HUB-FS Working Paper Series

FS-2019-J-002

中古マンションのプライシングモデルのための データクレンジング法

大槻 健太郎

一橋大学大学院経営管理研究科

横内 大介

一橋大学大学院経営管理研究科

First version: 2019 年 11 月 11 日

中古マンションのプライシングモデルのためのデータクレンジング法
A data cleansing method for the pricing model of the second hands properties database

Kentaro OTSUKI: Hitotsubashi University Business School 大槻 健太郎 *
Daisuke YOKOUCHI: Hitotsubashi University Business School 横内大介 †

This paper focused on a method of data cleansing for used condominium database that is open to the public by the Ministry of Land, Infrastructure, Transport and Tourism. By identifying and removing error records, which are not transaction outliers, it enables to analize this dataset appropriately. In this study, we proposed two methods to deal with serious error, that is suspect of "Fat finger error". In addition, as a result of noise removal, the analysis results using the hedonic model have improved significantly.

keywords: Data Cleansing, Outlier Detection, Residential Property, Property price

1 はじめに

2006 年からスタートした国土交通省の " 土地総合情報システム (1) " は WEB 上で公開されているデータベースであり，取引価格を含む不動産売買の記録を蓄積している（以下，国交省データベース）．このオープンな国交省データベースを利用すれば，ある地域の宅地やマンション価格の相場を調べたり，不動産市場の動向を把握したりと目的に応じた様々な分析が可能である．

我々の目的は国交省データベースに収録された実際の取引価格による中古マンションのプライシングモデルの推定である．ただし，国交省データベースに蓄積された情報は記入式のアンケートであるため，収録データからモデルの推定を行う際はいくつか注意すべき点がある．

まず不動産購入者から調査票を回収できない割合が相当数あるとされ，実際に行われた取引の全てを網羅しているわけではない点．そのためモデル推定を行う際はサンプルバイアスの可能性を認識する必要がある．

また，収録された取引情報についても間取りや建築年などの物件属性の一部が無回答だったり，桁の間違いを疑う異常値が含まれていたりする点が挙げられる．特に取引価格の桁間違いが疑われる異常値は，様々な取引要因から生じた外れ値とは全く性質が異なり，適切なプライシングモデルの推定を大きく阻害してしまう．

これまで国交省データベースから推定した不動産のプライシングモデルを用いた研究は Shimizu and Nishimura[1)]，唐渡 [2)]，早川・田島 [3)] など多くの例が挙げられる．一方，データセットに含まれる異常値の取り扱いは多くの場合，分析者に委ねられており，その取り扱いを示している例は少ない．しかしながら桁間違いによる異常値と通常の取引から生じた外れ値を識別することはそれほど単純ではない．また東京 23 区で取引された中古マンションの蓄積データを人間が 1 件ずつ確認することもあまり実際的とは言えない．したがってこうした異常値を効率的かつ汎用的に識別する方法を検討することはより正確なプライシングモデルの推定に有用である．

そこで本研究では国交省データベースに収録された中古マンションの取引を対象としてその売買データの記録状況と扱い方を整理したのち，桁間違いと考えられる異常値を半自動的に検出する手法を提案する．また本提案手法でクレンジングを施し，整備したデータセットを用いてプライシン

* 一橋大学大学院経営管理研究科博士課程
† 一橋大学・経営管理研究科

グモデルを推定した結果も示す.

2　先行研究

まず本研究ではデータに関連する用語を一般の用法に関わらず，次の意味で用いる.

“ データ ”はある現象を観察・調査して得られた結果を数値や文字列などで記述したもの.

“ データベース ”は収集した情報の検索や蓄積を容易に行うため，組織化して収録したデータの集まり.

“ データセット ”は何らかの分析を目的として複数物件のレコードを集め，各変量をカラムとした表形式を指す.

“ レコード ”とは一つの物件の取引における価格および物件属性の組を指す.

“ フィールド ”とはレコードに記された価格と物件属性それぞれのインプットを指す.

