

頻出部分グラフを用いた賃料分析

Rent Analysis using frequent subgraphs

尾崎知伸* 小黒淳斗
Tomonobu Ozaki Atsuto Oguro

日本大学 文理学部
College of Humanities and Sciences, Nihon University

Abstract: In this paper, we apply frequent subgraph mining and regression analysis to investigate the relationship between the room layout and the rent of apartments. We build a graph database having about 550 of floor plans, and extract characteristic substructures appearing frequently in the database. We prepare four types of target variables on the rent and apply different types of regression analyses, i.e., regression tree, model tree, support vector regression and multiple regression analysis, using extracted substructures as independent variables. Through the analysis, we reveal room layouts related to the rent.

1 はじめに

マンション等の賃貸物件の賃料は、専有面積や部屋の向き、間取り（部屋配置）、最寄り駅までの距離、周辺環境など、様々な要因によって決定される。これまでに、これらの要因に着目した賃料分析の結果が数多く報告されている。

文献 [1] では、占有面積や駅までの徒歩距離など、物件の各詳細情報が賃料にどれだけ影響を与えているか、重回帰分析を用いて分析している。また文献 [2] では、物件の情報を、部屋数や占有面積など物件そのものが持つ一般的な情報と、所在地や最寄り駅までの距離など外的要因に強く依存する情報に分類し、それぞれで構築されるモデルを統合することで賃料を推定している。一方、文献 [3] では、外部データとして地震に関する地域危険度調査結果を用い、建物の構造別、新旧の耐震基準別に物件所在地の危険度が賃料に与える影響を分析している。さらに文献 [4] では、物件の詳細情報といった客観的な情報に加え、実際に物件を見たときの感性などの主観情報も利用した分析を行っている。

間取りに着目した分析として、文献 [5, 6] が挙げられる。これらの研究では、間取り図を部屋をノード、そのつながりをエッジとするグラフ（以降、間取り図グラフと呼ぶ）で表現することで、構造的な側面からの分析を試みている。文献 [5] では、複数の間取り図グラフに共通して現れる部分グラフを抽出し、相関ルール分析を用いてそれらの関係を分析している。さらに各頻出部分グラフを含むか否かを属性とした回帰モデル

を構築し、部分グラフ、すなわち部屋配置が賃料に与える影響を考察している。また文献 [6] では、グラフに基づく間取りの類型化を行っている。

本論文では、文献 [5, 6] と同様、間取りに着目をした賃料分析を展開する。具体的には、種々の回帰分析を用い、複数の目的変数を設定するとともに、利用する部分グラフのクラスを限定することで、どのような部屋配置が賃料に影響を与えているかを考察する。

2 データセット

2.1 対象物件と詳細情報

本研究では「LIFULL HOME'S データセット」¹ を利用した。間取りに関して豊富なバリエーションを含む等の理由から、対象を（2階構造を持つ物件や間取り図を確認しにくい物件を除いた）東京 23 区内の 3 部屋マンション 547 件とした。表 1 に、対象データの賃料と主要な属性の統計値を示す。

本研究では、間取り図以外の物件情報から得られる情報を「詳細情報」と呼び、分析に利用する。詳細情報には、表 1 に示した駅までの徒歩距離、占有面積、部屋回数に加え、部屋の向きと備考情報に現れる頻出単語が含まれる。

部屋の向きは、カテゴリ変数であり、その値域は、東、西、南、北、南東、南西、北東、北西である。また、備考情報内の頻出単語に関しては、各単語を一つの変数とし、その単語を含むか否かを値とする。以下に利用した頻出単語の一覧を示す。

*連絡先：日本大学 文理学部 情報科学科
〒156-8550 東京都世田谷区桜上水 3-25-40
tozaki@chs.nihon-u.ac.jp

¹<http://www.nii.ac.jp/dsc/idr/lifull/homes.html>

表 1: 対象データに関する基本統計量

	平均	標準偏差
賃料	238304.8 円	176047.1
駅までの徒歩距離	641.1m	345.7
占有面積	96 平米	32.8
部屋階数	4.8 階	4.2

