博士論文

不動産価格予測モデルのための データクレンジング法と モデル更新アルゴリズムに関する研究

一橋大学大学院経営管理研究科
経営管理専攻
金融戦略・経営財務プログラム
BD18F001　大槻 健太郎
指導教員：横内 大介 准教授

序文

本研究の目的は不動産価格予測モデルの構築に必要なデータクレンジング，およびモデルの実運用上の課題を明らかにし，その１つの解決方法を提案することである．不動産価格に関する研究では，モデルの改良（関数形や説明変量の取り扱いなど）を中心的なテーマとして採り上げることが多い．しかしながら，不動産価格予測フローの入口と出口にあたる，データクレンジングや実運用上におけるモデル陳腐化への対応について採り上げた研究は少ない．本研究では予測に用いるモデル構築に必要なデータ収録値の浄化法，同一のモデルが適用可能な地域を特定するグルーピング手法，市況変化によるモデルの陳腐化を防ぐモデル更新のフレームワークに焦点を当て，現状の課題を整理するとともにそれらを解決する方法を提案する．なお，本研究で対象とする不動産は戸建てと比べ，相対的に物件の構造的な同質性が高いと考えられる中古マンションとする．また対象地域とする地域は取引量が多く，一定のサンプルサイズが確保できる東京 23 区とする．

目 次

第1章　序論

1.1　用語について

本研究において不動産価格予測モデル，中古マンション価格モデル，あるいは単にモデルと表記する際は統計モデルを指すものとする．第 2 章から第 4 章では議論に用いる統計モデルが線形回帰モデルであることを明示している．

つぎに本研究において予測とはモデルによる予測という意味で用いており，時間の経過を $\tau = 0, 1, 2, \cdots, t, \cdots$ とするとき，時点 t までに得られた中古マンションの価格 y と築年数などの属性データ x（多変量の場合は $\boldsymbol{x}$）をインサンプルとして中古マンション価格モデル $y = f_t(x)$ を推定し，時点 t 以降のデータをアウトオブサンプルとして予測値 $\hat{y}$ を得ることを指すものとする．つまり，予測精度とは特に断りが無いかぎり，予測において算出した予測値 $\hat{y}$ に対し，時間の経過によって実際に得られた観察値 y との差である予測誤差 $y - \hat{y}$ の大小を指すこととする．

1.2　研究の目的と背景

本研究の目的は WEB サイト上で数多く展開されている不動産査定サービスの予測精度を維持・改善するため，不動産価格予測モデルの運用上の課題を解決することである．具体的には不動産取引データベースのエラーレコードへの対応，売買取引の地域的な偏在によって生じる局所的なモデル予測精度の劣化，市況変化によるモデル陳腐化の 3 つを課題として採り上げている．

不動産価格に関する研究では，モデルの改良（関数形や説明変量の取り扱いなど）やモデルを用いた価格指数の構築を中心的なテーマとして採り上げることが多く，モデル構築の元になるデータの取り扱いに焦点を当てたものは少ない．不動産価格予測モデルの構築プロセスをデータの収集やデータベースの構築から始まり，モデルの推定を経てそのモデルのメンテナンスを終わりとするフローとして捉えるならば，本研究ではプロセスの入口と出口にあたる，データクレンジング，モデル更新アルゴリズムをテーマとし，実務上の課題に対する解決策を提案している．

なお，本研究で対象とする不動産は戸建てと比べ，相対的に物件の構造的な同質性が高いと考えられる中古マンションとする．また対象地域とする地域は取引量が多く，一定のサンプルサイズが確保できる東京 23 区とする．

研究目的の背景にある不動産査定サービスは日本では2015年ごろからソニー不動産（現SREホールディングス[1]）などの不動産仲介会社において提供が始まり，不動産物件のオーナーにとっては物件売却価格の目安を確認する手段として，そして不動産仲介会社にとっては媒介契約獲得の営業ツールとして双方にメリットがあるため，瞬く間に広まった．現在では不動産仲介大手から不動産ベンチャーまで数多くの会社がWEBサイト上で提供している．

こうしたサービスの開始以前にも不動産のオーナーにとっては保有する物件の現在の売却可能価格を知りたいというニーズ，物件の購入希望者にとっても自分の求める条件に合う物件の候補を集めて価格を比較したいニーズが存在していたと思われる．しかしながら，日本においては1990年代後半にWEBサイト上で不動産物件の検索サービスこそ提供され始めていたものの，取引価格となると取引当事者や不動産仲介に関わる立場でなければ知ることはできなかった．アメリカ，イギリス，フランス，オーストラリア，香港，シンガポールなど他の先進国（地域）においては不動産取引の取引価格が登記簿や税務署において公開され，閲覧可能であることから，日本では不動産流通市場における情報整備が遅れ，価格の透明性が無いとの指摘もされてきた（荒井 [1]）．

日本では不動産仲介業者であれば指定流通機構(REINS，Real Estate Information Network System，以降はレインズと呼ぶ）に登録された不動産情報を閲覧すれば，類似物件の売り出し価格や取引価格を参考として知ることができる[2]一方，一般の物件保有者や購入希望者にとっては不動産の取引価格の情報を得る手段はこれまでほとんど無かった．そもそも日本では不動産仲介業者は売主と買主の両方から仲介手数料を得ることができるため，かならずしも物件の売主と買主に対して正確な価格情報を提供するインセンティブが働かず，積極的に取引事例の情報が開示されることも無かった[3]．

このような不動産流通市場の価格の不透明性に対処するため，国土交通省は2006年より土地総合情報システム[4]というデータベース（以降は国交省データベースと呼ぶ）をWEB 上で公開している(国土交通省 [10])．国交省データベースは2005年第３四半期以降に売買された物件について，不動産登記簿に記録された物件購入者から取引価格を含む不動産取引情報をアンケート形式によって収集し，対象物件が容易に特定できないように加工して蓄積・公開している．

また2000年代後半以降，コンピュータの能力向上と情報通信技術の急速な普及により，ビッグデータの蓄積，大規模なデータの演算処理，ネットワークを通じた情報の共有が進んだため，それを活かした新しいタイプの不動産関連サービスが次々と生まれた(白川・大越 [14])．これらはプロップテック（PropTech）や不

[1] https://sre-group.co.jp/

[2] 一般媒介契約であればレインズへの登録義務が無いため，取引事例全てが網羅されているわけではない．

[3] 一般媒介契約であればレインズへの物件情報や成約処理を登録する義務は無いため，物件情報の囲い込みなどを行う悪質な仲介業者が発生する原因にもなっている．

[4] https://www.land.mlit.go.jp/webland/

動産テック（Real Estate Tech）と呼ばれ，その先駆けとなった米国では VR（仮想現実）を使用して遠隔で不動産物件の内見ができる Matterport，民泊の仲介プラットフォームを提供する Airbnb，シェアオフィス・コワーキングスペースを提供する WeWork など，それまでの不動産業界には無かったサービスを提供する企業が現れた．米国のプロップテック企業における代表的な４社を指す ZORC（ゾーク,「Zillow」「Opendoor」「Redfin」「Compass」の頭文字）は元々，不動産物件の売り手と買い手を繋ぐ情報提供と仲介をコアビジネスとして発展してきた．中でも筆頭格の Zillow[5]は全米のほぼ全ての物件の売買参考価格を過去の取引データを元に推計し，WEB の地図上に表示する「Zestimate」というサービスを展開している．日本でも 2015 年頃より AI による不動産査定を行うサービスが数多く開始され (谷山 [16])，不動産物件のオーナーは不動産市況の相場情報を得る機会や保有資産の売却想定価格を知る機会が得られるようになった．

学術研究においても飛躍的に性能が向上した計算機を活用することで，統計学や機械学習に関連した計算量の多い手法も容易に使えるようになったことから，テキスト解析技術や画像認識技術を用いた新しい切り口の研究などが行われている．たとえば服部ら [21] は間取り図の有無が賃貸物件の賃料に与える影響を線形回帰モデルと LightGBM を用いて検証している．また谷山ら [17] は日経不動産マーケット情報のニュース記事に基づく不動産センチメント指数，Google の検索結果に基づく不動産アテンション指数を開発するなど，テキスト情報や検索ボリュームを用いた新しい不動産景況指数を開発している．さらに不動産価格指数の算出や不動産価格予測モデルの改良といった従来からの不動産研究テーマにおいてもディープニューラルネットワークや勾配ブースティングモデルなど，より計算量の多いモデルが採り入れられるようになってきている．

不動産データへのアクセスが容易になり，不動産価格予測モデルの改良による不動産査定サービスの予測精度の向上が進む一方でこれまであまり採り上げられなかった課題も認識されるようになってきている．たとえば不動産査定サービスに用いる不動産価格予測モデルの推定根拠となる不動産データベースには入力値の欠損や誤り，住所や建物名の表記ゆれ，変量間の不整合など適切な分析を阻害する様々な問題点が数多く含まれる．また不動産データは新たな取引データが追加されるたび，それまでにない新しいタイプの入力エラーなどが混入することもあるため，一度行ったデータの処理を継続的にアップデートする必要も生じる．これまでこうした問題点への対処はデータベースを利用する分析者やサービス開発者のそれぞれに委ねられ，エラー値の削除や修正など，実際に施した処理の内容を詳細に明示されることが少なく，標準的な方法が確立されてこなかった．

また不動産査定サービスにおいては不動産価格予測モデルを運用する際に地域を細分化してモデルを推定することで一般的に予測精度は改善する．ただし，不動産取引の発生多寡は地域によってばらつきがあるため，地域を細分化しすぎると不動産価格予測モデルの更新頻度が高いエリアと低いエリアが生じてしまい，市

[5]https://www.zillow.com/

況の変化を取り込めないエリアが出てくる可能性がある．さらには，いかに精度の良い不動産価格予測モデルを開発しても，景気の変動，制度の変更，消費者行動の変化などを原因として，いずれはモデルが陳腐化し，予測精度が落ちてしまう．しかしながら，不動産価格予測モデルの運用場面において，いつモデルを見直すべきか，という意思決定に関する課題を採り上げた不動産研究は筆者の知る限り見当たらない．

こうした実務上の課題をデータサイエンスプロセス（たとえばSchutt, R and O'Neil, C [89] など, 次節の図1.3）の観点で捉えれば，データ浄化やデータ加工，アプリケーション化・意思決定といったステップでの課題として捉えられるが，不動産関連の先行研究ではあまり取り扱われることが無いテーマである．そこで本研究では不動産データのクレンジングや不動産価格予測モデル運用の意思決定の課題として採り上げ，一つの解決方法を提案する．

1.3 データサイエンスプロセスと不動産価格予測フローの対応付け

データサイエンスプロセスのフレームワークにはCRISP-DMやOSEMN，KDDプロセスなどが提唱されてきた．CRISP-DM(Cross-Industry Standard Process for Data Mining) はChapman et al. [41] が提唱したデータマイニングのための分野横断標準プロセスであり，ビジネス理解，データの理解，データの準備，モデリング，評価，実装の６つのステップでデータが持つ価値を実現できるとされる．またOSEMNはデータの取得（Obtain），データの浄化（Scrub），探索的データ解析（Explore），モデリング（Model），得られた示唆の提示（iNterpret）という５つのステップを表している（Hilary and Wiggins [81]）．Fayyad et al. [55] が提唱したKDDプロセス（Knowledge Discovery in Databases）はデータベースから知識発見を行うプロセスであり，データの選択，加工，変換，マイニング，評価の各ステップから成る．

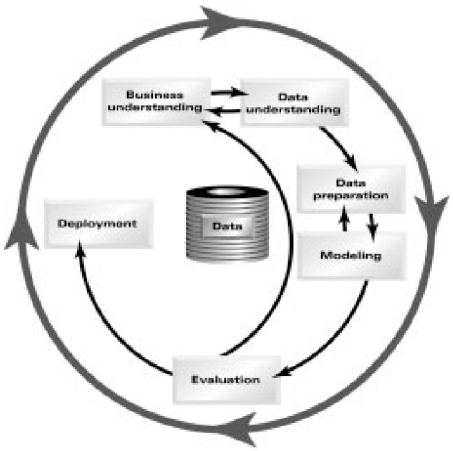

図 1.1: Chapman et al. [41] による CRISP-DM フロー

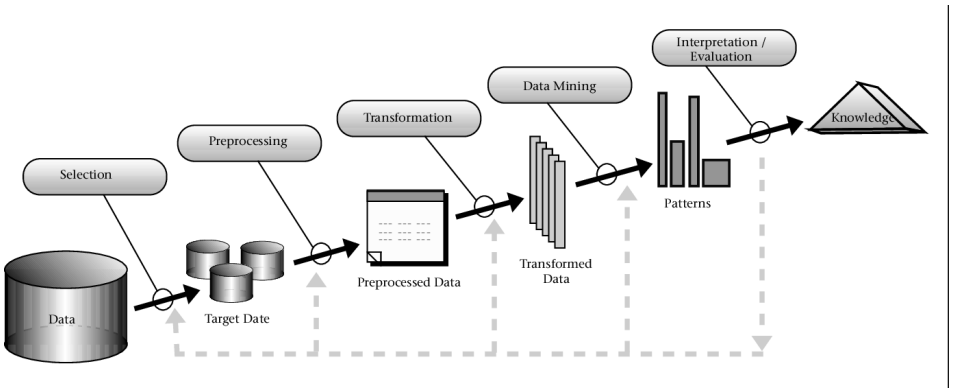

図 1.2: Fayyad et al. [55] による KDD プロセス

ただし，いずれのフレームワークでも基本的な考え方は大きく異なることはないため，本研究では O'Neil and Schutt [89] や横内・大槻・青木 [27] にて紹介されているデータサイエンスプロセス（図 1.3）に不動産価格予測フローを対応させ，それぞれのステップにおけるデータサイエンスと不動産価格に関する既存研究を紹介する．

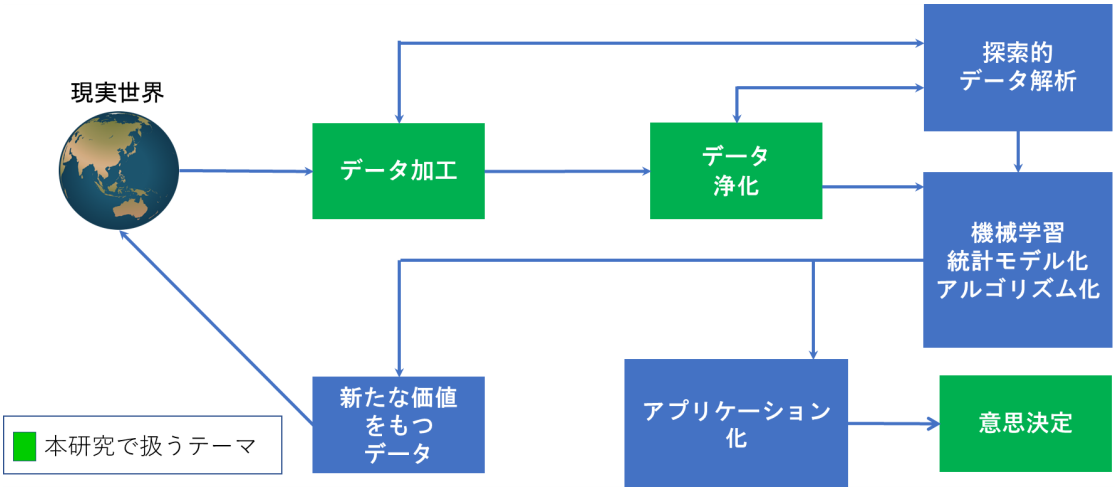

図 1.3: 出典：O'Neil and Schutt [89] ならびに横内・大槻・青木 [27] を参考として筆者作成

データ加工

現実世界で発生する人や組織の活動，自然現象や実験の結果などについて人の解釈できる内容を記号で記録したものが生データである（久保 [8]，O'Neil and Schutt [89]）．生データの収集過程では集める目的が決まっていて計画的に取得する場合だけでなく，業務の遂行などに伴って副次的に集められる場合もあるため，データの発生と取得の状況を把握し，データが現象をどれだけ偏りなく反映しているかを確認する必要がある（柴田 [13]）．収集した生データはデータを扱う目的に応じて適切な形に編集・加工され，データベースとして蓄積される．データベースを構築する際に考慮されるのがデータモデルであり，増永 [25] は大別してネットワークモデル，ハイアラキカルデータモデル，リレーショナルデータモデルの３つを挙げている．

不動産取引の生データは主に３つの種類があり，不動産仲介業者が自らの仲介業務のために蓄積したデータ，不動産仲介業者が宅地建物取引業法に基づいてレインズに登録するデータ，国土交通省が売買の成立した不動産登記簿上の買主からアンケートで収集するデータ（加工されたものが前述の国交省データベース）である．こうして収集された不動産データは各物件の延床面積や間取りなどの特徴をまとめて一つのレコードとし，リレーショナルデータの形式で収録されることが多い．国交省データベースも不動産物件の特徴をそれぞれのカラム，取引された各不動産物件をレコードとした矩形データで蓄積されている．

リレーショナルデータモデルの考案者である Codd [45] に従って中古マンションのリレーショナルデータを各変量のドメイン（定義域）$D_1, D_2, \cdots, D_p$ の直積集合の部分集合として表現すると次のとおり．

$$R \subset D_1 \otimes D_2 \otimes \cdots \otimes D_p \tag{1.1}$$

$$\begin{aligned} D_1 &= \{\text{m}^2\text{当たり単価}\} \\ D_2 &= \{\text{延床面積}\} \\ &\vdots \\ D_p &= \{\text{建物地上階数}\}. \end{aligned}$$

なお，R はリレーションを表す集合，$\otimes$ は直積を表す．それぞれのドメインから１つの値を取り出した組が不動産取引１件の観測に対応しており，実際に観測されたレコードの集合によってリレーショナルデータが形成される．また，あるドメインから実際に観察された値の束はデータベクトルとして捉えることができる．

データ浄化 (データクレンジング)

生データをデータベースとして取り込んだあとは重複したレコード，欠損値，異常値，不正な記録データなどをチェックするデータクレンジングのステップとなる．これまで外れ値の検出や欠損の処理は多くの研究がなされている．外れ値の検出に際しては正規分布を前提とした方法（Dixon [51] や Grubbs [61]，Hotelling [65]，Mahalanobis [78]），サンプル同士の距離を計測する LOF(Breunig et al. [37]) やサポートベクトルマシン（Vapnik [98]，Tax and Duin [95]）といった機械学習的な方法など様々なものが提案されている．また欠損の処理については単一代入法や多重代入法（Rubin [85]，Rubin [86]）など目的（観察サンプルの分布把握か母集団の推定か）や条件（欠損が単一の変量か複数の変量か）に応じて多くの方法が研究されている（高橋・渡辺 [15]，野間 [19]）．

不動産データに対して必要となるデータクレンジングの範囲は広い．まず物件の各属性情報に欠損値が多いうえ，入力ミスや収録ルールの無視によるエラーも相当な数が散見される．こうしたデータを分析やサービスに使えるデータとして整備するためには多くの作業と時間が必要になるが，不動産データに対するクレンジングはこれまでの研究では詳細に説明されることが少なく，標準的なデータクレンジングフローが確立されているとは言えない．不動産価格予測モデルの推定という点では大きく次の２点のデータクレンジングが課題となる．

1. 欠損値
   不動産データセットには各レコードの様々な変量において多くの欠損値が存在する．不動産レコードにおける変量の欠損値には大きく２つの種類がある．一つは物件の最寄り駅や用途地域のように住所ないしは緯度・経度に対して外部の地図データを組み合わせるなどの方法 (コールドデック補定) によって分析者が補完可能な変量である．もう一つは取引価格のように通常，取引当

事者や関係者以外には知りうる手段が無く，分析者が本来の値を補完することは不可能な変量である．

2. 外れ値
ある変量（たとえば取引価格）が極端に他の値と比べて外れた値を取るケースもあれば，間取りと延床面積を突き合わせると辻褄が合わないような変量間の不整合性というケースもある．また外れ値の意味を拡大解釈すれば，不動産取引データでは変量の定義に合わない入力値も散見される．たとえば物件名に仲介手数料についての記載があったり，管理会社の入力欄がブランクにもかかわらず備考欄に管理会社名の記載があったり，中古物件にもかかわらず建築年月が取引年月より後だったり，といった入力ルールの逸脱が観察される．さらには入力ルールの逸脱ではないものの，物件名やデベロッパー名などには表記のゆれが多数存在し，分析の目的によっては大きな支障が出る場合がある．

探索的データ解析

Tukey [96]，Tukey [97] が提案した探索的データ解析（EDA，exploratory data analysis）はデータに含まれる情報をいかに捉えるかを重視する考え方と方法論を指す．提案された当時，統計学で主流であったのは統計的推測法の開発のように理論構築が主でデータが従というコンセンサスであり，探索的データ解析はそれとは対称的なスタンスであった (柴田 [13])．

探索的データ解析では得られたデータに対していきなり統計モデルや機械学習手法を適用するのではなく，データの発生や取得の経緯を把握しながら外れ値を検出したり，むらのあるデータセットから均質性の高いデータを発見するためにデータを眺めたり，といったステップを重視する． 探索的データ解析は本来，図 1.3 のデータサイエンスプロセスにおけるデータ浄化のステップも含むが，ここではデータのブラウジングによる知識発見のステップとして採り上げる．

探索的データ解析ではデータビジュアライゼーションによってデータが示す状態（観察対象である現象）からモデル化につながるヒントを発見するため，箱ひげ図や幹葉表示などが考案された（Tukey [97]）．Inselberg [67] は高次元データの表現としてそれぞれのデータベクトルの座標軸を横に並べてそれぞれのレコードを折れ線で繋ぐ平行座標プロットを考案した．熊坂・柴田 [9]，Kumasaka and Shibata [75] の提案した TextilePlot は平行座標プロットを各軸の位置と尺度を適切に変換してそれぞれの折れ線が出来るだけ水平になるように改良し，データブラウジングに適したビジュアライゼーションツールとして発展させたものである．

不動産データにおいては箱ひげ図やヒストグラムなどを用いて外れ値を検出するだけでなく，取引価格に影響を与える変量の発見のためにデータビジュアライゼーションが行われる．不動産価格は様々な要因から形成されると考えられるた

め，不動産価格予測モデルへの採用候補となる説明変量は多くなりやすい．そこで適切に価格形成要因となる変量を捉えるため，相関行列のヒートマップや平行座標プロットが利用されたり（Dabreo [47], Hullman and Gelman [66], Li [79]），回帰木によって価格の説明に重要な変量を捉えたり（Fan et al. [54]）といったアプローチが用いられている．

機械学習・統計モデル化・アルゴリズム

不動産価格のモデル化は経済学の分野で Lancaster [77] や Rosen [84] などが理論的に整理したヘドニックアプローチを用いることが多い．ここではその理論的な背景には触れないが，ヘドニックアプローチは不動産に限るものではなく，商品 (財) の価格をその商品に含まれる便益や性能などの特性の集合とみなし，こうした特性に商品価格を回帰させるというアプローチである．不動産価格に対するヘドニックアプローチは金本 [5] や中村 [18] がその適用条件や問題点を解説している．また刈谷ら [7] は不動産の需要者の選好は人によって様々（非同質で多様）であり，それぞれ非同質な不動産を求めることから，Rosen [84] の設定した理論面の仮定には批判的な意見を主張している．

一方，不動産は車両などと比べても個別性が強いため，土地価格の予測モデルを推定するためには通常の最小二乗法よりロバスト MM 推定を用いる方が良いという報告もされている (Hannonen [63])．Chiodo [43] や清水 [90] はヘドニックアプローチによって推定した不動産価格予測モデルの説明変量には非線形性が存在していると主張している．

近年，不動産価格予測モデルには様々な機械学習手法や統計モデル，アルゴリズムを用いた研究がなされており，特に回帰木の発展形であるランダムフォレストや XGBoost，LightGBM を用いた不動産価格予測は精度の向上が著しい (Li et al. [80]，三田 [26]，福中 et al. [24], Sibindi et al. [87])．

新たな価値を持つデータ

不動産取引データから生み出され，新たな価値を持つ代表的なデータとしては不動産価格指数がある．有名な不動産価格指数には全米の主要都市の戸建て再販価格より算出されるケース・シラー住宅価格指数があり（Case and Shiller [38]，Case and Shiller [39]），景気動向を捉えるマクロ経済指標として使われている．再販価格を用いて価格動向を把握するアプローチは Bailey et al. [31] が提案したリピートセールス法と呼ばれる方法で，同じ物件の異時点での価格差によって指数を算出するため，個別性の強い不動産物件の品質の違いを考慮せずに指数を算出できる．米国では中古不動産の流通が一般的であるため，リピートセールス法は数多くの研究がなされている．

またサブプライムローン問題から端を発した2000年代後半の世界的な金融危機への危機感から，不動産（住宅）の価格動向を把握できる指標へのニーズが高まり，国際的に共通のルールに基づく指数の作成を目的として，欧州委員会統計局（Eurostat）や国際通貨基金（IMF），経済協力開発機構（OECD）などの国際機関によって住宅価格指数に関するハンドブックが作成された（Eurostat [53]）．このハンドブックでは住宅価格インデックスの算出に際し，1）住宅取引価格データの中央値を用いる方法，2）ヘドニックアプローチを用いる方法，3）リピートセールスを用いる方法，4）不動産鑑定価を用いる方法の4つを挙げている．また，ハンドブックにはこれらの方法を組み合わせてインデックスを算出する方法も紹介されている．こうした指数算出の流れを受け，国土交通省においても2012年8月より不動産価格指数（住宅）の算出と試験公表を開始し，2015年3月からは本格運用を行っている（国土交通省 [11]）．もともと中古不動産の流通が一般的でなく，再販される物件のデータが得にくい日本ではケース・シラー住宅価格指数のようにリピートセールス法を適用することが難しく，国土交通省による不動産価格指数ではヘドニックアプローチ（統計学的には線形回帰モデル）を採用し，各不動産物件ごとに異なる品質を調整して算出されている．

アプリケーション化・意思決定

探索的データ解析と統計モデル化などのフェーズを経たあとは最終的にサービスとして実装されたり，レポーティングによって意思決定の材料とされたりする．Provost and Fawcett [83] はビジネスにおけるデータサイエンスの究極の目標が意思決定の改善にあるとし，最も幅広くデータサイエンスが用いられるビジネスアプリケーションはオンライン広告，クロスセルのための商品推薦などのマーケティングであろう，と主張している．また，金融業や通信業において不正行為の検出のため，データドリブンなシステムによる自動的な意思決定を実装した例や大手小売業において災害時の品揃えについてデータに基づいて知見を得た例を紹介している．またデータサイエンスをビジネスの意思決定に活用するためのフレームワークをテーマとする研究も数多く行われている．Chiheb et al. [42] はビッグデータを活用して意思決定プロセスを改善するため，インテリジェンス，デザイン，選択，実施という４つのフェーズからなるフレームワークを提案している．

不動産データを用いたアプリケーション化の研究としてはLi et al. [79] が不動産検索サービスの利便性向上のため，物件情報と価格だけでなく，周辺環境の情報提供や類似の物件との比較を可能とするHouseSeekerシステムについて提案している．

1.4 各研究の位置付けと要旨

本研究は第 2 章から第 4 章まで 3 つの研究で構成される．各章はそれぞれデータサイエンスプロセスのうちのデータ浄化，データ加工，意思決定という 3 つのステップに対応し，それぞれ実務上の課題に対する解決策の一つを提案している．

データ浄化—エラーレコードの検出

第 2 章の研究ではデータ浄化ステップの外れ値検出のうち，数値エラーが発生しているレコードを特定することに焦点を当てている．不動産のデータは欠損や値の誤り，収録ルールからの逸脱など，いわゆる「汚い」データであるため，様々なデータ浄化の手続きが必要となるが，不動産関連の研究においてこうしたデータクレンジングをテーマとして採り上げた研究は少なく，日本の不動産データを用いた研究では筆者の知る限り見当たらない．

本研究では国土交通省が公開している国交省データベースの中古マンションデータを採り上げ，データクレンジングのポイントを整理している．その上で取引価格の桁間違い入力エラーを適切に検出するための 2 つの方法を提案している．

取引価格の桁間違いエラーは中古マンション価格モデルの推定に大きな影響を及ぼすにも関わらず，単純に取引価格の平米単価に上限値を設定してレコードを取り除くといった処理がしばしば行われる．しかしながら，たとえば東京２３区の中古マンションの場合は山手線の内側と外側の中古マンションでも大きく値段が異なることが分かっており，ある地域で桁間違いと判断される物件が別の地域では必ずしも桁間違いとは言えない．また築年数が経過していたり，最寄駅からの距離が離れていたり，といった物件の属性面も考慮する必要がある．

本研究では，取引価格に加えて地域性や物件属性を考慮した桁間違いの判定法，という中古マンション情報の意味的な要素に踏み込んだデータクレンジングの方法を提案している．本研究で提案した方法によって取引価格の桁間違いエラーとされるレコードを除去した結果，中古マンション価格モデルの推定結果が大幅に改善している．

なお，取引価格の桁間違いエラーが入力されているレコードは国交省データベースに特有のものではなく，大手不動産仲介会社などが蓄積している不動産データベースにも散見されるエラーであり，本研究で提案した手法は不動産データベース全般に対して効果がある．

データ加工—データ分割と再統合

第 3 章の研究目的は市況が変化する中で中古マンション価格モデルの予測精度を維持する，という実務的な問題の解決である．中古マンション価格モデルの予

測精度を改善するためには中古マンションの価格形成要因と考えられる変量を全てモデルに採用することが考えられる．ただし，ある程度大きな範囲（たとえば東京２３区）で１つのモデルを構築すると，マンションの立地条件ひとつとっても様々な要因を背景に取引されることから，価格形成要因となる変量を全て特定して，モデルに導入することは現実的に難しい．

予測精度を改善するためのもう一つの方法として様々な地域が混在するデータをできるだけ同質なデータに細分化したうえでモデルの推定と予測を行うアプローチがある．東京２３区の中古マンションデータであれば最寄り駅ダミーや町丁ダミーをモデルに導入することでもモデルの予測精度が良化する．しかし，中古マンション取引が活発な地域は偏在するため，過度に地域を細分化した場合，価格予測モデルの運用という観点から見ると，新たに取引が発生した地域の予測価格は更新されるものの，類似した近隣地域の予測価格は置き去りとなって更新されない．このため不動産査定の現場が一体として捉える地域のマンション価格の変化と，モデルが算出する当該地域の予測価格の間にしばしば大きな乖離が発生する．

本研究ではこの中古マンション価格モデルの予測精度の劣化問題に対処するため，モデル推定のための取引実績データをどの程度の地域まで分割すれば現場の感覚との乖離が無く，予測精度が劣化しないモデルの実装を実現できるかという，工学的な問題として読み替える．そして本研究ではこの問題を，想定する最小単位でのデータ分割と予測精度が落ちない形でのデータの再統合と捉えなおすことで解決している．データの再統合には中古マンション価格モデルの当てはまりを基準としたシンプルな方法を用い，予測精度の劣化問題に対処するう実例を示している．

本研究はデータサイエンスプロセスに則すれば不動産データベースに対してデータブラウジングを行い，類似したデータの範囲を捉えてモデル化する段階ともいえ，柴田 [12] が提唱するように探索的にデータを解析して均質性の高いレコードをまとめる，データ加工のステップに位置づけられる．

意思決定―モデル更新アルゴリズム

第 4 章では実務における不動産価格査定サービスを考慮した場合に不動産価格予測モデルの運用，より具体的には市況の変化がある中でモデルの予測精度を維持するために，モデルの更新をどう行っていくか，というサービス運用者の意思決定に関わるテーマを採り上げる．

中古マンション価格査定の実務では市況が変化する中で過去の取引データから推定した中古マンション価格モデルを用いて売却希望の物件の価格予測を行うため，直近の価格動向を織り込みつつ，いかに予測誤差を最小化するか，ということが重要な関心ごとである．ただし，個々の物件ごとに固有の事情を持ち，個別性が強い中古不動産は，査定対象となる全ての物件の予測誤差を最小化することは現実的では無く，一定の期間に取引された物件の平均的な予測誤差を最小化す

る，あるいは大多数の物件の予測を大きく外さないということが実務における目標となる．この目標を達成するためには市況を反映した最新のデータを十分なサンプルサイズで確保し，物件の属性や物件の所在地域に関する必要な情報を網羅することが理想であるが，時間が経過する中でも常のこれらの条件を維持しつづけることは難しい．

実務において現実にこの目標に対処するためには次に挙げる３つのアプローチが考えられる．

1. 中古マンション価格モデルの改良
   中古マンション価格に影響を及ぼす要素を説明変量としてモデルに取り込む，または適切な関数形を与えることで予測精度を改善することができる．ただし，採用するモデルによって結果の解釈性が落ちたり，オーバーフィッティングしたりする可能性があり，またデータに欠陥がある際はモデルの改良が難しくなる．

2. 中古マンション価格モデルの更新頻度の向上
   更新頻度が高いモデルであれば市況の変化によるモデルの陳腐化を回避し，予測精度を維持，向上できると考えられるが，更新時点ごとに十分なサンプルサイズを確保する必要がある．

3. 中古マンション価格モデルの推定に適したデータの期間設定
   市況の変化を取り込むために新しく観測された取引データを用いてモデルを更新する際に，モデル推定に用いるデータの期間をより適切な長さにすることで予測精度を向上できる余地がある．ただし，期間の長さの設定が難しく，安定した予測を行うためにはモデル推定にはある程度長い期間のデータが必要となる一方，あまり長期間のデータで推定すると直近の価格の変動を十分にモデルへ反映できないため，予測を大きく外す場合もある．

本研究では入手可能なデータの質や量に依存せずに取り組み可能であるという観点から，中古マンション価格モデルを推定するデータの期間の長さをテーマとする．

市況の変化に合わせて中古マンション価格モデルによる予測の精度を維持するには，時間の経過とともにモデルを更新していく必要がある．そのため，実務においてはモデルの更新に際し，一定の幅の期間経過ごとに新しいデータを追加しつつ，同時に同じ幅の古い期間のデータを削除する，いわゆる固定スライディングウィンドウを採用することが多い．しかしながら，固定スライディングウィンドウはモデルの推定に使用する期間の幅をどう設定するか，という点が難しい．もし期間の幅を長く取れば，サンプルサイズが大きく取れるため，中古マンション価格が緩やかに動く場合は予測精度が高くなり，長期的に安定する．しかしながら，中古マンション価格は経時的に少しずつ変わる場合だけでなく，短期間で大きく変動する場合もあるため，モデルの追従が遅れれば予測精度が低下する問題

を引き起こす．一方，期間の幅を短くすると，中古マンション価格の急な変動にも追従できるものの，モデルの推定に用いるサンプルサイズが小さくなり，予測精度が安定しない．

第 4 章では市況の変化に対応するべく，観察されるデータに応じて期間の幅を調整する可変ウィンドウを用いたモデル推定アルゴリズムを考案している．また実際に取引された中古マンションデータを用いて我々の提案する可変ウィンドウと既存手法である固定スライディングウィンドウを比較したところ，可変ウィンドウによって更新されたモデルは対象期間全般にわたって予測を大きく外すことなく，市況の変化に対する追従性を確保しつつ，予測の安定性を維持できることを確認している．

1.5 本論文の構成

本論文は５章から構成される．この章以降の本論文の構成は次のとおりである．

- 第２章 中古マンション価格モデルのためのデータクレンジング法
  データ浄化をテーマとした研究で日本不動産学会誌第 34 巻第 3 号の大槻・横内 [2] に論文が掲載されている．
- 第３章 東京２３区の中古マンション市場のデータ分割と統合
  データ加工（探索的データ解析も含む）をテーマとした研究で日本不動産学会誌第 36 巻第 4 号（近刊）の大槻・横内 [3] に論文が掲載予定である．
- 第４章 可変ウィンドウによる中古マンション価格モデルの更新
- 第５章 結論と今後の課題

第2章　中古マンション価格モデルのためのデータクレンジング法

2.1　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取引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価格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が挙げられる．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中古マンション価格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 区内だけでも 10 万件以上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は検出漏れや異常値判定にぶれが生じる可能性が高い．また時間の経過とともに新しい取引データが収録されるたびに新たなレコードに含まれる異常値の識別作業が発生するため，人間がそれまでの判定との整合性を取りながら異常値を検出する作業は負担が大きい.

