文章ベクトルを用いた機械学習による地価についての予測

中川 俊成†　　大槻 明‡

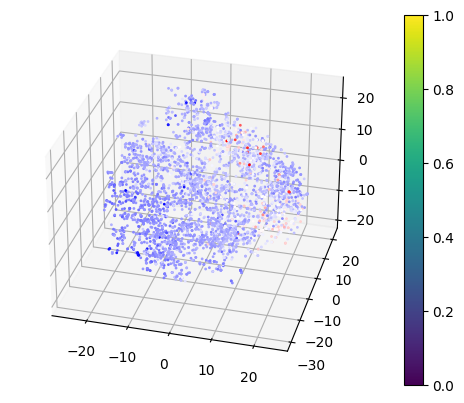
日本大学経済学部

**1.はじめに**

2019年以降国土交通省の土地総合情報システム上で公示地価の鑑定評価書のダウンロードが可能になって以来、大森・熊越[1]のように鑑定評価書のテキストマイニングを地価分析に応用した研究が登場し始めている。しかし鑑定評価書の文章データを特徴量とし、数的処理を行っているものは少ない。本研究ではオルタナティブデータである文章データを、文脈を考慮するSentence-BERTを用いてベクトル化し、駅との距離や地域経済的な量的変数と組み合わせ、変動率を教師データとする人手によらないラベル付与を経て、機械学習による法則性の発見の対象とすることで、従来の構造化データのみの分析では捉えることが難しい不動産への期待形成の実際などをモデルに組み込むことを目指し、予測精度と地理的な汎化性能の向上を図る。

**2.検証アプローチ**

図1 「地域要因の将来予測」3次元特徴量

大都市圏の地価変動に焦点を当て、予測を行う。

訓練・検証データは2019年1月1日時点(地価公示は毎年1月1日を基準にしている。)の東京都全域の公示地価鑑定評価書で地点数は宅地見込地、林地を除いた2555地点、鑑定評価数は5110サンプル(公示地価は1地点につき2つの鑑定評価を持つため)。検証データは20%でホールドアウト法を用いる。テストデータは、東京都のデータのままでは評価地点の地理的な要因や鑑定評価テキストが大きく変化しないため、大阪府全域の2022年1月1日時点の宅地見込地、林地を除いた1711地点、

3422サンプルの鑑定評価書を用いる。コロナの影響が強い2020年、2021年の鑑定評価書はそのイレギュラーさゆえ除外した。公示地価の当期から来期にかけての変動率、（つまり2019年データであれば、2020年、2022年データであれば2023年にかけての変動率）を目的変数とする。データの入手における指針として、当期以前に入手できるデータに限った。例えば、後述する市区町村別生産年齢人口については2020年調査分データを変数としたいところだが、それが公開されるのが2022年7月であり、2022年3月の地価公示より後のため、使用しない。このように、実際的な予測のため、情報公開のラグを考慮した。市区町村単位での説明変数は課税対象所得の変動率と市区町村別の1㎡あたり生産年齢人口とする。上記の指針のもと市区町村別の課税対象所得についての統計は2016～2017年、2019～2020年調査分のものを用いた。市区町村別の1㎡あたり生産年齢人口においても上記の指針のもと5年おきに行われる国勢調査の2020年分でなく2015年調査分の人口統計を用い、その市区町村の面積で除した。どちらもe-Statより取得。評価地点単位においては最寄り駅との距離、都道府県で最も乗降人数の多い駅との距離の二つと、本研究の特徴である、ベクトル化されたテキストも変数とする。ここでの文章データは鑑定評価書の「地域要因の将来予測」・「市場の特性」・「一般的要因」・「地域的要因」の4つの欄のテキストを使用する。ベクトル化の手法については、まずSentence-BERTでそれぞれのテキストを768次元のベクトルにする。園部勲氏が作成しHugging FaceとQiitaで情報公開されている日本語用Sentence-BERTモデル[2]を用いた。バッチサイズは512でベクトル化を行う。だがこのままでは次元数が過大になるため、t-SNEによって3次元に縮約する。例として「地域要因の将来予測」欄のテキストをプロットする(図1)。色は、目的変数である来期の変動率を表し、赤くなるほど来期の上昇率が高く、青いほど下降率が高くなる。右奥に赤い点が集中しており、左手前になるほど濃い青の点が多くなっているなどクラスターを形成していると見受けられるが、右奥にある赤い点には、減少をあらわす青い点も重なっていて、特徴量としての懸念点ともいえる。文章データは計4(欄)×3(次元)で12個の特徴量、説明変数となり、上記の量的変数と合わせて16個の変数で回帰タスクを行う。分析手法としてはXGBoost、LightGBMといった勾配ブースティング法、またランダムフォレスト法を用いる。それらの損失関数はRMSE、評価指標はMAE、RMSE、決定係数とする。比較検証のためテキストデータを使用する分析（a）とテキストデータを使用しない分析(b)を実施する。

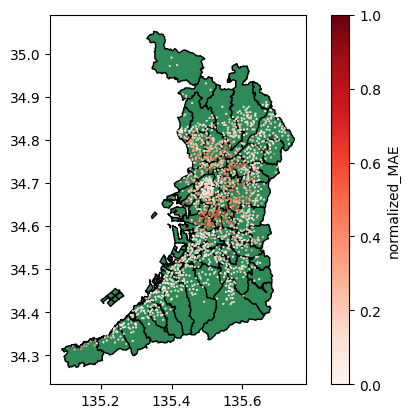
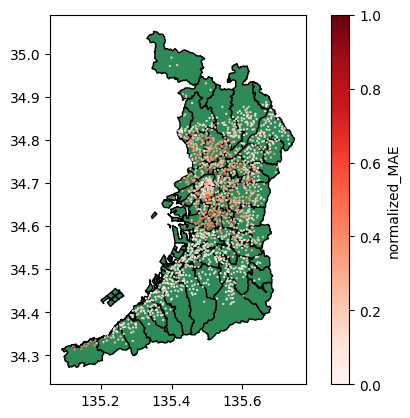
**3.結果**