桁間違いなどによる異常値の扱いはデータセットの外れ値に対する研究の中で扱われてきた. Barnett and Lewis[4] は外れ値の発生原因をそれぞれ，固有の要因による自然発生，不適切な計測手法（丸め誤差，収録ミスを含む）による計測エラー，不完全なデータ集計によるバイアスの 3 つに分類している.

一般個人や企業から収集したデータに含まれる外れ値については Chambers[5] が“ 代表性のある外れ値 ”と“ 代表性のない外れ値 ”という 2 つのタイプに分類し，前者は正しく収録された値で単一に存在するとは考えられない標本，後者は誤った値を持っており，何らかの意味で単一に存在する標本であるとし，代表性のある外れ値を含む場合のロバストなモデリングを紹介している.

不動産データの外れ値については Krause and Lipscomb[6] が取得した生データを整備し，分析可能なデータセットにする一連の手続きを“ Data Prepatation Process ”とし，データセットの統合やクレンジングの考え方を包括的にまとめた. 彼らはデータ測定のミスやデータベースへの誤ったインプットによる外れ値をデータエラーと定義し，通常の取引から生じる外れ値とは区別している. ただし，こうしたデータエラーの発生や影響の重大さを事前に把握できている場合は少なく，その識別は必ずしも明解で単純なプロセスではないと述べている.

3　データセットと内在するエラー

国交省データベースに収録された中古マンション取引のデータセットは表 1 の項目である. このデータセットに対するデータクレンジングのポイントを Krause and Lipscomb[6] を参考として紹介する.

欠損値 (Missing Data)

都道府県，市町村，取引価格，面積，取引時点以外の項目にはしばしば欠損値が存在する. 駅名や地区名などはほぼ網羅されている一方，改装有無や用途などは相対的に欠損値が多い.

ラベル付け (Labeling)

質的変数の項目は分析目的によって各レコードを識別し，分類するためのラベルとなる. たとえば用途や今後の利用目的という項目には“ 事務所 ”や“ 店舗 ”という事業性の用途が表記されたレコードがあり，居住用のレコードと区別することができる. また取引の事情等には“ 調停・競売等 ”や“ 関係者間取引 ”などと記載されたレコードが存在し，通常の価格形成と異なる取引要因があったものとして識別できる. また. 最寄り駅までの距離は徒歩 30 分未満は分単位だが，それ以降は 30 分毎に区分され，質的変数として収録されている. また建築年は 1945 年以前の物件では“ 戦前 ”という値が収録されている. これらは研究のデザインに応じて変数の変換やレコードの除去などが必要となる.

変数名	単位	例
種類		中古マンション等
都道府県名		東京都
市区町村名		世田谷区
地区名		赤堤
最寄駅. 名称		松原 (東京)
最寄駅. 距離	分	4
取引価格	円	27,000,000
間取り		１ＤＫ
面積	m^2	40
建築年		平成 15 年
建物の構造		ＲＣ
用途		住宅
今後の利用目的		住宅
都市計画		第１種低層住居専用地域
建ぺい率	%	50
容積率	%	100
取引時点		2010 年第３四半期
改装		未改装
取引の事情等		調停・競売等

表 1

データエラー (Data Error)

データセットにはいずれか，あるいは複数のフィールド値に何らかのミスが疑われるレコードが存在する．たとえば中古物件にも関わらず，取引時点より建築年が後のレコードや面積 35m^2 で間取りが４ＬＤＫという，広さと間取りが整合的でないレコードが含まれる．

本研究で扱う数値の桁間違いが疑われる異常値の検出は上記のデータエラーに該当する．具体的な例を挙げると，次の表 2 の物件 A，物件 B は一見してともに高級住宅街にある超高級物件に思える．実際に不動産売買仲介会社の WEB サイトを参考とすれば，最寄り駅が六本木で面積 150m^2 以上の築浅，駅至近物件であれば数億円の価格付け事例もあり，物件 B の価格はあり得る水準と判断できる．一方，物件 A については面積 160m^2 ではあるが築年数が 26 年を経ている．何らかの事情により実際に 12 億円で取引された可能性は否定できないものの，通常であれば一桁少ない 1 億 2 千万円の取引と考える方が妥当である．