分譲, 相談, 礼金, 火災, ペット, 仲介, 手数料, 保険, オートロック, 乾燥, エアコン, システムキッチン, 最上階, 飼育, 充実, フリーレント, 分割払い, 敷金, 採光, 焚, ガス, 保証, 定期, 高級, キャンペーン, 角, 楽器, クレジットカード, 閑静, 日当たり, エレベーター, フローリング

2.2 間取り図情報

間取り図は, 部屋の配置のみならず洗面台やトイレ等を表す種々の記号や LDK, WIC などの略語等, 多様な情報を含む. 本研究では, 画像として提供される間取り図を, 部屋をノード, その繋がりをエッジとする間取り図グラフへと変換した上で利用する. なお, 本研究ではノードラベルを 19 種, エッジラベルを 6 種準備し, 変換自体は独自に開発したツールを用いて人手により行った. 以下に, ノードラベル・エッジラベルを示す.

玄関ノード 玄関を表す. 入り口と部屋が直接繋がっている場合でも, 便宜上玄関ノードを設置する.

廊下ノード 玄関と部屋間の空間を表す.

洋室ノード 用途を限定しない一般的な洋居室を表す.

和室ノード 用途を限定しない一般的な和居室を表す.

ダイニングノード ダイニングを表す.

リビングノード リビングを表す.

キッチンノード (部屋と間仕切りがある) キッチンを表す. 対面キッチンやシステムキッチンなどは区別しない.

LD・LK・DK・LDK ノード リビング(L)・ダイニング(D)・キッチン(K)がそれぞれ, 1つの空間に間仕切りが無く存在する場合に利用する.

ベランダノード ベランダやバルコニーを表す.

洗面室ノード 洗面室を表す.

風呂ノード 浴室を表す.

トイレノード トイレを表す.

収納ノード 居室に隣接した収納を表す.

押入れノード 和室に隣接した押入れを表す.

ウォークインクローゼットノード 衣類等を収納するための部屋を表す.

窓ノード 部屋の壁に設けた開口部を表す.

ドアエッジ 部屋同士が開き戸でつながっていることを表す

引き戸エッジ 部屋同士が引き戸・襖でつながっていることを表す.

収納エッジ 部屋と収納ノードとをつなぐ.

ガラスエッジ 部屋と窓ノードとをつなぐ.

無しエッジ 空間を分ける仕切りに, ドアや引き戸が存在しない場合を表す.

壁エッジ 部屋と部屋が壁のみでつながっている, すなわち位置としては隣接しているが, 直接行き来することはできないことを表す.

本研究では, 対象物件の間取り図を間取り図グラフに変換したのち, それらに頻出する部分グラフ(部屋配置)を抽出し, 属性として利用する. すなわち, グラフデータベース D 中における部分グラフ P の頻度を $support(P, D) = |\{G \in D \mid P \subseteq G\}| / |D|$ と定義し, この値が一定値を超える部分グラフを属性とする. なお, 頻出部分グラフの抽出には gSpan アルゴリズム [7]² を利用した.

3 回帰分析による賃料予測

3.1 実験設定

本研究では, 以下に示す賃料に関する値を対象に, 回帰分析を行う.

(T_A) 賃料そのもの

(T_B) 賃料そのもの(説明変数に相場を含める)

(T_C) 賃料と相場との偏差

(T_D) 賃料の相場に対する比

²<https://www.cs.ucsb.edu/~xyan/software/gSpan.htm>

なお、賃料には共益費、管理費は含まないものとした。また相場は、相場情報³を参考に、区単位で決定した。

説明変数に関しては、すべての分析において「詳細情報」を利用する一方で、利用する頻出部分グラフに関していくつかの制限を設けることで、下記の10の組み合わせを準備した。

- (G₁) 部分グラフを使用しない
- (G₂) 中頻度の頻出部分グラフを利用
- (G₃) サイズが2以上の頻出部分グラフを利用
- (G₄) サイズが3以上の頻出部分グラフを利用
- (G₅) サイズが4以上の頻出部分グラフを利用
- (G₆) サイズが2以上のフリーパターンを利用
- (G₇) サイズが3以上のフリーパターンを利用
- (G₈) 低頻度のフリーパターンを利用
- (G₉) 中頻度のフリーパターンを利用
- (G₁₀) 高頻度のフリーパターンを利用