表1

| 手法 | MAE | RMSE | 決定係数 |
| --- | --- | --- | --- |
| XGBoost(1-a) | 2.8409 | 3.4918 | -2.0304 |
| XGBoost(1-b) | 2.7319 | 3.4244 | -1.9145 |
| LightGBM(2-a) | 2.659 | **3.3137** | **-1.7291** |
| LightGBM(2-b) | 2.5453 | 3.3716 | -1.8253 |
| Random Forest(3-a) | 2.8109 | **3.4852** | **-2.019** |
| Random Forest(3-b) | 2.7688 | 3.5625 | -2.1542 |

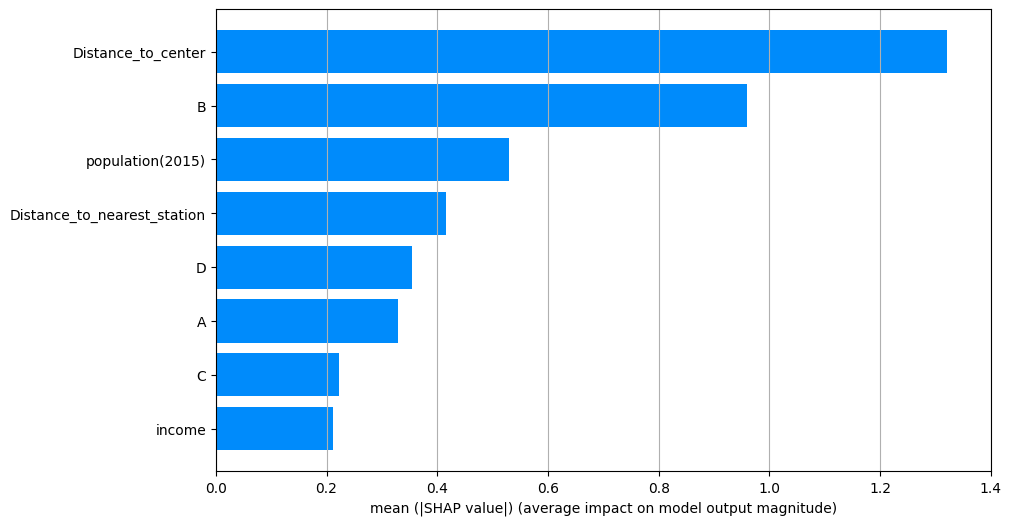
図2

(2-a) (2-b)



まずXGBoostでは全ての指標が(a)の方が悪かった。しかし、LightGBMとランダムフォレストではMAEは劣るもののRMSE、決定係数において(b)を精度において凌いだ。図2はLightGBMにおいて地点ごとのMAEを正規化し、実際の地図にプロットしたものだ。赤いほど、精度が悪くなっている。(b)は(a)を大阪都心部やその北部における精度で凌いでいるものの大阪市南部、東大阪市西部、守口市、門真市にかけての住宅地域、ベッドタウンにおいては精度が著しく悪くなっている。このような(b)の局所的な精度の大幅悪化という点で、(a)は優位に立ち、それがRMSEにも現れているのではないか。図3は変数ごとのSHAP特徴量重要度(LightGBMにおける実行結果)である。Aは「地域要因の将来予測」、Bは「市場の特性」、Cは「一般的要因」、Dは「地域要因」である。それらは別個の変数として扱われている3次元x,y,z成分の重要度を加算して算出した。

図3



「市場の特性」における記述の重要度が高いことが読み取れる。

一方で地点間で文章に差が出にくいであろう「一般的要因」の重要度は4つの中で最も低くなった。また「地域要因の将来予測」と「地域要因」が僅差で並んでいることには、どちらとも地域要因について述べたものであり、その類似性が重要度としても現れているといえる。

**4.おわりに**

マクロ経済の各地域への波及をあらわす特徴量を組み込むないしは、機械学習とは別に、VARなどの自己回帰モデルでマクロ経済変数の予測値を算出しておき、それと非構造化データの機械学習を組合せることも検討しておきたい。訓練・検証データも東京都だけであり、他の都道府県の評価書、そして地域の経済ニュース記事も学習すればより有用なのではないか。また地理的な固有表現が文章の割合の多くを占めることがあり、地理的な汎化性能を押し下げてしまうノイズの一因となった。地価に影響する表現のみを抽出すれば、より精度は向上するだろう。また既存研究では、人手または機械学習による極性表現の獲得によって単語ベースでの特徴量化を行っているものも多い[3][4][5]。比較検証としてそのような手法との比較も行わなければならないだろう。

[参考文献]

[1]大森 文彦, 熊越 祐介,Covid-19感染症流行を受けた東京都市圏のライフスタイル変化に関する考察,都市計画報告集,21巻,2号p.177-182(2022)

[2]https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2

[3]川崎拓海,穴田一,情報科学技術フォーラム講演論文集 (FIT),20th,第1分冊 p.99-100(2021)

[4]瞿雪吟,菅愛子,高橋大志,中国株式市場を対象とした金融極性辞書の構築と検証(2019)

[5]五島 圭一, 新谷 元嗣, 高村 大也, 景気単語極性辞書の構築とその応用,自然言語処理 29 巻,4号, p. 1233-1253 (2022)