これまで国交省データベースを用いて不動産価格のモデルを推定している研究は Shimizu and Nishimura [91]，唐渡 [6]，早川・田島 [22] 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.

[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)

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中古マンション価格モデルを推定した結果も示す．

2.2 先行研究

第 2 章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 ”とは一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 Lewis [32] 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 [40] 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 Lipscomb [74] 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 ”（以降，DPP と呼ぶ）と呼び，データセットの統合やデータクレンジングを包括的に紹介した上でデータクレンジングについてはラベル付

け，値エラー，欠損，DPP の記録というそれぞれの論点を整理し，解説している．なお，彼らは代表的な不動産関連学会誌における過去の不動産関連論文において，DPP プロセスの記述状況を評価するスコアを考案してサーベイを行っている．

2.3 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 4.1 の項目からなる．このデータセットに対して Krause and Lipscomb [74] を参考にデータクレンジングのポイントを紹介する．

表 2.1: 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		2010 年第３四半期
改装		未改装
取引の事情等		調停・競売等

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．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 2.2 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に見える．ここで実際に不動産売買仲介会社の WEB サイトを参考にすれば，最寄り駅が六本木一丁目で面積 150m^2 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 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，通常は一桁少ない 1 億 2 千万円の取引と考える方が妥当であろう．

表 2.2: 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
駅徒歩距離	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

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．

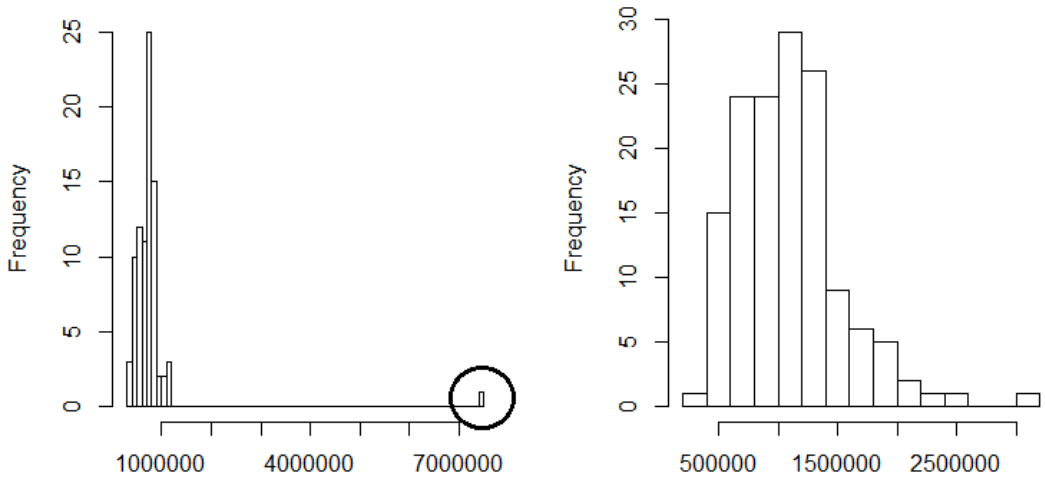

図 2.1: 最寄り駅が同じグループの m^2 当たり取引単価

図 2.1 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 m^2 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．

- m^2 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 $35m^2$ 未満の物件で取引価格が 8000 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 m^2 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 100m^2 を超えている場合は築年数によらず相当な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検出漏れや異常値判定にぶれが生じる可能性が高い．また時間の経過とともに新しい取引データが収録されるたびに新たなレコードに含まれる異常値の識別作業が発生するため，人間がそれまでの判定との整合性を取りながら異常値を検出する作業は負担が大きい．

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，通常は取引関係者以外には実際の取引価格が分からないため，本研究では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 (m^2) 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 2.1 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 20m^2 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，m^2 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．この正解レコードの詳細は付録 $A.1$ に掲載した．

2.4 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] $\mathrm{SIQR_u}$ 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 Tukey [97] 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 [20] 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 (既存研究と提案手法の比較例は付録 $A.2$ に掲載した)．

kimber [73] が提案した $\mathrm{SIQR_u}$ は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 $\mathrm{SIQR_u} := \mathrm{Q}_{(0.75)} - \mathrm{Q}_{(0.5)}$ と定義される．この $\mathrm{SIQR_u}$ による次の基準を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．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 [97] 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 [4] 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．

2.5 提案手法の適用結果と考察

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 2.2 は手法 1，図 2.3 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．

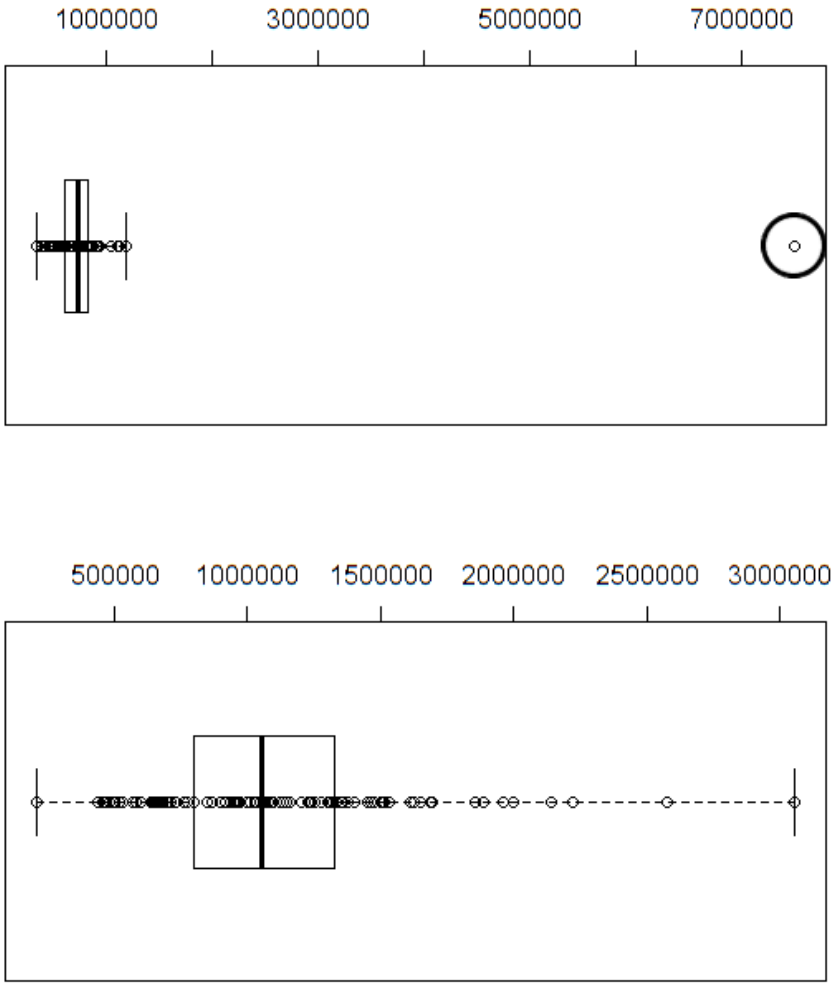

図 2.2: 手法 1 の適用結果

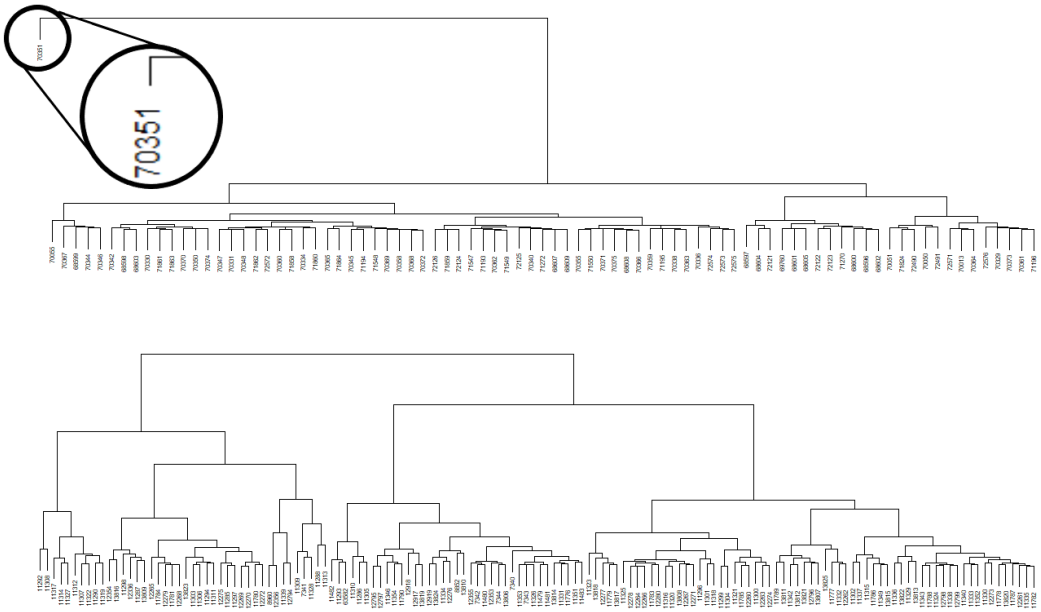

図 2.3: 手法 2 の適用結果．図中の数値はレコードの物件 ID を表す．

加えて各手法を適用した結果の個別例を表 2.3 に示した．

表 2.3: 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 $20m^2$ 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を取り上げた．この物件は 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 $\mathrm{SIQR_u}$ 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，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．

つぎに紹介した手法により異常値を検出した結果を表 2.4 に示す．表 2.4 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り，手法の判定の正確さを示す指標である．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り，正解（人間の異常値判定）の網羅性を示す指標として定義している．条件 A に該当するレコード数を $\#(A)$ と表すことにすれば，

$$\text{的中率} = \frac{\#(\text{人間の判定が異常値} \cap \text{提案手法が異常値})}{\#(\text{提案手法が異常値})} \tag{2.1}$$

$$\text{カバー率} = \frac{\#(\text{人間の判定が異常値} \cap \text{提案手法が異常値})}{\#(\text{人間が異常値})}. \tag{2.2}$$

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．

表 2.4: 各手法をデータセットへ適用した結果

属性	異常値検出数	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

なお，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は付録 $A.3$ に掲載した．

2.6 中古マンション価格モデルによるデータセットの改善効果

表 2.5: 線形回帰モデルの適用結果（括弧内は標準誤差，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	-5,198,847	-5,086,395	-5,118,729	-5,007,131	-5,197,480
	(1,090,939)	(635,185)	(613,661)	(673,909)	(611,263)	(676,042)
延床面積	759,817	728,995	725,807	728,970	724,692	730,080
	(3,138)	(1,828)	(1,767)	(1,942)	(1,761)	(1,948)
駅徒歩距離	-502,057	-477,979	-475,760	-477,148	-475,522	-477,394
	(20,042)	(11,666)	(11,272)	(12,398)	(11,236)	(12,428)
築年数	-491,624	-500,900	-499,504	-500,467	-500,803	-499,180
	(6,818)	(3,996)	(3,835)	(4,216)	(3,821)	(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

本節では桁の多い異常値レコードの除去が中古マンション価格モデル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[1)] など多くの論文で用いられている標準的な中古マンション価格モデルとして線形回帰モデルで検証を行う．推定に用いた線形回帰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線形回帰モデルを適用した結果を表 2.5 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．

2.7 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，標準的な中古マンション価格モデルである線形回帰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，線形回帰モデルのような標準的な中古マンション価格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，本研究の目的が中古マンション価格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減らし，的中率を向上させる余地はある．

一方，線形回帰モデルのフィッティング結果からは人間の判断と比べて，手法１，または手法１と手法２の併用の方が決定係数が良化しており，よりモデルの運用に望ましい異常値検出ができている可能性もある．

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，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本提案手法を適用しても，有効な結果が得られることは確認しているが，桁間違いエラーの発生要因に何らかの傾向が見られるかという点は継続して調査する必要がある．

第3章　東京２３区の中古マンション市場のデータ分割と統合

3.1　はじめに

第 3 章の研究では市況が変化する中で低下しやすい中古マンション価格モデルの予測精度をいかにして維持するか，という実務上の課題への工学的な解決を目的としている．

本章において中古マンション価格モデル，またはモデルとは本文 3.3 節の式 (3.1) に記載する統計モデルを指す．つぎに時間の経過を $\tau = 0, 1, 2, \cdots, t, \cdots$ とするとき，時点 t までに得られた中古マンションの価格と築年数などの属性データをインサンプルとして中古マンション価格モデル $y = f_t(x)$ を推定し，時点 t 以降のデータをアウトオブサンプルとして予測値 $\hat{y}$ を得ることを指す．また「予測精度」とは予測値 $\hat{y}$ に対応する観察値 y が得られる予測誤差 $y - \hat{y}$ を使用した MAPE(3.3 節の式 (3.2)) の値を指すものとする．

中古マンション価格モデルの予測精度を改善するためには予測対象となる中古マンションを考慮し，価格形成要因と考えられる変量を全てモデルに採用することが考えられる．ただし，東京２３区全体のようにある程度大きな範囲で１つのモデルを構築するとなると，マンションの立地条件ひとつとっても様々な要因を背景として取引されるため，価格形成要因となる変量を全て特定し，１つのモデルに導入することは現実的に難しい．

予測精度を改善するためのもう一つの方法として様々な地域が混在するデータを細分化し，できるだけ同質なデータに対してモデルの推定と予測を行うアプローチがあり，東京２３区の中古マンションデータであれば最寄り駅ダミーや町丁ダミーをモデルに導入しても予測精度は良化する．細分化のアプローチは柴田 [12] の著書における，データからモデルを構築する初手としてデータの非均質性の中に均質性を見出し，穏やかな変化を発見して簡潔な数式でモデリングする方法を参考としている．またモデルにダミー変数を用いる方法と発想は Suits [94] や林 [23] を参考とし，最寄り駅や町丁といった地域のダミー変数を用いることで中古マンションの買い手が当該地域のマンションをその価格で購入した意思決定の結果をモデルに反映できるという考えに基づいている．

しかし，過度に地域を細分化した場合，中古マンション取引が活発な地域は偏在するため，価格予測モデルの運用という観点から見ると，新たに取引が発生し

た地域の予測価格が更新されるものの，類似した近隣地域の予測価格は置き去りとなって更新されないことから，不動産査定の現場が一体として捉える地域のマンション価格の変化と，モデルが算出する当該地域の予測価格の間にしばしば大きな乖離が発生する．これが本研究で取り上げる中古マンション価格モデルの予測精度の劣化問題である．

この予測精度の劣化問題は，モデル推定のための取引実績データをどの程度の地域まで分割すれば現場の感覚との乖離が無く，予測精度が劣化しないモデルの実装を実現できるかという，工学的な問題として読み替える．そして，本研究ではこの問題を，想定する最小単位でのデータ分割と予測精度が落ちない形でのデータの再統合と捉えなおすことで解決する．たとえば，図3.1のように最小単位として最寄り駅単位のデータ分割を基礎にしたシステムを想定する．仮に類似した価格形成メカニズムをもつA，B，Cという３つの最寄り駅の取引データを１つのデータに再統合してモデルを推定した場合，A駅付近で新しい取引があったことを新グループの中に起こった事象として扱えることができるようになる．これは，実際には取引のなかったB駅，C駅でのマンション価格の予測にもA駅の新たな取引データの結果を反映できるようになることを意味する．もちろん，A，B，Cの３つを統合することによりモデルの予測性能が落ちるのであれば無意味なデータ統合であるが，モデルの予測性能があまり劣化しないのであれば１つの地域グループに再編することが可能となり，データの再統合によって劣化問題が解決できる．

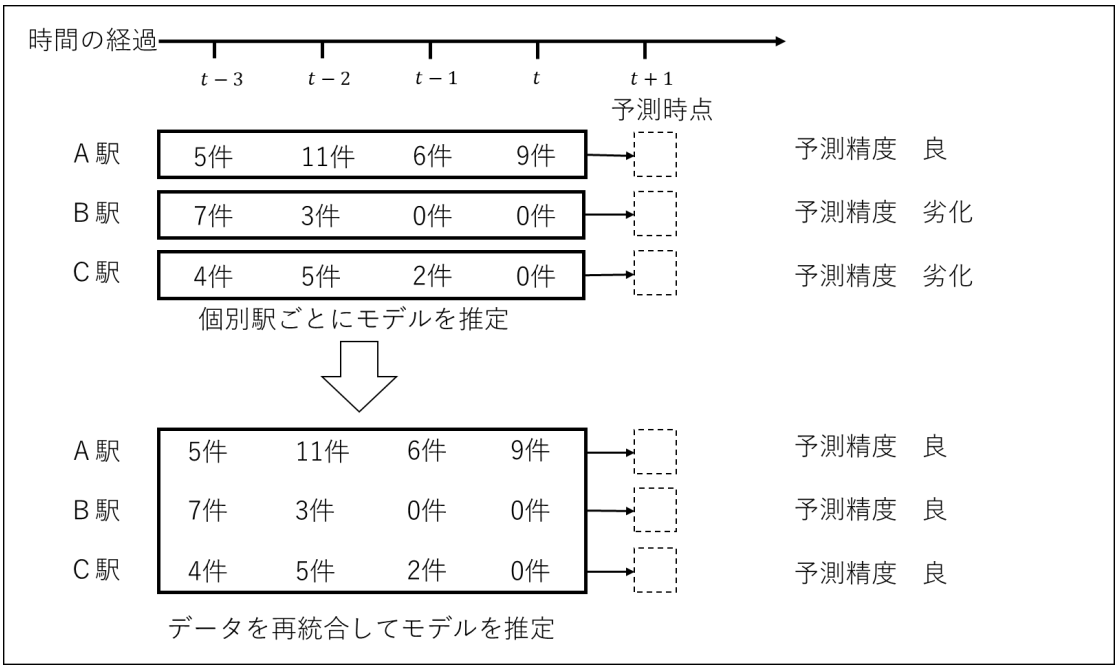

図 3.1: より大きな１つの地域へ再編する例

本稿では次の3.2節でデータの細分化に関する既存研究を紹介しつつ，本研究の

位置づけの違いを説明する．3.3 節では我々の提案するデータ細分化と再統合の方法を説明し，3.4 節では地域性の粒度の違いによる分割効果を比較する．3.5 節では細分化した地域を再統合する方法を実例によって説明し，3.6 節では予測精度の劣化問題に対処した例を示す．3.7 節を結論とする．

3.2 先行研究

住宅市場の研究においてデータの細分化はサブマーケットやマーケットセグメンテーションとして長く取り扱われてきたテーマである．ただし，本研究はヘドニックアプローチを裏付けとした経済学的なサブマーケットの定義や理論は議論せず，価格モデルの予測精度の劣化に対処するため，データの最適な粒度の発見フローを確立する，という工学的な位置づけに置く．この観点から既存研究を見ればデータへのアプローチの違いから大きく２つに分けられる．

まずデータに収録された物件に関する属性情報をもとに直接，細分化する研究が挙げられる．Dale-Johnson [48] はカリフォルニア州サンタクララ，Bourassa et al. [35] はシドニーとメルボルン，Bourassa et al. [36] はオークランド，Keskin and Watkins [72] はイスタンブールの住宅データを対象に築年数などの物件属性を用いて主成分分析や因子分析を行い，算出した主成分スコアや因子などを評価してデータを分割している．また Kauko et al. [71] は自己組織化マップと学習ベクトル量子化というニューラルネットワークによってヘルシンキの住宅データの細分化を試みている．これらの方法は物件属性の特徴を直接反映してデータの細分化ができるものの，物件属性を重み付けして組み合わせた合成変量（または因子）の解釈が難しく，不動産価格推定モデルの予測精度への効果も明確ではない．

もう一つの方法としてデータから推定した不動産価格のモデルを通じてデータを分割するアプローチが挙げられる．Schnare and Struyk [88] は地域性や部屋数などの違いによってデータを分割した場合としない場合のそれぞれで線形モデルを推定し，価格推定値の標準誤差の比較，ならびに残差のＦ検定によって細分化の可否を判断している．また Goodman and Thibodeau [58]，Goodman and Thibodeau [59] は地理的な隣接性を重視して，テキサス州ダラスの住宅市場を対象に小学校の学区や ZIP コードに基づく，隣接する地域境界を階層モデルに反映し，予測誤差やＦ検定，Ｊ検定を用いて分割可否を判定している．なお，彼らはその後の論文 Goodman and Thibodeau [60] で地理的に隣り合わない地域でもサブマーケットは形成しうるとも主張している．日本の研究では Islam and Asami [68] が東京２３区の住宅市場における空間的不均一性と依存性を指摘しつつ，空間スイッチング回帰を用いて東京の中心部と周辺の郊外の２つの地域に分けられることを示している．また研究目的がデータの細分化ではないものの，Shimizu and Nishimura [91] は路線ダミーを，Shimizu et al. [90] では区と路線ダミーを線形モデルに用いて地域性の違いをコントロールしている．このようにモデルを通じた細分化はモデルの推定

精度や予測精度の良化に直接つながるものの，細分化した結果の取引データの発生頻度，ひいてはモデルの更新頻度という観点は考慮されておらず，本研究が目的とする中古マンション価格モデルの予測精度の劣化問題は議論されていない．

本研究で提案する方法は，市況が変化する中でモデルの更新がされにくい地域の予測精度の劣化問題に対処するため，細分化したデータを再び統合する方法を取り上げている点が既存研究とは異なる．我々の方法を採ることで，データを再統合した後の地域グループはモデルの予測精度を劣化させないだけの新規発生データの取り込み頻度を維持することができる．

3.3 研究の方法

3.3.1 分析の方法

研究では 3.1 節にて説明した中古マンション価格モデルの予測精度の劣化問題に対して，地域単位でのデータ細分化により予測精度の改善を検証したのち，モデルの性能をなるべく維持した形でデータを再統合することによって解決する．この方法の検証のため，データを２つの期間に分け，過去のデータでモデルを推定し，後の時点のデータの価格予測を行う．ただし，不動産取引の発生という事象は容易にコントロールできないため，実際のデータ更新とそれによるモデルの再計算という点は取り扱わず，データの分割と再統合により，一定のサンプルサイズを確保し，予測の劣化問題に対処できるか調査する．

まず東京２３区の中古マンション取引データを異なる粒度で細分化する効果を検証するため，先に挙げた Shimizu and Nishimura [91]，Shimizu et al. [90] に倣い，モデルに地域単位のダミー変数を導入する方法を採用する．具体的にはまず東京２３区の中古マンションの価格水準へ大きく影響する，区，最寄り駅，町丁という異なる地域的な要素をそれぞれ中古マンション価格モデルへダミー変数として導入し，細分化によるモデルの推定精度と予測誤差への効果を比較する．

中古マンションは土地部分の価格への影響が大きい戸建てと比べ，物件の構造的な同質性が高いと考えられるため，構造的な要因による細分化は行わない．また，これらの地域性の違いをモデルへ反映するには地域ごとにモデルを推定したり，地域を表すダミー変数を他の説明変量の交差項として取り入れたり，と様々な方法が考えられるが，細分化した地域によっては，説明変量と比べてサンプルサイズが小さくなりすぎるケースを考慮し，説明変量の極端な増加を抑えつつ，検証結果の解釈性が高い方法としてシンプルに地域性を表すダミー変数をモデルへ用いる方法を採る．なお，区と町丁が行政界であるのに対し，最寄り駅は物件の地図上の位置から決まる属性という意味的な違いがあるものの，データの各取引レコードに一つ付される地域的なカテゴリ，という視点で捉えた場合は同質のものとみなせる．また中古マンション売買を扱う大手不動産会社のインターネットサイトでは住みたい街ランキングとして地域の代表に駅が用いられており，本稿

では物件の地域的な属性を表す要素として区と最寄り駅と町丁のダミーを同列に扱う．

次に細分化したデータを再び統合し，一定のサンプルサイズの確保が可能か調査する．本研究では実務での利用を想定し，よりモデルの仮定が少なく，説明力の高いシンプルな手法が望ましいことから，中古マンション価格モデルのあてはめ精度を基準とした方法を採用する．またデータの統合は異なる町丁単位で行うことも可能であるが，今回用いたデータは細分化した際のサンプルサイズがそれほど大きくないことから，異なる最寄り駅単位のデータを統合する．このように類似した地域を統合して新グループを形成することを本研究ではデータの再統合と呼ぶ．

具体的にデータの再統合は次の手順にて行う．

1. 最寄り駅が同じデータごとに細分化したのち，直近の取引データが少ない，またはデータが無い駅を予測精度の劣化が生じている可能性のある駅として特定する．

2. 特定した駅のデータに同じ路線の隣接駅のデータを加えて最寄り駅のダミー変数を加えたモデルとダミー変数を加えていないモデルを推定する．

3. 最寄り駅ダミー変数の有無によってモデルを比較し，決定係数にあまり差がなければそれらの最寄り駅はダミー変数を用いず，一つの地域とみなして差し支えないと判断する．今回用いたデータではデータの再統合基準として，決定係数の閾値を 5% に設定し，ダミー変数の有無で 5% 未満の差しか生じない場合は統合可能と判断する．

このデータの再統合によって一つの地域としてレコードの件数が一定以上あれば，予測精度を維持しながらより高い頻度でデータを取得し，モデルを更新することができ，劣化問題が解決できる．

3.3.2 データ

本研究で用いた中古マンションデータはＳＲＥホールディングス株式会社より提供を受けた実際の成約データである．

表 3.1: 使用するデータ

種類	平均値／収録例	標準偏差	最大値	最小値
成約年月日	2015-11-02	-	-	-
都道府県名	東京都	-	-	-
所在地名 1	品川区	-	-	-
所在地名 2	南大井２丁目	-	-	-
沿線略称	京急本線	-	-	-
駅名	大森海岸	-	-	-
駅徒歩（分）	7.1	4.4	44.0	1.0
成約価格	35,106,445	23,050,473	405,690,000	765,715
成約 m^2 単価	624,096	269,209	2,860,621	16,421
専有面積	55.71	21.17	300.00	7.99
築年月	1978-07-01	-	-	-
地上総階数	11	8	60	1

3.3.3 モデルと評価

中古マンション価格モデルの推定を通じて粒度の異なるデータへの分割効果，ならびにデータの統合効果を検証するため，地域性を表すダミー変数を用いた線形モデルを採用する．

$$y = \beta_0 + \sum_{i=1}^{4} \beta_i x_i + \sum_{l=1}^{k} \gamma_l z_l + \varepsilon. \tag{3.1}$$

y: 中古マンション m^2 当たり価格

x_i: 延床面積（m^2），駅徒歩距離（分），築年数（年），建物総階数

z_l: k 個の地域ダミー（区，最寄り駅，町丁によって数が異なる）

また本研究では予測精度を MAPE によって測るものとする．MAPE は n 件のデータの予測値を $\hat{y}_i, i = 1, \cdots, n$，観測値を y_i として次のとおり定義する．

$$\mathrm{MAPE} := \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|. \tag{3.2}$$

3.4 データ細分化の結果

3.4.1 推定したモデル

まず 2011 年 1 月から 2018 年 12 月までのデータに対し，区単位，最寄り駅単位，町丁単位のダミー変数を用いたそれぞれの場合，およびダミー変数を用いない場

