2020大会

ジャレフ・ジャーナル 12 (2020) 5−8

非構造化データによる不動産価格分析は有用か？

Is it useful to fit unstructured data to property price analysis?

植杉　大*
Dai Uesugi*

摂南大学経済学部*

本研究のポイントは，ヘドニックモデルによる不動産価格分析において，従来的な数値による構造化データに代わり，テキストデータによる非構造化データを用いている点である．テキストデータをトピックモデルの手法の一つである潜在ディリクレ配分（Latent Dirichlet Allocation）を用いて抽出されたトピックとして数値化し，ヘドニックモデルの説明変数として利用することにより，不動産価格が説明可能かを実験した．その結果，従来的な質的ダミー変数などの構造化データと比較して大きな説明力が観察され，その有用性が示された．

Keywords: トピックモデル（Topic Model），潜在ディリクレ配分（Latent Dirichlet Allocation），非構造化データ（Unstructured Data），ヘドニックモデル（Hedonic Model）

1. はじめに

周知のとおりビッグデータとは，従来的な公的機関が発表する統計のみならず，民間企業や消費者の経済行動をはじめとしたあらゆる行動記録，ソーシャルメディア等によって収集された，処理するにはあまりにも大きすぎるデータの総体と定義される．近年不動産テック企業がこのビッグデータの活用を推進しており，機械学習や AI，その他 VR などのハードを含めた不動産市場への新しいアプローチを行っている．

一方不動産研究，不動産計量分析において，ビッグデータは十分に活用されていない．例えば不動産価格分析において利用される説明変数は，一般に対象不動産の属性に関する数値データ(CBD までの距離･時間，駅までの距離・時間，面積 etc.)である．質的属性はダミー変数に変換して利用することはできるが，情報量としては断片的(鉄道会社ダミー，間取り，土地利用 etc.)である．これらのデータは構造化データ(structured data)という．一方，テキストや画像など，これまでの数値データとは異なるデータの種類を非構造化データ(unstructured data)といい，構造化データと比較して圧倒的に膨大かつ情報量も豊富である．

本研究の目的は，①不動産に係るテキスト等の非構造化データに対して，自然言語処理の統計モデルであるトピックモデルを適用して分析し，②ヘドニックモデルによる不動産価格分析において，テキスト等非構造化データから数値的に抽出されたトピック情報を説明変数として用いた場合，従来の構造化データと比較して，モデルの説明力や予測精度が向上するのかを検討することである．

２．不動産分野におけるテキスト分析

既存の不動産研究に関するテキスト分析は大きく 3 つに分類される．①帰納的内容分析(Inductive Content Analysis)，②センチメント分析(Sentiment Analysis)，③潜在意味解析を用いた分析である．

①は統計的手法ではなく，米国主要ジャーナルの文献サーベイを通じて，内容をいくつかのトピック(例えば，投資，鑑定，公共政策等)に縮約する方法である．この方法は一種「手作業」であって研究者の推論技術に依存しており，あくまでも「できるだけ」客観的に分析したもので主観性を完全に排除できていない(例えば，Jud(1996)，Dombrow and Turnbull(2004)を参照のこと)．