令和元年度
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 2 年 3 月

東京大学空間情報科学研究センター　助教　秋山 祐樹
東京大学空間情報科学研究センター　特任研究員　馬場 弘樹
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

近年、日本では人口・世帯数の減少や少子高齢化、大都市への人口移動などを背景に、全国的に空き家が増加し続けている。最新の「住宅・土地統計調査」（総務省統計局）によると、2018 年の日本全国の空き家数は約 846 万戸、空き家率は 13.6%となっており、空き家数、空き家率ともに過去の調査と比較しても、最も高い値となっている。特に、管理が不十分なまま放置されている状態の空き家（いわゆる「特定空家」）は、腐朽・破損による倒壊危険性の増大や、防犯性の低下、地域全体の魅力・活力の低下など、近隣住民だけでなく地域全体に深刻な影響をもたらす可能性があるとされる（浅見，2014）。

こうした背景を受けて、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）」が施行され、自治体は同法に基づいて空き家対策の取り組みを進めることになった。同法では特定空家等に対する措置だけではなく、空き家等の利活用の促進も重要な施策の一つとして位置づけられている。こうした状況を受けて、現在全国の自治体では空き家対策に関する計画を策定するなどの対策に取り組み始めている。こうした取り組みを実施するうえで必要になるものが、地域内の空き家に関する情報の把握であり、空き家の空間分布の把握に対するニーズは高まりつつある。また、同法の中では自治体全域を対象とした空き家に関する情報の把握やデータベースの整備などを自治体の努力義務としている。したがって、同取り組みを進めていくにあたり、空き家の空間分布を把握することが必要である。しかし自治体全域という広域を対象とした空き家の空間分布を把握するための主な手法は、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。そのため、自治体による現状の空き家の分布調査は、多くの時間、費用、労力を要している。また、個別目視に頼る方法では、把握した空き家分布の情報を更新する度に多大な予算や人員を確保する必要が生じてしまい、定期的かつ継続的な調査の実施と情報の更新が困難であることも大きな問題である。

1.2. 空き家分布推定の先行研究と本研究プロジェクトの位置づけ

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の把握の例がある。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アプローチとして、自治体が所有する公共データである住民基本台帳や、建物登記情報、水道利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、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、和歌山市で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を開発できるものと期待される。その結果、和歌山市において迅速かつ安価な空き家分布調査を、今後も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、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。

1.3. 本研究の目的

以上を鑑み、本研究の目的は和歌山市が保有する各種公共データ（今年度の場合、住民基本台帳、水道使用量情報、建物登記情報）と、空き家分布の調査データを活用することで、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を構築するとともに、同モデルの信頼性の検証を実施するものとする。また同モデルを用いて、和歌山市全域の空き家率を推定し、その結果を地図化した結果を作成するとともに、市が実施した空き家分布の調査データから得られる真値とモデルによる推定値を比較し、その違いの原因を考察する。さらに、同モデルに公的統計データを追加して活用する方法についても検討する。

本報告書の構成は以下の通りである。まず、2 章において和歌山市の各種公共データの概要とそれらの処理・統合の方法を説明する。次に、3 章において空き家分布推定のための基本モデルの構築方法についてまとめる。さらに、4 章では 3 章で構築したモデルを用いて推定した和歌山市全域の空き家分布推定結果を紹介するとともに、推定結果の信頼性の検証結果も紹介する。最後に、5 章において今年度成果のまとめるとともに、今年度成果に公的統計データを組み合わせて活用する方法についての検討、さらに今後、本研究を更に発展させていく上での検討事項についてまとめる。

2. 和歌山市公共データの概要と処理・統合の方法

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・水道情報データのファイルは CSV 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構成したものである。

2.1. 自治体保有データの概要

住基データ

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。

建物登記データ

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 2 の建物番号 24 の 1 行目は物件情報、建物、既存、番地、号となっており、建物番号 25 の 1 行目も同じ形で項目が並んでいる。2 行目以降も同じ規則が成立している。4 行目は主である建物情報の項目が並んでいる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ので、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

水道情報データ

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。

空き家実態調査データ

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。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. 自治体保有データの前処理と統合

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。

住所の表記ゆれの問題に対応する一つの方法として、「ジオコーディング」と呼ばれる住所を緯度・経度といった位置座標に変換する手法が挙げられる。座標はユニークな値なので、住所のような表記ゆれの問題は発生しない。この結果、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である。しかしながら、住所が適切に設定されていない場合や、住所の参照元となる住宅地図が新規開発などによる住所の変更・新設を捕捉しきれていない場合があり、一定程度の誤差が発生してしまう点に注意が必要である。今回、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。

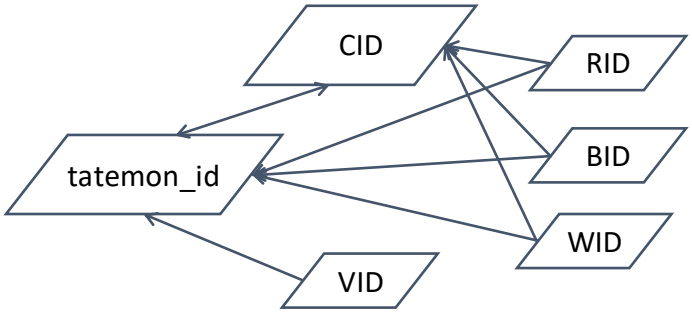
住所データを座標に変換する前に注意すべき点としては、CSV アドレスマッチングサービスは○○県○○市○○町○○番地という形式の住所でなければ正確には読み込めないため、水道情報・住基・建物登記データの住所を整形する必要がある。住所の整形作業の例として、省略してある都道府県名の付加や入力ミス等によって住所が文字化けするなどの個別の問題に対処する等の作業がある。この後、各公共データは CSV アドレスマッチングサービスを使って住所を座標に変換するが、建物登記データはその前に住所ごとに必要な項目の抽出やデータ整形作業を実施し、その後住所を座標に変換した。

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にデータを対応させる。この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、通常は一致しない座標が一定数存在する。そのため、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座標から見て最も近い建物ポリゴンにデータの値を付加するなどの方法で、建物ポリゴンに各

データの値を与えた。一方で、空き家実態調査データは既に緯度経度座標が割り当てられており、各建物ポリゴンに空き家判定の結果を与えた。上記の作業の結果、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、データ統合作業は完了する。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。

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ことである。このため、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ている。