合について，モデルを推計した結果とその当てはめ残差をそれぞれ表 3.2 に記載した．ダミー変数を用いない場合と比べ，いずれかのダミー変数を用いた場合の調整済決定係数は大きく改善した．ただし，ダミー変数を用いた場合の結果から，粒度を細かくするほど当てはまりは良化するが，使用するダミー変数の増加と比べ，その良化幅は逓減していることが分かる．

表 3.2: 地域ダミー変数の違いによるモデル比較 (各変量の数値は回帰係数，括弧内は標準誤差)

変量	区単位	最寄り駅単位	町丁単位	ダミー無し
切片	563,161	539,556	523,935	834,246
	(10,123)	(16,826)	(92,665)	(8,413)
延床面積	339	294	75	-378
	(76)	(73)	(83)	(101)
駅徒歩距離	-7,748	-7,145	-3,571	-14,369
	(379)	(388)	(627)	(482)
築月数	-773	-796	-782	-729
	(10)	(10)	(12)	(14)
地上総階数	5,925	7,248	7,737	7,539
	(225)	(241)	(320)	(283)
基準のダミー	葛飾区	お花茶屋	お花茶屋１丁目	-
（各係数は略）	他 22 変量	他 440 変量	他 2105 変量	
観察数	10,224	10,224	10,224	10,224
調整済 R^2	0.658	0.718	0.744	0.360

3.4.2 推定したモデルでの予測結果

前項で推定したそれぞれのモデルを用いて，2019 年 1 月から 2020 年 12 月までのデータに対する価格予測を行った．表 3.3 は推定したモデルで予測できた範囲，表 3.4 は MAPE である．表 3.3 の結果からダミー変数を最寄り駅単位で導入した場合と町丁単位で導入した場合は予測ができないレコードが発生している．また，より粒度の細かいダミー変数を用いるほどモデルで推定していない地域が増え，予測できないレコード数が逓増していることも分かる．一方，表 3.4 の結果を見れば，ダミー変数を用いた方が全体的に良い予測が得られていると判断できる．また町丁単位のダミー変数と最寄り駅単位のダミー変数を用いた予測結果を比較すれば，最寄り駅単位の方が良化している．これらの結果からデータの粒度の細分化はモデルの推定精度と説明力を高め，モデルによる予測精度も改善するが，過度な細分化はモデル推定時に過剰適合を起こしてしまい，予測において対象外となるダミー変数も発生しやすくなると考えられる．

表 3.3: 地域ダミーの違いによるモデルの予測数

	区単位	最寄り駅単位	町丁単位	ダミー無し
予測対象数	2,579	2,579	2,579	2,579
予測不可数	0	3	167	0
予測不可ダミー	0	3	132	0
予測可能数	2,579	2,576	2,412	2,579

表 3.4: 地域ダミーの違いによるモデルの予測誤差

	区単位	最寄り駅単位	町丁単位	ダミー無し
MAPE	19.3%	18.5%	19.0%	26.4%

3.5 データ統合の判定例

この節では最寄り駅単位で細分化したデータの統合を判定する例を示す．表 3.5 は東京メトロ東西線の葛西駅から茅場町駅まで，隣接する各駅のペアに対してデータの統合可否を判断した結果である．期間は 2011 年 1 月から 2018 年 12 月までのデータである．葛西駅から木場駅までの隣接する各ペア，ならびに門前仲町と茅場町のペアは最寄り駅のダミー変数を加えたモデルとダミー変数を加えないモデルとの間で決定係数にほとんど差がなく，データを再統合できる可能性があることを示唆している．

一方，木場駅と門前仲町駅のペアでは最寄り駅のダミー変数を加えたモデルとダミー変数を加えないモデルで決定係数の差が 12% 近く生じており，ダミー変数を用いて地域による違いを考慮する効果が大きく，葛西駅から木場駅までと門前仲町から茅場町は統合できないことが判定できる．

表 3.6 は実際に葛西駅から木場駅までの５駅を一つの地域として扱い，最寄り駅ダミー変数の有無を比較した．いずれの決定係数も同程度であり，ダミー変数の有無によってモデルで説明できる部分には差が無いと判断できる．

表 3.5: 隣接最寄り駅ペアによるデータの再統合の判定

変量	葛西∪西葛西	西葛西∪南砂町	南砂町∪東陽町	東陽町∪木場	木場∪門前仲町	門前仲町∪茅場町
切片	684,838	738,786	653,630	684,273	756,594	978,593
	(43,809)	(44,720)	(67,472)	(66,633)	(80,184)	(151,935)
延床面積	-993	-1,162	-965	-994	-1,678	-306
	(462)	(481)	(754)	(794)	(1,011)	(1,038)
駅徒歩距離	-4,944	-8,701	-6,803	-7,655	-8,428	-13,718
	(1,221)	(1,704)	(1,811)	(2,195)	(4,098)	(9,326)
築月数	-457	-576	-623	-668	-765	-1,083
	(40)	(52)	(67)	(70)	(108)	(155)
地上総階数	1,914	5,634	8,582	11,191	12,297	9,753
	(1,807)	(1,694)	(2,248)	(3,263)	(5,674)	(11,557)
西葛西	13,510					
	(12,151)					
南砂町		-23,195				
		(15,653)				
東陽町			40,193			
			(18,028)			
木場				7,906		
				(20,003)		
門前仲町					145,681	
					(32,995)	
茅場町						1,347
						(51,898)
観察数	122	122	128	117	73	55
調整済 R^2(駅ダミー)	0.542	0.635	0.602	0.572	0.578	0.479
調整済 R^2(無し)	0.541	0.632	0.589	0.575	0.463	0.489

表 3.6: データの再統合効果の確認

変量	駅ダミー有り	データの再統合
切片	674,276	706,503
	(36,266)	(32,555)
延床面積	-946	-1,073
	(396)	(377)
駅徒歩距離	-6,265	-6,762
	(1,064)	(1,049)
築月数	-579	-573
	(36)	(36)
地上総階数	6,594	5,745
	(1,434)	(1,320)
基準のダミー	葛西	-
(各係数は略)	他 4 変量	
観察数	299	299
調整済 R^2	0.555	0.551

3.6 データ統合の結果とその効果

データの再統合によって予測精度の劣化問題に対処している例として 2010 年代中頃からの価格上昇が大きい湾岸エリアと相対的に価格上昇が穏やかな城北エリアの２つを示す．

表 3.7 は東京メトロ有楽町線の新富町駅から辰巳駅まで，隣接する各駅のペアに対して前項の方法を用いてデータの再統合可否を判断した結果である．期間は 2011 年 1 月から 2015 年 12 月までのデータである．表 3.7 の結果によれば，新富町駅から豊洲駅までの隣接するペアは最寄り駅のダミー変数を加えたモデルとダミー変数を加えないモデルとの間で決定係数にほとんど差がなく，データを再統合できる可能性があることを示唆している．一方，豊洲駅と辰巳駅のペアでは最寄り駅のダミー変数を加えたモデルとダミー変数を加えないモデルで決定係数の差が 12% 近く生じており，データを再統合できないと判断できる．

表 3.8 はデータの統合効果を確認するために新富町駅から豊洲駅までの３駅を最寄り駅とする取引レコードに対し，最寄り駅のダミー変数を加えた場合とデータを再統合した場合のモデル推定の結果を示している．いずれの決定係数も同程度であり，ダミー変数の有無にかかわらず，モデルで説明できる部分には差が無いと判断できる．

表 3.9 は新富町駅，月島駅，豊洲駅を最寄り駅とするデータのサンプルサイズである．他の２駅と比べて新富町は 2012 年から 2015 年のサンプルが少なくなっている．

表 3.10 は 2011 年 1 月から 2015 年 12 月までのデータから推定したモデルによって 2016 年 1 月から 2016 年 12 月までのデータを予測した結果の MAPE である．左側は個別の駅ごとにそれぞれモデルを推定し，それぞれの駅のデータで MAPE を算出しているが，新富町は他の２駅と比べて MAPE の値が大きくなっている．右側はこれら３つの駅を一つの地域としてデータの再統合を行い，１つのモデルを推定して予測を行った上で比較のためにそれぞれの最寄り駅ごとに MAPE を計算している．新富町の MAPE は個別の最寄り駅ごとにモデルを推定する場合と比べて値が小さくなっていることが分かる．

表 3.7: 新富町駅から辰巳駅までの統合判定

変量	新富町∪月島	月島∪豊洲	豊洲∪辰巳
切片	626,012	613,660	494,683
	(64,217)	(52,646)	(134,588)
延床面積	64	1,020	650
	(819)	(690)	(1,410)
駅徒歩距離	-10,782	-17,899	-20,795
	(5,887)	(3,678)	(5,097)
築月数	-405	-278	-301
	(128)	(108)	(166)
地上総階数	7,523	7,109	5,900
	(1,192)	(801)	(1,254)
月島	36,300		
	(39,846)		
豊洲		4,191	
		(28,745)	
辰巳			208,175
			(50,011)
観察数	46	62	30
調整済 R^2(駅ダミー)	0.676	0.790	0.822
調整済 R^2(無し)	0.677	0.793	0.706

表 3.8: データを再統合したモデルの推定

変量	駅ダミー有り	データの再統合
切片	663,562	644,483
	(53,578)	(50,714)
延床面積	337	564
	(690)	(656)
駅徒歩距離	-14,487	-14,382
	(3,789)	(2,830)
築月数	-373	-401
	(100)	(94)
地上総階数	7,003	7,043
	(854)	(847)
新富町	-42,640	−
	(35,303)	
豊洲	-7,868	−
	(30,590)	
観察数	73	73
調整済 R^2	0.747	0.749

表 3.9: 各最寄り駅のサンプルサイズ

	2011	2012	2013	2014	2015	2016
新富町	3	0	3	2	3	5
月島	3	9	11	5	7	11
豊洲	4	4	3	2	14	14
計	10	13	17	9	24	30

表 3.10: 最寄り駅ごとのモデルとデータを再統合したモデルでの MAPE 比較

	MAPE	
	個別駅ごとのモデル	データを再統合したモデル
新富町	27.9%	21.0%
月島	15.2%	16.3%
豊洲	12.2%	11.4%

同様に表3.11は東武伊勢崎線の北千住駅から西新井駅まで，隣接する各駅のペアに対してデータの再統合可否を判断した結果である．期間は2011年1月から2015年12月までのデータである．

表 3.11 の結果においては小菅駅から梅島駅までの隣接ペアはデータを再統合できる可能性があることを示唆している．一方，北千住駅と小菅駅のペア，ならびに梅島駅と西新井駅のペアは決定係数の差が 5% 以上あるため，データを再統合できないと判断できる．特に北千住駅と小菅駅のペアでは決定係数は 25% 近く異なり，価格帯の隔たりも大きい．

表 3.12 はデータの再統合効果を確認するために小菅町駅から梅島駅までの３駅を最寄り駅とする取引レコードに対し，最寄り駅のダミー変数を加えた場合とデータを再統合した場合のモデル推定の結果を示している．いずれの決定係数も同程度であり，ダミー変数の有無にかかわらず，モデルで説明できる部分には差が無いと判断できる．

表 3.13 は梅島駅，五反野駅，小菅駅を最寄り駅とするデータのサンプルサイズである．他の２駅と比べ，梅島駅は 2015 年のサンプルが無い．

表 3.14 は表 3.10 と同様に 2011 年 1 月から 2015 年 12 月までのデータから推定したモデルによって 2016 年 1 月から 2016 年 12 月までのデータを予測した結果の MAPE である．左側は個別の駅ごとにそれぞれモデルを推定し，それぞれの駅のデータで MAPE を算出しているが，梅島駅は他の２駅と比べて MAPE の値が大きいが，右側のデータの再統合の結果を見れば梅島駅の MAPE は個別の最寄り駅ごとにモデルを推定する場合と比べて値が小さくなっていることが分かる．

表 3.11: 北千住駅から西新井駅までの統合判定

変量	北千住∪小菅	小菅∪五反野	五反野∪梅島	梅島∪西新井
切片	540,430	321,427	479,318	485,039
	(57,738)	(62,833)	(60,764)	(73,079)
延床面積	403	158	-394	-178
	(604)	(695)	(718)	(886)
駅徒歩距離	-14,191	-3,765	-6,952	-8,635
	(2,343)	(3,472)	(1,799)	(1,894)
築月数	-269	-466	-441	-511
	(79)	(94)	(73)	(82)
地上総階数	4,984	10,227	2,942	2,695
	(2,525)	(3,718)	(2,786)	(3,071)
小菅	-201,200			
	(29,118)			
五反野		30,681		
		(30,546)		
梅島			-15,394	
			(16,129)	
西新井				47,601
				(17,405)
観察数	39	23	41	51
調整済 R^2(駅ダミー)	0.760	0.699	0.607	0.623
調整済 R^2(無し)	0.411	0.699	0.608	0.570

表 3.12: データを統合したモデルの推定

変量	駅ダミー有り	データの再統合
切片	379,837	399,460
	(51,086)	(51,547)
延床面積	-13	-44
	(591)	(606)
駅徒歩距離	-6,699	-5,056
	(1,750)	(1,654)
築月数	-424	-430
	(70)	(72)
地上総階数	3,810	5,470
	(2,691)	(2,650)
五反野	60,582	
	(26,184)	
梅島	46,533	
	(25,156)	
観察数	47	47
調整済 R^2	0.608	0.577

表 3.13: 各最寄り駅のサンプルサイズ

	2011	2012	2013	2014	2015	2016
梅島	5	7	5	7	0	15
五反野	1	3	5	4	4	7
小菅	1	1	2	1	1	1
計	7	11	12	12	5	23

表 3.14: 最寄り駅ごとのモデルとデータを統合したモデルでの MAPE 比較

	MAPE	
	個別駅ごとのモデル	データを再統合したモデル
梅島	22.1%	19.7%
五反野	19.4%	19.9%
小菅	6.9%	3.0%

3.7 結論

本研究は，市況が変化する中で低下しやすい中古マンション価格モデルの予測精度をいかにして維持するか，という実務上の課題に対して，地域単位でのデータ細分化により予測精度の良化を検証したのち，モデルの予測精度を維持しながらデータを再統合する，という方法で解決している．

研究では東京２３区の中古マンション取引データを用いて物件の所在地域をベースとしたデータの細分化を実施し，細分化の程度とあてはめたモデルの推定精度と予測精度を比較することで，データの細分化とモデルフィッティングの精度との関係を数値的に調査した結果を示した．また，データ細分化が過ぎることによって生じる価格推定モデルの予測精度の劣化に対処する実例を示した．さらには予測精度の劣化問題が顕在化する前に類似した地域を統合しておくことで十分なサンプルサイズを確保し，モデル推定の精度とサンプルサイズの減少の間における適切なトレードオフを実現し，実務的な日々のデータ更新に耐えうる中古マンション価格モデルの構築に役立つと考える．

ただし，本研究で採用した，最寄り駅が同じレコードを統合して一つの地域を構築する方法は東京２３区の中古マンション市場を考慮したものであり，他の地域への応用は十分に検証する必要がある．また今回は，隣接する地域以外の統合を取り上げていないが，クラスタリングによる統合手法の研究を進めている．

さらに本研究ではデータを再統合した地域の時系列的な変化は考慮しておらず，再開発や新駅の開業などによって，統合した地域が変化する可能性もあるため，この点についても引き続きの課題とする．

第4章　可変ウィンドウによる中古マンション価格モデルの更新

4.1　はじめに

中古マンション取引の実務において，物件オーナーあるいは物件購入希望者が売買の見込み額を知るため，不動産会社各社の WEB サイトに用意された価格査定サービスを利用するようになって久しい．多くの価格査定サービスは機械学習や統計学の手法をベースとして築年数や最寄り駅までの距離，地域など，物件に付帯する属性情報から推定したマンション価格モデルを用いており，モデルの推定に際しては直近の価格動向を反映するため，現時点から過去に対してある幅を持った期間（以降はウィンドウと呼ぶ）の中で観察されたデータを用いることが多い．

本研究では期間の幅が一定の長さを持ち，時間の経過とともに新しいデータを追加ながら古いデータを削除するウィンドウのことを固定スライディングウィンドウと呼ぶ．固定スライディングウィンドウではウィンドウの長さについてあらかじめ単一の値を設定する必要があるが，後述するトレードオフ問題が存在するため，ウィンドウの長さの設定は容易ではない．一般にウィンドウを長く設定すると，モデル推定のためのサンプルサイズが大きくなるため，中古マンション価格が経時的に緩やかに動く場合は予測精度が高くなり長期的に安定する．しかしながら，金融政策の変更，都市の再開発，法律の変更など，様々な要因を背景として中古マンションの価格は短期間に大きく変動する場合があるため，モデルの追従が遅れれば予測精度が低下する問題を引き起こす．一方，ウィンドウを短くすると，中古マンション価格の急な変動に即応的に追従できるものの，モデルの推定に用いるサンプルサイズが小さくなるため，しばしば紛れ込む外れ値によって予測精度が安定しない．不動産会社の価格査定サービスが信頼され，安心して利用できるようにするためには，予測精度を安定させることが求められる．

そこで我々は長短の固定スライディングウィンドウでモデルを推定した場合のそれぞれのメリットを取り入れるため，観察されるデータに対して適応的にウィンドウの長さを決定する可変長のスライディングウィンドウ（以降は可変ウィンドウと呼ぶ）による中古マンション価格モデルの更新方法を提案する．

なお，第 4 章において中古マンション価格モデルおよび不動産価格モデルとはのちの 4.4.2 項で説明するような統計モデルを指すものとする．また本章において

中古マンション価格モデルによる予測とは 4.3 節で説明するようにある時点までに得られたデータによって推定した中古マンション価格モデルを用いて，その時点より先の時点で観察される説明変量によって観察されていない被説明変量を算出することを指し，予測精度とは予測において算出した予測値に対し，時間の経過によって実際に得られた観察値との差である予測誤差の大小を指すこととする．つまり，予測精度が高いとはこの予測誤差が小さいことを意味するものとする．

4.2 先行研究

不動産に関する研究において，固定スライディングウィンドウを採用して不動産価格モデルを更新している例としては Shimizu et al. [92]，Shimizu et al. [93]，Hill et al. [64] が挙げられる．彼らは市況の変化を反映した不動産価格モデルを通じて不動産価格指数を算出している．また，住宅価格の予測のために固定スライディングウィンドウによって多変量の線形回帰モデルを推定している例として Gao et al. [56] がある．ただし，日本の不動産を対象とし，不動産価格モデルの推定に用いるデータの期間の幅を調整する先行研究は筆者らが調査した限り見当たらない．

本研究で取り上げる可変ウィンドウのアイディアはネットワークトラフィックの予測やデータストリームマイニングなど，時間の経過とともに持続的にデータが観察され，変化の察知や予測が求められる分野で研究が進んでいる．これまでも Bifet and Gavalda [34]，Baig et al. [29]，Dalmazo et al. [49] など，多くの研究報告が挙げられている．Bifet and Gavalda [34] は時間の経過とともに新たなデータが追加される固定スライディングウィンドウを 2 つに区切り，古いウィンドウと新しいウィンドウの平均の差を算出し，その差が一定の閾値を超える場合に古いウィンドウ部分を切り捨てる可変ウィンドウの手法を提案している．彼らは電力使用量のデータを用いて分析を行い，可変ウィンドウを用いる予測は固定スライディングウィンドウを用いる予測より良好な結果が得られたと報告している．Baig et al. [29] はデータセンターのリソース管理のために深層学習を用いてリソースの予測に最適なウィンドウの長さを学習する手法を提案し，実際のデータセンターのワークロードデータを用いて，固定スライディングウィンドウと比較し，可変ウィンドウを用いる方法の方がリソースの予測精度が向上することを示している．Dalmazo et al. [49] はクラウドコンピューティングなどにおけるネットワークのトラフィック量を予測するため，時間の経過に対して動的にウィンドウの長さを選択するアルゴリズムを提案している．彼らはある時点でのウィンドウ内のトラフィック量データと次の時点でのウィンドウ内のトラフィック量データの平均と分散に基づいてウィンドウの長さを変更するアルゴリズムを提案し，固定スライディングウィンドウを用いる場合と比べ，トラフィック量の予測精度が改善することを示している．

一方，資産価格に関連する研究で可変ウィンドウを取り上げているのは Jeon and McCurd [69] があり，株価の実現ボラティリティ，さらに株価とコモディティの実

現相関をベイズモデルによって予測する際，モデル推定に用いるデータを時点によって変化させる方法を提案している．彼らはモデル推定に使うデータの重みづけを推定時毎に決定する方法とモデル推定に使うウィンドウの長さを推定時毎に決定する方法の２つを提案し，モデルの推定に固定スライディングウィンドウを用いる方法や観察されたデータの全てを用いる（新たなデータを取得するたびにウィンドウを拡張させる）方法と比べ，彼らの提案手法は予測誤差において，より良好な結果が得られたことを示している．

ただし，ここまで挙げた先行研究はいずれもトラフィック量や株価など，目的とする変量そのものによって予測を行う，単変量型のモデルである．したがって，本研究のように不動産価格をその不動産の属性情報で説明する多変量型のモデルには，先行研究の可変ウィンドウのアルゴリズムをそのまま適用することは出来ない．

Yoshida et al. [101] は長さの異なる複数の固定スライディングウィンドウを用いてそれぞれ線形回帰モデルを推定し，その予測誤差によって各固定スライディングウィンドウの重み付けを更新するアルゴリズムを提案している．彼らは日経平均株価の日次終値を用いて分析を行い，Bifet and Gavalda [34] の方法などに比べ，良好な予測精度が得られたことを報告している．ただし，彼らの方法は 1 期間に観察されるデータが 1 つであることを前提とした方法であり，中古マンション取引のようには 1 期間に観察されるデータが複数個ある場合には，彼らの重み付けの方法をそのまま適用することは難しい．また複数の固定スライディングウィンドウから推定したモデルの予測にそれぞれ重み付けする手法は時間の経過とともにウィンドウの数が増大して計算負荷も大きくなり，また予測結果の解釈も複雑になりやすい．

本研究の貢献は多変量型のモデルである中古マンション価格モデルの予測精度を維持するため，新たに観察されるデータに応じてモデル推定のウィンドウの長さを逐次的に決定する可変ウィンドウのスキームをシンプルな方法によって計算負荷の軽さを実現しながら考案した点，そして東京２３区の中古マンション取引データを用いて我々の提案する可変ウィンドウと既存の固定スライディングウィンドウでの比較分析を行っている点である．

4.3 問題の設定

本節では既存の方法である固定スライディングウィンドウによる中古マンション価格モデルの更新について説明したのち，固定スライディングウィンドウのウィンドウの長短によるメリットとデメリットのトレードオフをグラフを用いて示し，問題点を明らかにする．

まず本研究で用いる記号の導入を行う．時間の経過を $T = 0, 1, 2, \ldots, t, \ldots$ とし，各時点の最小単位の間隔を ΔT とする．このとき時点 $t-1$ から時点 t までの期間

$[t-1,t)$ に観察された中古マンション取引データの件数を n_t 件，築年数など中古マンションの属性 $\boldsymbol{x}_j$ と価格 y_j の組を $\{(\boldsymbol{x}_j, y_j)\}_{j=1}^{n_t}$ と表す．また時点 t において推定されるモデルを f_t，そのモデルを推定するデータが含まれるウィンドウを W_t と表記する．

固定スライディングウィンドウによるモデル推定では現在時点 t から過去方向に対し，あらかじめ定めた長さ l[1]のウィンドウ W_t，つまり期間 $[t-l,t)$ のデータを用いてモデル f_t を推定する．その後，時間が経過して時点 $t+1$ に至るとそれまでの経過期間 ΔT において新規に観察されたデータ n_{t+1} 件をウィンドウ W_t に追加し，同時に古い期間のデータ n_{t-l+1} 件を切り捨てることで，ウィンドウ W_t からウィンドウ W_{t+1} へスライドする．このウィンドウ W_{t+1} によってモデルを f_{t+1} に更新することでデータを通じて市況の変化をモデルに反映する．

図 4.1 では固定スライディングウィンドウによって中古マンション価格モデルを更新し，予測を行う例を示した．４期間分のデータで中古マンション価格モデルを推定する場合，図 4.1 の 3 つの矩形の上段のように時点 $t-5$ から時点 $t-1$ までの長さ $l=4$ のウィンドウ W_{t-1} の中で観察された $(n_{t-4}+n_{t-3}+n_{t-2}+n_{t-1})$ 件のデータでモデル f_{t-1} を推定する．次にこのモデル f_{t-1} を用いて時点 $t-1$ から時点 t に発生する売却希望の中古マンションの予測を行うため，中古マンションの属性である $\boldsymbol{x}_j, j=1,2,\ldots,n_t$ を用いて，予測値 $\hat{y}_j$ を得る．この中古マンションが成約すれば実際の観測値 y_j が得られるため，予測誤差 $y_j-\hat{y}_j$ を評価できる．その後，時点 t を超えるとウィンドウ W_t 内の $(n_{t-3}+n_{t-2}+n_{t-1}+n_t)$ 件のデータによって更新したモデル f_t を用い，時点 t から時点 $t+1$ までに発生する中古マンション価格の予測を行う．

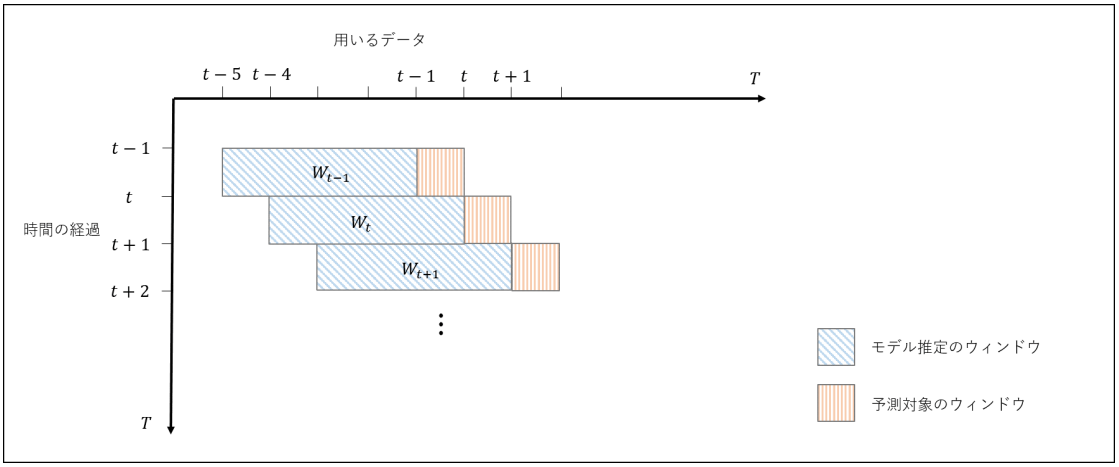

図 4.1: 固定スライディングウィンドウによるモデルの更新

固定スライディングウィンドウによって推定されたモデルは設定するウィンド

[1] $l=k\times\Delta T$，ただし，k は自然数とする．

ウの長さ l によって市況変化への対応に違いが出る．図 4.2 は東京都江東区の中古マンションデータに対し，ウィンドウの長さ $l = 0.5$ 年と $l = 5.0$ 年の固定スライディングウィンドウのそれぞれで価格モデルを推定し，そのモデルを用いて予測対象期間のデータに適用した際の予測誤差率の絶対値平均 (後の 4.5 節にて定義) を示している．ここに固定スライディングウィンドウのウィンドウ l が短い場合と長い場合におけるメリットとデメリットのトレードオフが表れている．つまり，2011 年から 2014 年までは $l = 5.0$ 年の固定スライディングウィンドウの方が予測誤差が小さく，相対的に安定した予測ができているが，2015 年から 2018 年までは $l = 0.5$ 年の固定スライディングウィンドウの方が予測誤差は小さい．これは，より短いウィンドウの固定スライディングウィンドウはモデル推定の際に使用されるデータのうち市況変化前の古いデータの割合が少なく，新しく得られるデータが中心となるため，市況の変化への対応が早いことによる．

この結果からそれぞれの予測時点によって市況が変化しており，予測誤差を小さくするウィンドウの長さは異なると言える．また市況の変化は地域によっても異なると想定されるため，我々は固定スライディングウィンドウにおける単一の最適なウィンドウの長さを求めるアプローチではなく，観察されたデータに応じてモデル推定のためのウィンドウの長さを適応的に変更する方法が望ましいと判断した．そこで本研究では新たに観察されたデータをその時点の中古マンション価格モデルによって説明することが困難になるとき，市況に変化が生じていると捉え，モデル推定に用いるウィンドウの長さを調整するスキームを考案する．

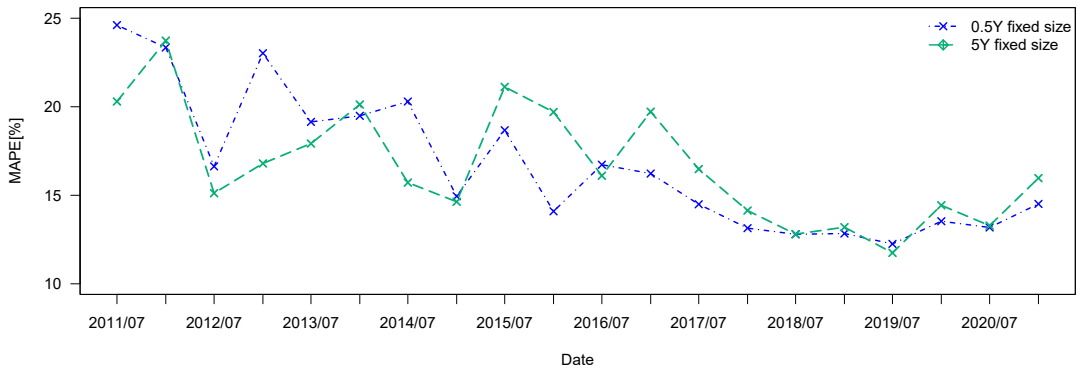

図 4.2: 東京都江東区の中古マンションデータで推定した価格モデルによる翌半年のアウトオブサンプルを予測した予測誤差率の絶対値平均（モデル推定に使用したインサンプルのウィンドウの長さは薄い点線が 0.5 年，太い長点線が 5.0 年），横軸は予測対象期間の終点

4.4 提案手法

4.4.1 可変ウィンドウの考え方

我々の提案する可変ウィンドウは図 4.3 のように観察されるデータに適応してウィンドウの長さを調整し，モデルを更新するスキームである．中古マンションの市況は短期間に変化する場合でも，たとえばある時点で一斉に全ての物件の取引価格が高騰ないしは暴落するということは現実的でなく，初めはいくつかの物件の取引価格に変化があり，他の物件の取引データに混ざって観察され始め，時間の経過とともに他の物件にもその変化が反映される，という動きが不動産実務において知られている．

こうした中古マンション市況の変化の特性を考慮し，新しく観察されるデータから市況の変化を検出する変化点判定ロジックと，どの程度まで直近のデータを重視すればモデルが市況の変化に追従可能かを判定するウィンドウの伸縮ロジックをスキームとして構成する．なお，中古マンションの市況の変化は価格のみの動きで捉えるのではなく，物件属性の違いを反映した中古マンション価格モデルを通じて判定する必要がある．次の項では本研究で用いるその価格モデルについて説明する．

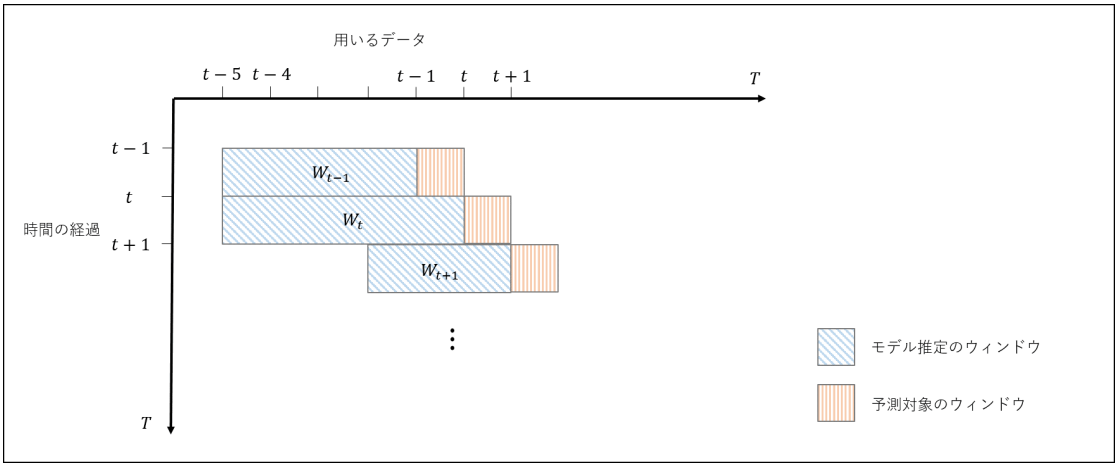

図 4.3: 可変ウィンドウによるモデルの更新

4.4.2 価格予測モデル

本研究では中古マンション価格の予測において初めて可変ウィンドウを採用する有効性の検証が主眼であるため，予測に用いる中古マンションの価格モデルは Shimizu and Nishimura [91] や早川・田島 [22] など，多くの不動産関連の研究で用いられる線形回帰モデルとする．

まず，中古マンションの物件価格を y，延床面積や築年数など d 個の物件属性を $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\top$ とする．このとき，

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \tag{4.1}$$

という線形回帰モデルが成り立つとする．なお $\beta_1, \ldots, \beta_d$ はそれぞれの x の回帰係数，β_0 は定数項，ϵ は誤差項を表しており，ϵ は平均 0, 分散 σ^2 の正規分布に従うものとする．

ここで n 組の中古マンションの成約データ $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ が観察できたとする．あとの項での説明を分かりやすくするため，観察されたデータおよび線形回帰モデルのパラメータを次のように表す:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & \boldsymbol{x}_1^\top \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_n^\top \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

したがって先の線形回帰モデル (4.1) は次のように書き直せる．

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}). \tag{4.2}$$

なお，$\boldsymbol{I}$ は単位行列である．

4.4.3 可変ウィンドウによるモデル更新のフロー

我々の提案手法は 2 つのロジックがあり，時間の経過に伴い，新たに観察されたデータから市況の変化を検知する変化点の判定ロジック，そして実際にどの程度直近のデータを重視すればモデルが市況の変化を捉えられるかを判定して過去のデータを切り離し，ウィンドウの長さを決定するロジックからなる．本研究では中古マンション価格の予測に線形回帰モデルを用いるため，市況の変化は線形回帰モデルの回帰診断に使用される Cook の距離（Cook [46]）を応用し[2]，新しく取得したデータが線形回帰モデルのパラメータ推定に与える影響を算出することによって判定する．また，ウィンドウの長さを決定するロジックにおいてはウィンドウの中でモデルが異なると判断される時点を F 検定を用いて決定する．以下にそのフローを示す．なお，4.3 節からの記号に加え，時点 t における可変ウィンドウのウィンドウ W_t の長さを l_t とし，ウィンドウ内のデータは $n(= n_{n_{t-l_t+1}} + \cdots + n_t)$ 件とする．

[2]正確には本研究で使用する Cook の距離は次の項で解説する Cook [46] を変形した距離を指すが，本稿の説明においては，同じように Cook の距離と表現しても大きな問題が生じないため，以降の説明では区別せずに Cook の距離と表現する．

Step1 新しく観察された時点 $t+1$ の n_{t+1} 組の取引データ $\{(\boldsymbol{x}_j, y_i)\}_{j=1}^{n_{t+1}}$ について，時点 t までのウィンドウ W_t に存在する n 組の取引データ $\{(\boldsymbol{x}_j, y_j)\}_{j=1}^{n}$ による線形回帰モデルを基準として，Cook の距離を算出する (図 4.4).

