人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第21回)
SIG-AM-21-14

部屋配置とその出現数に着目した二段階賃料推定

Two-stage rent estimation using frequent subgraphs in floor plans

長谷川 優也　尾崎 知伸 *
Yuya Hasegawa　Tomonobu Ozaki

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

Abstract: The rent for rental property is determined by various factors such as area and age. In this paper, a two step rent estimation model is proposed in which the estimated rent value from the basic attributes will be corrected based on the layout information. In the model, each rental property is represented as a graph of room layout, and the number of occurrences of partial layouts (or subgrahs) is employed as attributes. In the evaluation experiments, in addition to the verification of the estimation accuracy, we extracted critical layouts to improve the interpretability.

1　はじめに

マンション等の賃貸物件の賃料は，専有面積や築年数に加え，駅までの距離や部屋の向き，周辺環境，間取り（部屋配置）など，様々な要因によって決定される．一般に，専有面積が大きいほど賃料は高くなり，築年数が古いほど賃料は安くなるなど，いくつかの要因に関しては賃料への影響は明らかである．その一方で，賃料増減の明示的な根拠を示すことが難しい要因として，間取り（部屋配置）があげられる．例えば，専有面積や築年数，駅までの距離等の条件がまったく同じ場合でも，部屋配置が異なれば賃料も異なることは一般的であるが，どの様な部屋配置が（いくつ）あれば，どの程度賃料が高く（または安く）なるのかを明示的に説明することは容易ではない．仮に，賃料に大きな影響を与える部屋配置を特定することができれば，物件の貸し手に対しては賃料設定に関する新たな根拠を，借り手に対しては設定された賃料の妥当性判断基準を，また物件の設計者に対しては新たな検討要素をそれぞれ提供することが可能となり，その影響は大きいと考えられる．

これらのことを背景に，本論文では，部屋配置と賃料との関係を分析し，賃料に大きな影響を与える部屋配置を特定することを試みる．具体的には，専有面積や築年数などの基本的な属性から得られる賃料推定結果を間取り情報に基づき補正する，二段階賃料推定手法を提案する．その際，各間取りをグラフ化すると共に，新たな部分グラフ出現数の定義を提案し，物件（グラフ）中の部屋配置（部分グラフ）出現数を属性として利用する．詳しくは後述するが，提案手法では，家賃推定モデルを 2 回の教師付き学習（回帰分析）で構築することとなり，より精度の高い家賃推定が期待できる．加えて，部屋配置のみを利用して構築される補正モデルの近似ルールを抽出することで，賃料に影響を与える部屋配置の明示的な特定を実現する．

本論文の構成は以下の通りである．2 章で関連研究について言及する．3 章では，本研究に関する形式的な定義を与え，二段階賃料推定手法を提案する．その後 4 章で，間取り図のグラフ化を通じた属性の準備について説明する．5 章では，賃料の推定および賃料への影響が大きい間取りの特定に関し，それぞれ評価実験を行う．最後に 6 章でまとめを行い，今後の課題を述べる．

2　関連研究

これまでに種々の要因に着目した賃料の分析・推定に関する研究が報告されている [1, 2, 3, 4]．例えば文献 [1] では，重回帰分析を用い，物件の各詳細情報が賃料に与える影響を路線別に分析している．文献 [2] では，物件に対する詳細情報を専有面積や部屋数など物件そのものが持つ固有情報と，所在地や駅までの距離などの外的要因に依存する情報の 2 つに分類し，それぞれに対するモデルを構築・統合することで賃料を推定する手法を提案している．また外部データの利用に関しては，文献 [3] において，地震に関する地域危険度調査結果を用い，物件所在地の危険度が賃料に与える影響を建物の構造別，新旧の耐震基準別に分析している．文献 [4] では，物件詳細の客観的な情報に加え，実際に

*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
E-mail: tozaki@chs.nihon-u.ac.jp

物件を見たときの感性などの主観情報も利用し，重回帰分析やサポートベクトル回帰 [5] を用いて分析を行っている．

一方，間取りに着目した分析もいくつか報告されている [6, 7, 8]．これらの研究では，部屋を頂点，そのつながりを辺とするグラフを用いて部屋配置（間取り）を表現し，構造的な側面からの分析を試みている．例えば文献 [6] では，大型マンションを対象にグラフに基づく間取りを類型化した上で順序付けを行い，間取りが賃料に与える影響を分析している．また文献 [7] では，頻出部分グラフマイニング [9, 10] の枠組みを利用した分析を展開している．複数の間取り図（グラフ）に共通して現れる部屋配置（頻出部分グラフ）を抽出し，相関ルール分析を用いてそれらの関係を分析するとともに，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，部分グラフが賃料に与える影響を考察している．文献 [8] でも同様に，頻出部分グラフにおけるフリーパターン [11] を属性とした回帰木 [12] やモデル木 [13, 14] を構築することで，賃料に影響の強い部分グラフの抽出を試みている．

本論文では文献 [7, 8] と同様，頻出部分グラフマイニングの枠組みを利用した賃料推定を展開する．その際，属性値として，頻出部分グラフの出現の有無だけではなく，出現数を採用することを提案する．また部分グラフのみを利用して賃料の補正モデルを構築することで，より明示的な形で，部屋配置が賃料に与える影響を分析する．

