The 35th Annual Conference of the Japanese Society for Artificial Intelligence, 2021

機械学習と地理空間情報を活用した周辺の過去の賃料情報を用いたアパート賃料推定

Apartment rent prediction using neighborhood rent information based on machine learning and geospatial information

鶴田大 *1　豊島裕樹 *1
Masaru Tsuruta　Yuki Toyoshima

*1株式会社新生銀行
Shinsei Bank Ltd.

The outstanding balance of apartment loans over the medium to long term has been increasing at financial institutions. Financial institutions are faced with the challenge of properly assessing the profitability of apartments and managing the risk. In assessing profitability, it is necessary to accurately predict apartment rents. In this study, we show that the accuracy of rent prediction is improved by using latitude and longitude information and address information, and by using past rents of similar properties in the neighborhood with machine learning model. In addition, we use latitude and longitude information to incorporate geospatial information system(GIS) data, such as information on disasters to which each property belongs, and verify whether this contributes to improving the accuracy of rent predictions. Including the variables of past rents in the neighborhood to the model, we show that information on disasters do not improve the accuracy of rent prediction.

1. はじめに

個人のアパートやマンションのローンは, 借り手の相続税対策等を目的に, 特に地域銀行で増加してきている. [金融庁 17] では, 築年数の経過にともないアパート収支のみで返済資金を賄えない借り手が増える傾向にあることから, 金融機関は金利上昇や空室・賃料低下等のリスクを適切に評価し, わかりやすく伝える必要があるとされている. アパートの収支の評価においては, 空室や賃料等のいくつかの要素を検討する必要がある. 本研究では, アパート収支の評価時点にける足元の賃料を適切に評価する手法として機械学習を用い, 物件属性情報に加え, 緯度経度や住所に紐づく情報を活用し精度が改善することを示す. 緯度経度などの情報の活用においては, これらの情報を基準に近隣の同種物件の過去賃料を算出し, 説明変数として用いる. また, 緯度経度情報を活用し取得できる情報としてハザードマップのデータがある. 各物件が属するハザードマップの地理空間情報（GIS）データの紐づけを行い, 賃料の予測精度改善に貢献するか検証を行う.

2. 先行研究

家賃や中古マンション価格の推定に関する論文は, 従来多くの研究が行われている. 重回帰モデルが用いられているものとしては, [西 18] や [宗 18] などが挙げられる.

[清水 20] は, 戸建て住宅の物件価格の予測に線形回帰モデルや MLP(Multi-Layer Perceptron), ランダムフォレスト, 勾配ブースティング木を用いた比較を行っており, MLP の誤差の分布が広いことや, ハイパーパラメータの調整により勾配ブースティング木の予測精度が改善することを示している. また, 機械学習と緯度経度を用いた不動産の研究として, [李 20] はマンション価格の予測において, 建物や最寄駅の緯度経度の値を勾配ブースティング木の説明変数として活用することでモデルの予測精度が向上することを示している.

連絡先: 鶴田大, 株式会社新生銀行, 東京都千代田区外神田 3-12-8 住友不動産秋葉原ビル 11 階, TEL:080-7118-3682, Mail:Masaru.Tsuruta@Shinseibank.com

災害情報と不動産価格に関する研究は, [山形 10] や [佐藤 16] が挙げられる. [山形 10] は, マンション価格と災害リスクの関係について, 特に空間計量経済学のアプローチで空間自己相関を考慮した場合と, 考慮しない場合の比較を行い, 空間自己相関を考慮することで災害リスクなどの符号が変化することを示し, 空間自己相関を考慮することの重要性に言及している. 空間自己相関を考慮することで, 災害情報の符号が直感に合う結果となることを示している. これに対し本研究では, 賃料予測の問題において空間的な依存関係を緯度経度に基づくいくつかの変数で捉えている. [佐藤 16] は, 災害リスク情報について, 公示地価, 宅地取引価格, 賃料といった 3 つ不動産市場との関係を分析し, 災害リスクが不動産価格に反映されるには一定の閾値があることを示している. ただし, データの制約から市区町村単位の分析となっている.

3. 手法

3.1 データ

本研究では, 株式会社 LIFULL より取得した LIFULL HOME'S に掲載されている賃貸アパート物件の掲載情報を対象に分析を行う. 対象となる都道府県は, 全国の 47 都道府県である. 分析にあたってマンスリー賃貸および家具家電付き、定期借家権、分譲賃貸の物件を除外した. アパートを対象とするため, 建物階数が 3 階以下, 建物構造が木造と軽量鉄骨および情報を入力した不動産会社が本物件をアパートとみなす物件を対象としている. また, 賃料が 30 万円以下, 平米単価が 7000 円以下, 部屋面積が 15m^2 以上, 80m^2 以下の物件を対象とする. データの期間は, 2009 年 11 月から 2019 年 12 月までである.

