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

大槻 健太郎

OTSUKI Kentaro

一橋大学大学院経営管理研究科博士課程

School of Business Administration, Hitotsubashi University

横内 大介

YOKOUCHI Daisuke

一橋大学・経営管理研究科

School of Business Administration, Hitotsubashi University

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 us to analyze 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

データクレンジング、外れ値検知、居住用不動産、不動産価格

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、実際に売買された不動産取引の記録をその売買価格も含めて蓄積している（以下、国交省データベース）。

我々の目的は国交省データベースを用いて実際の取引価格から中古マンションプライシングモデルを構築することである。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、収録されたデータからモデルの構築を行う際は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、売買さ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の際はサンプルバイアスが生じている可能性を認識しておく必要がある。

つぎに収録された取引情報は間取りや建築年などの物件属性の一部が無回答であったり、桁の間違いを疑う異常値の混入があったりするため、適切なデータクレンジングを必要とする点が挙げられる。

特に取引価格の桁間違いによる異常値は、様々な取引要因から生じた価格の外れ値とは全く性質が異なり、実際の取引価格ではない。こうした異常値の混入は実際の

取引価格を用いた適切なプライシングモデルの構築を阻害する。ただし、桁間違いによる異常値と実際の取引から生じた外れ値を識別することは必ずしも単純ではなく、東京23区の数万件にもおよぶ中古マンションのレコードを人間が1件ずつ確認する作業も効率が良いとはいえない。

これまで国交省データベースを用いて不動産のプライシングモデルを推定している研究はShimizu and Nishimura¹⁾、唐渡²⁾、早川・田島³⁾など多くの例が挙げられる。ただし、データセットに含まれる異常値の取り扱いには分析者に委ねられており、その取り扱いを詳しく示している例はあまり多くない。

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取引を対象にその売買データの記録状況と扱い方を整理し、価格の桁間違いが疑われる異常値を機械的に検出する手法を提案する。また本提案手法でデータクレンジングを施したデータセットを用いてプライシングモデルを推定した結果も示す。

2 先行研究

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。

⁽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)

“データ”はある現象について観察・調査して得られた結果を数値や文字列などで記述した記録。

“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。

“レコード”とは1つの物件の取引における価格および物件属性の組。

“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。

“観測値”とはレコードに記された価格、その他物件属性についての数値または文字列。

“異常値”は実際に取引された情報とは異なる観測値。

“外れ値”は異常値以外に実際の取引情報に基づく観測値も含め、ある変量において他の多くの値から大きく逸脱した値。

桁間違いなどによって生じる異常値はこれまでデータの外れ値を研究する分野で取り扱われてきた。Barnett and Lewis⁴⁾ 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Chambers⁵⁾ 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を持ち、単独で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、何らかの意味で単独で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。Krause and Lipscomb⁶⁾ は不動産の生データを取得したのち分析可能なデータセットに整える一連の手続きを“Data Preparation Process”と呼び、データセットの統合や外れ値のクレンジングを包括的に紹介している。

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の項目からなる。このデータセットに対してKrause and Lipscombを参考にデータクレンジングのポイントを紹介する。

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。駅名や地区名に比べ、改装有無や用途は相対的に欠損値が多い。

ラベルによる分類 (Labeling)

質的変数の項目は分析目的によってレコードを識別し、適切に分類するラベルとなる。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”という事業性用途のレコードがあり、居住用のレコードと区別できる。

表1 国交省データセットの中古マンション等のレコード例

変量名	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		平成22年第3四半期
改装		未改装
取引の事情等		調停・競売等

取引の事情等は“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、その他の売買と異なる経緯で取引されたレコードとして識別できる。最寄り駅までの距離は徒歩30分未満は分単位であり、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。建築年は1945年以前の物件では“戦前”と収録されている。このように適切に分類されたラベルに基づき、研究デザインによって変数の変換やレコード除去を行い、データセットを整備する。

データエラー (Data Error)

データセットには物件属性のいずれか、あるいは複数の観測値に何らかのミスの混入を疑う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になるレコードや面積と間取りが不整合なレコードが含まれる。

本研究ではデータエラーに該当する価格の桁間違いが疑われる異常値の検出を扱う。具体的な例を挙げると、表2の物件A、物件Bは一見すればともに高級住宅街にある超高額物件と思える。実際に不動産売買仲介会社のWEBサイトを参考とすれば、最寄り駅が六本木一丁目で面積150m²以上の築浅、駅至近物件には数億円の価格付けがされた事例も発見できるため、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

表2 高額物件の事例

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	六本木一丁目
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

160m²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、通常は1桁少ない1億2千万円の取引と考える方が妥当であろう。

人間が桁間違いによる異常値か否かを識別する際は各地域ごとの相場観を考慮して判断を下すと考えられる。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を最寄り駅を同じくする物件グループの中で比較して異常値を判定する例を示す。