3 賃料推定の定式化

3.1 準備

物件賃料の予測モデル構築に利用する属性の全体集合を $X = X_{basic} \cup X_{graph} \cup \{x_{mp}\}$ と表記する．ここで属性 x_{mp} は最寄り駅単位で算出する物件に対する相場を表す．また X_{basic} は，専有面積や駅からの距離など物件に関する基本的な属性（以降，基本属性と呼ぶ）の集合を，X_{graph} は 4 章で導入する部屋配置に関する属性（以降，部屋配置属性と呼ぶ）の集合をそれぞれ表す．一方，予測対象である物件賃料属性を p と表記する．また，物件 t に対する属性 $a \in X \cup \{p\}$ の属性値を $a(t)$ と表記する．以上の準備を基に，本研究では，予測値（目的変数）の設定方法と利用属性（説明変数）が異なる複数の回帰モデルを考える．

最も基本的なモデルは，基本属性，部屋配置属性，相場のすべてを利用し，賃料を推定するモデル，すなわち

$$\widehat{p} = f_{bgm}\left(X_{basic}, X_{graph}, x_{mp}\right) \tag{1}$$

である．

第二の基本的モデルとして，賃料 p を直接推定するのではなく，相場との比 $\frac{p}{x_{mp}}$ を推定するモデル $f_{bg/m}$

$$\widehat{\frac{p}{x_{mp}}} = f_{bg/m}\left(X_{basic}, X_{graph}\right) \tag{2}$$

を考える．なおこの場合，物件 t に対する賃料は

$$\widehat{p(t)} = f_{bg/m}\left(X_{basic}(t), X_{graph}(t)\right) \times x_{mp}(t)$$

と推定することができる．

3.2 二段階賃料推定

賃料推定に関する回帰モデル f_{bgm} と $f_{bg/m}$ はそれぞれ，一つのモデルを用いて賃料を推定する．これに対し本論文では，二つのモデル，すなわち基本属性を用いた賃料推定モデルと間取り情報を用いた賃料補正モデルを組み合わせることを提案する．

提案する第一の推定モデル f_{bm+g} は，基本属性と相場を説明変数とするモデル f_{bm} の推定結果を，部屋配置属性を説明変数とするモデル f_g を用いて補正する構成をしており，形式的には下式で表現される．

$$\begin{aligned}\widehat{p} &= f_{bm+g}(X_{basic}, x_{mp}, X_{graph}) \\ &= f_{bm}\left(X_{basic}, x_{mp}\right) + f_g(X_{graph})\end{aligned} \tag{3}$$

本論文では，下記に示す方法により，f_{bm+g} を 2 段階に分けて推定することを提案する．

1. まず，目的変数を p とし，基本属性 X_{basic} と相場 x_{mp} のみを利用して f_{bm} を推定する．
2. その後，第二の推定として，賃料と推定値の残差 $p - f_{bm}(X_{basic}, x_{mp})$ を目的変数とし，部屋配置属性 X_{graph} を利用してモデル f_g を構築する．

この様に推定を 2 段階に分けることは，2 回の教師付き学習を行うことに相当し，推定精度の向上が期待できる．また補正モデルである f_g は，部屋配置属性 X_{graph} のみを利用して構築されるため，部屋配置が賃料に与える影響をより直接的に反映していることが期待できる．

本論文では f_{bm+g} に加え，相場との比を推定する二段階推定モデルとして，$f_{b/m+g}$ を提案する．

$$\begin{aligned}\widehat{\frac{p}{x_{mp}}} &= f_{b/m+g}\left(X_{basic}, X_{graph}\right) \\ &= f_{b/m}\left(X_{basic}\right) + \frac{1}{x_{mp}} f_g\left(X_{graph}\right)\end{aligned} \tag{4}$$

モデル $f_{b/m+g}$ の推定も，f_{bm+g} と同様の方法で実現する．すなわち，基本属性 X_{basic} を用いて相場との比

$\frac{p}{x_{mp}}$ を推定するモデル $f_{b/m}(X_{basic})$ を構築し，その後，残差を目的変数，部屋配置属性を説明変数としたモデル f_g を推定する．

4 部屋配置属性の構築

本章では，グラフとして表現される間取りの集合から，部屋配置属性 X_{graph} を抽出する方法，および各物件 t に対する属性値 $X_{graph}(t)$ の算出方法について形式的に説明する．

4.1 頻出部分グラフマイニングを用いた部屋配置属性の抽出

本論文では，関連研究 [6, 7, 8] と同様，部屋を頂点，その繋がりを辺とするグラフを用いて各部屋の間取りを定式化する．具体的には，12 種の頂点ラベル（玄関，廊下，居室，水回り，収納，ウォークインクローゼット，台所，ダイニングキッチン，リビングダイニング，リビングダイニングキッチン，ベランダ，窓）及び 5 種の辺ラベル（ドア，引き戸，ガラス，無し，収納）を用い，ラベル付き無向グラフとして間取りを表現する．なお「水回り」には洗面室，風呂，トイレを含める．また「窓」と「収納」は部屋ではないが，部屋配置に関する主要な要素として採用している．一方，辺に対する「無し」ラベルは，空間を分ける仕切りとしてドアや引き戸が存在しない場合に相当する．