以上より、図 1 のような統合フローによって tatemono_id に住基・建物登記・水道情報のいずれかが対応した場合、分析データとしてサンプルに加えた。



注: 建物固有 ID：tatemon_id、地番：CID、住基：RID、建物登記：BID、水道：WID、空き家：VID である。

図 1　データ統合のフロー

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。

住基データから作成した変数は次の通りである。

- 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。

建物登記データから作成した変数は次の通りである。

- 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_登記」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_登記」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。

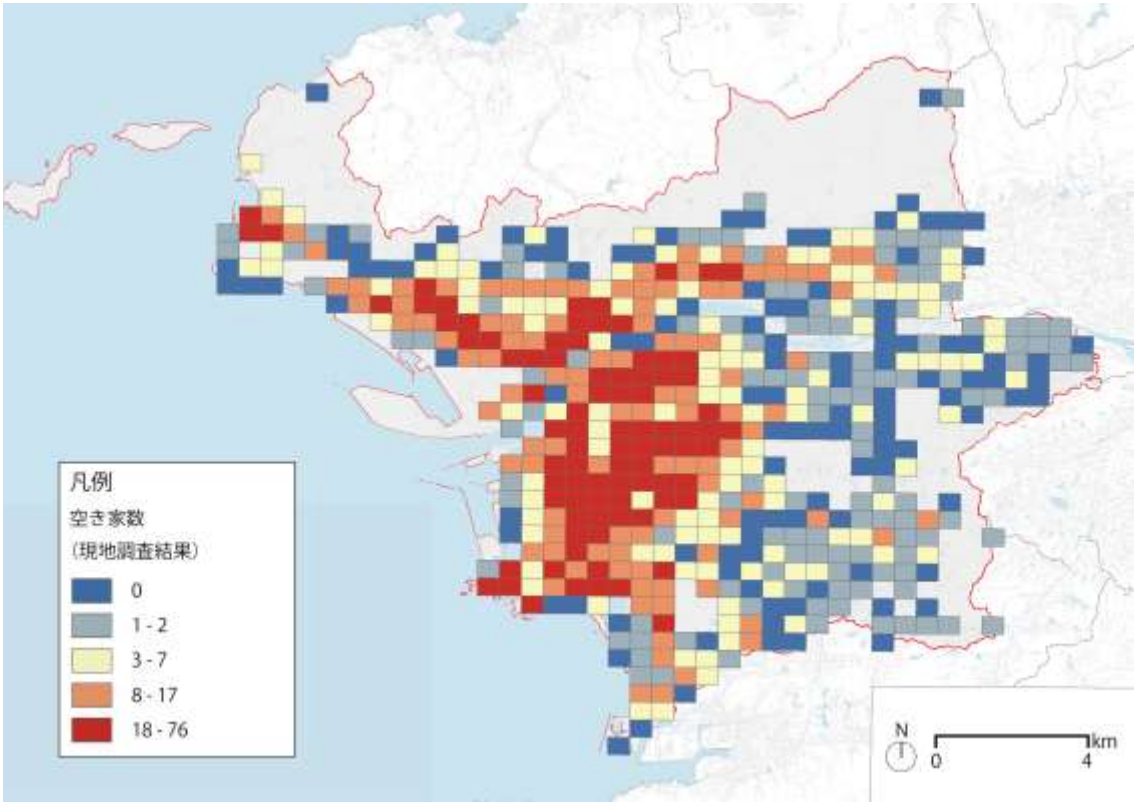


図 2　500m メッシュ別空き家現地調査結果（和歌山市全域）

水道情報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。

空き家実態調査から作成する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に格納されているため、特にデータ加工の必要はない。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、図 2 に 500m メッシュ[1]として表す。

2.3. 自治体保有データの統合結果

以上のような考えの元、提供頂いた和歌山市保有データについて統合を行った。統合作業には大きく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、もうひとつは建物へのデータの集約である。

まずジオコーディングの成果について説明する。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レベル、例えば号レベルや住区レベルなど、を出力する。今回、建物を完全に特定可能な号レベルでの緯度経度座標のみを利用した。表 2 は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%）が正確な緯度経度座標を得た。建物登記データは 142,401 件のうち、96,451 件（67.7%）が緯度経度座標を付与された。一見、この値は小さいように思えるが、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。水道情報データは 160,230 件のうち、122,035 件（76.2%）が緯度経度座標を付与された。また、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。

続いて、データがいずれかの建物に吸着した際の件数を説明する。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

[1] 本研究では国勢調査の 4 次メッシュを利用しているため、厳密には 500m メッシュではないが、簡便性を期して 500m メッシュと表記している。

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。

オコーディングに成功したデータを用いている。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のは、本研究で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に住む世帯のデータが除外されたためだと考えられる。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ていることがわかる。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。このように高い捕捉率になったのは、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負よされているのかについても表している。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%）であり、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。

3.　空き家分布推定の基本モデルの構築

3.1.　基本モデルの考え方

本節では、空き家分布推定モデルの変遷を概説し、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論文はいくつかあり、例えば Akiyama et al. (2020)によるクロス集計を使用する手法や、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は恣意的になる問題点があった。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し、欠損値への対応と最適な閾値設定を行えるようにした。今後、国勢調査・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、今後訓練データ割合を変更するなどして精度検証するプロセスを繰り返し、社会実装可能なモデルの構築を目指す（図 3)。

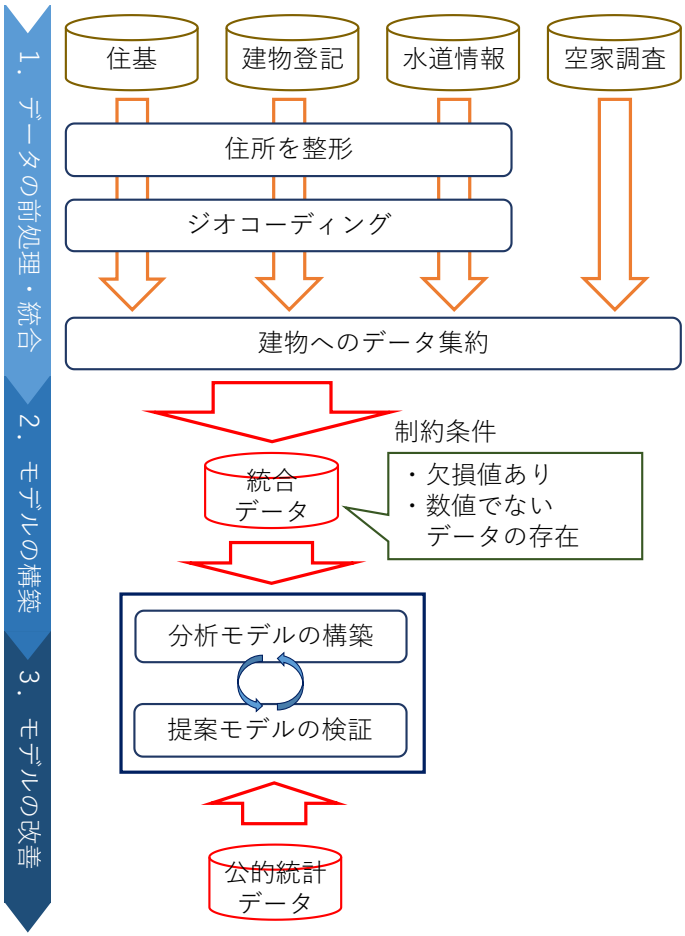


図 3　分析手順の概要

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、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