3.2 基本変数

掲載情報には, 部屋面積や間取りなどの物件の設備情報に加え, 住所や緯度経度といった近隣・周辺の特性, さらに最寄駅等の交通利便性にかかわる変数が入っている. 設備の情報としては, 面積, 構造, 築年月, 部屋の階数, 建物の階数といった情報が含まれる. さらに, バス・トイレ別やオートロック, TV モニター付きインターホンといった掲載情報に含まれる住宅内

のより詳細な特性を 0-1 変数として取り込む. 間取りの変数化については, 後述の緯度経度情報を活用した変数で述べる. 近隣・周辺の特性としては, 住所と緯度経度に加え, 小学校やコンビニ, スーパーからの距離などが含まれる. さらに, 住所情報を正規表現で分割を行うことで都道府県名や市区町村名を取得し変数として用いる. 交通利便性にかかわる変数としては, 最寄り駅, 最寄り路線, 最寄駅からの距離などが含まれる.

3.3 緯度経度情報を活用した変数

本研究では, 緯度経度情報を活用し, 各物件の周辺過去賃料の情報やハザードマップ, 公示地価の情報を活用する.

3.3.1 メッシュ単位の周辺賃料

緯度経度情報を活用した周辺賃料としては具体的に, 500m メッシュ, 1km メッシュ, 5km メッシュ, 10km メッシュの過去 1 年, 5 年, 全期間の賃料の中央値を用いる. メッシュについては, 全国で緯度経度情報を元に分割し, 各メッシュにナンバリングを行うことで識別を行っている. メッシュに加え, さらに専有面積や間取り, 築年数といった賃料への影響が大きい情報が類似する物件の賃料を参照するため, 元々カテゴリー変数である間取りに加え専有面積を $5m^2$ 単位, $10m^2$ 単位のカテゴリ変数化, 築年数を 5 年, 10 年単位のカテゴリ変数化を行う. これにより, [メッシュ]×[専有面積]×[築年月] といった同じメッシュ内の類似カテゴリ内の物件の過去賃料の中央値を, 評価対象物件の説明変数として用いる. 例えば 1 つの周辺賃料の計算例を挙げると, メッシュ ID が 500 の物件で専有面積 $18m^2$, 築年数 7 年であれば, 同じメッシュ ID が 500 内の専有面積が 15-$20m^2$, 築年数が 5-10 年の住戸の過去 1 年賃料の中央値を算出し, 対象物件の説明変数とする. さらに, 緯度経度情報を用いずとも, 市区町村や最寄り駅などが同一の物件の過去賃料を算出することができるため, これらの単位を基にした過去賃料の中央値も比較対象の変数として活用する. なお, 評価物件の過去 1 年の周辺賃料の計算頻度は四半期である. 例えば, 2019 年第 1 四半期の物件に対し, 2018 年第 1 四半期から 2018 年第 4 四半期の過去データを用い, 周辺賃料の算出を行う. なお, 過去賃料は 1 年, 5 年, 全期間を対象に算出する.

3.3.2 k 近傍法による各物件の周辺賃料

次に, 位置の近い物件の賃料情報を活用するため, kd-tree を用いた k 近傍法により効率的に各評価対象物件の周辺の N 件の賃料情報を平均し変数化する. メッシュ単位の周辺賃料と同様に, 築年数や面積に基づき学習データをいくつかのカテゴリに分け, カテゴリごとに緯度経度情報などを説明変数, 過去賃料を目的変数とした k 近傍法の学習を行う. 例えば, 2019 年第 1 四半期の物件に対し周辺過去賃料を計算するには, 2018 年第 1 四半期から 2018 年第 4 四半期の過去データを用い, 目的変数を賃料, 説明変数に緯度経度もしくは緯度経度と築年数、面積を用いた k 近傍法の学習を行う. 総当たりによる周辺 N 件の探索は, 計算量が増大となるため, kd-tree により近傍を探索する. この学習した k 近傍法のモデルに対し, 2019 年第 1 四半期の評価対象物件の緯度経度などの情報に基づき, k 近傍法の予測値を計算する. この予測値は, 緯度経度情報などから計算される近傍の過去 1 年間の平均賃料である. 厳密には, 緯度と経度でそれぞれ同じ値の間隔でも距離の単位が異なるが, ここでは簡便的に同じ尺度とみなし, 緯度経度の数字をそのまま用いる. また, 近傍 N 件の物件に対しては, 緯度経度から計算されるユークリッド距離の逆数で重みづけを行うことで, より遠いデータの影響を減らしている. 近傍の数は, 20 件と 30 件の 2 通りの計算を行い, 算出にあたっては, 過去の賃料を 1 年, 2 年, 4 年の範囲で計算している. メッシュ単位の周辺賃料と異なり, 学習データが多くなることで, kd-tree を用いても k 近傍法の計算時間を要するため, 相対的に短い期間での算出となっている.