頂点・辺ラベルの全体集合を $\mathcal{L}$ と表記し，ラベル付きグラフ $g = (V_g, E_g, \lambda_g)$ を，頂点集合 V_g と辺集合 $E_g \subseteq V_g \times V_g$，ラベル関数 $\lambda_g : V_g \cup E_g \to \mathcal{L}$ の 3 項組で表現する．2 つのグラフ $g = (V_g, E_g, \lambda_g)$ と $p = (V_p, E_p, \lambda_p)$ に対し，条件

1. $\forall u \in V_p\,[\lambda_p(u) = \lambda_g(f(u))]$
2. $\forall (u, v) \in E_p$
   $\exists (f(u), f(v)) \in E_g \; s.t. \; \lambda_p(u, v) = \lambda_g(f(u), f(v))$

を満たす単射関数 $f : V_p \to V_g$ が存在するとき，p を g の部分グラフと呼び $p \sqsubseteq g$ と表記する．n 個のグラフから構成されるデータベース $D = \{g_1, \cdots, g_n\}$ に対し，グラフ p の支持度を D 中で p を含むグラフの割合，すなわち

$$\mathrm{S}(p, D) = |\{g \in D \mid p \sqsubseteq g\}| / |D|$$

と定義する．利用者による頻度に関する閾値 $\sigma\,(0 < \sigma \leq 1)$ に対し，条件 $\mathrm{S}(p, D) \geq \sigma$ を満たすグラフ p を頻出部分グラフと呼ぶ．また D における頻出部分グラフの集合を

$$\mathcal{F}_D^\sigma = \{p \sqsubseteq g \mid g \in D, \mathrm{S}(p, D) \geq \sigma\}$$

と表記する．

賃貸物件 t に対し，その間取りを表すラベル付き無向グラフを $g(t)$ と表記する．また，賃貸物件集合 T に対し，それらの間取り図グラフの集合を $D_g = \{g(t) \mid t \in T\}$ と表記する．本論文では，D_g から得られる各頻出部分グラフ $p \in \mathcal{F}_{D_g}^\sigma$ を属性して利用する．すなわち $X_{graph} = \mathcal{F}_{D_g}^\sigma$ とする．

4.2 各物件に対する部屋配置属性の属性値

各頻出部分グラフ $p \in X_{graph}$ を属性とする場合，賃貸物件 t に対する p の属性値 $p(\,g(t)\,)$ が必要となる．属性値 $p(\,g(t)\,)$ として，グラフ $g(t)$ に対する p の支持度（出現数）を採用することが自然であると考えられるが，一般に，単一グラフ $g(t)$ における部分グラフ p の支持度は自明ではない．加えて，部屋配置を用いた賃料推定という応用を前提とした場合，これまでに提案されている支持度が必ずしも適しているとは限らない．以上のことを背景に，本論文では，単一グラフにおける部分グラフの新たな支持度を提案し，属性値計算に利用することを考える．以下，形式的な準備の後，既存の支持度 [15, 16] を導入し，次いで新たな支持度を提案する．

グラフ g と p に対し，p と同型である g の部分グラフ，すなわち条件 $o \sqsubseteq g \wedge p \sqsubseteq o \wedge o \sqsubseteq p$ を満たす部分グラフ $o = (V_o, E_o, \lambda_o)$ を g における p の出現と呼ぶ．また g における p の出現の全体集合を

$$O(p, g) = \{o \mid o \sqsubseteq g, p \sqsubseteq o, o \sqsubseteq p\}$$

と表記する．以下 $O(p, g)$ を用い，グラフ g に対する p の支持度を複数定義する．

(1) 出現の有無に基づく支持度

第 1 の支持度 $\mathrm{S}_{BIN}(p, g)$ は，グラフ g 中に p の出現が存在するか否かに基づくものであり，以下の様に定義される．

$$\mathrm{S}_{BIN}(p, g) = \begin{cases} 0 & (\,O(p, g) = \emptyset\,) \\ 1 & (\,\text{otherwise}\,) \end{cases}$$

支持度 $\mathrm{S}_{BIN}(p, g)$ は，グラフを用いた賃料推定に関する既存研究 [7, 8] を含め，幅広く用いられている．しかし，同じ部屋配置が 1 回しか現れない場合と 2 回以上現れる場合とを区別することができず，頂点・辺の数が小さい頻出部分グラフを属性とした場合，識別能力に関して問題があると考えられる．

(2) 極大独立集合に基づく支持度 [15]

第 2 の支持度 $\mathrm{S}_{MIS}(p,g)$ は，頂点を共有せずに g 中に配置することのできる出現 o の最大数であり，以下の様に定義される．

$$\mathrm{S}_{MIS}(p,g) = |MIS(p,g)|$$

ここで $MIS(p,g)$ は，g における p の各出現 $o \in O(p,g)$ を頂点，(g における) 頂点を共有する出現同士を結んだものを辺とするグラフ，すなわち

$$\big(O(p,g), \{(o_j,o_k) \mid o_j,o_k \in O(p,g), V_{o_j} \cap V_{o_k} \neq \emptyset\}, \lambda\big)$$

の極大独立集合を表す．

(3) 最制限頂点に基づく支持度 [16]

第 3 の支持度 $\mathrm{S}_{MRN}(p,g)$ は，p 中の頂点 $v \in V_p$ の写像先集合 ($\varphi_o(v)$ の適用結果) に対する最小要素数であり，以下の様に定義される．

$$\mathrm{S}_{MRN}(p,g) = \min_{v \in V_p} |\{\varphi_o(v) \mid o \in O(p,g)\}|$$