Step2 算出した Cook の距離で一定の閾値 v を超えるデータの数が，新しく取得した n_{t+1} 件に占める一定の割合 s と同じか下回る場合，既存のウィンドウ W_t の n 件に新規データ n_{t+1} 件を加えたウィンドウ W_{t+1} によって線形回帰モデルを更新し，Step1 に戻る（図 4.5 のように市況に大きな変化は無いと判断し，ウィンドウの長さは $l_{t+1} = l_t + \Delta T$ に伸長する）．一方，Cook の距離で一定の閾値 v を超えるデータが，新しく取得した n_{t+1} 件に占める一定の割合 s を超える場合は市況に大きな変化があると判断し，ウィンドウの長さを調整するために Step3 に進む.

Step3 モデルの推定に際して直近のデータを重視し，より古いデータの影響を取り除くため，既存のデータ n 件に新規データ n_{t+1} 件を加えたウィンドウの中で切り捨て時点の判定を行う．この際，図 4.6 に示すようにウィンドウ内に含まれる各時点で 2 つに区切ることを考え．2 つに区切った場合のモデルと区切らない場合のモデルの当てはめ残差によって F 統計量を算出し，それぞれの時点の F 統計量のうち，最も値が大きくなる時点より古いデータをウィンドウ内の切り捨て候補とする．その候補とした時点の F 統計量によって F 検定を行い，切り捨てるか否かを判定する．F 検定によって帰無仮説を棄却した際は，切り捨て候補となる時点より古いデータを切り捨てて Step4 へ進む．一方，F 検定によって帰無仮説を棄却できなければ，ウィンドウ内のデータを切り捨てせず，図 4.5 と同様の状態で上記 Step1 に戻る.

Step4 図 4.7 のように古いデータを切り捨てたウィンドウ W_{t+1} によって推定し直した線形回帰モデルによって時点 $t+1$ から時点 $t+2$ までの価格予測を行う.

以降の 4.4.4 節では上記 Step1 と Step2，4.4.5 節では Step3 について解説する.

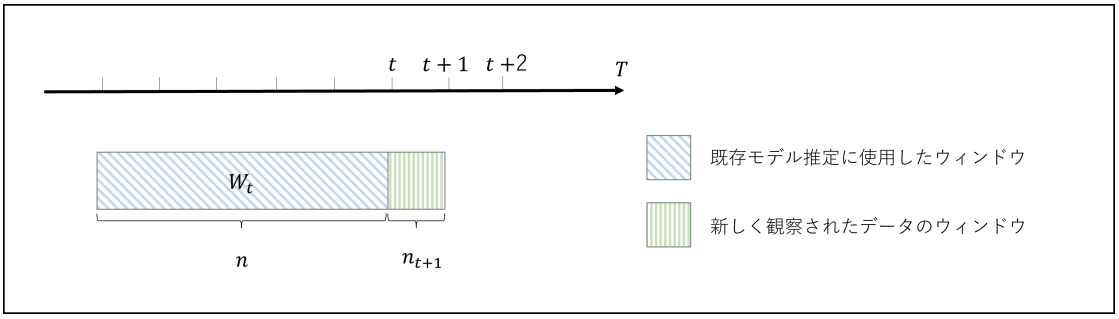

図 4.4: Step1．Cook の距離を算出

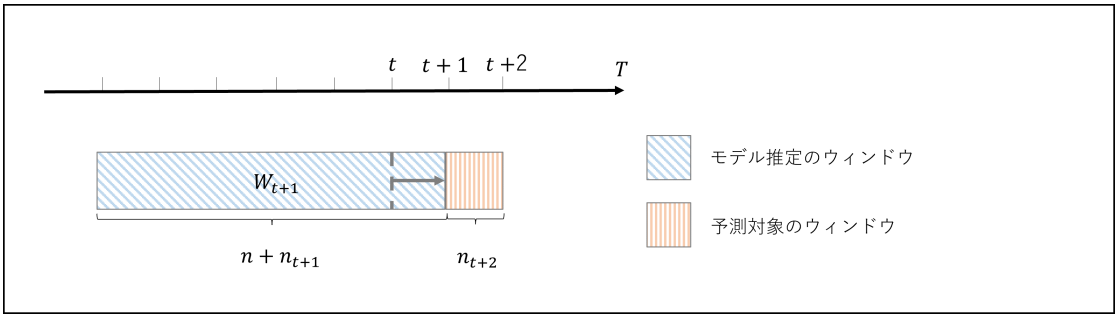

図 4.5: Step2. 市況に大きな変化が無いという判定の場合

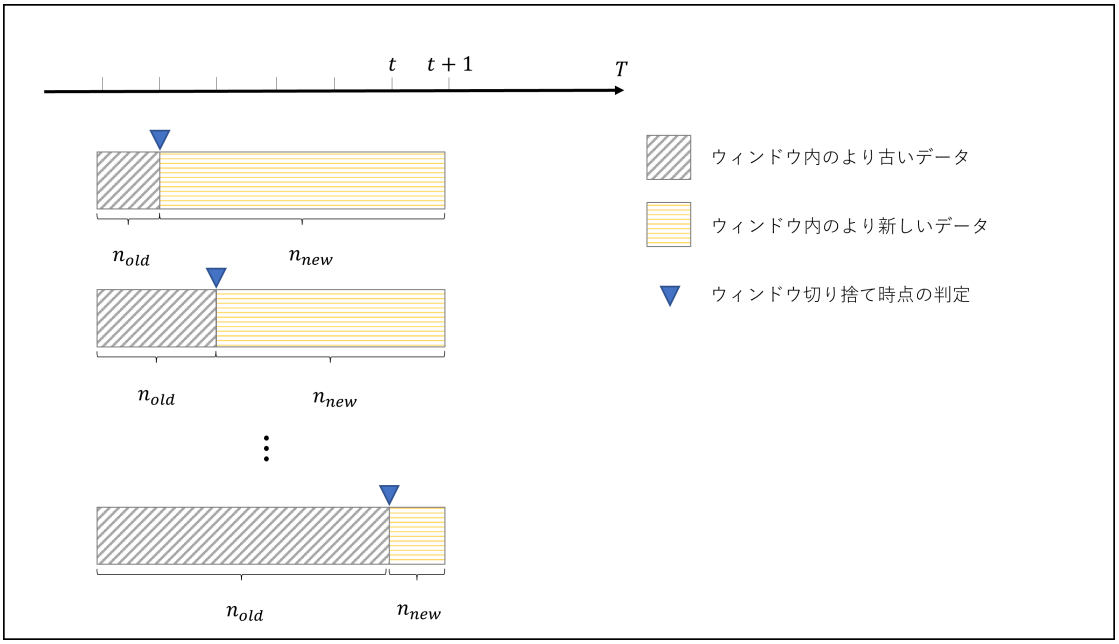

図 4.6: Step3. ウィンドウの切り捨て時点の判定

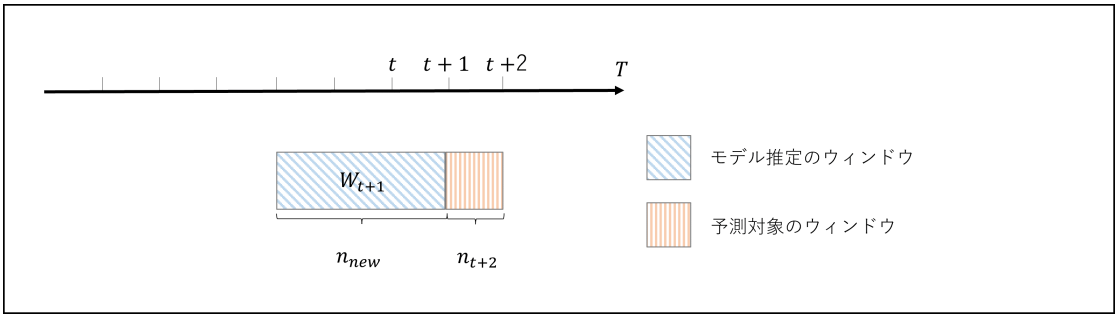

図 4.7: Step4. モデルを更新して次の時点の予測へ使用

4.4.4 変化点判定ロジック

本研究では「市況の変化」を「逐次的にデータをあてはめ続けた線形回帰モデルの陳腐化」と捉えることで，ウィンドウの縮小の要否を判定する．この線形回帰モデルの陳腐化とは，これまでのウィンドウの長さであてはめてきた線形回帰モデルでは新たに観察されたデータを説明することが困難な状態であるということを示しており，新たに観察されたデータは既存のモデルにとっての外れ値に相当するという見方ができる．そこで我々はいわゆるハット行列ないしその要素を用いた梃子比のアイディアによる外れ値診断が援用できると考えた．ハット行列を用いた外れ値診断の指標としては，DFFITS，DFBETAS，Delta-Beta 統計量，Cook の距離などがあるが (Belsley et al. [33]，Dobson [52], Neter et al. [76])，どの指標も定義やアイディアは類似していることから，本研究では Cook の距離をベースに新たな指標を構築することを試みた．

Cook. [46] は n 組のデータ $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ から式 (4.2) の線形回帰モデルを推定した回帰係数パラメータ $\hat{\boldsymbol{\beta}}$ から，あるサンプル i が与える影響 D_i を，サンプル i を除いて推計した回帰係数パラメータ $\hat{\boldsymbol{\beta}}_{(-i)}$ を用いて次の式による距離で評価している．

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(-i)} - \hat{\boldsymbol{\beta}})^\top \boldsymbol{X}^\top \boldsymbol{X} (\hat{\boldsymbol{\beta}}_{(-i)} - \hat{\boldsymbol{\beta}})}{(d+1)s^2}. \tag{4.3}$$

なお，s^2 は残差分散で $(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top (\boldsymbol{y} - \hat{\boldsymbol{y}})/(n-d-1)$ である．

この考え方を応用し，我々は時点 t までに得られた n 組のデータ $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ によって線形回帰モデルの回帰係数パラメータ $\hat{\boldsymbol{\beta}}_n$ を推定し，その後の時間の経過に伴って新たに得られた $n+1$ 組目の $(\boldsymbol{x}_{n+1}, y_{n+1})$ を既存の線形回帰モデルに取り込んだ際に回帰係数パラメータ $\hat{\boldsymbol{\beta}}_n$ に与える影響を距離として算出するため，式（4.3）を 1 つのサンプルを除くのではなく，1 つのサンプルを追加するよう次の形に書き換える．

$$\begin{aligned} D_{n+1} &= \frac{(\hat{\boldsymbol{\beta}}_{n+1} - \hat{\boldsymbol{\beta}}_n)^\top \boldsymbol{X}_n^\top \boldsymbol{X}_n (\hat{\boldsymbol{\beta}}_{n+1} - \hat{\boldsymbol{\beta}}_n)}{(d+1)s^2} \\ &= \left(\frac{y_{n+1} - \boldsymbol{x}_{n+1}^\top \hat{\boldsymbol{\beta}}_n}{s\{1 + \boldsymbol{x}_{n+1}^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{x}_{n+1}\}} \right)^2 \frac{\boldsymbol{x}_{n+1}^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{x}_{n+1}}{d+1}. \end{aligned} \tag{4.4}$$

ここで $\hat{\boldsymbol{\beta}}_n$ は $\boldsymbol{\beta}$ の推定値であり，$n+1$ 組目の $(\boldsymbol{x}_{n+1}, y_{n+1})$ をモデル推定に加えた場合の $\hat{\boldsymbol{\beta}}$ の変動は $\hat{\boldsymbol{\beta}}_{n+1} - \hat{\boldsymbol{\beta}}_n$ で表す．また，$\boldsymbol{X}_n^\top \boldsymbol{X}_n$ は正則行列とする．また $\hat{\boldsymbol{\beta}}_{n+1}$ を各データ点について算出することを避けるため，(4.4) の 2 段目の式を用いる．この式展開については本論文末の B.1．Cook の変形，へ記載する．

式 (4.4) の適用においては図 4.4 のように，時点 t までのウィンドウ W_t に内包された n_t 組のデータから推定したモデルに対し，時点 t から時点 $t+1$ まで観察された n_{t+1} 件の各データについて，式（4.4）を個別に適用してそれぞれ距離 D_{n+1} を

計算し，一定の閾値 v を超えるデータが新規データ n_{t+1} 件のうち一定の割合 s を超えるか否かによってモデルの陳腐化を判定する．

4.4.5 ウィンドウの伸縮ロジック

前項の Cook の距離による変化点判定において線形回帰モデルの陳腐化が無いと判定されれば，図 4.5 のように既存のモデル推定に使用したウィンドウのデータに新しく観察されたデータを加え，伸長したウィンドウによってモデルを推定し直す．一方，前項の Cook の距離による変化点判定によって線形回帰モデルが陳腐化していると判定された場合は，ウィンドウ内の古いデータを切り捨ててウィンドウの長さを縮小する．

この際，実際の市況変化は一部の中古マンション取引に反映されはじめ，徐々に他の中古マンション取引に広がっていくと考える方が自然であることから，新しく観察されたデータのみ用いてモデルを更新するではなく，図 4.6 のように既存のモデル推定に使用したウィンドウ内のどの時点より以前のデータを切り捨てるべきか判定する仕組みを取り入れる．我々はこの仕組みを Chow [44]，Bai and Perron [30], Julious [70] らのようにある期間内においてモデルの構造変化が生じている時点を特定する問題と捉え直すことで実現した．線形回帰モデルの構造変化を検出する研究は回帰係数に線形制約が存在するか否かを F 検定を用いて判定する内容であり，これまで数多くの研究が行われている．本研究では過度に複雑化せず，可変ウィンドウの中のいずれか 1 つの時点で構造変化が生じている可能性があるという前提に立ち，候補となる時点を F 統計量の大きさによって特定することとする．

具体的には時点 t までのウィンドウ W_t 内の n 件のデータに時点 $t+1$ までに新しく得られた n_{t+1} 件のデータを加え，n' 件とし，図 4.6 に示すようにウィンドウ内の各時点にて 2 つに区切り，古い期間のサンプル n_{old} 件と新しい期間のサンプル n_{new} 件に分ける．この期間の違いをダミー変数で区別する線形回帰モデル（式 (4.5)）を推定した場合と区別しない線形回帰モデル（式 (4.2) のサンプルサイズは n' と読み替える）を推定した場合のそれぞれの当てはめ残差によって F 統計量を計算する．ここでウィンドウ内の各時点について計算した F 統計量のうち，最も大きな値を取る時点をモデルの構造変化が生じた候補とし，有意水準 α によって F 検定を行う．

その結果，区切った時点の前後でモデルに違いは無い，という帰無仮説を棄却する場合はウィンドウ内の古い期間のデータ n_{old} を切り捨て，図 4.7 に示すようにウィンドウ W_{t+1} の長さを縮小してモデルの更新を行う．一方，帰無仮説を棄却できなかったときは図 4.5 のように既存のモデル推定に用いたウィンドウ W_t に新しいデータ n_{t+1} を加え，ウィンドウの長さを伸長して，モデルの更新を行う．

ウィンドウの長さの変更ロジックを示すため，観察データとモデルのパラメータを次のように表す:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_{n'} \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & \boldsymbol{x}_1^\top \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_{n'}^\top \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}, \boldsymbol{z} = \left[\begin{array}{c|c} \boldsymbol{0} & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{I}_{n_{new}} \end{array}\right], \boldsymbol{\gamma} = \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_d \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_{n'} \end{bmatrix}.$$

ここで $\boldsymbol{z}$ は要素が 0 の n' 次正方行列のうち，右下のブロックのみ n_{new} 次の単位行列である．ウィンドウのある時点の前後でモデルが異なると考える場合は式 (4.2) に基づけば

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z}\boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \quad N(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \tag{4.5}$$

と表せる．この式 (4.5) によるモデルの残差は次のとおり．

$$\mathrm{RSS}_k = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{z}\boldsymbol{X}\boldsymbol{\gamma}\|^2. \tag{4.6}$$

なお，k を式 (4.6) で推定する $\boldsymbol{\beta}$ と $\boldsymbol{\gamma}$ のパラメータ数とすれば，$k = 2(d+1)$ である．一方，ウィンドウ内でモデルの構造変化が生じていないと考える場合はモデルは式 (4.2) であるため，その当てはめ残差は次のとおり．

$$\mathrm{RSS}_d = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2. \tag{4.7}$$

これら式 (4.7) と式 (4.6) の残差ならびにモデルのパラメータ数から次の F 統計量を算出する．

$$F = \frac{(\mathrm{RSS}_d - \mathrm{RSS}_k)/(k-d-1)}{\mathrm{RSS}_k/(n'-k)}. \tag{4.8}$$

この F は自由度 $k-d-1, n'-k$ の F 分布に従う．

$$F \sim F_{k-d-1,n'-k}. \tag{4.9}$$

F 統計量はウィンドウ内に含まれる各時点で 2 つに分けた場合のそれぞれで計算し，最も大きな F 統計量について次の F 検定を行う．

$$H_0 : \ \boldsymbol{z} = \boldsymbol{0} \ \text{ vs. } \ H_1 : \ \boldsymbol{z} \neq \boldsymbol{0}. \tag{4.10}$$

ここでウィンドウ内に w 個の時点を含む場合，異なる時点の切り捨て判定を $w-1$ 回行うため，Bifet and Gavalda [34] でも取り上げられているように判定に用いる有意水準を $\alpha/(w-1)$ として設定し，帰無仮説の棄却可否によってウィンドウの中の古い期間のデータを削除可否を決定する．

なお，可変ウィンドウであっても，全く削除されずウィンドウの長さが伸び続けると新しく観察されるデータの重みが小さくなりすぎることから，ウィンドウの長さに上限値を設ける．これをウィンドウに含まれる時点の数の上限 W_{max} として設定する．またモデルの予測の安定性を考慮して下限 W_{min} を設定する．

4.5 分析の方法

4.5.1 データセット

本研究で用いる東京都の中古マンション取引のデータはSREホールディングス株式会社[3]が収集した2006年1月から2020年12月までの成約価格のデータセットである．我々は事前に東京都の中でも市況の変化による中古マンション価格の値動きが大きい地域と値動きが小さい地域が存在していることをデータセットを分析して確認している．図4.8と図4.9には江東区と大田区のそれぞれの地域において成約した中古マンションの平均m^2単価と件数を示す．江東区は大田区と比べ，2015年頃から価格が大きく上昇していることが見て取れる．そこで本研究の提案手法の効果を明らかにするため，値動きが大きい地域として江東区，値動きが小さい地域として大田区の2つの地域で評価を実施した．

なお，収集した中古マンションの成約データに対して，物件の所在情報や建築年月などに生じる欠損値の補完処理を施しているが，それでも欠損値が多い属性情報はモデルの変量として相応しくないため除外している．また成約価格に関しても大槻・横内 [2] で提案する方法によって，明らかに1桁大きい値が収録されているレコードは削除している．

表 4.1: データの基本統計量

江東区 （レコード数　993件）

変量名	単位	平均	最小	最大	標準偏差
専有面積	m^2	64.86	16.08	123.13	16.87
最寄駅までの徒歩	分	7.49	1.00	20.00	3.97
間取り部屋数		2.42	1.00	4.00	0.81
築年月	月数	223.70	1.00	605.00	144.89
成約価格	円	37,643,765	4,000,000	125,000,000	17,043,237
成約m^2単価	円	570,946	147,694	1,249,500	191,232

大田区 （レコード数　2,076件）

変量名	単位	平均	最小	最大	標準偏差
専有面積	m^2	57.63	7.99	217.5	19.24
最寄駅までの徒歩	分	7.32	1.00	22.00	3.95
間取り部屋数		2.30	1.00	5.00	0.82
築年月	月数	262.50	2.00	611.00	146.95
成約価格	円	30,717,753	3,100,000	128,000,000	13,696,535
成約m^2単価	円	528,566	143,421	1,043,580	141,025

[3]https://sre-group.co.jp/

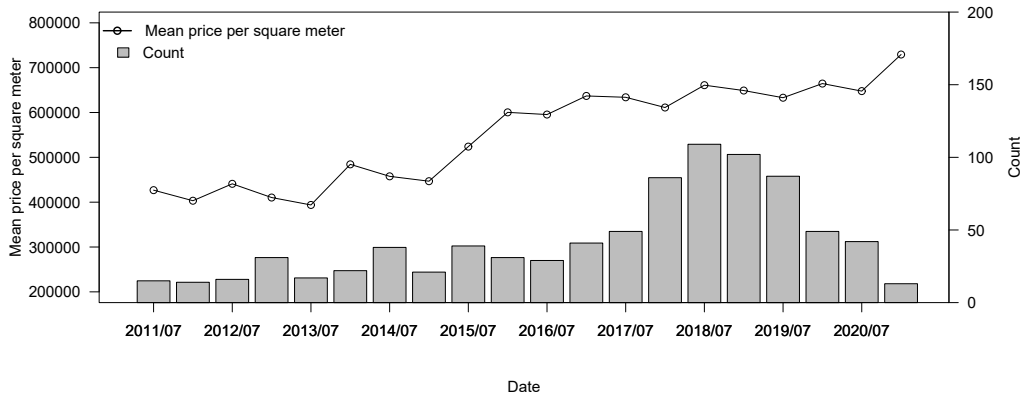

図 4.8: 江東区の平均 m^2 単価 (円単位，実線，左軸) と成約件数 (棒，右軸)

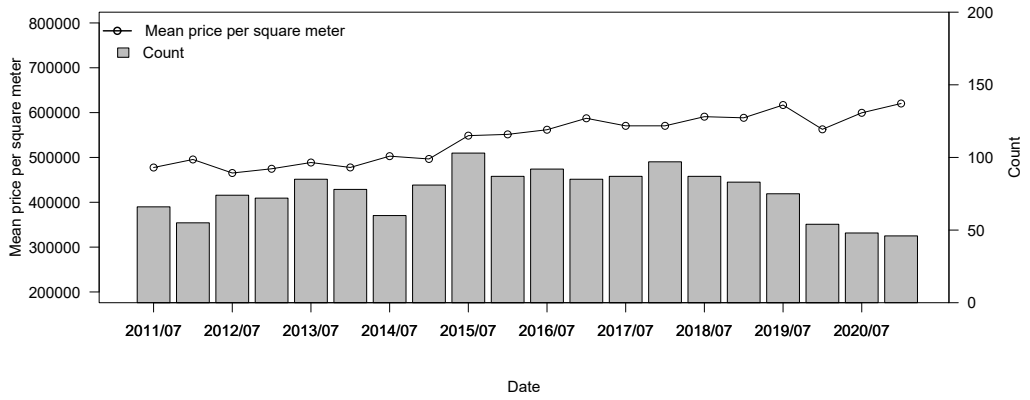

図 4.9: 大田区の江東区の平均 m^2 単価 (円単位，実線，左軸) と成約件数 (棒，右軸)

4.5.2 数値条件と精度指標

- 中古マンション価格モデルの変量

  線形回帰モデルの予測誤差によって，我々の提案する可変ウィンドウと既存手法である固定スライディングウィンドウの手法を比較する．式（4.1）にて説明した線形回帰モデルについて，分析に用いる変量は，

  y : 中古マンション m^2 当たり価格

  x_1 : 延床面積（m^2）

  x_2 : 最寄駅までの距離 (分)

  x_3 : 築年月（月数）

  x_4 : 間取り部屋数

x_5 : 建物総階数

とする．

- 可変ウィンドウの設定

表 4.2: 設定するパラメータ

パラメータ種類		値
Cook の距離に対する閾値	υ	90%
閾値超えデータの割合	s	10%
F 検定の有意水準	α	1%
ウィンドウの長さの上限	W_{max}	5.0 年
ウィンドウの長さの下限	W_{min}	0.5 年

実際のデータを用いた検証に際し，4.4 項で説明したパラメータの具体的な値は表 4.2 のとおり．なお，パラメータ (υ, s, α) の設定に際し，複数の閾値を比較した結果は本論文末の B.3. パラメータの設定，へ記載する．また時点 $T = 0, 1, 2, \ldots, t, \ldots$ の間隔 ΔT については，可変ウィンドウおよび固定スライディングウィンドウとも今回用いるデータのサンプルサイズを考慮し，0.5 年とする[4]．

- 予測精度の指標

予測精度を評価するため，式 (4.11)-(4.13) に定義する MAPE，予測誤差の 90%点，MER(誤差率中央値) を指標として用いた．

MAPE は次のとおり定義する．

$$\text{MAPE} := \frac{1}{n_{t+1}} \sum_{i=1}^{n_{t+1}} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100. \tag{4.11}$$

また，値 $z_1, z_2, \ldots, z_n$ を小さい順に並べた場合の $100q\%$ 点を $Q_{(q)}(\{z\}_{j=1}^{n})$ と表すことにする．予測誤差の 90%点を次のとおり定義し，評価に用いる．

$$90\%\text{点} := Q_{0.9}\left(\left\{ \left| \frac{y_i - \hat{y}_i}{y_i} \right| \right\}_{i=1}^{n_{t+1}} \right) \times 100. \tag{4.12}$$

MER は次のとおり定義する．

$$\text{MER} := Q_{0.5}\left(\left\{ \left| \frac{y_i - \hat{y}_i}{y_i} \right| \right\}_{i=1}^{n_{t+1}} \right) \times 100. \tag{4.13}$$

[4]各予測時点で観察可能なサンプルサイズによって 0.5 年より短い頻度とすることも可能である．

各指標は小さいほど予測精度が高いことと対応する．予測精度を評価する上で基本的にはMAPEを参照するが，MAPEは個別性の強い一部の物件の影響を受ける問題があるため，90%点とMERを併用する．90%点は大きく外れる場合に対応して，予測精度の安定性を測る指標となる．またMERは米国Zillow社[5]などでよく用いられる指標であり，実務では予測精度の良さを代表する指標となる．

4.6 分析と考察

まず江東区のデータに対して予測誤差を比較した結果を示す．図4.10に示す各予測時点のMAPEを見れば，2011年7月から2015年1月までは2.0年や5.0年の固定スライディングウィンドウと比べ，0.5年の固定スライディングウィンドウのMAPEは大きくなりやすく，かつ不安定な推移を示している．ただし，2015年7月はこの違いが逆転し，0.5年のウィンドウのMAPEが2.0年や5.0年のウィンドウと比べ，より低いMAPEを示すように変化する．特に5.0年の固定スライディングウィンドウは2016年頃から2018年頃まで他の長さのウィンドウと比べて相対的にMAPEが大きくでており，前節の図4.8の価格推移を考慮すれば，価格の大きな変動に追従できていないと考えられる．

一方，我々の提案手法である可変ウィンドウは2015年頃まで2.0年から5.0年の固定スライディングウィンドウとほぼ同程度のMAPEであるが，図4.11からも分かるように2015年7月時点でウィンドウの長さが短くなっており，0.5年の固定スライディングウィンドウのMAPEと同程度である．5.0年の固定スライディングウィンドウのMAPEの結果と比較すれば，可変ウィンドウは価格の大きな変動に追従できていると考えられる．

