博士学位請求論文要旨

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

1. 本論文の構成

本論文の構成は以下のとおり。

1　研究の目的と貢献

本研究の目的は WEB サイト上で不動産仲介各社より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，売買取引件数の地域的な偏在によ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研究は少ない．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モデルのメンテナンスを終わりとするフローで捉えれば，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．

2　本研究の背景と各研究の位置付け

本研究の背景にあるのは日本でも 2015 年ごろから不動産仲介会社から提供が始まった WEB サイト上の不動産査定サービスである．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って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を活かしたプロップテック（PropTech）や不動産テック（Real Estate Tech）と呼ばれる新しいサービスが次々と生まれた．その中でも不動産査定サービスは現在，不動産仲介大手から不動産ベンチャーまで多くの会社に提供されているサービスである．

ただし，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．

本研究ではこうした実務上の課題に対してデータサイエンス・プロセス（たとえば Schutt, R and O’Neil, C [1]，横内・大槻・青木 [2]）の観点から捉え，第 2 章から第 4 章までの 3 つの研究において，それぞれデータ浄化，データ加工，意思決定という各ステップの課題として整理し，解決策を提案している．

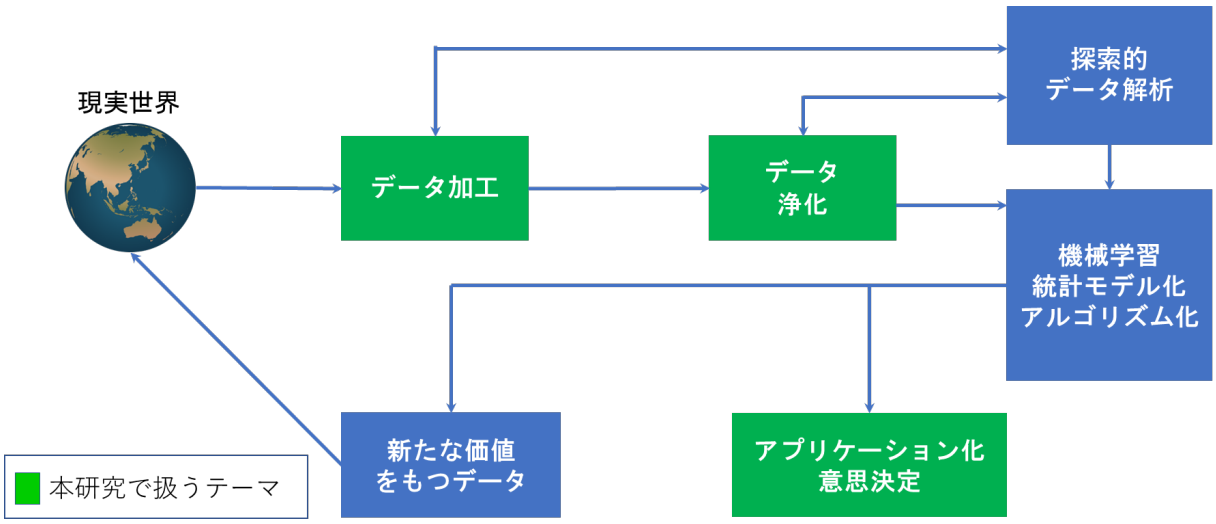

図 2.1　出典：O’Neil and Schutt [1] ならびに横内・大槻・青木 [2] を参考として筆者作成

3　データ浄化―エラーレコードの検出

第 2 章の研究では研究ではデータ浄化ステップのうち，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，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採り上げた研究は少なく，日本の不動産のデータに対する研究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されることもあるが，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，というケースがある．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点も考慮する場合がある．

本研究では取引価格だけでなく，地域性や中古マンションの物件属性を考慮に入れた桁間違いの判定を行うという中古マンション物件情報の意味的な要素に踏み込んだデータクレンジング方法を提案している．実際に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汎用性がある．

4　データ加工―データ分割と再統合

第３章の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予測対象となる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，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しても予測精度は良化することが分かっている．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．

この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している実例を示した．

本研究はデータサイエンス・プロセスに則せば柴田 [3] が提唱するように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，均質性の高いレコードを捉えてモデル化を行っているともいえ，探索的にデータを解析し，均質性の高いレコードをまとめるようなデータ加工を行っていると捉えられる．

5　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をどう行うか，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，モデルの更新と陳腐化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，ということは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直近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全て準備することが理想であるが，これらの条件を満たしたデータを揃え，時間が経過する中でも維持しつづけることは難しい．

実務においてこの目標に対処し，より良い予測を行うためには次に挙げるように大きく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み，また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，用いるモデル

によって解釈性が落ちたり，オーバーフィッティングする可能性もあり，またデータの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
   更新頻度が多いモデルであれば市況の変化に対するモデルの追従遅れを回避し，予測精度を向上することが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，安定した予測を行うためにはモデル推定には一定期間のサンプルサイズが必要となる．一方で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に反映できないため，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な方法として，中古マンション価格モデルを推定するために適切なデータの期間を設定することをテーマとする．

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スライディング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スライディング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，可変スライディングウィンドウによって更新されたモデルが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．

参考文献

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

[2] 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.

[3] 柴田里程, データリテラシー, 共立出版, 2001.