ここで $\varphi_o : V_p \to V_g$ は，出現 $o = (V_o, E_o, \lambda_o)$ に対し，条件

1. $\forall v \in V_p \Rightarrow \lambda_p(v) = \lambda_g(\varphi(v))$
2. $\forall (u,v) \in E_p \Rightarrow (\varphi(u), \varphi(v)) \in E_g$

を満たす頂点間の写像関数である．

$\mathrm{S}_{MIS}(p,g)$ 同様，$\mathrm{S}_{MRN}(p,g)$ も出現の重複を考慮した支持度である．$\mathrm{S}_{MIS}(p,g)$ は，出現 o の単位での重複を許さないのに対し，$\mathrm{S}_{MRN}(p,g)$ は，部分グラフ p の頂点の単位で，重複を無視する形でカウントを行っている．両支持度は，出現数を考慮するという点で，出現数の違いを区別できない $\mathrm{S}_{BIN}(p,g)$ の弱点を克服している．またそれぞれ支持度に関する逆単調性

$$\forall p \sqsubseteq q \to \mathrm{S}_{MIS}(p,g) \geq \mathrm{S}_{MIS}(q,g)$$
$$\forall p \sqsubseteq q \to \mathrm{S}_{MRN}(p,g) \geq \mathrm{S}_{MRN}(q,g)$$

を満たすことが知られている．

その一方で，部屋配置という対象の性質を考えた場合，重複を許容しないことが必ずしも良いとは限らない．例えば，キッチンに 1 つの部屋だけが繋がっているグラフ g^1 と，キッチンに 3 つの部屋が繋がっているグラフ g^3 に対し，“キッチン-居室” という部分グラフ p の支持度はすべて $\mathrm{S}_{MIS}(p,g^1) = \mathrm{S}_{MIS}(p,g^3) = \mathrm{S}_{MRN}(p,g^1) = \mathrm{S}_{MRN}(p,g^3) = 1$ となり，支持度の差はない．しかし，g^3 においてはキッチンへの導線が複数あり，また隣接した各部屋の利用方法にも複数の組み合わせが考えられることから，最も制約の強いキッチンだけを基準に支持度を考えることは必ずしも適切ではない．上記の議論を基に，本論文では，単一グラフにおける部分グラフの支持度として出現数に基づく支持度 $\mathrm{S}_{NOO}(p,g)$ および頂点数の比に基づく支持度 $\mathrm{S}_{ROV}(p,g)$ を提案する．

(4) 出現数に基づく支持度

本論文では，出現 $o \in O(p,g)$ の辺集合の種類数を支持度とすることを提案する．以下に提案する支持度 $\mathrm{S}_{NOO}(p,g)$ の形式的な定義を示す．

$$\mathrm{S}_{NOO}(p,g) = |\{E_o \mid (V_o,E_o,\lambda_o) \in O(p,g)\}|$$

$\mathrm{S}_{MIS}(p,g)$ や $\mathrm{S}_{MRN}(p,g)$ が，出現間で一部分でも重複を許容しないことに対し，$\mathrm{S}_{NOO}(p,g)$ では完全な重複以外を許容している．これにより，逆単調性 ($\forall p \sqsubseteq q \to \mathrm{S}_{NOO}(p,g) \geq \mathrm{S}_{NOO}(q,g)$) は成立しないが，出現の違いをより明確に表すことが可能となると考えられる．

(5) 頂点数の比に基づく支持度

支持度 $\mathrm{S}_{NOO}(p,g)$ は出現の重複を考慮しないため，一ヶ所だけが異なるような出現を多数持つような大きな部分グラフに対して不当に大きな値を与えてしまう可能性がある．この問題を軽減するために，出現に含まれる総頂点数を部分グラフの頂点数で正規化した支持度 $\mathrm{S}_{ROV}(p,g)$ を提案する．

$$\mathrm{S}_{ROV}(p,g) = \frac{|\{v \in V_o \mid (V_o,E_o,\lambda_o) \in O(p,g)\}|}{|V_p|}$$

出現の重複に対して正規化の考え方を適用することで，重複を許容しない $\mathrm{S}_{MIS}(p,g)$ や $\mathrm{S}_{MRN}(p,g)$ と，重複を許す $\mathrm{S}_{NOO}(p,g)$ との中間的な性質を持つ支持度が実現されることが期待できる．

5 評価実験

5.1 実験データと実験設定

実験には，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 [1]」を利用した．具体的には，間取り図を確認しにくい物件や二階建て構造を持つ物件等を除いた東京 23 区内の 2 部屋マンション 400 件および 3 部屋マンション 561 件の合計 961 件を選定している．

基本属性 X_{basic} として，データセット中で提供されている “徒歩距離”，“部屋面積”，“部屋階数”，“部屋の向き”，“部屋数” の 5 属性を採用した．各属性の統計量を表 1 に示す．なお “部屋の向き” は離散属性であり，その内訳は，東 110 件，西 88 件，南 448 件，北 4 件，南東 111 件，南西 120 件，北東 15 件，北西 8 件，不明 57 件である．

一方，物件 t の間取りに関するラベル付き無向グラフ $g(t)$ は，文献 [17] で準備・使用されたデータを援用した．また，部屋配置属性 X_{graph} の抽出には，頻出部分グラフマイナー gSpan[18] [2] を利用した．具体的に