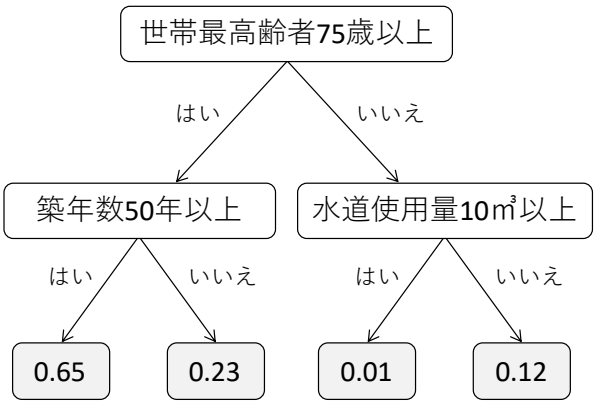


図 4　決定木のイメージ

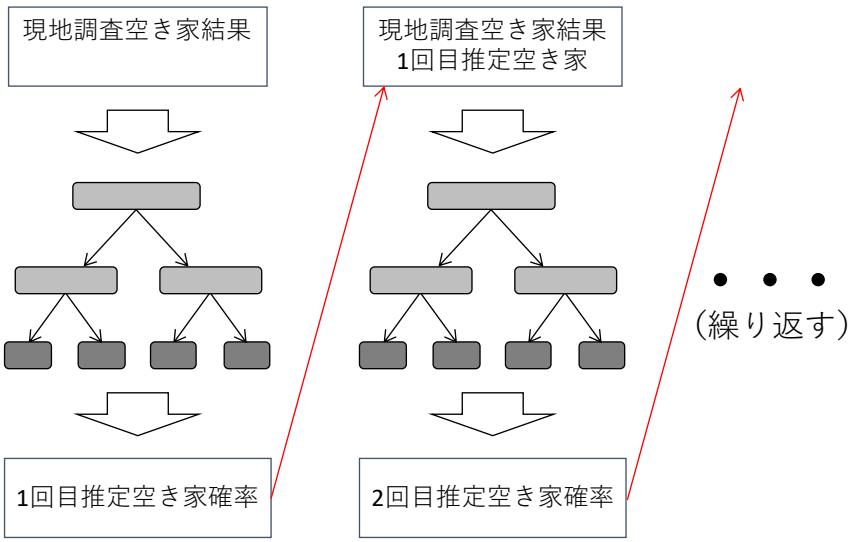


図 5　分析手法のイメージ

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。

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。

本研究では以上で紹介した XGBoost を住基・建物登記・水道情報データから統合して構築した分析データ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 500m 四方メッシュ（4 次メッシュ）や 250m 四方メッシュ（5 次メッシュ）で結果を作成した。

4. 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、推定空き家であり実際に空き家である件数は 1,044 件となる。

精度の検証で多用されるのが「正答率」、「真陽性率」、「偽陽性率」である。「正答率」は全データ件数のうち、正解で合ったものの比率、すなわち推定空き家で実際に空き家であったものかつ、推定空き家でなく実際には居住中の建物であったものの和を全体数で除する。正答率は 95.4% (25,279/26,509)と良好な数値が得られたといえる。しかしながら、空き家調査で注目されるのは現地調査で空き家と判定された際の正答率や居住中と判定された際の誤差率である。「真陽性率」は、現地調査で空き家と判定された際の正答率を表す指標であり、医療分野などで多用されている。一方、「偽陽性率」は現地調査で居住中と判定されたもののうち、間違って空き家と推定された建物の件数の比率である。本分析では、真陽性率と偽陽性率はそれぞれ 77.0% (1,044/1,356)、 3.7% (918/25,153)と算出された。この結果から、実際に空き家である建物を空き家と判定する方法については改善の余地がある一方、居住中の建物、すなわち非空き家の判定については誤差率 5%以下という高い信頼性で判別できるということを示している。

このように、市全域では基本モデルを用いて良好な結果を得られたが、地域ごとにどの程度誤差が生じるのかについての検討が必要であるため、次節において検討を行う。

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	1,044	312	1,356
	No	918	24,235	25,153
	合計	1,962	24,547	26,509

注：正答率 = 95.4% (25,279/26,509); 真陽性率 = 77.0% (1,044/1,356); 偽陽性率 = 3.7% (918/25,153)

4.2. 市全域での推定結果

はじめに、市全域での推定結果を示す。図 6 は 500m メッシュでの推定空き家数の誤答数、すなわち推定空き家と空き家現地調査結果の差分の絶対値を、図 7 は推定空き家数の誤答率を表す。ここで、図 7 の誤答率とは、推定空き家が実態と異なっていた場合の件数（誤答数）をメッシュ内総件数で除したものである。なお、実際の推定は建物単位で行われるが、個人情報保護の観点からメッシュ単位で再集計した。同様の理由で、メッシュ内のデータ件数が 3 件未満である場合、図から削除した。

市全域で推定空き家の誤答数をみると、和歌山市中心市街地付近で大きな値を取っており、山東地区などの山間部で小さな値を取っていることがわかる。さらに、中心市街地から紀ノ川を挟んだ北側は比較的新しい市街地になっているが、そこでも比較的大きな値を取っていることがわかる。ただし、誤答数の程度は中心市街地やその北部ニュータウン、加太地区の一部を除き 10 件未満に抑えられている。特定のメッシュ内で比率でなく総数を議論するのは、例えば間違って推定された空き家が集積している地区を特定する際に有益であるが、本来誤答数の程度はメッシュ内の対象建物件数に依存すると考えらえる。従って、推定空き家の誤答率についても考察を加える。

図 7 は推定空き家数の誤差率について地理的分布を可視化したものである。図 6 の分布と同様に中心市街地付近は誤答率が高いが、図 6 で値が小さかった山東地区周辺や加太地

