸着有無によってモデルを分類した結果、モデルの F 値は 0.760 となり、na モデルの F 値 0.746 より高い値を示した。このことから水道データの吸着有無によるモデル分類は、モデルの予測精度を高めるのに有効であると考えられる。

一方、水道データの吸着有無で分類された各モデルの F 値を見ると、吸着したモデルの F 値は 0.485、水道データが吸着していないモデルの F 値は 0.904 で、吸着していないモデルの F 値の方が高い値を示すことが分かった。もし水道の閉栓が空き家分類に重要な変数であったとすると、吸着したモデルの F 値はより高くなるはずである。したがって空き家の分類に予測上重要なのは、水道が閉栓しているか否かということより、むしろ水道データが吸着しているか否かということであると考えられる。この考察は、市が行う空き家実態調査の前提に影響を与える。現在の実態調査は水道閉栓情報等を基に空き家候補を割り出しているが、むしろ水道閉栓情報のない建物こそが空き家の候補になり得るからである。今回の結果は教師データとして用いた空き家実態調査の前提と齟齬を含む部分があり、その原因について来年度精査する必要がある。

5.2. 精度検証を踏まえた改善点の検討

以上のように、本研究で構築したモデルは一定程度の精度を保証し、空き家の分布推定モデルの重要な基盤となるものと期待される。ただし、本モデルは更なる改善を行う上で、具体的に以下 3 点を検討する必要がある。

1）正解データを少なくした上で、どの程度の精度を担保できるか。

本研究では訓練データと検証データの比率を 7:3 としたが、これを 6:4、5:5、4:6 など、様々な割合を適用させていくことが考えられる。例えば正答率 70～80%程度を担保するためには、どの程度のサンプル数が必要かを判断する必要がある。このような閾値を明らかにすることで、より少ない地区を対象とした現地調査から、正確な空き家分布の推定を行えるようになるものと考えられる。

2）特定地域のみに絞って正解データを利用した場合、他地域への外挿はどの程度可能か。

特定の地域（町丁目など）を検証データとして抽出し、残りの地域を訓練データとしてモデルを構築する方法が考えられる。例えば、秋山ほか（2018）は鹿児島市において用途地域ごとに分割したいくつかの地区を対象に現地調査（外観目視による空き家データの作成）を実施しそれを

訓練データとすることで分析を進めている。現地調査の効率化を考慮すると、一部の地区のみを集中的に現地調査し、その結果に基づいて他地区を予測する方が現実的であり、例えば用途地域だけでなく市街化区域と市街化調整区域で分割して訓練データを作成するなどの改善方法も考えられる。

3)水道以外のデータの吸着有無による分類でも、モデルの予測精度を上げられないか。

今年度の研究では、4.2 までに重要な変数として現れた水道データの吸着有無でモデルを分類することにより、モデルの予測精度向上を目指した。結果、モデルの予測精度は向上したが、同時に水道データは閉栓の有無によって空き家が予測されるのではなく、水道データ自体の吸着有無によって空き家が予測されたに過ぎないという結果を得られた。したがって、実際に空き家を予測する上で、他に重要な変数の存在が示唆される。来年度の研究では、こうしたほかの変数に対しても吸着有無によってモデルを分類することで、さらなるモデルの予測精度向上に加えて、どんな変数が予測上重要な変数であるのかを考察することが求められる。

5.3. 他の自治体への拡大の検討

和歌山市を対象としたこれまでの研究により、和歌山市の場合、公共データから空き家の分布をかなり高い精度で予測することができることが明らかとなった。そこで次のステップとしては、他の自治体（例えば和歌山県内の他の市町村）への適用が考えられる。具体的には和歌山市で構築したモデルを用いて、他の自治体の空き家の分布状況を推定したり、和歌山市では利用できなかった他の公共データを利用したモデルを構築したりする、といった展開が挙げられる。

なお、他の自治体へ空き家推定を展開していく上では、政府統計がかなり重要な役割を果たすものと期待される。今年度の分析から、政府統計の 1 つである国勢調査は自治体が保有するデータと比較して、空間的な解像度は低いものの、基本調査区ごとのスケールであれば空き家率を効果的に予測できることが明らかになった。公共データは自治体によってフォーマットが異なるため、データの前処理やデータ同士の結合処理を行う必要があるという課題が挙げられるが、政府統計は全国で統一されたフォーマットとなっているため、和歌山県全域、さらには日本全国に本研究を拡大していくことが容易であると考えられる。