[1] https://www.nii.ac.jp/dsc/idr/lifull/homes.html
[2] https://www.cs.ucsb.edu/~xyan/software/gSpan.htm

表 1: 基本属性の統計量

変数	賃料	徒歩距離	部屋面積	部屋階数
平均	232,936	691	87	4.8
標準偏差	269,021	404	42	4.7
最小値	55,000	80	31	0
中央値	158,000	640	77	3
最大値	2,050,000	2,720	3,56	43

表 2: 頻出部分グラフ数

サイズ	1	2	3	4	5
グラフ数	9	13	13	12	2

は，支持度パラメタを $\sigma = 450/961$ とし，49 の頻出部分グラフからなる集合 $X_{graph} = \mathcal{F}_D^{450/961}$ を導出した．サイズ（辺数）別の頻出部分グラフ数を表 2 に示す．

5.2 推定精度

提案する二段階推定手法を評価するために，賃料推定モデルの構築実験を行った．具体的には，対象となる 961 件のデータ D を 861 件からなる訓練データ集合 D_{train} と 100 件からなるテストデータ集合 D_{test} に分割し，D_{train} に線形重回帰モデルと XGBoost[19]，Random Forest[20]，LightGBM[21] をそれぞれ適用することで，賃料推定のための各モデル（f_{bgm}，$f_{bg/m}$，f_{bm}，$f_{b/m}$，f_g）を構築した．また推定精度の評価には，D_{test} に対する平均絶対誤差（Mean Average Error, MAE）

$$\frac{1}{|D_{test}|} \sum_{t \in D_{test}} \left| p(t) - \widehat{p(t)} \right| \tag{5}$$

を用いた．実験結果を表 3 に示す．

実験結果より，属性値として $\mathrm{S}_{MIS}(p,g)$ を採用したモデル $f_{b/m+g}$ を XGBoost を用いて推定した場合に MAE $= 3,768$ となり，最も精度が高いことが分かる．また，推定手法別モデル毎の MAE 平均値の上位 3 件は，XGBoost を用いた $f_{b/m+g}$（MAE 平均値 $4,080$），XGBoost を用いた f_{bm+g}（MAE 平均値 $10,978$），Random Forest を用いた f_{bm+g}（MAE 平均値 $11,696$）であった．

推定手法別にモデル f_{bgm} と f_{bm+g}，また $f_{bg/m}$ と $f_{b/m+g}$ の MAE 平均値をそれぞれ比較すると，線形重回帰モデルにおける $f_{bg/m}$ と $f_{b/m+g}$ 以外はすべて f_{bgm} より f_{bm+g}，また $f_{bg/m}$ より $f_{b/m+g}$ の方が精度が高いことが確認できる．モデル毎の MAE 平均値を算出すると f_{bgm} は 37,948，$f_{bg/m}$ は 40,166，f_{bm+g} は 26,162，$f_{b/m+g}$ は 23,202 となり，f_{bm+g} の MAE 平均値は f_{bgm} の約 69%，$f_{b/m+g}$ の MAE 平均値は $f_{bg/m}$ の約 58%となっていることが分かる．以上のことから，提案した二段階推定手法が有効に働いていることが確認できる．

次に，部屋配置属性に対する属性値の違いについて考察する．推定手法別支持度毎の MAE 平均値の上位 3 件は，$\mathrm{S}_{BIN}(p,g)$ を採用した XGBoost（MAE 平均値 $20,590$），$\mathrm{S}_{MRN}(p,g)$ を採用した XGBoost（MAE 平均値 $21,469$），$\mathrm{S}_{NOO}(p,g)$ を採用した XGBoost（MAE 平均値 $21,977$）であった．また，推定手法別モデル毎に MAE 値が最良となった回数を支持度毎に集計すると，$\mathrm{S}_{BIN}(p,g)$ は 3 回，$\mathrm{S}_{MIS}(p,g)$ は 6 回，$\mathrm{S}_{MRN}(p,g)$ は 2 回，$\mathrm{S}_{NOO}(p,g)$ は 2 回，$\mathrm{S}_{ROV}(p,g)$ は 3 回となり，差が小さいとは言え，$\mathrm{S}_{MIS}(p,g)$ が優れていることが示唆された．加えて，評価値が高かった XGBoost を用いたモデル $f_{b/m+g}$ に着目すると，精度が高い支持度は順に $\mathrm{S}_{MIS}(p,g)$，$\mathrm{S}_{NOO}(p,g)$，$\mathrm{S}_{MRN}(p,g)$，$\mathrm{S}_{ROV}(p,g)$，$\mathrm{S}_{BIN}(p,g)$ となり，属性値として出現を考慮した支持度が有効に働いていることが確認できる．

その一方で，支持度毎の MAE 平均値を算出すると，$\mathrm{S}_{BIN}(p,g)$ は 31,797，$\mathrm{S}_{MIS}(p,g)$ は 31,506，$\mathrm{S}_{MRN}(p,g)$ は 31,579，$\mathrm{S}_{NOO}(p,g)$ は 32,234，$\mathrm{S}_{ROV}(p,g)$ は 32,232 となり，平均という観点からは大きな差は確認できなかった．このことは，推定手法別モデル毎の $\mathrm{S}_{BIN}(p,g)$ に対する各支持度の勝率（$\mathrm{S}_{MIS}(p,g)$ は 0.50，$\mathrm{S}_{MIS}(p,g)$ は 0.56，$\mathrm{S}_{MIS}(p,g)$ は 0.44，$\mathrm{S}_{MIS}(p,g)$ は 0.44）からも確認ができる．以上の結果より，今回の実験では，属性値として出現数を考慮した支持度を採用することは一定の設定においては有効であるが，必ずしもすべての場合において推定精度の向上に寄与するとは限らないことが確認された．