属性	物件 A	物件 B
所在地	杉並区浜田山	港区六本木
最寄り駅	西永福	六本木 1 丁目
面積	160m^2	170m^2
駅までの距離	徒歩 10 分	徒歩 4 分
築年数	26 年	4 年
構造	RC	SRC
間取り	3LDK	3LDK
取引価格	12 億円	5 億 2 千万円

表 2

適切なプライシングモデルの構築において，こうした桁間違いによる異常値を含むレコードは除去が必要であるが，異常値か否かの判断は明確な線引きが難しい．たとえば価格が 3 億円の中古マンション取引のレコードは通常なら桁間違いの可能性を疑うが，都心の高級住宅街であれば十分あり得る水準となる．よって単純に取引価格の閾値を設定し，異常値と判定したレコードを除去する方法では実際の取引から生じた外れ値のレコードまで除外してしまうため，目的に適した方法とは言えない．

また人間がそれぞれの物件を目視して確認し，選別するという方法も効率的とは言えない．東京都 23 区の中古マンションデータを分析する場合，国交省データベースから期間 10 年分以上のデータを取得できるため，10 万件以上のレコードを確認しなくてはならない．また特定の地域だけのデータセットを分析の対象とする場合においても，人間が確認する方法では異常値を見落とす可能性がある．

そこでこれらの要件を踏まえて桁間違いの異常値を漏れなく効率的に検出するため，人間による判断と同様に地域の相場観を考慮し，異常値を含むレコードを抽出する機械的な手法を考案する．

なお数値に桁間違いが発生する可能性としては量的変数の取引価格，面積，最寄駅までの距離，築年数であるが，このうち面積は登記簿に記載されている専有部分の床面積 (m^2) の値であり，アンケートから値が収録されることはない．また築年数は取引時点と建築年の差から換算するため，桁間違いは生じない．最寄り駅までの距離についても距離（m）ではなく徒歩での所要時間（分）のため，数値の桁間違いは考えにくい．また距離が徒歩 30 分以上の場合は一定の幅で区切られた質的変量となるため，この誤りは生じない．結果として桁間違いが発生する可能性が高い変数は取引価格と判断できる．

取引価格のエラーは値の絶対値が大きい場合とは限らず，単身者向けの $20m^2$ 程度のマンションに 1 億円以上の価格が付くケースも含まれる．このため取引価格そのものではなく，m^2 当たりの取引単価の値に対して桁間違いを疑う異常値を検出する．

さらに取引価格の桁間違いは桁が大きい方への異常値と小さい方への異常値が想定されるが，プライシングモデルの推定に与える影響は前者の方がはるかに大きいため，本研究では桁が大きい方への異常値の識別に焦点を当てた．

4　クレンジング手法

まずデータセットに含まれる極端な値を持つ取引価格が実際の取引から生じた外れ値か桁間違いによる異常値かを識別する問題について，もし人間が目視で分類する際は各地域ごとの相場観を考慮して取引価格が異常値なのか否かを判断すると考えられる．実際に先に挙げた西永福の物件 A と六本木一丁目の物件 B に対し，最寄り駅ごとの物件グループを構築して桁間違いによる異常値を判定する例を示す．

図 4.1 はデータセットから特定の駅を最寄りとする物件グループを取り出し，その取引価格を縦軸にプロットしたもので左図は最寄り駅が京王井の頭線の西永福駅のグループ，右図は最寄り駅が東京メトロ南北線の六本木一丁目駅のグループである．これを見ると，明らかに西永福駅には 1 件の極端な外れ値が存在し，桁間違いの可能性が強く疑われる一方，六本木一丁目駅には数件の大きな値はあるものの，これらがすべて桁間違いを起こしているとは考えにくい．よって六本木一丁目には一定の高額物件が存在し得る地域性である，という判断が下せる．このようにして西永福の極端な外れ値が桁間違いによる異常値と判断する．