図1はデータセットから最寄り駅を同じとする物件グループを取り出し、そのm²当たり取引単価をヒストグラムで表している。左図は西永福駅、右図は六本木一丁目駅である。西永福駅グループには明らかに1桁多い可能性のある極端な値が存在する。一方、六本木一丁目駅グループ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であるとは考えにくい。よって西永福の物件Aを桁が多い異常値と判定する。

先の例でも説明したように人間が各取引物件レコードから桁間違いを判断する際は同じ最寄り駅のグループの中で次の基準で判定すると想定した。

- ・m²当たり単価をヒストグラムで並べ、上側に他の観測値と著しく異なる値があれば桁間違いの候補とする。
- ・桁間違いの候補から1桁落とした値が妥当か判定する。この際、築年数20年以上経過している物件は特に桁が多い異常値の可能性が高い。
- ・シングル向けと想定される面積35m²未満の物件で取引価格が8,000万円以上する場合は桁が多い異常値の可能性が高い。
- ・地区名、建築年、最寄り駅からの距離が同じで似た間取りの属性を持つ物件レコードがあれば、築年数の経過を考慮しつつ、物件同士のm²単価の整合性をみる。
- ・東京都心5区（千代田区、中央区、港区、新宿区、渋谷区）の最寄り駅において、徒歩5分以内で面積が100m²を超えている場合は築年数によらず相当な

図1 最寄り駅が同じグループのm²当たり取引単価

高額物件でも桁間違いでない可能性がある。

- ・不動産売買仲介会社のWEBサイトで類似した属性を持つ物件を参考とする。

ただし、東京都23区の中古マンションレコードは期間10年で10万件以上存在するため、列挙した基準で人間が1件ずつ物件を確認して異常値を選別する作業は効率が良いとは言えない。

そこで本研究では人間が判定した桁間違いの異常値をより効率的に検出するため、人間が判断を行う場合と同様に地域の相場観を考慮し、最寄り駅を同じくするグループ内での比較によって異常値のレコードを検出する機械的な手法を考案する。この際、人間が桁間違いとした判定を正解とみなし、考案した手法を評価する。

桁間違いに対して注意を払うべき量的変量は、取引価格、面積、最寄り駅までの距離、築年数である。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(m²)の値であり、アンケートの値は収録されない。築年数は取引時点と建築年の差から換算しているため、桁間違いは生じない。

最寄り駅までの距離は距離(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。一方、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のため、この誤りは生じない。よって桁間違いが発生する可能性が高い変量は取引価格と判断できる。

それぞれの最寄り駅グループごとの価格分布の形は非対称性が高く、多くは分布の右側が歪んでいる。先に挙げた図1はその一例である。そこで本研究では桁間違いのうち、桁が多い異常値の識別に焦点を当てる。

取引価格の桁が多い異常値は絶対値の大きさだけでは識別できず、単身者向けの20m²程度のマンションに1億円以上の価格がついている場合もある。このため取引価格そのものではなく、m²当たりの取引単価によって桁間違いを検出する。

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、129,264件のレコードを取得した。このデータセットに小節の冒頭で説明したデータクレンジングのポイントに基づき、欠損あるレコード

や用途が住宅でないレコードなどを削除した結果、本研究の検証に用いるレコードの数は101、405件となった。このうち桁が多い異常値に対する人間の判定基準から正解レコードとして168件を検出した。表3にその一部を抜粋した。

4 本クレンジング手法

人間が桁の多い異常値を判定する際に地域性を考慮することから、本研究で提案する手法は最寄り駅ごとに分けたグループに対して適用する。

我々はこのグループに対して m^2 当たり単価の分布形から異常値を検出する手法と m^2 当たり単価以外の物件属性である最寄り駅までの距離、築年数を考慮して異常値を検出する手法の2つを考案した。

[手法1] SIQR_uによる単変量の異常値検知

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中心とする位置に平均、尺度に標準偏差を用いて、各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標準偏差の値に影響を及ぼすというマスク効果が知られており、桁間違いのような大きな異常値を含む場合は正しく検出ができないことが多い。

より頑健な方法としてTukey⁷⁾の箱ひげ図を応用した

四分位範囲を用いる方法がある。野呂・和田⁸⁾はTukeyの考え方を応用し、平均や分散によるマスク効果のある手法と四分位範囲による頑健な手法を比較実証している。

一方、本研究で対象とする中古マンション価格の分布は前の節でも解説したように非対称であり、外れ値は分布の上側の裾に頻出していることが特徴である。野呂・和田における非対称分布に対する外れ値検出の方法では、第1四分位 - (四分位範囲の長さ) × 1.724を下限値、第3四分位 + (四分位範囲の長さ) × 1.724を上限値として、その範囲外を外れ値としている。