3.3.3 災害情報と公示地価

洪水の浸水, 津波の浸水, 土砂災害の警戒区域といったハザードマップの情報を国土交通省国土数値情報から取得し, 緯度経度情報を基準に物件に紐づけを行う. 洪水情報については, 平成 24 年の計画規模を用いる [*1]. 地域によっては複数河川の洪水の浸水域にあたる場合があるため, 物件に紐づいた浸水域のうち, 最大のものを変数として用いる. 土砂災害については, 土砂災害警戒区域の急傾斜地の崩壊, 土石流, 地滑りの 3 つの現象があるが, 場所によっては複数の現象が紐づく場合もあるため, いずれかの警戒区域に該当するかどうかの 0-1 変数としている. 津波については, 1m～2m といったように範囲での表示され, 各都道府県で公表している範囲が統一されていないため, 各範囲の上限と下限の平均値を変数とする [*2]. 公示地価も, 国土交通省国土数値情報から取得し, 緯度経度情報が含まれているため, 各物件に対し最近傍の公示地価を紐づけ, 説明変数として用いる.

3.4 分析単位

掲載情報は月次単位で取得しているため, 例えばある部屋で 2019 年 1 月, 2 月, 3 月と掲載があった際には 3 つのデータが存在する. この場合, この連続した 3 か月は, 1 つの空室期間であると考えられる. 実際に契約したと考えられる 3 月の掲載情報の賃料のみを予測するといったように, 入居が開始したと考えられる各一続きの掲載期間の最後の月の賃料が 1 つのサンプルとなるような調整を行う. これにより, 本研究の対象物件は 748,378 件, 部屋数は 2,297,929 戸, 述べサンプルサイズは, 3,670,349 件である.

3.5 機械学習モデル

本研究では, 機械学習手法の一種である勾配ブースティング木を用いる. 決定木にブースティングと呼ばれるアンサンブルモデルのアプローチを適用した手法である. ブースティングは, すでに学習済みの予測モデルに対し, 正しく分類できていない学習データを正しく分類できるよう新たな予測モデルを追加・構築していく手法である. ライブラリには LightGBM([Ke 17]) を使用する. 目的変数は対数賃料として, 目的変数を huber 関数とする. 主なハイパーパラメータとしては, 学習率を 0.1, 葉の数の最大値（num_leaves）を 13, 1 つの葉のノードの最小サンプル数 (min_data_in_leaf) を 3 としている.

3.6 検証方法

本研究では, 都道府県別にモデルを構築し, 各変数を追加することによる精度の改善効果を確認する. 2013～18 年の 5 年間の掲載情報を元にモデルを学習し, 将来時点の 2019 年のデータで精度を確認する. 学習データと検証データで同じ物件を含むことにより精度が過度に良くなる可能性があるため, 2018 年までに掲載がある物件について学習データと検証データで重複しないように 80%:20%に分割し, 80%のデータで学習を行う. 残りの 20%の物件の将来 2019 年時点のデータと, 2019 年から掲載が開始された物件を対象に検証を行う. 学習を行う際, 学習データ内で 5-fold に分割し, 4/5 のデータで学習を行い, 残りの 1/5 の検証データで early termination を行うこと

*1 都道府県によって, 5 段階と 7 段階のものがあるため, 同じ基準で影響を確認するため, 5 段階に読み替えを行っている.

*2 国土交通省国土数値情報では, 宮城県や愛知県といった海岸に面した都道府県でも津波の情報が無い場合がある. 本研究では取得可能な都道府県の情報のみ使用する.

を 5 回繰り返し, 学習された 5 つのモデルの平均値を最終的な推定結果として用いる. この際, 物件が重複しないような分割を行っている.