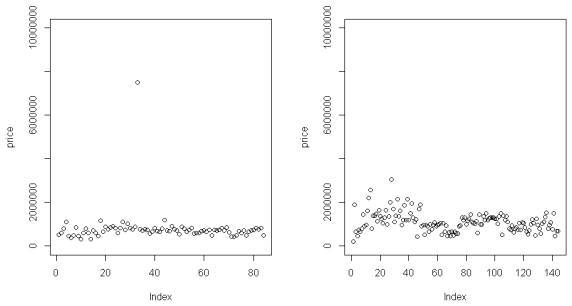

図 4.1

[手法 1] $SIQR_u$ による異常値検知

一般にある標本に含まれる観察値に対して外れ値か否か判定する際は標本全体を平均 0，分散 1 に標準化し，標準偏差を基準に当該観察値が中心から離れている度合いを計測する方法がよく用いられる．しかしながら，この方法は著しく大きな外れ値が混入した場合は平均や分散の計算に大きな影響を及ぼすというマスク効果が知られており，今回のように桁間違いによる異常値を含む場合は正しく検出できない可能性がある．より頑健な方法として Tukey[7] が紹介した箱ひげ図の四分位範囲幅 (IQR) を用いて異常値を検出する方

法もあるが，中古マンション価格の分布は一般に右に歪んだ形が多く，実際に高額で取引された外れ値と桁間違いによる異常値を識別できない可能性がある．また，野呂・和田 [8] は単峰だが非対称な分布に含まれる外れ値検出のため，対数変換を施した後に四分位範囲を用いる方法を提案した．ただし，この方法は事前に分布形についてある程度既知である必要であり，今回の最寄り駅毎の物件グループのようにそれぞれ分布形が異なる場合は適さない．

以上の理由から我々は kimber[9] が提案した $\mathrm{SIQR_u}$ を用いた．$\mathrm{SIQR_u}$ はサンプルの第 3 四分位からメディアンまでの範囲であり，サンプルの値を小さい順に並べて $100\alpha\%$ の点を $\mathrm{Q}_{(\alpha)}$ と示すと $\mathrm{SIQR_u} := \mathrm{Q}_{(0.75)} - \mathrm{Q}_{(0.5)}$ と定義され，サンプルが上方と下方に歪みのある非対称性な分布に強いとされている．この特徴を生かし，以下の基準で各最寄駅ごとのグループに適用し，地域ごとの相場観を勘案して桁間違いによる異常値の識別を行った．

$$\mathrm{Q}_{0.75} + k \cdot \mathrm{SIQR_u}$$

ここで k は正の定数である．桁が大きい方への間違いによる異常値を識別するため，$k = 10$ を採用した．

[手法 2] 階層的クラスタリングによる異常値検知

手法 1 のように m^2 当たりの取引単価のみに着目した単変量の検知手法とは異なり，他の量的変数である最寄り駅までの距離，築年数が取引価格に与えている影響を考慮し，階層的クラスタリングを用いて桁間違いが疑われる異常値を検出する．階層的クラスタリングは複数の変量ごとの距離に着目するため，変数間のバランスが悪い異常値検出が可能である他，結果の解釈もしやすい．

通常，階層的クラスタリングを用いる際も各変量の間で値の水準が大きく異なる場合はそれぞれの変量に対して標準化を施してから距離の計測を行うが，これは手法 1 で説明したようにマスク効果が生じ，異常値の検出力が鈍る可能性がある．より頑健な基準化を行うため，中心化のために中央値，尺度の調整には $\mathrm{SIQR_u}$ を用いた．

$$\frac{x_i - \mathrm{Q}_{0.5}}{c \cdot \mathrm{SIQR_u}}$$

分母の c は正の定数であり，$c = 2.5$ とした．これは Tukey[7] が提案した箱ひげ図の上側のひげの位置を参考とし，第 3 四分位から 1.5 倍の $\mathrm{SIQR_u}$，つまり中央値から 2.5 倍の $\mathrm{SIQR_u}$ までの範囲を正常な値の取るレンジと設定したためである．また階層的クラスタリングは金森・竹ノ内・村田 [10] を参考とし，統計ソフトウェアの `R` における `hclust` 関数を用いて Ward 法によって行い，桁間違いによる異常値か否かの判定は枝の長さ 4 を基準として判断した．