つまり、第1四分位以下と第3四分位以上の標本は同一に分布しているという仮定がおかれている。本研究で用いる価格データの分布の特徴を踏まえると、分布の裾が同一であるという仮定は不都合であることから、本研究では四分位範囲の長さの代わりに(第3四分位 - 中央値)を用いることにより、価格分布の中央値より右側の形状を強く反映した外れ値検出を行う。

kimber⁹⁾が提案したSIQR_uはサンプルの値を小さい順に並べて100 α %の点を $Q_{(\alpha)}$ と示すと

$$SIQR_u := Q_{(0.75)} - Q_{(0.5)}$$

と定義される。このSIQR_uによる次の基準を各最寄り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。

表3 人間の判断に基づく桁間違いレコード (168件)

物件ID	市区町村名	地区名	最寄駅・名称	取引価格	最寄駅・距離(分)	築年数	面積	間取り	平米当たり単価	建築年	建物の構造	取引時点	改装
1446	千代田区	五番町	市ヶ谷	640,000,000	1	32	90	3LDK	7,111,111	昭和54年	SRC	平成23年第2四半期	改装済
1623	千代田区	一番町	半蔵門	260,000,000	1	28	45	2LDK	5,777,778	昭和52年	RC	平成17年第3四半期	
1682	千代田区	神田駿河台	御茶ノ水	85,000,000	5	32	15	1K	5,666,667	昭和50年	SRC	平成19年第1四半期	未改装
1686	千代田区	神田駿河台	御茶ノ水	82,000,000	1	28	15	1K	5,466,667	昭和52年	SRC	平成17年第4四半期	未改装
3697	中央区	日本橋小網町	茅場町	200,000,000	3	11	20	1K	10,000,000	平成15年	SRC	平成26年第4四半期	未改装
4759	中央区	月島	勝どき	210,000,000	6	39	45	2DK	4,666,667	昭和46年	SRC	平成22年第2四半期	改装済
4797	中央区	月島	月島	120,000,000	6	39	25	2K	4,800,000	昭和46年	SRC	平成22年第4四半期	改装済
5003	中央区	佃	月島	550,000,000	3	8	60	2LDK	9,166,667	平成14年	RC	平成22年第1四半期	未改装
5181	中央区	日本橋蛸薬町	水天宮前	160,000,000	3	6	20	1K	8,000,000	平成16年	SRC	平成22年第3四半期	未改装
5235	中央区	日本橋馬喰町	馬喰町	160,000,000	0	6	25	1K	6,400,000	平成16年	SRC	平成22年第4四半期	未改装
5239	中央区	日本橋箱崎町	水天宮前	240,000,000	3	30	55	3DK	4,363,636	昭和55年	SRC	平成22年第4四半期	未改装
5394	中央区	湊	新富町(東京)	220,000,000	5	10	35	1DK	6,285,714	平成12年	RC	平成22年第3四半期	未改装
5543	中央区	築地	築地	99,000,000	4	30	45	2DK	2,200,000	昭和56年	SRC	平成23年第1四半期	未改装
6330	中央区	月島	勝どき	110,000,000	6	12	15	1K	7,333,333	平成8年	RC	平成20年第1四半期	改装済
6896	港区	赤坂	赤坂(東京)	960,000,000	5	7	65	1LDK	14,769,231	平成20年	RC	平成27年第3四半期	未改装
7395	港区	海岸	竹芝	150,000,000	1	36	20	1K	7,500,000	昭和54年	SRC	平成27年第3四半期	改装済
7775	港区	港南	天王洲アイル	790,000,000	7	9	75		10,533,333	平成17年	RC	平成26年第4四半期	未改装
8221	港区	芝	三田(東京)	590,000,000	4	6	65	2LDK	9,076,923	平成18年	RC	平成24年第3四半期	未改装
9135	港区	白金台	白金台	1,400,000,000	4	35	150	2LDK	9,333,333	昭和55年	SRC	平成27年第4四半期	未改装
11042	港区	元麻布	麻布十番	3,200,000,000	7	13	160	4LDK	20,000,000	平成14年	RC	平成27年第4四半期	改装済
11098	港区	高輪	泉岳寺	600,000,000	5	39	75	3LDK	8,000,000	昭和46年	SRC	平成27年第3四半期	未改装
11751	港区	芝	三田(東京)	100,000,000	3	32	30	1DK	3,333,333	昭和54年	SRC	平成23年第1四半期	未改装

(中略)