5.3 利用する部分グラフの制限

利用する部分グラフの影響を考察するため，部屋配置属性をサイズ 2 以下の部分グラフに限定した上でモデルの推定を行った．各モデルに対する平均絶対誤差を表 4 に示す．

今回の場合の MAE 最良値は，XGBoost を用いて推定した属性 $\mathrm{S}_{ROV}(p,g)$ を採用したモデル $f_{b/m+g}$ における $3,767$ であり，すべての頻出部分グラフを利用した場合とほぼ同様の値が得られた．また，推定手法別モデル毎の MAE 平均値の上位 3 件もすべての頻出部分グラフを利用した場合と同じであり，XGBoost を用いた $f_{b/m+g}$（MAE 平均値 $3,967$），XGBoost を用いた f_{bm+g}（MAE 平均値 $10,994$），Random Forest を用いた f_{bm+g}（MAE 平均値 $11,757$）であった．さらに，MAE 平均値を用いた二段階推定の効果について考

表 3: 実験結果：テストデータに対する平均絶対誤差（すべての部分グラフを利用した場合）

支持度 \ 関数	f_{bgm}	$f_{bg/m}$	f_{bm+g}	$f_{b/m+g}$	平均	f_{bgm}	$f_{bg/m}$	f_{bm+g}	$f_{b/m+g}$	平均
	Linear Regression					Random Forest				
$S_{BIN}(p,g)$	59,537	34,655	58,328	41,419	48,485	27,425	42,791	14,134	14,471	24,705
$S_{MIS}(p,g)$	59,894	35,936	53,382	44,831	48,511	26,189	44,585	10,497	12,074	23,336
$S_{MRN}(p,g)$	64,725	32,609	59,530	42,970	49,959	27,871	41,188	12,149	12,776	23,496
$S_{NOO}(p,g)$	55,978	35,076	58,101	46,906	49,015	33,075	46,919	11,064	12,523	25,895
$S_{ROV}(p,g)$	59,191	36,301	60,075	43,989	49,889	27,476	46,292	10,639	11,705	24,028
平均	59,865	34,915	57,883	44,023	49,172	28,407	44,355	11,696	12,710	24,292
	XGBoost					LightGBM				
$S_{BIN}(p,g)$	30,025	36,904	11,004	4,429	20,590	39,449	38,846	23,441	31,898	33,409
$S_{MIS}(p,g)$	33,568	43,632	10,590	3,768	22,890	30,461	38,706	23,947	32,033	31,287
$S_{MRN}(p,g)$	29,522	41,319	10,950	4,084	21,469	31,640	38,273	23,588	32,071	31,393
$S_{NOO}(p,g)$	28,276	44,402	11,214	4,016	21,977	31,540	39,054	25,454	32,143	32,048
$S_{ROV}(p,g)$	32,426	49,212	11,133	4,104	24,219	30,686	36,630	24,019	31,829	30,791
平均	30,763	43,094	10,978	4,080	22,229	32,755	38,302	24,090	31,995	31,786

察すると，線形重回帰モデルにおける $f_{bg/m}$ と $f_{b/m+g}$ 以外はすべて二段階推定の方が精度が高いことが確認できる．

属性値の違いに関しても，すべての頻出部分グラフを利用した場合と同様の傾向が確認できる．推定手法別支持度毎の MAE 平均の上位三件は，すべての頻出部分グラフを利用した場合と一致し，$S_{BIN}(p,g)$ を採用した XGBoost（MAE 平均値 19, 401），$S_{MRN}(p,g)$ を採用した XGBoost（MAE 平均値 20, 939），$S_{NOO}(p,g)$ を採用した XGBoost（MAE 平均値 21, 251）であった．また，精度が最も高い XGBoost を用いたモデル $f_{b/m+g}$ においては $S_{BIN}(p,g)$ の精度が最も低く，$S_{MRN}(p,g)$，$S_{NOO}(p,g)$，$S_{MIS}(p,g)$，$S_{ROV}(p,g)$ の順に精度が良くなっていることが確認できる．一方で，推定手法別モデル毎に MAE 値が最良となった回数を支持度毎に集計すると，$S_{BIN}(p,g)$ は 1 回，$S_{MIS}(p,g)$ は 4 回，$S_{MRN}(p,g)$ は 5 回，$S_{NOO}(p,g)$ は 2 回，$S_{ROV}(p,g)$ は 4 回となり，出現を考慮した支持度がより有効に働いていることが確認できる．

5.4 影響力の大きな部分グラフの抽出

賃料に大きな影響を与える部屋配置やその組み合わせを特定するために，推定したモデル f_g に対して，樹状モデルアンサンブルから近似ルール集合を抽出する手法である defragTrees[22][3] を適用した．モデル f_g は部屋配置属性のみを用いた残差調整モデルであり，予測値の大きいルールに現れる部屋配置の組み合わせが賃料に与える影響が大きいと考えられる．f_g から得られる近似ルール集合において，予測値が最大であるルールを図 1 と図 2 に示す．なお図 1 は，頻出部分グラフのサイズ上限を 2，支持度 $S_{MIS}(p,g)$ を採用した XGBoost によるモデル f_{bm+g} に対する結果である．一方図 2 は，同じく頻出部分グラフのサイズ上限を 2，支持度 $S_{ROV}(p,g)$ を採用した XGBoost によるモデル $f_{b/m+g}$ に対する結果である．