5 本クレンジング手法の適用結果と考察

手法 1 と手法 2 を用いて桁間違いによる異常値の検出力を比較するため，国交省データベースから 2005 年第 3 四半期から 2017 年第 3 四半期までに取引された東京 23 区の中古マンション等のデータセットを抽出し，小節 3 で説明したデータクレンジングのポイントに基づいて整備を行った．具体的には手法 1 ないし手法 2 を用いた異常値検出を行う際に，最寄り駅・名称，m^2 当たりの単価，最寄り駅までの距離，築年数に欠損が生じているレコードや建築年に “ 戦前 ” と記載あるレコードは不備があると判断し，データセットから除去した．また手法 2 で距離を計測する量的変数に用いるため，最寄り駅までの距離が 30 分未満のレコードのみ使用した．さらに用途や今後の目的に事業向け用途が記載されているレコードは住宅用物件とは属性が異なると考えて除外した．これらの処置を 129,264 件のデータセットに施した

ところ，比較検証に用いるレコード数は 101,405 件となった．

それぞれの手法を用いた識別結果を実際に収録されている物件例を用いて説明する．表 3 にはそれぞれの物件の属性と人間の判断，手法 1 の $SIQR_u$ を用いた検出手法，手法 2 の階層的クラスタリングを用いた検出手法の識別結果を示した．

まず，前節で例に取り上げた物件 A および B についてはいずれの手法でも人間の判断と同じように桁間違いによる異常値を抽出できた．実際に図 5.1 と図 5.2 は物件 A と B に手法 1 と手法 2 を用いた結果をそれぞれ表している．最寄り駅がそれぞれ西永福駅（いずれも上図）と六本木一丁目駅（いずれも下図）のデータであるが，西永福駅のデータは桁間違いによる異常値を疑う 1 レコードだけが飛び出している一方，六本木一丁目駅のグループには特段にとびぬけたレコードがなく分布しており，それぞれの駅ごとのレコードを目視で分類した場合と同様に上手く抽出されていることが分かる．

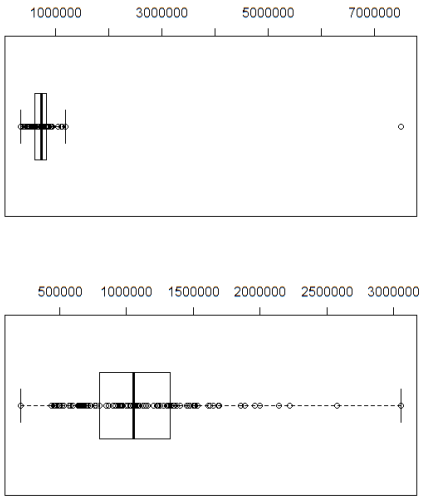

図 5.1

また物件 C は単身者向けの 20m^2 程度のマンションが 1 億円以上の価格で取引されるケースとして中延の物件を取り上げた．その結果，m^2 当たり取引単価が 9 百万円と極めて高額であり，いずれの手法でも桁間違いによる異常値として識別できた．