5.4. 将来推計の可能性

昨年度まで、そして今年度の研究成果は基本的に現状の空き家分布の状態を推定・把握しようとするものであるが、将来的な実用性や研究の拡張性を考慮すると「将来推計」、すなわち今後どの地域でどの程度空き家が増加する可能性があるか、ということを推定する手法の検討・開発も重要な研究課題となりうると考えられる。例えば現時点の政府統計や公共データに加えて、過去の政府統計や公共データを組み合わせてその時系列変化のトレンドを把握し、同様の時系列変化が今後も継続すると仮定することで（一般に「コーホート変化率法」と呼ばれる（厚生労働省, 2003 年））、将来推計がある程度可能になるものと期待される。また、現状の政府統計・公共データと将来推計人口（国立社会保障・人口問題研究所, 2017 年）を組み合わせることで、将来推計人口から地域ごとに将来減少する人口を、政府統計や公共データから世帯主年齢別の世帯数、そして性別年齢別の死亡率（人口動態統計）を掛け合わせる方法でも、ある程度の推定が可能になるものと期待される。

なお以上のような将来推計の課題は、推定精度の検証方法が挙げられる。これは現時点で結果が不明な将来の状況を推定するため、その結果の確からしさの検証ができないためである。この点については今後政府統計等で結果が明らかになった時点で検証したり、過去の統計から現在を推定し、その確からしさを検証したりすることで、検証できるものと考えられる。

今年度は以上の内容は実施できなかったが、以上のようにその可能性については検討を行うことが出来た。今後は以上の手法の開発や信頼性検証を行っていきたい。

5.5. 総括

以上のように、本研究を通して和歌山市が保有する公共データ（住民基本台帳、水道使用量情報、建物登記情報）と、市による空き家調査データ、また公的統計データ（国勢調査）を活用することにより、和歌山市全域の空き家分布状況を迅速、安価に推定するモデルの構築が実現した。また、同モデルの信頼性の検証も実施し、その結果、充分に高い信頼性であることも明らかとなった。さらに同モデルで得られた結果から、和歌山市全域の空き家率を推定し、その結果を可視化（地図化）することも実現した。加えて、同モデルに国勢調査を組み合わせて活用する方法、あるいは国勢調査のみから空き家率を推定するモデルの構築が実現した。

今後は、今年度構築したモデルを改善して信頼性の向上を目指すとともに、他の自治体（例えば和歌山県内の他の市町村）への展開も検討したい。また空き家分布の将来推計についても検討を進めていく予定である。

謝辞

本研究で使用した住宅地図（2019 年版 Zmap TOWNⅡ（株式会社ゼンリン）は、東京大学空間情報科学研究センターとの共同研究（No. 880）により使用したものである。ここに記して謝意を表したい。

参考文献

秋山祐樹・小川芳樹・仙石裕明・柴崎亮介・加藤孝明，「大規模地震時における国土スケールの災害リスク・地域災害対応力評価のためのミクロな空間データの基盤整備」，第 47 回土木計画学研究・講演集，CD-ROM(392), 2013.

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,744, 275-283, 2018.

浅見泰司，「都市の空閑地・空き家を考える」，プログレス，2014.

厚生労働省，「地域行動計画策定の手引き Ⅱ.人口推計」<https://www.mhlw.go.jp/topics/bukyoku/seisaku/syousika/030819/2b.html>，2003.（最終アクセス日：2021 年 3 月 25 日）

国立社会保障・人口問題研究所，「将来推計人口・世帯数」<http://www.ipss.go.jp/syoushika/tohkei/Mainmenu.asp>，2017.（最終アクセス日：2021 年 3 月 25 日）

西山弘泰，「宇都宮市における空き家の特徴と発生要因-宇都宮市空き家実態調査の結果から-」．駿台史学 153, 55-74, 2015.

Akiyama, Y., Ueda, A., Ouchi, K., Ito, N., Ono, Y., Takaoka, H. and Hisadomi, K., Estimating the Spatial Distribution of Vacant Houses using Public Municipal Data, *Geospatial Technologies for Local and Regional Development*, 165-183, 2020.