変数の組み合わせとしては, 基本変数で挙げた設備情報, 近隣・周辺の特性, 交通利便性にかかわる変数をすべての検証時に使用し, これらの基本変数に各提案変数を考慮することで精度がどのように変化するかを確認する. 周辺の過去賃料を用いない場合の比較対象として, 市区町村や最寄駅名の 0-1 ダミーを地域性として捉え, 緯度経度の値をそのままモデルの変数に加えた場合や, 市区町村や最寄り駅を基に算出した過去賃料と緯度経度の値の組み合わせの場合の学習を行う.

4. 結果

4.1 精度

表 1 に主な都道府県の MER(誤差率中央値) と全国平均の MER の結果を示す. まず, 緯度経度の現数値の追加のみの場合でも MER が低下する. 公示地価の追加によっても, 低下幅は大きくないが MER は低下する. MER の低下幅が大きいものとしては, 市区町村等の過去賃料, メッシュ単位の過去賃料と k 近傍法による変数を加えた場合である. これら 3 つの変数により MER が低下している [*3]. 災害情報に基づく変数は, 市区町村を 0-1 変数でとらえている場合, 0.1%程度 MER が低下するが, 緯度経度に基づく変数を加えると災害情報追加による MER の低下幅は減少し, ほぼ改善が無い結果となる. 表 1 に示したモデルの番号に基づきモデル 1 とモデル 9 の場合における変数の効果について, 次節で詳細な確認を行う.

4.2 変数の重要度

図 1 のモデル 9 で SHAP([Lundberg 17]) を用い計測した東京都の変数重要度の結果を図 2 に示す. それによれば k 近傍法により捉えた周辺賃料が相対的に高い結果となっている. k 近傍法の変数以外には, 面積, 部屋階数, 公示地価やバストイレ別といった過去周辺賃料では捉えられない変数が上位に位置する. さらに, 市区町村別の過去賃料は, 重要度が上位であるが, メッシュ単位の過去賃料は上位ではない結果となっている.

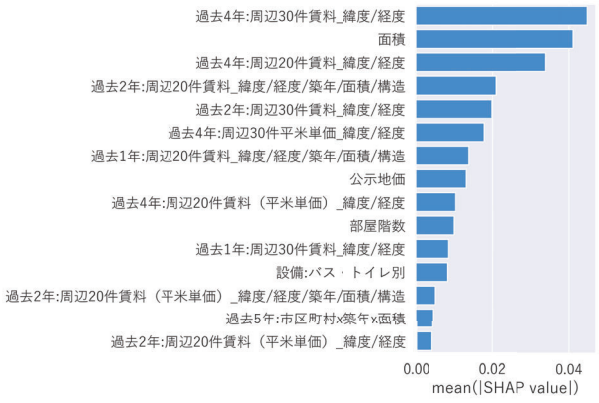

図 1: 東京都の SHAP による変数重要度（上位 15 変数）

4.3 SHAP による災害情報の追加効果の確認

災害情報は MER を低下させる効果が小さい結果となったが, 個別の変数がどのようにモデルの予測賃料に寄与しているか SHAP を用い確認する. 災害情報が比較的 MER の低下に効果のあるモデル 1 と, 最も MER が低いモデル 9 の場合で確認する. 全国の平均的な傾向を確認するため, 災害情報が無い場合の SHAP 値が各都道府県で 0 になるように引き算の調整を行い, この調整後 SHAP 値を各浸水ランクや浸水の深さ毎に全国で平均する. この結果を示したのが図 2 である. モデル 1 の場合, 洪水の浸水ランクが 11 から 13 の場合, SHAP 値が正であり, 浸水が無い場合と比較し賃料へプラスの効果がある結果となった.