[5]https://www.zillow.com/z/zestimate/

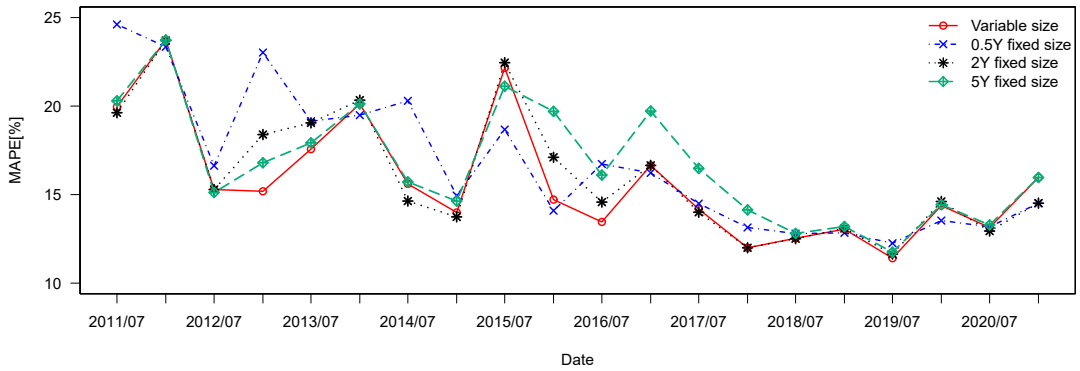

図 4.10: 江東区の MAPE（ウィンドウの長さは実線が提案手法，点線が 0.5 年，一点鎖線が 2.0 年，長点線が 5.0 年）

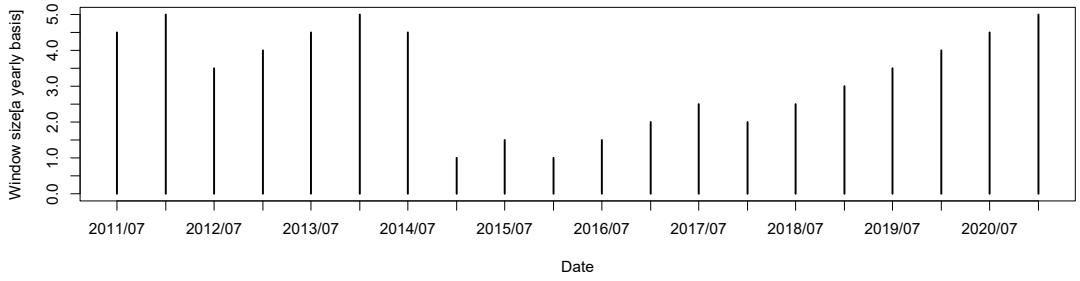

図 4.11: 江東区の可変ウィンドウの長さ (年単位)

表 4.3 によれば MER に関しては 2 番目に小さい結果となっており，90 %点に関して最も小さい結果となっている．

表 4.3: 江東区の MER(下端) と 90%点 (上端)

	MER	90%点
可変ウィンドウ	11.71%	28.60%
0.5 年固定ウィンドウ	11.23%	30.59%
2 年固定ウィンドウ	11.80%	29.70%
5 年固定ウィンドウ	12.24%	30.56%

次に大田区のデータに対して予測誤差を比較した結果を示す．図 4.12 を見ると，我々の提案する可変ウィンドウは全期間を通して MAPE が小さい状態を保って推移している．なお，全般的に大田区は江東区と比べてウィンドウの長さの違いによる MAPE のばらつきが小さい．ただし，2012 年 7 月，2017 年 7 月，2020 年 7

月は他のウィンドウと比べて 0.5 年の固定スライディングウィンドウの MAPE が大きくなっており，予測精度が不安定であることが分かる．また 2015 年 7 月から 2017 年 1 月まで 5 年の固定スライディングウィンドウの MAPE が相対的に悪化しており，江東区ほどではないものの図 4.9 の価格推移を考慮しても 2016 年頃からの価格上昇への追従が遅れていると考えられる．

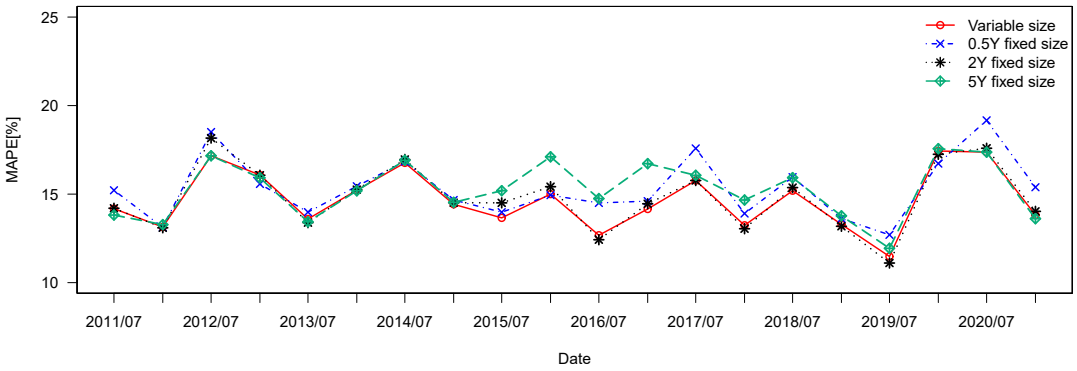

図 4.12: 大田区の MAPE（ウィンドウの長さは実線が提案手法，点線が 0.5 年，一点鎖線が 2 年，長点線が 5 年）

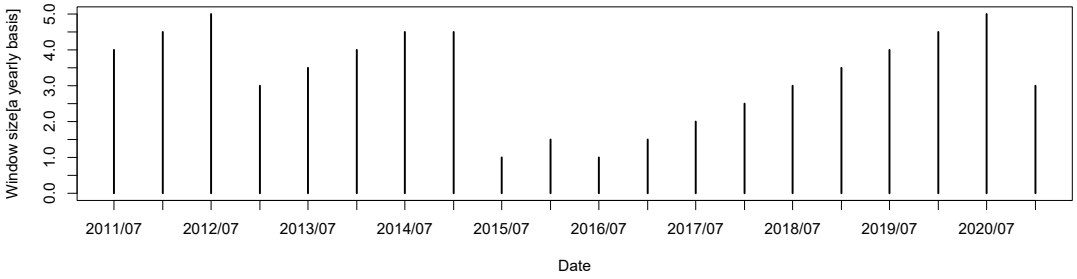

図 4.13: 大田区の可変ウィンドウの長さ (年単位)

表 4.4 のとおり，可変ウィンドウが両方とも最も低い結果が得られた．

表 4.4: 大田区の MER と 90%点

	MER	90%点
可変ウィンドウ	11.35%	28.94%
0.5 年固定ウィンドウ	11.58%	31.12%
2 年固定ウィンドウ	11.43%	29.09%
5 年固定ウィンドウ	12.30%	30.20%

ここまで得られた結果から，我々の提案する可変ウィンドウが江東区および大田区の両方，さらにいずれの期間においても固定スライディングウィンドウより小さい MAPE を達成している．また，より価格変動が大きい江東区の方が比較的変動が小さい大田区より，可変ウィンドウを採用することによる予測誤差の抑制効果が大きい結果となっており，価格の変動が急で大きい地域であるほど，本研究で提案する可変ウィンドウ方式を採用したモデル更新が有効であると考えられる．

4.7 結論

本研究では市況が変化する中で中古マンション価格モデルによる予測の精度を維持するため，固定スライディングウィンドウにおけるウィンドウの長さの設定問題を解決する方法として可変ウィンドウを用いたモデルの更新アルゴリズムを考案した．

また実際に我々の提案する可変ウィンドウと既存の固定スライディングウィンドウによって推定した線形回帰モデルを用いて，2011 年 1 月から 2020 年 12 月までの東京都江東区と大田区の中古マンションデータの予測を行い，可変ウィンドウを用いたモデルが期間全般にわたって予測を大きく外すことはなく，市況の変化に対する追従性を確保しつつ，予測の安定性を保てることを確認した．特に価格の変動が大きい江東区と比較的小さい大田区での結果の違いから，価格変動の大きい地域の方が可変ウィンドウを採用することによるメリットが大きいことが分かった．

なお，本研究ではこれまで取り上げられなかったモデル推定のためのウィンドウに焦点を当てたため，中古マンション価格モデルはシンプルな線形回帰モデルを採用したが，より予測精度の高いモデルへの改良余地（変量の追加，関数形の変更など）については今後の課題とする．

第5章　結論と今後の課題

結論

本研究では不動産仲介会社などのWEBサイトで提供されている不動産査定サービスを念頭に，統計モデルを用いた不動産価格予測における実務上の課題に対して解決策を提案している．具体的には不動産データベースに必要なデータクレンジング，不動産取引頻度の地域差による不動産価格予測モデルの劣化問題，市況変化による不動産価格予測モデルの陳腐化をテーマとして採り上げた．こうした課題は2000年代後半以降に不動産ビッグデータの蓄積が進んだことや高速なインターネット回線の普及により情報伝搬が迅速化したことなどを背景として，不動産査定サービスにおける価格予測に対し，より高い質が求められるようになって顕在化したものと考えられる．既存の不動産関連研究においてこうした実務上の課題が扱われている例は少なく，筆者の知る限り日本の不動産において採り上げているものは無い．

本研究の貢献は次の3つの課題をデータサイエンスプロセスに則した不動産価格予測フローの中で不動産データのクレンジングや意思決定という位置づけで捉え直し，それぞれ解決策を提案している点である．

- 第２章 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
  不動産のデータは欠損や値の誤り，収録ルールからの逸脱などが散見され，様々なデータ浄化の手続きが必要となるが，これまでの不動産関連研究においてデータクレンジングをテーマとして採り上げたものは少なく，日本の不動産のデータを用いた研究は筆者の知る限り見当たらない．そこで本研究では国土交通省の中古マンションデータベースを対象とし，価格の桁間違いエラーを検出する方法として地域性の違いに着目した２つのクレンジング方法を提案した．また不動産価格予測モデルに大きな影響を与える取引価格の桁間違いエラーの取り除くことによってモデルの当てはまりが大幅に改善することを確認した．

- 第３章 東京２３区の中古マンション市場のデータ分割と統合
  中古マンション価格の予測精度を改善するため，単純に地域を細分化して価格モデルを推定するとモデルの更新頻度が高い地域と低い地域が生じる．そこで地域を細分化して，中古マンション価格モデルの予測誤差が改善するこ

とを確認したのち，予測精度をできるだけ下げずに地域を再統合し，取引の発生頻度を保つ方法を提案した．価格形成が類似した隣接地域との取引発生頻度の違いを補完するためのデータ再統合によってモデルの更新頻度の地域差を解消し，予測精度の劣化が防げることを確認した．

- 第４章 可変ウィンドウによる中古マンション価格モデルの更新
  金融政策の変更，制度の変更，都市の再開発など様々な要因によって不動産市況は変化するため，どんなに予測精度の良い不動産価格モデルでもいずれは精度が悪化する．こうした市況の変化による中古マンション価格モデルの陳腐化を防ぐため，実務ではモデルを推定するデータを一定の長さにして入れ替える固定スライディングウィンドウが用いられる．本研究では固定スライディングウィンドウのウィンドウの長さに起因する問題に対処するため，観察されるデータに適応してウィンドウの長さを変更する可変ウィンドウを提案した．可変ウィンドウを採用したモデルは固定スライディングウィンドウを採用した場合と比べ，観察期間において予測を大きく外さないという結果が得られた．

今後の課題

本研究で提案した手法やその結果からさらに発展的な研究に広がる余地がある．

不動産データクレンジングの標準化とアーカイブ

より予測精度の高い中古マンション価格モデル，さらには土地や戸建てなど他の種類の不動産価格予測モデルを構築するためには収録されるデータへの理解を深めた上でデータクレンジングを行う必要がある．たとえば中古マンションの延床面積についてもデータベースによっては壁芯から計測している物件と内法で計測している物件が混在しているケースがある．築年数が古い中古マンションにも関わらず，高い価格で取引されている場合，価格の桁間違いエラーではなく建て替えによる容積率緩和を見込んだ取引という可能性もある．しかしながら，これまで分析者がアクセスできる不動産データベースが限られていたという背景もあり，不動産データに対する標準的なデータクレンジング指針や手法は確立されていない．

近年では国交省データベースも蓄積が進み，また大手不動産仲介会社から国立情報学研究所へ研究用のデータセットが提供されるなど，データ利用の環境が変わりつつある[1]．そこで不動産データにおいて頻出するデータクレンジングのポイントを類型化すれば，不動産データクレンジングの標準化に貢献できると考える．

[1] https://www.nii.ac.jp/dsc/idr/datalist.html

ここで不動産データクレンジングの標準化のために第2章で取り扱ったエラーレコードの特定をCodd [45] のリレーショナルデータの考えに従って捉え直す．まずデータ型や定義域などをドメインの属性情報として与えれば，データベクトルの個々の値をチェックすることができるため，たとえば価格にテキストデータが入力されている，といったシンタックス（syntax）の意味でのクレンジングが可能となる．それを一歩進めて各ドメイン間の関係性を活用すれば，本研究のように地域性を考慮して取引価格のエラーを判定したり，間取りに対する延床面積の値の妥当性を判定したり，と一つのドメインでは検出できないセマンティクス（semantics）を反映したデータクレンジングに拡張できる．

- シンタックスでのエラー検出
  たとえば価格，延床面積などは正数をとる（> 0）など，単独のドメインの定義域で検出可能なエラー．データの背後にある専門的な知識を必要としないことが多く，データベース構築の際に設定できることも多い．

- セマンティクスでのエラー検出
  明らかに判明するエラーとは言えず，単独のドメインの定義域では検出できない．

  – あるドメインの定義域が別のドメインの値によって変化する場合
    エラー検出の対象とするドメインに条件付き定義域を設定するともいえる．本研究で提案した手法は取引価格の桁間違いを地域という条件を使って検出している．条件付き定義域をどう発見するかという点でデータの背後にある専門的な知識が必要なこともあるが，検出する仕組みを作ればオートマチックなデータクレンジングが可能になる．

    一方，どのドメインの条件を使えば良いか専門家の知見でも自明でない場合は統計モデルや機械学習手法などを用いた手法が考えられる．本研究で提案した手法の一つは機械学習の手法を用いている．

  – あるドメインのエラー検出のための条件がデータに含まれていない場合
    そのままでは対応できないが，別のデータベースと統合することで検出できる可能性はある．

次に施したデータクレンジングの記録を保持するアーカイブ手法についても研究の余地がある．不動産の取引が発生し，新たなデータが収録されるたびにデータクレンジングが必要となるが，データを更新するたびに外れ値と判定したレコードを削除し続けている場合，それまでの外れ値を正しい値として採用するような大きな変化（たとえば新駅開業や再開発の影響など）を捉えられなくなるため，外れ値として削除した時点に遡って見直す必要が生じる可能性もある．

こうした変化に対処するためにはデータベースに対して施したクレンジングログを保持し，過去のデータベースを再現できる方法が必要になる．横内・柴田 [28],

Yokouchi and Shibata [100] はデータとその記述の一体化を企図して DandD（Data and Description）プロジェクトを進め，データを組織化する方法として DandD ルールを提唱している．DandD ルールを参考とし，データクレンジングの内容，つまり欠損値や異常値の背景とその判断理由，処理した内容をメタデータとしてデータセットに付与することで処理のブラックボックス化を避け，透明性と再現性を確保できると考える．また処置した記録を保持することでデータベースの更新時におけるメンテナンスも容易になる．さらにはデータ型などを合わせて属性情報に記録すれば計算機などに読み込んだ際に不動産価格モデルにおける変量の変換などが容易になるという利点もある．

モデルを通じた地域の統合とデータクレンジング

第 2 章では中古マンション取引データにおいて，実際の取引価格とは異なる取引価格の桁間違いというエラーの検出方法を提案したが，実際に取引された価格のレコードであっても不動産価格予測モデルの推定に悪影響を及ぼす外れ値が含まれている場合があり，こうしたレコードを検出する手法を考案することでモデルの予測精度を改善できる可能性がある．この目的のためにはモデルを通じて外れ値を検出することが考えられるが，地域の違いによってもモデルの当てはまりは異なるため，ある地域の中古マンション取引データに対してモデルの当てはまりが悪かった場合に特定の物件による外れ値であるのか，あるいは価格形成要因が異なる小地域が混在しているのかという点は判別が容易ではなく，レコードを１件ずつ見極める負担も大きい．

そこで第 4 章で提案した可変ウィンドウのアイディアを応用し，モデルの予測精度を基準に均質性の高い地域を特定しつつ，外れ値の検出を行うことで予測精度の改善を図る．具体的には線形回帰モデルにおける Cook の距離（式 4.4）とモデル推定ウィンドウ内の変化点検出ロジック (4.4.5) を応用し，時系列方向へウィンドウを伸縮するのではなく空間的な方向へウィンドウを拡縮させる（たとえば東京２３区であれば地下鉄の路線において最寄り駅が同じレコードの束に対し，隣の最寄り駅のレコードの束を新たに得られたデータと捉える）ことでそれぞれの地域における外れ値検出と類似した地域間のデータ統合を同時に行える可能性がある．図 5.1 は第 3 章でも採り上げた東京メトロ東西線の城東地区について最寄り駅単位で統合可否を判定する例である．モデルを推定する際に可変ウィンドウの考え方を使い，葛西を最寄り駅とするレコードの群からモデルを推定し，そのモデルを用いて西葛西を最寄り駅とする各レコードについて Cook の距離（式 (4.4)）を適用して外れ値か否か，さらに西葛西駅を葛西駅と統合できるかをモデル推定ウィンドウ内の変化点検出ロジック (4.4.5 項) で判定する．そこで葛西駅と西葛西駅が統合できた場合は均質的な一つの地域とみなし，同様に南砂町駅を最寄り駅とする各レコードの外れ値検出と統合判定を行う．その後も同様に東陽町駅を最寄り駅とする各レコードの外れ値検出と統合判定に続ける．

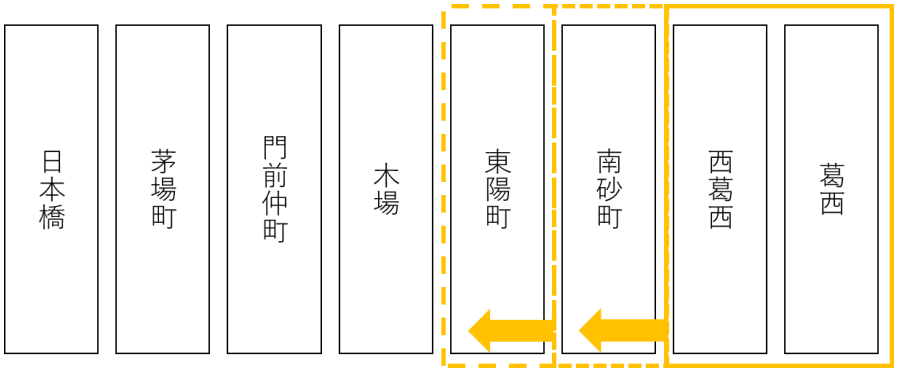

図 5.1: 東京メトロ東西線の城東地区について最寄り駅単位で統合可否を判定する例

この方法で形成された均質性の高い地域に対してデータブラウジングを行う中で共通の価格形成要因を見出せれば，よりモデルの予測精度改善に役立つファクターを発見できる可能性がある．本稿の第 3 章では地域性の違いによる価値を最寄り駅ダミーという質的変量を用いてモデルに導入しているが，再統合された地域に共通した要素から，地形や標高による快適さや交通ネットワークへのアクセスなど，地域ダミー変数による違いの根拠を発見できる余地がある．

モデルの改良

本研究では既存の研究でこれまで取り上げられることの少なかった中古マンションデータのデータクレンジングや意思決定をテーマとし，課題に対する解決策を提案することに重点を置いたため，シンプルな線形回帰モデルを用いている．柴田 [12] で解説されているようにデータブラウジングによってデータの中に緩やかな変化や線形的な変化が見つかれば線形回帰モデルが適用できる．

第 3 章ではデータの細分化と再統合によってデータの中の均質性を探索しているが，中古マンション，さらには戸建てなどの不動産データに拡張した場合は地域や条件によって線形回帰モデルの当てはまりがよくないケースが存在する．

そこで変量の追加や関数形の変更形など，モデルの改良によって予測精度を改善できる余地がある．変量の追加については地形に関する変量を確認した研究として早川・田島 [22] が東京都 23 区の中古マンション価格モデルに所在する町の平均標高を変量として追加し，標高が高いほど取引価格が高いと結論付けている．Ye et al. [99] は香港のマンションを対象としてモデルにマンションと最寄りの地下鉄駅との間の平均勾配の効果を加えて価格の違いを説明している．

また，関数の変更については清水 [90] が中古マンション価格モデルの説明変量の非線形性を検証するため，線形回帰モデルと一般化加法モデルなどの非線形モデルをアウトオブサンプルの予測誤差で比較し，市況の大きな変化が無ければ一

般化加法モデルなどによって説明変数の非線形性を考慮することが有効であると主張している．Garcia et al. [57] はスペイン・アリカンテの住宅，Xu et al. [102] はアメリカ・シカゴ郊外の住宅，福中ら [24] は日本の大都市圏の中古マンションデータに対し，線形回帰モデル，ランダムフォレスト，XGboost，LightGBM などを用いて当てはまりやアウトオブデータへの予測精度を比較した結果，XGboost や LightGBM が他の方法より良好なパフォーマンスを示したことを報告している．本研究の第 3 章や第 4 章のアイディアに対してこうした非線形性を表現できるモデルに改良することで線形回帰モデルの当てはまりが良くない地域においても，より良い予測精度を達成できる可能性がある．

ただし，いきなり複雑な関数形のモデルを用いると予測精度の改善は達成できたとしても，なぜそのモデルを用いた場合の予測精度が良くなるのか，あるいは予測精度が劣化した際にどのように改良すれば良いのかという点がブラックボックス化しやすく，実務での運用には課題が多い．そこでまずは線形回帰モデルの当てはまりが良くないデータについてその原因を探るアプローチを採用し，非線形性が生じている要因を明らかにする．具体的にはスプライン関数を用いた加法モデルをデータのブラウジングのために用い，モデルを推定した際の偏残差プロットを観察することによって，中古マンションの取引価格と説明変量である物件属性に非線形性が存在するか否か，また存在する場合はどのような地域や条件で発生するかを確認する．このようにモデルを通じて得られる知見に基づき，データへの理解を深めながらデータ分割や変量の追加を行うことでモデルの予測精度を改善する．

付 録A　第2章の補足

A.1　人間の判断に基づく桁間違いレコード

第 2 章で用いた人間の判断に基づく桁間違いレコード 168 件は次のとおり.

表 A.1: 人間の判断に基づく桁間違いレコード 168 件

No	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	市ケ谷	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	日本橋蛎殻町	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 3 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 3 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 22 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
12350	港区	麻布十番	麻布十番	130,000,000	2	6	20	１ＤＫ	6,500,000	平成 15 年	ＳＲＣ	平成 21 年第 1 四半期	未改装
12955	港区	海岸	日の出	220,000,000	1	7	30	１ＬＤＫ	7,333,333	平成 10 年	ＲＣ	平成 17 年第 4 四半期	未改装
13250	港区	芝公園	芝公園	240,000,000	1	26	50	２ＬＤＫ	4,800,000	昭和 54 年	ＳＲＣ	平成 17 年第 4 四半期	
13263	港区	白金	白金台	1,100,000,000	10	9	80	２ＬＤＫ	13,750,000	平成 10 年	ＳＲＣ	平成 19 年第 3 四半期	未改装
13426	港区	高輪	泉岳寺	360,000,000	5	10	40	１ＤＫ	9,000,000	平成 10 年	ＳＲＣ	平成 20 年第 1 四半期	改装済
13771	港区	南麻布	広尾	1,100,000,000	3	34	115	２ＬＤＫ	9,565,217	昭和 47 年	ＳＲＣ	平成 18 年第 3 四半期	
14313	新宿区	上落合	落合 (東京)	95,000,000	2	27	15	１Ｋ	6,333,333	平成元年	ＲＣ	平成 28 年第 1 四半期	未改装
14516	新宿区	北新宿	大久保 (東京)	140,000,000	7	43	25	１Ｋ	5,600,000	昭和 47 年	ＳＲＣ	平成 27 年第 1 四半期	改装済
15134	新宿区	荒木町	曙橋	1,300,000,000	1	8	140	３ＬＤＫ	9,285,714	平成 14 年	ＳＲＣ	平成 22 年第 2 四半期	未改装
16099	新宿区	中落合	落合南長崎	170,000,000	2	11	20	１Ｋ	8,500,000	平成 18 年	ＲＣ	平成 29 年第 1 四半期	未改装
16386	新宿区	下落合	高田馬場	150,000,000	5	41	45	１ＤＫ	3,333,333	昭和 44 年	ＳＲＣ	平成 22 年第 1 四半期	改装済
18910	新宿区	北新宿	大久保 (東京)	95,000,000	9	20	20	１Ｋ	4,750,000	平成元年	ＳＲＣ	平成 21 年第 1 四半期	未改装
19354	新宿区	歌舞伎町	東新宿	80,000,000	5	24	25	２ＬＤＫ	3,200,000	昭和 58 年	ＳＲＣ	平成 19 年第 4 四半期	未改装
19787	新宿区	富久町	曙橋	98,000,000	4	36	20	１Ｋ	4,900,000	昭和 47 年	ＳＲＣ	平成 20 年第 3 四半期	改装済
19941	新宿区	西新宿	新宿西口	89,000,000	1	26	15	１Ｋ	5,933,333	昭和 54 年	ＳＲＣ	平成 17 年第 4 四半期	未改装
19964	新宿区	西新宿	都庁前	84,000,000	6	25	20	１Ｋ	4,200,000	昭和 55 年	ＲＣ	平成 17 年第 4 四半期	改装済
19973	新宿区	西新宿	西新宿	200,000,000	4	7	20	１Ｋ	10,000,000	平成 13 年	ＲＣ	平成 20 年第 2 四半期	未改装
20170	新宿区	弁天町	牛込柳町	480,000,000	4	37	80	２ＬＤＫ	6,000,000	昭和 46 年	ＳＲＣ	平成 20 年第 2 四半期	改装済
21100	文京区	小日向	江戸川橋	530,000,000	7	17	80	３ＬＤＫ	6,625,000	平成 9 年	ＲＣ	平成 26 年第 3 四半期	未改装
21263	文京区	関口	江戸川橋	600,000,000	4	15	85	３ＬＤＫ	7,058,824	平成 13 年	ＲＣ	平成 28 年第 4 四半期	未改装
22502	文京区	本駒込	駒込	900,000,000	8	17	100	４ＬＤＫ	9,000,000	平成 8 年	ＳＲＣ	平成 25 年第 3 四半期	未改装
23113	文京区	本郷	春日 (東京)	200,000,000	2	8	20	１Ｋ	10,000,000	平成 14 年	ＲＣ	平成 22 年第 3 四半期	未改装
23161	文京区	本駒込	駒込	220,000,000	8	5	30		7,333,333	平成 17 年	ＲＣ	平成 22 年第 1 四半期	
23183	文京区	向丘	東大前	500,000,000	4	10	60	３ＬＤＫ	8,333,333	平成 12 年	ＳＲＣ	平成 22 年第 4 四半期	改装済
23337	文京区	小石川	茗荷谷	610,000,000	4	6	65	２ＬＤＫ	9,384,615	平成 17 年	ＳＲＣ	平成 23 年第 1 四半期	未改装
23601	文京区	音羽	江戸川橋	130,000,000	3	11	20	１Ｋ	6,500,000	平成 10 年	ＳＲＣ	平成 21 年第 3 四半期	改装済
23872	文京区	大塚	新大塚	73,000,000	5	12	15	１Ｋ	4,866,667	平成 7 年	ＲＣ	平成 19 年第 3 四半期	改装済
24082	文京区	関口	江戸川橋	400,000,000	7	3	55	２ＬＤＫ	7,272,727	平成 16 年	ＲＣ	平成 19 年第 4 四半期	未改装