94304	足立区	西保木間	竹ノ塚	180,000,000	22	5	55	2LDK	3,272,727	平成16年	RC	平成21年第3四半期	未改装
94801	足立区	島根	西新井	39,000,000	11	18	15	1R	2,600,000	昭和62年	RC	平成17年第3四半期	未改装
95245	足立区	保木間	竹ノ塚	70,000,000	19	14	50	3DK	1,400,000	平成6年	RC	平成20年第4四半期	未改装
95431	葛飾区	青戸	亀有	69,000,000	18	35	40	2DK	1,725,000	昭和55年	RC	平成27年第4四半期	未改装
97086	葛飾区	新宿	金町	110,000,000	18	19	55	2LDK	2,000,000	平成3年	SRC	平成22年第4四半期	未改装
97109	葛飾区	東金町	金町	110,000,000	9	26	40	2LDK	2,750,000	昭和59年	RC	平成22年第4四半期	未改装
97657	葛飾区	東金町	金町	270,000,000	14	12	95	4LDK	2,842,105	平成9年	RC	平成21年第4四半期	改装済
98022	葛飾区	宝町	お花茶屋	50,000,000	4	30	40	2DK	1,250,000	昭和52年	RC	平成19年第4四半期	未改装
98182	葛飾区	東立石	京成立石	430,000,000	12	16	55	3LDK	7,818,182	平成3年	RC	平成19年第4四半期	未改装
98251	葛飾区	堀切	堀切菖蒲園	62,000,000	10	18	35		1,771,429	昭和63年	RC	平成18年第4四半期	
98389	江戸川区	北葛西	西葛西	630,000,000	13	22	95	4LDK	6,631,579	平成6年	SRC	平成28年第1四半期	未改装
99695	江戸川区	船堀	船堀	500,000,000	1	24	70	3LDK	7,142,857	平成元年	SRC	平成25年第2四半期	未改装
100772	江戸川区	船堀	船堀	290,000,000	9	25	80	4LDK	3,625,000	昭和59年	RC	平成21年第4四半期	改装済
101050	江戸川区	西葛西	西葛西	58,000,000	15	16	25	1K	2,320,000	平成4年	RC	平成20年第1四半期	
101118	江戸川区	西小岩	小岩	360,000,000	10	9	75	3LDK	4,800,000	平成11年	SRC	平成20年第2四半期	未改装
101208	江戸川区	船堀	一之江	420,000,000	1	7	70	3LDK	6,000,000	平成13年	RC	平成20年第1四半期	未改装

$$Q_{(0.75)} + k \cdot \text{SIQR}_u$$

ここで k は正の定数である。本研究では経験的に得られた知見から $k=10$ を採用した。

〔手法2〕階層的クラスタリングによる多変量の異常値検知

m^2 当たりの取引単価だけでなく他の量的変量である最寄り駅までの距離、築年数も用いた階層的クラスタリングによって桁の多い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間の距離を反映できるため、築年数が古いあるいは最寄り駅までの距離が遠いにも関わらず m^2 当たりの取引単価が相対的に高い観測値の検出に役立つと考えられる。

階層的クラスタリングを用いる際に各変量の間で単位が異なるなど、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。より頑健な標準化を行うため、中心化のために中央値、尺度の調整には SIQR_u を用いた。

$$\frac{x_i - Q_{(0.5)}}{c \cdot \text{SIQR}_u}$$

分母の c は正の定数であり、 $c=2.5$ とした。これはTukey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 SIQR_u 、つまり中央値から2.5倍の 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経験的に枝の長さ4を基準として判断した。

5 提案手法の適用結果と考察

前節で紹介した手法により異常値を検出した結果を表4に示す。また手法1と手法2の一方が異常値と判定した場合および2つの手法の両方が異常値と判定した場合の結果も掲載した。

表4における的中率とは提案した手法が異常値と判定したレコードのうち、人間の判定でも異常値と判断したレコードの割合である。一方、カバー率は人間が異常値と判定したレコードのうち、提案手法が異常値と判定できた割合である。

この結果から手法1∪手法2はカバー率が1となり、人間の判定した異常値のレコードをすべて含む一方、的中率の高さでは手法1∩手法2が最も高くなることが分かる。

加えて各手法を適用した結果の個別例を表5に示した。

前節で例に取り上げた物件AおよびBはいずれの手法でも桁の多い異常値か否かの判定が人間の判定と一致した。図2は手法1、図3は手法2を用いた結果であり、いずれも上段が西永福駅を最寄り駅とするグループ、下段が六本木一丁目駅を最寄り駅とするグループである。西永福駅のグループは桁が多い異常値を疑う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、人間が判定した結果と同様に物件Aのみ異常値として検出した。

表4 各手法をデータセットへ適用した結果

属性	異常値検出率	的中率	カバー率
人間の判定	168	-	-
手法1	198	0.828	0.976
手法2	172	0.721	0.738
手法1∪手法2	236	0.712	1
手法1∩手法2	134	0.896	0.714

図2 手法1の適用結果

図3 手法2の適用結果。図中の数値はレコードの物件IDを表す。

表5 例示した個別物件の判定結果

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
最寄駅	西永福	六本木一丁目	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	7千3百万円	8千5百万円
人間の判定	誤	正	誤	誤	誤	誤
手法1	誤	正	誤	誤	誤	正
手法2	誤	正	誤	正	正	誤