なお、フリーパターンとは同一頻度を持つ部分グラフを同値類としたときの極小元、すなわち条件 $\neg \exists Q \subset P$ s.t. $\text{support}(P, D) = \text{support}(Q, D)$ を満たす部分グラフ P と定義される。今回は、同一頻度を持つ部分グラフを一つにまとめることで、回帰分析における多重共線性の問題を軽減することを狙いとしている。

3.2 実験結果

実験では、目的変数 $T_A \sim T_D$ と説明変数 $G_1 \sim G_{10}$ のすべての組み合わせに対し、4つの回帰分析手法

- 回帰木 (RT)[8]
- モデル木 (MT)[9, 10]
- サポートベクトル回帰 (SVR)[11]
- 重回帰 (MRA)

を適用し、それぞれの平均絶対誤差を算出した。なお、 T_C, T_D に関しては、相場の値から予測賃料を改めて算出した上で平均絶対誤差を計算している。また、実験には機械学習ライブラリ Weka⁴ を利用した。実験の集計結果を表2に示す。

表 2: 実験結果：平均絶対誤差の平均値・最良値

	平均	最良	
T_A	54035	35570	MT, G_2
T_B	50687	31098	MT, G_3
T_C	53617	35012	MT, G_3
T_D	42168	29896	RT, G_3
G_1	44219	30179	RT, T_D
G_2	49741	30108	RT, T_D
G_3	49586	29896	RT, T_D
G_4	48816	30292	RT, T_D
G_5	47182	31473	RT, T_D
G_6	56077	31680	MT, T_B
G_7	55008	32609	MT, T_D
G_8	62149	31618	MT, T_B
G_9	44106	31135	MT, T_B
G_{10}	44376	31490	MT, T_B
RT	37366	29896	$T_D - G_3$
MT	34884	30605	$T_D - G_2$
SVR	62672	41598	$T_D - G_9$
MRA	65581	41728	$T_D - G_9$

全体を通じ、数万円程度と平均絶対誤差が大きく、必ずしも予測精度の高いモデルが得られたわけではないが、実験結果より、いくつかのことが考察できる。まず目的変数に関しては、相場との比を用いる T_D が平均、最良ともに最も値が良い（誤差が小さい）ことが分かる。これは比を用いることで、目的変数の値域が小さくなったことが影響していると考えられる。次に利用する説明変数（頻出部分グラフ）に関しては、部分グラフを用いない G_1 が平均でトップ、最良でも3番目に精度が高く、必ずしも部分グラフが有効に機能しているとは言えない結果となった。一方、回帰モデルに関しては、ツリーベースの回帰モデルである回帰木とモデル木の精度が高い結果となった。このことは、一つのモデル（回帰式）で賃料を予測することは困難であり、適切な場合分けが必要であることを示唆していると考えられる。また、実験を通じて最も精度が高かった組み合わせは、サイズ2以上の頻出部分グラフ (G_3) を説明変数とし、相場との比 (T_D) を回帰木 (RT) でモデル化した場合であり、このことから、属性として多くの部分グラフを準備した上で、条件に応じてそれぞれ適切な属性を選択することが重要であることが伺える。

図1に、目的変数を T_D 、説明変数を G_3 としたときに得られた回帰木とモデル木の一部を示す。図より、いくつかの部分グラフが利用されていることが分かる。ここで回帰木における部分グラフ A は、1つの洋室ノードに2つの収納ノードが連結された部分グラフである。