No	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
24580	台東区	浅草橋	浅草橋	250,000,000	3	26	50	３ＬＤＫ	5,000,000	平成 3 年	ＳＲＣ	平成 29 年第 2 四半期	未改装
26332	台東区	駒形	浅草	250,000,000	2	6	45	１ＬＤＫ	5,555,556	平成 16 年	ＲＣ	平成 22 年第 4 四半期	未改装
28995	墨田区	本所	本所吾妻橋	60,000,000	9	25	20	１Ｋ	3,000,000	平成元年	ＲＣ	平成 26 年第 4 四半期	未改装
29423	墨田区	立花	東あずま	220,000,000	3	17	50	２ＬＤＫ	4,400,000	平成 5 年	ＲＣ	平成 22 年第 3 四半期	未改装
29704	墨田区	本所	本所吾妻橋	53,000,000	9	21	20	１Ｋ	2,650,000	平成元年	ＲＣ	平成 22 年第 2 四半期	未改装
30268	墨田区	菊川	菊川 (東京)	150,000,000	3	23	50	２ＬＤＫ	3,000,000	昭和 59 年	ＳＲＣ	平成 19 年第 1 四半期	未改装
31486	江東区	亀戸	亀戸	63,000,000	13	41	25	１ＤＫ	2,520,000	昭和 49 年	ＳＲＣ	平成 27 年第 3 四半期	
33088	江東区	亀戸	亀戸	59,000,000	16	36	25	１Ｋ	2,360,000	昭和 49 年	ＳＲＣ	平成 22 年第 4 四半期	未改装
34831	江東区	住吉	住吉 (東京)	360,000,000	6	6	65	２ＬＤＫ	5,538,462	平成 16 年	ＲＣ	平成 22 年第 2 四半期	改装済
36404	江東区	豊洲	豊洲	300,000,000	11	9	60	１ＤＫ	5,000,000	平成 12 年	ＲＣ	平成 21 年第 4 四半期	未改装
37283	江東区	東陽	東陽町	210,000,000	4	33	60	３ＬＤＫ	3,500,000	昭和 50 年	ＳＲＣ	平成 20 年第 2 四半期	未改装
37553	江東区	南砂	住吉 (東京)	200,000,000	14	27	55	２ＬＤＫ	3,636,364	昭和 56 年	ＳＲＣ	平成 20 年第 2 四半期	未改装
39820	品川区	小山台	不動前	150,000,000	6	22	20	１Ｋ	7,500,000	昭和 63 年	ＲＣ	平成 22 年第 2 四半期	未改装
40145	品川区	中延	中延	180,000,000	6	10	20	１Ｋ	9,000,000	平成 12 年	ＲＣ	平成 22 年第 2 四半期	未改装
40546	品川区	西五反田	五反田	320,000,000	1	31	50	２ＬＤＫ	6,400,000	昭和 54 年	ＳＲＣ	平成 22 年第 2 四半期	改装済
42270	品川区	南大井	大森 (東京)	150,000,000	9	4	20	１Ｋ	7,500,000	平成 17 年	ＲＣ	平成 21 年第 1 四半期	
42339	品川区	荏原	西小山	330,000,000	7	19	45	１ＤＫ	7,333,333	平成元年	ＲＣ	平成 20 年第 1 四半期	改装済
42432	品川区	上大崎	高輪台	77,000,000	9	23	15	１Ｋ	5,133,333	昭和 59 年	ＲＣ	平成 19 年第 4 四半期	未改装
43095	品川区	南品川	大井町	2,600,000,000	8	23	270		9,629,630	昭和 58 年	ＳＲＣ	平成 18 年第 3 四半期	
43448	目黒区	上目黒	中目黒	1,400,000,000	2	8	70	２ＬＤＫ	20,000,000	平成 21 年	ＲＣ	平成 29 年第 1 四半期	未改装
43931	目黒区	鷹番	学芸大学	120,000,000	5	29	15	１Ｋ	8,000,000	昭和 62 年	ＲＣ	平成 28 年第 2 四半期	未改装
44302	目黒区	東山	池尻大橋	220,000,000	4	10	20	１Ｋ	11,000,000	平成 14 年	ＲＣ	平成 24 年第 2 四半期	未改装
44332	目黒区	碑文谷	学芸大学	390,000,000	18	45	60	３ＬＤＫ	6,500,000	昭和 47 年	ＲＣ	平成 29 年第 1 四半期	未改装
44714	目黒区	五本木	祐天寺	110,000,000	7	8	35	４ＬＤＫ	3,142,857	平成 14 年	ＲＣ	平成 22 年第 1 四半期	改装済
45064	目黒区	緑が丘	自由が丘 (東京)	490,000,000	9	22	80	３ＬＤＫ	6,125,000	昭和 63 年	ＲＣ	平成 22 年第 1 四半期	未改装
45415	目黒区	下目黒	目黒	640,000,000	11	11	75	３ＬＤＫ	8,533,333	平成 10 年	ＳＲＣ	平成 21 年第 4 四半期	未改装
45693	目黒区	上目黒	祐天寺	92,000,000	8	21	15	１Ｋ	6,133,333	昭和 61 年	ＲＣ	平成 19 年第 3 四半期	未改装
47083	大田区	大森東	平和島	51,000,000	3	17	20	１Ｋ	2,550,000	平成 8 年	鉄骨造	平成 25 年第 2 四半期	未改装
47538	大田区	池上	池上	78,000,000	9	20	15	１Ｋ	5,200,000	平成 2 年	ＲＣ	平成 22 年第 2 四半期	未改装
49315	大田区	中馬込	馬込	92,000,000	2	12	20	１Ｋ	4,600,000	平成 17 年	ＲＣ	平成 29 年第 1 四半期	未改装
52302	大田区	南馬込	大森 (東京)	140,000,000	13	17	20	１Ｋ	7,000,000	平成 4 年	ＲＣ	平成 21 年第 2 四半期	未改装
52333	大田区	南六郷	雑色	290,000,000	11	19	75	３ＬＤＫ	3,866,667	平成 2 年	ＳＲＣ	平成 21 年第 1 四半期	改装済
52444	大田区	大森北	大森 (東京)	93,000,000	8	16	15	１Ｋ	6,200,000	平成 3 年	ＲＣ	平成 19 年第 4 四半期	未改装
52897	大田区	下丸子	鵜の木	270,000,000	8	13	135	４ＬＤＫ	2,000,000	平成 5 年	ＳＲＣ	平成 18 年第 1 四半期	改装済
53199	大田区	西蒲田	池上	64,000,000	2	16	15	１Ｋ	4,266,667	平成 3 年	ＳＲＣ	平成 19 年第 4 四半期	改装済
53348	大田区	東馬込	馬込	180,000,000	1	22	40	２ＤＫ	4,500,000	昭和 59 年	ＲＣ	平成 18 年第 1 四半期	改装済
53483	大田区	南馬込	西馬込	350,000,000	8	13	50	２ＬＤＫ	7,000,000	平成 6 年	ＲＣ	平成 19 年第 3 四半期	未改装
54924	世田谷区	砧	成城学園前	88,000,000	15	24	15		5,866,667	平成 3 年	ＲＣ	平成 27 年第 2 四半期	未改装
56614	世田谷区	三軒茶屋	若林 (東京)	62,000,000	10	25	15	１Ｋ	4,133,333	昭和 60 年	ＲＣ	平成 22 年第 4 四半期	改装済
57981	世田谷区	松原	明大前	85,000,000	2	30	30	１ＤＫ	2,833,333	昭和 57 年	ＳＲＣ	平成 24 年第 2 四半期	未改装
58742	世田谷区	成城	成城学園前	78,000,000	3	27	15	１Ｋ	5,200,000	昭和 59 年	ＲＣ	平成 23 年第 1 四半期	改装済
59018	世田谷区	東玉川	田園調布	380,000,000	6	15	50	２ＬＤＫ	7,600,000	平成 8 年	ＲＣ	平成 23 年第 1 四半期	未改装
60375	世田谷区	千歳台	祖師ケ谷大蔵	360,000,000	15	5	70	２ＬＤＫ	5,142,857	平成 15 年	ＲＣ	平成 20 年第 3 四半期	未改装
60451	世田谷区	等々力	尾山台	59,000,000	3	21	15	１Ｋ	3,933,333	昭和 59 年	ＲＣ	平成 17 年第 3 四半期	未改装
60639	世田谷区	松原	明大前	110,000,000	5	20	15	１Ｋ	7,333,333	昭和 63 年	ＳＲＣ	平成 20 年第 1 四半期	
60769	世田谷区	用賀	桜新町	74,000,000	10	19	15	１Ｋ	4,933,333	平成元年	ＲＣ	平成 20 年第 3 四半期	未改装
61548	渋谷区	渋谷	渋谷	1,100,000,000	11	11	105	３ＬＤＫ	10,476,190	平成 13 年	ＲＣ	平成 24 年第 4 四半期	未改装
61871	渋谷区	道玄坂	渋谷	74,000,000	7	37	20	１ＤＫ	3,700,000	昭和 54 年	ＳＲＣ	平成 28 年第 1 四半期	未改装
62092	渋谷区	西原	代々木上原	80,000,000	5	28	15	１Ｋ	5,333,333	昭和 60 年	ＲＣ	平成 25 年第 3 四半期	未改装
62991	渋谷区	笹塚	笹塚	90,000,000	6	34	35	１ＤＫ	2,571,429	昭和 51 年	ＲＣ	平成 22 年第 4 四半期	未改装
63821	渋谷区	幡ケ谷	幡ケ谷	160,000,000	8	22	30	２ＤＫ	5,333,333	平成元年	ＲＣ	平成 23 年第 4 四半期	未改装
63984	渋谷区	恵比寿	恵比寿	900,000,000	6	10	125	４ＬＤＫ	7,200,000	平成 11 年	ＳＲＣ	平成 21 年第 2 四半期	未改装
64273	渋谷区	本町	初台	140,000,000	6	29	35	１ＤＫ	4,000,000	昭和 55 年	ＲＣ	平成 21 年第 1 四半期	未改装
64573	渋谷区	笹塚	笹塚	75,000,000	5	21	15	１Ｒ	5,000,000	昭和 59 年	ＲＣ	平成 17 年第 4 四半期	未改装
65062	渋谷区	本町	西新宿五丁目	79,000,000	6	20	15	１Ｋ	5,266,667	昭和 60 年	ＳＲＣ	平成 17 年第 3 四半期	未改装
65379	中野区	新井	中野 (東京)	52,000,000	12	27	10	１Ｋ	5,200,000	昭和 60 年	ＲＣ	平成 24 年第 3 四半期	未改装
66629	中野区	本町	西新宿五丁目	330,000,000	6	44	50	２ＬＤＫ	6,600,000	昭和 47 年	ＳＲＣ	平成 28 年第 3 四半期	改装済
67442	中野区	若宮	鷺ノ宮	480,000,000	4	13	80	３ＬＤＫ	6,000,000	平成 10 年	ＲＣ	平成 23 年第 1 四半期	未改装
67747	中野区	中央	新中野	130,000,000	4	7	20	１Ｋ	6,500,000	平成 12 年	ＲＣ	平成 19 年第 3 四半期	未改装
67793	中野区	中央	中野坂上	69,000,000	4	22	15	１Ｋ	4,600,000	昭和 59 年	ＲＣ	平成 18 年第 4 四半期	未改装
68109	杉並区	阿佐谷南	阿佐ケ谷	74,000,000	3	33	15	１Ｋ	4,933,333	昭和 59 年	ＲＣ	平成 29 年第 1 四半期	改装済
69034	杉並区	久我山	久我山	80,000,000	2	41	25	１ＤＫ	3,200,000	昭和 49 年	ＳＲＣ	平成 27 年第 3 四半期	
69436	杉並区	井草	井荻	470,000,000	8	12	85	４ＬＤＫ	5,529,412	平成 10 年	ＲＣ	平成 22 年第 1 四半期	改装済
70351	杉並区	浜田山	西永福	1,200,000,000	10	26	160	３ＬＤＫ	7,500,000	平成元年	ＲＣ	平成 27 年第 3 四半期	改装済
70638	杉並区	堀ノ内	方南町	89,000,000	5	20	15	１Ｋ	5,933,333	平成 6 年	ＲＣ	平成 26 年第 3 四半期	未改装
71831	杉並区	成田東	南阿佐ケ谷	310,000,000	6	24	50	２ＬＤＫ	6,200,000	昭和 60 年	ＲＣ	平成 21 年第 2 四半期	未改装
71941	杉並区	和田	東高円寺	76,000,000	3	17	15	１Ｋ	5,066,667	平成 4 年	ＲＣ	平成 21 年第 3 四半期	未改装
72642	杉並区	堀ノ内	方南町	210,000,000	10	20	45	２ＤＫ	4,666,667	昭和 62 年	ＲＣ	平成 19 年第 4 四半期	未改装
72998	豊島区	池袋本町	北池袋	78,000,000	11	31	15		5,200,000	昭和 59 年	ＲＣ	平成 27 年第 1 四半期	
73341	豊島区	北大塚	大塚 (東京)	240,000,000	4	10	35	１ＬＤＫ	6,857,143	平成 15 年	ＲＣ	平成 25 年第 2 四半期	未改装
74674	豊島区	要町	千川	100,000,000	4	14	20	１Ｋ	5,000,000	平成 8 年	ＲＣ	平成 22 年第 2 四半期	改装済
75189	豊島区	高田	早稲田 (都電)	85,000,000	4	33	30	１ＤＫ	2,833,333	昭和 52 年	ＲＣ	平成 22 年第 3 四半期	未改装
75193	豊島区	千早	要町	100,000,000	8	25	15	１Ｒ	6,666,667	昭和 60 年	ＲＣ	平成 22 年第 2 四半期	未改装
75695	豊島区	池袋本町	北池袋	100,000,000	5	11	15	１Ｋ	6,666,667	平成 10 年	ＲＣ	平成 21 年第 2 四半期	改装済

No	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
75739	豊島区	北大塚	大塚 (東京)	300,000,000	8	14	50	２ＬＤＫ	6,000,000	平成７年	ＳＲＣ	平成 21 年第２四半期	改装済
75754	豊島区	巣鴨	巣鴨	280,000,000	12	12	50	２ＬＤＫ	5,600,000	平成９年	ＳＲＣ	平成 21 年第３四半期	未改装
75780	豊島区	高田	高田馬場	86,000,000	6	24	15	１Ｋ	5,733,333	昭和 60 年	ＲＣ	平成 21 年第４四半期	未改装
75828	豊島区	西巣鴨	西巣鴨	57,000,000	4	20	20	１Ｋ	2,850,000	平成元年	ＲＣ	平成 21 年第２四半期	未改装
75858	豊島区	東池袋	新大塚	130,000,000	2	36	30	３ＤＫ	4,333,333	昭和 48 年	ＳＲＣ	平成 21 年第１四半期	未改装
76109	豊島区	駒込	駒込	560,000,000	5	2	55	２ＬＤＫ	10,181,818	平成 17 年	ＲＣ	平成 19 年第４四半期	未改装
76213	豊島区	高田	目白	120,000,000	12	12	20	１Ｋ	6,000,000	平成８年	ＲＣ	平成 20 年第３四半期	未改装
76276	豊島区	西池袋	池袋	230,000,000	5	9	30	１ＤＫ	7,666,667	平成 11 年	ＳＲＣ	平成 20 年第２四半期	改装済
77935	北区	王子	王子	78,000,000	9	18	15	１Ｒ	5,200,000	平成４年	ＳＲＣ	平成 22 年第２四半期	改装済
78450	北区	西ケ原	上中里	94,000,000	9	23	20	１Ｋ	4,700,000	昭和 61 年	ＲＣ	平成 21 年第４四半期	改装済
78741	北区	田端	田端	190,000,000	7	6	30	１ＬＤＫ	6,333,333	平成 14 年	ＲＣ	平成 20 年第２四半期	未改装
80001	荒川区	東日暮里	三ノ輪	68,000,000	5	24	25	１ＤＫ	2,720,000	昭和 61 年	ＲＣ	平成 22 年第３四半期	改装済
80448	荒川区	西尾久	尾久	64,000,000	3	21	20	１Ｋ	3,200,000	昭和 62 年	ＲＣ	平成 20 年第４四半期	改装済
80626	荒川区	南千住	南千住	170,000,000	9	9	60	３ＬＤＫ	2,833,333	平成 11 年	ＲＣ	平成 20 年第４四半期	未改装
84972	板橋区	南町	要町	180,000,000	7	15	35	１ＤＫ	5,142,857	平成６年	ＳＲＣ	平成 21 年第２四半期	改装済
85123	板橋区	板橋	板橋区役所前	210,000,000	7	9	60	３ＬＤＫ	3,500,000	平成 11 年	ＲＣ	平成 20 年第４四半期	未改装
85395	板橋区	高島平	新高島平	48,000,000	12	17	15	１Ｋ	3,200,000	平成２年	ＲＣ	平成 19 年第１四半期	未改装
85699	板橋区	氷川町	板橋区役所前	120,000,000	2	2	20	１Ｋ	6,000,000	平成 17 年	ＳＲＣ	平成 19 年第３四半期	未改装
86965	練馬区	大泉町	大泉学園	250,000,000	28	12	65	３ＬＤＫ	3,846,154	平成 10 年	ＲＣ	平成 22 年第４四半期	未改装
88249	練馬区	早宮	豊島園	42,000,000	14	25	15	１Ｋ	2,800,000	昭和 63 年	ＲＣ	平成 25 年第４四半期	未改装
89079	練馬区	桜台	新桜台	55,000,000	2	22	15	１Ｋ	3,666,667	平成元年	ＲＣ	平成 23 年第４四半期	未改装
89206	練馬区	関町北	武蔵関	58,000,000	9	36	35	１ＤＫ	1,657,143	昭和 50 年	ＲＣ	平成 23 年第３四半期	未改装
89641	練馬区	関町北	武蔵関	39,000,000	9	34	15	１Ｋ	2,600,000	昭和 50 年	ＲＣ	平成 21 年第３四半期	未改装
90123	練馬区	関町北	武蔵関	77,000,000	3	22	25	１Ｋ	3,080,000	昭和 61 年	ＲＣ	平成 20 年第３四半期	未改装
90140	練馬区	関町北	武蔵関	95,000,000	5	14	20	１Ｋ	4,750,000	平成５年	ＲＣ	平成 19 年第２四半期	未改装
93069	足立区	弘道	五反野	40,000,000	6	23	15	１Ｋ	2,666,667	昭和 62 年	ＲＣ	平成 22 年第２四半期	改装済
94011	足立区	足立	五反野	61,000,000	7	21	15	１Ｒ	4,066,667	昭和 63 年	ＲＣ	平成 21 年第１四半期	未改装
94208	足立区	竹の塚	竹ノ塚	97,000,000	15	18	65	３ＬＤＫ	1,492,308	平成３年	ＲＣ	平成 21 年第１四半期	未改装
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２四半期	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２四半期	
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３四半期	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

A.2　既存研究と提案手法の比較例

既存研究である野呂・和田[8)]と本提案手法の判定を比較する．比較の条件と例として採り上げた最寄り駅別レコードの価格分布については次のとおり．

表 A.2: 比較の条件

既存手法 A	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 B	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 C	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 1 四分位 −1.724× 四分位範囲 上限値 = 第 3 四分位 +1.724× 四分位範囲
既存手法 C ’	常用対数変換後に手法 C 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 四分位 $+10 \times \mathrm{SIQR_u}$
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 ∪ 提案手法 2	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

表 A.3: 最寄り駅別のレコードの価格分布

	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

それぞれの手法を適用した判定結果の例は次のとおり.

表 A.4: 判定は１が異常値，０が正常値

人間の判断	既存手法A	既存手法B	既存手法C	既存手法C’	提案手法1	提案手法2	手法1∪手法2	手法1∩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	間取り	平米当たり単価	建物の構造	改装
0	0	0	0	1	0	0	0	0	7340	港区	麻布台	六本木一丁目	3,100,000	3	15	33	１Ｋ	206,667	ＳＲＣ	未改装
0	0	1	0	0	0	0	0	0	11290	港区	六本木	六本木一丁目	100,000,000	3	45	5	１ＤＫ	2,222,222	ＲＣ	改装済
0	1	1	1	0	0	0	0	0	11292	港区	六本木	六本木一丁目	490,000,000	4	190	5	３ＬＤＫ	2,578,947	ＳＲＣ	未改装
0	1	1	1	0	0	0	0	0	11308	港区	六本木	六本木一丁目	520,000,000	4	170	4	３ＬＤＫ	3,058,824	ＳＲＣ	未改装
0	0	0	0	1	0	0	0	0	68605	杉並区	永福	西永福	15,000,000	3	45	40	２ＬＤＫ	333,333	ＳＲＣ	未改装
0	0	0	0	1	0	0	0	0	69760	杉並区	永福	西永福	10,000,000	4	30	36	１ＤＫ	333,333	ＳＲＣ	
1	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	３ＬＤＫ	7,500,000	ＲＣ	改装済
0	0	1	1	0	0	0	0	0	17360	新宿区	百人町	新大久保	32,000,000	3	25	3	１Ｋ	1,280,000	ＲＣ	改装済
0	0	1	0	0	0	0	0	0	17362	新宿区	百人町	新大久保	18,000,000	4	15	20	１Ｋ	1,200,000	ＲＣ	
0	0	1	1	0	0	0	0	0	17363	新宿区	百人町	新大久保	26,000,000	3	20	8	１Ｋ	1,300,000	ＲＣ	改装済
0	0	1	0	0	0	0	0	0	17367	新宿区	百人町	新大久保	24,000,000	3	20	9	１Ｋ	1,200,000	ＲＣ	未改装
0	0	1	1	0	0	0	0	0	17369	新宿区	百人町	新大久保	31,000,000	3	25	2	１Ｋ	1,240,000	ＲＣ	未改装
0	0	1	1	0	0	0	0	0	17370	新宿区	百人町	新大久保	32,000,000	3	25	2	１Ｋ	1,280,000	ＲＣ	未改装
0	0	1	1	0	0	0	0	0	17371	新宿区	百人町	新大久保	32,000,000	3	25	2	１Ｋ	1,280,000	ＲＣ	未改装
0	0	1	1	0	0	0	0	0	17372	新宿区	百人町	新大久保	33,000,000	3	25	2	１Ｋ	1,320,000	ＲＣ	未改装
0	0	1	1	0	0	0	0	0	17373	新宿区	百人町	新大久保	33,000,000	3	25	2	１Ｋ	1,320,000	ＲＣ	未改装
0	0	1	1	0	0	0	0	0	17377	新宿区	百人町	新大久保	32,000,000	3	25	2	１Ｋ	1,280,000	ＲＣ	未改装
0	0	1	1	0	0	0	0	0	17379	新宿区	百人町	新大久保	31,000,000	3	25	2	１Ｋ	1,240,000	ＲＣ	
0	0	1	0	0	0	0	0	0	17380	新宿区	百人町	新大久保	30,000,000	3	25	2	１Ｋ	1,200,000	ＲＣ	未改装
0	0	1	1	0	0	0	0	0	54862	世田谷区	北沢	下北沢	80,000,000	5	65	15	３ＬＤＫ	1,230,769	ＲＣ	改装済
0	0	0	0	1	0	0	0	0	54866	世田谷区	北沢	下北沢	35,000,000	3	80	32	２ＬＤＫ	437,500	ＳＲＣ	未改装
0	0	1	1	1	0	0	0	0	55716	世田谷区	北沢	下北沢	860,000	2	55	28	２ＬＤＫ	15,636	ＳＲＣ	未改装
0	0	1	1	1	0	0	0	0	56249	世田谷区	代沢	下北沢	100,000,000	5	70	5	３ＬＤＫ	1,428,571	ＲＣ	未改装
0	0	1	1	1	0	0	0	0	56251	世田谷区	代沢	下北沢	110,000,000	5	75	3	３ＬＤＫ	1,466,667	ＲＣ	
0	0	0	0	1	0	0	0	0	59432	世田谷区	代沢	下北沢	28,000,000	12	70	19	３ＬＤＫ	400,000	ＲＣ	未改装
0	0	1	1	1	0	0	0	0	59951	世田谷区	北沢	下北沢	25,000,000	3	75	23	３ＬＤＫ	333,333	ＳＲＣ	未改装

A.3　提案手法１と提案手法２の判定結果の違い

提案手法１と提案手法２の判定結果の違いは次のとおり.

表 A.5: 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	手法１∪手法２	手法１∩手法２	No	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	建物の構造	改装
0	1	0	1	0	275	千代田区	神田淡路町	淡路町	150,000,000	3	2	70	３ＬＤＫ	2,142,857	ＳＲＣ	未改装
0	1	0	1	0	280	千代田区	神田淡路町	淡路町	260,000,000	3	2	115	４ＬＤＫ	2,260,870	ＳＲＣ	未改装
0	1	1	1	1	287	千代田区	神田淡路町	御茶ノ水	140,000,000	7	9	50	１ＬＤＫ	2,800,000	ＲＣ	未改装
0	1	1	1	1	703	千代田区	九段南	九段下	440,000,000	8	13	150	２ＬＤＫ	2,933,333	ＲＣ	未改装
0	0	1	1	0	1251	千代田区	内神田	淡路町	8,000,000	4	31	15	１Ｒ	533,333	ＳＲＣ	改装済
1	1	1	1	1	1446	千代田区	五番町	市ケ谷	640,000,000	1	32	90	３ＬＤＫ	7,111,111	ＳＲＣ	改装済
1	1	1	1	1	1623	千代田区	一番町	半蔵門	260,000,000	1	28	45	２ＬＤＫ	5,777,778	ＲＣ	
1	1	0	1	0	1682	千代田区	神田駿河台	御茶ノ水	85,000,000	5	32	15	１Ｋ	5,666,667	ＳＲＣ	未改装
1	1	0	1	0	1686	千代田区	神田駿河台	御茶ノ水	82,000,000	1	28	15	１Ｋ	5,466,667	ＳＲＣ	未改装
1	1	1	1	1	3697	中央区	日本橋小網町	茅場町	200,000,000	3	11	20	１Ｋ	10,000,000	ＳＲＣ	未改装
0	0	1	1	0	4598	中央区	八丁堀	銀座一丁目	19,000,000	13	9	25	１Ｋ	760,000	ＲＣ	
1	1	1	1	1	4759	中央区	月島	勝どき	210,000,000	6	39	45	２ＤＫ	4,666,667	ＳＲＣ	改装済
1	1	1	1	1	4797	中央区	月島	月島	120,000,000	6	39	25	２Ｋ	4,800,000	ＳＲＣ	改装済
1	1	1	1	1	5003	中央区	佃	月島	550,000,000	3	8	60	２ＬＤＫ	9,166,667	ＲＣ	未改装
1	1	1	1	1	5181	中央区	日本橋蛎殻町	水天宮前	160,000,000	3	6	20	１Ｋ	8,000,000	ＳＲＣ	未改装
1	1	1	1	1	5235	中央区	日本橋馬喰町	馬喰町	160,000,000	0	6	25	１Ｋ	6,400,000	ＳＲＣ	未改装
1	1	1	1	1	5239	中央区	日本橋箱崎町	水天宮前	240,000,000	3	30	55	３ＤＫ	4,363,636	ＳＲＣ	未改装
1	1	1	1	1	5394	中央区	湊	新富町 (東京)	220,000,000	5	10	35	１ＤＫ	6,285,714	ＲＣ	未改装
0	0	1	1	0	5465	中央区	銀座	銀座	96,000,000	5	27	105	３ＬＤＫ	914,286	ＳＲＣ	
1	0	1	1	0	5543	中央区	築地	築地	99,000,000	4	30	45	２ＤＫ	2,200,000	ＳＲＣ	未改装
1	1	1	1	1	6330	中央区	月島	勝どき	110,000,000	6	12	15	１Ｋ	7,333,333	ＲＣ	改装済
1	1	1	1	1	6896	港区	赤坂	赤坂 (東京)	960,000,000	5	7	65	１ＬＤＫ	14,769,231	ＲＣ	未改装
1	1	1	1	1	7395	港区	海岸	竹芝	150,000,000	1	36	20	１Ｋ	7,500,000	ＳＲＣ	改装済
1	1	1	1	1	7775	港区	港南	天王洲アイル	790,000,000	7	9	75		10,533,333	ＲＣ	未改装
1	1	1	1	1	8221	港区	芝	三田 (東京)	590,000,000	4	6	65	２ＬＤＫ	9,076,923	ＲＣ	未改装
1	1	1	1	1	9135	港区	白金台	白金台	1,400,000,000	4	35	150	２ＬＤＫ	9,333,333	ＳＲＣ	未改装
0	0	1	1	0	9212	港区	新橋	御成門	28,000,000	10	13	25	１ＤＫ	1,120,000	ＲＣ	未改装
0	1	0	1	0	9293	港区	台場	お台場海浜公園	93,000,000	2	8	45	１ＬＤＫ	2,066,667	ＲＣ	未改装
0	1	0	1	0	10258	港区	三田	麻布十番	350,000,000	5	6	75	３ＬＤＫ	4,666,667	ＲＣ	未改装
1	1	1	1	1	11042	港区	元麻布	麻布十番	3,200,000,000	7	13	160	４ＬＤＫ	20,000,000	ＲＣ	改装済
1	1	0	1	0	11098	港区	高輪	泉岳寺	600,000,000	5	39	75	３ＬＤＫ	8,000,000	ＳＲＣ	未改装
0	1	0	1	0	11208	港区	六本木	六本木	1,100,000,000	6	11	170		6,470,588	ＳＲＣ	
0	0	1	1	0	11417	港区	東麻布	神谷町	62,000,000	9	39	90	３ＬＤＫ	688,889	ＲＣ	改装済
0	1	1	1	1	11595	港区	港南	品川	200,000,000	6	8	70	２ＬＤＫ	2,857,143	鉄骨造	改装済
1	1	1	1	1	11751	港区	芝	三田 (東京)	100,000,000	3	32	30	１ＤＫ	3,333,333	ＳＲＣ	未改装
1	1	0	1	0	12350	港区	麻布十番	麻布十番	130,000,000	2	6	20	１ＤＫ	6,500,000	ＳＲＣ	未改装
1	1	1	1	1	12955	港区	海岸	日の出	220,000,000	1	7	30	１ＬＤＫ	7,333,333	ＲＣ	未改装
1	1	1	1	1	13250	港区	芝公園	芝公園	240,000,000	1	26	50	２ＬＤＫ	4,800,000	ＳＲＣ	
1	1	1	1	1	13263	港区	白金	白金台	1,100,000,000	10	9	80	２ＬＤＫ	13,750,000	ＳＲＣ	未改装
0	1	0	1	0	13307	港区	白金	白金高輪	170,000,000	2	2	60	１ＬＤＫ	2,833,333	ＲＣ	未改装
1	1	0	1	0	13426	港区	高輪	泉岳寺	360,000,000	5	10	40	１ＤＫ	9,000,000	ＳＲＣ	改装済
1	1	1	1	1	13771	港区	南麻布	広尾	1,100,000,000	3	34	115	２ＬＤＫ	9,565,217	ＳＲＣ	
0	0	1	1	0	13856	新宿区	市谷加賀町	牛込柳町	130,000,000	7	24	60	４ＬＤＫ	2,166,667	ＲＣ	未改装
1	1	1	1	1	14313	新宿区	上落合	落合 (東京)	95,000,000	2	27	15	１Ｋ	6,333,333	ＲＣ	未改装

表 A.5: 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	手法 1 ∪ 手法 2	手法 1 ∩ 手法 2	No	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	建物の構造	改装
1	1	0	1	0	14516	新宿区	北新宿	大久保 (東京)	140,000,000	7	43	25	1K	5,600,000	SRC	改装済
1	1	1	1	1	15134	新宿区	荒木町	曙橋	1,300,000,000	1	8	140	3LDK	9,285,714	SRC	未改装
0	0	1	1	0	15664	新宿区	大京町	国立競技場	24,000,000	4	34	70	2LDK	342,857	SRC	未改装
0	1	1	1	1	16069	新宿区	中井	中井	58,000,000	2	7	30	2LDK	1,933,333	RC	改装済
1	1	1	1	1	16099	新宿区	中落合	落合南長崎	170,000,000	2	11	20	1K	8,500,000	RC	未改装
1	1	1	1	1	16386	新宿区	下落合	高田馬場	150,000,000	5	41	45	1DK	3,333,333	SRC	改装済
0	0	1	1	0	18608	新宿区	西新宿	都庁前	25,000,000	6	5	20	1K	1,250,000	SRC	未改装
1	1	0	1	0	18910	新宿区	北新宿	大久保 (東京)	95,000,000	9	20	20	1K	4,750,000	SRC	未改装
1	0	1	1	0	19354	新宿区	歌舞伎町	東新宿	80,000,000	5	24	25	2LDK	3,200,000	SRC	未改装
1	1	1	1	1	19787	新宿区	富久町	曙橋	98,000,000	4	36	20	1K	4,900,000	SRC	改装済
1	1	0	1	0	19941	新宿区	西新宿	新宿西口	89,000,000	1	26	15	1K	5,933,333	SRC	未改装
1	1	1	1	1	19964	新宿区	西新宿	都庁前	84,000,000	6	25	20	1K	4,200,000	RC	改装済
1	1	1	1	1	19973	新宿区	西新宿	西新宿	200,000,000	4	7	20	1K	10,000,000	RC	未改装
1	1	1	1	1	20170	新宿区	弁天町	牛込柳町	480,000,000	4	37	80	2LDK	6,000,000	SRC	改装済
1	1	0	1	0	21100	文京区	小日向	江戸川橋	530,000,000	7	17	80	3LDK	6,625,000	RC	未改装
1	1	0	1	0	21263	文京区	関口	江戸川橋	600,000,000	4	15	85	3LDK	7,058,824	RC	未改装
0	1	1	1	1	21700	文京区	千駄木	本駒込	42,000,000	12	8	15	2LDK	2,800,000	RC	未改装
1	1	0	1	0	22502	文京区	本駒込	駒込	900,000,000	8	17	100	4LDK	9,000,000	SRC	未改装
0	1	0	1	0	22525	文京区	本駒込	巣鴨	140,000,000	6	4	55	3LDK	2,545,455	RC	改装済
0	0	1	1	0	22856	文京区	湯島	上野広小路	23,000,000	2	10	20	1K	1,150,000	RC	未改装
1	1	1	1	1	23113	文京区	本郷	春日 (東京)	200,000,000	2	8	20	1K	10,000,000	RC	未改装
1	1	1	1	1	23161	文京区	本駒込	駒込	220,000,000	8	5	30		7,333,333	RC	
1	1	1	1	1	23183	文京区	向丘	東大前	500,000,000	4	10	60	3LDK	8,333,333	SRC	改装済
1	1	1	1	1	23337	文京区	小石川	茗荷谷	610,000,000	4	6	65	2LDK	9,384,615	SRC	未改装
1	1	0	1	0	23601	文京区	音羽	江戸川橋	130,000,000	3	11	20	1K	6,500,000	SRC	改装済
1	1	0	1	0	23872	文京区	大塚	新大塚	73,000,000	5	12	15	1K	4,866,667	RC	改装済
1	1	0	1	0	24082	文京区	関口	江戸川橋	400,000,000	7	3	55	2LDK	7,272,727	RC	未改装
1	1	1	1	1	24580	台東区	浅草橋	浅草橋	250,000,000	3	26	50	3LDK	5,000,000	SRC	未改装
0	0	1	1	0	25458	台東区	台東	末広町 (東京)	4,500,000	8	28	25	1DK	180,000	SRC	改装済
0	0	1	1	0	25695	台東区	根岸	鶯谷	35,000,000	2	9	25	1K	1,400,000	RC	未改装
1	1	1	1	1	26332	台東区	駒形	浅草	250,000,000	2	6	45	1LDK	5,555,556	RC	未改装
0	0	1	1	0	26839	台東区	今戸	浅草橋	26,000,000	16	14	60	2LDK	433,333	SRC	改装済
0	0	1	1	0	27092	台東区	池之端	京成上野	19,000,000	9	21	20	1DK	950,000	SRC	改装済
0	0	1	1	0	28676	墨田区	東墨田	八広	15,000,000	6	27	15	1K	1,000,000	RC	未改装
0	1	0	1	0	28711	墨田区	東向島	東向島	22,000,000	6	9	15	1K	1,466,667	RC	
0	1	0	1	0	28755	墨田区	東向島	東向島	26,000,000	7	3	20	1K	1,300,000	RC	
0	1	0	1	0	28888	墨田区	文花	小村井	21,000,000	5	9	15	1K	1,400,000	RC	未改装
1	1	0	1	0	28995	墨田区	本所	本所吾妻橋	60,000,000	9	25	20	1K	3,000,000	RC	未改装
1	1	1	1	1	29423	墨田区	立花	東あずま	220,000,000	3	17	50	2LDK	4,400,000	RC	未改装
1	1	0	1	0	29704	墨田区	本所	本所吾妻橋	53,000,000	9	21	20	1K	2,650,000	RC	未改装
1	1	1	1	1	30268	墨田区	菊川	菊川 (東京)	150,000,000	3	23	50	2LDK	3,000,000	SRC	未改装
1	1	0	1	0	31486	江東区	亀戸	亀戸	63,000,000	13	41	25	1DK	2,520,000	SRC	
0	1	1	1	1	32692	江東区	東雲	辰巳	92,000,000	6	9	70		1,314,286	RC	
1	1	0	1	0	33088	江東区	亀戸	亀戸	59,000,000	16	36	25	1K	2,360,000	SRC	未改装
1	1	1	1	1	34831	江東区	住吉	住吉 (東京)	360,000,000	6	6	65	2LDK	5,538,462	RC	改装済
1	1	1	1	1	36404	江東区	豊洲	豊洲	300,000,000	11	9	60	1DK	5,000,000	RC	未改装
0	0	1	1	0	36581	江東区	有明	有明 (東京)	27,000,000	11	3	60	2LDK	450,000	RC	未改装
1	1	1	1	1	37283	江東区	東陽	東陽町	210,000,000	4	33	60	3LDK	3,500,000	SRC	未改装
1	1	1	1	1	37553	江東区	南砂	住吉 (東京)	200,000,000	14	27	55	2LDK	3,636,364	SRC	未改装
0	1	1	1	1	38182	品川区	勝島	大井競馬場前	21,000,000	8	4	20	1K	1,050,000	RC	改装済
1	1	1	1	1	39820	品川区	小山台	不動前	150,000,000	6	22	20	1K	7,500,000	RC	未改装