物件Cは単身者向けの20m²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を取り上げた。この物件はm²当たり取引単価が9百万円と極めて高額であり、いずれの提案手法でも桁間違いによる異常値と識別し、人間の判定と一致した。

次の物件Dと物件Eは手法1が人間の判定と同じように異常値と識別した一方、手法2は異常値に分類しなかった。設定した枝の長さによる面もあるが、概して階層的クラスタリングはSIQR_uを用いた手法と比べ、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見受けられた。

物件Fは他の物件と比べて判定が難しいが、人間の判定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、m²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、複合的な観点が必要な場合に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ことが分かった。

なお手法2は各レコードの変量間の距離で識別を行っているため、数値の桁が多い間違いの異常値だけでなく、極端に桁が少ない間違いのレコードも異常値として識別するケースが散見される。

6 ヘドニック分析によるデータセットの改善効果

本節では桁の多い異常値レコードの除去がプライシングモデルのフィッティングに与える効果を示す。このためShimizu and Nishimura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。

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

P ：マンション取引価格

x_i ：延床面積 (m²)、最寄り駅までの距離 (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6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ほか、決定係数は20%以上改善した。

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機械的に識別するための手法を提案した。この手法によって人間の判定に近い異常値の検出が可能となった。

提案手法を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に適用し、標準的なプライシングモデルであるヘドニックモデルを推定した結果、処置前よりモデルの決定係数が20%以上向上するなど、顕著な改善が見られた。

また提案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に類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。

なお提案手法のいずれを用いるべきかという点に関しては、本研究の目的がプライシングモデルに悪影響を与える異常値を可能な限り除去することであるため、手法1と手法2のいずれか一方で異常値判定されたレコードを除去する方法を結論とする。ただし、引き続き改良を重ねることで正常な値を異常値と判定してしまう割合を

表6 ヘドニックモデルの適用結果（括弧内は標準偏差、駅ダミーは省略）

変量	処置前	人間の判断	手法1	手法2	手法1∪手法2	手法1∩手法2
切片	-6,623,775 (1,090,939)	-5,198,847 (635,185)	-5,086,395 (613,661)	-5,118,729 (673,909)	-5,007,131 (611,263)	-5,197,480 (676,042)
延床面積	759,817 (3,138)	728,995 (1,828)	725,807 (1,767)	728,970 (1,942)	724,692 (1,761)	730,080 (1,948)
最寄駅距離	-502,057 (20,042)	-477,979 (11,666)	-475,760 (11,272)	-477,148 (12,398)	-475,522 (11,236)	-477,394 (12,428)
築年数	-491,624 (6,818)	-500,900 (3,996)	-499,504 (3,835)	-500,467 (4,216)	-500,803 (3,821)	-499,180 (4,228)
観察数(n)	101,405	101,237	101,207	101,233	101,169	101,271
調整済R ²	0.482	0.720	0.732	0.695	0.733	0.694

減らし、的中率を向上させる余地はある。

今回は提案手法を桁が多い異常値の検出に適用したが、桁が少ない異常値に対しても応用できる。しかしながら桁が少ない異常値については事故などの訳有り物件と著しい使用劣化による低価格物件の差異を十分に識別する情報がなく、また除去すべきか否かは分析目的にも依存するため、検証は今後の研究課題とする。また国交省データセット以外のデータセットに対しても提案手法を適用し、汎用性の高さを検証することも引き続き課題とする。

謝辞

本研究を進める中で貴重なご意見をいただいた匿名査読者のみなさまに感謝を申し上げます。

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) 野呂竜夫、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.21-30.
- 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。

表7 既存研究である野呂・和田と本提案手法の比較（判定は1が異常値、0が正常値）

既存手法①	平均値と標準偏差に基づくレンジによる検出レコード 下限値 = 平均値 - 3 × 標準偏差 上限値 = 平均値 + 3 × 標準偏差
既存手法②	中央値と四分位範囲に基づくレンジによる検出レコード 下限値 = 中央値 - 2.224 × 四分位範囲 上限値 = 中央値 + 2.224 × 四分位範囲
既存手法③	分布の歪みを考慮したレンジによる検出レコード 下限値 = 第④分位 - 1.724 × 四分位範囲 上限値 = 第③分位 + 1.724 × 四分位範囲
既存手法④	常用対数変換後に手法③を適用した検出レコード
提案手法1	本文記載の方法による検出レコード 閾値 = 第③分位 + 10 × SIGMA
提案手法2	階層的クラスタリングを用いた検出レコード
提案手法1 U 提案手法1	提案手法1と提案手法2のいずれか一方で検出されたレコード
提案手法1 ∩ 提案手法2	提案手法1と提案手法2の両方で検出されたレコード