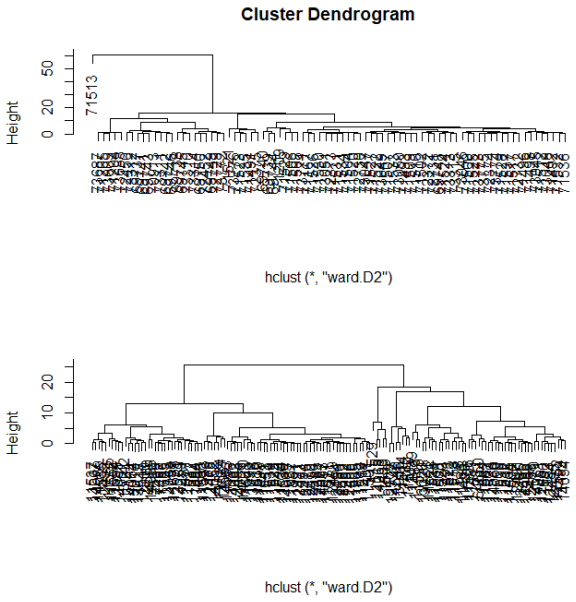

図 5.2

次に物件 D と E は物件 A や C ほど極端な m^2 当たり単価ではない物件例を取り上げた．手法 1 は人間の判断と同じように異常値と識別した一方，手法 2 では異常値に分類しなかった．階層的クラスタリングの枝の長さの設定にもよるが，概して階層的クラスタリングは $SIQR_u$ を用いた方法に比べ，桁間違いによる異常値が疑われるレコードを見逃すケースがあった．

一方，物件 F については他の物件と比べて明快ではないが，人間の判断では異常値と分類した．これに対し，手法 1 の $SIQR_u$ は異常値と認識せず，手法 2 の階層的クラスタリングが異常値と分類する結果となった．物件 F を人間が判断する場合，m^2 当たり取引単価だけでなく，築年数なども考慮して桁間違いによる異常値か否かを判断するが，こうした複合的な要素があるレコードに対しては単変量で識別する手法 1 ではなく，多変量で識別する手法 2 の分類性能が上手く機能する．

属性	物件 A	物件 B	物件 C	物件 D	物件 E	物件 F
所在地	杉並区浜田山	港区六本木	品川区中延	文京区関口	豊島区大塚	豊島区高田
最寄り駅	西永福	六本木 1 丁目	中延	江戸川橋	新大塚	早稲田 (都電)
延床面積	160m^2	170m^2	20m^2	85m^2	15m^2	30m^2
駅までの距離 (徒歩・分)	10	4	6	4	5	4
築年数（年）	26	4	10	15	12	33
構造	RC	SRC	RC	RC	RC	RC
間取り	3LDK	3LDK	1K	3LDK	1K	1DK
取引価格	12 億円	5 億 2 千万円	1 億 8 千万円	6 億円	73 百万円	85 百万円
人間の判断	誤	正	誤	誤	誤	誤
手法 1	誤	正	誤	誤	誤	正
手法 2	誤	正	誤	正	正	誤

表 3

識別結果から m^2 当たり取引単価のみでほぼ判断できる桁間違いによる異常値検出には手法 1 の $\mathrm{SIQR_u}$ が適しており，築年数なども考慮した異常値を検出する際は手法 2 の階層的クラスタリングが補完的に機能することが分かった．一方，手法 2 は各レコードの変量間の距離で分類を行っているため，数値の桁を大きい方に間違えている異常値だけでなく，極端に桁が小さい取引価格のレコードも異常値に識別しているケースが散見される．なお本研究の目的を考えれば，分析するデータセットからプライシングモデルに悪影響を与える桁間違いによる異常値を可能な限り除去し，適切なデータセットを整えられる手法が最も望ましいため，手法 1 と手法 2 を組み合わせ，いずれか一方の手法で誤収録と判断されたレコードを異常値を分類するほうが理に適う．

これら各手法の処理を最寄り駅ごとにグループ化したデータセットにそれぞれ適用したところ，処置前 101,405 件のレコードに対して手法 1 は 198 件，手法 2 は 172 件，手法 1 と手法 2 の両方を適用した場合は 236 件の異常値を検出した．

6 ヘドニック分析を通じた本クレンジング手法の妥当性の検証

本節では検出した桁間違いを疑う異常値レコードを元のデータセットから除去した効果がプライシングモデルのフィッティングに与える影響を示す．このため Shimizu and Nishimura[1] など多くの論文で用いられている標準的なプライシングモデルとしてヘドニックモデルで検証を行う．推定に用いたヘドニックモデルの説明変数は以下のとおり．

$$P = \beta_0 + \sum_{i=1}^{3} \beta_i x_i + \sum_{l=1}^{469} \gamma_l z_l + \epsilon$$

P: マンション取引価格
x_i: 延床面積（m^2），最寄駅までの距離 (分)，築年数（年）
z_l: 最寄り駅ダミー

異常値レコードの除去前後でヘドニックモデルを適用した結果を表 4 に示す．いずれの手法でも取引価格の上方に位置する極端な値が除去され，切片項の係数が緩やかになった他，決定係数は 20% 以上の改善が見られた．