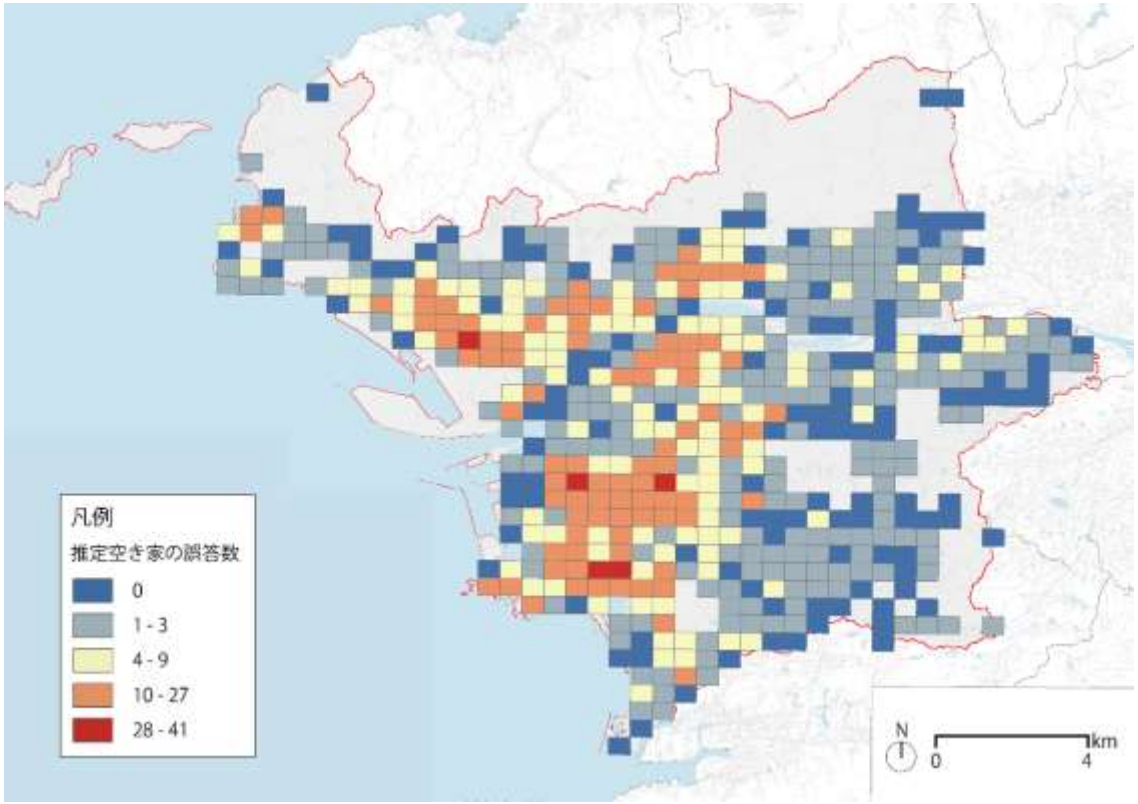


図 6　500m メッシュ別推定空き家の誤答数

区周辺などでも一部高い誤答率になっている。これは、2 つの可能性が考えられる。第一に、中心市街地では空き家形成の要因が複雑であり、対象建物件数が多くても誤答が生じてしまう可能性である。第二に、加太地区や山東地区は対象建物件数が少ないためにメッシュ間でばらつきが大きくなり、結果的に誤答率の高いメッシュが生じてしまう可能性である。ただし、推定空き家の誤答率の分布は集積して分布している訳ではなく、高い誤答率の周辺に低い誤答率のメッシュが存在するなど、空間的異質性が高いといえる。従って、次節で誤答数、誤答率の傾向に基づき地区を指定し、詳細な空間的傾向を把握する。

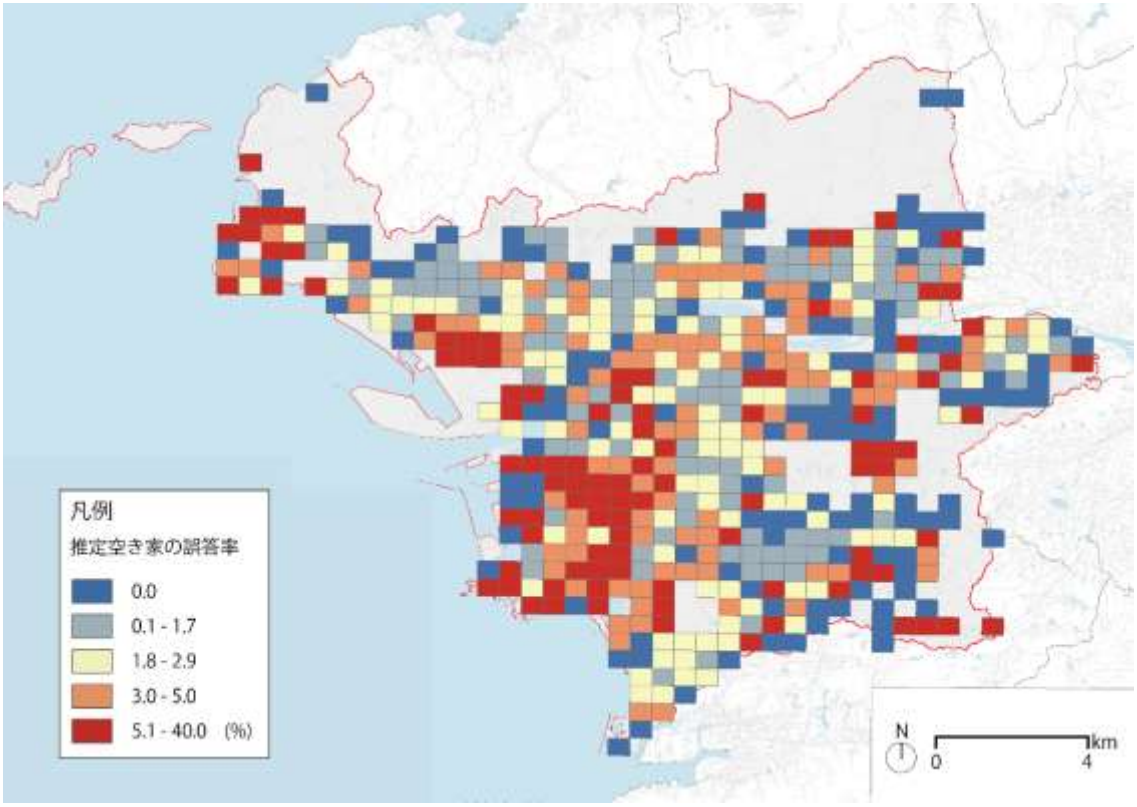


図 7　500m メッシュ別推定空き家の誤答率

4.3.　地区別の推定結果

本節では、誤答数、誤答率の傾向が顕著であった加太地区、山東地区を 250ｍメッシュ[2]で詳細に空間的な傾向把握を行った。なお、図 8 は対象地区（および中心市街地周辺）の位置を示したものである。