表 A.5: 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	手法 1 ∪ 手法 2	手法 1 ∩ 手法 2	No	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	建物の構造	改装
1	1	1	1	1	40145	品川区	中延	中延	180,000,000	6	10	20	１Ｋ	9,000,000	ＲＣ	未改装
1	1	1	1	1	40546	品川区	西五反田	五反田	320,000,000	1	31	50	２ＬＤＫ	6,400,000	ＳＲＣ	改装済
1	1	0	1	0	42270	品川区	南大井	大森 (東京)	150,000,000	9	4	20	１Ｋ	7,500,000	ＲＣ	
1	1	1	1	1	42339	品川区	荏原	西小山	330,000,000	7	19	45	１ＤＫ	7,333,333	ＲＣ	改装済
1	1	0	1	0	42432	品川区	上大崎	高輪台	77,000,000	9	23	15	１Ｋ	5,133,333	ＲＣ	未改装
1	1	1	1	1	43095	品川区	南品川	大井町	2,600,000,000	8	23	270		9,629,630	ＳＲＣ	
0	0	1	1	0	43117	品川区	八潮	品川シーサイド	27,000,000	23	25	70	３ＬＤＫ	385,714	ＳＲＣ	未改装
1	1	1	1	1	43448	目黒区	上目黒	中目黒	1,400,000,000	2	8	70	２ＬＤＫ	20,000,000	ＲＣ	未改装
1	1	1	1	1	43931	目黒区	鷹番	学芸大学	120,000,000	5	29	15	１Ｋ	8,000,000	ＲＣ	未改装
1	1	1	1	1	44302	目黒区	東山	池尻大橋	220,000,000	4	10	20	１Ｋ	11,000,000	ＲＣ	未改装
1	1	1	1	1	44332	目黒区	碑文谷	学芸大学	390,000,000	18	45	60	３ＬＤＫ	6,500,000	ＲＣ	未改装
1	1	1	1	1	44714	目黒区	五本木	祐天寺	110,000,000	7	8	35	４ＬＤＫ	3,142,857	ＲＣ	改装済
1	1	1	1	1	45064	目黒区	緑が丘	自由が丘 (東京)	490,000,000	9	22	80	３ＬＤＫ	6,125,000	ＲＣ	未改装
1	1	1	1	1	45415	目黒区	下目黒	目黒	640,000,000	11	11	75	３ＬＤＫ	8,533,333	ＳＲＣ	未改装
1	1	1	1	1	45693	目黒区	上目黒	祐天寺	92,000,000	8	21	15	１Ｋ	6,133,333	ＲＣ	未改装
0	0	1	1	0	46010	目黒区	目黒	不動前	18,000,000	23	22	40	３ＤＫ	450,000	ＲＣ	未改装
1	0	1	1	0	47083	大田区	大森東	平和島	51,000,000	3	17	20	１Ｋ	2,550,000	鉄骨造	未改装
0	0	1	1	0	47307	大田区	大森南	昭和島	15,000,000	13	10	15		1,000,000	ＲＣ	未改装
1	1	0	1	0	47538	大田区	池上	池上	78,000,000	9	20	15	１Ｋ	5,200,000	ＲＣ	未改装
0	1	1	1	1	48809	大田区	千鳥	千鳥町	28,000,000	3	24	15	１Ｋ	1,866,667	ＲＣ	未改装
1	1	0	1	0	49315	大田区	中馬込	馬込	92,000,000	2	12	20	１Ｋ	4,600,000	ＲＣ	未改装
0	1	0	1	0	49998	大田区	羽田	穴守稲荷	17,000,000	4	27	15	１Ｋ	1,133,333	ＲＣ	未改装
0	1	0	1	0	50078	大田区	東糀谷	穴守稲荷	25,000,000	7	2	20	１Ｋ	1,250,000	ＲＣ	未改装
1	1	0	1	0	52302	大田区	南馬込	大森 (東京)	140,000,000	13	17	20	１Ｋ	7,000,000	ＲＣ	未改装
1	1	1	1	1	52333	大田区	南六郷	雑色	290,000,000	11	19	75	３ＬＤＫ	3,866,667	ＳＲＣ	改装済
1	1	0	1	0	52444	大田区	大森北	大森 (東京)	93,000,000	8	16	15	１Ｋ	6,200,000	ＲＣ	未改装
1	1	1	1	1	52897	大田区	下丸子	鵜の木	270,000,000	8	13	135	４ＬＤＫ	2,000,000	ＳＲＣ	改装済
0	1	1	1	1	53074	大田区	田園調布	田園調布	72,000,000	14	6	25	３ＬＤＫ	2,880,000	ＲＣ	
1	1	0	1	0	53199	大田区	西蒲田	池上	64,000,000	2	16	15	１Ｋ	4,266,667	ＳＲＣ	改装済
1	1	0	1	0	53348	大田区	東馬込	馬込	180,000,000	1	22	40	２ＤＫ	4,500,000	ＲＣ	改装済
1	1	1	1	1	53483	大田区	南馬込	西馬込	350,000,000	8	13	50	２ＬＤＫ	7,000,000	ＲＣ	未改装
1	1	0	1	0	54924	世田谷区	砧	成城学園前	88,000,000	15	24	15		5,866,667	ＲＣ	未改装
0	1	1	1	1	55328	世田谷区	桜	宮の坂	21,000,000	3	29	25	１Ｋ	840,000	ＲＣ	改装済
0	0	1	1	0	55332	世田谷区	桜	宮の坂	13,000,000	3	21	30	２ＬＤＫ	433,333	ＲＣ	未改装
0	0	1	1	0	56144	世田谷区	世田谷	世田谷	82,000,000	1	41	50	３ＬＤＫ	1,640,000	ＳＲＣ	改装済
1	1	1	1	1	56614	世田谷区	三軒茶屋	若林 (東京)	62,000,000	10	25	15	１Ｋ	4,133,333	ＲＣ	改装済
1	1	1	1	1	57981	世田谷区	松原	明大前	85,000,000	2	30	30	１ＤＫ	2,833,333	ＳＲＣ	未改装
1	1	0	1	0	58742	世田谷区	成城	成城学園前	78,000,000	3	27	15	１Ｋ	5,200,000	ＲＣ	改装済
0	0	1	1	0	58800	世田谷区	宮坂	宮の坂	50,000,000	7	9	95	３ＬＤＫ	526,316	ＲＣ	未改装
1	1	1	1	1	59018	世田谷区	東玉川	田園調布	380,000,000	6	15	50	２ＬＤＫ	7,600,000	ＲＣ	未改装
1	1	1	1	1	60375	世田谷区	千歳台	祖師ケ谷大蔵	360,000,000	15	5	70	２ＬＤＫ	5,142,857	ＲＣ	未改装
1	1	1	1	1	60451	世田谷区	等々力	尾山台	59,000,000	3	21	15	１Ｋ	3,933,333	ＲＣ	未改装
1	1	1	1	1	60639	世田谷区	松原	明大前	110,000,000	5	20	15	１Ｋ	7,333,333	ＳＲＣ	
1	1	1	1	1	60769	世田谷区	用賀	桜新町	74,000,000	10	19	15	１Ｋ	4,933,333	ＲＣ	未改装
0	1	1	1	1	60818	渋谷区	上原	代々木上原	160,000,000	8	13	55	３ＬＤＫ	2,909,091	ＲＣ	未改装
1	1	1	1	1	61548	渋谷区	渋谷	渋谷	1,100,000,000	11	11	105	３ＬＤＫ	10,476,190	ＲＣ	未改装
1	1	1	1	1	61871	渋谷区	道玄坂	渋谷	74,000,000	7	37	20	１ＤＫ	3,700,000	ＳＲＣ	未改装
1	1	1	1	1	62092	渋谷区	西原	代々木上原	80,000,000	5	28	15	１Ｋ	5,333,333	ＲＣ	未改装
1	1	1	1	1	62991	渋谷区	笹塚	笹塚	90,000,000	6	34	35	１ＤＫ	2,571,429	ＲＣ	未改装
1	1	1	1	1	63821	渋谷区	幡ケ谷	幡ケ谷	160,000,000	8	22	30	２ＤＫ	5,333,333	ＲＣ	未改装
1	1	1	1	1	63984	渋谷区	恵比寿	恵比寿	900,000,000	6	10	125	４ＬＤＫ	7,200,000	ＳＲＣ	未改装

表 A.5: 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	手法 1 ∪ 手法 2	手法 1 ∩ 手法 2	No	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	建物の構造	改装
0	0	1	1	0	64026	渋谷区	神山町	代々木公園	480,000,000	11	39	230	３ＬＤＫ	2,086,957	ＲＣ	未改装
1	1	1	1	1	64273	渋谷区	本町	初台	140,000,000	6	29	35	１ＤＫ	4,000,000	ＲＣ	未改装
1	1	1	1	1	64573	渋谷区	笹塚	笹塚	75,000,000	5	21	15	１Ｒ	5,000,000	ＲＣ	未改装
1	1	0	1	0	65062	渋谷区	本町	西新宿五丁目	79,000,000	6	20	15	１Ｋ	5,266,667	ＳＲＣ	未改装
1	1	1	1	1	65379	中野区	新井	中野 (東京)	52,000,000	12	27	10	１Ｋ	5,200,000	ＲＣ	未改装
1	1	0	1	0	66629	中野区	本町	西新宿五丁目	330,000,000	6	44	50	２ＬＤＫ	6,600,000	ＳＲＣ	改装済
1	1	1	1	1	67442	中野区	若宮	鷺ノ宮	480,000,000	4	13	80	３ＬＤＫ	6,000,000	ＲＣ	未改装
1	1	1	1	1	67747	中野区	中央	新中野	130,000,000	4	7	20	１Ｋ	6,500,000	ＲＣ	未改装
1	1	1	1	1	67793	中野区	中央	中野坂上	69,000,000	4	22	15	１Ｋ	4,600,000	ＲＣ	未改装
1	1	1	1	1	68109	杉並区	阿佐谷南	阿佐ケ谷	74,000,000	3	33	15	１Ｋ	4,933,333	ＲＣ	改装済
1	1	1	1	1	69034	杉並区	久我山	久我山	80,000,000	2	41	25	１ＤＫ	3,200,000	ＳＲＣ	
1	1	1	1	1	69436	杉並区	井草	井荻	470,000,000	8	12	85	４ＬＤＫ	5,529,412	ＲＣ	改装済
1	1	1	1	1	70351	杉並区	浜田山	西永福	1,200,000,000	10	26	160	３ＬＤＫ	7,500,000	ＲＣ	改装済
1	1	0	1	0	70638	杉並区	堀ノ内	方南町	89,000,000	5	20	15	１Ｋ	5,933,333	ＲＣ	未改装
1	1	1	1	1	71831	杉並区	成田東	南阿佐ケ谷	310,000,000	6	24	50	２ＬＤＫ	6,200,000	ＲＣ	未改装
1	1	1	1	1	71941	杉並区	和田	東高円寺	76,000,000	3	17	15	１Ｋ	5,066,667	ＲＣ	未改装
0	1	0	1	0	72527	杉並区	西荻北	西荻窪	59,000,000	11	4	30	３ＬＤＫ＋Ｓ	1,966,667	ＲＣ	未改装
1	1	0	1	0	72642	杉並区	堀ノ内	方南町	210,000,000	10	20	45	２ＤＫ	4,666,667	ＲＣ	未改装
1	1	1	1	1	72998	豊島区	池袋本町	北池袋	78,000,000	11	31	15		5,200,000	ＲＣ	
1	1	0	1	0	73341	豊島区	北大塚	大塚 (東京)	240,000,000	4	10	35	１ＬＤＫ	6,857,143	ＲＣ	未改装
1	1	1	1	1	74674	豊島区	要町	千川	100,000,000	4	14	20	１Ｋ	5,000,000	ＲＣ	改装済
1	0	1	1	0	75189	豊島区	高田	早稲田 (都電)	85,000,000	4	33	30	１ＤＫ	2,833,333	ＲＣ	未改装
1	1	1	1	1	75193	豊島区	千早	要町	100,000,000	8	25	15	１Ｒ	6,666,667	ＲＣ	未改装
1	1	1	1	1	75695	豊島区	池袋本町	北池袋	100,000,000	5	11	15	１Ｋ	6,666,667	ＲＣ	改装済
1	1	0	1	0	75739	豊島区	北大塚	大塚 (東京)	300,000,000	8	14	50	２ＬＤＫ	6,000,000	ＳＲＣ	改装済
1	1	1	1	1	75754	豊島区	巣鴨	巣鴨	280,000,000	12	12	50	２ＬＤＫ	5,600,000	ＳＲＣ	未改装
1	1	1	1	1	75780	豊島区	高田	高田馬場	86,000,000	6	24	15	１Ｋ	5,733,333	ＲＣ	未改装
1	1	1	1	1	75828	豊島区	西巣鴨	西巣鴨	57,000,000	4	20	20	１Ｋ	2,850,000	ＲＣ	未改装
1	1	0	1	0	75858	豊島区	東池袋	新大塚	130,000,000	2	36	30	３ＤＫ	4,333,333	ＳＲＣ	未改装
1	1	0	1	0	76109	豊島区	駒込	駒込	560,000,000	5	2	55	２ＬＤＫ	10,181,818	ＲＣ	未改装
1	1	1	1	1	76213	豊島区	高田	目白	120,000,000	12	12	20	１Ｋ	6,000,000	ＲＣ	未改装
1	1	1	1	1	76276	豊島区	西池袋	池袋	230,000,000	5	9	30	１ＤＫ	7,666,667	ＳＲＣ	改装済
0	1	1	1	1	77665	北区	西ケ原	上中里	44,000,000	7	9	55	２ＬＤＫ	800,000	ＲＣ	未改装
0	0	1	1	0	77711	北区	赤羽北	北赤羽	76,000,000	7	14	70	４ＬＤＫ	1,085,714	ＲＣ	未改装
1	1	1	1	1	77935	北区	王子	王子	78,000,000	9	18	15	１Ｒ	5,200,000	ＳＲＣ	改装済
0	0	1	1	0	77979	北区	滝野川	飛鳥山	7,000,000	3	26	20	１ＤＫ	350,000	ＲＣ	未改装
0	0	1	1	0	78093	北区	堀船	栄町 (東京)	11,000,000	6	31	40	１ＤＫ＋Ｓ	275,000	ＳＲＣ	未改装
1	1	1	1	1	78450	北区	西ケ原	上中里	94,000,000	9	23	20	１Ｋ	4,700,000	ＲＣ	改装済
1	1	1	1	1	78741	北区	田端	田端	190,000,000	7	6	30	１ＬＤＫ	6,333,333	ＲＣ	未改装
0	0	1	1	0	79070	荒川区	西尾久	荒川遊園地前	14,000,000	6	29	45	３ＬＤＫ	311,111	ＲＣ	未改装
0	0	1	1	0	79388	荒川区	東日暮里	荒川一中前	41,000,000	5	4	65	３ＬＤＫ	630,769	ＲＣ	未改装
0	0	1	1	0	79969	荒川区	東尾久	赤土小学校前	8,600,000	11	20	25	１Ｋ	344,000	ＲＣ	未改装
1	1	0	1	0	80001	荒川区	東日暮里	三ノ輪	68,000,000	5	24	25	１ＤＫ	2,720,000	ＲＣ	改装済
0	0	1	1	0	80270	荒川区	西尾久	荒川車庫前	3,200,000	4	9	70	３ＬＤＫ	45,714	ＲＣ	未改装
1	1	1	1	1	80448	荒川区	西尾久	尾久	64,000,000	3	21	20	１Ｋ	3,200,000	ＲＣ	改装済
0	0	1	1	0	80466	荒川区	西尾久	上中里	18,000,000	13	21	55	３ＬＤＫ	327,273	ＲＣ	未改装
1	1	1	1	1	80626	荒川区	南千住	南千住	170,000,000	9	9	60	３ＬＤＫ	2,833,333	ＲＣ	未改装
0	1	0	1	0	81766	板橋区	坂下	蓮根	19,000,000	12	6	15	１Ｋ	1,266,667	ＲＣ	
0	1	0	1	0	82054	板橋区	高島平	高島平	16,000,000	4	8	15	１Ｋ	1,066,667	ＲＣ	未改装
0	1	0	1	0	83229	板橋区	舟渡	浮間舟渡	23,000,000	6	8	20	１Ｋ	1,150,000	ＲＣ	未改装
1	1	1	1	1	84972	板橋区	南町	要町	180,000,000	7	15	35	１ＤＫ	5,142,857	ＳＲＣ	改装済

表 A.5: 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	手法 1 ∪ 手法 2	手法 1 ∩ 手法 2	No	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	建物の構造	改装
1	1	1	1	1	85123	板橋区	板橋	板橋区役所前	210,000,000	7	9	60	３ＬＤＫ	3,500,000	ＲＣ	未改装
1	1	1	1	1	85395	板橋区	高島平	新高島平	48,000,000	12	17	15	１Ｋ	3,200,000	ＲＣ	未改装
1	1	1	1	1	85699	板橋区	氷川町	板橋区役所前	120,000,000	2	2	20	１Ｋ	6,000,000	ＳＲＣ	未改装
1	1	1	1	1	86965	練馬区	大泉町	大泉学園	250,000,000	28	12	65	３ＬＤＫ	3,846,154	ＲＣ	未改装
0	1	1	1	1	86968	練馬区	高松	光が丘	20,000,000	9	10	15	１Ｋ	1,333,333	ＲＣ	未改装
0	1	0	1	0	88225	練馬区	羽沢	氷川台	22,000,000	7	10	20	１Ｋ	1,100,000	ＲＣ	
1	1	1	1	1	88249	練馬区	早宮	豊島園	42,000,000	14	25	15	１Ｋ	2,800,000	ＲＣ	未改装
1	1	1	1	1	89079	練馬区	桜台	新桜台	55,000,000	2	22	15	１Ｋ	3,666,667	ＲＣ	未改装
1	1	1	1	1	89206	練馬区	関町北	武蔵関	58,000,000	9	36	35	１ＤＫ	1,657,143	ＲＣ	未改装
1	1	0	1	0	89641	練馬区	関町北	武蔵関	39,000,000	9	34	15	１Ｋ	2,600,000	ＲＣ	未改装
1	1	0	1	0	90123	練馬区	関町北	武蔵関	77,000,000	3	22	25	１Ｋ	3,080,000	ＲＣ	未改装
1	1	1	1	1	90140	練馬区	関町北	武蔵関	95,000,000	5	14	20	１Ｋ	4,750,000	ＲＣ	未改装
0	0	1	1	0	90281	練馬区	立野町	吉祥寺	47,000,000	26	11	75	３ＬＤＫ	626,667	ＲＣ	改装済
0	1	0	1	0	90933	足立区	綾瀬	綾瀬	23,000,000	9	9	20	１Ｋ	1,150,000	ＲＣ	未改装
0	1	0	1	0	91531	足立区	加平	北綾瀬	15,000,000	6	27	15	１Ｋ	1,000,000	ＳＲＣ	未改装
0	1	1	1	1	92644	足立区	舎人	見沼代親水公園	35,000,000	1	2	40	１ＬＤＫ	875,000	ＲＣ	未改装
1	1	1	1	1	93069	足立区	弘道	五反野	40,000,000	6	23	15	１Ｋ	2,666,667	ＲＣ	改装済
0	0	1	1	0	93288	足立区	六木	八潮	5,700,000	20	26	45	２ＬＤＫ	126,667	ＲＣ	未改装
1	1	1	1	1	94011	足立区	足立	五反野	61,000,000	7	21	15	１Ｒ	4,066,667	ＲＣ	未改装
1	1	0	1	0	94208	足立区	竹の塚	竹ノ塚	97,000,000	15	18	65	３ＬＤＫ	1,492,308	ＲＣ	未改装
1	1	1	1	1	94304	足立区	西保木間	竹ノ塚	180,000,000	22	5	55	２ＬＤＫ	3,272,727	ＲＣ	未改装
0	0	1	1	0	94673	足立区	小台	田端	16,000,000	28	29	80	４ＬＤＫ	200,000	ＲＣ	未改装
1	1	1	1	1	94801	足立区	島根	西新井	39,000,000	11	18	15	１Ｒ	2,600,000	ＲＣ	未改装
0	0	1	1	0	94809	足立区	新田	東十条	12,000,000	29	20	50	３ＤＫ	240,000	ＲＣ	未改装
1	1	0	1	0	95245	足立区	保木間	竹ノ塚	70,000,000	19	14	50	３ＤＫ	1,400,000	ＲＣ	未改装
1	1	1	1	1	95431	葛飾区	青戸	亀有	69,000,000	18	35	40	２ＤＫ	1,725,000	ＲＣ	未改装
0	1	0	1	0	96174	葛飾区	立石	京成立石	27,000,000	7	25	20	１Ｋ	1,350,000	ＲＣ	
1	1	1	1	1	97086	葛飾区	新宿	金町	110,000,000	18	19	55	２ＬＤＫ	2,000,000	ＳＲＣ	未改装
1	1	0	1	0	97109	葛飾区	東金町	金町	110,000,000	9	26	40	２ＬＤＫ	2,750,000	ＲＣ	未改装
1	1	0	1	0	97657	葛飾区	東金町	金町	270,000,000	14	12	95	４ＬＤＫ	2,842,105	ＲＣ	改装済
1	1	0	1	0	98022	葛飾区	宝町	お花茶屋	50,000,000	4	30	40	２ＤＫ	1,250,000	ＲＣ	未改装
1	1	1	1	1	98182	葛飾区	東立石	京成立石	430,000,000	12	16	55	３ＬＤＫ	7,818,182	ＲＣ	未改装
1	1	1	1	1	98251	葛飾区	堀切	堀切菖蒲園	62,000,000	10	18	35		1,771,429	ＲＣ	
1	1	1	1	1	98389	江戸川区	北葛西	西葛西	630,000,000	13	22	95	４ＬＤＫ	6,631,579	ＳＲＣ	未改装
1	1	1	1	1	99695	江戸川区	船堀	船堀	500,000,000	1	24	70	３ＬＤＫ	7,142,857	ＳＲＣ	未改装
1	1	1	1	1	100772	江戸川区	船堀	船堀	290,000,000	9	25	80	４ＬＤＫ	3,625,000	ＲＣ	改装済
1	1	1	1	1	101050	江戸川区	西葛西	西葛西	58,000,000	15	16	25	１Ｋ	2,320,000	ＲＣ	
1	1	1	1	1	101118	江戸川区	西小岩	小岩	360,000,000	10	9	75	３ＬＤＫ	4,800,000	ＳＲＣ	未改装
1	1	1	1	1	101208	江戸川区	船堀	一之江	420,000,000	1	7	70	３ＬＤＫ	6,000,000	ＲＣ	未改装

A.4 提案手法１と提案手法２の R コード

提案手法１の $SIQR_u$ のコードを A.1 に示す.

プログラム A.1: $SIQR_u$

```
1
2 k=10
3
4 rbscale.SIQR <- function(setdata){
5     maxthreshold <-(quantile(setdata,.75)-quantile(setdata
          ,.50))*k+quantile(setdata,.75)
6     setdata > maxthreshold
7 }
8
9 outliersSIQR=NULL
10
11 for(s in unique(station.num$station)){
12     set.data <- subset(input.data, station %in% s)
13     fence <- rbscale.SIQR(set.data$perprice)
14     outliersSIQR <- as.data.frame(rbind(outliersSIQR,set.data[
          fence,]))
15     rm(set.data)
16 }
```

ループ部分は次のコードでもよい.

```
1 install.packages("dplyr")
2 library("dplyr")
3
4 outliersSIQR <- input.data %>% group_by(station) %>%
     summarise(across(colnames(.data),.fns = list(~.x)) ,
     outlier = rbscale.SIQR(perprice)) %>% filter(outlier == T)
      %>% as.data.frame
```

提案手法 2 の階層的クラスタリングのコードを A.2 に示す.

プログラム A.2: Hierarchical clustering

```
c=2.5

rbscale.std <- function(setdata){
    IQR<-(quantile(setdata,.75)-quantile(setdata,.50))*c
    (setdata-median(setdata))/IQR
}

outliers=NULL
branchlength=4
dd=c("perprice","minutes","age") #select variables

for(s in unique(station.num$station)){
    set.data <- subset(input.data, station %in% s)
    dist.data1 <- apply(set.data[,dd], 2, rbscale.std)
    hdata1 <- dist(dist.data1)
    hdataclust1 <- try(hclust(hdata1,method="ward.D2"))

    if (class(hdataclust1) == "try-error") {
  rm(set.data,dist.data1,hdata1,hdataclust1)
  next
    }else{
        hdataclust1merge <- as.data.frame(hdataclust1$merge)
        outcan1 <- hdataclust1$height[which(hdataclust1merge
            [,1]<0)]>branchlength
  if(sum(outcan1==TRUE)==0){
      rm(set.data,dist.data1,hdata1,hdataclust1)
  next
  }else{
      outpoint1 <- hdataclust1merge[which(hdataclust1merge
          [,1]<0)[outcan1],]
      outliers <- as.data.frame(rbind(outliers,set.data[-
          outpoint1[outpoint1 <0],]))
  }
    }
}
```

付 録B　第4章の補足

B.1　Cookの距離 [46] の変形

時間の経過とともに新たに得られた $n+1$ 組目の $(\boldsymbol{x}_{n+1}, y_{n+1})$ が既存の線形回帰モデルに与える影響を算出するため，既存のモデル推定の根拠となる n 件のデータに対して，データ点を 1 つ追加する場合として，Cook [46] によって提案された本文中の式（4.3）を次の形に変形する．なお，$\boldsymbol{X}_n^\top \boldsymbol{X}_n$，$\boldsymbol{X}_{n+1}^\top \boldsymbol{X}_{n+1}$ とも正則行列とする．

$$D_{n+1} = \frac{(\hat{\boldsymbol{\beta}}_{n+1} - \hat{\boldsymbol{\beta}}_n)^\top \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} (\hat{\boldsymbol{\beta}}_{n+1} - \hat{\boldsymbol{\beta}}_n)}{(d+1)s^2} \tag{B.1}$$

$$\hat{\boldsymbol{\beta}}_{n+1} - \hat{\boldsymbol{\beta}}_n = (\boldsymbol{X}_{n+1}^\top \boldsymbol{X}_{n+1})^{-1} \boldsymbol{x}_{n+1} (y_{n+1} - \boldsymbol{x}_{n+1}^\top \hat{\boldsymbol{\beta}}_n). \tag{B.2}$$

ここで右辺の逆行列部分の $\boldsymbol{X}_{n+1}^\top \boldsymbol{X}_{n+1}$ について，

$$\boldsymbol{X}_{n+1}^\top \boldsymbol{X}_{n+1} = \boldsymbol{X}_n^\top \boldsymbol{X}_n + \boldsymbol{x}_{n+1} \boldsymbol{x}_{n+1}^\top \tag{B.3}$$

であるから，逆行列の補題 (Hager [62]) を適用して次の形が得られる:

$$\begin{aligned} (\boldsymbol{X}_{n+1}^\top \boldsymbol{X}_{n+1})^{-1} &= (\boldsymbol{X}_n^\top \boldsymbol{X}_n + \boldsymbol{x}_{n+1} \boldsymbol{x}_{n+1}^\top)^{-1} \\ &= (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} - \frac{(\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{x}_{n+1} \boldsymbol{x}_{n+1}^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1}}{1 + \boldsymbol{x}_{n+1}^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{x}_{n+1}}. \end{aligned} \tag{B.4}$$

したがって，

$$\begin{aligned} (\boldsymbol{X}_{n+1}^\top \boldsymbol{X}_{n+1})^{-1} \boldsymbol{x}_{n+1} &= (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{x}_{n+1} - \frac{(\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{x}_{n+1} \boldsymbol{x}_{n+1}^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{x}_{n+1}}{1 + \boldsymbol{x}_{n+1}^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{x}_{n+1}} \\ &= \frac{(\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{x}_{n+1}}{1 + \boldsymbol{x}_{n+1}^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{x}_{n+1}}. \end{aligned} \tag{B.5}$$

よって，

$$\hat{\boldsymbol{\beta}}_{n+1} - \hat{\boldsymbol{\beta}}_n = \frac{(\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{x}_{n+1}}{1 + \boldsymbol{x}_{n+1}^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{x}_{n+1}} \quad (y_{n+1} - \boldsymbol{x}_{n+1}^\top \hat{\boldsymbol{\beta}}_n). \tag{B.6}$$

したがって，

$$\begin{aligned} D_{n+1} &= \frac{(\hat{\boldsymbol{\beta}}_{n+1} - \hat{\boldsymbol{\beta}})^\top \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} (\hat{\boldsymbol{\beta}}_{n+1} - \hat{\boldsymbol{\beta}})}{(d+1)s^2} \\ &= \left(\frac{y_{n+1} - \boldsymbol{x}_{n+1}^\top \hat{\boldsymbol{\beta}}_n}{s\{1 + \boldsymbol{x}_{n+1}^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{x}_{n+1}\}} \right)^2 \frac{\boldsymbol{x}_{n+1}^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{x}_{n+1}}{d+1}. \quad \square \end{aligned} \tag{B.7}$$

B.2　可変ウィンドウの長さと MAPE の結果

4.6 節で示した東京都江東区と東京都大田区のデータを用いた MAPE の結果を示す.