図 1 と図 2 より，各部屋配置の支持度に関して上限と下限が設定されていることが分かる．また図 2 における最後の条件 $[\,(\text{収納}) \overset{\text{収納}}{\longleftrightarrow} (\text{玄関})\,] < 1.00$ は，玄関に収納がないことを表している．これらのことより，特定の部屋配置が多ければ多いほど良いというわけではなく，適当な量が存在することが伺える．

図 1 には，“収納と窓のある居室” や “収納がある廊下” など，収納に関する条件が多く，収納が賃料を決定する一つの要因となっていることが考えられる．このことは図 2 にも当てはまる．さらに図 2 では，“居室” を含む部屋配置グラフが数多く含まれていることが分かる．両ルールにおいて共通する部屋配置は

(窓) $\overset{\text{ガラス}}{\longleftrightarrow}$ (ベランダ) $\overset{\text{ガラス}}{\longleftrightarrow}$ (窓)，(収納) $\overset{\text{収納}}{\longleftrightarrow}$ (廊下)，
(収納) $\overset{\text{収納}}{\longleftrightarrow}$ (居室) $\overset{\text{ガラス}}{\longleftrightarrow}$ (窓)，(居室) $\overset{\text{ガラス}}{\longleftrightarrow}$ (窓)，
(収納) $\overset{\text{収納}}{\longleftrightarrow}$ (居室) $\overset{\text{ドア}}{\longleftrightarrow}$ (廊下)

の 5 つであり，この結果からも，“居室” と “収納” が中心的な役割を果たしていることが示唆される．

[3] https://github.com/sato9hara/defragTrees

表 4: 実験結果：テストデータに対する平均絶対誤差（サイズ 2 以下の部分グラフのみを利用した場合）

支持度 \ 関数	f_{bgm}	$f_{bg/m}$	f_{bm+g}	$f_{b/m+g}$	平均	f_{bgm}	$f_{bg/m}$	f_{bm+g}	$f_{b/m+g}$	平均
	Linear Regression					Random Forest				
$S_{BIN}(p,g)$	59,425	33,785	58,155	38,297	47,415	29,061	43,169	16,310	14,081	25,655
$S_{MIS}(p,g)$	55,475	32,955	60,620	42,233	47,821	25,551	45,511	10,173	12,467	23,425
$S_{MRN}(p,g)$	60,391	31,764	56,938	39,097	47,048	33,045	38,541	12,162	13,349	24,274
$S_{NOO}(p,g)$	58,077	34,964	58,326	37,405	47,193	25,290	44,376	10,443	12,479	23,147
$S_{ROV}(p,g)$	60,115	33,547	58,401	41,122	48,296	26,990	44,927	9,699	12,333	23,487
平均	58,697	33,403	58,488	39,631	47,555	27,987	43,305	11,757	12,942	23,998
	XGBoost					LightGBM				
$S_{BIN}(p,g)$	22,658	39,505	11,125	4,316	19,401	32,777	38,584	23,862	31,794	31,754
$S_{MIS}(p,g)$	30,024	41,988	10,547	3,793	21,588	29,657	37,206	24,506	31,912	30,820
$S_{MRN}(p,g)$	29,271	39,227	11,231	4,025	20,939	31,826	38,399	24,032	31,610	31,467
$S_{NOO}(p,g)$	29,256	40,669	11,099	3,980	21,251	31,521	37,535	27,113	31,752	31,980
$S_{ROV}(p,g)$	30,075	46,901	10,969	3,767	22,928	31,269	38,129	23,772	31,897	31,267
平均	28,257	41,658	10,994	3,976	21,221	31,410	37,971	24,657	31,793	31,458

$$
\begin{aligned}
+8{,}828 \Leftarrow\ & [\,(\text{窓}) \overset{\text{ガラス}}{\longleftrightarrow} (\text{ベランダ}) \overset{\text{ガラス}}{\longleftrightarrow} (\text{窓})\,] \geq 1 \\
\wedge\ & [\,(\text{居室}) \overset{\text{ガラス}}{\longleftrightarrow} (\text{窓})\,] \geq 1 \\
\wedge\ & [\,(\text{収納}) \overset{\text{収納}}{\longleftrightarrow} (\text{居室}) \overset{\text{ガラス}}{\longleftrightarrow} (\text{窓})\,] \geq 1 \\
\wedge\ & [\,(\text{収納}) \overset{\text{収納}}{\longleftrightarrow} (\text{居室}) \overset{\text{ドア}}{\longleftrightarrow} (\text{廊下})\,] < 4 \\
\wedge\ & [\,(\text{収納}) \overset{\text{収納}}{\longleftrightarrow} (\text{廊下}) \overset{\text{ドア}}{\longleftrightarrow} (\text{水回り})\,] < 3 \\
\wedge\ & [\,(\text{収納}) \overset{\text{収納}}{\longleftrightarrow} (\text{廊下})\,] < 3
\end{aligned}
$$