はじめに、図 9 は中心市街地周辺における 250m メッシュ別の推定空き家の誤答数、図 10 は 250m メッシュ別の推定空き家の誤答率を表す。全体を見ると、中心市街地周辺では

[2] 本研究では国勢調査の 5 次メッシュを利用しているため、厳密には 250m メッシュではないが、簡便性を期して 250m メッシュと表記している。

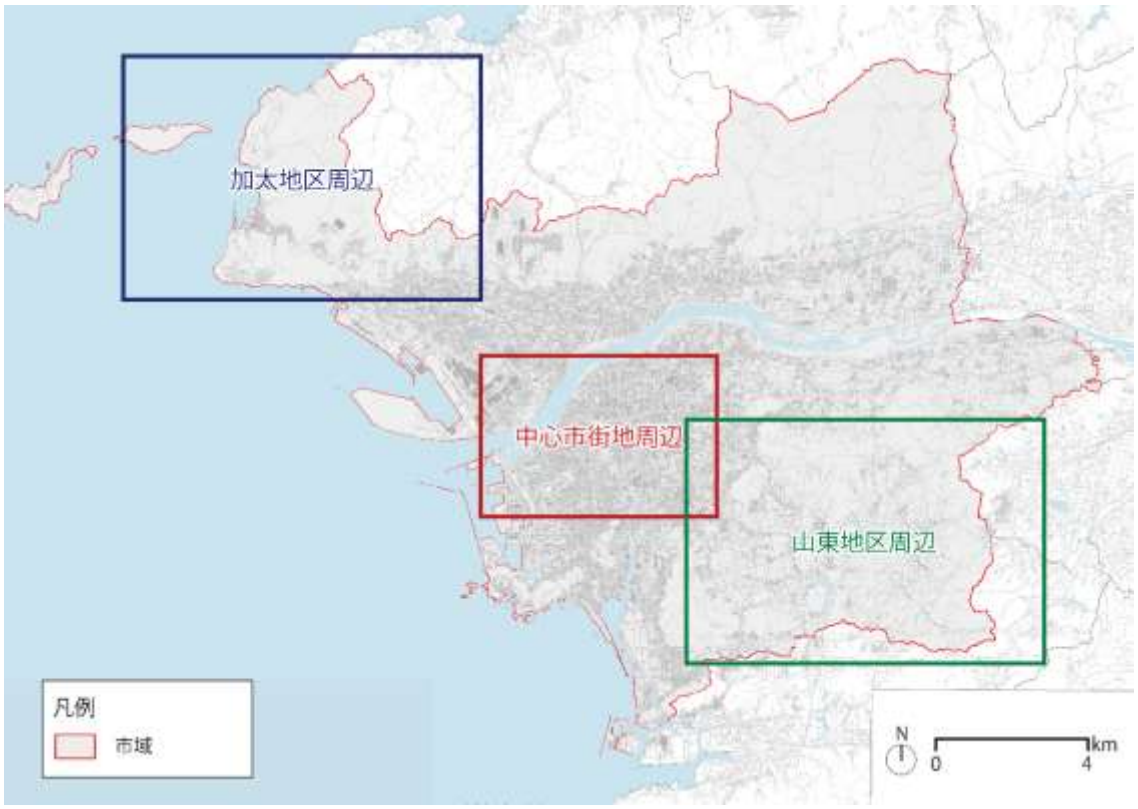


図 8　推定対象地区

誤答数が 10 を超えるメッシュが比較的多いが、誤答率でみると 25%を超えるメッシュは少ない。図 9 より、和歌山城南部で誤答数が多いことがわかる。当該地区周辺は住宅地が分布しており、住宅の絶対数が多くなるために 250m メッシュあたりの誤答数が多くなると推察される。図 10 で該当地区の誤答率が小さいことからもその傾向を示すことができる。

次に、図 11 は加太地区における 250m メッシュ別の推定空き家の誤答数、図 12 は 250m メッシュ別の推定空き家の誤答率である。誤答率の高い 2 つの橙色のメッシュの内、北東に位置するメッシュは、その大半が山間部である。メッシュ内の対象建物件数は 5 軒で、いずれも住宅密集地から少し外れた山沿いに位置していた。また、南西に位置するメッシュに関しても、その大半が山間部と田園地帯であり、対象建物件数は 3 軒のみであった。

続いて、図 13 は山東地区における 250m メッシュ別の推定空き家の誤答数、図 14 は 250m メッシュ別の推定空き家の誤答率である。山東地区には誤答率が 25.1%以上のメッシュが 3 つ含まれている。加太地区と同様に、全てのメッシュにおいて、面積の半分以上は山間部で、対象建物件数はいずれも 3 軒であった。

以上より、誤答数の大きいメッシュでは集計建物数が大きく、一方で誤答率の高いメッシュでは、集計建物数が少ないことが分かる。すなわち、メッシュ内の建物数が少ないメッシュでは、誤答数が 1〜2 件あっただけでも、その誤答率が大きくなってしまうため、誤答率のばらつきが大きくなることに注意が必要である。