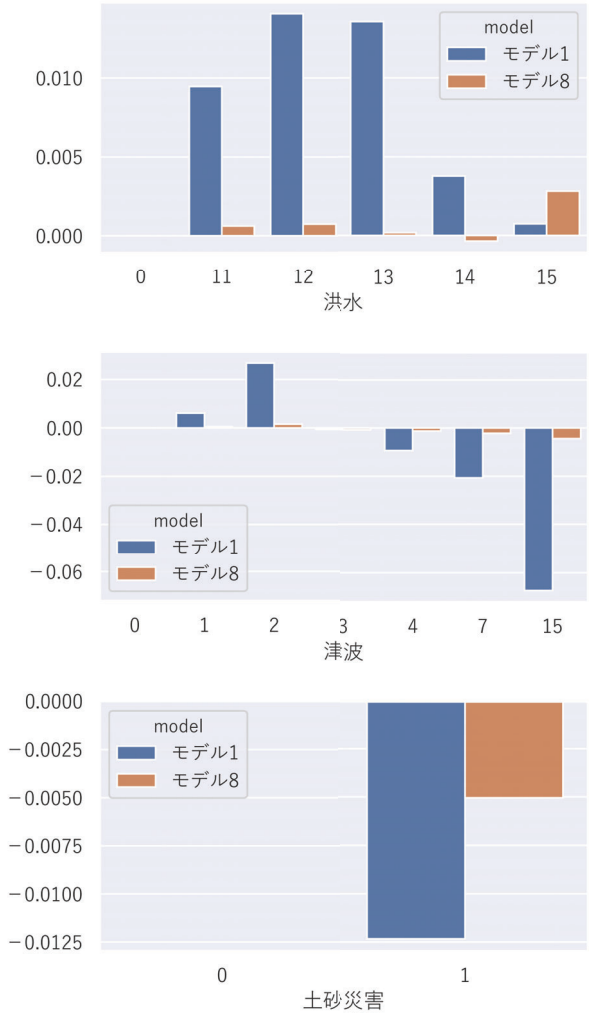

図 2: 洪水の浸水ランク, 津波の浸水深さ, 土砂災害の警戒区域と調整後 **SHAP** 値（全国平均）の関係.

この傾向を個別都道府県で確認するため, 図 3 に物件の多い 1 都 3 県で個別に洪水の浸水域と調整後の SHAP 値の関係を示す. 神奈川県や千葉県で浸水ランクが 11 から 13 の場合, SHAP 値が正であり, 賃料へプラスに寄与する結果となった. 浸水域に該当する地域を確認すると, 例えば神奈川県の場合, 川崎駅や横浜駅の周辺といった利便性が高く賃料が高いエリアが浸水域になっている. また千葉県では, 市川市で同じような影響が確認される. このように, 都道府県によっては, 地域性を市区町村や最寄り駅の 0-1 変数といった広い単位でとらえた場合, 洪水による浸水の賃料への効果は期待されるものと逆になることが生じうる. 一方で, 図 2 において, モデル 9 の SHAP 値は, その絶対値が大きく低下していることが分かる.

*3 本研究では, 周辺の過去賃料算出の対象に, 予測時に入手可能な場合, 評価対象と同じ物件や同じ部屋の過去賃料も含めている. 含めずに検証を行った場合, MER の低下幅は小さくなる.

表 1: **変数別のモデル精度：誤差率中央値.** 市区町村等の過去賃料は市区町村や最寄り駅単位の過去賃料の中央値, 緯度経度の過去賃料はメッシュ単位に基づく過去賃料の中央値, k 近傍は k 近傍法により算出した物件周辺の過去平均賃料をそれぞれ変数として使用.

モデル	市区町村等	緯度経度	公示地価	k 近傍	災害	埼玉県	千葉県	東京都	神奈川県	長野県	静岡県	福岡県	全国平均
1	0-1					0.0772	0.0835	0.0898	0.0886	0.0619	0.0683	0.0621	0.0608
2	0-1				○	0.0758	0.0780	0.0858	0.0856	0.0610	0.0672	0.0619	0.0596
3	0-1	原数値				0.0464	0.0540	0.0492	0.0556	0.0521	0.0525	0.0507	0.0489
4	0-1	原数値	○			0.0457	0.0526	0.0480	0.0544	0.0528	0.0521	0.0501	0.0482
5	0-1	原数値	○		○	0.0456	0.0525	0.0479	0.0544	0.0526	0.0524	0.0496	0.0482
6	過去賃料	原数値	○			0.0432	0.0491	0.0451	0.0522	0.0449	0.0493	0.0421	0.0434
7	過去賃料	過去賃料	○			0.0342	0.0397	0.0392	0.0422	0.0337	0.0414	0.0352	0.0373
8	過去賃料	過去賃料	○	○		0.0297	0.0340	0.0318	0.0363	0.0285	0.0376	0.0298	0.0339
9	過去賃料	過去賃料	○	○	○	0.0294	0.0337	0.0317	0.0364	0.0287	0.0376	0.0300	0.0338