Akiyama, Y. and Ogawa, Y., “Development of Building Micro Geodata for Earthquake Damage Estimation”, *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 5528-5531, 2019.

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.

Yamashita, S. and Morimoto, A., “Study on Occurrence Pattern of the Vacant Houses in Local Hub City”, *Transactions of CPIJ*, 50, 932-937, 2015.

付録

XGBoost による推定値の算出方法

XGBoost は決定木の中でも回帰木を利用しており、建物ごとの空き家か否かの判定は「空き家確率」として最終的に出力される。本モデルは決定木を逐次的に学習させていくものであり、t番目の木を学習させるためには、$t-1$番目までの全ての木の情報を用いる。ただし、木の数が大きくなるにつれ、誤差が小さくなるため、その改善の余地が少なくなっていく。

空き家の推定確率を算出する際、大まかな流れは以下の通りである。

- 決定木をT本作成する。そのため、以下の流れを繰り返し行う。
- 決定木は、分岐を繰り返すことで作成し、その際に特徴量（例えば、住基の建物内人員数など）の閾値を設定する。
- 閾値設定は全ての候補を調べ、分岐させた際に最適な葉の重みを設定したとき、損失関数の減少が最大になるものを選択する。
- 上記の決定木の作成により予測値を更新する。

続いて、損失関数に基づく最適な重み付け値について簡潔に述べる。当モデルにおいて、t番目の決定木をf_tとおくと、t番目の単純な形式での損失関数は$\sum_{i=1}^{I} l\left(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)\right)$と表せる。ただし、$x_i$は$i$番目の入力データ、$y_i$は$i$番目の出力データ、$\hat{y_i}^{(t-1)}$は$t-1$番目までの木を用いた$i$番目出力データの予測値（分類の場合には確率値）である。式の意味するところは、$t-1$番目までの決定木をベースとして、$f_t(x_i)$を加味することでt番目の損失をさらに減少させることを意味する。ただし、上記の損失関数では過学習してしまう恐れがあるため、罰則項を加えて修正し、t番目の損失関数$L^{(t)}$を、

$$L^{(t)}(f_t) = \sum_{i=1}^{I} l\left(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)\right) + \gamma T + \frac{1}{2}\lambda \|w\|^2$$

と表す。ただし、Tは最終ノードの数を、wは最終的な出力値の集合を示している。ここで、最終ノード数が増えるほど、出力値の種類が多くなるほど学習データに対応出来るようになるが、それによる過学習を抑えるため、γ, λはそれぞれ罰則項の調整パラメータとして設定する必要がある。$\sum_{i=1}^{I} l\left(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)\right)$の推定に際して、$f_t(x_i)$まわりの Taylor 展開を二次の項まで行うことで近似解を求められる。最終的に、$L^{(t)}$を最小化することにより、j番目の最適な重み付け値w_j^*を求めることができる。

XGBoost はパラメータチューニングを必要とするため、グリッドサーチによるチューニングを行う。木の深さの最大値を表す max_depth は 2—8、子ノードでのデータの重み付け合計値の最小値を表す min_child_weight は 1—3、各木でのランダム抽出される列の割合を表す colsample_bytree は 0.5—1.0、各木でのランダムな抽出を表す subsample は 0.5—1.0 の範囲でそれぞれ試行する。

モデルの評価関数としてはエラー率$error_i$を採用しており、下記の式で表される。

$$error_i = \frac{1}{N}\sum_{i=1}^{I} |y_i - [\hat{y_i}]|$$

ただし、y_i: 空き家ダミー(空き家である場合 1 、そうでない場合 0)、$\hat{y}_i$: 推定空き家確率、| ・ |: 括弧内の絶対値、[・]: 括弧内が 0.5 以上であれば 1、そうでなければ 0 を返す関数、N: データ数である。パラメータを変化させた際に、訓練データと検証データでクロスバリデーションを行い、検証データのエラー率が改善しなくなるまで行う。その後、エラー率の最も低いパラメータセットをチューニングパラメータとして採用する。

このように、XGBoost は一般の関数を想定しており、チューニングパラメータを適切に設定することで精度の高い予測値を実現している。なお、XGBoost では欠損値の有無も決定木の分岐条件に含むことができるため、複数のデータを組み合わせることで欠損値が多くなる問題にも対応している。