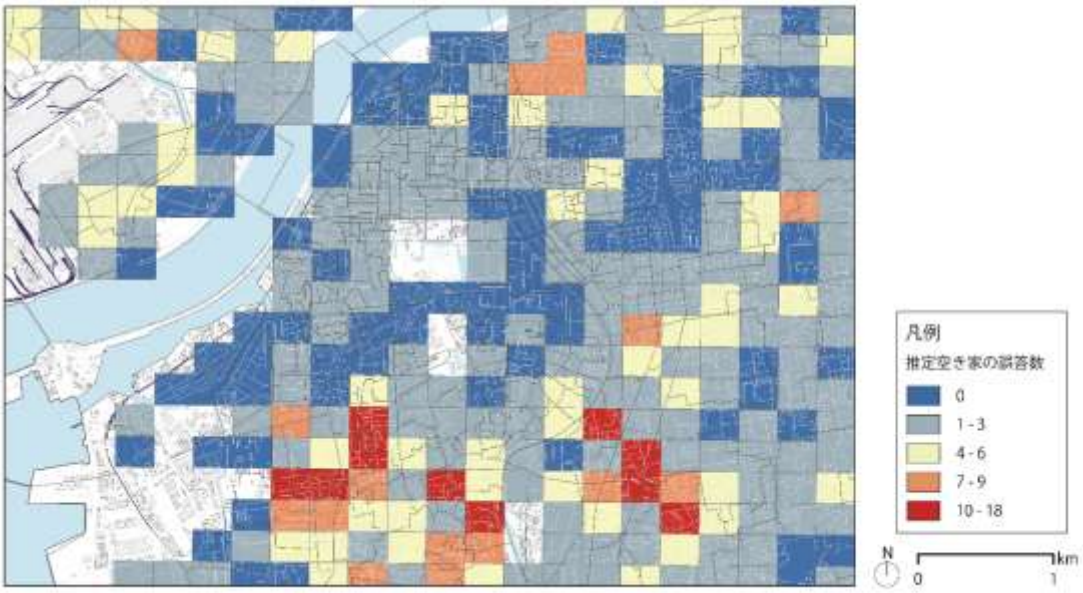


図 9　250m メッシュ別中心市街地内推定空き家の誤答数

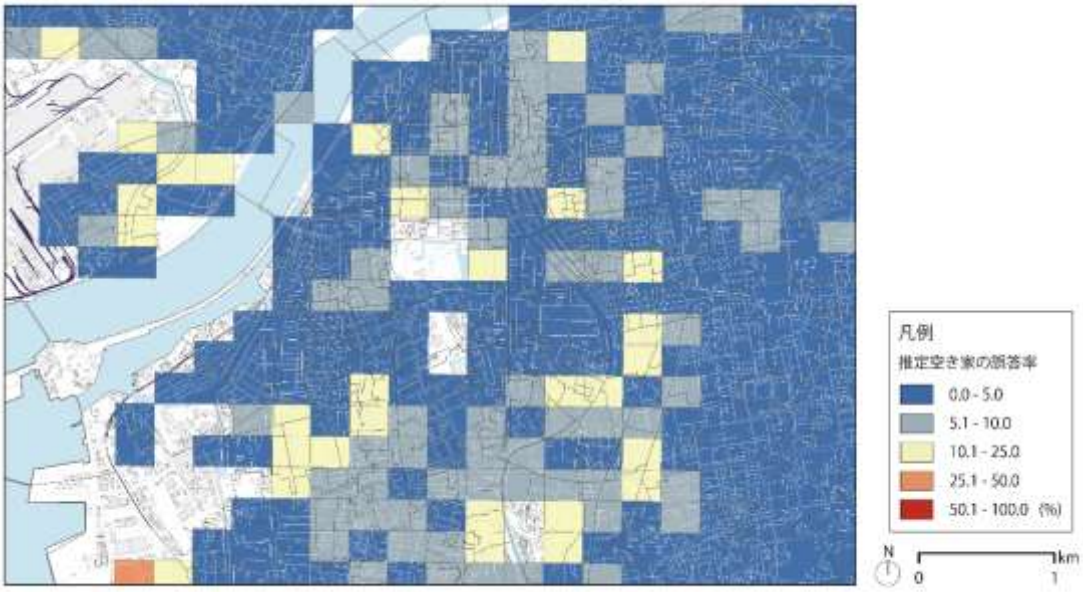


図 10　250m メッシュ別中心市街地内推定空き家の誤答率

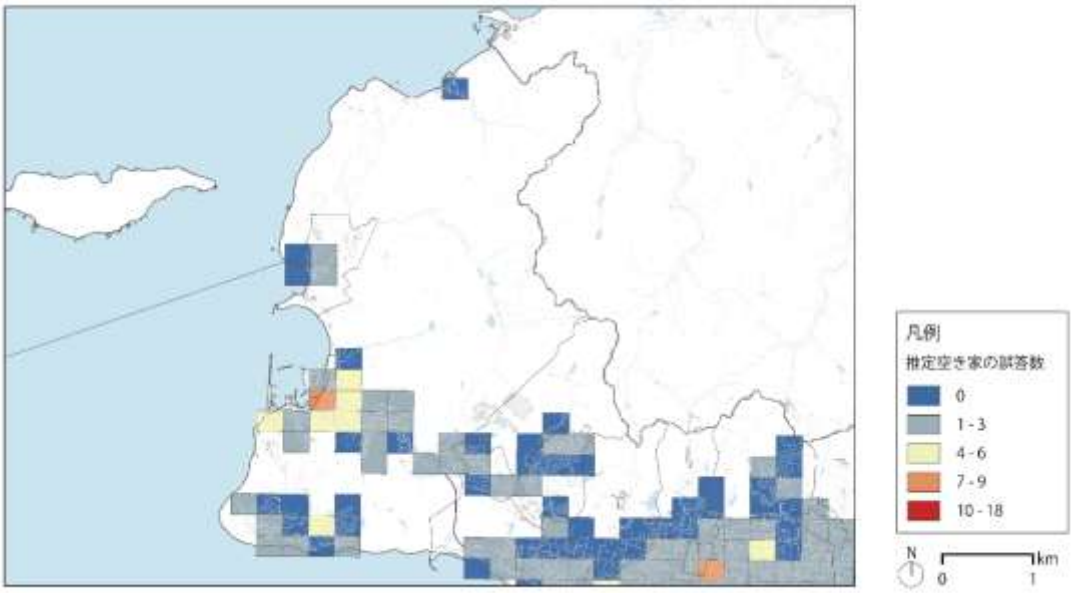


図 11　250m メッシュ別加太地区推定空き家の誤答数

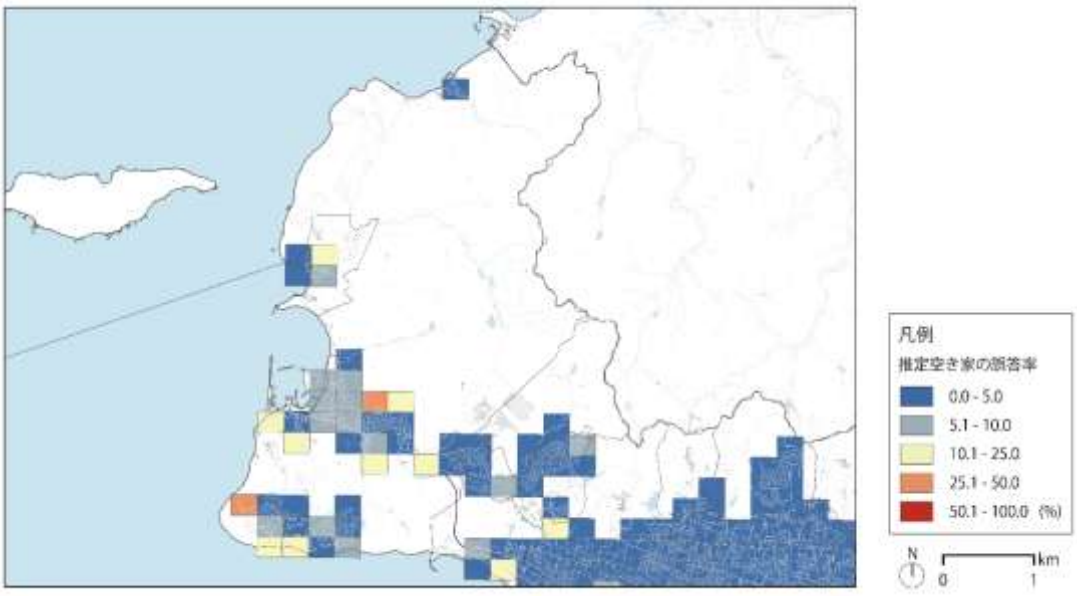


図 12　250m メッシュ別加太地区推定空き家の誤答率

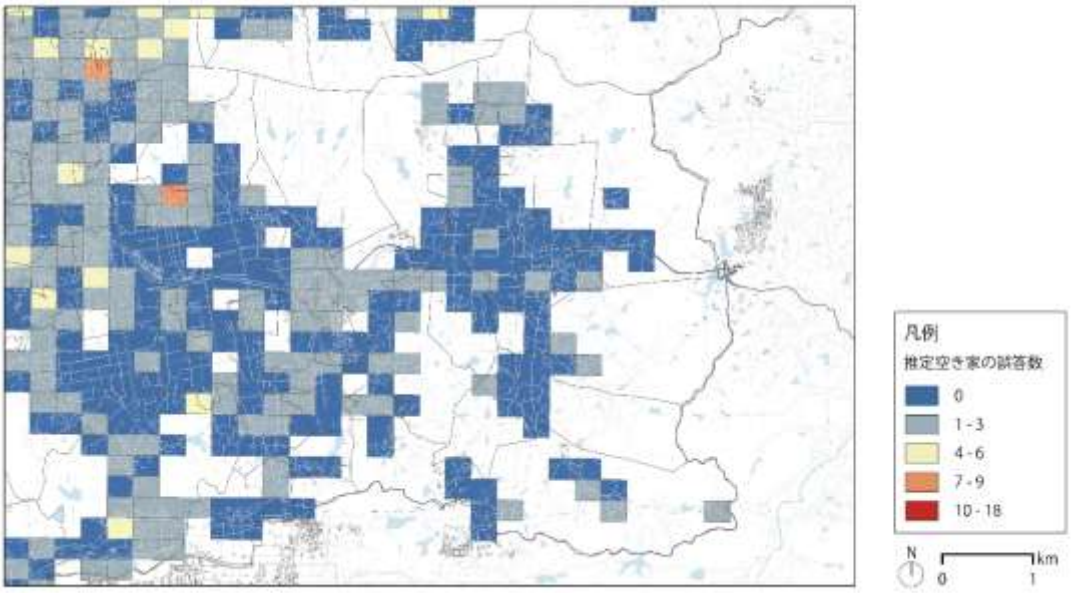


図 13　250m メッシュ別山東地区推定空き家の誤答数

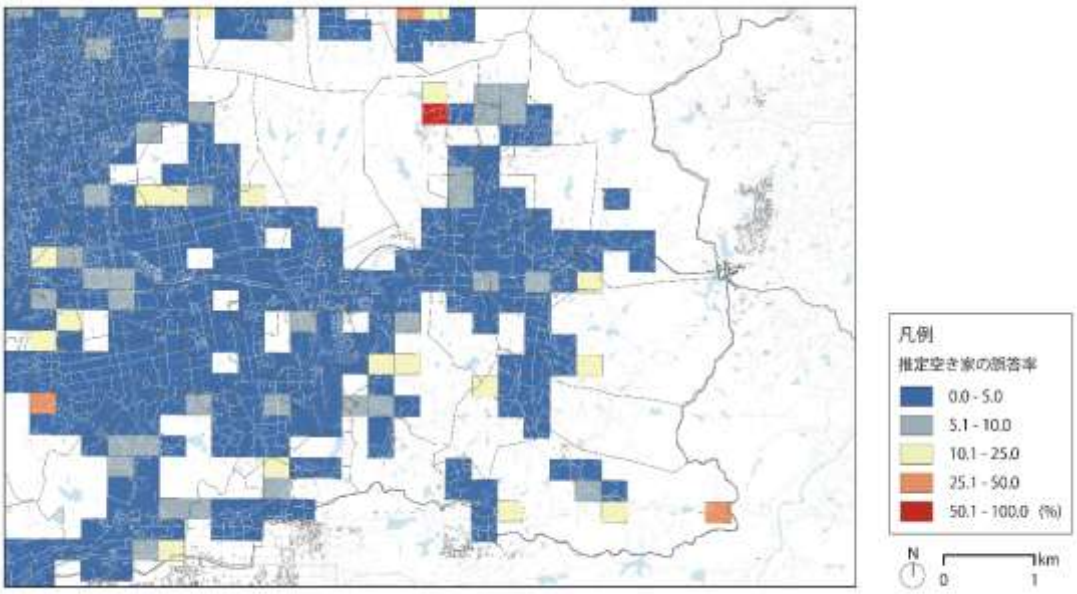


図 14　250m メッシュ別山東地区推定空き家の誤答率

5. まとめと今後の検討事項

5.1. まとめ

本報告書では和歌山市の保有するデータ（住基、建物登記、水道情報）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するようなモデルの構築が可能であることがわかった。以下、自治体保有データの利用、前処理、モデル構築、検証において得られた知見である。

第一に、分析結果から勾配ブースティング木の一種である XGBoost の有用性が明らかになった。当該手法は決定木ベースで欠損値に対応できるため、住基、建物登記、水道情報など複数の公共データを掛け合わせた際、いずれかのデータしか建物に吸着しないようなケースも扱える。これは、利用可能なデータを増やすだけでなく、欠損値の偏りによる推定結果のバイアスを是正することにもつながる。

第二に、検証データの結果から、正答率は 95.4% (25,279/26,509)と良好な数値が得られた。さらに、現地調査で空き家と判定されたものを基準とする真陽性率は 77.0% (1,044/1,356)と改善の余地があるものの、現地調査で居住中（非空き家）と判定された偽陽性率は 3.7% (918/25,153)となり、充分に高い精度が保証された。この結果から、少なくとも居住中の建物を判別する際には誤差率 5%以下で判別可能であるということが示された。