	六本木一丁目	西永福	新大久保	下北沢
サンプルサイズ	144	84	173	46
歪度	1.09	8.30	0.85	0.02
尖度	5.55	73.76	3.52	4.50
歪度（常用対数変換後）	-0.47	2.93	-0.01	-4.55
尖度（常用対数変換後）	3.98	22.03	2.96	27.48
Shapiro-Wilk test p 値	0.0000	0.0000	0.0000	0.0814
Shapiro-Wilk test（常用対数変換後） p 値	0.0231	0.0000	0.0602	0.0000

人間の判断	既存手法①	既存手法②	既存手法③	既存手法④	提案手法1	提案手法2	提案手法1 U 提案手法2	物件ID	市区町村名	地区名	最寄駅・名称	取引価格	最寄駅・距離(分)	面積	築年数	間取り	平米当たり単価	常用対数変換後平米当たり単価	建築年	建物の構造	取引時点	改装
0	0	0	0	1	0	0	0	7340	港区	麻布台	六本木一丁目	3,100,000	3	15	33	1 K	206,667	5,315	昭和58年	S R C	平成28年第4四半期	未改装
0	0	1	0	0	0	0	0	11290	港区	六本木	六本木一丁目	100,000,000	3	45	5	1 D K	2,222,222	6,347	平成24年	R C	平成29年第2四半期	改装済
0	1	1	1	0	0	0	0	11292	港区	六本木	六本木一丁目	490,000,000	4	190	5	3 L D K	2,578,947	6,411	平成24年	S R C	平成29年第2四半期	未改装
0	1	1	1	0	0	0	0	11308	港区	六本木	六本木一丁目	520,000,000	4	170	4	3 L D K	3,058,624	6,486	平成24年	S R C	平成28年第2四半期	未改装
0	0	0	0	1	0	0	0	68605	杉並区	永福	西永福	15,000,000	3	45	40	2 L D K	333,333	5,523	昭和49年	S R C	平成22年第2四半期	未改装
0	0	0	0	1	0	0	0	69760	杉並区	永福	西永福	10,000,000	4	30	36	1 D K	333,333	5,523	昭和49年	S R C	平成22年第2四半期	未改装
1	1	1	1	1	1	1	1	70351	杉並区	浜田山	西永福	1,200,000,000	10	160	26	3 L D K	7,500,000	6,875	平成元年	R C	平成27年第3四半期	改装済
0	0	1	1	0	0	0	0	17360	新宿区	百人町	新大久保	32,000,000	3	25	3	1 K	1,280,000	6,107	平成26年	R C	平成29年第1四半期	改装済
0	0	1	0	0	0	0	0	17362	新宿区	百人町	新大久保	18,000,000	4	15	20	1 K	1,200,000	6,079	平成9年	R C	平成29年第1四半期	未改装
0	0	1	1	0	0	0	0	17363	新宿区	百人町	新大久保	26,000,000	3	20	8	1 K	1,300,000	6,114	平成21年	R C	平成29年第1四半期	改装済
0	0	1	0	0	0	0	0	17367	新宿区	百人町	新大久保	30,000,000	3	25	2	1 K	1,200,000	6,079	平成20年	R C	平成28年第4四半期	未改装
0	0	1	1	0	0	0	0	17369	新宿区	百人町	新大久保	31,000,000	3	25	2	1 K	1,240,000	6,093	平成26年	R C	平成28年第4四半期	未改装
0	0	1	1	0	0	0	0	17370	新宿区	百人町	新大久保	32,000,000	3	25	2	1 K	1,280,000	6,107	平成26年	R C	平成28年第4四半期	未改装
0	0	1	1	0	0	0	0	17371	新宿区	百人町	新大久保	32,000,000	3	25	2	1 K	1,280,000	6,107	平成26年	R C	平成28年第4四半期	未改装
0	0	1	1	0	0	0	0	17372	新宿区	百人町	新大久保	33,000,000	3	25	2	1 K	1,320,000	6,121	平成26年	R C	平成28年第4四半期	未改装
0	0	1	1	0	0	0	0	17373	新宿区	百人町	新大久保	33,000,000	3	25	2	1 K	1,320,000	6,121	平成26年	R C	平成28年第4四半期	未改装
0	0	1	1	0	0	0	0	17377	新宿区	百人町	新大久保	32,000,000	3	25	2	1 K	1,280,000	6,107	平成26年	R C	平成28年第4四半期	未改装
0	0	1	1	0	0	0	0	17379	新宿区	百人町	新大久保	31,000,000	3	25	2	1 K	1,240,000	6,093	平成26年	R C	平成28年第4四半期	未改装
0	0	1	1	0	0	0	0	17380	新宿区	百人町	新大久保	30,000,000	3	25	2	1 K	1,200,000	6,079	平成26年	R C	平成28年第4四半期	未改装
0	0	1	1	0	0	0	0	54862	世田谷区	北沢	下北沢	80,000,000	5	65	15	3 L D K	1,230,769	6,090	平成13年	R C	平成29年第3四半期	改装済
0	0	0	0	1	0	0	0	54866	世田谷区	北沢	下北沢	35,000,000	3	80	32	2 L D K	437,500	5,641	昭和57年	S R C	平成26年第1四半期	未改装
0	0	1	1	1	0	0	0	55716	世田谷区	北沢	下北沢	860,000	2	55	28	2 L D K	15,636	4,194	昭和57年	S R C	平成22年第2四半期	未改装
0	0	1	1	1	0	0	0	56249	世田谷区	代沢	下北沢	100,000,000	5	70	5	3 L D K	1,428,571	6,155	平成24年	R C	平成28年第2四半期	未改装
0	0	1	1	1	0	0	0	56251	世田谷区	代沢	下北沢	110,000,000	5	75	3	3 L D K	1,466,667	6,166	平成24年	R C	平成27年第4四半期	未改装
0	0	0	0	1	0	0	0	59432	世田谷区	代沢	下北沢	28,000,000	12	70	19	3 L D K	400,000	5,602	平成28年	R C	平成21年第3四半期	未改装
0	0	1	1	1	0	0	0	59951	世田谷区	北沢	下北沢	25,000,000	3	75	23	3 L D K	333,333	5,523	昭和57年	S R C	平成17年第4四半期	未改装