表 B.1: 江東区の可変ウィンドウの長さと MAPE

予測時点		2011/7	2012/1	2012/7	2013/1	2013/7	2014/1	2014/7	2015/1	2015/7	2016/1
ウィンドウの長さ l_t		4.5 年	5.0 年	3.5 年	4.0 年	4.5 年	5.0 年	4.5 年	1.0 年	1.5 年	1.0 年
MAPE	可変	19.96%	23.73%	15.29%	15.19%	17.57%	20.12%	15.61%	13.99%	22.16%	14.72%
	0.5 年	24.61%	23.35%	16.63%	23.02%	19.14%	19.49%	20.30%	14.94%	18.67%	14.09%
	2 年	19.62%	23.70%	15.28%	18.39%	19.05%	20.34%	14.64%	13.74%	22.45%	17.11%
	5 年	20.30%	23.73%	15.12%	16.80%	17.92%	20.12%	15.72%	14.63%	21.12%	19.70%

予測時点		2016/7	2017/1	2017/7	2018/1	2018/7	2019/1	2019/7	2020/1	2020/7	2021/1
ウィンドウの長さ l_t		1.5 年	2.0 年	2.5 年	2.0 年	2.5 年	3.0 年	3.5 年	4.0 年	4.5 年	5.0 年
MAPE	可変	13.46%	16.65%	14.18%	12.00%	12.53%	13.05%	11.41%	14.38%	13.14%	15.97%
	0.5 年	16.73%	16.23%	14.49%	13.14%	12.80%	12.84%	12.25%	13.53%	13.18%	14.51%
	2 年	14.58%	16.65%	14.00%	12.00%	12.51%	13.09%	11.64%	14.61%	12.94%	14.52%
	5 年	16.10%	19.72%	16.48%	14.14%	12.81%	13.20%	11.75%	14.43%	13.28%	15.97%

表 B.2: 大田区の可変ウィンドウの長さと MAPE

予測時点		2011/7	2012/1	2012/7	2013/1	2013/7	2014/1	2014/7	2015/1	2015/7	2016/1
ウィンドウの長さ l_t		4.0 年	4.5 年	5.0 年	3.0 年	3.5 年	4.0 年	4.5 年	4.5 年	1.0 年	1.5 年
MAPE	可変	14.18%	13.14%	17.17%	16.08%	13.59%	15.22%	16.77%	14.42%	13.66%	15.00%
	0.5 年	15.22%	13.12%	18.51%	15.57%	13.98%	15.45%	16.84%	14.67%	13.98%	14.94%
	2 年	14.20%	13.10%	18.16%	16.07%	13.39%	15.25%	16.96%	14.52%	14.51%	15.42%
	5 年	13.82%	13.28%	17.17%	0.159	0.134	0.152	0.169	0.145	0.152	0.171

予測時点		2016/7	2017/1	2017/7	2018/1	2018/7	2019/1	2019/7	2020/1	2020/7	2021/1
ウィンドウの長さ l_t		1.0 年	1.5 年	2.0 年	2.5 年	3.0 年	3.5 年	4.0 年	4.5 年	5.0 年	3.0 年
MAPE	可変	12.68%	14.16%	15.77%	13.24%	15.19%	13.29%	11.48%	17.43%	17.37%	13.82%
	0.5 年	14.50%	14.60%	17.58%	13.90%	15.96%	13.62%	12.70%	16.72%	19.17%	15.39%
	2 年	12.42%	14.45%	15.77%	13.04%	15.37%	13.18%	11.11%	17.25%	17.59%	14.03%
	5 年	14.75%	16.72%	16.06%	14.67%	15.93%	13.77%	11.94%	17.57%	17.37%	13.61%

B.3　可変ウィンドウに設定するパラメータ

本文中の可変ウィンドウによるモデル更新フローで用いた 3 つのパラメータ，変化点判定のために Cook の距離を変形して算出した閾値，その閾値を超えたサンプルが予測対象のデータに占める割合，ウィンドウの長さを調整するための F 検定の有意水準についてそれぞれ υ は $(0.9, 0.7, 0.5)$，s は $(0.1, 0.3, 0.5)$，α は $(0.1, 0.01)$ の合計 18 通りの組み合わせで MER と 90%点を比較した結果を示す．また固定スライディングウィンドウの結果も比較の参考として併記する．本文中で設定した $(\upsilon, s, \alpha) = (0.9, 0.1, 0.01)$ は江東区と大田区の両地域で MER と 90%点が良好な組み合わせを採用している．

表 B.3: 江東区のパラメータ比較

υ	s	α	MER	90%点
0.9	**0.1**	**0.01**	11.71%	28.60%
0.9	0.1	0.1	11.75%	28.81%
0.9	0.3	0.01	11.88%	29.97%
0.9	0.3	0.1	11.85%	29.11%
0.9	0.5	0.01	11.52%	30.09%
0.9	0.5	0.1	11.52%	30.09%
0.7	0.1	0.01	11.62%	29.00%
0.7	0.1	0.1	11.75%	29.09%
0.7	0.3	0.01	11.88%	29.32%
0.7	0.3	0.1	11.75%	28.81%
0.7	0.5	0.01	11.88%	29.97%
0.7	0.5	0.1	11.88%	29.97%
0.5	0.1	0.01	11.62%	29.00%
0.5	0.1	0.1	11.75%	29.09%
0.5	0.3	0.01	11.62%	29.00%
0.5	0.3	0.1	11.75%	29.09%
0.5	0.5	0.01	11.75%	29.09%
0.5	0.5	0.1	11.84%	29.09%
0.5 年固定			11.24%	33.18%
2 年固定			11.72%	29.27%
5 年固定			12.24%	30.56%

表 B.4: 大田区のパラメータ比較

υ	s	α	MER	90%点
0.9	**0.1**	**0.01**	11.35%	28.94%
0.9	0.1	0.1	11.26%	29.00%
0.9	0.3	0.01	11.32%	28.75%
0.9	0.3	0.1	11.32%	28.75%
0.9	0.5	0.01	11.32%	28.75%
0.9	0.5	0.1	11.32%	28.75%
0.7	0.1	0.01	11.38%	29.44%
0.7	0.1	0.1	11.28%	29.89%
0.7	0.3	0.01	11.26%	29.40%
0.7	0.3	0.1	11.24%	29.69%
0.7	0.5	0.01	11.32%	28.75%
0.7	0.5	0.1	11.32%	28.75%
0.5	0.1	0.01	11.38%	29.44%
0.5	0.1	0.1	11.38%	29.89%
0.5	0.3	0.01	11.38%	29.44%
0.5	0.3	0.1	11.28%	29.89%
0.5	0.5	0.01	11.27%	28.77%
0.5	0.5	0.1	11.38%	29.00%
0.5 年固定			11.58%	31.12%
2 年固定			11.43%	29.09%
5 年固定			12.30%	30.20%

B.4 可変ウィンドウスキームの R コード

可変ウィンドウの伸縮のための F 検定のコードは次のとおり.

プログラム B.1: Ftest

```
ftest <- function(res_r,res_k) {
  dfr <- df.residual(res_r) # n-r-1
  dfk <- df.residual(res_k) # n-k-1
  nmrtr <- (sum(resid(res_r)^2) - sum(resid(res_k)^2))/(dfr-
     dfk) # (RSS^(r)-RSS^(k))/(k-r-1)
  dnmntr <- sum(resid(res_k)^2)/dfk # RSS^(k) / (n'-k)
  fvalue <- nmrtr/dnmntr # Fstat.
  pvalue <- pf(fvalue,dfr-dfk,dfk,lower.tail=F)
  list(fvalue=fvalue,df1=dfr-dfk,df2=dfk,pvalue=pvalue)
}
```

可変ウィンドウスキームの検証に用いたコードは次のとおり.

プログラム B.2: modelupdating

```
upsilon <- 0.9 # 0.7 #0.5
s <- 0.1 #0.1 #0.3 #0.5
alpha <- 0.01 #0.1

features <- c("unit_size"
              ,"walk_minute"
              ,"Months"
              ,"total_story"
              ,"room_count"
              ,"per_price"
)

rm(X,y,hat0,res0,x1,pred_error,pred_errorl,hat1,sd1,
   pred_cook_d,new_data,initial_data,threshold,lm_result,i,j,
   k,res_r,res_k)

initial_data <- updating_data[contract_date >= "2006-01-01" &
    contract_date < updating_period[5],]
lm_result <- lm(per_price~.,data = initial_data[,features,
   with = F])
threshold <- quantile(cooks.distance(lm_result),probs =
   upsilon)
k <- 2

pred_error_list <- NULL
elapsed_process <- data.frame(matrix(rep(NA, length(
   updating_period[5:(length(updating_period)-1)])*6), ncol
```

```
    =6,nrow=length(updating_period[5:(length(updating_period)
    -1)])))
colnames(elapsed_process) <- c("elapsedtime","sample","
    bwindow","awindow","thrshld","r2")

for(i in 5:(length(updating_period)-1)){
  X <- cbind(rep(1,nrow(initial_data)),initial_data[,features
      , with = F])
  X <- na.omit(X)
  y <- as.matrix(X[,"per_price"])
  X <- as.matrix(X[,-ncol(X),with =F])
  hat0 <- X \%*\% solve((t(X) \%*\% X)) \%*\% t(X)
  res0 <- y - hat0 \%*\% y
  new_data <- updating_data[contract_date >= updating_period[
      i] & contract_date < updating_period[i+1],]
  x1 <- cbind(rep(1,nrow(new_data)),new_data[,c(features,"
      property_id"), with = F])
  x1 <- na.omit(x1)
  pred_error <- x1$per_price - predict(lm_result,x1[,features
      , with = F])
  pred_errorl <- cbind(as.data.frame.Date(updating_period[i
      +1]),property_id=x1$property_id,pred_error,x1$per_price,
      pred_error/x1$per_price)
  pred_error_list <- rbind(pred_error_list,pred_errorl)
  x1 <- x1[,-ncol(x1),with=F]
  x1 <- as.matrix(x1[,-ncol(x1),with =F])
  hat1 <- x1 \%*\% solve((t(X) \%*\% X)) \%*\% t(x1)
  sd1 <- sqrt((t(res0) \%*\% res0) / (nrow(X)-ncol(x1)))
  pred_cook_d <- ((pred_error/(rep(sd1,nrow(x1))*(rep(1,nrow(
      x1)) + diag(hat1))))^2*diag(hat1))/ncol(x1)

  #print(updating_period[i+1])

  elapsed_process[(i-4),1] <- updating_period[i+1]
  elapsed_process[(i-4),2] <- paste(sum(threshold <=
      pred_cook_d),"/",length(pred_cook_d),sep = "")
  elapsed_process[(i-4),3] <- length(unique(
      initial_data$period))

  initial_data <- rbind(initial_data,new_data)

  if(sum(threshold <= pred_cook_d)/length(pred_cook_d) > s |
      length(unique(initial_data$period)) > 10){
    if(length(unique(initial_data$period)) < 2){
      break
    }
```

```
55      f_test_p <- NULL
56      ctime <- k
57      for(j in k:(i+1)){
58        initial_data$judge <- as.numeric(
             initial_data$contract_date >= updating_period[j])
59        if(nrow(unique(initial_data[initial_data$judge==1,"
             period"])) < 2){
60          initial_data <- initial_data[,-"judge"]
61          break
62        }
63        res_r <- lm(per_price~.,data = initial_data[,features,
             with = F])
64        res_k <- lm(per_price~.*judge,data = initial_data[,c(
             features,"judge"), with = F])
65
66        if(is.null(f_test_p)){
67          f_test_p <<- ftest(res_r,res_k)$pvalue
68          next
69        }
70
71        if(ftest(res_r,res_k)$pvalue < f_test_p){
72          f_test_p <<- ftest(res_r,res_k)$pvalue
73          ctime <<- j
74        }
75      }
76      if(f_test_p > (alpha/(length(unique(initial_data$period))
           -1)) & length(unique(initial_data$period)) < 10 ){
77      } else {
78        initial_data <- initial_data[contract_date >=
             updating_period[ctime],]
79        k <- ctime+1
80      }
81    }
82
83    elapsed_process[(i-4),4] <- length(unique(
         initial_data$period))
84    elapsed_process[(i-4),5] <- threshold
85    lm_result <- lm(per_price~.,data = initial_data[,features,
         with = F])
86    threshold <- quantile(cooks.distance(lm_result),probs =
         upsilon) #0.7 #0.9
87    elapsed_process[(i-4),6] <- summary(lm_result)$adj.r.square
88  }
89
90  colnames(pred_error_list) <- c("updating_period","property_id
       ","pred_error","per_price","pred_error/per_price")
```

```
91 elapsed_process$elapsedtime <- as.Date(
      elapsed_process$elapsedtime, origin = "1970-01-01")
92
93 graph_data <- tapply(pred_error_list$pred_error,
      pred_error_list$updating_period,function(x)mean(abs(x)))
94 graph_data_mape <- tapply(pred_error_list$‘pred_error/
      per_price‘,pred_error_list$updating_period,function(x)mean
      (abs(x)))
95 graph_data_mer <- tapply(pred_error_list$‘pred_error/
      per_price‘,pred_error_list$updating_period,function(x)
      median(abs(x)))
96 graph_data_90 <- tapply(pred_error_list$‘pred_error/per_price
      ‘,pred_error_list$updating_period,function(x)quantile(abs(
      x),probs=0.9))
```

謝辞

本論文の作成にあたり，指導教官である一橋大学大学院横内大介准教授には研究への取り組み方，分析の切り口や論文のまとめ方など，様々な面でご指導を頂きました．また研究以外においても博士課程への入学，研究と仕事との両立，研究からビジネスアイディアへの発展など，幅広い内容にも親身に相談に乗っていただきました．深く感謝いたします．

同大学院の中川秀敏教授と中村信弘教授には学位論文審査および最終試験の審査をお引き受けくださり，ありがとうございます．心より御礼申し上げます．

同大学院の本多俊毅教授，宮川大介教授にはファカルティセミナーでの発表等を通じ，専門的な観点からご質問とご意見を頂きました．また大橋和彦教授，伊藤彰敏教授，野間幹晴教授，鈴木健嗣教授には講義やセミナーを通じて多くの示唆をいただきました．心より御礼申し上げます．

SRE ホールディングス株式会社には本論文で採り上げたテーマを進めるために貴重なデータを提供いただきました．心より感謝いたします．

参考文献

[1] 荒井俊行, 不動産取引価格およびその関連情報の公開・開示の促進について, 土地総合研究, 2013.

[2] 大槻健太郎, 横内大介, 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, 2020, 日本不動産学会誌, **34(3)**, p.101-108.

[3] 大槻健太郎, 横内大介, 東京２３区の中古マンション市場のデータ分割と統合 (印刷中), 日本不動産学会誌, 2023, **36(4)**.

[4] 金森敬文, 竹ノ内高志, 村田昇,『Rで学ぶデータサイエンス パターン認識』, 共立出版, 2009.

[5] 金本良嗣, ヘドニック・アプローチによる便益評価の理論的基礎, 土木学会論文集, 1992, **449**, p. 47-56.

[6] 唐渡広志, 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, 2016, **92**, p.17-20,

[7] 刈谷武昭, 小林裕樹, 清水千弘, 賃貸・分譲住宅の価格分析法の考え方と実際-ヘドニック・アプローチと市場ビンテージ分析-, プログレス, 2016.

[8] 久保拓弥, データ解析のための統計モデリング入門 一般化線形モデル・階層ベイズモデル・MCMC, 岩波書店, 2012.

[9] 熊坂夏彦，柴田里程, Textile Plot 環境, 統計数理, 2007, **55(1)**, p.47-68.

[10] 国土交通省, 不動産取引価格情報の公開について, 都市住宅学, 2009, **66**, https://www.jstage.jst.go.jp/article/uhs/2009/66/2009_50/_pdf/-char/ja (参照 2022-12-21).

[11] 国土交通省, 不動産価格指数（住宅）の作成方法, https://www.mlit.go.jp/common/001360416.pdf (参照 2022-12-21).

[12] 柴田里程, データリテラシー, 共立出版, 2001.

[13] 柴田里程, データ分析とデータサイエンス, 近代科学社, 2015.

[14] 白川慧一，大越利之, Real Estate Tech サービス提供の実態と地方圏における活用可能性に関する研究, 国土交通省, 平成 28 年度国土政策関係研究支援事業 最終報告書, https://www.mlit.go.jp/common/001259693.pdf, (参照 2022-12-21)

[15] 高橋将宣, 渡辺美智子, 欠測データ処理 -R による単一代入法と多重代入法-, 共立出版, 2017.

[16] 谷山智彦, "テクノロジーの活用による不動産市場の活性化", 不動産政策研究 各論 II 不動産経済分析, 東洋経済新報社, 2018, p.194-207.

[17] 谷山智彦, 本間純, 川口有一郎, 不動産市場における情報伝搬-価格先行指標としてのニュース記事とインターネット検索量-, ジャレフ・ジャーナル, 2014, **7**, p.1-16.

[18] 中村良平, ヘドニック・アプローチにおける実証分析の諸問題, 土木学会論文集, 1992, **449**, p,57-66.

[19] 野間久史, 連鎖方程式による多重代入法, 応用統計学, 2017, **46(2)**, p.67-86

[20] 野呂竜夫, 和田かず美, 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, 2015, **72**, p.41-54.

[21] 服部凌典, 岡本一志, 柴田淳司, 賃料予測モデルにおける間取り図の影響分析, 日本知能情報ファジイ学会誌, 2021, **33(2)**, p.640-650.

[22] 早川季歩, 田島夏与, 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, 2017, **99**, p.96-101.

[23] 林知己夫, データ解析の考え方, 科学基礎論研究, 1989, **19(2)**, p.81-87.

[24] 福中公輔, 橋本武彦, 橋本明広, 栗田一生, 三田匡能, 情報の非対称性の解消に向けた中古マンション価格推定の取り組み, デジタルプラクティス, 2020, **11(3)**, p.475-488.

[25] 増永良文, リレーショナルデータベースの基礎, オーム社, 1990.

[26] 三田匡能, 高価格物件を考慮した中古マンションの売買価格予測, 日本行動計量学会大会抄録集, **48**, 2020.

[27] 横内大介, 大槻健太郎, 青木義充, はっきりわかるデータサイエンスと機械学習, 近代科学社, 2020.

[28] 横内大介・柴田里程, インターデータベース − DandD インスタンスのエージェント化 − , 統計数理, 統計数理研究所, 2001, **49(2)**, p.317-331.

[29] Baig, S. u. R., Iqbal, W., Berral, J. L. and Carrera, D., Adaptive sliding windows for improved estimation of data center resource utilization, *Future Generation Computer Systems*, 2020, **104**, p.212-224.

[30] Bai, J. and Perron, P., Estimating and Testing Linear Models with Multiple Structural Change, *Econometrica*, 1998, **66(1)**, p.47-78.

[31] Bailey, M. J., Muth, R. F. and Nourse, H. O., A Regression Model for Real Estate Price Index Construction, *Journal of American Statistical Association*, 1963, **58**, p.933 − 942.

[32] Barnett, V. and Lewis, T., *Outliers in Statistical Data*, 3rd ed., John Wiley and Sons, Chichester, 1994.

[33] Belsley, D. A. Kuh, E. and Welsh, R. E., *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.*, John Wiley & Sons, New York, 1980.

[34] Bifet, A. and Gavalda, R., Learning from timechanging data with adaptive windowing, *In Proceedings of the 2007 SIAM international conference on Data Mining*, 2007, p.443—448.

[35] Bourassa, S. C., Hamelink, F., Hoesli, M. and Mac-Gregor, B. D., Defining housing submarkets., *Journal of Housing Economics*, 1999, **8**, p.160-183.

[36] Bourassa, S. C., Hoesli, M. and Peng, V. S., Do housing submarkets really matter?, *Journal of Housing Economics*, 2003, **12**, p.12-28.

[37] Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J., "LOF: identifying density-based local outliers", In *Proceedings of 2000 ACM SIGMOD International Conference on Management of Data*, 2000, ACM Press, p.93–104.

[38] Case, K. E. and R. J. Shiller, Prices of Single Family Homes since 1970: New Indexes for Four Cities, *New England Economic Review*, 1987. p.45 − 56.

[39] Case, K. E. and R. J. Shiller, The Efficiency of the Market for Single − Family Homes, *The American Economic review*, 1989, **79(1)**, p.125 − 137.

[40] Chambers, R.L., Outlier Robust Finite Population Estimation, *Journal of the American Statistical Association*, 1986, **81**, p.1063-1069.

[41] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R., *CRISP-DM 1.0 step-by-step data mining guide*, SPSS, 2000.

[42] Chiheba, F., Boumahdia, F. and Bouarfa, H., A New Model for Integrating Big Data into Phases of Decision-Making Process, *Procedia Computer Science*, 2019, **151**, p.636–642.

[43] Chiodo, A. J., Hernández-Murillo, R. and Owyang, M. T., Nonlinear Effects of School Quality on House Prices, *Federal Reserve Bank of St. Louis Review*, 2010, **92(3)**, p.185-204.

[44] Chow, G. C., Tests of Equality Between Sets of Coeeficients in Two Linear Regressions, *Econometrica*, 1960, **28**, p.591-605.

[45] Codd, E. F., A Relational Model of Data for Large Shared Data Banks, *Communications of the ACM*, 1970, **13(6)**, p.377-387.

[46] Cook, R. D., Detection of Influential Observations in Linear Regression, *Technometrics* (American Statistical Association), 1977, **19**, p.15-18.

[47] Dabreo, S., Rodrigues, S., Rodrigues, V. and Shah, P., Real Estate Price Prediction, *International Journal of Engineering Research and Technology*, 2021, **10(4)**, p.644-649.

[48] Dale-Johnson, D., An alternative approach to housing market segmentation using hedonic price data, *Journal of Urban Economics*, 1982, **11**, p.311-332.

[49] Dalmazo, B. L., Vilela, J. P. and Curado, M., Online traffic prediction in the cloud:a dynamic window approach, *INTERNATIONAL JOURNAL OF NETWORK MANAGEMENT*, 2016, **26(4)**, p.269-285.

[50] Deypir, M., Sadreddini, M. H. and Hashemi, S., Towards a variable size sliding window model for frequent itemset mining over data streams, *Computers & Industrial Engineering*, 2012, **63**, p.161–172.

[51] Dixon, W. J., Analysis of extreme values, *Annals of Mathematical Statistics*, 1950, **21(4)**, p.488-506.

[52] Dobson, A. J. *An Introduction to Generalized Linear models*, 2nd ed., CHAPMAN & HALL/CRC, London, 2001.

[53] Eurostat, Handbook on Residential Property Prices Indices (RPPIs), 2013, Available online at: https://ec.europa.eu/eurostat/documents/3859598/5925925/KS-RA-12-022-EN.PDF (accessed 21 December 2022).

[54] Fan, G. Z., Ong, S. E. and Koh, H. C., Determinants of House Price: A Decision Tree Approach, *Urban Studies*, 2006, **43(12)**, p.2301-2316.

[55] Fayyad, U., Shapiro, G. P. and Smyth, P., From Data Mining to Knowledge Discovery in Databases, *AI Magazine*, 1996, **17(3)**, p.37-54.

[56] Gao, G., Bao, Z., Cao, J., Qin, A. K., Sellis, R. and Wu, Z., Location-Centered House Price Prediction: A Multi-Task Learning Approach, *ACM Transactions on Intelligent Systems and Technology*, 2022, **13(2)**, p.1-25.

[57] Garcia, R. T. M., Lopez, M. F. C. and Sanchez, V. R. P., Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times, *Land*, 2022, **11(11)**, p.1-32.

[58] Goodman, A. C. and Thibodeau, T. G., Housing market segmentation, *Journal of Housing Economics*, 1998, **7**, p.121-143.

[59] Goodman, A. C. and Thibodeau, T. G., Housing market segmentation and hednic prediction accuracy, *Journal of Housing Economics*, 2003, **12**, p.181-201.

[60] Goodman, A. C., and Thibodeau, T. G., The Spatial Proximity of Metropolotan Area Housing Submarkets, *REAL ESTATE ECONOMICS*, 2007, **35(2)**, p.209-232.

[61] Grubbs F.E., Sample criteria for testing outlying observations, *Annals of Mathematical Statistics*, 1950, **21**, p.27-58.

[62] Hager, W. W., Updating the inverse of a matrix, *SIAM Review*, 1989, **31(2)**, p.221-239.

[63] Hannone, M., Predicting Urban Land Prices: A Comparison of Four Approaches, *International Journal of Strategic Property Management*, 2008, **12**, p.217-236.

[64] Hill, R. J., Scholz, M., Shimizu, C. and Steurer, M., An evaluation of the methods used by European countries to compute their official house price Indices, *Economie et Statistique / Economics and Statistics*, 2018, **500-501-502**, p.221-238.

[65] Hotelling, H., The generalization of Student's ratio, *Annals of Mathematical Statistics*, 1931, **2(3)**, p.360–378.

[66] Hullman, J. and Gelman, A., Designing for Interactive Exploratory Data Analysis Requires Theories of Graphical Inference, *Harvard Data Science Review*, 2021, **3(3)** Avairable online at: https://arxiv.org/pdf/2104.02015.pdf (accessed 21 December 2022).

[67] Inselberg, A., The Plane with Parallel Coordinates, *The Visual Computer*, 1985, **1**, p.69–91.

[68] Islam, K. S. and Asami, Y., Addressing structural instability in housing market segmentation of the used houses of Tokyo, Japan, *Procedia – Social and Behavioral Sciences*, 2011, **21**, p.33–42.

[69] Jeon, Y. and McCurdy, T. H., Time-Varying Window Length for Correlation Forecasts, *econometrics*, 2017, **5(4)**, p.1-29.

[70] Julious, S. A., Inference and estimation in a changepoint regression problem, *The statistician*, 2001, **50**, p.51-61.

[71] Kauko, T. Hakfoort, J. and Hooimeijer, P., Capturing housing market segmentation: an alternative approach based on neural network modeling, *Housing Studies*, 2002, **17**, p.875-894.

[72] Keskin, B. and Watkins, C., Defining spatial housing submarkets: Exploring the case for expert delineated boundaries, *Urban Studies*, 2017, **54(6)**, p.1446-1462.

[73] Kimber, A. C., Exploratory Data Analysis for Possibly Censored Data from Skewed Distributions, *Applied Statistics*, 1990, **39(1)**, p.21-30.

[74] Krause, A. and Lipscomb, C., The Data Preparation Process in Real Estate: Guidance and Review, *Journal of Real Estate Practice and Education*, 2016, **19(1)**, p.15-42.

[75] Kumasaka, N. and Shibata, R., High-dimensional data visualisation: The textile plot, *Computational Statistics and Data Analysis*, 2008, **52(7)**, p.3616-3644.

[76] Kutner, M., Nachtsheim, C., Neter, J. and Li, W., *Applied Linear Statistical Models*, 5th ed., McGraw - Hill/Irwin, New York, 2004.

[77] Lancaster, K. J., A New Apporoach to Consumer Theory, *Journal of Political Economy*, 1966, **74(2)**, p.132-157.

[78] Mahalanobis, P. C., "On the generalised distance in statistics", In *Proceedings of the National Institute of Sciences of India*, 1936, **2(1)**, p.49–55.

[79] Li, M., Bao, Z., Sellis, T., Yan, S. and Zhang, R., HomeSeeker: A visual analytics system of real estate data, *Journal of Visual Languages and Computing*, 2018, **45**, p.1-16.

[80] Li, T., Akiyama, T. and Tasaka, Y., "Constructing a highly accurate price prediction model in real estate investment using LightGBM", 2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR), 2021.

[81] Mason. H. and Wiggins, C., A Taxonomy of Data Science, Available online at: https://sites.google.com/a/isim.net.in/datascience_isim/taxonomy (accessed 21 December 2022).

[82] Ma, X., Hummer, D., Golden, J. J., Fox, P. A., Hazen, R. M., Morrison, S. M., Downs, R. T., Madhikarmi, B., L., Wamg C. and Meyer, M. B., Using Visual Exploratory Data Analysis to Facilitate Collaboration and Hypothesis Generation in Cross-Disciplinary Research, *SPRS International Journal of Geo-Information*, 2017, **6(11)**.

[83] Provost F, Fawcett T., Data science and its relationship to big data and data-driven decisionmaking, *Big Data*, 2013, **1(1)**, p.51-59.

[84] Rosen, S., Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition,*Journal of Political Economy*, 1974, **82**, p.34-55.

[85] Rubin, D. B., "Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonrepsonse",In *Proceedings of the Survey Research Merhods Section of rhe American Statistical Association*, 1978, p.20-34.

[86] Rubin, D. B., *Multiple Imputation for Nonresponse in Surveys*, Wiley: New York, 1987.

[87] Sibindi, R., Mwangi, R. W., Waititu, A. G., A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices, *Engineering Reports*, Available online at: https://onlinelibrary.wiley.com/doi/epdf/10.1002/eng2.12599 (accessed 28 December 2022).

[88] Schnare, A. B. and Struyk, R. J., Segmentation in Urban Housing Market, *Journal of Urban Economics*, 1976, **3(2)**, p.146–166.

[89] Schutt, R and O ' Neil, C., *Doing Data Science: Straight Talk from the Frontline*, O ' Reilly: New York, 2013.

[90] Shimizu, C., Karato, K. and Nishimura, G. N., Nonlinearity of housing price structure, *International Journal of Housing Markets and Analysis*, 2014, **7**, p.459-488.

[91] Shimizu, C. and Nishimura, K. G., Pricing structure in Tokyo metropolitan land marketsand its structural changes: pre-bubble, bubble, and post-bubble periods, *Journal of Real Estate Finance and Economics*, 2007, **35(4)**, p.475-496.

[92] Shimizu, C., Nishimura, K. G. and Watanabe, T., Housing Prices in Tokyo: A Comparison of Hedonic and Repeat Sales Measures, *Journal of Economics and Statistics* 2010, **230(6)**, p.792-813.

[93] Shimizu, C., Takatsuji, H., Ono, H. and Nishimura, K. G., Structural and temporal changes in the housing market and hedonic housing price indices: A case of the previously owned condominium market in the Tokyo metropolitan area, *International Journal of Housing Markets and Analysis*, 2010, **3(4)**, p.351-368.

[94] Suits D, B., Use of Dummy Variables in Regression Equations, *Journal of the American Statistical Association*, 1957, **52**, p.548-551.

[95] Tax, D., Duin, R., "Data domain description using support vectors", In *Verleysen, M. (eds.), Proceedings of the European Symposium Artifcial Neural Networks*, 1999. D. Facto, Brussel, p.251-256.

[96] Tukey, J. W., The Future of Data Analysis, *Annals of Mathematical Statistics*, 1962, **33**, p.1-67.

[97] Tukey, J. W., *Exploratory Data Analysis*, Addison-Wesley, 1977.

[98] Vapnik, V. N., *The Nature of Statistical Learning Theory*, Springer-Verlag, Berlin, 1995.

[99] Ye, V. Y. and Becker, C. M., The (literally) steepest slope: spatial, temporal, and elevation variance gradients in urban spatial modelling Get access Arrow, *Journal of Economic Geography*, 2018, **18(2)**, p.421–460.

[100] Yokouchi, D. and Shibata, R., DandD: Client server system, *Proceedings in COMPSTAT 2004*, Physica-Verlag, Prague, 2004.

[101] Yoshida, S., Hatano, K., Takimoto, E. and Takeda, M., Adaptive Online Prediction Using Weighted Windows, *IEICE Transactions*, 2011, **94-D(10)**,p.1917–1923.

[102] Xu, K. and Nguyen, H., Predicting housing prices and analyzing real estate markets in the Chicago suburbs using machine learning, 2022, Available online at: https://arxiv.org/abs/2210.06261