第三に、推定空き家の誤答数と誤答率について、その地理的分布を観察したところ、中心市街地でその値が大きくなっていることがわかった。これは、中心市街地での空き家の形成過程が複雑であるため、十分な訓練データがあっても一定程度の誤答が生じてしまうと推察される。さらに、山東地区や加太地区の一部は、誤答数は少なかったものの、誤答率は高い値となった。これは、山東地区や加太地区では対象建物件数が少ないため、メッシュ間でのばらつきが大きくなってしまい、誤答率の高いメッシュや低いメッシュが混在する結果となったと考えらえる。

5.2. 精度検証を踏まえた改善点の検討

以上のように、本研究で構築した基本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と考えられる。ただし、本モデルは自治体での活用という社会実装を行う上では更なる改善と検討が必要である。例えば、以下の点が挙げられる。

・正解データを少なくしてもどの程度の精度を担保できるか。
・特定地域のみに絞って正解データを利用した場合の、他地域への外挿はどの程度可能か。
・公的統計を利用することでの特徴量をどの程度充実させることができるか。

まず、一つ目については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。例えば正答率を 7〜8 割に担保するためには、どの程度のサンプル数が必要かを判断する必要がある。このような閾値を確定することにより、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。

続いて二つ目では、特定の町丁目を検証データとして抽出し、残りの町丁目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して分析を進めている。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。

最後に三つ目では、次節で詳述する公的統計データを追加して、空き家推定の精度がどの程度変わるかを検討する。ただし、これはデータ構造が階層的であるため、その取り扱いに注意する必要があり、データの前処理の方法が重要になると考えられる。さらに、公的統計のみで分析を行うことで、住基や建物登記情報、水道情報などを容易に得られないような自治体においても簡易的に空き家分布推定を行うことができるため、他の自治体への展開に際して重要な変数になるものと考えられる。基本モデルを用いた推定では住基データ、建物登記データ、水道情報データのいずれかが利用可能であれば、空き家分布推計が可能である。他の自治体においても、それらを利用して空き家推定を行えばよいが、現実的にはこれらの自治体保有のデータを利用するためには、個人情報保護審査会の審査を通過する必要があり、その審査は 1 年、あるいはそれ以上の時間を要する場合がある。その上、個人情報保護審査会を経ても、これらのデータが利用可能になるとも限らない。そこで、公的統計のみを用いた手法を検討・開発するという展開が考えられる。その場合も、現地調査を実施することで、公的統計のみから推定した結果がどの程度の精度を担保できるかを検証可能である。

5.3. 公的統計データを組み合わせた活用方法の検討

本報告書の執筆時点で、国勢調査（平成 17、22、27 年）と住宅・土地統計調査（平成 15、20、25 年）の調査票情報が利用可能となっている。これらの公的統計を活用して、来年度以降は 2 つの分析を行うことを予定している。

第 1 に、本研究で作成した和歌山市の公共データを用いた空き家分布の推定モデルに、以上 2 種類の公的統計から作成した変数（特徴量）を追加し、モデルの推定精度を向上させることを検討する。2 つの公的統計には、和歌山市の公共データだけでは捉えられない変数（居住期間、人口、高齢化率等）が数多く含まれている。これらを、国勢調査では調査区ごと（あるいは基本単位区ごと）に、住宅・土地統計調査では調査単位区[3]ごとに集計し

[3] 調査区、基本単位区、調査単位区は、国勢調査または住宅・土地統計調査における地域区分である。国勢調査では、市区町村を細分化した最小の地域単位として、20～30 世帯から成る基本単位区が設定されており、2 つ以上の基本単位区を組み合わせたものを調査区という。また、住宅・土地統計調査では以下の手順で調査単位区を設定している。(1) 直近の国勢調査から約 20 万の調査区を抽出する、(2) 抽出された調査区のうち、70 住戸を超える調査区については分割して単位区を設定し、70 住戸以下の調査区について

た値を新たな変数として、和歌山市の公共データと組み合わせ、空き家分布の推定精度がどの程度向上するのかを検証する。また、本研究の分析を和歌山市以外の自治体で行うには、当該自治体の公共データが必要となるが、自治体の中には今回の分析で用いた公共データの変数の一部を調査していない自治体も存在するものと考えられる。このような場合、公共データだけを用いた分析では十分な推定精度が確保できない可能性があるため、公的統計によって推定精度を底上げすることが望ましい。

第 2 に、自治体の公共データを用いることなく、公的統計を用いた推定手法により、どこまでの推定精度を確保できる手法を開発できるかを検討する。本研究で作成した和歌山市に対応したデータの整形プログラムは、そのまま他の自治体に適用できるものではないものと考えられる。そのため、今回の分析を他の自治体で行う際には、分析前のデータ整形が大きな障害となるものと予想される。自治体内部でのデータ整形が困難な場合、当該作業を外部に委託することになるものと考えられるが、その場合、個人情報保護審査会での審査に多大な時間を要することになるため、分析自体を断念する恐れがある。そのため、公的統計だけから作成したモデルでどこまで精度の高い推計ができるか検証する必要がある。推定精度の検証は、まずは和歌山市の保有する空き家実態調査のデータと照らし合わせる形で行う。そして、一定の精度を確保できることが分かった段階で、今度は県内の他の自治体にモデルを当てはめることを検討する。その際の精度検証は、自治体が空き家調査を行っていればその結果と照合し、そうでなければ現地調査により実施する予定である。

5.4. 総括

以上のように、本研究を通して和歌山市が保有する各種公共データ（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、同モデルに公的統計データを組み合わせて活用する方法、あるいは公的データのみから空き家率を推定する方法についての可能性についての検討も実施した。

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。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活用したモデルを構築したりする、などの展開が挙げられる。また今年度は検討だけに留まった公的統計データの具体的な活用についても実施していきたいと考えている。

は調査区をそのまま単位区とする、(3) 設定された単位区から調査単位区を抽出する。

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.
- 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, in press, 2019.
- Chen, T., and Guestrin, C., “Xgboost: A scalable tree boosting system”, In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794), 2016.
- 浅見泰司,「都市の空閑地・空き家を考える」, プログレス, 2014.
- 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.
- 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 の推定値の算出方法

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。

令和元年度

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 2 年 3 月

東京大学空間情報科学研究センター　　助教　秋山 祐樹
〒277-8568　千葉県柏市柏の葉 5-1-5　総合研究棟 4 階 404 号室

東京大学空間情報科学研究センター　　特任研究員　馬場 弘樹
〒277-8568　千葉県柏市柏の葉 5-1-5　総合研究棟 4 階 451 号室

和歌山県データ利活用推進センター　　主事　徳冨 智哉
〒640-8203　和歌山県和歌山市東蔵前丁 3 番 1 7　南海和歌山市駅ビル 5 階