表8 提案手法1と提案手法2の判定の違い（判定は1が異常値、0が正常値）

人間の判断	提案手法1	提案手法2	提案手法1 U 提案手法2	物件ID	市区町村名	地区名	最寄駅・名称	取引価格	最寄駅・距離(分)	築年数	面積	間取り	平米当たり単価	建築年	建物の構造	取引時点	改装
0	1	0	1	275	千代田区	神田淡路町	淡路町	150,000,000	3	2	70	3 L D K	2,142,857	平成25年	S R C	平成27年第4四半期	未改装
0	1	0	1	280	千代田区	神田淡路町	淡路町	260,000,000	3	2	115	4 L D K	2,260,870	平成25年	S R C	平成27年第2四半期	未改装
0	1	1	1	287	千代田区	神田淡路町	御茶ノ水	140,000,000	7	9	50	1 L D K	2,800,000	平成15年	R C	平成24年第1四半期	未改装
0	1	1	1	703	千代田区	九段南	九段下	440,000,000	8	13	150	2 L D K	2,933,333	平成16年	R C	平成29年第1四半期	未改装
0	0	1	1	1251	千代田区	内神田	淡路町	8,000,000	4	31	15	1 R	533,333	昭和54年	S R C	平成22年第1四半期	改装済
1	1	1	1	1446	千代田区	五番町	市ヶ谷	640,000,000	1	32	90	3 L D K	7,111,111	昭和54年	S R C	平成23年第2四半期	改装済
1	1	1	1	1623	千代田区	一番町	半蔵門	260,000,000	1	28	45	2 L D K	5,777,778	昭和52年	R C	平成17年第3四半期	未改装
1	1	0	1	1682	千代田区	神田駿河台	御茶ノ水	85,000,000	5	32	15	1 K	5,666,667	昭和50年	S R C	平成17年第4四半期	未改装
1	1	0	1	1686	千代田区	神田駿河台	御茶ノ水	82,000,000	1	28	15	1 K	5,466,667	昭和52年	S R C	平成17年第4四半期	未改装
1	1	1	1	3697	中央区	日本橋小網町	茅場町	200,000,000	3	11	20	1 K	10,000,000	平成15年	S R C	平成26年第4四半期	未改装
0	0	1	1	4598	中央区	八丁堀	銀座一丁目	19,000,000	13	9	25	1 K	760,000	平成16年	R C	平成25年第2四半期	未改装
1	1	1	1	4759	中央区	月島	勝どき	210,000,000	6	39	45	2 D K	4,666,667	昭和46年	S R C	平成22年第2四半期	改装済
1	1	1	1	4797	中央区	月島	月島	120,000,000	6	39	25	2 K	4,800,000	昭和46年	S R C	平成22年第2四半期	改装済
1	1	1	1	5003	中央区	佃	月島	550,000,000	3	8	60	2 L D K	9,166,667	平成14年	R C	平成22年第1四半期	未改装
1	1	1	1	5181	中央区	日本橋蛸殻町	水天宮前	160,000,000	3	6	20	1 K	8,000,000	平成16年	S R C	平成22年第3四半期	未改装
1	1	1	1	5235	中央区	日本橋馬喰町	馬喰町	160,000,000	0	6	25	1 K	6,400,000	平成16年	S R C	平成22年第4四半期	未改装
1	1	1	1	5239	中央区	日本橋箱崎町	水天宮前	240,000,000	3	30	55	3 D K	4,363,636	昭和55年	S R C	平成20年第4四半期	未改装
1	1	1	1	5394	中央区	湊	新富町(東京)	220,000,000	5	10	35	1 D K	6,285,714	平成12年	R C	平成22年第3四半期	未改装
0	0	1	1	5465	中央区	銀座	銀座	96,000,000	5	27	105	3 L D K	914,286	昭和59年	S R C	平成23年第3四半期	未改装
1	0	1	1	5543	中央区	築地	築地	99,000,000	4	30	45	2 D K	2,200,000	昭和56年	S R C	平成23年第4四半期	未改装
1	1	1	1	6330	中央区	月島	勝どき	110,000,000	6	12	15	1 K	7,333,333	平成8年	R C	平成20年第1四半期	改装済