図 1: モデル f_{bm+g} から得られたルールの例

$$
\begin{aligned}
+18{,}902 \Leftarrow\ & [\,(\text{窓}) \overset{\text{ガラス}}{\longleftrightarrow} (\text{ベランダ})\,] < 2.67 \\
\wedge\ & [\,(\text{窓}) \overset{\text{ガラス}}{\longleftrightarrow} (\text{ベランダ}) \overset{\text{ガラス}}{\longleftrightarrow} (\text{窓})\,] < 1.46 \\
\wedge\ & [\,(\text{窓}) \overset{\text{ガラス}}{\longleftrightarrow} (\text{居室})\,] \geq 1.86 \\
\wedge\ & [\,(\text{水回り}) \overset{\text{ドア}}{\longleftrightarrow} (\text{居室})\,] < 1.71 \\
\wedge\ & [\,(\text{収納}) \overset{\text{収納}}{\longleftrightarrow} (\text{居室})\,] \geq 1.79 \\
\wedge\ & [\,(\text{収納}) \overset{\text{収納}}{\longleftrightarrow} (\text{居室}) \overset{\text{ドア}}{\longleftrightarrow} (\text{廊下})\,] \geq 1.24 \\
\wedge\ & [\,(\text{収納}) \overset{\text{収納}}{\longleftrightarrow} (\text{居室}) \overset{\text{ガラス}}{\longleftrightarrow} (\text{窓})\,] \geq 1.89 \\
\wedge\ & [\,(\text{収納}) \overset{\text{収納}}{\longleftrightarrow} (\text{玄関})\,] < 1.00
\end{aligned}
$$

図 2: モデル $f_{b/m+g}$ から得られたルールの例

6 まとめ

本論文では，賃貸物件の賃料に大きな影響を与える部屋配置を特定することを目的に，単一グラフにおける部分グラフの支持度に関する新たな定義を与えると共に，基本的な属性から得られる賃料推定結果を間取り情報に基づき補正する二段階賃料推定手法を提案した．また実験により，提案手法が推定精度の向上に有効であることを確認すると共に，初期的な結果ではあるが，注目すべき部屋配置の特定を行った．

今後の課題としては，多様な部屋配置を考慮するためにより低頻度な頻出部分グラフを利用することがあげられる．加えて，より効果的に賃料への影響が強い部屋配置を特定するため，傾向スコア分析 [23] や対比集合抽出 [24]，樹状モデルアンサンブル分析手法 [25] 等を併用することを予定している．

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．本研究の一部は，JSPS 科研費 17K00315 の助成を受けたものです．

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，筑波大学社会工学コモンズ 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] A. J. Smola and B. Schoelkopf : A tutorial on support vector regression, NeuroCOLT2 Technical Report NC2-TR-1998-030, 1998.

[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 $100m^2$ 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] 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.

[8] 尾崎 知伸，小黒 淳斗：頻出部分グラフを用いた賃料分析，人工知能学会第 111 回知識ベースシステム研究会，pp.13-16, 2017.

[9] D. J. Cook and L. B. Holder eds. : *Mining Graph Data*, Wiley-Interscience, 2005.

[10] T. Washio, J. N. Kok and L. DeRaedt eds.: *Advances in Mining Graphs, Trees and Sequences*, IOS Press, 2006.

[11] Z. Zeng, J. Wang, J. Zhang, and L. Zhou: FOGGER : An Algorithm for Graph Generator Discovery, *Proc.* of 12th International Conference on Extending Database Technology, pp.517–528, 2009.

[12] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone : *Classification and regression trees*, Wadsworth, 1984.

[13] J. R. Quinlan : Learning with Continuous Classes, *Proc.* of the 5th Australian Joint Conference on Artificial Intelligence, pp.343-348, 1992.

[14] 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.

[15] M. Kuramochi and G. Karypis : Finding frequent patterns in a large sparse graph, *Data Mining and Knowledge Discovery*, Vol.11, No.3, pp.243–271, 2005.

[16] B. Bringmann and S. Nijssen : What is Frequent in a Single Graph? *Proc.* of the 5th International Workshop on Mining and Learning with Graphs, pp.1–4, 2007.

[17] K. Kamihori, T. Shimano, A. Oguro and T. Ozaki: Towards discovery of human's recognition mechanisms for complex structured images, Proc. of the Third International Workshop on Skill Science, pp.26–27, 2016.

[18] X. Yan and J. Han : gSpan : Graph-based Substructure Pattern Mining, *Proc.* of the 2002 IEEE International Conference on Data Mining, pp.721–724, 2002.

[19] T. Chen and C. Guestrin : XGBoost: A Scalable Tree Boosting System, *Proc.* of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.785–794 , 2016

[20] L. Breiman : Random Forests, *Machine Learning*, Vol.45, Issue 1, pp.5–32, 2001.

[21] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu : LightGBM: A Highly Efficient Gradient Boosting Decision Tree, Advances in Neural Information Processing Systems 30, 2017.

[22] S. Hara and K. Hayashi : Making Tree Ensembles Interpretable: A Bayesian Model Selection Approach *Proc.* of the 21st International Conference on Artificial Intelligence and Statistics, pp.77–85, 2018.

[23] 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.

[24] S. D. Bay and M. J. Pazzani : DetectingChange in Categorical Data: Mining Contrast Sets, *Proc.* of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.302–306, 1999.

[25] H. Deng : Interpreting Tree Ensembles with inTrees, International Journal of Data Science and Analytics, 2018.