³<http://www.homes.co.jp/chintai/price/>

⁴<http://www.cs.waikato.ac.nz/ml/weka/>

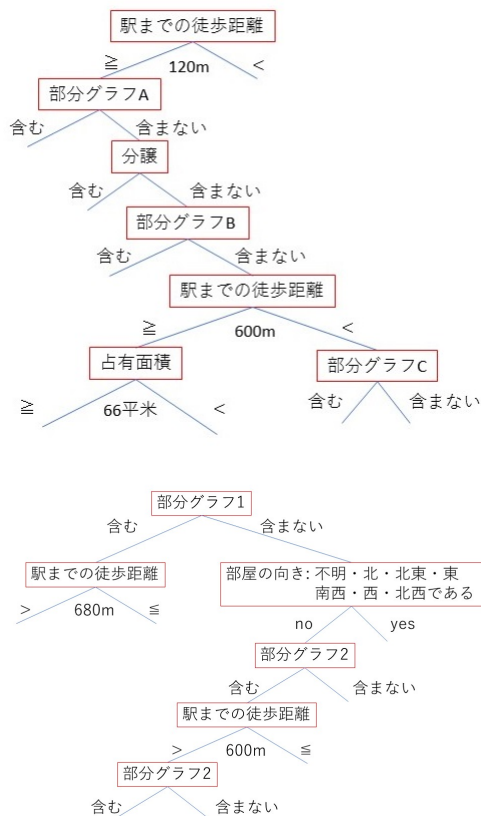


図 1: 得られた回帰木（上）とモデル木（下）の一部

またモデル木における部分グラフ 1 は、廊下、トイレ、収納の 3 ノードがこの順で連結された部分グラフである。前者は 2 つの収納がある洋室を、また後者は独立した収納のあるトイレを表しており、収納の豊富さが一つの基準となる可能性があることが伺える。

4 おわりに

本研究では、東京 23 区の 3 部屋マンションを対象に、物件が持つ詳細情報および部屋配置を説明変数とする賃料に関する回帰モデルを複数構築し、精度の面から比較を行った。またその過程を通じ、賃料に影響を与えていると考えられるいくつかの部屋配置を特定できる可能性を確認した。

今後の課題として、傾向スコア分析 [12] 等を利用し、各部屋配置が賃料に与える影響の定量化等を検討している。

謝辞： 本研究では、株式会社 LIFULL が国立情報学研究所の協力により研究目的で提供している「LIFULL HOME'S データセット」を利用した。

参考文献

- [1] 阿部成治, 石崎幸司: 首都圏における民間賃貸住宅家賃の重回帰分析, 都市住宅学 (19), pp.39-44, 1997.
- [2] 金春愛, 黄嘉平, 住田潮, 盧韶南: マクロ・ミクロ統合に基づく不動産賃料推定モデルの開発, 筑波大学社会工学commons Discussion Paper Series, No.1182, 2007.
- [3] 山鹿久木, 中川雅之, 齊藤誠: 地震危険度と家賃-耐震対策のための政策的インプリケーション, 日本経済研究 (46), pp.1-21, 2002.
- [4] 瀧澤重志, 材木敦史, 加藤直樹, 具源龍: 新橋に立地するオフィスビルの感性評価を考慮した賃料分析, 日本建築学会計画系論文集 (627), pp.1053-1059, 2008.
- [5] 瀧澤重志, 吉田一馬, 加藤直樹: グラフマイニングを用いた室配置を考慮した賃料分析: 京都市郊外の 3LDK を中心とした賃貸マンションを対象として, 日本建築学会環境系論文集 (623), pp.139-146, 2008.
- [6] 花里俊廣, 平野雄介, 佐々木誠: 首都圏で供給される民間分譲マンション 100 m^2 超住戸の隣接グラフによる分析, 日本建築学会計画系論文集 (591), pp.9-16, 2005.
- [7] X. Yan and J. Han : gSpan: Graph-Based Substructure Pattern Mining, Proc. of the 2002 IEEE International Conference on Data Mining, 2002.
- [8] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone : Classification and regression trees, Wadsworth, 1984.
- [9] J.R. Quinlan : Learning with Continuous Classes, Proc. of the 5th Australian Joint Conference on Artificial Intelligence, pp.343-348, 1992.
- [10] Y. Wang and I. H. Witten : Induction of model trees for predicting continuous classes, Proc. of Poster papers of the 9th European Conference on Machine Learning, pp.128-137, 1997.
- [11] A.J. Smola and B. Schoelkopf : A tutorial on support vector regression, NeuroCOLT2 Technical Report NC2-TR-1998-030, 1998.
- [12] 星野崇宏: 調査観察データの統計科学—因果推論・選択バイアス・データ融合, 岩波書店, 2009.