(中略)

1	1	1	1	1	93069	足立区	弘道	五反野	40,000,000	6	23	15	1 K	2,666,667	昭和62年	R C	平成22年第2四半期	改装済
0	0	1	1	1	0	93288	足立区	六木	5,700,000	20	26	45	2 L D K	126,667	昭和62年	R C	平成25年第1四半期	未改装
1	1	1	1	1	0	94011	足立区	足立	61,000,000	7	21	15	1 R	4,066,667	昭和63年	R C	平成21年第1四半期	未改装
1	1	1	0	1	0	94208	足立区	竹の塚	97,000,000	15	18	65	3 L D K	1,492,308	平成3年	R C	平成21年第1四半期	未改装
1	1	1	1	1	0	94304	足立区	西保木間	180,000,000	22	5	55	2 L D K	3,272,727	平成16年	R C	平成21年第3四半期	未改装
0	0	1	1	1	0	94673	足立区	小台	16,000,000	28	29	80	4 L D K	200,000	昭和54年	R C	平成20年第3四半期	未改装
1	1	1	1	1	1	94801	足立区	島根	39,000,000	11	18	15	1 R	2,600,000	昭和62年	R C	平成17年第3四半期	未改装
0	0	1	1	1	0	94809	足立区	新田	12,000,000	29	20	50	3 D K	240,000	昭和61年	R C	平成18年第3四半期	未改装
1	1	1	0	1	0	95245	足立区	保木間	70,000,000	19	14	50	3 D K	1,400,000	平成6年	R C	平成20年第4四半期	未改装
1	1	1	1	1	1	95431	葛飾区	青戸	69,000,000	18	35	40	2 D K	1,725,000	昭和55年	R C	平成27年第4四半期	未改装
0	1	0	1	1	0	96174	葛飾区	立石	27,000,000	7	25	20	1 K	1,350,000	平成元年	R C	平成26年第4四半期	未改装
1	1	1	1	1	0	97086	葛飾区	新宿	110,000,000	18	19	55	2 L D K	2,000,000	平成3年	S R C	平成22年第4四半期	未改装
1	1	1	0	1	0	97109	葛飾区	東金町	110,000,000	9	26	40	2 L D K	2,750,000	昭和59年	R C	平成22年第4四半期	未改装
1	1	1	0	1	0	97657	葛飾区	東金町	270,000,000	14	12	95	4 L D K	2,842,105	平成9年	R C	平成21年第2四半期	改装済
1	1	1	0	1	0	98022	葛飾区	宝町	50,000,000	4	30	40	2 D K	1,250,000	昭和52年	R C	平成19年第4四半期	未改装
1	1	1	1	1	1	98182	葛飾区	東立石	430,000,000	12	16	55	3 L D K	7,818,182	平成3年	R C	平成19年第4四半期	未改装
1	1	1	1	1	1	98251	葛飾区	堀切	62,000,000	10	18	35		1,771,429	昭和63年	R C	平成18年第2四半期	未改装
1	1	1	1	1	1	98389	江戸川区	北葛西	630,000,000	13	22	95	4 L D K	6,631,579	平成6年	S R C	平成28年第1四半期	未改装
1	1	1	1	1	1	99695	江戸川区	船場	500,000,000	1	24	70	3 L D K	7,142,857	平成元年	S R C	平成25年第3四半期	未改装
1	1	1	1	1	1	100772	江戸川区	船場	290,000,000	9	25	80	4 L D K	3,625,000	昭和59年	R C	平成21年第4四半期	改装済
1	1	1	1	1	1	101050	江戸川区	西葛西	58,000,000	15	16	25	1 K	2,320,000	平成4年	R C	平成20年第1四半期	未改装
1	1	1	1	1	1	101118	江戸川区	西小岩	360,000,000	10	9	75	3 L D K	4,800,000	平成11年	S R C	平成20年第2四半期	未改装
1	1	1	1	1	1	101208	江戸川区	船場	420,000,000	1	7	70	3 L D K	6,000,000	平成13年	R C	平成20年第1四半期	未改装