7 結論

本研究では国交省データベースの中古マンショ

変数	処置前		手法 1		手法 2		手法 1+ 手法 2	
	係数	標準誤差	係数	標準誤差	係数	標準誤差	係数	標準誤差
切片	-6,623,775	(1,090,939)	-5,086,395	(613,661)	-5,118,729	(673,909)	-5,007,131	(611,263)
延床面積	759,817	(3,138)	725,807	(1,767)	728,970	(1,942)	724,692	(1,761)
最寄り駅距離	-502,057	(20,042)	-475,760	(11,272)	-477,148	(12,398)	-475,522	(11,236)
築年数	-491,624	(6,818)	-499,504	(3,835)	-500,467	(4,216)	-500,803	(3,821)
（駅ダミーは省略）								
観察数 (n)	101,405		101,207		101,233		101169	
調整済 R^2	0.482		0.732		0.695		0.733	

表 4

ン価格データセットを整理し，桁間違いが疑われる取引価格の異常値を半自動的に識別するための手法を提案した．この手法により，プライシングモデルの推定に悪影響を与えるレコードを除去し，適切なデータセットの構築が可能となった．

この手法を 10 万件以上ある東京都 23 区の中古マンション価格データセットに処置し，ヘドニックモデルを適用して処置の前後で比較した結果，決定係数が 20 %以上向上するなど，顕著な改善が見られた．

また今回の手法による異常値の適切な除去は，ヘドニックモデルのような標準的なプライシングモデルの性能向上に効果があるだけでなく，機械学習によるプライシング AI に頻発する過適合問題に対しても有効な対策となりうるだろう．

なお今回はプライシングモデルの推定に大きく影響する桁が大きい方への間違いが疑われる異常値を対象としたが，桁が小さい方への間違いに対しても本研究で提案した手法を適用できる．しかしながら桁が小さい方への間違いが疑われる異常値は事故などの訳有り物件や著しい使用劣化により成立した低価格との差異を十分に識別する情報がなく，除去すべきか否かは分析目的にも依存するため，その取り扱いは今後の研究課題とする．

注

(1) 国土交通省土地総合情報システム「不動産取引価格情報ダウンロード」http://www.land.mlit.go.jp/webland/download.html(2018 年 4 月 4 日閲覧)

参考文献

1) Shimizu, C. and Nishimura, G.N. (2007), “ Pricing Structure in Tokyo Metropolitan Land Markets and its Structural Changes:Pre-bubble, Bubble, and Post-bubble Periods ”, *Journal of Real Estate Financial Economics*, 35, pp.475-496.
2) 唐渡広志 (2016),「ヘドニック・アプローチを利用した不動産価格指数の推定方法とその問題点」, 都市住宅学第 92 号, pp.17-20
3) 早川季歩・田島夏与 (2017),「都心高額住宅地の成立条件：東京区における中古マンション等取引価格情報を用いた実証分析」, 都市住宅学第 99 号, pp.96-101.
4) Barnett, V. and Lewis, T., (1994), *Outliers in Statistical Data*, John Wiley and Sons.
5) Chambers, R.L., (1986). “ Outlier Robust Finite Population Estimation ”. Journal of the American Statistical Association, 81, pp.1063-1069.
6) Krause, A. and C. Lipscomb. (2016). “ The Data Preparation Process in Real Estate: Guidance and Review. ”. Journal of Real Estate Practice and Education, 19(1), pp.15-42.
7) Tukey, J.W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
8) 和田かず美（2015）「統計実務におけるレンジチェックのための外れ値検出方法」, 統計研究彙報第 72 号, 総務省統計研修所, pp.41-54.
9) Kimber, A. C. (1990). “ Exploratory Data Analysis for Possibly Censored Data from Skewed Distributions, ”. Applied Statistics, 39, pp.2130.
10) 金森敬文, 竹ノ内高志, 村田昇（2009）『R で学ぶデータサイエンス　パターン認識』, 共立出版.