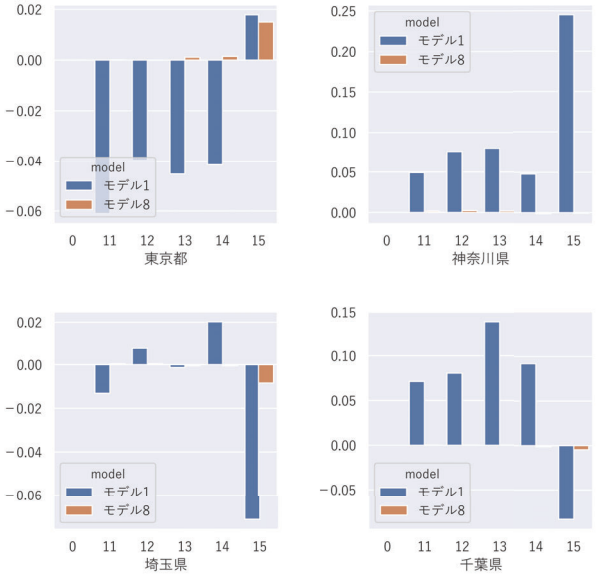

図 3: **1 都 3 県の洪水の浸水域と調整後 SHAP 値の関係.**

図 1 の津波の場合については, 浸水が深くなるほど SHAP 値が低下し, 賃料が低下する効果がモデル 1 でも 9 でも確認される. ただし, モデル 9 ではその効果が大きく減少する結果となった. 図 1 の土砂災害については, 警戒区域で賃料が低下する効果がモデル 1 でも 9 でも確認される. モデル 9 の SHAP 値の絶対値が低下するものの影響は残る結果となっている.

5. 結論

本研究では, アパート賃料の予測において, 地理空間情報を活用した周辺の過去の賃料情報をメッシュ単位や k 近傍法を用い捉えることで, 全国で予測精度が向上することを示した. 特に k 近傍法を用い捉えた対象物件の周辺過去賃料がモデル上重要な変数となった. 一方で, 地理空間情報に基づき, 災害情報を物件に紐づけることで, 物件ごとの災害情報に基づく賃料への影響を確認し, 予測精度が改善するかを検証したが, 災害情報の予測精度の改善効果は低い結果となった. 特に, 周辺の過去賃料情報を緯度経度情報に基づき取り込んだうえでモデル精度を確認すると, 改善効果は小さくなり, 過去の周辺賃料情報に既に災害情報が考慮されてしまい捉えられている可能性がある. SHAP を用いた確認で, 土砂災害情報は周辺の過去賃料情報を緯度経度情報に基づき取り込んだうえでも賃料に与える効果が残っているが, 土砂災害は紐づく物件の割合が低いため, モデルの全体の精度改善に与える効果は少ない結果となった.

謝辞

株式会社 LIFULL には, LIFULL HOME'S のデータの分析結果の発表について許諾を頂いたことに謝意を表する. また, 本稿の執筆にあたり, 株式会社新生銀行の嶋田康史氏にはご助言をいただいた. ここに謝意を表する.

参考文献

[金融庁 17] 金融庁: 平成 28 事務年度 金融レポートの主なポイント, 2017.

[佐藤 16] 佐藤慶一, 松浦広明, 田中陽三, 永松伸吾, 大井昌弘, 大原美保, 廣井悠: 災害リスク情報と不動産市場のヘドニック分析, ESRI Discussion Paper Series, 2016.

[清水 20] 清水千弘: 不動産テック (FinTech ライブラリー), 朝倉書店, 2020.

[宗 18] 宗健, 新井優太: 富裕層及び団地の集積が家賃に与える影響, 都市住宅学, 103, 126-131, 2018.

[西 18] 西颯人, 浅見泰司, 清水千弘: 住宅設備と賃料, CSIS Discussion Paper No.153, 2018.

[山形 10] 山形与志樹, 村上大輔, 瀬谷創, 堤盛人: 環境・災害リスク指標とマンション価格のマルチレベルモデルによる空間計量経済分析, 土木計画学研究・講演集, 43, 2010.

[李 20] 李天琦, 秋山卓也, 田坂祐太: LightGBM を用いた不動産投資における高精度価格推定モデルの構築, 人工知能学会全国大会論文集, 34, 2020.

[Ke 17] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Liu, T. Y.: Lightgbm: A highly efficient gradient boosting decision tree, Advances in neural information processing systems, 30, 3146-3154, 2017.

[Lundberg 17] Lundberg, S., Lee, S. I. : A unified approach to interpreting model predictions, 31st Conference on Neural Information Processing Systems, 2017.